SUNDAY through THURSDAY / FEBRUARY 17, 18, 19, 20, and 21, 2019

2019 IEEE INTERNATIONAL SOLID-STATE CIRCUITS CONFERENCE

2019 DIGEST OF TECHNICAL PAPERS

INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS • UNIVERSITY OF PENNSYLVANIA

ISSCC®

IEEE

SSCS

VOLUME SIXTY-TWO
ISSN 0193-6530

IEEE SOLID-STATE CIRCUITS SOCIETY

2019 IEEE INTERNATIONAL SOLID-STATE CIRCUITS CONFERENCE

DIGEST OF TECHNICAL PAPERS

First Edition

February 2019

IEEE Catalog Number CFP19ISS-PRT

2019 IEEE International Solid-State Circuits Conference

DIGEST OF TECHNICAL PAPERS

Copyright and Reprint Permission:
Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For reprint or republication permission, email to IEEE Copyrights Manager at pubs-permissions@ieee.org. All rights reserved. Copyright ©2019 by IEEE.

PREPRESS IN THE UNITED STATES OF AMERICA
by S³ iPublishing, Inc.
Lisbon Falls, Maine

PRINTED IN THE UNITED STATES OF AMERICA
by Penmor Lithographers
Lewiston, Maine

VOLUME 62

IEEE Cat. No. CFP19ISS-PRT
ISBN Softbound 978-1-5386-8531-0
Library of Congress Number 81-644810
ISSN 0193-6530

Publisher and Managing Editor: Laura C. Fujino
Technical Editors: Jason H. Anderson, Leonid Belostotski, Dustin Dunwell, Vincent Gaudet, Glenn Gulak, James W. Haslett, David Halupka, Kenneth C. Smith

REFLECTIONS....4
The Saratoga Group – Three Decades of History....5
FOREWORD....8
AWARDS....27

PAPER SESSIONS

1 Plenary Session....10
2 Processors....40
3 Nyquist Rate ADCs....56
4 Power Amplifiers....72
5 Image Sensors....92
6 Ultra-High-Speed Wireline....110
7 Machine Learning....128
8 DC-DC Converters....144
9 High-Frequency Transceivers for Radar and Communications....160
10 Sensor Interfaces....178
11 Diagnostics....188
12 Emerging Technologies....200
13 Non-Volatile Memories....208
14 Machine Learning & Digital LDO Circuits....220
15 Power for 5G, Wireless Power, and GaN Converters....236
16 Frequency Synthesizers....254
17 Technologies for Human Interaction & Health....274
18 Analog Techniques....292
19 Adaptive Digital & Clocking Techniques....310
20 Noise-Shaped & VCO-Based ADCs....326
21 4G/5G Transceivers....342
22 Physiological Monitoring....358
23 DRAM....376
24 SRAM & Computation-in-Memory....386
25 Circuits Enabling Security....398
26 Frequency Generation & Interference Mitigation....406
27 Energy Harvesting & DC/DC Control Techniques....418
28 Techniques for Low-Power & High-Performance Wireless....436
29 Quantum & Photonics Technologies....454
30 Advanced Wireline Techniques....472

TUTORIALS

TUTORIALS 1-10....493

FORUMS

F1 Sub-6GHz 5G Radio Circuits and Systems: From Concepts to Silicon....496
F2 Memory-Centric Computing from IoT to Artificial Intelligence and Machine Learning....498
F3 The Complete Advanced Driver-Assistance System (ADAS) Sensing Network....501
F4 Intelligence at the Edge: How Can We Make Machine Learning More Energy Efficient?....504
F5 56Gb/s to 112Gb/s and Beyond – Design Challenges and Solutions in Wireline Communications....507
F6 The Right Tool for the Job: Application Optimized Data Converters....510

EVENING EVENTS

EE1 Student Research Preview: Short Presentations with Poster Session....512
EE2 How to Save Lives With Circuits....514
EE3 Making a Career Choice....516
EE4 Industry Showcase....517
EE5 Moving to 'The Dark Side'!....519
EE6 How Can Hardware Designers Reclaim the Spotlight?....521

SHORT COURSE

SC Integrated Phased Arrays: Theory, Practice, and Implementation for 5G and Beyond....524

INDEX TO AUTHORS....526
COMMITTEES....530
CONFERENCE LAYOUT....534
2020 CALL FOR PAPERS....535
CONFERENCE TIMETABLE....536

Reflections

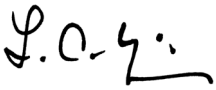
What you see before you this year, is the result of many many years of continuous iterative refinement of the submission process and information processing. Again, this year, we continue to provide a simplified printed Digest, in which the continuation pages (typically including a micrograph and occasionally summary data) are not included, but are available in the Digest download and in IEEE Xplore. In order to be more-green, we continue to use partially recycled paper.

Once more, this year, we have a technical editorial group (listed below) under the direction of a managing editor (Laura Chizuko Fujino). Again, we emphasize full technical and language editing of most papers, as the need dictates.

In recognition of the large amount of work leading to the Digest open before you, I wish to acknowledge a great many individuals: Eugenio Cantatore, Un-Ku Moon, Jan van der Spiegel, Anantha Chandrakasan, members of the ITPC, and all of the authors, for their individual contributions; Brad Phillips, and Mira Digital Publishing, for Web-based and other preparatory support, including continuing improvement and facilitation of the paper-review and pre-voting process in a new double-blind world, as well as preparation and implementation of the downloads available at the Conference; Steve Bonney, and S^3 iPublishing, for author and Session-Chair interaction, for figure layout, for paper formatting, for pre-press preparation, for printer interfacing, and for general assistance; Melissa Widerkehr, and Widerkehr and Associates, for general interfacing, problem-solving, and coordination; Jason Anderson (University of Toronto), Leo Belostotski (University of Calgary), Dustin Dunwell (Huawei Technologies), Vincent Gaudet (University of Waterloo), Glenn Gulak (University of Toronto), David Halupka (Intel), James Haslett (University of Calgary), and K. C. Smith (University of Toronto), as technical editors for heroic effort on the usual tight schedule.

My sincere thanks to you all!

P.S. This year marks the 30-year anniversary of the formation of the Saratoga Group, largely composed of graduate-student volunteers from the University of Toronto. The following few pages describes some of the history of this novel group.



Laura Chizuko Fujino
ISSCC Director of Publications and Presentations

February 2019

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

The Visuals Supplement and The Saratoga Group – Three Decades of History

Laura Chizuko Fujino
University of Toronto, Toronto, Canada

Preamble:
ISSCC 2019 marks the 30th year of the Saratoga Group, a group annually dedicated to the in-conference real-time creation of the conference record (named the Slide Supplement, and now the Visuals Supplement) that first appeared in a book version for ISSCC 1990.

The idea of providing attendees with all presentation material arose as a result of a variety of comments made in an attendee survey conducted for ISSCC 1989. The survey crystallized ideas which had been expressed for many years earlier concerning a heartfelt need to reconcile two problems that had long bothered the attendees: The first was dissatisfaction with those who snapped photos of the slides during presentations; the second was the often-heard lament that the audience was shown marvelous things that were not recorded or depicted in the Digest.

Since the data reduction of the survey was assigned to K.C. Smith, Laura Fujino did all of it, and was graciously allowed to present the findings at the August Executive Committee meeting in 1989. Amid a lot of discussion of various topics raised by the Survey, a decision was made by the Executive Committee that something needed to be done, and that a solution lay in some mechanism to provide copies of all the presentation slides to all attendees, after the Conference. Subsequent to the meeting in which Laura Fujino was present only for the Survey presentation, David Pricer as the Executive-Committee Chair was assigned to seek some mechanism for the resolution of this problem. Shortly thereafter, he contacted Laura Fujino with an invitation to join the Executive Committee with the role of addressing this challenge! Within a few weeks, a concept emerged on how to proceed: The existence of a photocopier with a 35mm slide projector attachment was identified from which medium-size paper copies of the slides could be made. As well, this photocopier had a paper to paper enlarging/reduction facility, which though of limited range could be used in multiple passes to acquire a paper image of the desired size.

A production process was conceived as follows: Soon after the completion of the talk, the author's slide carousel was used to produce a set of paper images from their 35mm slides; later these images were sized using the photocopier to provide paper images that were combined on an eight-per-page presentation, by a manual cut-and-paste process; this was combined with the title, author data, and abstract, in a nominal two-page format with extensions at the back of the compiled book, called the "*Slide Supplement to the Digest of Technical Papers*".

Correspondingly, during the 1990 Conference held at the San Francisco Hilton, a group of volunteers and one part-time casual employee met in the "Saratoga Room" to implement this process. The group included: John Eggert (Digest and Supplement printer) and his wife Shirely, John Wuorinen (Digest Editor) and his wife Susan, Nancy Pricer, wife of Dave, Henry Osborne (casual convention services employee), K.C. Smith, all under the direction of Laura Fujino. Thus, thereafter, this group was called the "Saratoga Group". Since the process was machine limited it took very long days to maintain a reasonable schedule with two operators (Henry and Laura). John Eggert led the manual final layout process, while Nancy Pricer collected the slides and sized paper copy, and the others were involved primarily with image sizing. (It was pleasing to note that no blood was spilled during the cut-and-paste process!) The final product, the ISSCC 1990 Slide Supplement (193 pages) was printed, bound, and mailed to all attendees, about one month after the Conference.

Later Developments:
The acceptance of the first edition of the Slide Supplement was sufficiently strong that we were encouraged to continue the logical development of the original concept. For the next two years, ISSCC 1991 and 1992, the process continued with some modification in the "Saratoga Room" at the San Francisco Hilton. But, for ISSCC 1993, operations were moved to the San Francisco Marriott, where the Conference has remained to this day.

For ISSCC 1991 and 1992, the major change was replacement of some of the initial volunteers and casual employee in the "Saratoga Group" by graduate-student volunteers from the University of Toronto.

For ISSCC 1994, following a change of the printer, Pat Duplessis joined the team in charge of manual page layout for which he introduced the use of a waxer for image adhesion. This process continued with one major change, until the introduction of electronic projection in 2001: The major change, introduced at ISSCC 1993, was to use author provided hard copy of their 35mm slides to simplify image production but requiring a slide-to-paper checking process, which in turn required more volunteer students, but was otherwise faster and did not rely as much on unreliable equipment. For ISSCC 1997, Steve Bonney joined Pat Duplessis for page layout. For ISSCC 1998, 1999, and 2000, we replaced much of the photo reduction by scanning the author's hard copy, and thereafter, sizing and page layout by computer.

For ISSCC 2000, an experiment was conducted with electronic projection (replacing 35 mm slides by computer-driven projectors) in three sessions, necessitating some additional image processing for Supplement production. For ISSCC 2001, electronic projection for day sessions became the norm. This included the use of multiple computers for back up and reliability, with the need for more student volunteers to operate them, and provide set up and checking. At the same time, the name "Slide Supplement" was changed to "Visuals Supplement", and the task of page layout was assigned solely to professionals.

The Graduate-Student Volunteer Group:
Today, the graduate-student volunteers, the major part of the current Saratoga Group is involved in a diverse set of tasks: Before the Conference they set up and check the laptops of which three are used for each regular session to provide projection, projection back-up, and audio recording, as required; organize Speaker and Committee Registration material; handle corresponding Speaker and Committee Registration; set-up equipment for Speaker Rehearsal; assist with Speaker Rehearsals; assist with Plenary Speaker Rehearsal; unpack and check Award plaques, organize and transport to the ballroom; man the Press desk; interact with unionized A/V staff; check slides with Speakers prior to their presentation; operate the computer projection system for Tutorials, Forums, Plenary Session, Regular Sessions, Evening Sessions, Short Course; assist speakers during the question and answer period; operate the recording system for the Plenary Session, Tutorials, and Short Course; act as videographers for various events; take photos of on-going events; complete data reduction for Tutorials, Forums, Short Course, Regular Sessions; help with crisis intervention.

Over the past 30 years, the number of graduate-student volunteers annually has ranged from 3 for ISSCC 1991 through a peak of 22 for ISSCC 2008 (to support the audio recording of the entire Conference), to 17 since ISSCC 2011.

An interesting vignette concerning the performance of the volunteer students occurred early on in the context of their handling of electronic-projection systems. Virtually at the moment that student-operated projection began, there were many reports from attendees concerning a magical process that they had observed

during the question and answer periods of sessions that they had attended: Their common comment was that miraculously, during the formation of the question by a questioner that a highly relevant slide illustrating the subject of the question and often its answer would appear suddenly on the screen. Their subsequent question to me was what marvelous artificial intelligence software had we acquired that allowed such instant insight and response! My retort was that it was not magically artificial, but real human, the consequence of care in selecting and assigning of the graduate-student volunteers: Each of them was highly-selected from amongst senior members of a large graduate-student body, typically in the PhD program, or at the end of a Masters degree from a range of circuit-related specializations; further they were assigned to sessions that were in their area of research interests. Thus, they were amongst the most astute listeners in the audience!

It is interesting to note that a large fraction of these Saratoga-Group graduate-student volunteers have gone on to very successful careers in the solid-state area, both in industry and academia. Increasingly, one has seen them presenting a paper at ISSCC! For example, this year, marking the 30th anniversary there are 6 papers with co-authorship by an alumnus of the Saratoga Group. Moreover, one is giving a Short Course, and as well, is receiving the ISSCC 2018 Lewis Winner Outstanding Paper Award. Another interesting phenomenon, the result of the three decade history, is the fact that many volunteers are second-generation members, their advisor having been an earlier volunteeer.

Conclusion:

Upon reflection of the past 30 years, one is surprised that such a process evolved through technological change, personnel change, and vagaries of attendance variation. Beyond the annual appearance of increasingly higher quality of slide material, including videos, animation, and so on, the process has influenced an enormous number of young lives, hundreds of individuals whose outlook on life is different because of their week or so of frenetic, yet focused activity at ISSCC which many identify as much akin to start-up dynamics!

On the facing page, you will find a list of members of the Saratoga Group over this 30-year period of hyper activity, with the years of their membership. This list is dominated by student-volunteer members, but includes others, some of whom are volunteers, while others are assigned professionals.

Reminiscences:

In recognition of the significance of the success of this 30-year adventure, some Saratoga Group alumni have been motivated to reveal some secrets of the past. All five were group leaders in their own time!

Reminiscences by Vincent Gaudet[1] (1997-2000 and 2003), Warren Gross[2] (1999-2003)

We both spent several years with the Saratoga Group, and were there at an important time of transition away from projection by caroussel and towards the new electronic age of Powerpoint. In the old days, Saratoga Group members spent most of their time behind closed doors, painstakingly verifying that the actual (physical) slides that were presented corresponded to what we had on file, allowing us to put together a true record of what was presented, for distribution after the Conference. This all changed around 2000, and we were honoured to be the two Group members who oversaw the very first Powerpoint presentation at ISSCC, for a groundbreaking microprocessor design. We manned the laptops at the back of the room, advancing slides at the request of the presenter, with the aid of a sound signal. During the question period, we quickly understood that a question pertained to a particular slide, and discreetly moved the presentation to that point, thus starting a new ISSCC tradition, and a far more visible role for the Saratoga Group!

Reminiscences by Kostas Pagiamtzis[3] (2001-2006)

Being a member of the Saratoga Group was a defining aspect of my graduate school studies. The deal was that in return for volunteering in the group, you got to visit San Francisco and attend ISSCC. What a deal! But, being part of the Saratoga Group was hard. Among other tasks, we spent long hours preparing and running the multiple computers during the sessions. Our goal when running these session computers was to make the presentations appear to run perfectly from the point of view of the audience, regardless of any behind-the-scenes glitches. Although any individual failure was a rare event, with about 200 presentations compressed in 5 parallel sessions over three days, multiple "rare" failures were common. Many problems were solved by having a backup laptop that could be swapped into production in the case of a crash or lock-up. We made special note of slow-loading slides, pre-loading them on the backup laptop and swapping the live and backup machines in real-time to avoid a slide delay or hiccup. All this was done, while paying close attention to the technical presentation, so that we were able to select and display the relevant slide during question period. The tasks were physically and mentally taxing, but Laura arranged for the most valuable reward possible to engineering graduate students ... abundant free food! So, the trade-off was long hours in return for being at the foremost circuits conference and getting free meals! It was a good deal!

I was part of the Saratoga Group in three roles: as a student volunteer, as the lead student volunteer and, after graduation, as a technical editor. Perhaps the most memorable role was being the lead student also known as "the king", a tongue-in-cheek title for the lead volunteer given that he actually has no power over the other students. After all, who actually has the power to influence the behavior of graduate students? Even academic advisors often fail to conjure this power but there is someone who has it! Laura Fujino. She is the dynamo powering the Saratoga Group. Laura, along with KC Smith, lead the Saratoga Group by example. You can't help but give your best effort in the shared goal when supporting with them. My strongest memories of the Saratoga Group are of relationships with fellow students and the relationship with Laura and KC. To this day I count many fellow student members of the Saratoga Group as friends and colleagues. I also count Laura and KC as friends and, for me, their friendship and guidance is the most lasting legacy of the Saratoga Group.

Reminiscences by Andrew Shorten[4] (2011-2017), Mario Milicevic[5] (2011-2017)

Leading the Saratoga volunteer team from 2012 to 2016 was one of the most rewarding experiences of graduate school. Three floors beneath the SF Marriott Marquis, our 16-hour days at ISSCC were fueled by high doses of adrenaline during Q&A slide changes, unexpected equipment failures, impromptu photo sessions, and on-the-fly plenary slide edits: “Survived again” quickly became the team’s mantra year after year as we swiftly improvised solutions to technical hiccups, leaving attendees none-the-wiser as to the effort required to pull off what appeared to be a flawless Conference.

But that was the easy stuff. The real adventure was in the months leading up to the Conference, when we had to assemble the “right team” to make ISSCC happen. With about one third of the team graduating each year, we had to constantly be on the lookout for new volunteers. We learned the hard way after our first year that we needed to interview prospective team members. We would look for incoming graduate students in the electronics and computer research groups who were broad-based, self-motivated and technically savvy. As volunteers, they would be rewarded with the opportunity to help run the flagship conference of the IEEE Solid-State Circuits Society. As group leaders, we did our part to make sure that everyone performed well to earn this opportunity.

By Day 2 at ISSCC, the team is usually exhausted, having enjoyed the live music a little too hard at the Foley’s piano bar the previous night. At around this point, newcomers are hit with the realization that ISSCC is a marathon, and that if they’re going to survive, they’ll need to make new friends. This is when the magic begins to unfold. Team members who back home sit only a few cubicles away from each other, finally start to talk! Dressed in full suits—a rather unusual sight for an engineering graduate student—they open up to each other and form amazing friendships.

While the adrenaline of slide changes and pointer-mouse replicator glitches keeps the team going at the Conference, it is the friendships and camaraderie that bring them back to ISSCC, year after year, even long after they have graduated. We are honored to have been given the opportunity to make this magic happen, and we look forward to reuniting with old friends at each year’s Conference.

[1]currently Professor & Chair, ECE, University of Waterloo, Waterloo, Canada
[2]currently Professor & Chair, ECE, McGill University, Montreal, Canada
[3]currently Principal Analog Engineer, Alphawave, Toronto, Canada
[4]currently Hardware Development Engineer, Apple, Cupertino, CA
[5]currently Staff Communication Systems Engineer, MaxLinear, Carlsbad, CA

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Three Decades of Participation

The Saratoga Group Membership from 1990 to 2019

2008	Mohammed Abdalla
2002	Sherif Abdalla
2009-2011	Karim Abdelhalim
2005	Mohamed Abdulla
2004-2008	Imran Ahmed
2019	Utku Alakusu
2007	Mehdi Alimadadi
2008, 2009	David Alldred
2013, 2014	Guy Alter
2014, 2015	Nadeesha Amarasinghe
1997	Jason Anderson
2006-2010	Ricardo Aroca
2003	Igor Arsovski
2003-2006	Navid Azizi
2015	Michael Baechle
2019	Mohammad Bagherbeik-Tabrizi
2016, 2017	Robert Baker
1995, 1996	Srinivasa Banna
2007-2011, 2015	Kevin Banovic
2007-2012	Mike Bichan
1996-1998	Jason Bickford
1994	Adi Bonen
1997-2019	Steve Bonney
1993	Bertil Brandin
2007-2019	Nancy Bush
2005-2009	Trevor Caldwell
1999, 2000	Anthony Carusone
2001	David Cassan
2010, 2011	Pearl Cao
2005-2007	Theodoros Chalvatzis
2003	Trevis Chandler
2012-2015, 2017	Jingxuan Chen
2017-2019	Paul Chen
1995	Peter Chen
2017	Yu Ting Chen
2008	Horace Cheng
1996, 2000	Ruth Cherian
2000	Sarah Cherian
2003	Dickson Cheung
1992, 1994	Raymond Chik
2014, 2015	Stephen Alexander Chin
2004	Jeffrey Chow
1992	Terry Choy
1994	Jeremy Cooperstock
2019	Jeremy Cosson-Martin
2002, 2004-2007	Ahmad Darabiha
1997	Sandy Decker
2005, 2006	Tod Dickson
1991	Ralph Duncan
2007-2013	Dustin Dunwell
1994-2000	Pat Duplessis
2017, 2018	Gairik Dutta
1990, 1991, 1993	John Eggert
1990, 1991	Shirley Eggert
2002, 2005	Yadollah Eslami
2001	Tooraj Esmailian
2002	Kamran Farzan
2009	William Feng
2001	Ted Fill
2009, 2010	Armin Fomani
2013, 2014	Ying Ying Fu
1990-2019	Laura Fujino
2013-2015	Michal Fulmyk
2006, 2007	Sean Garcia
2008, 2009	Ruslana Gelman
1997-2000, 2003	Vincent Gaudet
2016-2018	James Gentry
2004-2006	Dan Gerken
2004	Ahmed Gharbiya
2011-2013	Chris Gooch
1999-2003	Warren Gross
2018, 2019	Nikita Gusev
2005, 2007	Afshin Haftbaradaran
2004-2012	David Halupka
2002	Anas Hamoui
2008	Adam Hart
2018	Peter Hermansen
2013-2015	Robert Hesse
2007-2009	Derek Ho
1999, 2001	Steve Hranilovic
2018, 2019	Hsuan Hsiao
2010	Safeen Huda
2009, 2014	Alija Husic
2001-2006	Marcus van Ierssel
2004	Joel Ironstone
1998	Nadine Jackson
2009, 2011	Hamed Jafari
1991	Steve Jantzi
2019	Violet (Wanlin) Jiang
2014	Ge Jin
2008, 2009	Tony Kao
2004-2006	Rafal Karakiewicz
2015, 2016	Samira Karimelahi
2008	Mehdi Khanpour
2015	Alhassan Khedr
2015-2019	Jin Hee Kim
2012	Neno Kovacevic
1992	Karen Kozma
2012	Ivan Krotnev
2006	Fumie Kunimatsu
2006	Ian Kuon
2007	Mohamed Kwokgy
2018, 2019	Camilo Andres Tejeiro Lara
2005-2009	Katya Laskin
2001	Agustin Lebron
2003	Dora Lee
1992, 1993	Edward Lee
2012	Junmin Lee
2005-2017	Bertram Leesti
1998	Oliver Leung
1993	Kam-Wing Li
2018, 2019	Rophina Li
2013	Joshua Liang
1995	Tracy Liao
2013-2016, 2018	Sevil Zeynep Lüleç
1997-1999	Dave McCausland
2007-2009	Scott McLeod
2011-2017	Mario Milicevic
2001-2003	Shahriar Mirabbasi
2016	Vincent Mirian
2017, 2018	Danial Mohammadi
1999	Elizabeth Morelli
2019	Paula Miller
2019	Samantha Murray
2015	Rosanah Murugesu
2016-2018	Javid Musayev
2006-2008	Akram Nafee
1995	Cuong Nguyen
2006	Sean Nicholson
2016-2019	Gerard O'Leary
1998	Javad Omidi
1990	Henry Osborne
2001-2006	Kostas Pagiamtzis
2006, 2007	Samir Parikh
2018	Dhruv Patel
2009	Dimpesh Patel
2006, 2007	Jennifer Pham
2007-2019	Brad Philips
1997	Jason Podaima
2008	Gail Prestidge
1990-1996	Nancy Pricer
2015, 2016	Wahid Rahman
2001-2008	Jennifer Rodrigues
2000	Jonathan Rogers
2007, 2008	Bill Romer
2011	Alain Rousson
2009	Martin Rozee
2016-2019	Daniel Rozhko
1996	Ricardo Saad
2001, 2002	Saman Sadr
1992	Reza Safaee-Rad
1995-1997	Mazen Saghir
2004-2008	Steve Samson
2016	Navid Sarhangnejad
1995	Angela Schultz
1993-1995	Kenneth Schultz
2007	Mahdi Shabani
2007-2009	Shahriar Shahramian
2010	Shayan Shahramian
2012	Alireza Sharif-Bakhtiar
1996, 1998	Ali Sheikholeslami
2012	Ravi Shivnaraine
2012	Stefan Shopov
2011-2017	Andrew Shorten
2000	Jim Small
1990-2019	Kenneth Smith
2010	Vadim Smolyakov
2013-2015	Dawei Song
2009	William Song
2009	Soroush Tabatabaei
2008, 2010	Tina Tahmoureszadeh
2018, 2019	Ian David Taras
2019	Camilo Tejeiro
2011-2013	Clifford Ting
2012	Anita Tino
2009	Alex Tomkins
2005-2007	Olivier Trescases
1991	Patty Trnka
2012, 2013	Colin Tse
2005-2012	Aleksey Tyshchenko
2013, 2014	Jasmina Vaslijevic
2012-2015	Aynaz Vatankhahghadim
2012	Jeffrey Wang
2010, 2011	Jing Wang
2012	Luke Wang
2004	Robert Wang
2007	Guowen Wei
2011, 2013-2015	Yue (Victor) Wen
2003	Joyce Wong
2001	Laurie Wood
2002, 2003	Sarah Wood
1991	Steve Wood
1990	John Wuorinen
1990	Susan Wuorinen
2004, 2008	Navid Yaghini
2007-2011	Kentaro Yamamoto
2010	Hemesh Yasotharan
2007, 2008	Kenneth Yau
2015-2017	Yue Yin
2016-2018	Jingshu Yu
2004	Ricky Yuen
2014	Mohammad Meysam Zargham
2017-2019	Saba Zargham
2010, 2011	Guangzhao (Andy) Zhang
2015, 2016	WeiJia Zhang
2019	Yi Fan (Danny) Zhang

Foreword

Envisioning the Future

Eugenio Cantatore
Eindhoven University of Technology
Eindhoven, The Netherlands
ISSCC International Technical Program Chair

I wish you a warm welcome to the 2019 International Solid-State Circuits Conference! This year the Conference continues its tradition of showcasing excellent highly-advanced work in integrated circuits and systems from industry, universities, and research institutions around the world. Of the accepted technical papers, 49% are from academe, 27% are from industry, 18% are joint submissions from industry with academe or research institutions, and 6% are from research institutions. Continuing the practice introduced in 2017, the ISSCC paper selection followed a double-blind review process with anonymized manuscripts.

The 2019 Conference theme is "Envisioning the Future": Relentless innovation has always been the driving force in integrated circuits and systems. This year's theme reflects the need to continuously monitor, investigate, and nurture new ideas, to ensure a sustainable future for our industry and its research drivers. These innovations can include disruptive technologies, such as artificial intelligence (AI), machine learning (ML), quantum computing, and the efficient integration of photonics with electronics. Alternatively innovations may take a more evolutionary path toward higher performance, such as 5G (in wireless) and 112Gb/s (in wireline). This year, ISSCC, the flagship conference in solid-state circuits, will focus especially on all these innovations, as you will recognize in the Plenary Session, in the Technical Sessions, in the Educational Events (both Tutorials and Forums), in the Demonstration Session, and in the Industry Showcase.

The Conference officially begins with the Plenary Session, where four innovators will share their vision on diverse challenges and opportunities in integrated circuits and systems. The first talk by Yann LeCun (Facebook and New York University) will deal with the revolution in computer comprehension of images, audio, and text, which has been enabled by deep learning (DL). He will discuss the imperative need to abandon the currently common paradigms of supervised and reinforcement learning, to emphasize self-supervised learning in AI. The following talk by Hoi-Jun Yoo (KAIST) will focus on AI in mobile devices and the need to achieve greatly improved power efficiency in such applications. Special attention will be given to the challenge of enabling learning on mobile platforms, and to the role that mimicking the brain can have to make possible ever-improving AI implementations in resource-limited systems. The third talk by Meint Smit (Eindhoven University of Technology) shifts gears, focusing on the integration of photonics with electronics, one of the most promising directions for innovation in technology. Integration of photonics with electronics will have a far-reaching impact, enabling unprecedented communication speeds, together with high energy efficiency, and novel sensor applications. The final plenary by Vida Ilderem (Intel), will discuss the groundbreaking architectures and circuit innovations required to make 5G a reality, and how these innovations will enable data-rich services and sophisticated cloud applications, delivered faster and with lower latency.

The ISSCC 2019 Technical Program consists of 197 outstanding technical papers including 4 system-focused invited talks, distributed over 30 thematic sessions. In addition to these regular sessions, the Conference continues to promote educational activities in other ways: a series of ten 90-minute introductory tutorials, as well as six full-day expert-oriented forums, and a short course.

This year, the 4 invited system-focused talks emphasize systems applications linked to silicon innovation: Jim Kahle (IBM Research) will discuss architectures and technologies of their AI supercomputers, Summit and Sierra (Monday afternoon); Chih-Ming Hung (MediaTek) will present the technological challenges in developing automotive surround-view radar systems (Tuesday morning); Masahiro Fujita (Sony) will discuss the synergies between AI and robotics (Tuesday afternoon); and Gerrit den Besten (NXP Semiconductors) will describe full-duplex communication interfaces over single twisted pair from 10Mb/s to 10Gb/s for automotive applications (Wednesday afternoon).

On Sunday, February 17, ten Tutorials will cover a wide range of topics: "Fundamentals of Integrated Radars", "Fundamentals of Power-Conversion Topologies", "Advances and Prospects for In-Memory Computing", "Fundamentals of Efficient IoT Microcontrollers", "Noise-Shaping in Data Converters", "Basics of Clock-and-Data-Recovery Circuits", "Hardware Security – from Basics to ASICs", "Current-Sensing Techniques", "Calibration Techniques for Wireless Transceivers", and "Low-Noise Sensor Interfaces". These Tutorials are given by top experts in their field, and offer an opportunity to experience an introduction to each topic area and an overview of the recent primary developments.

The goal of our 6 all-day Forums is to update experts on the most-recent advanced developments in their fields, thanks to contributions by renown technical leaders: On Sunday, February 17th, two forums will be held in parallel: "Sub-6GHz 5G Radio Circuits and Systems: From Concepts to Silicon" and "Memory-Centric Computing from IoT to Artificial Intelligence and Machine Learning".

On Thursday, February 21st, four forums will be offered in parallel: "The Complete Advanced Driver-Assistance System (ADAS) Sensing Network", "Intelligence at the Edge: How Can We Make Machine Learning More Energy Efficient?", "56Gb/s to 112Gb/s and Beyond – Design Challenges and Solutions in Wireline Communications", and "The Right Tool for the Job: Application-Optimized Data Converters". Also on Thursday is an all-day Short Course on "Integrated Phased Arrays: Theory, Practice, and Implementation for 5G and Beyond", which will present a broad overview of phased arrays, including enabling circuits, up-scaling trends, and future directions.

As usual, the Conference will feature a variety of evening events that combine education on timely topics and entertainment. On Sunday evening, a workshop co-sponsored by the IEEE Solid State Circuits Society (SSCS) on "How to Save Lives with Circuits" highlights the impact of circuits on healthcare-related industry, encompassing system architecture, security, and circuit design aspects. The workshop will feature keynote presentations and an interactive panel. Also on Sunday evening, the "Student Research Preview" will present selected graduate-student research projects from around the globe, and will include an introductory talk, "Machine Learning: The Next Big Opportunity for Chip Designers", by Marian Verhelst, KU Leuven.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

On Monday, the evening program will feature the “Industry Showcase”, which highlights how advances in silicon circuits, SoCs, and systems, are fueling exciting innovations in industry. The event will begin with a keynote talk, “Accelerated Platforms: The Future of Computing”, by Bill J. Dally (NVIDIA and Stanford University), followed by interactive demonstrations in the Atrium level. The Showcase participants were chosen through an open-submission voting process by members of the Industry Showcase Committee.

Just before the Industry Showcase, between 7 and 8 p.m., a new ISSCC initiative, a panel “Making a Career Choice”, is available for students who are wondering what career path to take. The event will include distinguished panelists representing various career options, including academia, research, industry, and start-ups.”

On Tuesday evening, the program presents two entertaining and inspiring panel discussions: “Moving to The Dark Side” will provide a forum to debate the relationship between business and management on one side (the “dark” one) and engineering on the other. An interesting discussion on technical innovation, financial constraints, career paths, and business, will animate this panel. The second panel, “How Can Hardware Designers Reclaim the Spotlight?”, will discuss the fact that circuit components are becoming increasingly commoditized, seemingly losing relevance in the larger scheme of things. The panel will explore this trend and evaluate new opportunities where hardware innovation can play a central role.

On Monday and Tuesday evenings, two Demonstration Sessions will be held during the social hours. These sessions provide an opportunity to observe live demos of devices described in nearly 50 selected papers from the technical program, and to engage in face-to-face discussions with the authors.

The high standards that we associate with ISSCC are a result of the dedicated volunteer work of the International Technical-Program Committee (ITPC). This year, the ITPC consists of 158 members from industry and academe, organized into eleven technical subcommittees. Each member has spent a significant amount of time soliciting excellent submissions, reviewing the submitted papers, organizing evening sessions and educational events, preparing the Advance Program, Press-Kit, and Digest material, and performing session-chair duties. I am deeply thankful to all ITPC members, and in particular to the Subcommittee Chairs for their leadership in overseeing these tasks: Kofi Makinwa (Analog), Michael Flynn (Data Converters), Byeong-gyu Nam and Tom Burd (Digital Architectures & Systems), Edith Beigné (Digital Circuits), Chris Van Hoof (Imagers/MEMS/Medical/Displays), Jonathan Chang (Memory), Yogesh Ramadass (Power Management), Piet Wambacq (RF), Makoto Nagata (Technology Directions), Stefano Pellerano (Wireless), and Frank O'Mahony (Wireline). Also, I sincerely thank the leadership of the Regional Committees: Kostas Doris and Yiannos Manoli from the European Region, and Tai-Cheng Lee and Makoto Takamiya from the Far-East Region. Additionally, I have enormously benefited from the help of the ISSCC 2019 Program Vice-Chair, Un-Ku Moon, and the ISSCC 2018 Program Chair, Alison Burdett.

ISSCC would not be possible without the help of many other individuals that deliver excellent support work behind the scenes. It is my pleasure to acknowledge Melissa Widerkehr and Widerkehr and Associates for their precious support with Conference operations and arrangements. I am grateful to Brad Phillips and MIRA Digital Publishing for their assistance with the electronic manuscript submission, pre-voting, and assembly of the Advance Program, as well as the Digest, and to Steve Bonney and S^3 iPublishing for page layout and facilities coordination. As well, I am deeply thankful to the Technical Editors: Jason Anderson, Leo Belostotski, Dustin Dunwell, Vincent Gaudet, Glenn Gulak, James Haslett, and Dave Halupka both as an Editor and multi-media-coordinator. Very special thanks go to Laura Fujino and Kenneth Smith for their truly invaluable help with all kind of aspects of the Conference, including awards, paper submission process, preparation of the Advance Program, Press Kit, Digest,Tutorial and Short Course USB , as well as diverse editing and presentation preparation. Many thanks to Denis Daly for his leading role in the Press Kit preparation and for organizing the Student-Research Preview, to Ali Sheikholeslami for his coordination of the Tutorials and Short Course, to Dan Friedman for organizing the Short Course, to Andreia Cathelin for leading the Forums, to Uming Ko for organizing the Demonstration Sessions, to Trudy Stetzler for managing the Conference website and A/V coordination, and to Bryant Griffin and John Weinmann for their financial oversight for the Conference.

Furthermore, I gratefully acknowledge the exceptional help provided by the volunteer graduate students from the University of Toronto, who, thanks to their technical expertise, ensure the orderly conduct of the presentations in each session, and support countless other behind-the-scene activities.

A very special acknowledgement goes to Anantha Chandrakasan, the ISSCC Senior Technical Advisor, for his invaluable support in inviting excellent speakers, for inspiring me during the many years serving on the Technical Program Committee, and for advising in countless other aspects of the Conference. Finally, my deep appreciation and gratitude go to Jan van der Spiegel, the ISSCC Conference Chair, for his kind and inspired leadership. I am confident he will maintain, and strengthen even more, the tradition of excellence provided to Conference attendees.

I really look forward to seeing you at the Conference and wish you an inspiring ISSCC 2019!

Eugenio Cantatore
ISSCC 2019 International Technical-Program Chair

Session 1 Overview

Plenary Session — Invited Papers

Chair: ***Jan van der Spiegel***
University of Pennsylvania, Philadelphia, PA
ISSCC Conference Chair

Associate Chair: ***Eugenio Cantatore***
Eindhoven University of Technology, Eindhoven, The Netherlands
ISSCC International Technical Program Chair

The Plenary Session will begin with welcome remarks, and an introduction from the Conference Chair, Jan van der Spiegel. Then, Eugenio Cantatore, the Technical Program Chair, will provide an overview of ISSCC 2019. Next, the Plenary Session will feature four keynote speakers, who are renowned leaders in their respective fields. In line with the Conference theme, ***"Envisioning the Future",*** they will overview important innovation trends that are predicted to have a strong impact on the further development of Integrated circuits and systems: Artificial Intelligence (AI) and Machine Learning (ML), integration of photonics and electronics, and 5G. An Awards Ceremony that recognizes major technical and professional accomplishments, presented by the IEEE, Solid-State-Circuits Society (SSCS), and ISSCC, will take place following the second plenary talk.

The series of four plenaries starts with a presentation on Deep Learning:
In this first talk, Yann LeCun, VP & Chief AI Scientist at Facebook and Silver Professor of Computer Science, New York University, analyzes the present state of Deep Learning (DL) from an historical perspective. In his analysis, he identifies important future trends: from more self-supervised learning, to dynamic networks whose architecture changes for each new sample; from number representations optimized for inference on low-power hardware to very large networks with very sparse activations; from new operators (such as fast K-nearest neighbors) to networks that manipulate annotated graphs instead of tensors.

The evolution of AI is bringing it from cloud computing to the edge and to mobile devices. The second talk by Hoi-Jun Yoo, Professor of Electrical Engineering at KAIST and the Director of the SDIA Research Center, focuses on this trend. The most important requirements for edge and mobile Deep Neural Networks (DNNs) are low power consumption, reduced precision, and limited memory size. To optimize the energy efficiency of DNN implementations, a holistic approach must be taken: It must simultaneously consider neural network architectures, hardware architectures, and even circuits. Several other aspects of AI in resource-limited devices (such as local learning and brain-mimicking functions) are addressed.

The next two plenary talks focus on the future of communication technologies:
The integration of photonic functions in photonic chips, and the creation of new systems that put together integrated electronics and photonics is one of the most compelling emerging technologies today. Meint Smit, Professor of Photonic Integration at the Eindhoven University of Technology, provides an overview of the state-of-the-art in InP- and Si-based photonics, reviewing similarities and differences, and discussing in detail membrane-based InP photonic technologies supporting efficient integration with electronics.

5G is the first wireless technology that enables the variety and volume of data that is emerging with the Internet of Things, optimizing the network both for human-human and for machine-machine communication. Vida Ilderem, VP and Director of Wireless Communication Research (WCR) at Intel Labs, discusses in her talk why 5G represents an inflection point, where wireless communication technology is driven by applications and expected use cases. She reviews what is 5G and what are its key requirements, and highlights the disruptive architectures and technology innovations required to make 5G and beyond a reality.

With this line-up of speakers, this plenary session is intended to assist you in evolving your personal vision of the future. We hope that you will enjoy these inspiring and motivating presentations.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

1.1 Deep Learning Hardware: Past, Present, and Future — 8:45 AM

Yann LeCun, *Facebook AI Research & New York University, New York, NY*

Deep learning has caused revolutions in computer understanding of images, audio, and text, enabling new applications such as information search and filtering, autonomous driving, radiology screening, real-time language translation, and virtual assistants. But almost all these successes largely use supervised learning, which requires human-annotated data, or reinforcement learning, which requires too many trials to be practical in most real-world situations. In contrast, animals and humans seem to learn vast amounts of background knowledge about the world through mere observation and occasional actions in a self-supervised manner. Making progress in self-supervised learning is the main challenge of AI for the next decade. Success may result in machines with some level of common sense. But they will be built around deep learning architectures that are considerably larger than current ones, requiring vastly more powerful hardware than what we have today.

1.2 Intelligence on Silicon: From Deep Neural Network Accelerators to Brain-Mimicking AI-SoCs — 9:20 AM

Hoi-Jun Yoo, *KAIST, Daejeon, Korea*

Deep learning is influencing not only the technology itself but also our everyday lives. Formerly, most AI functionalities and applications were centralized on datacenters. However, the primary platform for AI has recently shifted to mobile devices. With the increasing demand on mobile AI, conventional hardware solutions face their ordeal because of their low energy efficiency on such power hungry applications. For the past few years, dedicated DNN accelerators inference have been under the spotlight. However, with the rising emphasis on privacy and personalization, ability to learn on mobile platform is becoming the second hurdle for "on-device AI." Going hand in hand, the brain mimicking is also a highlighted field in AI. Fundamentals of neuromorphic architecture start from the synapses and the neurons, which are realized with custom memories including non-volatile memories. The brain mimicking AI is not restricted to the naïve neuron implementation, but it extends further to mimic the behavioral characteristics such as "connectome" or "visual attention." In conclusion, mobile learning and brain mimicking will be the two horses driving the carriage called AI, thus opening up new requirements for the next generation deep learning SoCs.

ISSCC, SSCS, IEEE AWARD PRESENTATIONS — 9:55 AM

BREAK — 10:20 AM

1.3 Integration of Photonics and Electronics — 10:40 AM

Meint K. Smit, *Eindhoven University of Technology, Eindhoven, The Netherlands*

The application market for Photonic Integrated Circuits (PICs) is rapidly growing. Photonic Integration is the dominant technology in high bandwidth communications and is set to become dominant in many fields of photonics, just like microelectronics in the field of electronics. Photonic integrated circuits offer compelling performance advances in terms of precision, bandwidth and energy efficiency. To enable uptake in new sectors, the availability of highly standardized (generic) photonic integration platform technologies is of key importance as this separates design from technology, reducing barriers for new entrants. The major platform technologies today are InP-based monolithic integration and Silicon Photonics. In this perspective paper we will describe the current status and future developments of InP-based generic integration platforms.

1.4 5G Wireless Communication: An Inflection Point — 11:15 AM

Vida Ilderem, *Intel, Hillsboro, OR*

The 5G era is upon us, ushering in new opportunities for technology innovation across the computing and connectivity landscape. 5G presents an inflection point where wireless communication technology is driven by application and expected use cases, and where the network will set the stage for data-rich services and sophisticated cloud apps, delivered faster and with lower latency. This paper will highlight the disruptive architectures and technology innovations required to make 5G and beyond a reality.

PRESENTATION TO PLENARY SPEAKERS — 11:50 AM

CONCLUSION — 11:55 AM

1.1 Deep Learning Hardware: Past, Present, and Future

Yann LeCun, Facebook AI Research and New York University, New York, NY

Abstract
Historically, progress in neural networks and deep learning research has been greatly influenced by the available hardware and software tools. This paper identifies trends in deep learning research that will influence hardware architectures and software platforms of the future.

1 The Past

Modern AI is powered by deep learning (DL), whose origins go back to early experiments with electronic neural nets in the 1950s. DL is based on four simple ideas: (1) complex functions can be efficiently constructed by assembling simple parameterized functional blocks (such as linear operators and point-wise non-linearities) into multi-layer computational graphs; (2) The desired function can be learned from examples by adjusting the parameters; (3) the learning procedure minimizes an objective function through a gradient-based method; (4) the gradient can be computed efficiently and automatically through the back-propagation algorithm (backprop, for short), which is nothing more than a practical application of chain rule to compute the partial derivatives of the objective with respect to all the parameters in the system by propagating signals backwards through the network. The key advantage of DL is that it alleviates the need to hand-design a feature extractor, as would be required with traditional machine learning and pattern-recognition methods. When trained for a particular task, DL systems automatically learn multi-layer hierarchical representations of the data that are suitable for the task. One may wonder why such a simple idea as backprop was not popularized until the late 1980s, and why such a natural idea as DL did not become widely used until the early 2010s, fueling the recent wave of interest in AI.

1.1 The Unreasonable Influence of Hardware and Software Tools on Progress

Several lessons can be drawn from the 60+ year history of neural networks: (1) new ideas seem limited by the available hardware (and software); (2) specialized hardware does not enable new methods; (3) biological inspiration is fruitful but can be a trap. From the 1950s to the 1980s, neural-network models used binary neurons: the McCullloch-Pitts neuron which computes a weighted sum of its inputs and passes it through a sign function. This was largely because multiplications were prohibitively expensive at the time, whether implemented in analog electronics, in digital circuits, or in software. With binary inputs, additions are sufficient to compute the neurons'weighted sums. One reason backprop did not emerge until the late 1980s is that it required the use of neurons with continuous non-linearities (such as sigmoids), which did not become practical until workstation performance approached one million floating-point multiply-accumulate operations per second.

More puzzling, while the basic techniques of DL have been around since the late 1980s, what caused it to lose popularity within the research community in the mid-1990s? (1) the performance of computers at the time; (2) the small number of applications for which collecting large labeled datasets was cost effective; (3) the effort involved in developing flexible neural net simulators; (4) the reluctance of many research institutions at the time to distribute open source software. What sparked its sudden resurgence around 2013? There are four main factors: (1) improved methods; (2) larger datasets with many samples and many categories; (3) Low-cost TFLOPS-class general-purpose GPUs (GPGPUs); (4) open-source libraries with interpreted language front-ends (Torch, Theano, cuda-convNet, Caffe). The first three of these enabled record-breaking results in image recognition and speech recognition, while the last one allowed these results to be easily replicated because they incorporated all the engineering "tricks" necessary to get DL models to work. Arguably, prior to this, computer vision and ML research was limited to what was implementable easily in Matlab.

Interestingly, there were attempts to build dedicated hardware architectures for neural networks before the advent of GPGPUs. But none of them were successful, in part because they lacked flexibility [1, 2] and/or were designed for particular types of neural networks that had no proven practical use [3, 4, and references therein]. What has accounted for the success of GPGPU for DL is their wide availability, generality, programmability, and well-supported software stacks.

The important lesson is that *Hardware capabilities and software tools both motivate and limit the type of ideas that AI researchers will imagine and will allow themselves to pursue. The tools at our disposal fashion our thoughts more than we care to admit.*

In other words, what the AI hardware and software communities will produce over the next few years will shape AI research for the coming decades.

1.2 Neural-Network Hardware of Yesteryear

The history of neural networks is inextricably linked with hardware. The original 1957 Perceptron was a dedicated analog computer whose weights were implemented with motorized potentiometers [5], and the rival model, Adaline, was implemented with electrochemical "memistors" [6]. Early adaptive equalizers based on the Adaline algorithms used relays to represent weights [7]. Interest in neural networks had waned following the publication of Minsky and Papert's book"Perceptrons"in 1969 [8], and for two decades, hardware development for "trainable" systems was confined to adaptive-filter applications.

Then, in the mid 1980s, A second wave of interest in neural networks took off, following work on Hopfield networks, Boltzmann Machines, and the popularization of backprop [9, 10, 11]. In 1985, the Adaptive Systems Research Department was created at Bell Laboratories in Holmdel, NJ, under the leadership of Lawrence D. Jackel. This group, which I joined in 1988, spent a decade developing a series of neural-network chips. The evolution of their technology gives us an idea of the constraints brought about by hardware considerations. For example, an analog vector-matrix multiplication can simply be implemented as a resistor array, with separate rows for positive and negative coefficients. In 1986, the group build a 6×6 micron, 12×12 resistor array using e-beam lithography (see Figure 1.1.1). But it quickly became clear that non-programmability was a major limitation and that line amplifiers, I/O circuitry, and signal conversion would defeat the purpose of manufacturing very small resistors [1]. Next, the group built a 54-neuron mixed analog-digital chip. Each neuron had 54 ternary weights (–1, 0, +1) with analog summing. The chip could perform simple feature extraction on binary images, but its speed was limited by I/O bandwidth [2].

When I joined in late 1988, I developed the first convolutional neural net (ConvNet or CNN for short) and obtained excellent results on handwritten character recognition tasks (zip codes) [12, 13]. The ConvNet architecture was loosely inspired by that of the ventral pathway in the visual cortex and was designed to process data that comes to us in the form of an array (possibly, multidimensional) in which nearby values are correlated: image, video, audio, text, and so on. ConvNets are composed of two main types of stages: convolution stages and pooling stages (see Figure 1.1.2). In a ConvNet designed for image recognition, the input, output, and intermediate layer activations are all 3-dimensional arrays (often called tensors). Each "slice" of the input tensor is a color channel. Each slice of the first layer output (called a feature map) is obtained by performing a discrete convolution of each of the input slices with different convolution kernel (also called filters). The results are added and passed through a half-wave rectification non-linearity (also called a ReLU for Rectified Linear Unit). The coefficient of the convolution kernels, whose size if typically 5×5 or 3×3 are subject to learning. A convolutional filter followed by a ReLU detects a particular motif regardless of its location on the input and produces outputs that are equivariant to shifts (that is when the input shifts, the output shifts accordingly). The subsequent pooling layer reduces the spatial resolution of the feature maps by aggregating values within a neighborhood using a max or L_p norm operation. Pooling windows are stepped by more than one pixel, resulting in a lower-resolution feature map. Pooling makes the representation robust to small shifts in the location of distinctive features.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Multiple alternated stages of convolutions and pooling with increasing numbers of feature maps and decreasing spatial resolutions are stacked, producing outputs that are influenced by a large part of the input image. ConvNet are often trained in a supervised manner using a form of stochastic gradient descent (SGD). An image is shown to the network and the output is computed. The output is compared to the desired output (representing a label for the image) using a loss function. The gradient of this loss with respect to all the filter coefficients in the network is computed through the *back-propagation procedure* which propagates gradient signals *backwards* through a version of the network in which all arrows have been reversed and in which each operator multiplies these signals by its Jacobian matrix. The essential characteristic of ConvNets is their ability to learn hierarchical representations of the signal automatically.

Because of the prevalence of convolutions in image analysis, the next generation of Bell Labs chips, called Net32k, contained 256 neurons, each with 128 ternary weights, capable of 320×10^9 synaptic operations per second [14]. Multiple rows could be combined with power-of-two weights to form 64 neurons with 128 4-bit weights. An important new feature was the presence of shift registers to allow fast convolution with multiple filters, while minimizing access to external memory. Almost simultaneously, another chip called ANNA (Analog Neural-Network Accelerator) was built specifically to run convolutional networks [15]. It contained 64 neurons with 64 weights each. The weights were stored on capacitors and refreshed from an external RAM through a DAC with 6-bit accuracy. The activations were represented digitally by 3 bits. The synapses were essentially multiplying DACs, and the sum was done in analog. The chip ran at 20MHz and was capable of 4×10^9 synaptic operations per second. Just as in Net32k, ANNA had shift registers that allows it to run realistic ConvNet architectures with 130,000 connections in about 1ms, or about 500 times faster than the best workstations at the time, with comparable accuracy on zip code digit recognition [16]. Both Net32k and ANNA came close to being used in commercially deployed applications (postal envelopes and bank-check reading), but were eventually dropped for software implementations on floating-point DSPs. The lessons one can draw from this history is that: (1) bringing exotic fabrication technologies and architectural concepts to market is difficult; (2) system-level performance is considerably more important than raw speed. It took only two years after ANNA for FPGAs to become sufficiently powerful to run ConvNets at the same speed using non-conventional number representations [17].

By the mid-1990s, ConvNet applications had been widely deployed by AT&T and its subsidiaries for reading documents, and by the late 1990s, somewhere between 10% and 20% of all bank checks in the US were automatically read by the ConvNet-based system developed by our team [18]. But by the mid 1990s, interest in neural networks had waned in the machine-learning research community and a second "neural network winter" set in, that was to last over a decade.

The reasons for this winter are somewhat linked with all of hardware, software, and data. Neural networks required a lot of data, but, in the pre-Internet days "large" datasets were available only for a few tasks, such as handwriting and speech recognition. Second, computer hardware was limited. Furthermore, the workhorse of ML research, Unix workstations from SUN or SGI, were typically capable of only 10MFLOPS. Correspondingly, training a simple ConvNet for handwritten character recognition could take weeks. Third, software tools had to be built from the ground up. This was before the pre-eminence of Matlab, Python, and other interactive languages suitable for numerical work. Léon Bottou and I started developing a neural-network simulator with a home-grown Lisp-dialect front end in 1987. This system, called SN, eventually inaugurated the idea of neural networks as computational graphs of standard parameterized modules with automatic differentiation [19], a concept eventually picked up by frameworks like Theano, Torch, Caffe, and now TensorFlow, and PyTorch. This was before it was common-place for companies to release code in open source, and it is not until 2002 that it was open-sourced under the name Lush, as we were leaving AT&T. By then, few people were interested in neural networks.

At the tail end of this second winter, efforts to implement ConvNets on FPGA resumed at my NYU lab with the Xilinx Virtex4-based CNP project in 2008 [20], and the Virtex6-based NeuFlow project in 2010 [21]. NeuFlow revived the idea of dataflow stream processing in which a number of configurable operators can be dynamically assigned different functions, such as convolution, non-linear mapping, pooling/subsampling or arithmetic operations, all interconnected with suitable FIFOs to chain multiple operations without having to write intermediate results to external memory. Performance when running a ConvNet was roughly 150×10^9 operations per second, consuming 10W. Dataflow stream processors have become a popular design for ConvNet accelerators, such as the Eyeriss project [23, 24] which exploits the high level of data reuse in convolution operations.

Around 2003, Geoffrey Hinton (University of Toronto), Yoshua Bengio (University of Montréal) and I (having just left industry and joined NYU), with funding from the Canadian Institute for Advanced Research, started a series of research projects, workshops, and summer schools with a deliberate goal of reviving the interest of the ML community in neural networks. By 2007, enough researchers became interested again, following compelling new results in unsupervised layer-wise learning of very deep networks [25]. That is when the domain was rebranded as Deep Learning. Later, innovations such as using rectifying non-linearities (ReLU) instead of sigmoids and using dropout for regularization enabled purely supervised backprop training of very deep networks (with a dozen layers). By 2009, groups at Microsoft, Google, and IBM, were achieving significant reductions in error rates in speech recognition by replacing acoustic models based on Gaussian Mixture Models with deep networks [26]. Within 18 months, these systems were deployed commercially on Android phones and other platforms. Simultaneously, Collobert and Weston, then at NEC Labs, showed that a type of ConvNet architecture could produce vector representations of words that yielded record-breaking results on various natural language processing tasks [27, 28]. But the NLP community was initially skeptical and deep learning did not become dominant in NLP until quite recently.

Meanwhile, ConvNets were starting to produce record-breaking results on a number of image recognition tasks, such as semantic segmentation [29] and pedestrian detection [30], but the results were largely ignored by the computer vision community. Then in late 2012, Alex Krizhevsky working in Geoffrey Hinton's lab produced a very efficient implementation of ConvNets on GPUs. He was not the first person to do so (researchers at Microsoft, Stanford, IDSIA, and other places had done it before), but his implementation was efficient for very large and very deep ConvNet, allowing the team to win the ImageNet competition (object recognition in images with 1000 categories, and 1.3M training samples), reducing the top-5 error rate from 25% to 16% [31]. This sent shockwaves through the computer vision community. The team made its code available in open source, and within 2 years, almost everyone in the field was using ConvNets.

The lesson from this experience is that, *The availability of suitable hardware and simple-to-use open-source software is critical to the wide adoption of a class of methods by the community. Good results are essential but not sufficient.*

2 The Need for DL Hardware

Is DL-specific hardware really necessary? The answer is a resounding yes. One interesting property of DL systems is that the larger we make them, the better they seem to work. While this property is true for networks trained with supervised learning, the trend is to rely increasingly on unsupervised, self-supervised, weakly supervised or multi-task learning, for which larger networks perform even better. The demands on DL-specific hardware will undoubtedly increase.

There are five use cases with different hardware requirements: (1) DL research and development; (2) off-line training of DL models for production; inference on servers in data centers; (4) inference on mobile devices and embedded systems; (5) on-line learning on servers and mobile devices.

2.1 DL Research and Development: The requirement is for HPC-type multi-node machines, each hosting multiple GPUs or other flexible/programmable devices with fast 32-bit floating-point performance (FP-32). The communication network must be high bandwidth and low latency to allow for the parallelization of training large models on large datasets. Using FP-32 is necessary because one must be sure that when an experiment fails, it is not because of a lack of numerical accuracy. This use case requires high flexibility, programmability, extensive libraries, and efficient MPI-style communication libraries. Price and power consumption are relatively secondary to performance and flexibility. For most types of DL models, parallelization can be done easily by distributing multiple training samples across processing elements, GPUs, or cluster nodes. However, batching multiple training samples at the level of a GPU is neither desirable, nor always possible. The best architectures are those that can be saturated with the smallest batch of samples.

2.2 Off-Line Training of Well-Understood Models: Once a model architecture has been tested and deployed, the process of retraining it periodically as new data becomes available can be optimized. It is possible to perform routine training on specialized hardware with reduced-precision arithmetics, such as 16-bit float (supported by NVIDIA GPUs and Google's TPU), fixed point, or more exotic number systems. Requirements for low-latency training and high levels of parallelization are less stringent than in R&D scenarios.

2.3 Inference on Servers: Today, much of the computational load of DL systems is spent running neural networks in data centers. However, the volume is growing quickly. For example, Facebook performs 3×10^{14} predictions per day (some of which are run on mobile devices). Many are relatively "simple" neural networks with sparse inputs used for newsfeed and advertisement ranking and for text classification [32]. But a lot of computation goes into larger ConvNets for image, video, and speech understanding, as well as for language translation. Every day, users upload 2 to 3 billion photos on Facebook. Each photo goes through a handful of ConvNets within 2 seconds of being uploaded. A large ConvNet trunk extract features used for generic tagging, objectionable content filtering (nudity and violence), search, and so on. Other ConvNets perform OCR of text in images (to detect hate speech), face detection and recognition, caption generation for the visually impaired, and a few other tasks. As communication services, such as live video, further expand, large-scale spatio-temporal ConvNets are being deployed to perform action recognition [33], speech recognition for subtitling, and language translation [34], all in real time with minimum latency. For this use case, power consumption and cost are important, flexibility and raw performance are secondary, and communication latency is unimportant. The ideal architecture is a specialized DL-inference accelerator sitting in a standard data-center server node. Since much of the computation is spent performing convolution, a convolutional net accelerator working on individual samples (not batches) is ideal. The requirements of the automotive industry for autonomous driving systems are somewhat similar, with considerably more stringent requirements on latency.

2.4 Inference on Mobile and Embedded Devices: The ubiquity of smartphones, and the upcoming availability of self-contained wearable devices for augmented reality (AR) and virtual reality (VR) are putting heavy demands on DL-inference accelerators with very-low power consumption. Real-time tasks require that the DL system be run on the device without the latency of a round-trip to a server. Applications include feature tracking and 3D reconstruction for AR, object segmentation/recognition, OCR in natural scenes, real-time language translation, and speech-based virtual assistants. Beyond mobile and wearables, low-cost DL chips will appear in cameras, appliances, autonomous surveillance and ground maintenance systems, and toys.

2.5 Rethinking Arithmetics: Given the robustness of DL systems to arithmetic imprecision, there is a distinct possibility that using unconventional number representations can improve efficiency for inference on servers and embedded devices. For example, [35] shows that an 8-bit logarithmic number system combined with a Kulisch accumulator leads to significant reduction in power consumption, while yielding negligible degradation in accuracy for a ResNet-50 network trained on ImageNet. Extended to 16 bits, this number system shows significant reduction in both power and silicon area over standard 16 bit floating point formats.

3 Present and Future Deep-Learning Architectures

The word "architecture" in the context of DL designates the graph of functional modules, not the architecture of the underlying hardware. Typical DL architectures are composed of a number of basic modules: multiple convolutions in 1D, 2D, and 3D; linear operators (matrices); linear operators applied to sparse inputs (word embedding lookup tables for NLP); divisive normalization; element-wise functions; pooling/subsampling; element-wise operators; bilinear operators (multiplicative interactions for attention); and so on. Low-level operations are often performed on a batch of multiple samples, simply because parallelization is simple. But there is no algorithmic reason to batch multiple samples. Thus, much of the DL R&D activity goes into designing architectures that are appropriate for a class of problems. Such popular families of architectures include ConvNets, multi-layer LSTM, Transformer Networks, and architectures with "attention" (multiplicative interactions).

3.1 DL Architectures Today

Video, image, and speech recognition, as well as language translation and NLP, use a variety of ConvNet architectures. Figure 1.1.3, from [22], shows the top-1 accuracy on ImageNet of various ConvNet designs as a function of number of operations. In such practical applications, much of the computation is spent performing convolutions in the lower layers. The upper layers are typically less compute bound but more memory bound [24]. The spatio-temporal resolution of layer activations typically diminishes in the upper layers, but the number of channels or feature types typically increases.

In computer-vision applications, the trend is to apply ConvNets to an entire image so as to detect, segment, and recognize objects of any size at any location [37, 38, 39, 40, 62]. In such networks, all the layers are convolutional, though some of the top layers, sometimes called "fully connected, can be viewed as convolutions with a 1×1 kernel.

An increasingly popular class of ConvNet architectures for image segmentation, reconstruction, and object detection are the so-called Feature U-Net, Feature Pyramid Network, RetinaNet, and variants [39, 40, 41]. They can be viewed as a ConvNet encoder topped by a "reverse ConvNet" decoder whose role is to produce an image at the same resolution as the input (see Figure 1.1.4). They contain skipping connections from each layer of the encoder to the layer of the corresponding resolution in the decoder. The number of applications of this type of architecture is likely to increase in image annotation, autonomous driving, and medical image analysis [42].

Similar architectures are used in the context of image generation and video prediction [43, 46]. Video prediction is a subject of wide interest because it may allow future systems, such as robots or self-driving cars, to predict what is going to happen in their environment and to plan accordingly [44, 45].

In translation and language understanding, the Transformer Network architecture is increasingly popular [36]. It makes extensive use of multiplicative interactions.

3.2 Architectural Elements of Future DL Systems

We are witnessing an evolution in the types of architecture proposed by DL researchers, which may determine what hardware will be required in the near future. Generally speaking, the evolution is towards more sophisticated network architectures, *dynamic* network architectures that change with each new input in a data-dependent way, inputs and internal states that are not regular tensors, but are graphs whose nodes and edges are annotated with numerical objects (including tensors).

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

3.2.1 Dynamic Networks, Differentiable Programming: A relatively recent concept in DL is the idea of dynamic networks. Regular DL systems use a static network of parameterized modules. But, in increasingly many applications, the network architecture is dynamic and changes for every new data point. In effect, dynamic DL systems can be seen as the execution trace of a program, with conditionals and loops that are input-dependent. DL frameworks such as PyTorch record a "tape" of this execution trace, which can be played backwards to back-propagate gradients through the program. This method is known as "autograd". The phrase "differentiable programming" designates the process of writing a program with calls to parameterized functions that automatically compute the gradient of the function's output with respect to the parameters, allowing the function to be finalized through learning. Dynamic networks are particularly useful in a variety of applications: for natural-language processing, for data that does not come in the form of a fixed-sized tensor, for systems that need to activate parts of a large network on demand in a data-dependent way (such as the Multi-scale DenseNet architecture [47] shown in Figure 1.1.4) and for "reasoning" networks whose output is another network specifically designed to answer a particular question [48, 49] (see Figure 1.1.5).

3.2.2 Neural Networks on Graphs: One of the most exciting recent developments in DL is neural networks on graphs [50]. Many problems are difficult to represent with fixed-size tensors or variable-length sequences of tensors, but are better represented by graphs whose arcs and nodes are annotated by tensors. This suggests the use of networks of differentiable modules whose inputs and outputs are annotated graphs. The idea goes back to Graph Transformer Networks, built to recognize character strings [18]. But recent incarnations of graph neural networks have been applied to 3D meshes, social networks, gene-regulation networks, and chemical molecules. Convolution operations can easily be defined on irregular graphs: they are defined as diagonal operators in the eigenspace of the graph Laplacian, which is a generalization of the Fourier transform. We foresee an increase in the usage of such networks for a wide variety of applications, which are likely to violate the assumptions of current DL hardware.

3.2.3 Graph Embedding Networks: Increasingly, DL is used for large-scale embedding of knowledge bases. For example, using a large knowledge graph composed of triplets (subject, relation, object), such as ("Barak Obama", "was born in", "Hawaii") one may train a network to rate such triplets or to predict one of the elements from the other two. A special case of this consists in learning a vector for each object and subject, such as a simple scalar-valued operation between the vectors (distance) will predict the presence or absence of a particular relation between the object or subject. These methods, applied on a large scale, are particularly efficient for recommender systems, and can use hyperbolic metric spaces to represent hierarchical categories [51].

3.2.4 Memory-Augmented Networks: To endow DL systems with the ability to reason, they need a short-term memory, to be used as an episodic memory, or a scratchpad/working memory. For example, if a system is to answer questions about a series of events (described as a text), it must be able to store the story in a memory and retrieve the relevant bits to answer a particular question. This led to the memory-network architecture [52, 53] in which a recurrent neural nework is augmented by what amounts to a differentiable associative memory circuit (see Figure 1.1.5). This associative memory can be quite large and requires finding the nearest neighbors to a key vector very efficiently. As DL systems are increasingly used for high-level cognitive tasks, such memory modules will become commonplace and very large, requiring hardware support.

3.2.5 Complex Inference and Search: Most of today's DL systems simply produce an output given an input. But complex reasoning requires that the output variable actually be an *input* to a scoring network whose scalar output (akin to energy) indicates the incompatibility between the input and an output proposal. An inference procedure must search for the output value that minimizes the energy. This type of model is called an *energy-based model* [54]. If the energy-minimizing inference procedure is gradient-based, inference hardware will need to support back-propagation.

3.2.6 Sparse Activations: As the size of DL systems grows, it is likely that the modules' activations will become increasingly sparse, with only a subset of variables of a subset of modules being activated at any one time. This is akin to how the brain represents information: on average, neurons in the brain are at 2% of their maximum activation, and most neurons are quiet most of the time, which is good for power dissipation. Examples of explicitly sparse networks already exist (for processing volumetric imaging data [56]).

3.2.7 Overall:

New architectural concepts such as dynamic networks, graph data, associative-memory structures, and inference-through-minimization procedures are likely to affect the type of hardware architectures that will be required in the future.

4 The Revolution will not be Supervised

With all the hype around the new AI and DL, the way machines learn today is vastly less efficient than the way humans and animals learn. Almost all practical applications of DL use supervised learning (SL), in which the system is fed the desired output during training, with a tiny minority using reinforcement learning (RL). Most humans are capable of learning to drive a car in about 30 hours of training without ever causing accidents. In contrast, current model-free RL methods would likely require millions of hours of practice, with numerous accidents, for an autonomous car to learn to drive. This is not a problem in easy-to-simulate fully-observable environments with discrete state, such as the game of go or chess. But, it does not work in the real world! Obviously, our current learning paradigms are missing a key ingredient.

One hypothesis is that this missing ingredient is self-supervised learning. The bulk of learning in humans and animals is self-supervised: we learn enormous amounts of background knowledge about how the world works by observation in the first days, weeks, and months of life. In particular, we learn intuitive physics and the properties of the physical world. By the age of 9 months, babies understand object permanence, stability, animate vs inanimate objects, stability, gravity, inertia, and so on. The ability to predict what is going to happen in the world is what allows us to learn to drive without causing accidents: our world model allows us to anticipate the consequences of our actions, to maintain the car on the road, and to avoid disasters.

The idea of self-supervised learning is to train a machine to predict any subset of its input from other subsets (with a possible overlap between the subsets). For example, given a 6-frame video clip, one could train a DL system to predict the last two frames from the first four.

Why should SSL be more efficient than either RL or SL? In RL, the system produces an output (often an action or sequence of actions) and gets in return a single scalar value representing the "reward" for this action. Learning a complex task in this scenario requires a very large number of trials, and a large number of errors. While the process works fine for fully-observable games (such as chess and go) where millions of trials can be generated through self-play, it is largely impractical in the real world. A model-free RL system would require millions of hours of driving and numerous crashes to train a car to drive itself. The number of trials required is large because the feedback from the environment is information-poor. In SL, the system is given the correct answer, generally in the form of a target output vector. While this is less information-poor than in RL, it still requires a lot of training samples to capture the essence of the problem. On the other hand, SSL asks the machine to predict a large amount of information in the form of a high-dimensional signal (such as a whole video frame). More complex models with more parameters can be learned with a given number of samples or trials. The main difficulty is that predicting the future of a video is not achievable exactly because the world is not entirely predictable. If one uses a least-square criterion to train a video predictor, the resulting predictions are blurry frames: an average of all the possible futures. To make sharp predictions, one must have a set of latent variables that, when passed through a predictor, parameterize the set of plausible predictions. One technique used to train such models is Generative Adversarial Networks (GAN) [59], which for training uses two networks simultaneously: a *generator* that makes predictions using

observations and a source of random vectors drawn from a known distribution, with a *discriminator* whose role is to produce a scalar energy indicating whether a generated prediction is plausible or not. The discriminator is trained to distinguish real data (low energy) from generated predictions (high energy). The generator trains itself to produce predictions that the discriminator cannot tell are fake. To do so, the generator uses the gradient of the discriminator's output energy with respect to its input to compute how to modify its predictions, and thereby modify its parameters. Variations of GANs have produced stunning results in image generation [61, 62]. Other latent-variable generative models, such as Variational Auto-Encoders [60] and regularized latent variable models [45] have also produced good results.

One hope is that training a system to predict videos will allow it to discover much of the hidden regularities, geometry, and physics of the world, such as the fact that the scenery changes in particular ways as the camera moves, and that certain objects occlude others and can move independently. Such predictions can be done in pixel space [43, 44], or in higher-level representations (such as instance segmentation maps obtained by a pre-trained system [46]).

The use of predictive models that not only predict the evolution of the environment, but also predict the consequences of actions, is key to reducing the number of trials a system needs to learn a skill. I predict that self-supervised latent-variable predictive models will be the centerpiece of intelligent systems based on model-predictive control and model-based reinforcement learning for such applications as robotic grasping and manipulation [44] and autonomous driving [45]. Figure 1.1.6 shows a latent-variable predictive model that predicts a visual representation of the surroundings of a car. This kind of model can be used to predict multiple scenarios of how surrounding cars are going to move, and to plan a driving policy accordingly.

If self-supervised learning eventually allows machines to learn vast amounts of background knowledge about how the world works through observation, one may hypothesize that some form of *machine common sense* could emerge! One form of common sense is our ability to fill in the blanks, using our knowledge of the structure and constraints of the world.

Future DL systems will largely be trained using a form of self-supervised learning. These systems will be much larger than they are today, because the amount of data with which they can be trained (e.g. raw video) is essentially unlimited. Such systems will eventually be trained to acquire vast amounts of background knowledge so as to acquire a form of common sense. New high-performance hardware will be required to enable such progress.

5 Requirements for Future DL Hardware and Software

5.1 How Will DL Software Evolve?

Clearly, what is needed is a software framework for *differentiable programming* that is both interactive, flexible, dynamic, and efficient. Although frameworks such as PyTorch, TensorFlow, and others are moving in that direction, the main obstacle is that people love Python, largely because of its gigantic set of libraries. But Python is very slow and memory hungry. It is often impractical to develop high-volume applications or embedded applications that rely on Python at runtime. However, for static compute graphs, there is no issue: one can export the graph to adhere to a standard format, such ONNX (Open Neural Net Exchange), and use one of the numerous ONNX-compliant backends. On the other hand, for dynamic networks, there are two main options: One is to provide a compiler for a sufficiently large subset of Python that can produce Python-independent executables for DL (such as Torch.Jit in the recently-released PyTorch-1.0 [64]). This may also require an auxiliary domain-specific language to specify low-level numerical operations (on tensors and graphs) such as Tensor Comprehensions [55]; A second option is to design a suitable compilable language from scratch. It would have to be interactive and dynamic, have safe parallelism, and use type inference as much as possible, perhaps something resembling Julia or Skip [63] with good support for scientific computing. However, dedicated user's desire to access the vast repository of Python libraries will limit its potential adoption.

5.2 Hardware for Training

One problem is that sparsity, architecture dynamicity, and modules that manipulate non-tensor data (graphs), break the assumption that one can perform computation on batches of identically-sized samples. Unfortunately, with current hardware, batching is what allows us to reduce most low-level neural network operations to matrix products, and thereby to reduce the memory access-to-computation ratio. Thus, we will need new hardware architectures that can function efficiently with a batch size of one. As well, handling sparse structured data is another requirement. Increasingly, input data will come to us in a variety of forms, beyond tensors, such as graphs annotated with tensors and symbols.

Down the line, one can imagine architectures and learning algorithms that favor sparse activations in the network. When most units are off most of the time, it may become advantageous to make our hardware *event driven*, so that only the units that are activated consume resources. Such sparse networks, such as Submanifold Sparse ConvNets (implemented in software) have been shown to be very effective for processing sparse data, such as 3D scenes, which are represented by voxel arrays that are largely empty [56]. Sparse activation is one of the features that makes the brain so power-efficient.

5.3 Hardware for Inference

While demand for data-center and cloud-based inference will grow, future DL applications will increasingly run on mobile phones, wearables, home apppliances, vehicles, IoT devices, and robots. Applications in augmented and virtual reality and telepresence will require extremely low-power ASICs for DL inference for such things as real-time/low-latency object tracking, 3D re-construction, instance labeling, facial reconstruction, predictive compression and display.

In the short and medium term, the bulk of the computation will be convolutions. Since batching is out of the question, hardware will have to exploit the regularities of convolutions instead of being mere matrix-product engines.

Ultimately, the solution to power constraints may well be the exploitation of sparse activations, perhaps using event-based computation. In any case, it may exploit the use of exotic number representations (the 8-bit logarithmic representation of [35]).

6 The Long Term Outlook

In the long run, could we see a return to analog implementations? Perhaps programmable resistor technology will become sufficiently compact, reliable, durable, and configurable for DL applications. But since this would require one unmovable physical memory cell per parameter in the network, only activations could be circulated (assuming they are converted to digital representation), and hardware multiplexing would be limited to sections that share weights (as in the ANNA chip). It is very unclear whether analog implementations provide any power dissipation advantages over digital, and current evidence seems to point in the opposite direction.

A number of authors have been advocating architectures with spiking neurons. Unfortunately, the performance of spiking neuron circuits seems considerably inferior to that of traditional digital architectures for realistic ConvNet-type networks [57]. Current learning algorithms do not take advantage of the peculiarities of spiking networks, and no spiking-neuron learning algorithms has been shown to come close to the accuracy of backprop with continuous representations.

The important trends discussed in this paper include: (1) more self-supervised learning, resulting in larger network architectures; (2) dynamic network resulting from *differentiable programs* whose architecture changes for each new sample; (3) the need for hardware that is efficient for batch-size 1, implying the end of reliance of matrix products as the lowest-level operator; (3) exotic number representation for inference on low-power hardware; (4) very large networks with very sparse activations, that new architectures could exploit for power reduction; (5) new operators such as fast K-nearest neighbors for

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

(differentiable) associative-memory modules; (6) networks that manipulate annotated graphs instead of tensors. However, chances are that the bulk of the computation in future DL systems will still consist primarily of convolutions.

References

[1] L.D. Jackel, R.E. Howard, H.P. Graf, B. Straughn, J.S. Denker, "Artificial Neural Networks for Computing". Journal of Vacuum Science & Technology B: Microelectronics Processing and Phenomena, 4(1), pp. 61-63, 1986.

[2] H. Graf, P. de Vegvar, "A CMOS Associative Memory Chip Based on Neural Networks", *ISSCC*, pp. 304- 305, 1987.

[3] G. Indiveri, et al., "Neuromorphic Silicon Neuron Circuits", *Frontiers in Neuroscience*, 5, p. 73, 2011.

[4] S.B. Furber, F. Galluppi, S. Temple, L. Plana, "The Spinnaker Project", Proceedings of the IEEE, 102(5), pp. 652-665, 2014.

[5] F. Rosenblatt, "The Perceptron, A Perceiving and Recognizing Automaton (Project Para)". Cornell Aeronautical Laboratory, 1957.

[6] B. Widrow, W.H. Pierce, J.B. Angell. "Birth, Life, and Death in Microelectronic Systems", IRE Trans. Mil. Electron., 1051(3), pp. 191- 201, 1961.

[7] R.W. Lucky, "Automatic Equalization for Digital Communication", Bell System Technical Journal, 44(4), pp. 547-588, 1965.

[8] M. Minsky, S.A. Papert, "Perceptrons: An Introduction to Computational Geometry". MIT press, 1969.

[9] J.J. Hopfield. "Neural Networks and Physical Systems with Emergent Collective Computational Abilities". Proceedings of the National Academy of Sciences, 79(8), pp. 2554-2558, 1982.

[10] G.E. Hinton, T.J. Sejnowski, "Optimal Perceptual Inference", Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 448-453, June 1983.

[11] D.E. Rumelhart, G.E. Hinton, R.J. Williams, "Learning Representations by Back-Propagating Errors. Nature, 323(6088), pp. 533, 1986.

[12] Y. LeCun, B.E. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition", Neural Computation, 1(4), pp. 541-551, 1989.

[13] Y. LeCun, B.E. Boser, J.S. Denker, D. Henderson, R.E. Howard, W.E. Hubbard, L.D. Jackel, "Handwritten Digit Recognition with a Back-Propagation Network", NIPS, pp. 396-404, 1989.

[14] H.P. Graf, R. Janow, D. Henderson, R. Lee, "Reconfigurable Neural Net chip with 32K Connections", Advances in Neural Information Processing Systems, pp. 1032-1038, 1991.

[15] B.E. Boser, E. Sackinger, J. Bromley, Y. Le Cun, L.D. Jackel, "An Analog Neural Network Processor with Programmable Topology". IEEE Journal of Solid-State Circuits, 26(12), pp. 2017-2025, 1991.

[16] E. Sackinger, B.E. Boser, J. Bromley, Y. LeCun, L.D. Jackel, "Application of the ANNA Neural Network Chip to High-Speed Character Recognition", IEEE Transactions on Neural Networks, 3(3), pp. 498-505, 1992.

[17] J. Cloutier, E. Cosatto, S. Pigeon, F.R. Boyer, P.Y. Simard, "VIP: An FPGA-Based Processor for Image Processing and Neural Networks", Proc. of Int. Conf. Microelectronics for Neural Networks, pp. 330-336, 1996.

[18] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, "Gradient-Based Learning Applied to Document Recognition", Proceedings of the IEEE, 86(11), pp. 2278-2324, 1998.

[19] L. Bottou, P. Gallinari, "A Framework for the Cooperation of Learning Algorithms", Advances in Neural Information Processing Systems, pp. 781-788, 1991.

[20] C. Farabet, C. Poulet, Y. LeCun, "An FPGA-Based Stream Processor for Embedded Real-Time Vision with Convolutional Networks, ICCV Workshops, pp. 878-885, September 2009.

[21] C. Farabet, Y. LeCun, K. Kavukcuoglu, E. Culurciello, B. Martini, P. Akselrod, S. Talay, "Large-Scale FPGA-Based Convolutional Networks", R. Bekkerman, M. Bilenko, J. Langford (Eds.), "Scaling up Machine Learning: Parallel and Distributed Approaches", pp. 399-419, Cambridge University Press, 2011.

[22] A. Canziani, A. Paszke, E. Culurciello, "An Analysis of Deep Neural Network Models for Practical Applications, arxiv: 1605.07678, 2017.

[23] Y.H. Chen, T. Krishna, J.S. Emer, V. Sze, "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks". IEEE Journal of Solid-State Circuits, 52(1), pp. 127-138, 2017.

[24] Y.H. Chen, J. Emer, V. Sze, "Eyeriss v2: A Flexible and High-Performance Accelerator for Emerging Deep Neural Networks". arXiv:1807.07928, 2018.

[25] G.E. Hinton, R.R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks", Science, 313(5786), pp. 504-507, 2006.

[26] G.E. Hinton, L. Deng, D. Yu, G.E. Dahl, A.R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The shared views of four research groups", IEEE Signal Processing Magazine, 29(6), pp. 82-97, 2012.

[27] R. Collobert, J. Weston, "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. ICML, pp. 160-167, 2008.

[28] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, "Natural Language Processing (almost) from Scratch", Journal of Machine Learning Research, pp. 2493-2537, August 2011.

[29] C. Farabet, C., Couprie, L. Najman, Y. LeCun, "Scene Parsing with Multiscale Feature Learning, Purity Trees, and Optimal Covers", ICML arXiv:1202.2160, 2012.

[30] P. Sermanet, K. Kavukcuoglu, S. Chintala, Y. LeCun, "Pedestrian Detection with Unsupervised Multi-Stage Feature Learning", CVPR pp. 3626-3633, 2013.

[31] A. Krizhevsky, I. Sutskever, G.E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks", Advances in Neural Information Processing Systems, pp. 1097-1105, 2012.

[32] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, "Bag of Tricks for Efficient Text Classification", Proc 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Vol. 2, pp. 427-431, 2017.

[33] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, "A Closer Look at Spatiotemporal Convolutions for Action Recognition:, Proc. Computer Vision and Pattern Recognition, pp. 6450-6459, 2018.

[34] M. Ott, S. Edunov, D. Grangier, M. Auli, "Scaling Neural Machine Translation". arXiv:1806.00187, 2018.

[35] J. Johnson, "Rethinking Floating Point for Deep Learning", ArXiv:1811.01721, 2018.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, "Attention is all you need" NIPS, pp. 5998-6008, 2017.

[37] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, "Overfeat: Integrated Recognition, Localization and Detection Using Convolutional Networks". Proc. ICLR, arXiv:1312.6229, 2014.

[38] K. He, G. Gkioxari, P. Dollr, R. Girshick, R. (2017, "Mask R-CNN", Proc. ICCV, pp.2980-2988, October 2017.

[39] T.Y. Lin, P. Dollr, R.B. Girshick, K. He, B. Hariharan, S.J. Belongie, "Feature Pyramid Networks for Object Detection" CVPR, Vol. 1, No. 2, p. 4, 2017.

[40] T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollr, "Focal Loss for Dense Object Detection", Proc. ICCV, arXiv:1708.02002, 2017.

[41] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", International Conference on Medical Image Computing and Computer-Assisted Intervention", pp. 234-241, October 2015.

[42] P. Jaeger, S. Kohl, S. Bickelhaupt, F. Isensee, T.A.Kuder, H.-P. Schlemmer, K. Maier-Hein, "Retina U-Net: Embarrassingly Simple Exploitation of Segmentation Supervision for Medical Object Detectio", arXiv:1811.08661, 2018.

[43] M. Mathieu, C. Couprie, Y. LeCun, "Deep Multi-Scale Video Prediction Beyond Mean Square Error", ICLR, arXiv:1511.05440, 2016.

[44] C.Finn, I. Goodfellow, S. Levine, "Unsupervised Learning for Physical Interaction Through Video Prediction", Advances in Neural Information Processing Systems, pp. 64-72, 2016.

[45] M. Henaff, A. Canziani, Y. LeCun, "Model-Predictive Policy Learning with Uncertainty Regularization for Driving in Dense Traffic" To appear in 2019.

[46] P. Luc, C. Couprie, Y. LeCun, J. Verbeek, "Predicting Future Instance Segmentations by Forecasting Convolutional Features". ECCV, arXiv:1803.11496, 2018.

[47] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, K.Q. Weinberger, "Multi-Scale Dense Networks for Resource Efficient Image Classification". ICLR, arXiv:1703.09844, 2018.
[48] J. Johnson, et al, "Inferring and Executing Programs for Visual Reasoning" ICCV, pp. 3008-3017, 2017.
[49] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, K. Saenko, "Learning to Reason: End-to-End Module Networks for Visual Question Answering", ICCV arxiv:1704.05526, 2017.
[50] M.M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, P. Vandergheynst, "Geometric Deep Learning: Going Beyond Euclidean Data", IEEE Signal Processing Magazine, 34(4), pp. 18-42, 2017.
[51] M. Nickel, D. Kiela, "Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry", arXiv:1806.03417, 2018.
[52] S. Sukhbaatar, J. Weston, R. Fergus, "End-to-End Memory Networks", Advances in Neural Information Processing Systems, pp. 2440-2448, 2015.
[53] A. Miller, A.Fisch, J. Dodge, A.H. Karimi, A. Bordes, J. Weston, "Key-Value Memory Networks for Directly Reading Documents", ArXiv:1606.03126, 2016.
[54] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, F. Huang, "A Tutorial on Energy-BasedLlearning.in Bakir et al (Eds), *Predicting Structured Data*, MIT Press, 2006.
[55] N. Vasilache, O. Zinenko, T. Theodoridis, P. Goyal, Z. DeVito, W.S. Moses, S. Verdoolaege, A. Adams, A. Cohen, "Tensor Comprehensions: Framework-Agnostic High-Performance Machine Learning Abstractions". arXiv:1802.04730, 2018.
[56] B. Graham, M. Engelcke, L. van der Maaten, "3D Semantic Segmentation with Submanifold Sparse Convolutional Networks", CVPR, pp 18-22, 2018.
[57] C. Farabet, R. Paz, J. Prez-Carrasco, C. Zamarreo, A. Linares-Barranco, Y. LeCun, E. Culurciello, T. Serrano-Gotarredona, B. Linares-Barranco, "Comparison Between Frame-Constrained Fix-Pixel-Value and Frame-Free Spiking-Dynamic-Pixel ConvNets for Visual Processing:, Frontiers in Neuroscience, 6, 32, 2012.
[58] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y Li, A. Bharambe, L. van der Maaten, "Exploring the Limits of Weakly Supervised Pretraining" ECCV, arXiv:1805.00932, 2018.
[59] I. Goodfellow, et al, "Generative Adversarial Nets", NIPS, pp. 2672-2680, 2014.
[60] D.P. Kingma, M. Welling, "Auto-Encoding Variational Bayes", ICLR. arXiv:1312.6114, 2014.
[61] T. Karras, T. Aila, S. Laine, J. Lehtinen, "Progressive Growing of Gans for Improved Quality, Stability, and Variation", ICLR. arXiv:1710.10196, 2018.
[62] A. Brock, J. Donahue, K. Simonyan, "Large Scale Gan Training for High Fidelity Natural Image Synthesis". arXiv:1809.11096, 2018.
[63] http://github.com/facebookresearch/maskrcnn-benchmark
[64] http://www.skiplang.com
[65] https://pytorch.org

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

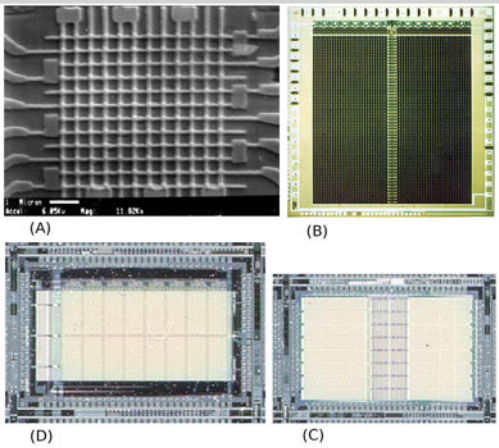


Figure 1.1.1: Early neural network chips from Bell Labs. (A) 1986: 12-resistor array, 6×6 microns [1]; (B) 1987: 54×54 analog array with programmable ternary weights [2]; (C) 1991: Net32K Convolver 256×128 programmable ternary weight array with FIFOs for convolutions [14]; (D) 1991: ANNA ConvNet chip 64×64 array with 6-bit weights and 3-bit activations [15].

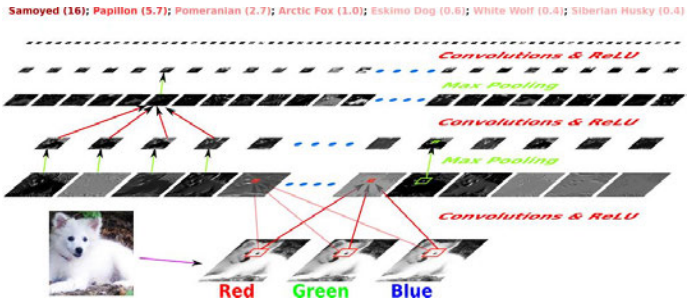


Figure 1.1.2: An example of Convolutional Network architecture for image recognition. Not all layers are represented [37].

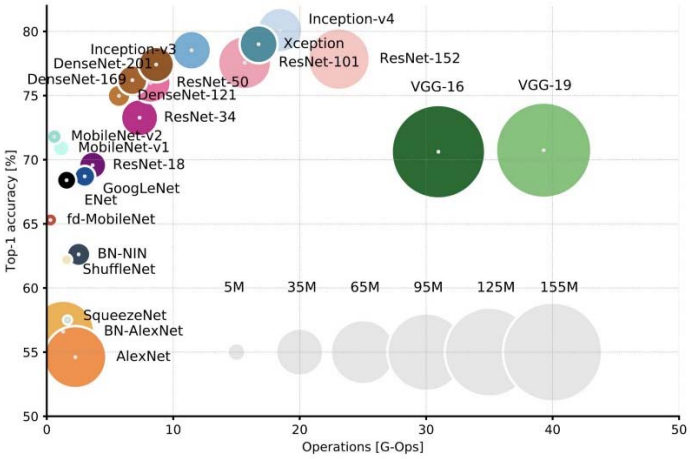


Figure 1.1.3: Top-1 accuracy on ImageNet versus number of operations for one pass of various ConvNet architectures. Circle size indicates the number of parameters [22].

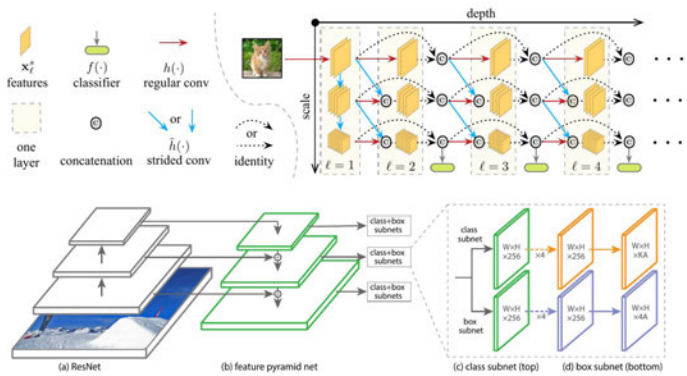


Figure 1.1.4: (Top) Multi-Scale DenseNet with conditional computation for accelerated results [47]. (Bottom) RetinaNet architecture for image semantic segmentation [40].

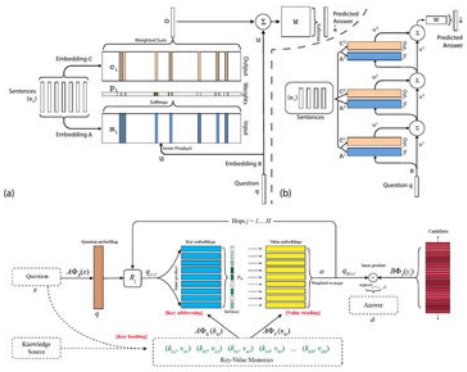


Figure 1.1.5: (Top) Memory Network architecture [52]; (Bottom) Key-Value Memory Network architecture for question answering [53]. Both architectures contain a central processing network connected with a "soft" associative memory circuit that stores facts. The memory module is a "soft" associative memory circuit in which the "address"vector is compared with each key vector through a dot product, producing scalar matching scores. The scores are normalized to sum to one. The output is a linear combination of the stored value vectors, weighted by the normalized scores.

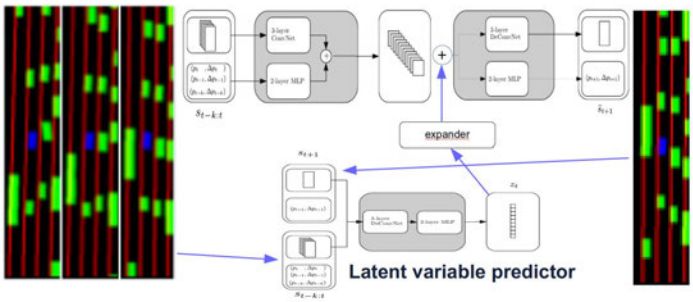


Figure 1.1.6: An example of self-supervised learning. A latent-variable model predicts how surrounding cars will move relative to the ego car (in the center). The model takes a few past frames and predicts the future relative positions of other cars, conditioned on a vector of latent variables. It is trained using data collected from traffic cameras overlooking roads. Different samplings of the latent variable produce different futures. This model can be used to plan or to train an artificial driver to minimize the probability of collision.

1.2 Intelligence on Silicon: From Deep-Neural-Network Accelerators to Brain Mimicking AI-SoCs

Hoi-Jun Yoo, *KAIST, Daejeon, Korea*

1. Introduction

While, currently, Artificial-Intelligence technology is affecting all industrial paradigms, it is also impacting lifestyle of all society. AI technology is widely used in most information hardware, software, and networking, underlying all consumer technology: smartphones, home appliances, and the Web.

Figure 1.2.1 shows 3 main application areas for AI technology: Cloud, Edge, and Mobile. The first is Cloud AI where intelligent applications are processed by server machines at data centers with results sent to edge controllers (such as speakers) and through to mobile devices (such as smartphones). As AI applications become more popular, servers need to incorporate dedicated AI accelerators to speed up the ever-increasing demand for AI processing with low energy consumption [1, 2].

In addition to AI processing in servers, recently many edge devices such as CCTV or AI speakers (as depicted in Figure 1.2.1) are required to process AI algorithms directly themselves, as so-called "on-device AI" [3]. This is needed, for example, in the case of an edge CCTV camera since most of the data which it collects and sends to the server is mundane, within infrequent critical events. Thus, it is apparent that sending all of the useless data to the server is a waste of network communications and server computation power. Accordingly, in this case, if the edge device has intelligent functions to extract only the meaningful data and send only the selected data to the server, the communication traffic can be reduced with much quicker response from the server. But, the need to minimize communication and latency extends beyond the Edge to the mobile devices, themselves. Correspondingly, mobile devices (such as smartphones, automobiles, and drones) will perform AI functions with their own processors avoiding round-trip access to the Edge or server. For these applications, real-time operations are required to avoid obstacles or collision with fast decision making with local intelligence even when they are moving at high speed. This is especially critical for smartphones which increasingly process high quality AI functions, such as face recognition and gesture recognition for user authentication, or natural user interface (UI), all with ultra-low power consumption. Smartphones are of particular concern because their security function is always-on.

In order to satisfy AI processing demands, high-performance silicon processors have already been developed to provide AI functions to the servers [1, 2, 4]. Currently, convolutional neural network (CNN) algorithms are accelerated using dedicated parallel processing elements (PE) to enhance convolution matrix multiplication, and to reduce external memory access. More recently, other deep-neural-network algorithms, such as recurrent neural network (RNN) and Restricted Boltzmann Machine (RBM) have been realized as accelerator chips [5]. As well, many DNN accelerators have been introduced as NPU (neural processing unit) and DPU (deep learning processing unit) or silicon IPs, taking the hardware intensive CNN operations from the main CPU to reduce the computation burden on the main processor [6].

However, the requirements for edge and mobile DNN accelerators are quite different from DNN accelerators used for server applications. Since many users want to run different AI functions and AI algorithms in the server, the server accelerator should support various types of DNN algorithms with high precision and large external memory. In contrast, the mobile accelerator usually supports only a limited DNN configuration and an algorithm for a specific application in real-time, with low power consumption, reduced number of precision, and limited memory size.

From an applications point of view, there will be (at least) two types of AI systems in the future: The first of these is an AI system replacing the human in the control loop (as in current Advanced Driver Assistance System (ADAS) applications) shown in Figure 1.2.2(a). Most current AI chips are targeting this category aiming to precisely control the machines just as well (or even better than) a human would. The second kind of AI system exhibits humanistic intelligence to collaborate symbiotically with human beings as illustrated in Figure 1.2.2(b). It structures the AI system to be more social, more like a pet or a human being, rather than like a machine which is unaffected by its user's emotions. In this situation, the AI system needs sensing ability to facilitate interaction with human users intelligently and emotionally, such as face recognition, emotion recognition, gesture recognition, and communication in natural language. (More details about such human-like intelligence will be explained in Section 3-C.)These social abilities are required particularly in mobile devices, and in contrast with AI in autonomous control, should be continuously active since its time of need is not predictable. Thus, low power consumption with high performance is critical, and energy efficiency rather than simple performance, is the most important parameter. In the following sections, the details of possible chip implementation schemes will be presented.

In this paper, we will examine the recent trends demonstrated in actual chip implementations of CMOS DNN accelerators; this is intended to present a sharp contrast with conventional mere descriptions of proposed unimplemented DNN chip architectures [7, 8, 9]. In Section 2, real chip implementation is emphasized for mobile AI applications using real chip DNN accelerators designed for high energy efficiency and high flexibility. In Section 3, the need for local learning will be introduced with new learning algorithms and hardware architectures. In particular, deep reinforcement learning will be emphasized as the next major application area for local learning. As well, humanistic intelligence will be introduced as a way to enhance the interaction between AI systems and human beings. In Section 4, we will describe a brain-mimicking AI SoC using two different approaches: One applies neuromorphic computing employing new non-volatile memories and mixed-mode circuits to mimic the neuron level activities of the brain. The second employs the connectome brain map to directly mimic functions of parts of the brain, and their interconnections. And finally, in Section 5 conclusion will be drawn.

2. CMOS DNN Accelerator ASICs

There have been many proposals for DNN accelerators [6] and many hardware architectures. As well, FPGA optimization schemes have been developed for DNN accelerators for high performance inference [6, 10]. At the same time, DNN algorithms are still evolving: A few software-centric works (such as SqueezeNet, MobileNets, and ShuffleNet) have been proposed to reduce the number precision without accuracy loss allowing reduced on-chip memory size to enable the DNN operations on existing mobile CPUs [5, 6, 10, 11, 12, 13]. For example, MobileNets can be installed on mobile devices because it is as accurate as VGG16 [14] while being 32 times smaller and 27 times less compute intensive. But the application of these software-centric techniques is very limited and performance degrades unbearably if they are applied to different applications from the optimized one. In this respect, FPGAs may be the way to utilize the most recent advanced algorithms, but with excess power consumption and limited performance compared to that of a custom ASIC. Recently, DNN accelerator architectures incorporate the flexibility of dynamically reconfigurable hardware architectures to meet the changing requirements of the new algorithms. It is important to note that in addition to the acceleration provided by the DNN algorithm, it also allows for greater energy efficiency. In the remainder of Section 2, the trends of the real chip implementation issues will be examined with emphasis on the energy efficiency and the reconfigurable architectures.

2.1 Energy Efficiency and Accuracy

DNN accelerators are optimized for convolutional computation with on-chip parallel processors, which also integrate the on-chip memory to store the weights and images, reducing access to the external memory. High-performance DNN accelerators have been developed for server applications and are currently successfully employed [1, 2, 4, 10]. However, the available approaches cannot obtain high energy efficiency (computation capability per unit power). Energy efficiency is particularly important in mobile DNN processors, along with the high computation capability of DNN algorithms. In order to enhance the energy efficiency without accuracy loss, not only the hardware but also the algorithms themselves should be optimized. The frequently used co-optimization schemes are Data Reuse, Reduced Number Precision, Weight Quantization, Sparsity Exploitation, Tensor Decomposition, Pruning, Approximate Computation, and Teacher-Net Learning [6]. These techniques can reduce the power consumption, chip complexity, on-chip memory size, and the off-chip memory access, but there may be a trade-off between the inference accuracy and chip complexity.

In addition to the DNN configuration and chip architecture level optimizations, circuit level optimizations are also useful for real chip implementation to reduce the power consumption and increase computation speed [15]. For example, the complicated multiplier can be replaced with a Look-up Table with quantization schemes to increase energy efficiency [16]. As well, the cell circuitry of the on-chip SRAM can be customized to speed up the vertical cell access [17], and sub-threshold logic can be used to enhance energy efficiency [17]. Potentially, an NVM (Non-Volatile Memory) rather than SRAM can enhance the energy efficiency further, as will be discussed in Section 3.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Figure 1.2.3 shows the performance-inference power efficiency of the reported DNN chips. 10TOPS/W is the criterion of high energy efficiency, and only the ASIC approach can meet the criterion. In addition to the above hardware-neural network co-optimization schemes, adopting small weight widths of 1bit and 4bits can lead to more than 10TOPS/W efficiency. Reducing the bit length to binary weights (BinaryConnect) and binary inputs (BNNs) will reduce the complexity of chips and reduce computation time, but accuracy degradation is unavoidable as a result [18]. Therefore, the programmability of bit lengths is required to secure the accuracy in addition to energy efficiency.

Figure 1.2.4 illustrates the trade-off between programmability and energy efficiency. Even though a GPU is not so versatile as a CPU, it is more effective in graphics-intensive applications, and an NPU, less programmable than a GPU, is more efficient in a convolution-intensive application. Since the programmability of an NPU is restricted compared to that of a GPU, there is concern that it may not accommodate new DNN algorithms in accordance with the evolution of algorithms. Recently, the flexibility of DNN accelerators has been enhanced with reconfigurable hardware, making it possible to accommodate various DNN algorithms on a single accelerator chip.

2.2 High Flexibility

Figure 1.2.5(a) shows 8 reported DNN chips as a function of energy efficiency and programmability. Programmability or hardware flexibility is as important as energy efficiency in DNN accelerators for optimizing mobile applications. A flexible computing architecture can customize the hardware architecture to provide exactly the required DNN processing capability needed. Therefore, it provides the programmability of hardware, just like FPGAs do for DNN accelerators while speeding up the processing time by an order-of-magnitude over fixed-instruction-set CPU with far less power consumption. There are two approaches to increase flexibility of DNN accelerators. The first is to use heterogeneous architectures, while the second is to adopt reconfigurable architectures.

Two parameters can be utilized to enhance the flexibility of DNN processors, the neural network hardware architecture and bit precision length with either coarse grain or fine grain as shown in Figure 1.2.5(b). Our first example is to integrate diverse types of DNNs, in this case, CNN and RNN, together in coarse grain [19]. Since the CNN has different optimal architecture from the RNN hardware, heterogeneous accelerators are designed for CNN and RNN separately. In addition, RNN and fully-connected layer (FCL) can share the same hardware architecture, RNN hardware can be reconfigured as the FCL. With software programming, the chip can be reconfigured to select the DNN architecture such as a CNN only or CNN+FCL or CNN+FCL+RNN [19]. The second approach is to reconfigure the bit precision in coarse grain with fine-grain architecture or the architecture in coarse grain with fine-grain bit precision [20]. For the given neural network, the bit precision of the weights and feature maps can vary 1 to 4bits layer by layer to reduce the required computation without significant impact on accuracy. In addition, both the DNN architectures and the bit length can be reconfigured together. For example, coarse bit length selection between 8bits and 16bits for PE functions, together with fine-grain hardware reconfiguration (such as the number of PE allocation) can be performed with software according to neural network types (CNN, RNN, and FCL), and it can be extended to use binary/ternary weights [21]. The third example is the most-advanced reconfigurable architecture, for which details are shown in Figure 1.2.6. Figure 1.2.6(a) shows both DNN architecture and the bit precision to be reconfigured dynamically in fine grain to further enhance the flexibility and the energy efficiency of the DNN accelerators [22]. Figure 1.2.6(b) explains its bit-serial architecture to vary the bit precision continuously from 1bit to 16bits. The network architecture can be changed dynamically as the fine-grain reconfigurable processing (DRP) even layer by layer to make it more versatile.

3. Local Learning and Intelligence

Learning is regarded as the basis of intelligence. Currently, learning for DNNs (that is weight determination) is done remotely, typically in the cloud server and downloaded to edge/mobile devices. It is a popular concept that learning is so computationally intense that only a high-performance server can perform it properly using available big data. However, the demand for local learning is increasing and a hardware solution for it is required.

3.1 Edge and Mobile Learning

Currently, DNN learning demonstrably takes a long time; For example, ResNet50 employs 1.2M images for learning, taking 11 days for training using a single Nvidia P40 GPU. Many DNN accelerators have been developed to speed up the training time in the cloud server [1, 2, 4]. A server-centric system as shown in Figure 1.2.7 uploads a large amount of data from mobile devices to the cloud server through a long wide area network, resulting in uncontrolled latency and high data-transmission energy consumption. Also, privacy may not be protected because even private data must be sent to the cloud server to use the intelligent applications. In addition, run-time training is required to filter out malicious or inconsistent data to protect the security of the system from adversarial attacks. Moreover, training of an edge or mobile device is necessary to realize the quick adaptation to the new environment, while reducing energy consumed by communications and the cloud server. For example, a pruned network can be re-trained on a mobile device with mobile learning for personalization and robustness [23]. Also, federated learning is proposed to protect privacy by keeping the sensitive data on the mobile devices while the learning task is solved by a loose federation of participating mobile devices which are coordinated by a central server [24].

Recently, local learning accelerators have been proposed for real-time object tracking and low-power local adaptation. A new algorithm, Feedback Alignment, is used to integrate on-line learning and inference on a single chip, making on-line learning possible in a mobile environment for real-time accurate object tracking [25]. Figure 1.2.7(b) provides an example of the datapath of a local CNN learning-accelerator chip. Weights are stored in a column buffer (C-buf), and feature maps are broadcasted to PEs and error maps are moved to the PEs from output memory to generate weight gradients. The weight gradients are aggregated temporally on accumulation registers of each PE, and then the results are stored to C-buf. As shown in the figure, the weight-gradient data path is reversed compared to that of the inference. It utilizes data sparsity optimally, together with variable number precision, enabling local mobile CNN learning at high speed and with low power [26].

3.2 Deep Reinforcement Learning

Reinforcement learning (RL), as illustrated in Figure 1.2.8, enables the AI system to find its optimal next operation in the unknown environment by repetitive self-learning. RL needs a reward to learn its optimal operation. Sometimes, the reward is clear, such as when a robot falls when it steps on irregular terrain. Of course, sometimes, in the real world, the reward cannot be clearly defined. In this case, the reward can be predicted in three possible ways: use demonstrated human behaviors as the reward (imitation learning, inverse reinforcement learning (IRL)); decide on the reward by itself (self-supervised learning, model-based learning); or obtain the reward with the help of other modalities.

The policy function of DQL (Deep Q-Learning) uses DNN for on-chip learning which enables the AI system with DQL to update its policy function. AI robot and the autonomous car will utilize Deep RL to make decisions much as a human being does, in a new environment different from the one in which it learned. In this case, remote learning may suffer from unpredictable latency, and local learning is mandatory to obtain secured RL with on-site data in real-time, such as for car-driving as illustrated in Figure 1.2.8. The Figure illustrates a multi-step process of DRL with four functions: the AI system in the car observes its state using its sensors; the policy function is implemented as the DNN; the DNN receives the state (how fast the car moves, whether there are pedestrians, the street conditions, and so on) as its input and generates the next action; and the AI system calculates the reward as the feedback from the resulting output action. Following a few thousand action cycles and the collection of the observed states, the policy-update process starts to modify the weights in the DNN. With the newly modified policy, the action is regenerated and the policy update is repeated. Usually, this process requires massive computation and memory size, requiring a full function DRL chip with RNN learning or FC-layer learning as reported in [27].

3.3 Humanistic Intelligence

As explained in Figure 1.2.2, AI systems and human beings will collaborate in the near future, on the basis of mutual understanding. The required trust and understanding are acquired by the AI system through local learning and recognition of the intent and emotion of the user. In this case, the AI system can be regarded as a pet or a companion with humanistic intelligence so that users can communicate or even share their emotions with the AI system. Learning

techniques, such as DRL, will be the key technology for the personalization and collaboration of the AI system. But, as well, the human user can also learn to adapt their behavior over time. For example, the ever-changing emotions of the user can act as a reward for the DRL process, with the robot learning and modifying its behavior by recognizing the user's emotions [28]. Figure 1.2.9 provides examples of an emotion recognition DNN accelerator [22], as well as a gesture recognition accelerator based on CNN and RNN [19]. Human gestures, such as clapping illustrated on the left side of the Figure, must be deciphered according to the temporal sequence of human poses. The RNN receives the output of the CNN to recognize the gesture. The emotion can be detected based on the shapes of the facial muscles using CNN. The right side of the Figure shows an example of emotion recognition (choosing one of 7 emotion categories), using a chip for facial frontalization and CNN [29].

4. Brain-Mimicking AI SoCs

There are two different ways to mimic the brain functions as shown in Figure 1.2.10. The first is the neuromorphic approach which tries to mimic the neuronal behavior, and the second is the functional modeling approach which mimics the cognitive behavior of a part or all of the brain.

4.1 Neuromorphic Approach

The neuromorphic approach of Figure 1.2.10 can be divided again into two methods, synapse centric and neuron centric. The first one is to mimic the basic "synaptic" behavior such as spike neural network or pulsed neural network with analog-digital mixed-mode circuits. Information can be conveyed by spike timing, for example, using the interval between spiking signals (such as spike-timing-dependent plasticity (STDP)), providing large information transmission capacity [30, 31]. SRAMs are used to store the digital values of the weights which are converted into analog, and multiplied with the input pulses to obtain output spiking signals. The design consumes less power than current Von Neumann architectures, but as a novel technique, it lacks high performance AI functions due to their less developed AI algorithms and limited integration scale due to SRAM cell size (100 to 200F^2, 10 times more than DRAM NVM cells).

If the synapse can be realized with a single device rather than a complicated circuit, the integration level can be easily scaled up with reduced power consumption. The non-volatile memories with 4F^2 to 12F^2 cell size such as Memristor (using Resistive RAM (RRAM)), Phase Change RAM (PCRAM), and Magnetic RAM (MRAM) have been actively studied as synaptic devices. If such a device can be integrated with neuronal circuits, non Von Neumann architectures such as biological neuronal computation can be realized, minimizing power consumption needed to transfer weight values from memory to the synapse [32, 33, 34]. Furthermore, in principle, the Memristor can also store analog weights so that a compact analog spiking neural network could be easily realized. Such Processor in Memory (PIM) (or Computation in Memory (CIM)) technology is a promising direction for the realization of AI functions. However, such cell-device technology is not yet mature, and only limited array sizes have been realized in actual silicon for small applications including character recognition or low-resolution face recognition [32].

The neuron-centric approach is to create an analog circuit which mimics the behavior of the "neuronal" cell faithfully [35, 36, 37, 38, 39, 40, 41]. Their attempts of reproducing biological synaptic or neuronal behaviors have resulted in CMOS circuits, requiring more than 10 transistors. Correspondingly, their integration though providing a faithful reproduction of neuronal behavior, is limited by the large number of transistors per cell. However, using this scheme, new AI-supporting functions such as low-energy event-detection sensors can be developed with a limited number of neurons [42].

4.2 Functional-Model Approach

The functional-model approach mimics the behavior of a part of the brain (or of the whole brain). Cognitive psychology studies the computational models of human behavior, along with the advances in anatomy show that networks of the brain (connectomes) play an essential role in brain intelligence [43]. Thus the hippocampus can be mimicked as the key to "Artificial General Intelligence" [44]. In addition, brain functions such as "episodic memory" can be modeled to speed up AI learning mimicking the past-experience memory [45]. Even the whole brain functions can be modeled as a hierarchical combination of multiple machine learning functions each mimicking a functional module of the brain [46]

Another interesting brain-mimicking example concerns "visual attention". Our eyes receive a huge stream of visual data (10^8 to 10^9 bits/second). Nevertheless, its processing is possible because the brain's visual attention function confines visual data for possible high-level cognitive processing, such as object recognition or scene interpretation. Visual attention has been modeled from psychological, neurobiological, and computational perspectives, and especially, from the anatomy and functionality of the human visual system front end. This explains how, when, and why, human select behaviorally relevant image regions. A saliency map can be realized using a neural network based on the Treisman's "Feature Integration Theory" [47], with its computational architectures already developed for possible applications [48, 49]. Based on this theoretical study, single and multiple visual attention has been implemented on silicon, as shown in Figure 1.2.12 [50, 51]. There are two types of attention, bottom-up and top-down. Bottom-up attention is mainly based on characteristics of a visual scene (stimulus-driven) whereas top-down attention (goal-driven) is related to cognitive phenomena such as memory, expectations, reward, and current goals [52]. Another important type of attention illustrated in Figure 12 is spatio-temporal attention which depends on the changes due to egocentric movements or dynamics of the surrounding world [53]. Recently, concepts of visual attention, such as spatial [54], channel-wise [55], and zoom-in [56], have been applied to CNNs to improve performance for object classification. Furthermore, residual connections are added to various attention algorithms to improve accuracy [57, 58].

Brain-mimicking AI can be realized, not simply with machine-learning modules (including DNNs) to reproduce the local functional blocks of the brain, but also using cognitive architecture based on the neurobiological connectome and knowledge from the cognitive behavioral experiments.

5. Conclusions

Most of DNN researches have been associated with software solutions, with recent proposals for improved DNN processors. However, these designs are inappropriate for mobile applications and often unsuited for silicon-chip implementations. As well, for human-computer interaction, social intelligence measures (such as emotion recognition and gesture recognition) must be available with low-power consumption.

For true progress in AI, DNNs and their hardware architectures must be examined from the point-of-view of silicon-chip implementation, especially, for edge and mobile applications. In such applications, energy efficiency is very important, and a holistic optimization must be undertaken involving, neural network architectures, hardware architectures, and even circuits. In addition, to overcome deficiencies of high energy efficiency silicon DNN, the reconfigurability of hardware is necessary. For this, both neural network hardware architecture and bit-precision must be re-configurable at both fine grain and coarse grain levels, for more versatility and higher energy efficiency.

Learning is the most important attribute of AI. Until now, all learning is done primarily on high-powered servers at remote data centers. But, as energy costs and latency force intelligent functions to move to the edge and mobile devices (as "on-device AI"), local learning provides personalization and privacy of learning functions. Moreover, reinforcement learning will be used for fine tuning of intelligent functions needed for local and very detail data. The humanistic intelligence factors such as emotion sharing and intent recognition also require local learning, and low power, making high-performance learning processors mandatory.

The brain can be mimicked either structurally (as an architecture) or functionally (behaviorally). Some neuromorphic approaches try to mimic the spiking signals of synapses and neurons. Such a synapse can be realized with memory, typically SRAM. However, in order to reduce area and power consumption, non-volatile memories such as RRAM (Memristor), PC-RAM, and MRAM are being actively studied. Neuronal behavior can be modeled with analog CMOS circuits using as many different pulse shapes as needed for biological neurons for information transfer and processing. Some designs try simply to understand biological behavior of the brain based on neuromorphic VLSI, while others are used simply for a particular application such as in low-energy event-detection camera. Another interesting approach is to mimic the functionality of the brain based on a computational model combining knowledge of cognitive science and "connectome". An interesting and active topic is the visual attention of the brain. The computational model of the brain's visual attention is used to select only the salient region of the image to speed up processing with the limited computational resources of the chip. Bottom-up attention, top-down attention, and spatio-temporal attentions have been explained.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

The brain continues to inspire our search for Machine Intelligence. Correspondingly, AI and DNN research will likely make more progress by utilizing brain-mimicking.

References

[1] N. P. Jouppi et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit," ISCA, Vol. 17 1–12, 2017.

[2] NVIDIA Corporation, "Tesla V100 Data Center GPU," 2017. Available at: https://www.nvidia.com/en-us/data-center/tesla-v100/

[3] MIT Technology Review, "On-Device Processing and AI Go Hand-in-Hand," 2018. Available at: https://www.technologyreview.com/s/610421/on-device-processing-and-ai-go-hand-in-hand/

[4] S. M. Tam et al., "SkyLake-SP: A 14nm 28-Core Xeon® Processor," ISSCC, pp. 34-36, 2018.

[5] X. Zhang et al., "Shufflenet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," arXiv, 1707.01083, 2017.

[6] V. Sze et al., Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," Proc. of the IEEE, Vol. 105, No. 12, pp. 2295-2329, 2017.

[7] Y. Chen et al., "DaDianNao: A Machine-Learning Supercomputer," MICRO, pp. 609-622, 2014.

[8] A. Parashar et al., "SCNN: An Accelerator for Compressed-sparse Convolutional Neural Networks," ISCA, Vol. 45, No. 2, pp. 27-40, 2017.

[9] S. Venkataramani et al., "SCALEDEEP: A Scalable Compute Architecture for Learning and Evaluating Deep Networks," ISCA, pp. 13-26, 2017.

[10] S. Han, "Accelerating Inference at the Edge," Hotchips tutorial, 2018

[11] A. Howard et al., "Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv, 1704.04861, 2017.

[12] M. Sandler et al., "Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation," arXiv, 1801.04381, 2018.

[13] F. N. Iandola et al., "Squeezenet: Alexnet-Level Accuracy with 50x Fewer Parameters and <0.5MB Model Size," arXiv, 1602.07360, 2016.

[14] K. Simonyan et al., "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv, 1409.1556, 2015.

[15] H. J. Yoo, "Mobile/Embedded DNN and AI SoCs," Symp. on VLSI Technology and Circuits, Short Couse, June 2017.

[16] J. Lee et al., "A 21mW Low-Power Recurrent Neural Network Accelerator with Quantization Tables for Embedded Deep Learning Applications," A-SSCC, pp. 237-240, 2017.

[17] K. Bong et al., "A 0.62mW Ultra-Low-Power Convolutional-Neural-Network Face-Recognition Processor and a CIS Integrated with Always-On Haar-Like face Detector," ISSCC, pp. 248-249, 2017.

[18] M. Courbariaux et al., "Binaryconnect: Training Deep Neural Networks with Binary Weights During Propagations," arXiv, 1511.00363, 2015.

[19] D. Shin et al., "DNPU: An 8.1TOPS/W Reconfigurable CNN-RNN Processor for General-Purpose Deep Neural Networks," ISSCC, pp. 240-241, 2017.

[20] S. Yin et al., "A 1.06-to-5.09 TOPS/W Reconfigurable Hybrid-Neural-Network Processor for Deep Learning Applications," Symp. on VLSI Circuits, pp. 26-27, 2017.

[21] K. Ueyoshi et al, "QUEST: A 7.49 TOPS multi-purpose log-quantized inference engine stacked on 96 MB 3D SRAM using inductive coupling technology in 40 nm CMOS", ISSCC, pp.216-218, 2018.

[22] J. Lee et al., "UNPU: A 50.6TOPS/W Unified Deep Neural Network Accelerator with 1b-to-16b Fully-Variable Weight Bit-Precision," ISSCC, pp. 218–219, 2018.

[23] P. S. Chandakkar, Y. Li, P. L. K. Ding and B. Li, "Strategies for Re-Training a Pruned Neural Network in an Edge Computing Paradigm," IEEE EDGE, pp. 244-247, 2017.

[24] J. Konecny et al., "Federated learning: Strategies for Improving Communication Efficiency," arXiv, 1610.05492, 2016.

[25] D. Han et al., "A 141.4 mW Low-Power Online Deep Neural Network Training Processor for Real-time Object Tracking in Mobile Devices," ISCAS, pp. 1-5, 2018.

[26] J. Lee et al., " LNPU: A 25.3TFLOPS/W Sparse Deep-Neural-Network Learning Processor with Fine-Grained Mixed Precision of FP8-FP16," ISSCC, Paper7.7, 2019.

[27] C. Kim et al., "A 2.1 TFLOPS/W Mobile Deep RL Accelerator with Transposable PE Array and Experience Compression," ISSCC, Paper 7.4, 2019

[28] iMotions, A. S. (2015). Affectiva iMotions Biometric Research Platform. from https://imotions.com/affectiva/

[29] S. Kang et al, "B-Face: 0.2 mW CNN-Based Face Recognition Processor with Face Alignment for Mobile User Identification" Symp. On VLSI Circuits, pp. 137-138, 2018.

[30] F. Akopyan et al., "TrueNorth: Design and Tool Flow of a 65mW 1 Million Neuron Programmable Neurosynaptic Chip," IEEE TCAD, Vol. 34, No. 10, pp. 1537–1557, 2015.

[31] M. Davies et al., "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," IEEE Micro, Vol. 38, No. 1, pp. 82-99, 2018.

[32] P. Yao et al., "Face Classification Using Electronic Synapses," Nature Communications, Vol. 8, p. 15199, 2017.

[33] S. Park et al., "Electronic System with Memristive Synapses for Pattern Recognition," Scientific reports, Vol. 5, p. 10 123, 2015.

[34] W. Chen et al., "A 65nm 1Mb Nonvolatile Computing-in-Memory ReRAM Macro with Sub-16ns Multiply-and-Accumulate for Binary DNN AI Edge Processors," ISSCC, pp. 494-496, 2018.

[35] S. Choudhary et al., "Silicon Neurons that Compute," ICANN, pp. 121–128, 2012.

[36] G. Indiveri et al., "ReRAM-Based Neuromorphic Computing," Resistive Switching: From Fundamentals of Nanoionic Redox Processes to Memristive Device Applications, pp. 715-730, 2016.

[37] R. Serrano-Gotarredona et al., "CAVIAR: A 45k Neuron, 5M Synapse, 12G Connects/s AER Hardware Sensory–Processing–Learning–Actuating System for High-Speed Visual Object Recognition and Tracking," IEEE Trans. on Neural Networks, Vol. 20, No. 9, pp. 1417-1438, 2009.

[38] S. B. Furber et al., "The SpiNNaker Project," Proc. of the IEEE, Vol. 102, No. 5, pp. 652-665, 2014.

[39] S. Joshi et al., "Scalable Event Routing in Hierarchical Neural Array Architecture with Global Synaptic Connectivity," IEEE CNNA, pp. 1-6, 2010.

[40] B. V. Benjamin et al., "Neurogrid: A Mixed-Analog-Digital Multichip System for Large-Scale Neural Simulations," Proc. of the IEEE, Vol. 102, No. 5, pp. 699-716, 2014.

[41] FACETS project, "Fast Analog Computing with Emergent Transient States," Available at: http://www.facets-project.org

[42] S. C. Liu et al., "Event-Based Neuromorphic Systems," New York, NY, USA: Wiley, 2014.

[43] H. S. Seung, "Reading the Book of Memory: Sparse Sampling Versus Dense Mapping of Connectomes," Neuron 62 (1), 17-29, 2009.

[44] K. L. Stachenfeld et al., "The Hippocampus as a Predictive Map," Nature Neuroscience, Vol. 20, No. 11, .pp. 1643, 2017.

[45] D. Hassabis et al., "Neuroscience-Inspired Artificial Intelligence," Neuron, Vol. 95, No. 2, pp. 245–258, 2017.

[46] N. Arakawa et al., "The Whole Brain Architecture Initiative," ICONIP, pp. 316-323, 2016.

[47] S. Lee et al., "Familiarity Based Unified Visual Attention Model for Fast and Robust Object Recognition," Pattern Recognition, Vol. 43, pp. 1116–1128, 2010.

[48] A. Treisman, G. Gelade,"A Feature-Integration Theory of Attention," Cognitive Psychology, Vol. 12, No. 1, pp. 97–136, 1980.

[49] L. Itti et al. "A Model of Saliency Based Visual Attention for Rapid Scene Analysis," IEEE TPAMI, 20(11):1254–1259, 1998.

[50] K. Kim et al., "A 125GOPS 583mW Network-on-Chip Based Parallel Processor with Bio-inspired Visual-Attention Engine," ISSCC, pp. 308-615, 2008.

[51] J. Kim et al., "An Attention Controlled Multi-Core Architecture for Energy Efficient Object Recognition," Signal Processing: Image Communication, Vol. 25, No. 5, pp. 363-376, 2010.

[52] S. Lee et al., "A 345mW Heterogeneous Many-Core Processor with an Intelligent Inference Engine for Robust Object Recognition," IEEE JSSC, Vol. 46, No. 1, pp. 42–51, 2011.

[53] J. Oh et al., "A 320mW 342GOPS Real-Time Moving Object Recognition Processor for HD 720p Video Streams," ISSCC, pp. 220–224, 2012.

[54] M. Jaderberg et al., "Spatial Transformer Networks," NIPS, pp. 2017-2025, 2015.

[55] B. Zhou et al., "Learning Deep Features for Discriminative Localization," CVPR, pp. 2921-2929, 2016.

[56] J. Fu et al., "Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition," CVPR, pp. 4476-4484, 2017.

[57] F. Wang et al., "Residual Attention Network for Image Classification," CVPR, pp. 6450-6458, 2017.

[58] L. Chen et al., "SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning," CVPR, pp. 6298-6306, 2017.

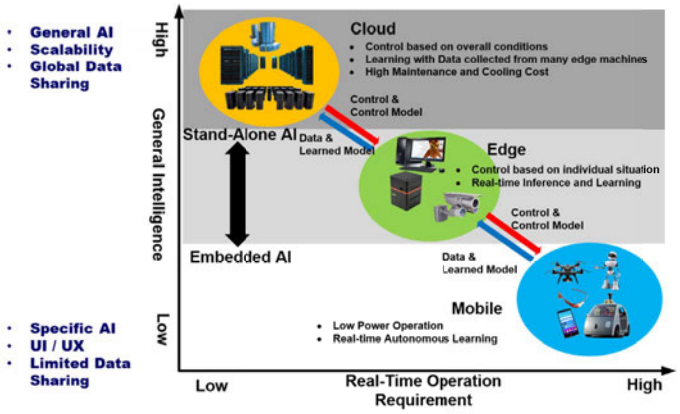


Figure 1.2.1: AI Applications.

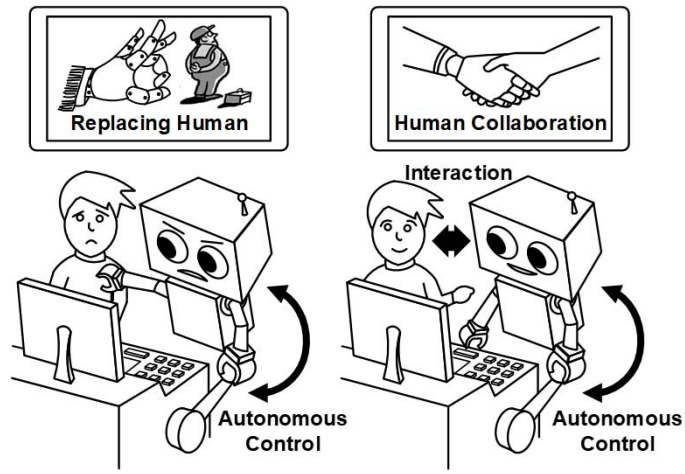


Figure 1.2.2: Application Types of AI systems.

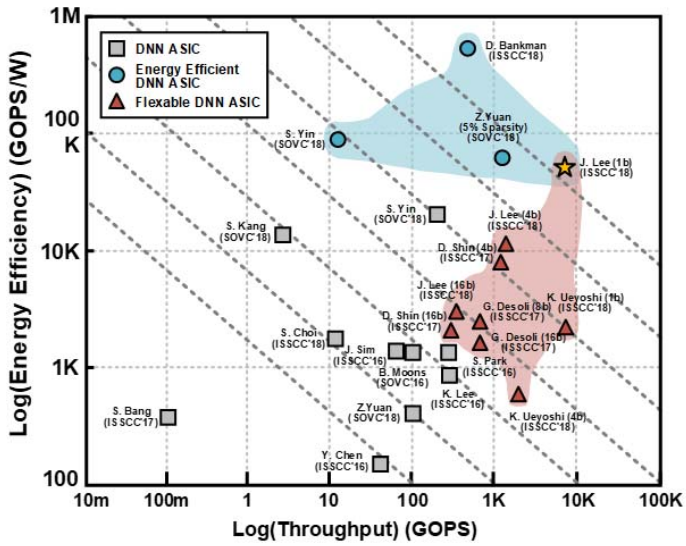


Figure 1.2.3: Energy efficiency of reported DNN chips.

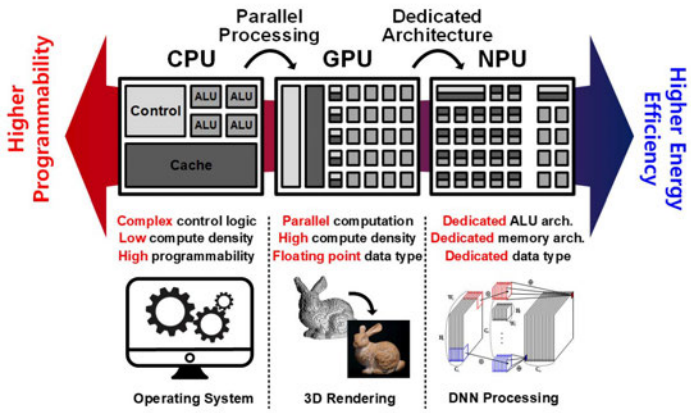


Figure 1.2.4: Comparison of processors: CPU, GPU, and NPU.

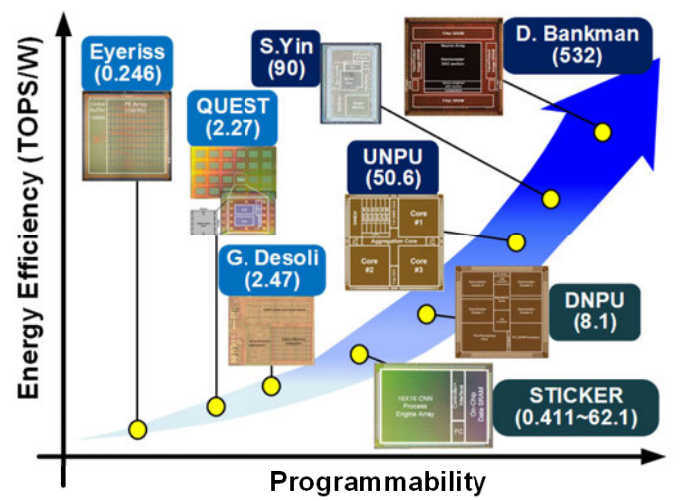


Figure 1.2.5(a): Energy Efficiency and Programmability of DNN Solutions.

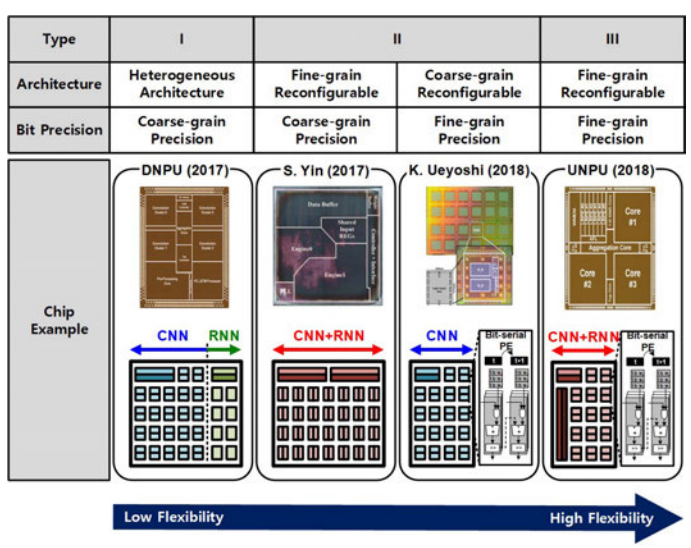


Figure 1.2.5(b): Reconfigurability and Precision variation of DNN Solutions.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

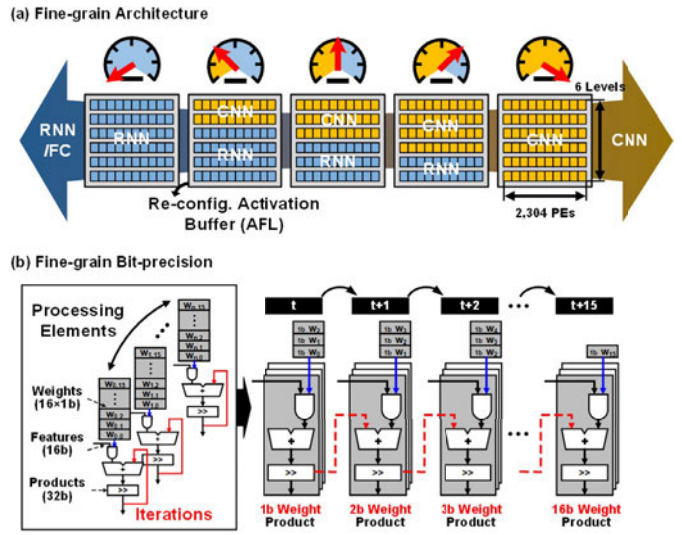


Figure 1.2.6: Fine-Grain Dynamic Reconfigurable DNN Processor with Fine-Grain Precision Variation.

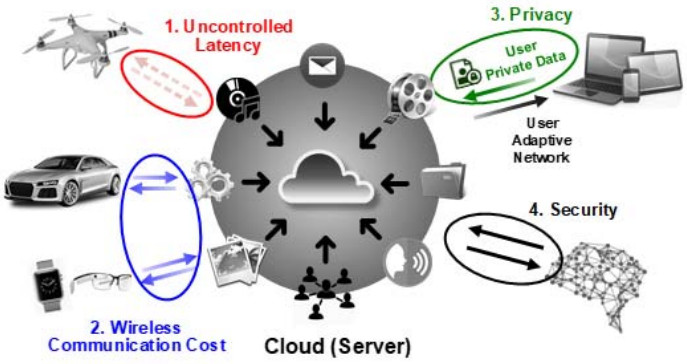


Figure 1.2.7(a): Server-Centered DNN Training Problems.

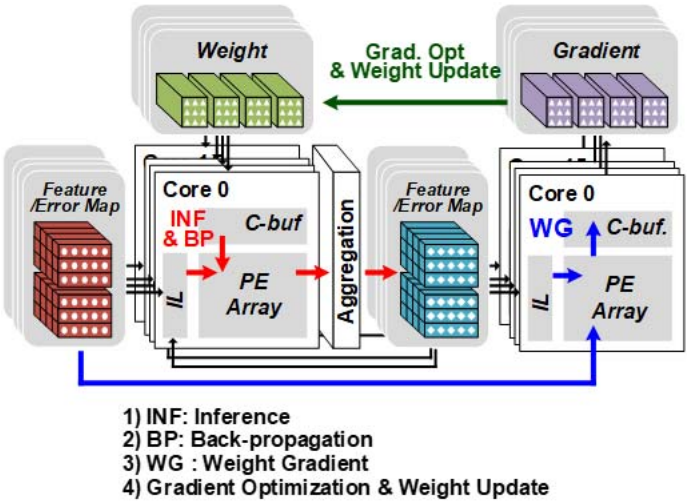


Figure 1.2.7(b): Architecture for On-Chip Learning.

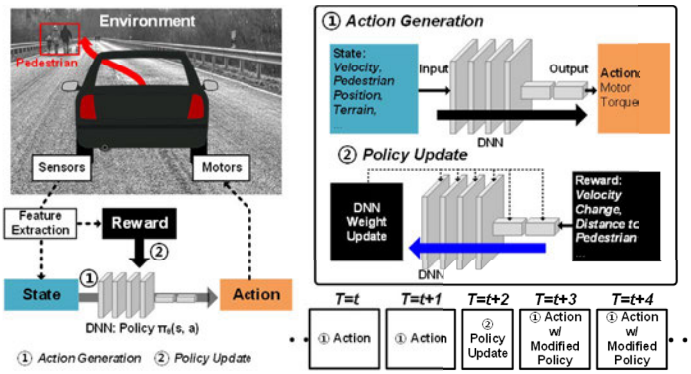


Figure 1.2.8: Online Learning for Deep Reinforcement Learning.

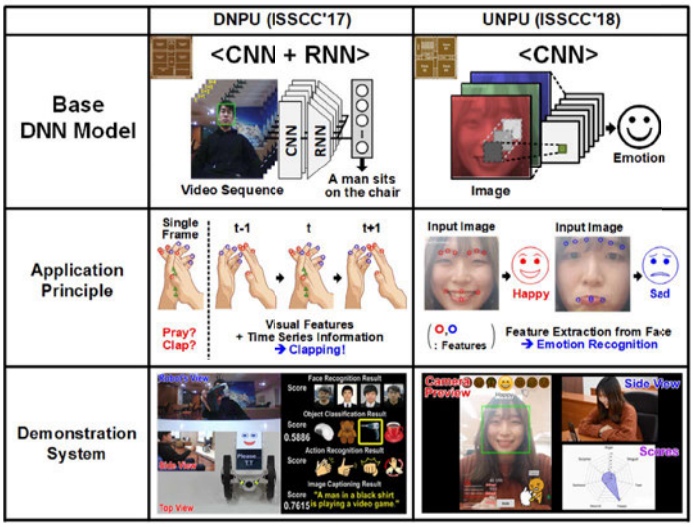


Figure 1.2.9: Gesture Recognition and Emotion Recognition.

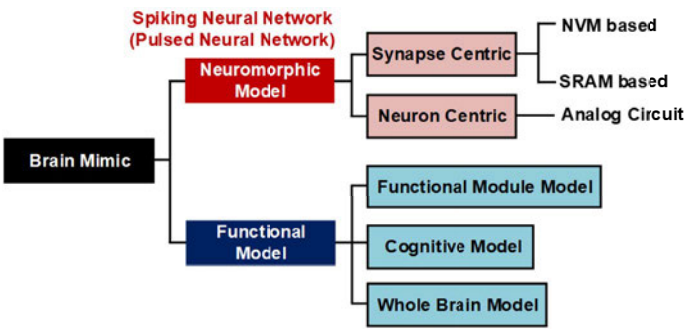


Figure 1.2.10: Types of Brain-Mimicking Approaches.

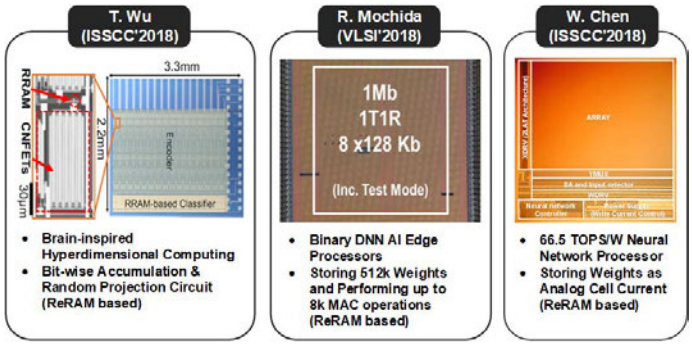


Figure 1.2.11: DNN Accelerator based on NVM PIM architectures.

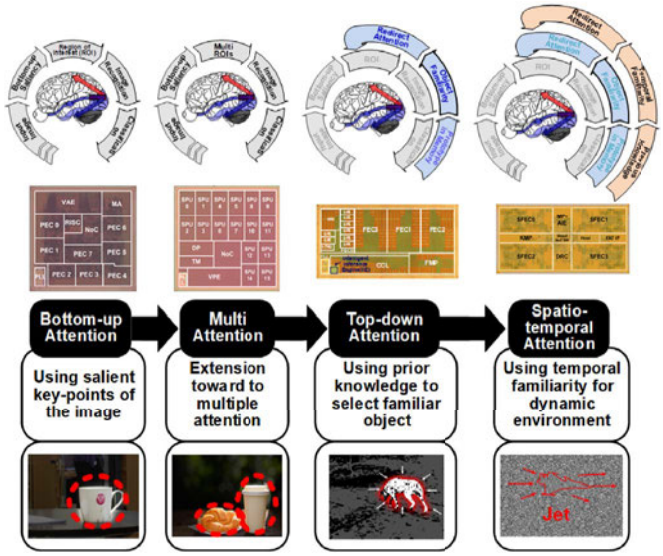


Figure 1.2.12: AI-SoCs based on Brain-Mimicking Visual Attention.

INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS
ISSCC
UNIVERSITY OF PENNSYLVANIA

ISSCC AWARDS

2018 Lewis Winner Award for Outstanding Paper

"A Fully Integrated Scalable W-Band Phased-Array Module with Integrated Antennas, Self-Alignment and Self-Test"

Shahriar Shahramian[1], Mike Holyoak[2], Amit Singh[1], Bahar Jalali Farahani[1], Yves Baeyens[1]

[1]Bell Laboratories, New Providence, NJ
[2]LGS Innovations, Florham Park, NJ

2018 Anantha P. Chandrakasan Distinguished-Technical-Paper Award

"A 16Gb 18Gb/s/pin GDDR6 DRAM with Per-Bit Trainable Single-Ended DFE and PLL-Less Clocking"

Young-Ju Kim, Hye-Jung Kwon, Su-Yeon Doo, Yoon-Joo Eom, Young-Sik Kim, Min-Su Ahn, Yong-Hun Kim, Sang-Hoon Jung, Sung-Geun Do, Chang-Yong Lee, Jae-Sung Kim, Dong-Seok Kang, Kyung-Bae Park, Jung-Bum Shin, Jong-Ho Lee, Seung-Hoon Oh, Sang-Yong Lee, Ji-Hak Yu, Ji-Suk Kwon, Ki-Hun Yu, Chul-Hee Jeon, Sang-Sun Kim, Min-Woo Won, Gun-hee Cho, Hyun-Soo Park, Hyung-Kyu Kim, Jeong-Woo Lee, Seung-Hyun Cho, Keon-Woo Park, Jae-Koo Park, Yong-Jae Lee, YongJun Kim, Young-Hun Seo, Beob-Rae Cho, Chang-Ho Shin, Chan-Yong Lee, YoungSeok Lee, Yoon-Gue Song, Sam-Young Bang, YounSik Park, Seouk-Kyu Choi, Byeong-Cheol Kim, Gong-Heum Han, Seung-Jun Bae, Hyuk-Jun Kwon, Jung-Hwan Choi, Young-Soo Sohn, Kwang-Il Park, Seong-Jin Jang

Samsung Electronics, Hwaseong, Korea

2018 Jan Van Vessem Award for Outstanding European Paper

"A 128-Pixel 0.56THz Sensing Array for Real-Time Near-Field Imaging in 0.13µm SiGe BiCMOS"

Philipp Hillger[1], Ritesh Jain[1], Janusz Grzyb[1], Laven Mavarani[1], Bernd Heinemann[2], Gaëtan Mac Grogan[3], Patrick Mounaix[4], Thomas Zimmer[5], Ullrich R. Pfeiffer[1]

[1]Universität Wuppertal, Wuppertal, Germany; [2]IHP, Frankfurt (Oder), Germany
[3]Institut Bergonié, Bordeaux, France; [4]CNRS, Talence, France
[5]Université Bordeaux, Talence, France

2018 Takuo Sugano Award for Outstanding Far-East Paper

"A PUF Scheme Using Competing Oxide Rupture with Bit Error Rate Approaching Zero"

Meng-Yi Wu, Tsao-Hsin Yang, Lun-Chun Chen, Chi-Chang Lin, Hao-Chun Hu, Fang-Ying Su, Chih-Min Wang, James Po-Hao Huang, Hsin-Ming Chen, Chris Chun-Hung Lu, Evans Ching-Sung Yang, Rick Shih-Jye Shen

eMemory, Hsinchu, Taiwan

2018 ISSCC Award for Outstanding Forum Presenter

"Energy Efficient Nyquist-Rate ADCs"

Klaas Bult

Analog Design Consult B.V., Bosch en Duin, The Netherlands

2018 ISSCC Award for Outstanding Forum Presenter

"RF, mm-Wave and Fiber-Optics Design in FDSOI CMOS Technologies"

Sorin P. Voinigescu

University of Toronto, Toronto, Canada

2018 Evening Session Award

"Lessons Learned – Great Circuits That Didn't Work – (Oops, If Only I Had Known!)"

Organizers: **Phillip Restle,** IBM T. J. Watson Research Center, Yorktown Heights, NY
Kostas Doris, NXP, Eindhoven, The Netherlands
Vivek De, Intel, Hillsboro, Oregon
Paul Ferguson, Analog Devices, Wilmington, MA

Moderator: **Tom Lee,** Stanford University, Stanford, CA

Panelists: **Bram Nauta,** *University of Twente, Enschede, The Netherlands*
Nicky Lu, *Etron Technology, Hsinchu, Taiwan*
Shanthi Pavan, *Institute of Technology, Madras, Chennai, India*
David J. Allstot, *University of California, Berkeley, CA*
Chris Mangelsdorf, *Analog Devices, San Diego, CA*
Jon Strange, *MediaTek Wireless, West Malling, United Kingdom*

2018 Demonstration Session Certificate of Recognition

"A 1/4-inch 3.9Mpixel Low-Power Event-Driven Back-Illuminated Stacked CMOS Image Sensor"

Oichi Kumagai[1], Atsumi Niwa[1], Katsuhiko Hanzawa[2], Hidetaka Kato[1], Shinichiro Futami[1], Toshio Ohyama[1], Tsutomu Imoto[1], Masahiko Nakamizo[1], Hirotaka Murakami[2], Tatsuki Nishino[1], Anas Bostamam[1], Takahiro Iinuma[1], Naoki Kuzuya[1], Kensuke Hatsukawa[3], Frederick Brady[2], William Bidermann[2], Toshifumi Wakano[1], Takashi Nagano[1], Hayato Wakabayashi[2], Yoshikazu Nitta[1]

[1]Sony Semiconductor Solutions, Atsugi, Japan
[2]Sony Electronics, San Jose, CA
[3]Sony LSI Design, Atsugi, Japan

2018 Demonstration Session Certificate of Recognition

"A Multimode 76-to-81GHz Automotive Radar Transceiver with Autonomous Monitoring"

Brian P. Ginsburg[1], Karthik Subburaj[2], Sreekiran Samala[1], Karthik Ramasubramanian[2], Jasbir Singh[2], Sumeer Bhatara[2], Sriram Murali[2], Dan Breen[1], Meysam Moallem[1], Krishnanshu Dandu[1], Saket Jalan[2,3], Neeraj Nayak[1], Rittu Sachdev[2], Indu Prathapan[2], Karan Bhatia[1], Tim Davis[1], Eunyoung Seok[1], Harikrishna Parthasarathy[2], Rohit Chatterjee[2], Venkatesh Srinivasan[1], Vito Giannini[1,4], Anil Kumar[2], Ross Kulak[1], Shankar Ram[2], Pankaj Gupta[2], Zahir Parkar[2], Sachin Bhardwaj[2], Rakesh Y.C.[2], Rajagopal K. A.[2], Arun Shrimali[2], Vijay Rentala[1]

[1]Texas Instruments, Dallas, TX
[2]Texas Instruments, Bangalore, India
[3]Intel, Bangalore, India
[4]Uhnder, Austin, TX

2018 Demonstration Session Certificate of Recognition

"UNPU: A 50.6TOPS/W Unified Deep Neural Network Accelerator with 1b-to-16b Fully-Variable Weight Bit-Precision"

Jinmook Lee, Changhyeon Kim, Sanghoon Kang, Dongjoo Shin, Sangyeob Kim, Hoi-Jun Yoo

KAIST, Daejeon, Korea

2018 Demonstration Session Certificate of Recognition

"A 9.02mW CNN-Stereo-Based Real-Time 3D Hand-Gesture Recognition Processor for Smart Mobile Devices"

Sungpill Choi, Jinsu Lee, Kyuho Lee, Hoi-Jun Yoo

KAIST, Daejeon, Korea

2019 Silkroad Award

"A 48-GHz, 5.6-mW Gate-Level-Pipelined Multiplier Using Single-Flux-Quantum Logic"

Ikki Nagaoka, Nagoya University, Nagoya, Japan

ISSCC 2018 Student-Research Preview (SRP) Poster Award

"A 4.54uW, 0.77pJ/c.s. Dual Quantization-Based Capacitance-to-Digital Converter for Acceleration, Humidity and Pressure Sensing in 0.18μm CMOS"

Sujin Park, KAIST, Daejeon, Korea

ISSCC 2018 Student-Research Preview (SRP) Poster Award
(Honorable Mention)

"A 2.26μW 15fps CMOS Vision Sensor with Mixed-Signal Cooperative Fast/Slow Motion Sensing and Region-of-Interest Imaging"

Xiaopeng Zhong
Hong Kong University of Science and Technology, Hong Kong

ISSCC 2018 Student-Research Preview (SRP) Poster Award
(Honorable Mention)

"A 320uV-Output Ripple and 100ns-Settling Time at 0.5V Supply Digital-Low-Dropout Regulator Using Multi-Level Gate Voltages and a Fast-Decision PD Detector"

Younghyun Lim
Ulsan National Institute of Science and Technology (UNIST), Korea

ISSCC 2018 Student-Research Preview (SRP) Poster Award
(Honorable Mention)

"A Coin-Battery-Powered Regulation-Free 2.4-GHz Bluetooth Low-Energy Transmitter"

Xingqiang Peng, University of Macau, Macau

IEEE SOLID-STATE CIRCUITS SOCIETY AWARDS

2017 Journal of Solid-State Circuits Best Paper Award

"A 28-GHz 32-Element TRX Phased-Array IC With Concurrent Dual-Polarized Operation and Orthogonal Phase and Gain Control for 5G Communications"

Bodhisatwa Sadhu[1], Yahya Tousi[1], Joakim Hallin[2], Stefan Sahl[4], Scott K. Reynolds[1], Örjan Renström[3], Kristoffer Sjögren[2], Olov Haapalahti[3], Nadav Mazor[4], Bo Bokinge[3], Gustaf Weibull[2], Håkan Bengtsson[3], Anders Carlinger[3], Eric Westesson[5], Jan-Erik Thillberg[3], Leonard Rexberg[3], Mark Yeck[1], Xiaoxiong Gu[1], Mark Ferriss, Duixian Liu, Daniel Friedman[1], Alberto Valdes-Garcia[1]

[1]IBM T. J. Watson Research Center, Yorktown Heights, NY, [2]Ericsson, Lindholmen, Sweden; [3]Ericsson, Kista, Sweden, [4]IBM Research, Haifa, Israel; [5]Ericsson, Lund, Sweden

IEEE Technical Field Awards

2019 IEEE Donald O. Pederson Award in Solid-State Circuits

Laurence Nagel
Omega Enterprises Consulting

"For the development and demonstration of SPICE as a tool to design and optimize electronic circuit."

2019 IEEE Fellows

Elad Alon, University of California at Berkeley
"For contributions to mixed-signal integrated circuit design and Methodology"

Lucien Breems, NXP Semiconductors
"For contributions to over-sampled converters using complementary metal-oxide-semiconductor technology"

Meng-fan Chang, National Tsing Hua University
"For contributions to static and nonvolatile memories for embedded Systems"

Christian Enz, Ecole Polytechnique Fédérale de Lausanne
"For contributions to low-power analog circuit design"

Maysam Ghovanloo, Georgia Institute of Technology
"For contributions to implantable wireless integrated circuits and Systems"

Hossein Hashemi, University of Southern California
"For development of radio-frequency and optical phased-array integrated circuits"

Nitin Jain, Anokiwave
"For leadership in the development of physics-based models for mm-wave System-on-Chip ICs"

Chris Hyung-Il Kim, University of Minnesota
"For contributions to on-chip circuit reliability evaluation and Characterization"

Lee-Sup Kim, KAIST
"For contributions to energy-efficient multimedia processor Architectures"

Farinaz Koushanfar, University of California, San Diego
"For contributions to hardware and embedded systems security and to privacy-preserving computing"

Pui-in Mak, University of Macau
"For contributions to radio-frequency and analog circuits"

Katsufumi Nakamura, Analog Devices
"For contributions to integrated circuits for digital imaging"

Samar Saha, Prospicient Devices
"For contributions to compact modeling of silicon field-effect transistors"

Shaojun Wei, Tsinghua University
"For leadership in integrated circuits engineering of smart cards and reconfigurable devices"

Jared Zerbe, Apple
"For contributions to the development of high performance serial interfaces"

Lin Zhong, Rice University
"For contributions to the development of energy-efficient driver circuits for organic light-emitting diodes"

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

1.3 Integration of Photonics and Electronics

Meint Smit,
with Kevin Williams, Jos van der Tol
Institute for Photonic Integration (formerly COBRA Institute),
Eindhoven University of Technology, Eindhoven, The Netherlands

Preamble

The market for photonic integrated circuits (PICs) is rapidly growing. Photonic integration which is now the dominant technology in high-bandwidth and long-distance telecommunications is increasingly applied to shorter distances within data centers. Now, it is set to become also dominant in many other fields: PICs offer compelling performance advances in terms of precision, bandwidth, and energy efficiency. To enable uptake in new sectors, the availability of highly standardized (generic) photonic-integration-platform technologies is of key importance, as this separates design from technology, reducing barriers for new entrants. Another major challenge is low-cost energy-efficient integration of photonics with the electronic circuitry that is used for driving and controlling the photonic IC and processing its information. Today, the major platform technologies are indium phosphide (InP)-based monolithic integration and silicon (Si)-based photonics. InP technology offers integration of the full suite of photonic components, including lasers, optical amplifiers, and high-performance modulators. While Si photonics offers better compatibility with CMOS process facilities, it lacks the most important photonic building blocks: lasers and optical amplifiers. In this paper, we describe the current status and directions for future developments of InP-based generic integration, and we compare the potential of InP photonics and Si photonics for integration with controlling electronics. In what follows, we will focus in Section 1 on similarities and differences between InP and Si photonics. In Section 2, we will give a concise overview of the present status of this technology and how it compares with Silicon photonics. In sections 3 and 4 we will discuss membrane-based technologies which support efficient integration with electronics.

1.0 Introduction: Similarities and Differences Between Photonic and Electronic Integration

From the beginning, in the first publications on integrated photonics (known then as "integrated optics"), it was assumed that photonic integration would follow the same path as microelectronic integration [1, 2]. Indeed, at this early stage, we see a lot of similarities. Figure 1.3.1 shows a clear exponential increase of photonic-chip complexity (measured as the number of components integrated on a single chip), similar to Moore's Law in electronics [3].

Today, the most sophisticated circuits including lasers, modulators, detectors, and multiplexers have more than 1,500 components on one chip [4]. Currently, the number of components per chip is constrained by electrical connectivity and thermal management. For the future, membrane technologies, addressed later in this paper, offer a route to the size and energy reductions required for higher-density integration.

1.1 Semiconductor Integration Technology

Figure 1.3.2 highlights some of the milestones which have been reached in both electronic and photonic integration. Their development follows largely the same path, with a delay of about 30 years for photonics. The starting point for microelectronic integration was the invention of the transistor [5, 6], a compact semiconductor amplifier which replaced the bulky vacuum tube. For photonic integration it was the invention of the semiconductor laser, which replaced the bulky gas and doped-glass solid-state lasers [7]. In the early years both the transistor and the semiconductor laser were used as discrete components. In 1958 Jack Kilby succeeded to integrate a circuit containing several transistors in a silicon substrate [8]. Later, other technologies were explored, but in the eighties CMOS became dominant. In photonics a first integrated circuit consisting of a laser integrated with a modulator was reported in 1987 [9].

1.2 Generic Integration Technologies

In the early years of both microelectronic and photonic integration design and technology development were closely connected, and chip designs were technology specific. With increasing circuit complexity, this close connection between design and technology became increasingly difficult. Correspondingly, an important step in the development of semiconductor integration technology was the introduction of generic integration processes, technologies that allowed for decoupling of design and technology by offering chip designers a small set of well-defined standardized building blocks with which they could design a broad range of application-specific circuits. This approach to microelectronics, first introduced by Carver Mead and Lynn Conway [10], caused a revolution in IC-design and opened the way to CMOS-VLSI. An important advantage of designing with standardized building blocks is that they enable reuse of components and sub-circuits that are made available to designers through a so-called process design kit (PDK). Another advantage is that a number of designs from various designers can be combined on a so-called multi-project wafer (MPW), which leads to a large reduction in the costs of prototyping. In photonics, this approach was pioneered by the COBRA research institute, now known as the Institute for Photonic Integration, at TU Eindhoven [11]. Later, in 2008, it was adopted by the ePIXnet Network of Excellence [12], which brought Europe the world's first photonic MPW runs in indium phosphide (InP) and silicon photonics.

Today, InP-based monolithic integration and Si photonics are the two major integration technologies in photonics. In both of these access to MPW runs is currently offered by industrial and semi-industrial foundries. The advantage of Si Photonics is that it is produced in a CMOS foundry, which provides a well-controlled and rapidly scalable fabrication environment. As the substrates for Si are significantly larger than those for InP, the potential for scaling of Si to high volumes seems superior. But, Si photonics is lacking light sources and amplifiers which limits its potential for large scale integration. To resolve this, work on integration of InP-based light sources and amplifiers on silicon substrates is underway, but formidable manufacturing challenges remain. Overall, InP-based monolithic integration currently offers the most comprehensive range of photonic functionality with high-energy-efficiency quantum-well lasers and modulators, as well as optical amplifiers, detectors, and a range of passive components for creating interferometers, combiners, and modulators. An important milestone on the road to mature integration technology is the reduction of killer-defect density to a level < 1 /cm^2, for silicon electronics this milestone was reached in 1987, and for InP photonics 23 years later, in 2010 [13].

1.3 Differences Between Electronic and Photonic Integration Technology

Although microelectronic and photonic integration show a lot of similarities, there are also marked differences. Photonic building blocks are larger and active building blocks, such as optical amplifiers and modulators, operate at much higher power levels than transistors. However at the sub-system level, the electronic amplifiers currently operate with three to four times more power than the photonic devices to which they are connected. The footprints for the basic active components themselves are comparable. As well, the electronic-driver areas are often dominated by passive components such as resistors, capacitors, and the input/output connections, and furthermore, photonic circuits include redundant chip area for cross-talk mitigation, waveguide bends, and electrical connections. A more appropriate basis for comparison of high-performance active photonic devices is with RF and analog electronics. There is also an important difference in terms of operational capacity and technology maturity; the higher wafer throughputs in electronic circuit manufacturing enables a faster manufacturing learning curve when introducing new technology nodes.

Nanophotonic technology offers an order-of-magnitude power-dissipation-level reduction as well as footprint reduction [14,15]. Both become critical as designs become thermally-constrained. While photonics will not scale to the same component densities seen in CMOS electronics, the application of membrane-based technologies creates a roadmap to the level of tens of thousands of components per chip, as indicated in Figure 1.3.1. This raises new challenges in terms of mechanical, thermal, and electrical design.

2.0 Generic InP-Based Integration Technology

An extensive description of InP-based generic integration technology is given in [16]. Here, we will provide a brief overview of this technology including some recent progress. Figure 1.3.3 shows a relationship between generic electronic and photonic integration technology. Electronic circuits are composed of a very small set of basic building blocks: transistors, capacitors, and resistors, and electrical interconnection lines. In photonics, we can do something similar. Light has an amplitude, a phase and a polarization. These properties can be

manipulated with an optical amplifier, a phase modulator, and a polarization converter, respectively. With a process in which we can integrate these basic building blocks with waveguides and passive components, we can realize a large variety of circuits, including pulsed and tunable lasers, transmitters and receivers, and a variety of sensor-readout circuits.

In 2008, TU Eindhoven started providing external designers with open access to its generic foundry process in the framework of the ePIXnet network of excellence, which had adopted a foundry model for photonics [12]. In the next decade, a number of large R&D projects (EuroPIC, PARADIGM, MEMPHIS) worked on transferring the foundry approach from the university to an industrial environment. Presently, two European foundries provide commercial access to InP-based generic photonic integration processes. Thus far, they are the only open-access generic InP-foundries in the world providing pre-specified building blocks in Process Design Kits. Figure 1.3.4 shows the basic building blocks which are available on these photonic integrated circuit (PIC) platforms.

2.1 Process Design Kits (PDK)

A PDK for photonic integrated circuits is very similar to those for traditional CMOS, containing information about the fabrication process, design rules, mask layers, models for simulation, and predefined components. But, due to the nature of integrated photonic components, many of the predefined components come in the form of sophisticated modules for generating the mask-layout. These are somewhat comparable to parameterized cells (PCells) as used in electronics, but more advanced as these modules can be optically defined and take material and process properties from the PDK to generate and optimize these components. Further, photonic PDKs contain templates for packaging and testing of the designed PICs. These templates define the electrical and optical input and output ports, such that they comply with standard packaging and test equipment. This adds to lowering access barriers even further to make PIC technology part of the mainstream semiconductor industry.

PDK development in both InP-based as well as other material systems such as silicon photonics started from photonic design software because of the different requirements that PICs impose on the mask-layout and simulation software (for example, with a need for smooth curves and optical simulations). Today, it can be observed that photonic design is being integrated into traditional electronic design automation (EDA) software such as tools from Mentor, Cadence, and Synopsys. These EDA vendors offer a large pool of capabilities which are also relevant for PICs. With increasing complexity of PIC designs, the need for verification with design-rule checking (DRC), and layout versus schematic (LVS) is becoming mandatory, and not all current PDA tools provide such functionality. A remaining challenge is the need to address co-simulation of electronics and photonics as is required in lasers, detectors, high-speed modulators, and modulator drivers.

2.2 Present-Day Capabilities of Photonic InP Foundries

The example shown in Figure 1.3.5, a widely tunable laser, illustrates many of the present capabilities of available InP foundry platforms. It consists of a ring resonator with a semiconductor optical amplifier (SOA), for providing gain, and three cascaded asymmetric Mach-Zehnder interferometer (AMZI) filters, for providing wavelength selection. The three Mach-Zehnder filters each consist of two so-called multi-mode interference (MMI) couplers and two electro-optic phase modulators. The filters have periodic wavelength responses with a ratio 1:2:4, which can be shifted with the phase modulators. By appropriate control of the phase shifters, the laser wavelength can be tuned over more than 70nm, as shown at the bottom right of the figure. The laser signal is coupled out of the ring with an MMI power splitter. In total, the laser consists of 15 building blocks: an SOA, 7 phase shifters, and 7 MMI couplers. Output power can be increased by including a booster output amplifier. The overall size of the laser is 1.5×3.5 mm^2; 4 of them will fit into a MPW-cell of 4×5 mm^2.

Several hundreds of different MPW cell designs have already been processed by the two InP foundries in more than 30 MPW runs, with PIC designs for telecommunications, data communications, microwave-photonic applications, fiber sensor readouts, metrology, medical diagnostics, gas sensing, and automotive applications.

These foundries provide optical amplifiers with 70 to 90cm^{-1} gain and a user defined length, various types of tunable lasers, electro-optical phase modulators for 20Gb/s operation, thermo-optic and current injection modulators, high responsivity photodetectors (with low dark currents and >30GHz bandwidth), high and low confinement waveguides (with losses < 2dB/cm), and a variety of passive devices: wavelength demultiplexers, MMI couplers and reflectors, spot-size converters, and polarization converters. A more detailed description of the building blocks available in the InP foundries in 2018 is provided in the JePPIX[1] roadmap 2018 [17].

Recent innovations include the introduction of 192nm scanner lithography for precision waveguide fabrication [18] in low-excess-loss (< 0.2dB) arrayed waveguide gratings [19]. High accuracy methods to automate in-line measurements for gain [20], loss [21], and intra-cavity reflections [22] have been devised to enable both process optimizations and statistically representative component models. High speed electroabsorption modulators have now been demonstrated, with 55GHz bandwidth [23], and 112Gb/s operation with polarization multiplexed electroabsorption modulators [24]. In the future, improvements of platform capabilities are envisaged through enhanced efficiency, precision, and bandwidth of building blocks, and introducing statistical performance data into the PDK.

3. Wafer-Scale Integration of Photonics and Electronics

As InP integrated photonic circuits increase in complexity, the wiring density between photonic and electronic components and thermal loading become more critical. Semiconductor lasers are typically a few hundreds microns long for widely tunable lasers, with the possibility of reducing to 100µm within integrated circuits [25]. Techniques to create shorter electrically pumped lasers and light sources with lower energy consumption are addressed in section 4. Eliminating non-functional area is attractive for increasing functionality, but leads to challenges in terms of optical, electronic [26], and thermal [27] crosstalk, as well as optical, electrical, and thermal connectivity. A broad range of strategies have been identified for chip and wafer scale connection, and the integration of photonic and electronic circuits. These range from the full embedding of a subset of key photonic functions into electronics, through to 3D integration, hybrid integration, and heterogeneous integration. Figure 1.3.6 illustrates a few of the approaches that are being explored.

The conceptually simplest approach from the electrical perspective is to integrate electronics and photonics without the light source in one chip (Figure 1.3.6d), with hybrid connection to an InP laser as depicted in Figures 1.3.6e and 1.3.6f. Hybrid integration is defined here as the assembly of pre-fabricated chips, involving either a sub-micron alignment requirement for chip to substrate assembly, or the use of larger chips incorporating mode adaptation. The red circles indicate the positions were accurate alignment and efficient coupling is needed. In this approach, high-speed RF connections between modulators, detectors, and drivers can be combined within the same chip. However, this does not address the challenges associated with photonic connection and impairments introduced with imperfect interfaces in the photonic circuits, and there are open questions on how scalable such approaches will be with many light sources and amplifiers per chip. 3D integration using through-silicon vias offers two dimensional grids of bond pads and a route for thousands of electrical connections, but the use of large bond pads limits component density on the photonic side, and potentially on the electronic side, as well. Many of the proposed solutions using Si waveguides use dielectric claddings, which will obstruct heat removal.

Heterogeneous integration offers a highly scalable route to combine photonics and electronics, and can be implemented at the wafer scale. Figure 1.3.6g shows an example in which InP dies with an unprocessed-active-layer stack are bonded on top of a processed silicon photonics layer. The InP substrate is removed after bonding, and the active components (lasers or optical amplifiers) can be processed at the wafer scale. Because the processing is accurately aligned to the underlying silicon photonics layer, so are the lasers. However, vertical coupling between the active InP and the passive silicon layer introduces losses and potentially reflections, as well. Adding electronics as a third layer in this scheme increases the complexity of processing and is further down the roadmap.

An alternative approach is heterogeneous integration by top-down bonding of processed InP wafers or dies on CMOS wafers, removing the InP substrate and wafer scale processing of the electrical interconnections with vias through the

[1]JePPIX is the Joint European Platform for Photonic Integration of Components and Circuits, a consortium in which Europe's key players in InP and TriPleX-based Photonic Integration are cooperating on the establishment of an open-access foundry infrastructure (www.jeppix.eu).

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

optical and thermal-isolating bonding layer, without the need for enlarged bond pads, saving surface area. The approach is called IMOS (InP Membrane On Silicon), as depicted in Figures 1.3.6h and 1.3.6i. In this arrangement, double-sided cooling is required to transfer heat from both the photonic and the electronic layer. The difference with Figure 1.3.6g is that the photonic circuit is not partitioned between an active InP and a passive silicon layer which must be processed separately. All optical interfaces are implemented within the monolithic photonic circuit and can, therefore, exhibit low loss and low reflection. The electrical connections can be reduced to the micron-scale thickness of the InP integrated photonic membrane and the bonding layer, thus reducing parasitic capacitance and increasing energy efficiency.

Membrane-based photonic integrated circuits offer a powerful means to create a photonic plane over an electronic control plane as the substrate can be removed to reduce parasitics, enabling thermal and electrical vias, and ensuring the most intimate heatsink connections. As processing of the light source and full photonic functionality is separated from the electronics, the photonic circuits can be created without compromising the electronics. Similarly, the electronic fabrication flow is not compromised by the photonics.

Co-designed photonics and electronics is anticipated to enable a step change in both efficiency and function. The shorter connections between detectors and amplifiers enable lower parasitic loss [28] and energy efficiencies through the removal of impedance matching networks. The closer placement of electronic and photonic chips already allows for higher numbers of electronic connections and functions, such as digital to analog conversion to be performed in the optical domain [29, 30]. So far these techniques have been studied for single elements and devices, enabling connections between co-designed chips which are placed in close proximity, but larger circuits implementing on-chip multiplexing will require more electrical connections. One example is sensor and communication chips which increasingly employ multiple wavelength channels. While this is straight forward to implement from a photonic perspective, there are considerable challenges from an electronic perspective due to higher levels of control complexity and crosstalk.

Figure 1.3.9 shows a cross-section for the methods under development. A polymer layer (Benzocyclobutene (BCB)) is used for the purposes of both planarization and adhesion, but also provides optical, thermal, and electrical isolation. Creating vias for electrical signal routing and thermal management requires the removal of the substrate. This photonic-electronic integration, also known as photronic integration, uses processed PIC wafers and ASIC wafers. Si BiCMOS wafers provide mechanical support for the InP wafer membrane which is several microns in thickness.

In this approach, the full suite of generic building blocks is available, requiring no re-engineering of the photonics. Similarly no bespoke processes are required from the electronics, offering a “zero-change” integration approach. The fabrication challenges are therefore primarily at the interface, creating fine pitch vias. Design challenges address the re-engineering of co-designed circuits to take full advantage of circuit-level design freedom. With such intimate integration, we can avoid the use of transmission lines, resistive terminations, and uncontrolled inductances, for example, from bond wires, avoiding the capacitance of bondpads. The heterogeneously integrated combination of state-of-the-art industry-sourced photonics and electronics is expected to deliver considerable speed and energy improvements over co-packaged systems. Moreover, it offers the possibility of creating “smart photonic building blocks”, such as lasers and modulators which are integrated with monitor diodes and electronic control circuitry, maintaining the building blocks at the correct operating point despite variations in temperature and fabrication process parameters.

4. Silicon-Based Nanophotonic InP-Membrane Platform (IMOS)

Further innovation is feasible by moving to a nanophotonic integration platform. This enables the monolithic combination of the new functions achievable with sub-wavelength photonic components with the full range of active devices such as lasers, amplifiers, and efficient modulators and detectors. This technology evolution is captured in Figure 1.3.10 and described below.

The membranes used in the photronics approach have the same micron-scale waveguide features as used in current manufactured PIC products. However, removing the substrate creates considerably more design flexibility for future technology nodes. As can be seen from Figures 1.3.9 and 1.3.10, the photonic waveguides can now be completely clad between low index materials such as BCB, SiO_x, SiN_x, or even air. This enables far tighter confinement of light, so that the waveguides and the components based on them can become much smaller, as illustrated in Figure 1.3.10. This enables a higher integration density, and offers the potential for new forms of mode adaptation, and the introduction of precision surface gratings for coupling light in and out of the PIC.

A nanophotonic IMOS-node might become the successor of the microphotonics IMOS-node described in the previous section. It will support higher integration densities and better energy efficiency for short-range communications as is required in data centers. An important difference with the microphotonic IMOS technology is that we can no longer build on mature InP-foundry technology, as the thin (200 to 300nm) membranes require complete redevelopment of the foundry process, and all of its building blocks. At TU Eindhoven, we are working on development of a nanophotonic IMOS technology [15].

Figure 1.3.11 shows an example of a laser developed in this platform [31]. It uses short photonic crystal reflectors for creating the laser resonator. The laser shown is a few hundred microns long, but optimalization of the gain stack will allow for efficient lasers which are an order-of-magnitude smaller and significantly more efficient than current InP lasers. They are expected to produce more than 100μW output power, which is sufficient for high-speed short-range communication.

Optical wireless is a particularly promising application domain for IMOS technology, as wireless transceivers can benefit from the combination of high-performance active components, such as lasers and detectors, as well as from precision beam forming enabled by nano-photonic gratings. A novel integrated optical wireless receiver/transmitter has recently been presented [32] where the device is realized on an InP membrane platform where active and passive components are integrated monolithically (Figure 1.3.12). It can be reconfigured to either receiver mode or transmitter mode, by simple control of the operation mode of a photodetector (short SOA). In receiver mode, 17.4Gb/s orthogonal frequency division multiplexing (OFDM) signal transmission was demonstrated in an indoor optical wireless system, illustrating the potential of this concept. Array integration of high density surface emitters in combination with precision, high-speed control of light (in terms of amplitude and phase) offers opportunities for dynamic beam forming in communications, ranging, and imaging.

5. Conclusion

While photonic integration has a long history which lags that of microelectronics by thirty years, there is increasing evidence that the technology is set to follow the similar trajectory in the coming years. With the recent establishment of generic foundry technologies and the enabled multi-project wafer services, comparisons with microelectronics are both timely and instructive. The application of generic methodologies to the main platforms has allowed separation of design innovation and technology development which enables a broader uptake of the technology and diversity in circuit innovation. Continued pressure on performance motivates the development of new technologies for higher speed, higher precision, and energy efficiency. Circuits produced with wafer-scale technologies will benefit from large-scale manufacturing methods. We have highlighted promising new technologies to introduce photonic-electronic integration, and to transition to integrated nanophotonics for high density and richer functionality.

Acknowledgements

The authors acknowledge support from the NWO Research School for Integrated Nanophotonics, the EC H2020 WIPE project and the NWO Photronics project. A. Meighan, M. Spiegelberg and T. de Vries are acknowledged for the photronics experiments and the photograph of the first membrane attachment of InP to Silicon BiCMOS. V. Pogoretskyi is acknowledged for processing a first IMOS MPW wafer.

References

[1] S. E. Miller, “Integrated optics: An Introduction,” Bell Syst. Tech. J., Vol. 48, pp. 2059-2069, Sept. 1969.

[2] P.K. Tien, “Integrated Optics and New Wave Phenomena in Optical Waveguide”, Rev. of Modern Physics, Vol. 49, No. 2, pp. 361-420, 1977.

[3] M.K. Smit, J. van der Tol and M.T. Hill, "Moore's Law in Photonics", Laser & Photonics Reviews, 6, 1, pp. 1–13 (2012).

[4] J. Summers et al., Monolithic InP-Based Coherent Tansmitter Photonic Integrated Circuit with 2.25 Tbit/s Capacity, Electronics Letters Vol. 50, pp. 1150-1152, 2014.

[5] US patent 2524035 J. Bardeen et al, "Three-Electrode Circuit Element Utilizing Semiconductive Materials" oldest priority 26.02.1948.

[6] US patent 2569347 W. Shockley, "Circuit Element Utilizing Semiconductive Material" oldest priority 26.06.1948.

[7] Zh. I. Alferov, V. M. Andreev, E. L. Portnoi, M. K. Trukan, Fiz. Tekh. Poluprovodn. 3, 1328, 1969.

[8] Jack S. Kilby, https://en.wikipedia.org/wiki/Integrated_circuit.

[9] M. Suzuki, Y. Noda, H. Tanaka, S. Akiba, Y. Kushiro, H. Isshiki, “Monolithic Integration of InGaAsP/InP Distributed Feedback Laser and Electroabsorption Modulator by Vapor Phase Epitaxy”, Journal of Lightwave Technology, Volume 5, Issue 9, pp. 1277-1285, September 1987.

[10] Carver Mead and Lynn Conway, *Introduction to VLSI Systems*, Addison-Wesley Longman Publishing Co., Inc. Boston, MA, ISBN:0201043580, USA ©1979.

[11] M.K. Smit, “InP Photonic Integrated Circuits”, Proc. 15th LEOS Annual Meeting, Glasgow, paper ThX1 (invited), Nov. 2002.

[12]Towards a Foundry Model for Micro- and Nanophotonic ICs, A vision for Europe, vision document of the ePIXnet Steering Committee, March 2007.

[13] F. Kish et al., “System-on-Chip Photonic Integrated Circuits”, IEEE Journal of Selected Topics in Quantum Electronics, 24, 1, Jan.-Feb. 2018.

[14] J.J.G.M van der Tol., Jiao,Y., Shen, L., Millan-Mejia, A.J., Pogoretskiy, V., van Engelen, J.P. & Smit, M.K., “Indium Phosphide Integrated Photonics in Membranes”, IEEE Journal of Selected Topics in Quantum Electronics. 24, 1, 9 p., 6100809, 2018.

[15] J.J.G.M. van der Tol, Y. Jiao, K.A., “InP Photonic Integrated Circuits on Silicon”, *Silicon Photonics*, Lourdudoss, S., Chen, R. T., Jagadish, C. (eds.). Elsevier, p. 189-219 23 p. (Semiconductors and Semimetals; vol. 99), Sep 2018.

[16] M.K. Smit et al., An Introduction to InP-Based Generic Integration Technology”, Semiconductor Science and Technology, Volume 29, Number 8, http://iopscience.iop.org/article/10.1088/0268-1242/29/8/083001, 2014.

[17] JePPIX roadmap 2018, “The Road to a Mulit-Billion Euro Market in Integrated Photonics”, the JePPIX consortium, www.jeppix.eu/roadmap2018.

[18] L.M. Augustin, R. Santos, E. den Haan, S. Kleijn, P.J.A. Thijs, S. Latkowski, D. Zhao, W. Yao, J. Bolk, H. Ambrosius, S. Mingaleev, A. Richter, A. Bakker, T. Korthorst, “InP-Based Generic Foundry Platform for Photonic Integrated Circuits”, IEEE Journal of Selected Topics in Quantum Electronics, 24, 1, 6100210, 2018.

[19] J. Bolk, H. Ambrosius, R. Stabile, S. Latkowski, X. Leijtens, E. Bitincka, L. Augustin, D. Marsan, J. Darracq, K. Williams, '
“Deep UV Lithography pProcess in Generic InP Integration for Arrayed Waveguide Gratings”, IEEE Photonics Technology Letters, 30, 3, pp. 1222-1225, 2018.

[20] D. Pustakhod, K. Williams, X. Leijtens, “Fast and Robust Method for Measuring Semiconductor Optical Amplifier Gain”, IEEE Journal of Selected Topics in Quantum Electronics, 24, 1, 9, 8004439, 2018.

[21] D. Pustakhod, K. Williams, X.J.M. Leijtens, “Method for Polarization-Resolved Measurement of Electroabsorption”, IEEE Photonics Journal, 10(2):6600611, 2018.

[22] D. Zhao, D. Pustakhod, K. Williams, X. Leijtens, “High Resolution Optical Frequency Domain Reflectometry for Analyzing Intra-Chip Reflections”, IEEE Photonics Technology Letters. 29, 16, pp. 1379-1382 4 p., 7967818, 2017.

[23] M. Trajkovic, F. Blache, H. Debregeas, L.M. Augustin, E. den Haan, K.A. Williams, X.J.M. Leijtens, “55GHz EAM Bandwidth and Beyond in InP Active-Passive Photonic Integration Platform”, Conference on Lasers and Electro-Optics, San Jose, post-deadline session, 2018.

[24] M. Baier, F.M. Soares, Z. Zheng, M. Schell, “112Gb/s PDM-PAM4 Generation and 80 km Transmission Using a Novel Monolithically Integrated Dual-Polarization Electro-Absorption Modulator InP PIC”, European Conference on Optical Communication, 2017.

[25] K. Nakahara et al., "Direct Modulation at 56 and 50Gb/s of 1.3-μm InGaAlAs Ridge-Shaped-BH DFB Lasers", IEEE Photonics Technology Letters, 27, 5, pp. 534-536, 2015.

[26] W. Yao et al., "Performance Degradation of Integrated Optical Modulators Due to Electrical Crosstalk", Journal of Lightwave Technology, 34(13), pp. 3080-3086, 2016.

[27] G. Gilardi et al., "Observation of Dynamic Extinction Ratio and Bit Error Rate Degradation Due to Thermal Effects in Integrated Modulators", Journal of Lightwave Technology, 33, 11, pp. 2199-2205, 2015.

[28] S. Saeedi, "A 25 Gb/s 3D-Integrated CMOS/Silicon-Photonic Receiver for Low-Power High-Sensitivity Optical Communication", Journal of Lightwave Technology, 34, 12, 2924, 2016.

[29] M. Vanhoecke et al., "Segmented Optical Transmitter Comprising a CMOS Driver Array and An InP IQ-MZM for Advanced Modulation Formats", Journal of Lightwave Technology, Vol. 35, pp. 862-867, 2017.

[30] A. Aimone et al., "DAC-Free Ultra-Low-Power Dual-Polarization 64-QAM Transmission with InP IQ Segmented MZM Module", postdeadline paper Th5C.6 Proceedings Optical Fiber Communications Conference, 2016.

[31] V. Pogoretskiy, J. van der Tol , A. Higuera-Rodriguez, M. Smit, Y. Jiao, “Integrated Photonic Crystal DBR Laser in an InP Membrane Platform”, Proc. PIERS Conference, Singapore, 2017.

[32] Y.Jiao, Z. Cao, L. Shen, J. J. G. M. van der Tol, A.M.J. Koonen, “Membrane-Based Receiver/Transmitter for Reconfigurable Optical Wireless Beam-Steering Systems”, IEEE Journal of Selected Topics in Quantum Electronics. 24, 1, 6 p., 6100506, 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

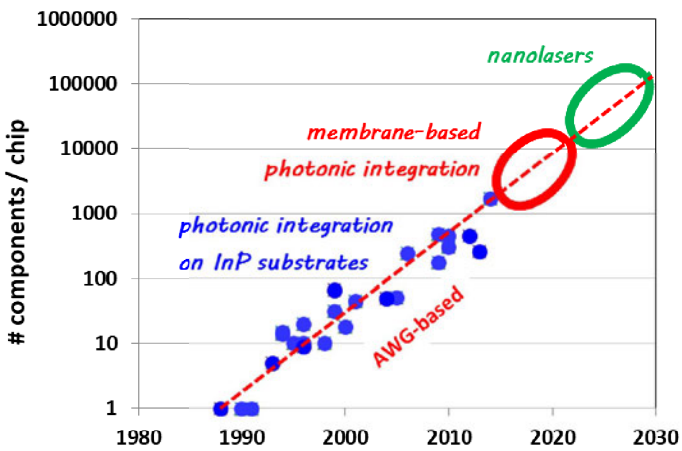


Figure 1.3.1: Moore's Law for photonics.

	Electronics	Photonics
Invention of key component (transistor/semiconductor laser)	1947	1969
Semiconductor integration technology	1958	1987
Generic integration technology (MPWs*)	1979	2008
Killer-random-defect densities reported < 1/cm^2	1987	2010

Figure 1.3.2: Important milestones in the development of electronic and photonic integration. *multi-project wafer

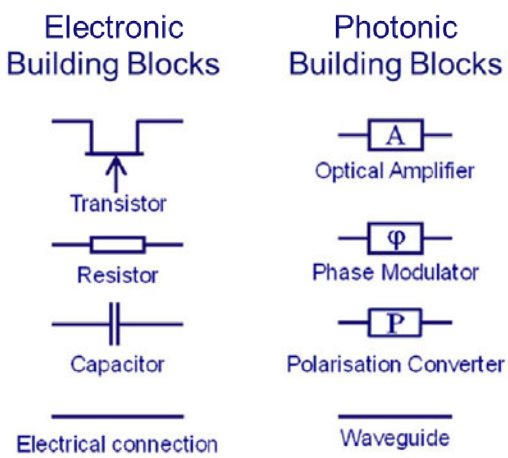


Figure 1.3.3: Basic building blocks in generic electronic and photonic integration technology.

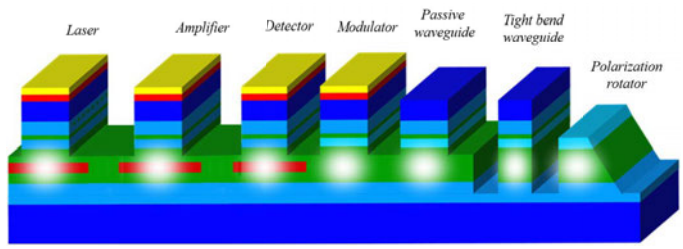


Figure 1.3.4: Building blocks on open-access InP PIC platforms.

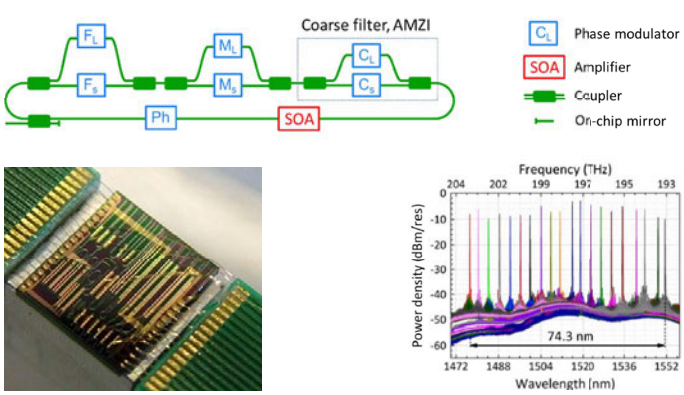


Figure 1.3.5: Circuit layout of a widely tunable laser and a photo micrograph of an array of lasers fabricated in one of the JePPIX foundry processes (@SMART Photonics).

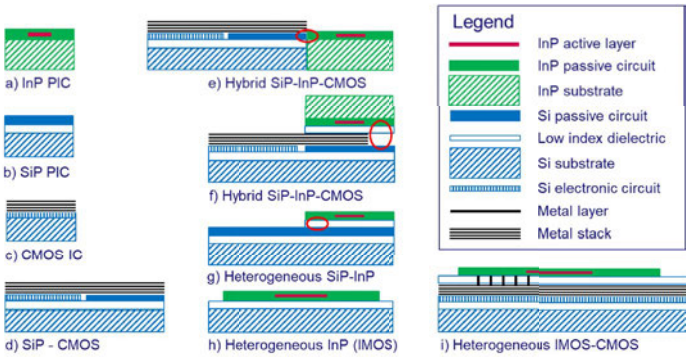


Figure 1.3.6: Examples of hybrid and heterogeneous integration of photonic and electronic ICs.

	Optical connections	Electronic connections	Thermal connections	Scaling potential
Monolithic Si electronic photonic integration	× Off-chip laser no amplifiers	✓ Ultrahigh density	✓ Silicon heatsink	✓ Embedded in electronics
3-D integration technology	× Off-chip laser no amplifiers	× Bond pad limited	× Silica insulator	× Assembly challenges
Hybrid integration Chips on carrier	× Optical vias at the assembly interface	× Bond pad limited	× Silica insulator	× Assembly challenges
Heterogeneous InP and electronics integration	✓✓ On-chip laser and amplifiers	✓ High density	✓ InP heatsink	✓✓ Energy efficient InP optoelectronics

Figure 1.3.7 is an illustration of two wafers, an InP and a Si BiCMOS, which were co-designed and bonded on top of each other.

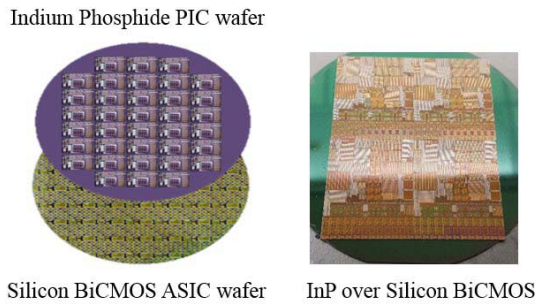


Figure 1.3.8: InP photonic wafer attachment to a Si electronic wafer (left) to enable intimate, high density interconnection of an InP membrane over Silicon BiCMOS (right).

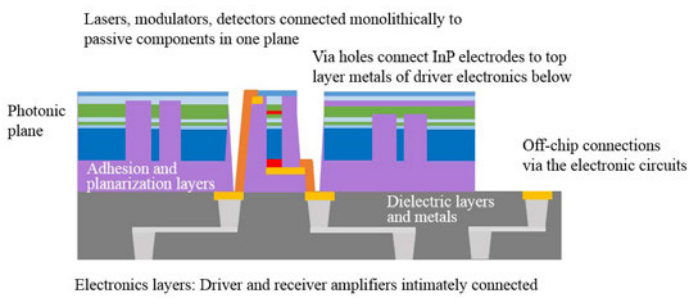


Figure 1.3.9: Photronics integration: the heterogeneous integration of generic InP integrated photonics with microelectronics shown in cross-section.

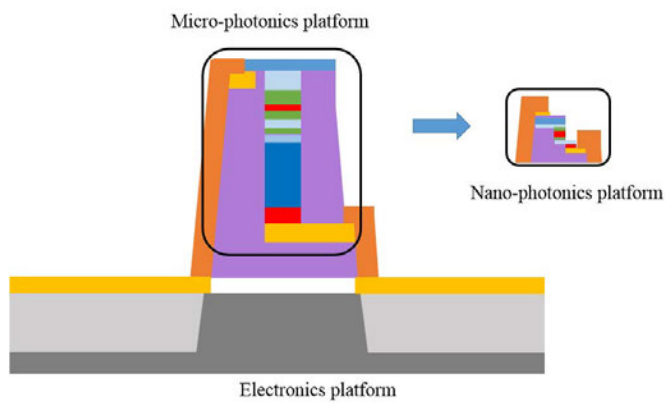


Figure 1.3.10: Transitioning from photronics to integrated nano-photonics to enable order-of-magnitude reduction in waveguide cross-sections.

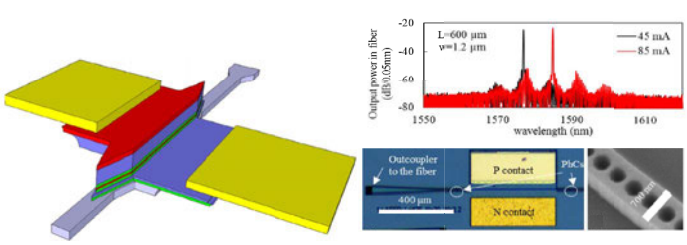


Figure 1.3.11: A photonic crystal reflector laser realized in the IMOS platform. left: the optical amplifier; bottom right: the laser(left) and the photonic crystal reflectors (right); top right: spectra for two pumping currents.

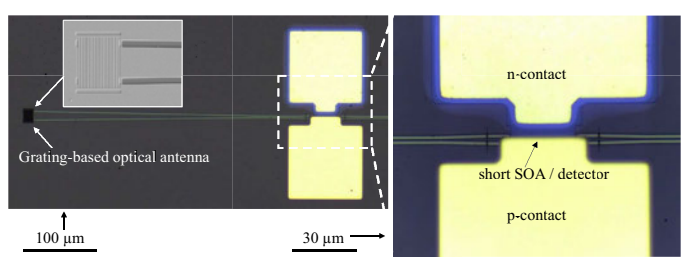


Figure 1.3.12: Integrated reconfigurable optical wireless transmitter/receiver: Left: The transparent SOA gate (in transmitter mode) / photodetector (in receiver mode) integrated with grating-based optical antenna. Right: Zoom in picture of the SOA/detector.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

1.4 5G Wireless Communication: An Inflection Point

Vida Ilderem, *Intel, Hillsboro, OR*

Abstract

The 5G era is upon us, ushering in new opportunities for technology innovation across the computing and connectivity landscape. The advent of the Internet of Things has resulted in the generation of massive amount of data, thus creating both challenges and opportunities for industry. Today's networks will not be able to handle the variety and volume of data that is emerging. 5G presents an inflection point where the network will set the stage for data-rich services and sophisticated cloud applications, delivered faster and with lower latency, and where the wireless communication technology is driven by a multitude of applications and expected use cases. Therefore, 5G will optimize not only the network for human-human communication, but it will also deliver a solution for machine-type communications. This paper will review what is 5G and what are its key requirements, and will highlight the disruptive architectures and technology innovations required to make 5G and beyond a reality.

1. Introduction

The advent of the Internet of Things (IoT), where things and devices are becoming more intelligent and connected, requires networks to become faster, smarter, and more agile in order to handle the unprecedented increase in volume and complexity of data traffic. The magnitude of this data challenge is expected to see personal data use reach 1.5GB per day, hospital data use to reach 4000GB per day, and connected factory use to reach an astounding 1 million GB per day [1].

Next-generation wireless communication, 5G, will require intelligence and flexibility across mobile networks and devices; the rollout of 5G from the cloud to the edge will transform the network to handle the data wave we expect in the coming years. From device to data center, Intel is delivering advanced capabilities, intelligent architectures, and powerful solutions that will unlock the future of 5G, see Figure 1.4.1. The wireless communication journey started with the First Generation (1G) analog voice communication. 2G introduced digital voice communication and text messaging. 3G enabled data services resulting in an affordable mobile Internet. The rapid penetration of broadband Internet access due to multimedia traffic workload led to 4G for improved capacity and higher data rates for mobile users. 5G is not simply an evolution of 4G. It presents an inflection point, not only due to major improvements required over 4G/LTE (Long Term Evolution) data rates and capacity, but also 5G will be the first wireless standard to address the inclusion of a massive number of machines/things in the network, some of which will require lower latency and higher reliability. Many pieces have to come together to successfully realize and deploy 5G. The following sections will review the need for network transformation, 5G key requirements for enhanced mobile broadband, and machine-type communication including spectrum requirements, key 5G enabling technologies, and architectures touching on Intel's research in architecture, beamforming, interference cancellation, antennas, RF design, and modems, finally concluding with a summary.

2. 5G Network Transformation

The need for new network topologies is leading to innovations in dense heterogeneous networks, and efficient use of existing spectrum along with introduction of new spectrum.

Ultra-Dense Networks (UDN) is a new frontier and one of the key enablers for 5G which will shift our paradigms towards dynamic networks that are user/machine centric. UDN is comprised of three vectors: air interface, new spectrum, and network efficiency as depicted in Figure 1.4.2 [2]. Air interface covers new coding schemes, full duplex, and full-dimensional multi-input multi-output (MIMO) antennas with 3 to 5× improvement in capacity. New spectrum covers licensed, unlicensed, and shared-license, resulting in 3 to10× capacity increase. Network efficiency covers cell densification, device-device communication, joint multi-cell scheduling, and Wi-Fi offload, to name a few, with a potential of 40 to 50× capacity improvement. In summary, to increase the network capacity by about 1000×, and deliver quality of experience (QoE), innovations in massive MIMO, small cells, new spectrum such as mm-wave, multi-radio access technology (RAT) aggregation, and full duplex are required. Wireless-network densification is, however, not without its challenges such as interference and backhaul capacity, which are active research areas.

5G is designed to address machine-machine communication, referred to as machine-type communication (MTC), in addition to evolving for a better QoE for human communication. The topology for MTC will require agile flexible and high-throughput networks. MTC must address massive number of objects/devices that can impact network traffic, along with mission-critical applications that are sensitive to latency and/or reliability. Interactive virtual reality is an example of an application that requires a few milliseconds (ms) of end-end latency, while autonomous driving will need both reliable and low latency network. A new air interface, 5G NR (New Radio), is being standardized to address MTC use cases.

In IoT applications, the sensed data is mostly generated at the edge of the network, which requires a closer look at the role of centralized cloud computing in the network. Therefore, one of the observed trends is for compute and storage to move closer to where the data is being generated. Multi-access edge computing (MEC) enables distributed computing at the edge where we can intelligently partition complex workloads between edge and cloud to provide the best QoE for the user on various device types. MEC combined with 5G networking will be key to meet the bandwidth, latency, and reliability requirements of a massive number of connected IoT devices [3]. Examples include virtual- or augmented-reality applications, autonomous driving, and so on.

Thus, 5G requires a new network architecture to support a massive number of devices, efficient platforms, and mobile broadband through optimizing cellular network capacity, coverage, throughput, QoE, and spectrum sharing and efficiency. 5G enables devices that will be empowered with anytime/anywhere access to cloud/edge computing resources at 50× the speed, with 10× less latency, and 1000× the capacity of today's network via software-defined virtualized 5G networks. In the next section, 5G use cases and requirements to make this transformation a reality are reviewed.

3. 5G and Its Requirements

5G addresses three classes of communication services, see Figure 1.4.3: enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable low-latency communication (URLLC). Each class of service has a divergent set of requirements from data rates to latency with different traffic characteristics as depicted in Figure 1.4.4, supporting applications above and beyond the traditional high speed connection. These varied and sometimes conflicting requirements have driven 5G to leverage different frequency spectrum, with <1GHz for mMTC, and <6GHz, in selected mm-wave bands for eMBB and URLLC. Even though eMBB is still an important use case for mobile operators around the world, emergence of other usage scenarios such as those classified under mMTC and URLLC enabled by 5G has opened the door for a myriad of non-traditional applications.

A challenge for 5G is the complexity and variety of its use cases. For instance, URLLC is required for autonomous systems, healthcare, and emergency services. mMTC is an enabler for smart cities, smart agriculture, and manufacturing, whereas eMBB delivers to augmented/virtual reality, mobile office, and entertainment. Each of these market segments have their own unique requirements and needs for bandwidth, response time, energy efficiency, and cost. Enabling these services will require wireless platforms and networks to sense, interpret, and act in real-time. Standardization efforts are underway to provide guidance on the essential functional support needed to cope with the wide collection of use cases 5G will make possible.

3.1 Spectrum – Requirements and Status

Spectrum is the center piece of any generation of wireless communication. The diverse 5G requirements mandate the streamlining of spectrum utilization, and exploring spectrum segments that were not previously available. Therefore, the emerging network capitalizes on a variety of radio interfaces across licensed, licensed shared, and unlicensed spectrum in low-, mid-, and high-frequency bands. Different physical characteristics of spectrum leads to different applications deployed in a suitable spectrum [4]. Low frequencies (those below 1GHz) provide obstacle penetration and non-line of sight (NLOS) operation in cluttered environments, whereas high frequencies (those in mm-wave) are needed for applications requiring very large bandwidths within a relatively short range. For applications where low frequencies do not provide sufficient bandwidth, for example eMBB in mixed line-of-sight (LOS)/NLOS situations, a spectrum range in between (for instance those in the 3 to 6GHz band) would be optimum.

A new air interface, 5G NR, is required to take full advantage of the advanced capabilities of 5G networks. One key feature that differentiates 5GNR from 4G is

self-contained frame structure as well as dynamic time division duplex (TDD) with instant cross-link measurement, which is critical to support the upcoming diverse requirements of connecting both people and machines/things, especially for low latency applications. Channel sizes, and other building blocks of 5GNR specifications are customized to the environment where these different frequency ranges would be used. The 3GPP Release 15 standard is the first specification of 5G NR, which defines the fundamental air interface and network architecture of 5G.

4. 5G Key Technology Enablers

5G use cases and real-time-data analytics are driving system, architecture, and technology requirements, while leveraging process nodes for density, energy efficiency, and performance. Due to the diverse set of applications, no single frequency band (low, mid, high) can meet all 5G requirements. It is widely acknowledged that Moore's Law enables new devices with higher functionality and complexity while controlling power, cost, and size. 5G enabling technologies can significantly benefit from not only process scaling, but also from advanced packaging for passive/antenna integration. In the following subsections, Intel's research solutions to address these 5G challenges are highlighted.

4.1 Low-Power Architectures

Several wireless protocols address the needs of IoT devices, including low-power long-range (LoRa), NB-IoT, Wi-Fi, Bluetooth, and so on. The key metrics for this market segment are energy efficiency and latency, depending on the application. One way to address energy efficiency is through re-architecting the radio at the system level. In both Wi-Fi and cellular protocols, the main radio is always listening or pinging the base station to check for data communication. Wake-up radio (WuR) is a new concept to optimize platform radio power consumption. WuR is a companion radio in Wi-Fi that serves as a wake-up for the main radio, and it is not meant to transfer or access data. Figure 1.4.5 depicts simulated savings for a Wi-Fi wake-up receiver with two different latencies. We can expect power savings of about 7× at 5 sec latency and over 220× at 100 msec latency [5]. Figure 1.4.6 shows the differences in architecture between a Wi-Fi and a cellular WuR (C-WuR). In the case of C-WuR the main radio is reused; however, the baseband is modified to implement the wake-up protocol. C-WuR power savings can be a combination of low-complexity C-WuR digital block, faster sync, and maintaining baseband receiver in deep sleep mode.

As the carrier frequency increases, the effective antenna aperture must be maintained to preserve acceptable link budgets. At mm-wave frequencies this often requires arrays of antennas, which also introduce directionality and hence the need for phased array systems to electronically steer the beam and preserve spatial coverage. Reducing the area and power consumption of each individual transceiver in the phased array is critical to allow scalability and system integration. For the same total isotropic radiated power, the power of each power amplifier (PA) decreases with the square of total number of elements in the array, potentially resulting in better overall efficiency and linearity, while enabling integration in low-cost technologies, such as CMOS. Direct-conversion (zero-IF) architectures can also lead to low power and low area transceiver designs, since no IF mixing stages with large resonant coils are required. Intel has designed a scalable E-band phased-array radio transceivers [6] in our 22FFL process [7]. The proposed architecture and design provides a low-cost solution for fully-integrated phased-array systems with substantial scalability at silicon, package, and board level. This solution paves the way for fully-integrated digital beamforming phased-array systems, which will further boost network capacity.

4.2 Beam-Management Architecture

There are several "Gigabit"-class wireless solutions which manifest very substantial differences in terms of how they perform in practical deployments. The mm-wave spectrum from 24 to 86GHz can serve for both backhaul and network access scenarios, and paves the way for higher capacity and throughput usages in part due to a wider available bandwidth (800MHz to 14GHz). However, these frequency ranges are not without challenges, such as propagation particular characteristics due to atmospheric effects on link availability [8], and LOS/NLOS and blockage issues. Therefore, propagation and channel characteristics, beam management, MIMO/smart antennas, and system architecture must be studied and optimized. To take advantage of the higher available bandwidth, there is a need for wideband high-performance circuits and antennas. Beam management, including channel acquisition, beamforming, beam steering, and beam tracking is an important area of research to allow for spatial diversity and managing of interference, specifically in dense areas of use.

Figure 1.4.7 depicts three architectures for beamforming: analog, hybrid, and digital. Analog beamforming is commonly implemented in mm-wave phased array systems because of its simplicity and low power consumption. However, it may limit the beam-tracking capability of the mm-wave system since it is implemented at RF to avoid array-wide local-oscillator (LO) distribution. Furthermore, analog beamforming is not well suited to MIMO applications, where the array should steer, with full-array gain, in multiple directions simultaneously. In the hybrid architecture, the array is divided into subarrays, each with its own baseband and RF chain. This architecture permits each subarray to produce a unique beam, although the smaller subarrays may limit the performance relative to that of a full array. Finally, full-digital beamforming provides low latency beam management and optimal beamforming using all available antenna information for digital processing and storage. However, power consumption of this architecture is large compared to its analog beamforming counterpart due to the need for analog-digital converters (ADCs) and digital-interface power dissipation. Therefore, we propose a digital beam management architecture that combines a direct-conversion frontend, high performance low power ADCs, and blind spatial compression to reduce interface throughput requirements. Antenna-element information is preserved for digital beam tracking algorithms to provide near-optimum low-latency beamforming. RF circuitry is implemented in scaled CMOS, permitting low-power digital processing. The solution results in an overall power consumption that is competitive with analog beamforming while offering improved performance.

4.3 Self-Interference Cancellation

Simultaneous transmit (Tx) and receive (Rx) signals in the same carrier frequency band, also known as full duplex (FD), is an active area of research aimed at doubling spectral efficiency. Figure 1.4.8(a) depicts the challenge for in-band FD and demonstrates how the Tx signal can impact the Rx signal. To address this issue, we have developed an echo-cancellation technique that is effective in cancelling this in-band self-interference [9]. Self-interference cancellation (SIC) needs to optimize the antenna or circulator isolated Rx/Tx antennas, along with the analog/RF SIC to address the Tx echo, and the digital SIC to address the remaining echo cancellation, see Figure 1.4.8(b). Our solution has resulted in a maximum of ~150dB SIC with an average SIC of 130dB using a 2×2 MIMO. Our closed-loop technique continuously updates the system parameters without requiring a special-training signal and synchronizations, such as a orthogonal frequency division multiplexing (OFDM) boundary. This results in fast and continuous tracking even during random data transmission; hence, it is agnostic to any air interface.

4.4 Antenna/Frontend Module

Development of multi-band/multi-mode mobile terminals in combination with implementation of MIMO operation for communication, present considerable challenges for efficient antenna-system integration in a wireless platform. MIMO antennas are a key technology ingredient to improve spectral efficiency, and thus network capacity and coverage. On the infrastructure side, MIMO dimensions will be large (sometimes referred to as massive MIMO (mMIMO)) to serve multiple users. On the terminal/edge side, the size of the MIMO is limited by the device/terminal form factor. One advantage of operation at mm-wave frequencies is a reduction in the antenna size due to the shorter wavelengths. Various beamforming networks (multi user-MIMO (mu-MIMO), mMIMO), as well as antenna-array antenna-element topologies at sub-6GHz, and mm-waves are being developed for 5G advancement [10,11]. Researchers are also exploring various approaches to embed sub-6GHz, as well as mm-wave bands, within the tightly-integrated small-form-factor and challenging material environments of a smartphone. For instance, we have designed a quad-feed differential broadside antenna topology that exhibits >35dB of cross-polarization discrimination (XPD), which is capable of achieving more than 14GHz of bandwidth, demonstrating an antenna architecture integrated with CMOS RFIC [12].

Design and integration of the antenna system, RF frontend module (FEM), and CMOS RFIC are, as well, becoming more complex and challenging. Traditionally, antennas, FEM, and RFICs are designed independently, and assembled with interconnects and matching networks to create a wireless-communication system. Figure 1.4.9 presents a conventional mm-wave phased-array module cross-section with passive networks, power combiners, and impedance transformers. In this design, lossy interface networks and large impedance-transformation ratios degrade the total power and impact the system efficiency. Co-design of these components is desirable to optimize the effects of FEM on beam patterns and to avoid any performance degradation [13]. For example, the frontend integration and transceiver-performance-fluctuation effects on beam

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

patterns and steering antenna arrays can be identified with such a co-design approach. A dual-polarized (DP) MIMO Tx/Rx CMOS chip integrated with impedance-matched DP-broadside stacked patch antenna structure is shown in Figure 1.4.10 [14]. A proximity-coupled backside-excitation scheme is used to achieve an 8GHz antenna bandwidth on both vertical and horizontal polarizations. The simulated DP radiation patterns including wire bonds, matching circuitry, and antenna feeds show a 90° spatial 3dB bandwidth per polarization.

4.5 RF Design

4G/LTE evolution with more bands and tighter specifications presented challenges to form factor, power, and cost. 5G exacerbates these challenges through introduction of more diverse frequency bands and wider bandwidths. Figure 1.4.11 depicts a generic model of a radio. For instance, the wide-bandwidth and high-throughput mm-wave systems must not only provide power-optimized RF and baseband, but must also address the heavy burden on the power dissipation at the input/output (I/O) data interface between the ADC and the baseband processor at the receiver. Therefore, the fabrication process must simultaneously provide high-performance analog and RF/mm-wave transistors, combined with high-density and low-leakage digital transistors. While Moore's Law continues to benefit digital baseband performance, new processes must also provide RF-specific optimizations in order to preserve wideband high-linearity and highly tunable transceiver design.

An excellent balance between integration density and RF/mm-wave performance has been demonstrated by FinFET technologies. The tri-gate transistor addresses short-channel effects, such as drain-induced barrier lowering (DIBL) and increased device leakage. The benefits are immediately apparent when comparing the I-V curves of planar devices, Figure 1.4.12 [15]. However, these deeply scaled CMOS can exhibit worse thermal dissipation compared to bulk CMOS due to higher local junction temperatures, referred to as self-heating [16], which must be taken into consideration during the design. In the absence of self-heating, energy-efficient designs must co-optimize transconductor efficiency and RF performance, resulting in device operations below peak F_t/F_{max} values of the process. In addition to transistor performance requirements such as gain, leakage, F_t, and F_{max}, the quality factor for the passives and overall system parasitic due to interconnect and layout must also be addressed. We have leveraged Intel's 22FFL process to design and optimize an state-of-the-art CMOS E-Band PA with peak gain of 16.7dB, and 26.3% power-added efficiency, and a 11mW low noise amplifier (LNA) with 20dB gain and 4dB noise figure [17].

In summary, the feasibility of designing mm-wave frontend blocks in a deeply scaled CMOS process is demonstrated. In combination with faster and denser digital circuits, this technology will provide a platform for implementing low-cost, fully-integrated mm-wave phased-array system-on-chips (SoCs) for next-generation wireless systems.

4.6 Modems

5G requires new modem technologies that can accommodate a wide range of use cases, air interfaces, and device form factors. It needs modems that can connect big/small things, fast/slow things, and consumer/industrial/machine things. Additionally, these modems will need to accommodate the previous wireless generation air-interface protocols, such as LTE and 3G.

The modem architecture must address the peak data-rate requirements, while the power requirements are a strong function of the use case. For instance, Internet browsing or voice calls require lower power envelopes, whereas a video download needs a higher data rate. Therefore, the modem design mandates co-optimization of power-performance and of course cost. The modem SoC must focus on dynamic voltage and frequency scaling (DVFS), power gating, network-on-chip-SoC subsystem state management, and so on. New circuit techniques, such as near-threshold-voltage (NTV) operation to maximize active energy efficiency becomes very important [18]. Advanced baseband algorithms such as error correction codes (ECC) such as polar coding and low density parity check (LDPC) could become complex to implement. Therefore, both architecture and circuit design must be co-optimized along with silicon process technology. One of the key blocks of a modem is memory, which makes density, power, and performance important parameters to balance. Logic power/performance must be optimized during the operation, whereas memory-power optimization becomes more important during sleep modes. Various design techniques can be used to optimize the memory size/power [19]. Consequently, application of Moore's Law for the modem is crucial where density/power/performance needs to be co-optimized. Integration of cellular and connectivity protocols will require an extensive amount of validation and testing to ensure that when we integrate all these radios, they will operate to deliver the desired QoE.

5. Summary

5G will not just reshape the wireless, computing, and cloud ecosystem, it will impact nearly every market segment with its ability to render everything smart and connected. Wireless networks must transform to become more powerful, agile, and intelligent, to realize the potential for IoT, and to enable richer experiences throughout daily life. Smart systems and future edge devices need to have capabilities that will require energy-efficient solutions at all levels, including architecture, computation, communication, sense/actuation, power management, and security, to name a few.

5G will encompass a broad variety of cellular and connectivity options including new spectrum such as mm-wave. A wide variety of technologies are required to enable 5G from the infrastructure to the terminal/edge devices. Moore's Law scaling is very critical to provide the power, performance, and cost requirements for this network transformation.

Acknowledgement

I would like to thank my Wireless Communication team and the Next Generation & Standards team for their support and contribution to this paper. I would also like to acknowledge and thank Dr. Vivek De for his support and input.

References

[1] Amalgamation of analyst data and Intel analysis, and VNI (Visual Networking Index) Global Traffic Forecast.
[2] Qian (Clara) li, *et al.*, "5G Network Capacity – Key Elements and Technologies," IEEE Vehicular Technology Magazine, vol 9, no. 1, p71, March 2014.
[3] ETSI White Paper on "MEC on 5G Networks", June 2018.
[4] 5G Americas White Paper on 5G Spectrum Recommendation, April 2017.
[5] E. Alpman *et al.*, "802.11g/n Compliant Fully Integrated Wake-Up Receiver With –72-dBm Sensitivity in 14-nm FinFET CMOS," in *IEEE Journal of Solid-State Circuits*, vol. 53, no. 5, pp. 1411-1422, May 2018.
[6] S. Pellerano, et al., "A Scalable 71-76GHz 64-Element Phased-Array Transceiver Module with 2x2 Direct-Conversion IC in 22nm FinFET CMOS Technology," *ISSCC Dig. Tech. Papers*, pp. 174-175, Feb. 2019.
[7] B. Sell, *et al.* "22FFL: A High Performance and Ultra Low Power FinFET Technology for Mobile and RF Applications," IEDM, p. 685, December 2017.
[8] T. S. Rappaport, *et al.*, "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!" IEEE Access, pp. 335-349, 2013(1).
[9] Y.S Choi, et al." Simultaneous Transmission and Reception: Algorithm, Design and System Level Performance" IEEE Transactions on Wireless Communications, Dec 2013.
[10] L-P. Shen *et al.,* "Dual-Polarized Wideband Multi-Beam Arrays in Wireless Communications," 2015 IEEE International Symposium on Antennas and Propagation (AP-S), pp.1544-1545, 2015.
[11] N. Ojaroudiparchin *et al.*, "Multi-Layer 5G Mobile Phone Antenna for Multi-User MIMO Communications," 2015 23rd Telecommunications Forum Telfor (TELFOR), pp. 559-562, 2015.
[12] Chintan Thakkar, *et al.* "A 42Gb/s 4.8pJ/b 60GHz Digital Tx with 12b/symbol Polarization MIMO," *ISSCC Dig. Tech. Papers*, pp. 172-173, Feb. 2019.
[13] D. Choudhury, "mmW Antenna Integrated Front-End IC-Module Co-Design and Testing for 5G Applications," IEEE IMS2017 workshop WSC presentation, Honolulu-HI; June 4, 2017.
[14] S. Daneshgar *et al.*, "A 27.8Gb/s 11.5pJ/b 60GHz Transceiver in 28nm CMOS with Polarization MIMO," *ISSCC Dig. Tech. Papers*, pp. 166-167, Feb. 2018.
[15] M. Bohr. "Intel 22nm Tri-Gate Announcement, Apr. 2011, https://www.intel.com/content/www/us/en/silicon-innovations/standards-22nm-3d-tri-gate-transistors-presentation.html,
[16] M. Shrivastava *et al.*, "Physical Insight Toward Heat Transport and an Improved Electrothermal Modeling Framework for FinFET Architectures," in IEEE Transactions on Electron Devices, vol. 59, no. 5, pp. 1353-1363, May 2012.
[17] W. Shin *et al.*, "A Compact 75 GHz LNA with 20 dB Gain and 4 dB Noise Figure in 22nm FinFET CMOS Technology," RFIC 2018.
[18] Jain, S. *et al*, "A 280mV-to-1.2V wide-operating-range IA-32 processor in 32nm CMOS", *ISSCC Dig. Tech. Papers*, pp. 66-68, Feb. 2012.
[19] Khellah, M *et al.,* "A 256-Kb Dual-VCC SRAM Building Block in 65-nm CMOS Process With Actively Clamped Sleep Transistor," IEEE Journal of Solid-State Circuits, Volume 42, Issue 1, Page(s):233 – 242, Jan. 2007.

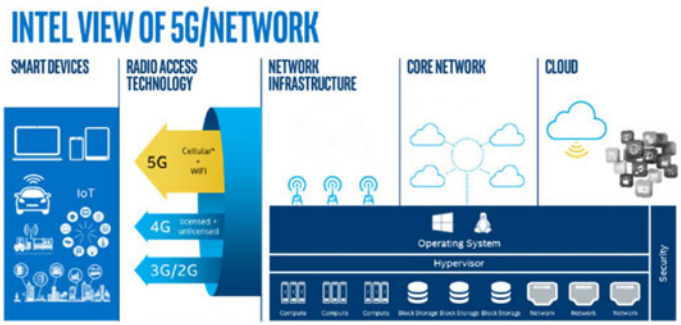


Figure 1.4.1: 5G network.

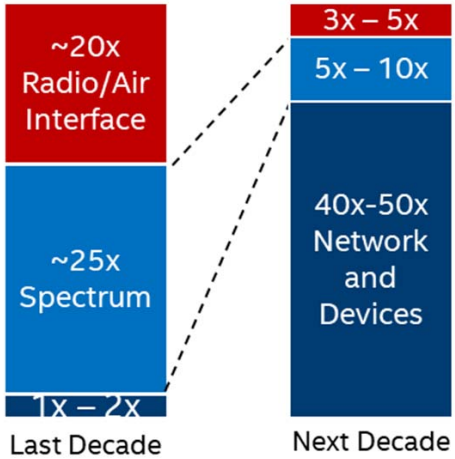


Figure 1.4.2: UDN technology enablers.

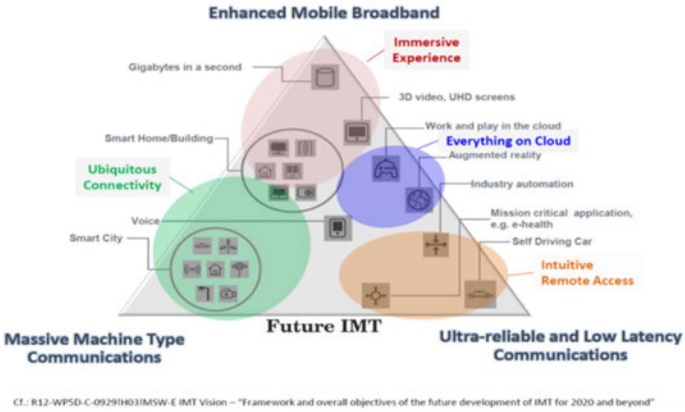


Figure 1.4.3: 5G service categories.

ITU and 5G Requirements
ITU-R M.[IMT.VISION]

Attribute	IMT-Adv. 4G	IMT-Future 5G
Achievable Rates (Mbps)	1Gbps	10-50Gbps
Connection Density		10^6-10^7/km^2
Mobility & Coverage	350km/h	500km/h
Energy Efficiency	1x	50-100x
Spectral Efficiency	1x	5-15x
Latency	10ms	1ms

Selected 5G Requirements (ITU WP5D – July 2014)

Figure 1.4.4: 5G key requirements.

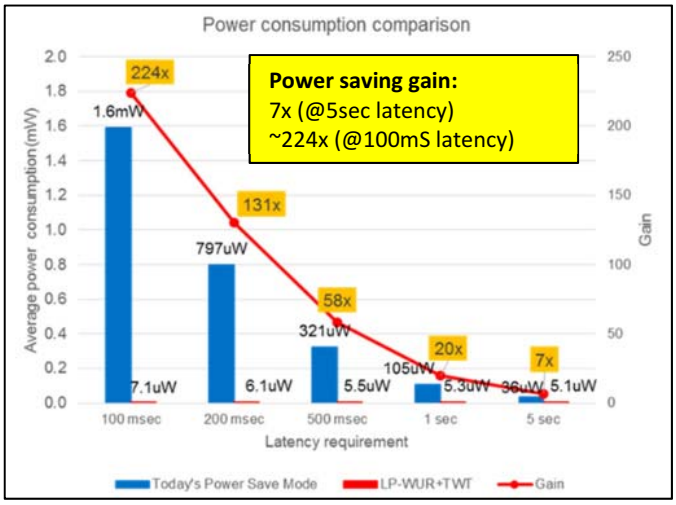


Figure 1.4.5: Wake-up radio simulation results.

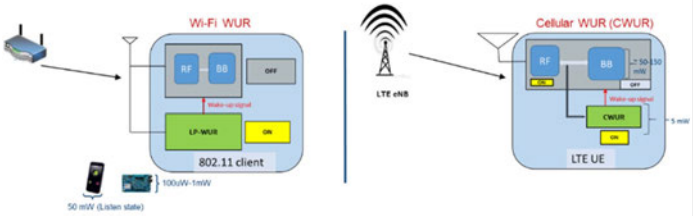


Figure 1.4.6: Architectures for Wi-Fi and cellular WuRs.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

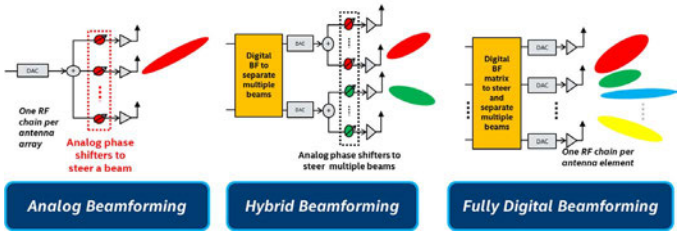


Figure 1.4.7: Architectures for beamforming.

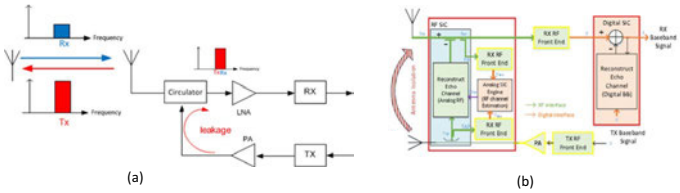


Figure 1.4.8: (a) Challenge for in-band full duplex; (b) Echo self-interference cancellation solution.

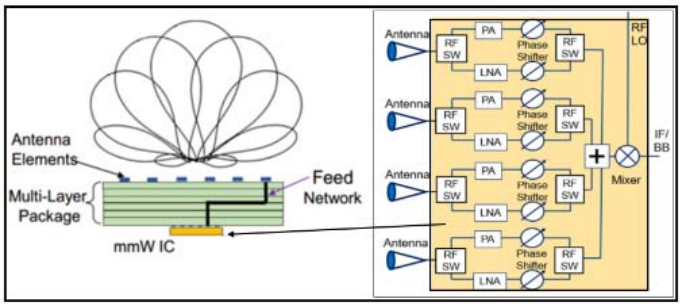


Figure 1.4.9: An example of antenna-integrated mm-wave phased-array module [13].

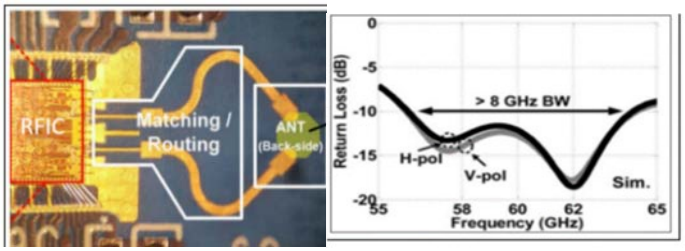


Figure 1.4.10: CMOS-IC integrated with dual-polarized broadside antenna [14].

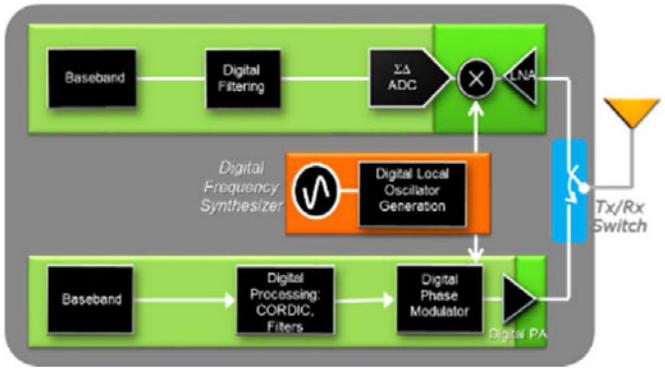


Figure 1.4.11: Generic picture of a digital radio.

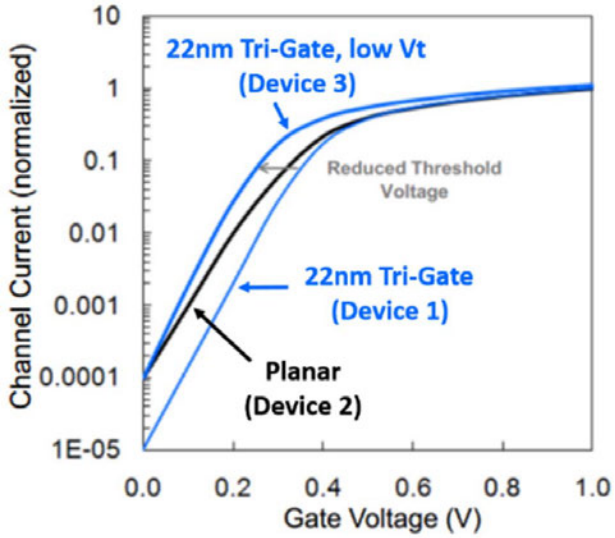


Figure 1.4.12: I-V characteristic of planar and Tri-gate devices [15].

Session 2 Overview: *Processors*

DIGITAL ARCHITECTURES AND SYSTEMS SUBCOMMITTEE

Session Chair: Ingrid Verbauwhede
KU Leuven, Belgium

Session Co-Chair: James Myers
Arm, Cambridge, United Kingdom

Interim Subcommittee Chair: Thomas Burd, AMD, Santa Clara, CA

Processors continue to diversify and specialize for an ever-wider range of applications. The first paper describes a high-performance supercomputer, followed by low-power streaming vision and post-quantum cryptographic processors. A versatile platform for collaborative robotics comes next, followed by very specialized mixed-signal graph processor and a CMOS annealer. The final paper showcases a multi-core automotive MCU with real-time virtualization.

INVITED PAPER **1:30 PM**

2.1 Summit and Sierra: Designing AI/HPC Supercomputers

J. A. Kahle, IBM Research, Austin, TX

In invited Paper 2.1, IBM describes two supercomputers, Summit and Sierra, built upon IBM POWER9™ CPUs and NVIDIA Volta™ GPUs. The Summit supercomputer contains 9216 CPUs and 27648 GPUs, achieving 200 PFLOPS in 15MW.

2:00 PM

2.2 A 978GOPS/W Flexible Streaming Processor for Real-Time Image Processing Applications in 22nm FDSOI

S. Smets, KU Leuven, Heverlee, Belgium

In Paper 2.2, KU Leuven presents a 22nm 6.4mm^2 coarse-grain reconfigurable streaming processor with 330 streaming elements, performing real-time dense optical flow in 30fps VGA at 10.7mW.

2:30 PM

2.3 An Energy-Efficient Configurable Lattice Cryptography Processor for the Quantum-Secure Internet of Things

U. Banerjee, Massachusetts Institute of Technology, Cambridge, MA

In Paper 2.3, MIT shows an IoT targeted lattice processor supporting Kyber, NewHope, R. EMBLEM and LIMA Round 1 post-quantum cryptography proposals, in a 0.28mm^2 chip in 40nm CMOS.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

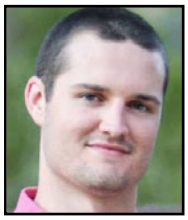
3:15 PM

2.4 A Distributed Autonomous and Collaborative Multi-Robot System Featuring a Low-Power Robot SoC in 22nm CMOS for Integrated Battery-Powered Minibots

V. Honkote, Intel, Bangalore, India

In Paper 2.4, Intel presents a platform including a 22nm SoC with path planning and motion control acceleration, for collaborative robotic applications. Wide-range voltage-frequency scaling delivers 37 to 238mW at 80 to 365MHz.

3:45 PM



2.5 A 40×40 Four-Neighbor Time-Based In-Memory Computing Graph ASIC Chip Featuring Wavefront Expansion and 2D Gradient Control

L. R. Everson, University of Minnesota, Minneapolis, MN

In Paper 2.5, the University of Minnesota introduces a mixed-signal time-domain 65nm application-specific processor for shortest-path problems, scaling across multiple chips and achieving 559 million traversed edges per second at 10^5× lower energy than an FPGA. The processor operates at 1.79ns per node with a peak power of 26.4mW.

4:15 PM

2.6 A 2×30k-Spin Multichip Scalable Annealing Processor Based on a Processing-In-Memory Approach for Solving Large-Scale Combinatorial Optimization Problems

T. Takemoto, Hitachi, Sapporo, Japan

In Paper 2.6, Hitachi describes a 40nm CMOS annealing processor for combinatorial optimization problems, supporting 50% higher spin count and multichip interfaces. The 24mm^2 chip achieves a 22μs annealing time, which is 26000× faster than a CPU.

4:45 PM

2.7 A 28nm 600MHz Automotive Flash Microcontroller with Virtualization-Assisted Processor for Next-Generation Automotive Architecture Complying with ISO26262 ASIL-D

N. Otsuki, Renesas Electronics, Kodaira, Japan

In Paper 2.7, Renesas reveals a 94mm^2 4-core, 600 MHz automotive control processor, supporting virtualization with 0.44μs context switching latency in 28nm CMOS with e-FLASH.

2.1 Summit and Sierra: Designing AI/HPC Supercomputers

James A. Kahle[1], Jaime Moreno[2], Dan Dreps[3]

[1]IBM Research, Austin, TX
[2]IBM Research, Yorktown Heights, NY
[3]IBM Systems and Technology, Austin, TX

The Summit and Sierra Supercomputer Systems, deployed in 2018 at the Department of Energy (DOE) National Laboratories, Oak Ridge (ORNL) and Lawrence Livermore (LLNL), provide a significant increase in computing capability relative to the predecessors, and represent a major step in the path to Exascale computing. Arising from the DOE CORAL program (https://asc.llnl.gov/CORAL/), Summit and Sierra were listed as #1 and #2, respectively, in the November 2018 version of the Top500 list (https://www.top500.org/lists/2018/11/).

Supercomputers such as Sierra and Summit have unique design challenges in network bandwidth, communication patterns, synchronous coordination and parallel computation not found in traditional cloud deployments. While Supercomputers and Data Centers are both large collections of interconnected compute nodes installed in a single location, they differ in the style of computing they perform. In particular, Supercomputers are characterized for synchronous computation on "bare metal" resources, coordinated parallel work, message passing among parallel tasks, intensive computation periods that can last many hours or even days so that job duration may be longer than the reliability window. In contrast, general Data Centers use virtualized images, interacting components driven by external events, information exchange used to drive work, leading to massive short-lived computation pieces and software structures designed for failures.

Driven by the high efficiencies demanded by large supercomputer machines and the slowdown of Moore's law of scaling, the CORAL systems address the design challenges with a heterogeneous approach. Applications have a diverse set of needs that can be addressed with strong CPUs for analytics capabilities and powerful GPUs for massively parallel sections. Examples of analytics capabilities well suited to CPUs are exemplified by complex codes with data-dependent paths, lots of indirection and pointer chasing, and dependency on latency of the memory subsystem, as found in oil and gas reservoir simulation, AI and graph analytics. Examples of massively parallel compute which are well suited to GPUs are exemplified by simple kernels, dense FP operations, and simple data access patterns as found in oil and gas seismic simulation, financial value at risk and image analytics. Tying diverse compute engines together through a high bandwidth interconnect and state-of-the-art memory systems was key in meeting application needs.

To address the explosion of data and associated computational demands, IBM has created a data centric systems approach to these demands with four guiding principles: 1) Minimize data motion, 2) enable compute at all levels of the system hierarchy, 3) modularity of system components, 4) application-driven design. With OpenPOWER [1], IBM created an environment for modulatory with its partners: NVIDIA supplied the GPU for highly parallel work, and Mellanox provided the high-speed interconnect for the system.

Figure 2.1.1 describes the scale and compute power of these two systems, whereas Fig. 2.1.2 provides an overall overview of the components in them. To illustrate the data movement capabilities of the larger Summit system, the data starts in a GPFS parallel file system comprised of 32,648 10TB hard disks in 39 Racks. Data flows to the compute nodes via a network with a cross sectional bandwidth of 115.2PB/s. For each compute node (see Fig. 2.1.3), there is 25GB/s of injection bandwidth feeding two POWER9 processors, each with 150GB/s of memory BW to 512 GB of DRAM. A high speed NVLink 2.0 with 150GB/s of bandwidth feeds the Tesla V100 GPUs that have 900GB/s of B/W to the 16 GB of HBM2 memory.

The POWER9 processor [2-3], depicted in Fig. 2.1.4, supports high single-thread computation, runs the operating system, and coordinates data movement for the system. The high-speed POWER bus supports memory movement between the system DRAM and the NVLINK. The heterogeneous mix between the CPU and GPU allows a diverse set of applications to efficiently run on the system.

To keep the processor node at a suitable temperature, a cold-plate technology was deployed [4]. Warm water first enters the system through rack-level connectors, flows over the POWER9 processors and then to the Tesla V100 GPUs. The rest of the components in the node are cooled with forced air that can be further cooled with a rear door heat exchanger. The CPU and GPUs consume 300W each, leading to a 2.4KW 2U drawer, 58.5KW rack, for a total system power under 15MW.

To maintain optimal performance with the 300W power budget, the POWER9 processor implements dynamic clock and voltage control with an embedded PowerPC 405 on-chip controller and 20 additional microcontrollers distributed around the chip. Multiple sense points monitor for any voltage droop and can lengthen the clock pulse within a few cycles to avoid any timing conditions. While maintaining minimal timing margins, the processor achieves consistent performance at an application basis.

A key feature of CORAL nodes is the high-performance NVLink 2.0 bus [5] depicted in Fig. 2.1.5, which connects the CPU with the GPUs. NVLink 2.0 uses 25Gb/s coherent I/O, enabling seamless GPU/CPU integration, key to effectively achieving higher performance in a productive programming environment with full coherency between the CPU and GPU.

To enable the capabilities described above for NVLink 2.0, two copies of a 1.24Tb X24 bidirectional processor link macro were integrated into the POWER9 chip, in 14nm SOI, running at 25.8Gb per lane, using 6.3 mm^2 and consuming 4.3W. The X24 link has 24 lanes of RX and TX respectively. One low-noise LC PLL drives two resonant clock distributions to the RX bank and TX bank. This physical layout increases isolation of TX to RX. Both the RX and TX run off the same 1.0V supply, while the PLL has an AV_{DD} of 1.5V. Multiple copies of this core are used on a single die. The RX PHY is compatible with several short-reach industry standards. The core has full diagnostics, eye monitor capability and some enhanced features like TDR embedded.

The 25.8Gb/s transmitter achieves 1.3pJ/b at 1.0V supply with active pre-cursor and post-cursor feed-forward equalization (FFE), using a full rate source series terminated driver. The serializer utilizes a resonant clock to combine 16b of data into serial main data and FFE data. The FFE data can consist of pre-cursor, post-cursor, or both, depending on the needs of the product. FFE can be disabled if not needed by the channel, reducing power of the pre-drive buildup path. A simplified block diagram of the transmitter is depicted in Fig. 2.1.5.

Two classes of benchmarks were specified in the CORAL program: *scalable science* benchmarks, which represent single applications expected to run across the entire system (e.g., across 4200 or 4600 nodes) for which performance scalability is a major challenge, and *throughput* benchmarks, which represent the simultaneous execution of multiple instances of the same application or a set of different applications, each using 216 or 192 nodes.

Furthermore, two versions of each benchmark were specified in the CORAL program: *baseline*, which only allowed adding high-level code directives for performance (such as OpenMP or OpenACC program directives), and *optimized,* which allowed code restructuring, architecture-specific code optimizations, or any other coding technique to increase performance. *Baseline* is intended to give a metric of "out-of-the-box" performance of the systems, whereas *optimized* targets demonstrate the ultimate capabilities of the systems.

Figure 2.1.6 depicts performance measurements on the benchmarks and their geometric mean. These values are improvement ratios with respect to the prior-generation systems installed at the DOE labs, namely IBM BlueGene/Q at LLNL and Cray Titan at ORNL. The measurements collected show the *optimized* code performing 5.9-to-7.8× better than the reference systems (geometric mean across the benchmarks in each class), whereas the *baseline* code delivers 2.4-to-6.0× better performance, amply fulfilling the goals of the CORAL program. Additional performance improvements should be expected from further application tuning based on greater experience using the systems. Moreover, usage of early, partial, systems has already resulted in significant performance gains on scientific applications [6]. Moreover, early experiments leveraging NVLink in applications that require moving large data in/out from the GPUs are also reporting major performance gains, for example [7].

References:

[1] https://openpowerfoundation.org/
[2] C. Gonzalez, et al., "The 24-Core POWER9 Processor with Adaptive Clocking, 25-Gb/s Accelerator Links, and 16Gb/s PCIe Gen4," *IEEE JSSC*, vol. 53, no. 1, pp. 91-101, 2018.
[3] B. Thompto, "POWER9: Processor for the Cognitive Era," *IEEE Hot Chips 28*, pp. 1-19, 2016.
[4] S. Chun, et al., "IBM POWER9 Package Technology and Design," *IBM J. of Research and Development*, vol. 62, no. 4/5, 2018.
[5] IBM-POWER9-NPU team, "Functionality and Performance of NVLink with POWER9 Processors," *IBM J. of Research and Development*, vol. 62/4-5, 2018.
[6] T. Straatsma, "Early Application Results on Summit", *ORNL Smoky Mountains Conference*, Sept. 2018.
[7] https://developer.ibm.com/linuxonpower/2018/07/27/tensorflow-large-model-support-case-study-3d-image-segmentation/

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

	Summit	Sierra
Peak Performance	**200 Petaflops**	**125 Petaflops**
Number of Nodes	4608	4320
Node Performance	43 Teraflops	29 Teraflops
Memory per node	512 GB DDR4, 96 GB HBM2	256 GB DDR4, 64 GB HBM2
NV Memory per node	1600 GB	1600 GB
Total System Memory	**11.1PB** DDR4+HBM2+NV	**9PB** DDR4+HBM2+NV
Compute nodes	**9,216 IBM POWER9™ CPUs 27,648 NVIDIA Volta™** GPUs	**8640 IBM POWER9™ CPUs 17280 NVIDIA Volta™** GPUs
File System	250 PB, 2.5 TB/s, GPFS™	156 PB, 1.5 TB/s, GPFS™
Power Consumption	**15 MW**	**12 MW**
Interconnect	Mellanox EDR 100G InfiniBand	
Operating System	Red Hat Enterprise Linux (RHEL) version 7.4	

Figure 2.1.1: Summit and Sierra: next-generation AI and HPC platform.

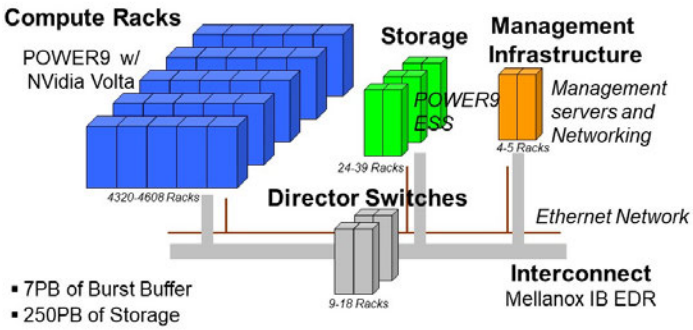


Figure 2.1.2: The complete architecture build out.

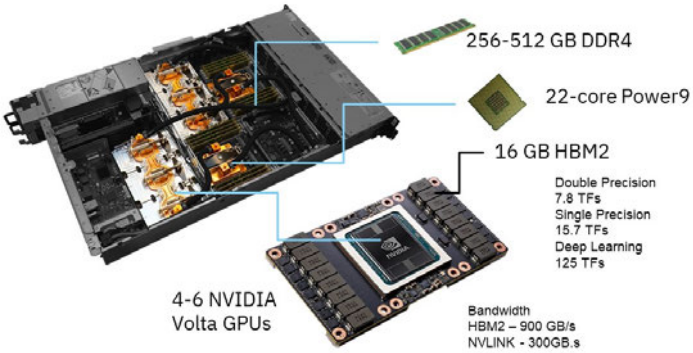


Figure 2.1.3: Compute planar – heterogeneous fat node.

- 24 cores Physical Design
  - 22 cores for yield optimization in CORAL
- 14nm FinFET SOI technology
  - 8 billion transistors
- PCI-Express 4.0
- NVLink 2.0
- Large Caches
  - L1I/D: 32 KiB per core, 8-way set assoc.
  - L2: 258 KiB per core
  - L3: 120 MiB eDRAM, 20-way

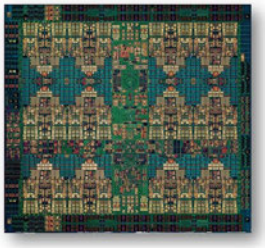


Figure 2.1.4: IBM POWER9 processor.

- Next generation of GPU/CPU integration
- Key to achievable acceleration performance
- High productivity programming environment

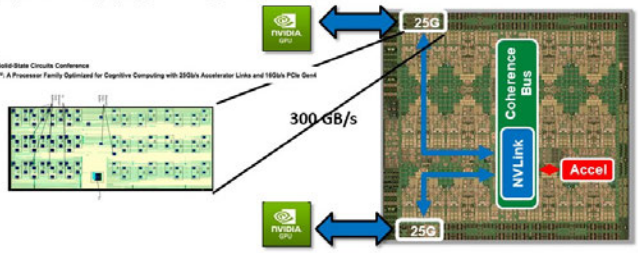


Figure 2.1.5: NVLink 2.0 25G coherent I/O.

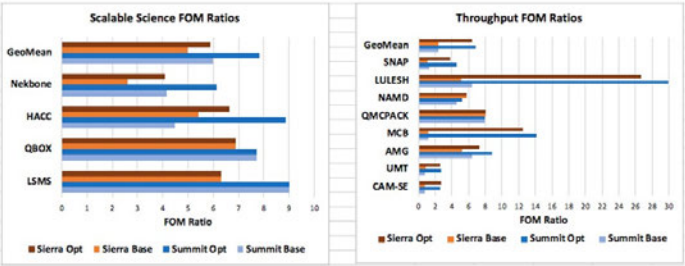


Figure 2.1.6: Performance results.

2.2 A 978GOPS/W Flexible Streaming Processor for Real-Time Image Processing Applications in 22nm FDSOI

Sander Smets[1], Toon Goedemé[1], Anurag Mittal[2], Marian Verhelst[1]

[1]KU Leuven, Heverlee, Belgium, [2]San Jose, CA

The recent trend towards high frame rate, high resolution image processing necessitates a solution for the growing Von Neumann bottleneck. Bandwidth pressure is relieved in traditional image processors by tiling the image and storing pixel patches in a local scratchpad memory [1]. Yet, this only partially alleviates the memory bottleneck, due to repeated fetches for overlapping patches and the scratchpad access becoming the main bottleneck (Fig. 2.2.1). Streaming architectures [2] eliminate this overhead by streaming the image pixels into an array of processing elements and transferring intermediate data results to a next processing element instead of sending them back to memory. Crucial in the efficiency of such streaming image processing architectures is the line buffering strategy, required to merge intermediate results in the stream with previously computed results. Existing state-of-the-art architectures either have dedicated, application-specific line buffer instances which are inflexible towards various image processing workloads [1], or achieve flexibility by routing streams back to centralized scratchpad memories, again resulting in a memory bottleneck around the centralized buffers, and introducing the need for inefficient on-chip-networks between the processing elements [3]. This paper introduces the concept of flexible, instruction-programmable streaming processing with embedded configurable-delay FIFOs to enable line buffering for a wide range of computer vision algorithms on a single platform. Providing a deep reconfigurable pipeline, this design facilitates the mapping and execution of complex vision tasks, like dense optic flow in real-time (30fps VGA) at low power (10.7mW), marking a 5.8× improvement over the state-of-the-art.

The architecture (Fig. 2.2.2) consists of 330 single-cycle processing elements (PEs) and 90 configurable delay FIFOs, which are each flexible in the instruction they execute, and in the way they are concatenated, forming a deeply pipelined processing fabric handling the so-called pixel streams. Every PE or FIFO processes a new input every clock cycle and passes its result to the next PE or FIFO in the pipeline. The configuration of the units and the connections between them determines how pixels travel through the pipeline and how they are processed.

The PE, FIFO and crossbar configurations are provided in an instruction stream preceding the pixel stream. During the programming phase, indicated by a dedicated program signal, all PEs revert to a pass-and-store mode during which they pass the instruction stream to their successive PEs. The instruction stream has the same format as the data stream, allowing the signal interconnection fabric to be re-used, resulting in very limited configurability overhead. To ensure the instruction stream passes by all instruction registers, the program signal configures the crossbars in a default pattern that passes by every single PE. When the program signal goes low, the PEs store their current instruction in an instruction register.

Every individual PE consists of a 16b or 32b fixed-point ALU, an instruction register (IR) and a guard control unit (GC). The ALU selects its inputs from 5 input channels, which consist of 2 outputs of the previous PE, 2 inputs from the beginning of the PE unit chain (e.g. for data reuse within a PE chain), and a constant value locally stored in the PE (e.g. for a filter weight). The operation performed by the ALU and settings of internal multiplexers are determined by the PE's local instruction register, which remains frozen for the complete processing stream. This eliminates all activity overhead in instruction fetching and decoding. Finally, the guard control unit enables on-line control flow decisions by guarding operation outputs based on 4b flags travelling along with the pixels, enabling conditional instruction execution. The PEs are clustered in groups of 4 pipelined instances which connect to a hierarchical configurable interconnect network, enabling concatenation of PEs into longer pipelines and various configurations. Overhead of unused PEs is minimized through clock gating and bypassing.

The distributed memories for line buffering total 690KB of SRAM, spread over 90 configurable-delay FIFOs. Each of these buffers is clustered with one PE, hence extending the PE instruction set with the capability to specify a configurable delay. This clustering is beneficial, as several computer vision kernels only require a single operation between successive buffers. Each of the FIFOs can delay a pixel stream up to 1024 cycles, buffering up to one line of HD frames without chaining FIFOs together. Unused SRAM blocks are placed in sleep mode for the full execution duration to minimize power consumption.

Crucial to streaming fabrics is their ability to map and concatenate a wide range of image processing compute kernels found in typical image processing pipelines, as listed in Fig. 2.2.1. Fig. 2.2.3 demonstrates how 3 of these kernels can be mapped on the presented fabric: horizontal and vertical Sobel filters and integral image construction. The extension from the first 2 kernels to larger or 2D filters is straightforward. For the third kernel, integral image construction, the first PE computes the running sum over a single line in the image and the PE with FIFO sums the current running sum with the result of the previous line. The reset of the values of the running sums at the start of a new line is implemented using a conditional operation using the guard control unit. Many of such basic image processing kernels can be concatenated in one long streaming pipeline across different interconnected PE units. We illustrate this for an optic flow application, using pyramidal Lucas-Kanade (LK) dense optical flow consisting of many concatenated image kernels (Fig. 2.2.4). The complete multi-scale algorithm is split into 7 passes over the streaming fabric (3× subsampling, 4× single-scale LK evaluation, warping performed off-chip). During subsampling, an 11-tap anti-aliasing Sobel filter is applied on both input images, occupying 52 out of 330 PEs and 20 out of 90 FIFOs. LK evaluation occupies 219 out of 330 PEs and 39 out of 90 FIFOs, accounting for 150 effective operations per clock cycle, with only 39 SRAM accesses. In between passes, the fabric is reconfigured with the next instruction stream.

The design was implemented in a 22nm FDSOI technology with an active area of 4.9mm^2 and runs at 200MHz at 0.72V and room temperature. Fig. 2.2.7 provides chip details and a die micrograph. Fig. 2.2.4 shows measurement results of the chip while running the LK dense optic flow at a 30MHz clock, which suffices for real-time VGA processing. Note that reconfiguring the architecture and clearing the pipeline only blocks execution around 3% of the time and only accounts for 2% of the overall energy consumption.

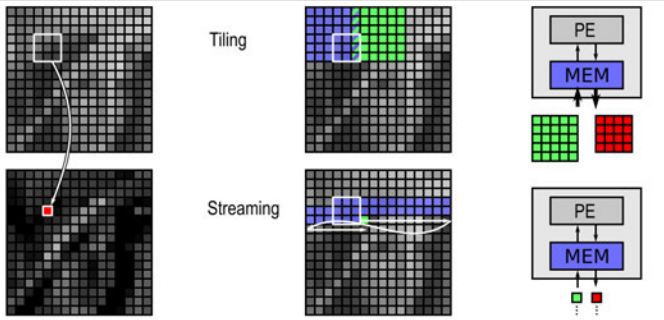
Figure 2.2.5 shows how throughput and energy efficiency can be traded off by modulating the supply and body bias voltages for full utilization (all PEs chained in a 330-tap filter) and during LK evaluation. The minimal energy point, at 36MHz, a supply of 480mV and a bias voltage of -150mV, achieves a peak energy efficiency of 978GOPS/W. Towards higher resolutions and/or frame rates, the body bias allows tuning the dynamic vs. leakage power ratio. At 220MHz (30fps, Full HD LK flow) a 450mV bias voltage enables a 7% power reduction with respect to zero-bias.

Figure 2.2.6 compares the presented design to recent streaming architectures and optic flow implementations. The demonstrated peak efficiency of 978GOPS/W at 36MHz remains within a factor 3 compared to non-programmable solutions. As a result of the configurable line buffering scheme, the architecture scales to high resolutions at a low energy cost. In terms of energy efficiency, this work marks a 5.8× improvement for optic flow, enabling its usage for autonomous navigation on energy-constrained platforms.

Acknowledgements:
The authors would like to thank Research Foundation Flanders (FWO) for supporting the work of S. Smets by a Doctoral Fellowship, and GlobalFoundries for chip fabrication support.

References:
[1] Y. Sasagawa and A. Mori, "High-level Video Analytics PC Subsystem Using SoC With Heterogeneous Multicore Architecture," *IEEE JSSC*, vol. 51, no. 4, pp. 1051-1059, 2016.
[2] F. Barranco, et al., "Parallel Architecture for Hierarchical Optical Flow Estimation Based on FPGA", *IEEE TVLSI*, vol. 20, no. 6, pp. 1058-1067, 2012.
[3] X. Fan, et al., "Stream Processing Dual-Track CGRA for Object Inference," *IEEE TVLSI*, vol. 26, no. 6, pp. 1098-1111, 2018.
[4] Z. Li, et al., "A 1920 x 1080 25fps, 2.4TOPS/W Unified Optical Flow and Depth 6D Vision Processor for Energy-Efficient, Low Power Autonomous Navigation", *IEEE Symp. VLSI Circuits*, pp. 135-136, 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE



Kernel \ Application	LK Flow	SGM Depth	CNN Class.	GMG BS
Pixelwise ops.	✓		✓	✓
Sobel filter	✓			
2D filter			✓	
Integral image	✓			
Morphological ops. (dilation, erosion...)				✓
Census transform		✓		

Figure 2.2.1: Overview of patch-based computer vision kernels and their applications.

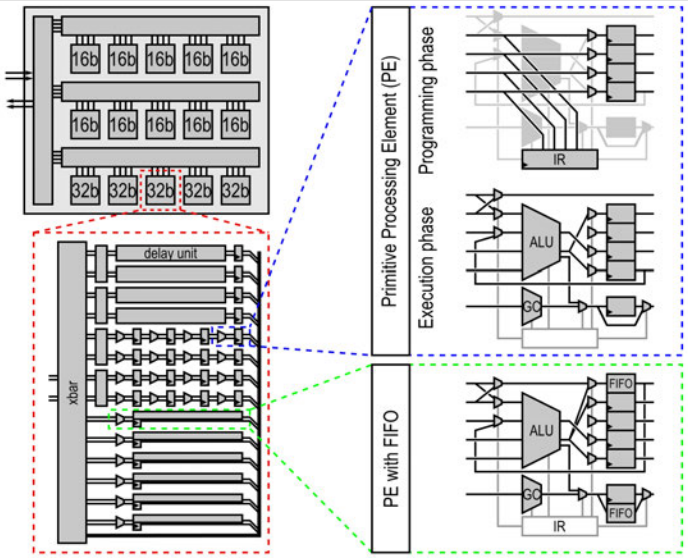


Figure 2.2.2: Top-level architecture.

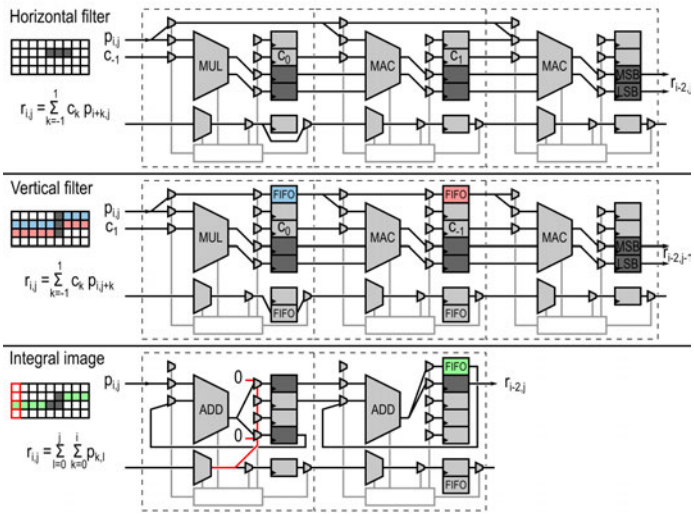


Figure 2.2.3: Mapping of typical computer vision kernels on the streaming fabric.

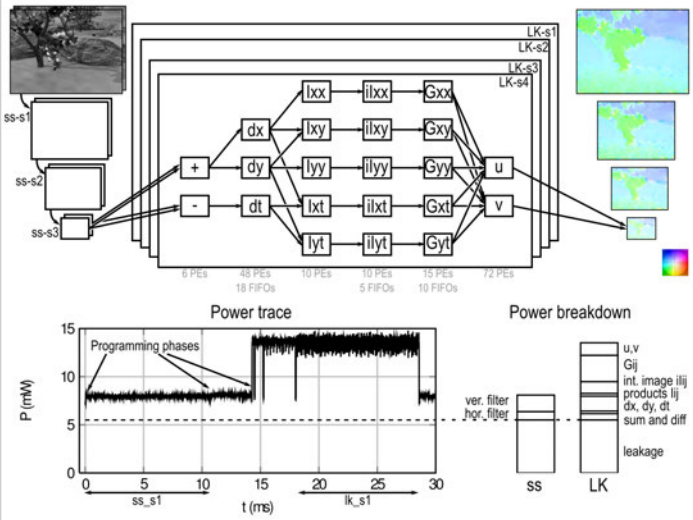


Figure 2.2.4: Measured power consumption for Lucas-Kanade dense optical flow.

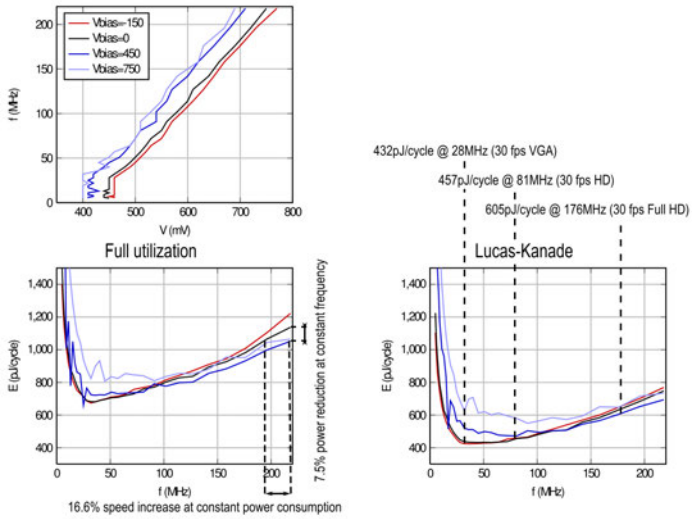
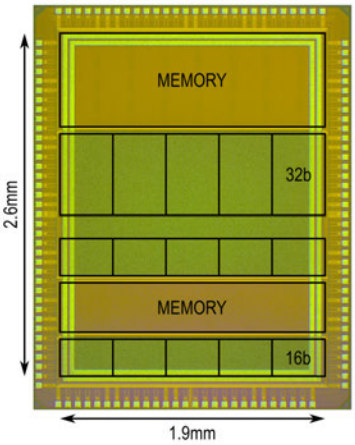


Figure 2.2.5: Body bias tuning and measured efficiency.

	TVLSI '12 [2]	JSSC '16 [1]	TVLSI '18 [3]	VLSI '18 [4]	**This work**
Technology	90nm FPGA	28nm CMOS	55nm CMOS	28nm CMOS	**22nm FDSOI**
Application	LK flow	Sparse LK flow	AlexNet	NG-fSGM flow	**LK flow**
Processing approach	Streaming (distributed mem, fixed buffer sizes)	Patch-based	Streaming (central memory)	Patch-based	**Streaming (distributed mem, config buffer sizes)**
Programmability	Fixed datapath	SoC with fixed datapath for flow accelerator	Programmable CGRA	Fixed datapath	**Programmable CGRA**
Image size and throughput	640x480, 32 fps	1920x1080, 30 fps	227x227, 35 fps	640x480, 30 fps 1920x1080, 25 fps	**640x480, 30 fps 1920x1080, 30 fps**
On-chip memory	252 KB	2 MB	108 kB	196 KB	**690 KB**
Chip area	N/A	143.95 mm^2	5.19mm^2	9.3mm^2	**4.9mm^2**
Peak performance	N/A	3.5 TOPS	77.4 GOPS	1.82 TOPS	**145 GOPS**
Efficiency	N/A	1871 GOPS/W	50.78 GOPS/W	2.4 TOPS/W	**978 GOPS/W**
Power consumption	N/A	1878 mW	841 mW	62 mW @ 30 fps VGA 760 mW @ 25 fps FHD	**10.7 mW @ 30 fps VGA 101.4 mW @ 30 fps FHD**
Normalized energy	N/A	56.8 nJ/pix	N/A	6.73 nJ/pix @ 30 fps VGA 14.66 nJ/pix @ 25 fps FHD	**1.16 nJ/pix @ 30 fps VGA 1.63 nJ/pix @ 30 fps FHD**

Figure 2.2.6: Performance comparison.



Technology	22nm FDSOI
Die area	2.2 x 2.9 mm^2
Active area	1.9 x 2.6 mm2
Supply voltage	0.8 V (nom)
Bias voltage	-150 - 900 mV
Clock frequency	5 - 220 MHz
Power consumption for 30 fps VGA optic flow	10.73 mW @ 30 MHz

Figure 2.2.7: Annotated chip photograph and chip summary.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

2.3 An Energy-Efficient Configurable Lattice Cryptography Processor for the Quantum-Secure Internet of Things

Utsav Banerjee[1], Abhishek Pathak[2], Anantha P. Chandrakasan[1]

[1]Massachusetts Institute of Technology, Cambridge, MA
[2]Indian Institute of Technology Delhi, New Delhi, India

Modern public key protocols, such as RSA and elliptic curve cryptography (ECC), will be rendered insecure by Shor's algorithm [1] when large-scale quantum computers are built. Therefore, cryptographers are working on quantum-resistant algorithms, and lattice-based cryptography has emerged as a prime candidate [1]. However, high computational complexity of these algorithms makes it challenging to implement lattice-based protocols on resource-constrained IoT devices, which need to secure data against both present and future adversaries. To address this challenge, we present a lattice cryptography processor with configurable parameters, which enables up to two orders of magnitude energy savings and 124K-gate reduction in system area through architectural optimizations. The ASIC demonstrates multiple lattice-based protocols proposed in Round 1 of the NIST post-quantum standardization process.

Figure 2.3.1 provides an overview of the "Learning with Errors" (LWE) problem which forms the basis of several lattice-based schemes. The LWE hardness assumption states that it is computationally difficult to determine the secret vector s, given the matrix A and the vector b = As + e, where all arithmetic is modulo a small integer q, and s and e are short vectors sampled from a discrete distribution. This hardness is preserved even in the presence of quantum adversaries. The two most commonly used variants of LWE are Ring-LWE and Module-LWE, which operate on polynomials instead of vectors for efficiency, both of which can be accelerated using our processor.

Figure 2.3.2 shows the system block diagram, along with details of a typical Ring-LWE computation. A 24KB LWE cache interfaces with a modular arithmetic unit to perform polynomial operations, including the number-theoretic transform (NTT). An energy-efficient Keccak-f[1600] core, used for hashing and pseudo-random number generation (PRNG), drives the discrete distribution sampler. The LWE cache, the Keccak core and the sampler have dedicated clock gates, which can be independently configured for fine-grained power savings. The processor is equipped with a 1KB instruction memory. Custom instructions are applied to implement various lattice-based algorithms. The two most important computations required in all protocols are sampling and convolution. The polynomials are generated, or "sampled", either uniformly through rejection sampling or from a discrete distribution, typically binomial, with a carefully chosen standard deviation. Computing convolution of two polynomials involves transforming to the NTT domain followed by coefficient-wise multiplication and an inverse transform.

The hardness of the LWE problem is directly related to the statistical properties of the sampled polynomials. This makes an accurate and efficient sampler a critical component of any lattice crypto implementation. Sampling accounts for about 70% of the computational overhead in software implementations of lattice-based protocols [2]. Fig. 2.3.3 describes an energy-efficient discrete distribution sampler which reduces this overhead and provides up to two orders of magnitude energy savings over assembly-optimized software. Samplers post-process pseudo-random bit strings to generate numbers from a specified distribution, thus making an efficient PRNG a key requirement for energy savings. Hardware implementations of three standard PRNGs with full data-path architectures were profiled on our test chip, and SHA-3 (SHAKE) was observed to be 2× and 3× more energy-efficient than ChaCha20 and AES, respectively. Therefore, our PRNG consists of a 24-cycle 34K-gate Keccak-f[1600] core, which can be configured in different SHA-3 modes and consumes 0.89nJ per round. Our Keccak core processes its 1600b state in parallel, avoiding expensive register shifts and multiplexing required in serial architectures. The associated area overhead is very small, since the PRNG accounts for only 9% of the total processor area. Rejection sampling for primes with high rejection probability can be a bottleneck in LWE-based protocols. For faster rejection sampling, the rejection bound is set as a multiple of the prime modulus q [3] followed by Barrett reduction, providing up to 43% energy savings compared to conventional rejection. Our binomial sampler takes two k-bit chunks ($k \leq 32$, configurable) from the PRNG and computes the difference of their Hamming weights (HW) to generate a sample with standard deviation $\sigma = \sqrt{(k/2)}$. This method is 16× more energy-efficient than the conventional Knuth-Yao (KY) sampler [1, 4], and is also constant-time, eliminating potential timing side-channels.

Polynomial operations, such as NTT and convolution, account for about 30% of the computations. However, the associated memory and logic together occupy more than 75% of the total hardware area. Hardware architectures for NTT, first proposed in [1], consist of SRAM banks for storing polynomials along with a modular arithmetic unit to perform the butterfly computations. These memories are typically implemented using two-port [1] or four-port [4] RAMs, which can pose large area overheads in resource-constrained devices. To reduce this area, we implement the constant geometry NTT [5] and split each polynomial among 4 single-port RAMs, as shown in Fig. 2.3.4. Regular memory access patterns of the constant geometry NTT allow butterfly inputs and outputs to ping-pong between these single-port RAMs without any read or write hazards. This NTT architecture provides ~124K-gate area savings compared to the traditional approach, while still having enough memory to accommodate multiple polynomials required in lattice-based algorithms. The constant factors ω and ψ used in NTT-based negative-wrapped convolution are related as $\omega = \psi^2$ and $\omega^{-i} = \omega^{N-i}$, which is used to compress pre-computed tables stored in the NTT constants RAM by 38%. The butterfly, with a 24b data-path and configurable modulus q, is implemented as a unified Cooley-Tukey (CT) + Gentleman-Sande (GS) structure, eliminating the need for expensive bit reversals. The multiplier and adder/subtractor in the butterfly are re-used for coefficient-wise modular operations on polynomials.

Figure 2.3.5 demonstrates the configurability of our processor by benchmarking NIST Round 1 post-quantum protocols [6], such as Kyber, NewHope, R-EMBLEM and LIMA. Our hardware can be configured for polynomials of length (N) 64 to 2048, modulus q up to 24b, and discrete distributions with varying standard deviations, thus allowing the processor to tune the security level to provide energy scalability. When executing the Kyber-768 and NewHope-1024 key exchange schemes, our design is respectively 28× and 37× more energy-efficient than Cortex-M4 software, after accounting for voltage scaling. Moreover, post-quantum key exchange using our processor is 30× more energy-efficient than state-of-the-art pre-quantum ECC-based key exchange at the same pre-quantum security level.

Figure 2.3.6 compares this work with software implementation on ARM Cortex-M4, as well as previous work in custom hardware design for lattice-based cryptography. The proposed single-port RAM-based NTT architecture makes our design more area-efficient than [4]. Although the use of multiple parallel butterflies can reduce NTT energy [4], we have used a single butterfly since NTT is only a small fraction of the total computation. An energy-efficient SHA-3 core along with our fast sampling architecture provides 28× energy savings in binomial sampling compared to [4]. This work also demonstrates complete lattice-based protocols, while achieving more than an order of magnitude improvement in energy-efficiency over software.

The chip was fabricated in a 40nm LP CMOS process and supports voltage scaling from 1.1V down to 0.68V. All hardware measurements are reported at 12MHz and 0.68V. Our lattice cryptography processor occupies 106K NAND gate equivalents (GE) and uses 40.25KB of SRAM. It has an average power of 516μW when performing the NewHope post-quantum key exchange. Through architectural and algorithmic optimizations, this work demonstrates practical hardware-accelerated quantum-resistant lattice-based cryptographic protocols that can be used to secure resource-constrained IoT devices of the near future.

Acknowledgements:
The authors would like to thank Texas Instruments for funding this work, and the TSMC University Shuttle Program for chip fabrication support.

References:
[1] I. Verbauwhede, et al., "Circuit Challenges from Cryptography," *ISSCC*, pp. 428-429, 2015.
[2] T. Oder, et al., "Practical CCA2-Secure and Masked Ring-LWE Implementation," *IACR Trans. on CHES*, pp. 142-174, 2018.
[3] S. Gueron, et al., "Speeding up R-LWE Post-Quantum Key Exchange," *IACR Cryptology ePrint Archive*, Report 2016/467, 2016.
[4] S. Song, et al., "LEIA: A 2.05mm² 140mW Lattice Encryption Instruction Accelerator in 40nm CMOS," *IEEE CICC*, 2018.
[5] M. Pease, "An Adaptation of the Fast Fourier Transform for Parallel Processing," *Journal of the ACM*, vol. 15, no. 2, pp. 252-264, 1968.
[6] NIST, "Post-Quantum Cryptography Standardization," NIST PQC Round 1, 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

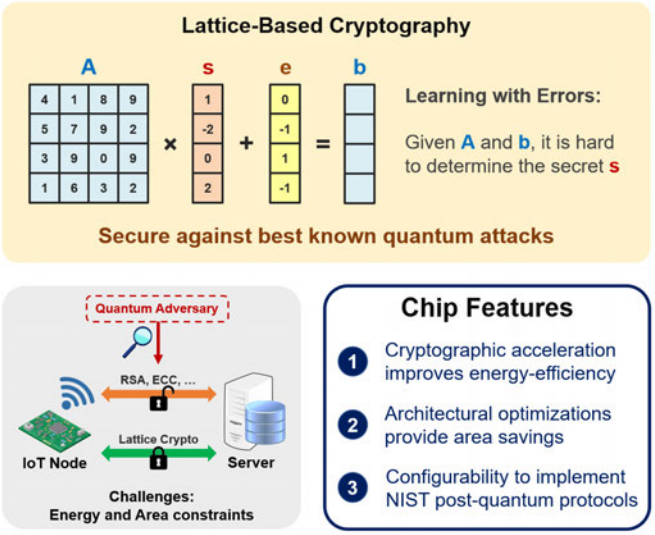


Figure 2.3.1: Quantum-resistant security for IoT networks; lattice-based cryptography, challenges and proposed hardware solutions.

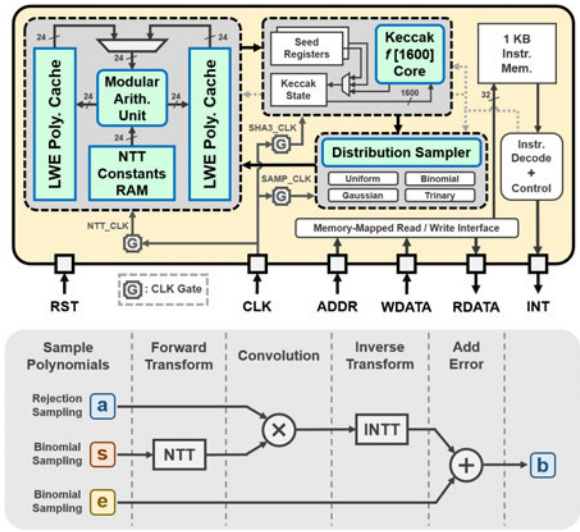


Figure 2.3.2: System diagram along with overview of a typical Ring-LWE computation.

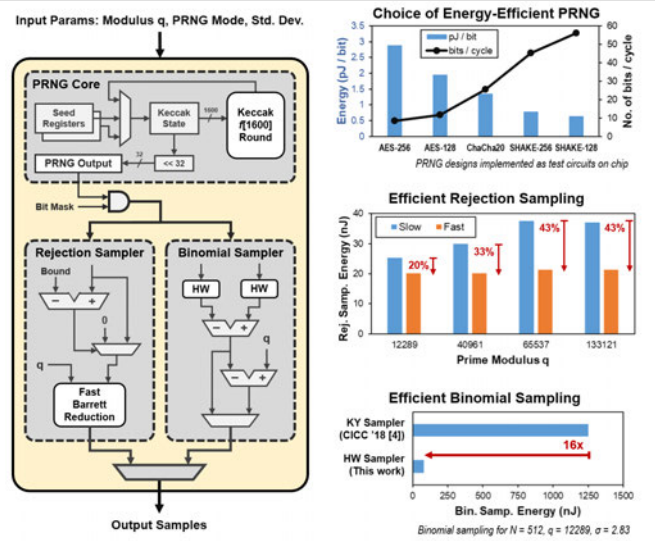


Figure 2.3.3: Architecture of discrete-distribution sampler with efficient PRNG and sampling.

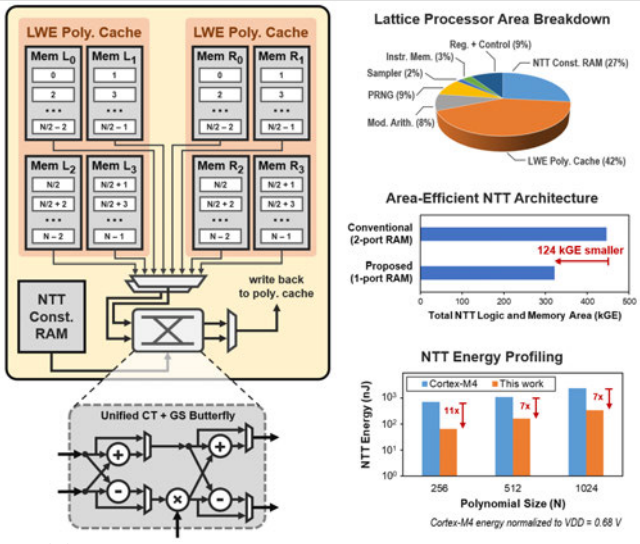


Figure 2.3.4: Proposed single-port RAM-based area-efficient NTT architecture with processor area breakdown and NTT energy profiling.

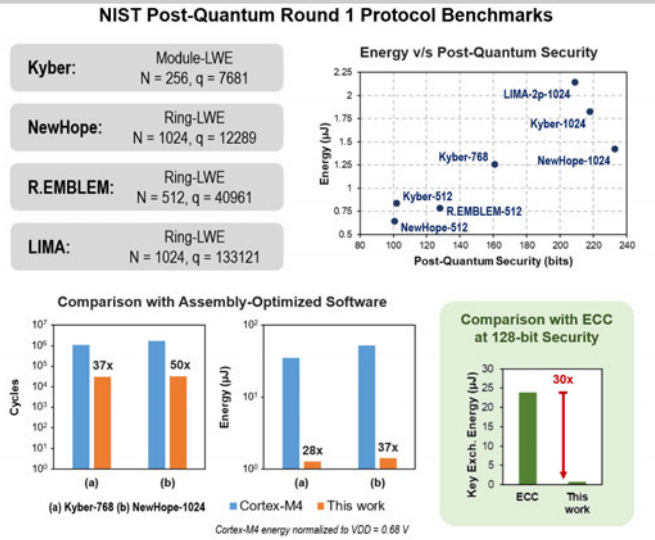
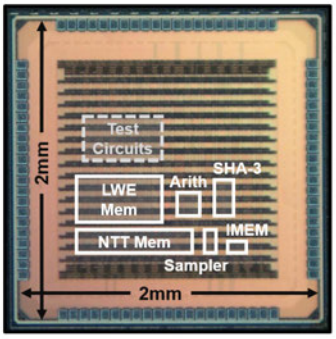


Figure 2.3.5: Configurability of the lattice cryptography processor along with NIST Round 1 post-quantum protocol benchmarks.

	Cortex-M4	ISSCC'15 [1] [a]	CICC'18 [4]	**This work**
Technology	–	130 nm	40 nm	40 nm
Supply Voltage (V)	3.0	1.2	0.9	0.68 – 1.1
Frequency (MHz)	100	500	300	12 – 72
Total Area (mm^2)	–	–	2.05	**0.28**
Logic Gates	–	–	–	106k
Supported Lattice Crypto Primitives	All (Software)	Ring-LWE	Ring-LWE	**Ring-LWE and Module-LWE**
Supported Lattice Parameters	All (Software)	N: 256 q: 7681	N: 64 to 2048 q: 32-bit config.	N: 64 to 2048 q: 24-bit config.
NTT Performance (N = 256, q = 7681)				
NTT Cycles	22031	1700 [b]	160	1288 [b]
NTT Energy (nJ)	13.5 x 10^3	–	31	**63.4**
Binomial Sampling Performance (N = 512, q = 12289)				
Sample Cycles	155872	–	3704 [c]	1009 [d]
Sample Energy (nJ)	95.8 x 10^3	–	1250	**44.4**

[a] Post-synthesis data reported for [1]
[b] NTT cycles in [1] and this work include the multiplication of coefficients with powers of ψ
[c] [4] uses ChaCha20 as PRNG for sampling
[d] This work uses SHAKE-256 as PRNG for sampling and cycles include the generation of pseudo-random bits

Figure 2.3.6: Comparison with Cortex-M4 software and hardware lattice cryptography accelerators.



Chip Specifications	
Technology	40 nm LP CMOS
Supply voltage	0.68 – 1.1 V
Package	64-pin QFN
Die size	2 mm x 2 mm
Cryptographic Core	
Total area	0.28 mm^2
Logic gates	106k (NAND2 equiv.)
SRAM	40.25 KB
Max. frequency	12 MHz at 0.68 V 72 MHz at 1.1 V
Lattice parameters	N: 64 to 2048 q: 24-bit configurable
Hash function	SHA-3-256 / 512
PRNG	SHAKE-128 / 256
Avg. power	516 μW at 0.68 V & 12 MHz (for NewHope-1024)

Figure 2.3.7: Chip micrograph and performance summary.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

2.4 A Distributed Autonomous and Collaborative Multi-Robot System Featuring a Low-Power Robot SoC in 22nm CMOS for Integrated Battery-Powered Minibots

Vinayak Honkote[1], Dileep Kurian[1], Sriram Muthukumar[1], Dibyendu Ghosh[1], Satish Yada[1], Kartik Jain[1], Bradley Jackson[2], Ilya Klotchkov[2], Mallikarjuna Rao Nimmagadda[1], Shreela Dattawadkar[1], Pranjali Deshmukh[1], Ankit Gupta[1], Jaykant Timbadiya[1], Ravi Pali[1], Karthik Narayanan[1], Saksham Soni[1], Saransh Chhabra[1], Praveen Dhama[1], N. Sreenivasulu[1], Jisna Kollikunnel[1], Sureshbabu Kadavakollu[1], Vijay Deepak Sivaraj[1], Paolo Aseron[2], Leonid Azarenkov[2], Nancy Robinson[2], Arun Radhakrishnan[3], Mikhail Moiseev[2], Ganeshram Nandakumar[1], Akhila Madhukumar[4], Roman Popov[2], Kamakhya P. Sahu[1], Ramesh Peguvandla[1], Alberto Del Rio Ruiz[3], Mukesh Bhartiya[1], Anuradha Srinivasan[2], Vivek De[2]

[1]Intel, Bangalore, India; [2]Intel, Hillsboro, OR; [3]Intel, Guadalajara, Mexico
[4]Intel, Toronto, Canada

Multi-robot systems, working collectively to accomplish complex missions beyond the capability of a single robot, are required for a wide range of applications such as search-and-rescue, precision agriculture and industrial automation. The collective behavior of a multi-robot system is governed by its continuous interactions with the physical environment, inter-robot information exchange, and intelligent decision making. Energy-efficient performance, size, weight and scalability of the key sensor, compute, communication, control and actuator components [1-3] of the individual robot platform are critical for realizing efficient advanced systems. In this paper, we demonstrate a *distributed, autonomous and collaborative* multi-robot system featuring integrated, battery-powered, crawling and jumping minibots for an example search-and-rescue application. We present a 16mm^2 low-power Robot SoC in 22nm CMOS that is integrated in the cm-scale minibot platform along with audiovisual and motion sensors, battery, low-power wireless communication and motion actuator components (Fig. 2.4.7).

In our example search-and-rescue application (Fig. 2.4.1), 4 minibots collaboratively navigate and map an unknown area without a central server or human intervention, detecting obstacles and finding paths around them, avoiding collisions, communicating among themselves, and delivering messages to a base station when a human is detected. Each minibot platform integrates: (i) camera, LIDAR and audio sensors for real-time perception and navigation; (ii) a low-power Robot SoC for sensor data fusion, localization and mapping, multi-robot collaborative intelligent decision making, object detection and recognition, collision avoidance, path planning and motion control; (iii) low-power ultra-wide-band (UWB) radio for anchorless dynamic ranging and inter-robot information exchange; (iv) long range radio (LoRa) for robot-to-base-station critical message delivery; (v) battery and PMIC for platform power delivery and management; (vi) 64MB pseudo-SRAM (PSRAM) and 1GB flash memory; and (vii) actuators for crawling and jumping motions.

The Robot SoC compute and control die (Fig. 2.4.2) integrates: (i) a real-time (RT) subsystem with an x86 host processor for sensor data grab and preprocessing; (ii) a Tensilica DSP processor for localization, mapping, collision avoidance and collaborative intelligent decision making; (iii) dedicated path planning (PP) and motion control (MC) hardware accelerators; (iv) an always-on (AON) subsystem with audio accelerator for human voice detection; (v) a memory subsystem with 8KB ROM, 2.86MB SRAM and 370KB register file (RF); (vi) an applications subsystem with a x86 processor and CNN accelerator for object detection and recognition; and (vii) die-to-die simplex link (D2DSL) interface to external chips and FPGAs in the platform.

Individual functions of a real-time search-and-rescue application are mapped onto the Robot SoC (Fig. 2.4.2). The complex application, running continuously for several hours with battery-powered minibots, demands execution of many performance-critical, compute-heavy and memory-intensive tasks with extreme energy efficiency. Hence, dedicated accelerator blocks are integrated on the SoC for the critical path planning and motion control functions.

The RT subsystem is designed to meet the strict real-time compute and deterministic interrupt latency requirements – particularly for sensor data acquisition/processing. RT supports simultaneous interactions with a large set of peripherals and sensors and enables the execution of time-critical dynamic-ranging capability essential for realizing a fully autonomous multi-robot system.

The audio accelerator in the AON subsystem receives 4-channel microphone input for detecting and identifying human voice. The tightly coupled memory (TCM) in the memory subsystem constitutes the next level of cache for the x86 application processor with 32KB L1 data and instruction caches. Cache misses result in data movements through the external PSRAM interface. Each bank of the ROM, SRAM and RF arrays in the memory subsystem can be power gated to minimize memory power at runtime.

Path planning by a robot to move efficiently from the current position to a target location, considering all the obstacles along the way, is extremely compute and memory intensive, and is typically off-loaded to a cloud server in state-of-the-art small form-factor robot platforms with limited compute and memory resources. In order to achieve smooth navigation and efficient mission execution in real time search-and-rescue in our battery-powered minibot platforms, we have integrated an energy- and area-efficient PP accelerator engine in the Robot SoC (Fig. 2.4.3). The PP accelerator implements (i) space discretization to determine the optimal direction of path exploration quickly, and (ii) a continuous space approach to accurately identify narrow paths for critical search-and-rescue navigation, to achieve effective path planning with limited compute and memory resources. The current robot position and the obstacle locations obtained from the LIDAR sensor are first mapped onto the entire area of interest divided into major grids. From the current position, neighboring grids are clipped by the clipper module and local targets assigned. Neighboring grids containing obstacles are sub-divided into finer grids. The transformed obstacle map with respect to the current robot position is then stored in a linked list to optimize memory usage without compromising accuracy. Thus, continuous obstacle-free spaces are utilized effectively and narrow feasible paths around obstacles are identified efficiently, while avoiding multiple readings of the complete obstacle map. Parameterized trajectory generators (PTG) find the edges connecting the current robot position to local targets. In conventional implementations, all of the reference trajectories are typically stored in memory. Here, the trajectories are computed by the PTGs *in real time on the fly,* thus avoiding look-up tables for storing pre-computed trajectories. Once the exploration is complete, the start position is backtracked from the target location, and the nodes and edges connecting the start and end points are stored in TREE memory.

The MC accelerator (Fig. 2.4.3) enables complete closed loop autonomous control for any minibot motion and helps meet the real-time operation requirements. It provides the capability to control with high-precision multiple-motion actuators of different types in parallel. The PWM module, configurable in one-shot or continuous mode, outputs four 10b resolution pulses to the motor drivers based on controller commands. The encoder module determines directionality and the filter module implements a moving average filter with configurable window size for required accuracy. A fully configurable PID controller is used where parameters can be changed on the fly and the sampling interval can be set to provide precise timing control resulting in accurate odometry.

Eight SoC power states and event-based transitions are implemented as a state machine in hardware. The SoC consumes 14.9mW, 30.3mW, 19.8mW and 36.9mW, in AL, AO, AS and AA power states, respectively (Fig. 2.4.4). Platform memory organizations across the SoC memory subsystem, PSRAM and flash, and memory utilizations for different compute and communication functions in the example application are shown in Fig. 2.4.4. Measurements demonstrate efficient and accurate operation of the PP & MC accelerators in the Robot SoC (Fig. 2.4.5). The PP accelerator operates at 75 to 160MHz (0.6 to 1V) and the MC accelerator operates at 0.5 to 4MHz (0.65 to 0.8V), consuming 6 to 37.5mW and 1 to 6mW, respectively. The Robot SoC operates over a wide voltage-frequency range of 80 to 365MHz (0.6-1V), consuming 32 to 238mW, and achieving minimum energy consumption of 0.29nJ at 0.65V, 140MHz. Platform measurements (Fig. 2.4.6) for the end-to-end search-and-rescue application for the fully distributed, autonomous and collaborative battery-powered, cm-scale minibots demonstrate accurate and reliable real-time energy-efficient performance, enabled by major improvements in energy consumption and memory usage achieved by the Robot SoC with integrated PP and MC accelerators, compared to an Edison-based reference platform.

References:
A. Amravati, et al., "A 55nm Time-Domain Mixed-Signal Neuromorphic Accelerator with Stochastic Synapses and Embedded Reinforcement Learning for Autonomous Micro-Robots," *ISSCC*, *pp.* 124-126, 2018.
X. Zhang, et al., "A Multi-Chip System Optimized for Insect-Scale Flapping-Wing Robots," *IEEE Symp. VLSI Circuits*, *pp.* C152-C153, 2015.
A. Suleiman, et al., "Navion: A Fully Integrated Energy-Efficient Visual-Inertial Odometry Accelerator for Autonomous Navigation of Nano Drones," *IEEE Symp. VLSI Circuits*, *pp.* 133-134, 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

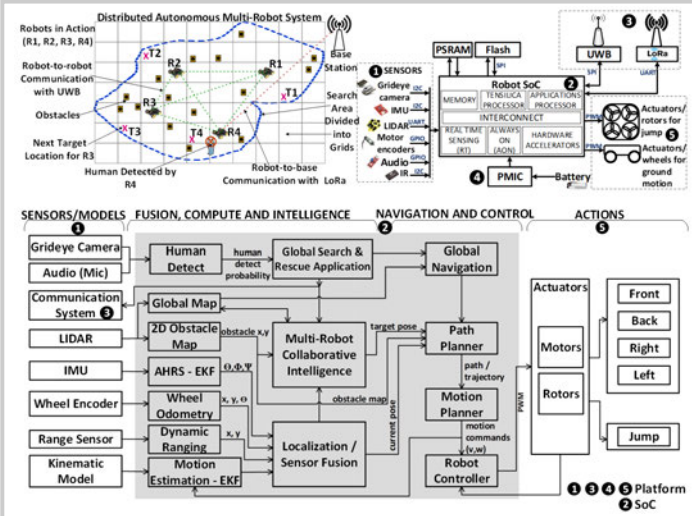


Figure 2.4.1: System architecture diagram and algorithmic flow for search and rescue application of the distributed, autonomous and collaborative multi-robot system.

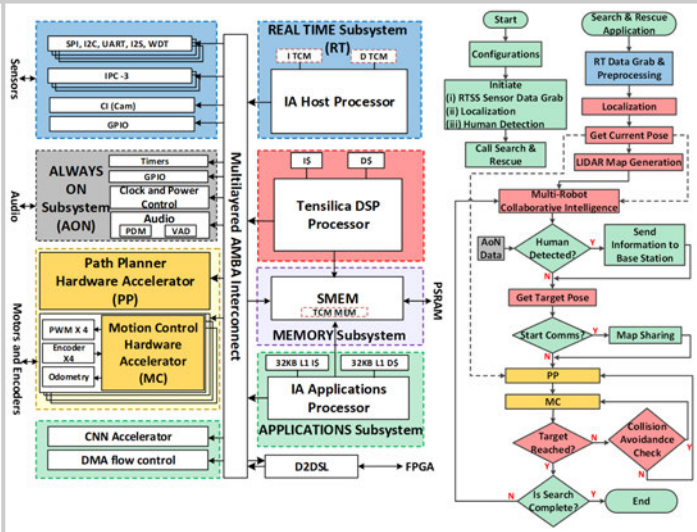


Figure 2.4.2: Robot SoC microarchitecture details and algorithm flow for search and rescue application.

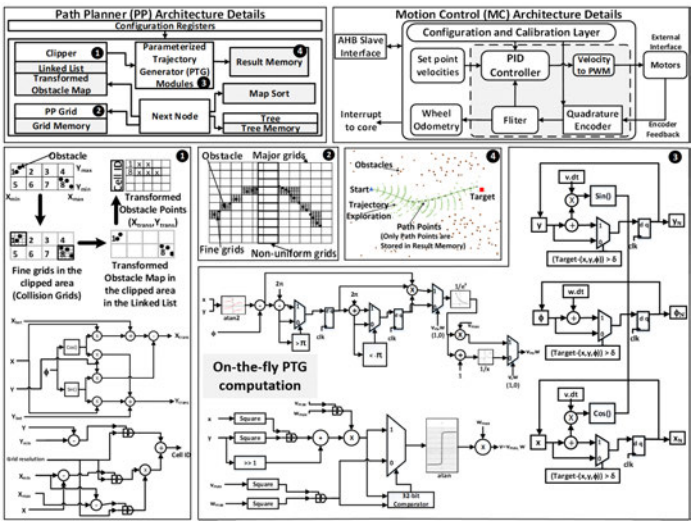


Figure 2.4.3: Architectural details and key features of the path planner (PP) and motion control (MC) hardware accelerators.

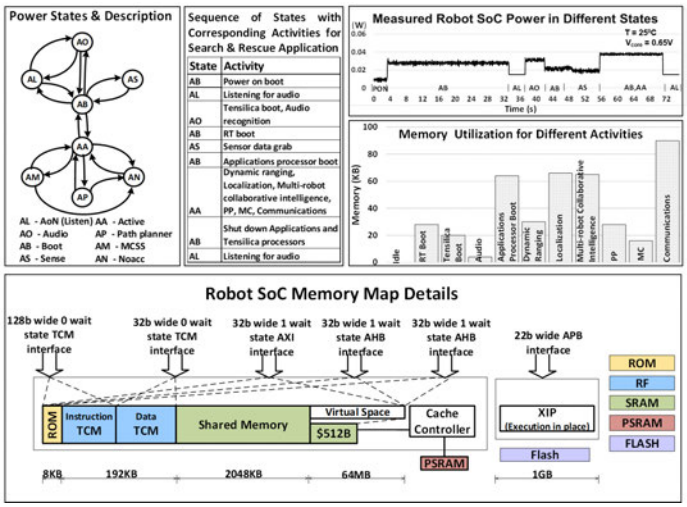


Figure 2.4.4: Robot SoC and system power management features; Robot SoC measured power and memory for a sample application; Robot SoC memory organization and mapping.

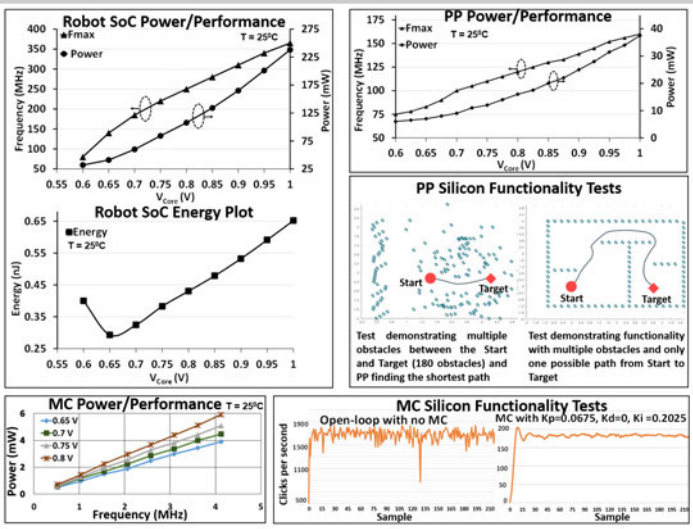
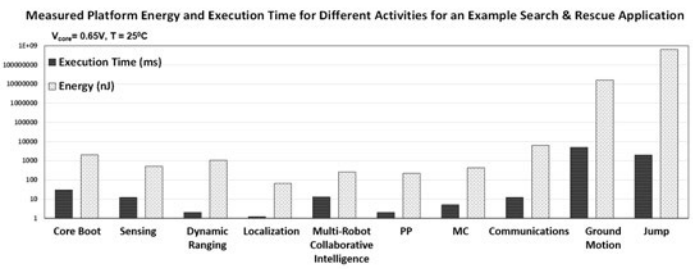


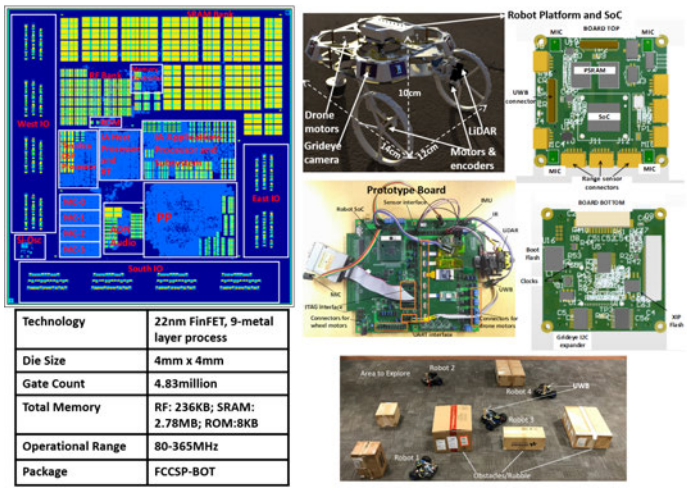
Figure 2.4.5: Measured voltage, frequency, energy data and functionality demonstration for key IP blocks in the Robot SoC.



Functionality	Memory (KB) Robot SoC	Memory (KB) Edison	Execution Time (ms) Robot SoC	Execution Time (ms) Edison	Energy (nJ) Robot SoC	Energy (nJ) Edison
Localization	62	429	0.5	1.3	2.75E+04	2.02E+05
Collaborative Intelligence	61	66	11	13.6	8.91E+05	3.40E+06
Path Planner	28	20145	2	5.3	2.40E+04	9.09E+05
Motion Control	18	25	0.011	0.49	4.40E+01	7.02E+04
Total for All the Functionalities with Search & Rescue Application	560	24000	37	98.9	3.00E+06	3.25E+07

	This work	[1]	[2]	[3]
Technology	22nm	55nm	40nm	65nm
Area	16mm²	3.4mm²	5.4mm²	20mm²
Supply voltage	0.6-1V	0.4-1V	0.6-0.9V	Unavailable
Operational range	80-365MHz	0.0005-40MHz	30-220MHz	83.3MHz
Power	37mW	690uW	41.6mW	24mW
Circuit	Digital	Time domain mixed signal	Digital	Digital
Multi-robot collaboration	Yes	No	No	No

Figure 2.4.6: Measured system-level data for performance, execution time and comparison with Edison platform.



Technology	22nm FinFET, 9-metal layer process
Die Size	4mm x 4mm
Gate Count	4.83million
Total Memory	RF: 236KB; SRAM: 2.78MB; ROM:8KB
Operational Range	80-365MHz
Package	FCCSP-BOT

Figure 2.4.7: Robot SoC chip micrograph, platform, form factor and representative multi-robot collaboration pictures.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

2.5 A 40×40 Four-Neighbor Time-Based In-Memory Computing Graph ASIC Chip Featuring Wavefront Expansion and 2D Gradient Control

Luke R. Everson, Sachin S. Sapatnekar, Chris H. Kim

University of Minnesota, Minneapolis, MN

Single-source shortest path (SSP) problems have a rich history of algorithm development [1-3]. SSP has many applications including AI decision making, robot navigation, VLSI signal routing, autonomous vehicles and many other classes of problems that can be mapped onto graphs. Conventional algorithms rely on sequentially traversing the search space, which is inherently limited by traditional computer architecture. In graphs which become very large, this slow processing time can become a bottleneck in real world applications. We propose a time-based ASIC to address this issue. Our design leverages a dedicated hardware implementation to solve these problems in linear time complexity with superior energy efficiency. A 40×40 four-neighbor grid implements a wavefront (WF) expansion with a first-in lockout mechanism to enable traceback. Outside the array, a programmable resistive ladder provides bias voltages to the edge cells, which enables pulse shaping reminiscent of the A* algorithm [3].

Figure 2.5.1 provides a high-level schematic of the chip. The chip is structured to model a graph with a regular Manhattan grid structure. Each of the 1600 vertices have four connections to its neighbors in the cardinal directions (N, S, E, and W). The chip functions by propagating a pulse between the vertices through the edges. The time it takes for a pulse to travel from each cell is proportional to the distance, or cost, to travel through that edge. Each vertex operates autonomously; it senses and stores the direction of the input, prevents other pulses from overwriting it, and propagates it to neighboring stages. The first pulse, or set of simultaneous pulses, represents the fastest way to reach that cell. Since the first pulse is the only pulse latched in each cell, tracing the pulse chain back to the start will reveal the shortest path. Although the core was initially designed for SSP, each evaluation contains all shortest paths to the start node.

Figure 2.5.2 describes the schematic of the vertex and corresponding timing diagram. The functionality of the cell will be described with an example of two pulses arriving: North and then South. First, a global enable signal, *EN*, is asserted enabling the core. Additionally, not shown in this figure, a pulse is started from somewhere in the array. The four inputs are merged together in a detection circuit to determine if a pulse has arrived in the cell. In the example, *IN<N>* will flow through and latch P_{IN}, or Pulse Input. P_{IN} is compared with the input from each of the four directions. If P_{IN} is asserted and the input is not from the direction that asserted P_{IN}, Pulse Latch, in this example *PL<{S, E, W}>*, will assert. *PLb* is connected to the SRAM DMA, which will flip the *IP<{S, E, W}>* SRAM that will be read out after evaluation during the path traceback. *IP<3:0>* shows how the vertex stores the input pulse. Initially all four bits are cleared. The bits that do not correspond to the first input are flipped with the DMA circuit. This notation corresponds to the Colormap Readout Key in Fig. 2.5.4. Finally, the pulse is propagated to the neighbors that were not responsible for latching the cell and have a connection stored in their respective local SRAMs. Connections are predetermined based on the description of the graph at runtime.

Figure 2.5.3 shows the edge schematic. Each edge consists of one current-starved inverter, one standard inverter, and four binary weighted bits of capacitor loading. The delay is modulated by the 4b weight stored locally in the SRAM and the voltage bias applied to each branch of the first inverter. The delay of each unit, defined as a vertex and an edge, is shown for different bias voltages and sets of weights. The top left colormap shows the simulated delays for each unit with the biases V_{X3} and V_{Y3} applied at the closed switches. V_{H3} and V_{V3} are larger than $\{V_{H1}, V_{H2}\}$ and $\{V_{V1}, V_{V2}\}$. V_{H3} and V_{V3} are applied through a transmission gate controlled by the scan bits. All other points are held open which yields a linear voltage increase from the edges at each stage until target stage 30, illustrated in the lower left figure.

Figure 2.5.4 details examples from the SSP application and collision avoidance (CA), or minimal edge-effect problem. CA is useful if there is an incentive to avoid obstacles, such as self-driving cars or drone navigation. In this application, the pulse is started simultaneously from the sides of the obstacles. Where the WFs meet, shown in white, signifies the path that maximized the distance between the obstacles. In SSP, we show two outputs generated from the same map under different conditions. In this application, a map is shown with blockages shown as black blocks. The WF is initiated in the upper left and it propagates down and across the core. The top figure does not have a voltage gradient applied and each edge has the same weight. This gives the WF a very regular pattern as it traverses the map. The bottom figure has a voltage gradient applied that is weakest in the top-left corner and strongest in the bottom-right corner. The key difference between the two outputs is shown in lower-right, where the yellow cells signify that the upper half supersedes the lower half. Without the applied bias the pulses nearly arrived at the same time (barrier between blue and orange), however, with the voltage gradient, the pulse traverses the blockages faster.

This ASIC is not constrained to solving problems that conform to a 40×40 grid. Fig. 2.5.5 highlights the scalability of this core via a four-core example with a single blockage spanning three cores. First, core 0 is evaluated and the WF reaches cores 3 and 1, both at two points, which will be used to start the pulse in subsequent evaluations. Next, core 3 is evaluated, but the pulse does not uncover much of the map due to the complete vertical blockage. After this, core 1 is evaluated and it contacts core 2 in two locations: the bottom right corner and directly above the blockage. These two cells are used to start the evaluation for core 2. Finally, it is revealed that core 2 and core 3 share an unexplored boundary, and core 3 is re-evaluated to fully uncover the obstacle. It is important to note that multicore still solves the SSP problem in linear time as the grid size increases.

Motivated by [1], we set out to recreate the optics experiments in Manhattan geometries to illustrate the utility of our hardware architecture. Shown in the lower half of Fig. 2.5.6 is a sampling of the core readout at different time points during a single evaluation of a two-slit experiment. The pulse starts to the left and above of the first slit. The WF traverses the first block and then passes into the second block. Next, the WF reaches the middle of the second boundary and then spreads out until it reaches the two-slits and generates two leading WF. This behavior mimics what is reported in [1] and has interesting consequences for future physical explorations of novel applications with low-power CMOS. The comparison table in Fig. 2.5.6 is mainly used for general comparisons, since to our knowledge, this is the first ASIC for graph traversal. The peak power is quoted when all 156 perimeter vertices are evaluating corresponding to the pulse originating from the center, equating to 183.1μW/vertex. 55% of the power is due to SRAM access storing the pulse information in-situ. Compared to state-of-the-art FPGA [4], μProcessor, CPU, and GPU [5] implementations, our core has roughly 6 orders of magnitude superior energy efficiency. In this work, we described an in-memory computing ASIC graph processor and highlighted the versatile applications and scalability of the proposed chip.

Acknowledgements:
This research was supported in part by the National Science Foundation under award number CCF-1763761.

References:
[1] C. Y. Lee, "An Algorithm for Path Connections and Its Applications", *IRE Transactions on Electronic Computers*, vol. EC-10, no. 3, pp. 346-365, Sept. 1961.
[2] E. W. Dijkstra, "A Note on Two Problems in Connexion with Graphs", *Numerische Mathematik,* vol. 1, pp. 269–271, 1959.
[3] P. E. Hart, N. J. Nilsson and B. Raphael, "A Formal Basis for the Heuristic Determination of Minimum Cost Paths," in *IEEE Trans. on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100-107, 1968.
[4] S. Zhou, C. Chelmis and V. K. Prasanna, "High-Throughput and Energy-Efficient Graph Processing on FPGA," *IEEE International Symp. on Field-Programmable Custom Computing Machines*, pp. 103-110, 2016.
[5] Y. Zhou and J. Zeng, "Massively Parallel A* Search on a GPU," Conference on Artificial Intelligence (AAAI), pp. 1248-1254, 2015.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

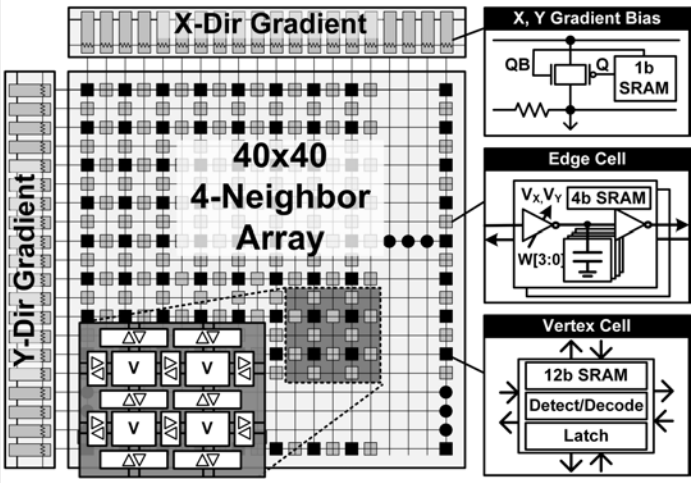


Figure 2.5.1: Proposed 40×40 graph ASIC chip for solving single-source shortest path problems based on 2-dimensional wavefront expansion.

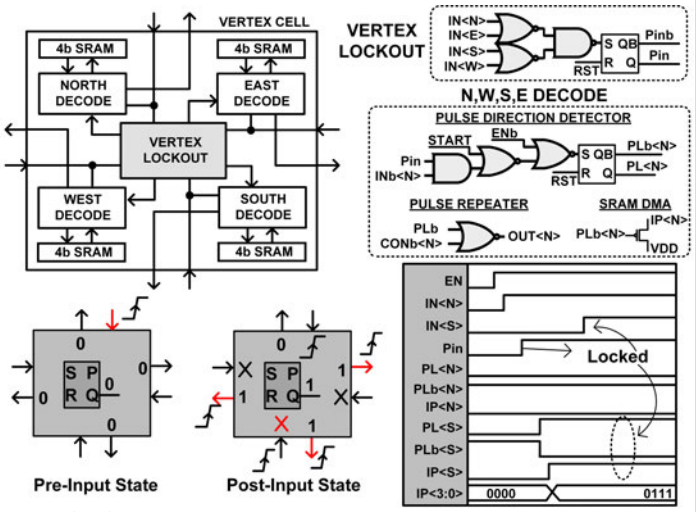


Figure 2.5.2: Details of the vertex cell, including the implementation of lockout and decode logic. Timing diagram shows lockout procedure with pre- and post-input state of vertex cell.

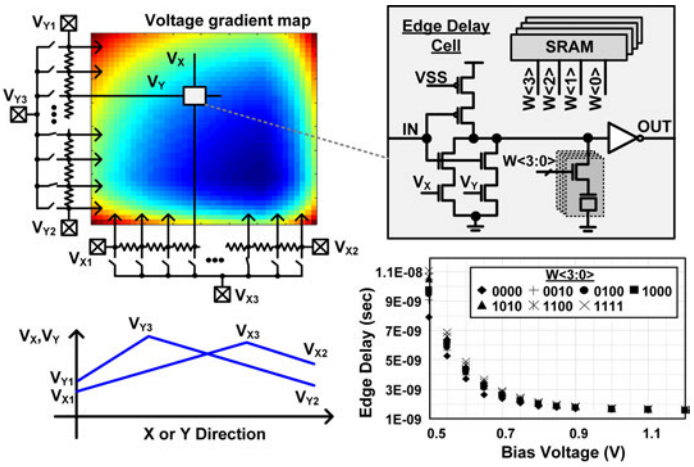


Figure 2.5.3: Voltage gradient map for proposed graph chip and schematic of edge cell with digital and analog delay control options; analog control voltage profiles in x and y directions; measured delay vs. bias voltage for different digital capacitor loads.

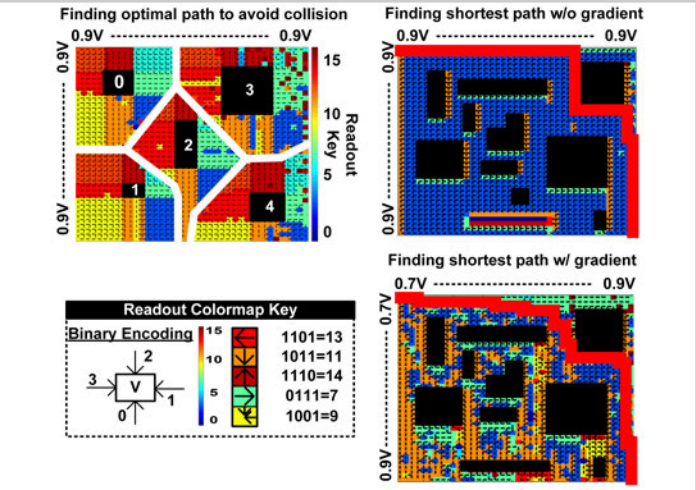


Figure 2.5.4: Test chip measurement results. Left: example of Collision Avoidance application where black boxes represent obstacles and white lines denote optimal path to avoid obstacles. Right: A* shortest path application shows routes with and without voltage gradient to accelerate wavefront.

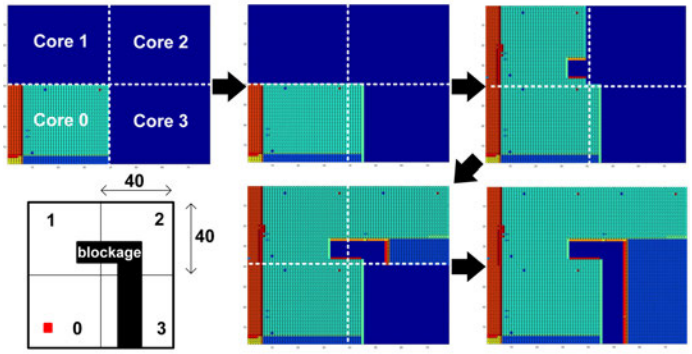


Figure 2.5.5: Four-core example shown to highlight scalability of the graph ASIC for larger maps. 4 cores can be stitched together or alternatively, a single core can be reused 4 times. The start node of next core is determined by the point of first impact on shared edge. Even with blockages spanning multiple cores, the framework can successfully traverse the entire map.

Architecture	This Work	FPGA [4]	µProcessor	CPU [5]	GPU [5]
Product	ASIC	Xilinx Virtex	ARM Cortex-M0	Intel Xeon E5630	NVIDIA Tesla K20c
Technology	65nm	20nm	40nm	32nm	28nm
Voltage	1.2V	-	1.1V	0.7-1.35V	-
Peak Power	26.4mW	24.22W	127µW	20W/core	225W
Throughput [MTEPS]	559	731	$5.34*10^{-4}$	0.83	9.0
Energy per Node	0.328pJ*	33nJ	89.1nJ	24.1µJ	25µJ
Normalized Energy	1x	10^5	$2.7x10^5$	$1.19x10^6$	$2.3x10^7$

*55% from SRAM Program (does not include cache access energy), Energy/Node=Unit Delay*Unit Power
MTEPS = Million Traversed Edges Per Second

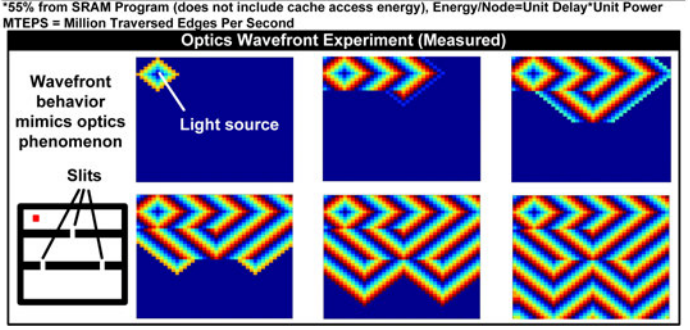
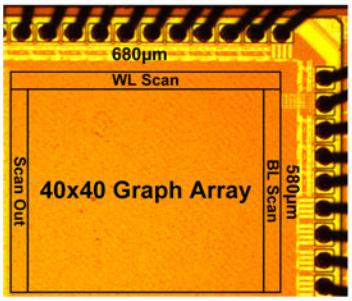


Figure 2.5.6: Comparison table shows 5 orders of magnitude improved energy efficiency over traditional implementations (upper). Measured results of scientific computation (i.e. optics wavefront propagation in two-slit experiment) (lower).



Applications	A* shortest path, obstacle avoidance, scientific computation (optics)
Technology	65nm LP CMOS
Architecture	Time-based
# of Vertices	1600
# of Edges	6400
Edge Resolution	4b + Analog Gradient
Voltage	1.2V
Peak Power	26.4mW
Delay per Node	1.79ns @ [V_B=.9V, V_{DD}=1.2V]
Power per Node	183.1μW
Energy per Node	0.238pJ

Figure 2.5.7: Die photo and chip summary.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

2.6 A 2×30k-Spin Multichip Scalable Annealing Processor Based on a Processing-In-Memory Approach for Solving Large-Scale Combinatorial Optimization Problems

Takashi Takemoto[1], Masato Hayashi[2], Chihiro Yoshimura[2], Masanao Yamaoka[2]

[1]Hitachi, Sapporo, Japan
[2]Hitachi, Tokyo, Japan

The last decade has seen impressive progress in the development of a new computer architecture, commonly known as annealing processor [1, 2]. An annealing processor provides a fast means for finding the ground state of an Ising model; thus, it can efficiently solve NP-hard combinatorial optimization problems [3]. In addition to quantum annealers based on superconducting circuits [1], annealing processors based on CMOS technology have received increased interest and are being developed on the basis of simulated annealing (SA) [2]. However, these CMOS annealing processors (CMOS-APs) have room for improvement, such as: i) expanding the bit widths of coefficients, and ii) increasing the number of spins handled by the processor. To address these challenges, a CMOS-AP based on the processing-in-memory approach (where CMOS circuits and an SRAM are tightly coupled [4]) has been developed. Its key features are threefold: a spin operator (processing local memory) which provides coefficients with expandable bit width and fast parallel spin updates according to the Gibbs distribution; a low-latency inter-chip interface (I/F) connecting two Ising chips, resulting in an increased number of spins; and a highly integrated spin circuit which directly connects the spin operator with the SRAM cell. Installed in a 2×30k spin system, the CMOS-AP demonstrates the capability for multi-chip operation with energy efficiency 1.75×10^5 higher than running SA on a CPU.

Annealing processors solve combinatorial optimization problems based on the calculation flow shown in Fig. 2.6.1. The cost function of the optimization problem is given by the Ising model consisting of σ_i, $i = 1,...,N$ spin variables, which can take the values –1 or 1. The energy of a spin configuration is given by:

$$H(\boldsymbol{\sigma}) = -\sum_{i,j=1,i<j}^{N} \sigma_i J_{ij} \sigma_j - \sum_{i=1}^{N} h_i \sigma_i ,$$

where coefficients J_{ij} and h_i represent spin-spin couplings and external magnetic fields, respectively. In the annealing process, each spin is iteratively updated according to an annealing schedule, whose temperature drops gradually. The resulting spin configuration of minimal energy represents the optimal solution of the optimization problem. There are two approaches for implementing annealing processors. The CPU-based approach stores spin values and coefficients in a discrete memory. In contrast, the processing-in-memory approach [4] stores them on an on-chip memory in the annealing processor. The CPU-based approach requires $N^2\times k$-communications between the CPU and memory for handling k annealing steps. Thus, its run-time increases significantly as the size of the problem increases. In contrast, the processing-in-memory approach only requires local communication between adjacent spins; it thus permits construction of a large-scale spin system where the communication time increases linearly in system size N.

Figure 2.6.2 shows a block diagram of the proposed CMOS-AP chip, which consists of an Ising core, a random pulse generator (gen.), an inter-chip I/F, and a control unit. The topology of the Ising model forms a square King's graph with 30,976=176×176 spins. Spin updates inside the Ising core are synchronized with a 100MHz clock. The annealing process performs two functions: (a) based on the spin couplings, the spin state is updated towards lower energy, while (b) thermal fluctuation based on the Gibbs distribution allows escape from local minima (annealing process in Fig. 2.6.1). Function (a) evaluates $H(\sigma)$ locally and finds the lower-energy state by majority voting of adjacent spin values, using the coefficients $J_X[2:0]$ and $h_0[2:0]$. Function (b) is reproduced by the exclusive-OR (EOR) of a random pulse, Rand_G, and the coefficients, leading to a spin state of higher energy when Rand_G is 1. Rand_G is generated by comparing the output of 32b XorShift with threshold P_{TH} ("Threshold gen." in Fig. 2.6.2).The ratio of the temperature-induced spin flips (mark ratio in Rand_G) gradually decreases and finally becomes 0 according to the temperature schedule in the threshold gen. block. Updating spins iteratively until the temperature becomes 0 gives a near-optimal spin configuration with low energy.

SA permits simultaneous spin updates for those which are disconnected from one another. To achieve a fast annealing time, the CMOS-AP adopts a sparse King's graph. However, even if such a graph is selected, as shown in the spin update rule in Fig. 2.6.3, at least four clock cycles are required to update all spins, because the spins are divided into four groups. The waiting time between updates of each spin is efficiently used to realize a larger-spin CMOS-AP, namely, a CMOS-AP with inter-chip connections and a shared spin circuit (Fig. 2.6.4). The inter-chip I/F, connecting CMOS-AP chips A and B (without affecting the runtime and results of the annealing process) is shown in the block diagram in Fig. 2.6.3. On chip A, the 176b spin data, i.e., 2×88b of spin data from groups SP1 and SP3 on the right side of the Ising core, are split into 8×22b chunks and transmitted through the 100Mb/s LVDS I/F to chip B in sequence. In chip B, the received spin data is reassembled into 2×88b form and fed into dummy spins consisting of flip-flops. The spin data is simultaneously transmitted and received in both directions between chips A and B. As shown in the timing chart in Fig. 2.6.3, the dummy spins are updated every 11 clock cycles. This implies low latency, since the spins inside the Ising core are updated only twice during this period.

To create a large-scale spin system, it is necessary to downsize the spin circuit, which is the base unit of the CMOS-AP. To effectively use the spin operator, which dominates the area of the spin circuit, it is shared among four spins according to the above-mentioned spin-update rule (Fig. 2.6.4). The updated spin values are directly written to the memory cells through the tristate buffers. In addition, the read coefficients are directly connected to the selector through a spin-to-spin read-write circuit for updating the spin value. The gate width of the cell transistors is adjusted to enlarge the hold margin against the imbalance of the cell-storage-node loads, which are caused by parasitic load capacitances in the spin-to-spin read-write circuit. The short bitline with 56 memory cells enables reduction of the write-amplifier load and a stable sense readout, using only the inverter. As a result, the spin circuit area is 1,225μm^2/4spins, which is more than 50% smaller than the unshared approach.

The CMOS-AP is fabricated in 40nm CMOS process. The chip size is 4.3×5.5mm. As shown in Fig. 2.6.5 (top), chips A and B are mounted on an evaluation board along with a control FPGA. As a first evaluation result, we measured one-spin flip ratios, as shown at the bottom of Fig. 2.6.5. Both for strongest and weakest coupling condition, where adjacent coefficients sum to 27 and 1, respectively, the measured spin-flip ratios agree well with the theoretical values based on the Gibbs distribution over a wide temperature range.

In the top of Fig. 2.6.6 we compare annealing results of the 2×30k spin system to optimized SA on a CPU, using problems with randomly allocated coefficients. The annealing time (without I/O time) is 22μs (=22 clock cycles×100 annealing steps), which is 2.6×10^4-times faster than that of the CPU. The CMOS-AP further operates much faster than our previous chip [2], which took 10ms for a similar problem. This speedup is attributed to the improved sampling of the Gibbs distribution (Fig. 2.6.5). The relative difference between final energy states, obtained from the CPU and CMOS-AP, is less than 1.2%. The lower left of Fig. 2.6.6 shows the transition of spin configurations when solving a max-cut problem revealing an "ABCD" pattern. In the intermediate temperature state (ii), a border line between chips is observed due to inter-chip connection latency. This border line disappears in the final low-temperature state (iii), where spin flips become rare and latency becomes negligible. As shown in the lower right of Fig. 2.6.6, for systems with 60k spins, the power efficiency including I/O time is 1.75×10^5-times higher than running the "SG3" algorithm on the CPU. Photos of the chip and the credit-card-sized evaluation board are shown in Fig. 2.6.7.

Acknowledgements:
This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References:
[1] M. Johnson, et al., "Quantum Annealing with Manufactured Spins," *Nature*, vol. 473, pp. 194-198, 2011.
[2] M. Yamaoka, et al., "20k-spin Ising chip for Combinational Optimization Problem with CMOS Annealing," *ISSCC*, pp. 432-433, 2015.
[3] A. Lucas, "Ising Formulations of Many NP Problems," *Front. Phys.*, vol. 2, no. 5, pp. 1-15, 2014.
[4] A. Biswas and A. P. Chandrakasan, "Conv-RAM: An Energy-Efficient SRAM with Embedded Convolution Computation for Low-Power CNN-Based Machine Learning Applications," *ISSCC*, pp. 488-489, 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

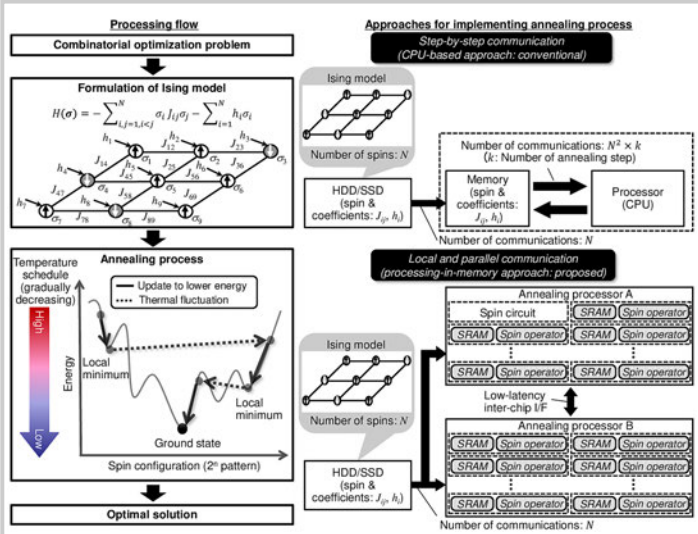


Figure 2.6.1: Processing flow of annealing processor (left) and approach for implementing the annealing process (right).

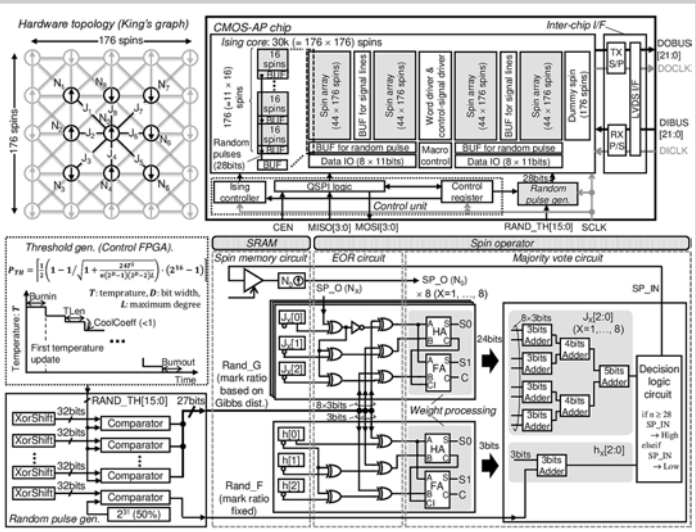


Figure 2.6.2: Block diagram of one CMOS-AP (top) and spin circuit for N_0 spin (bottom).

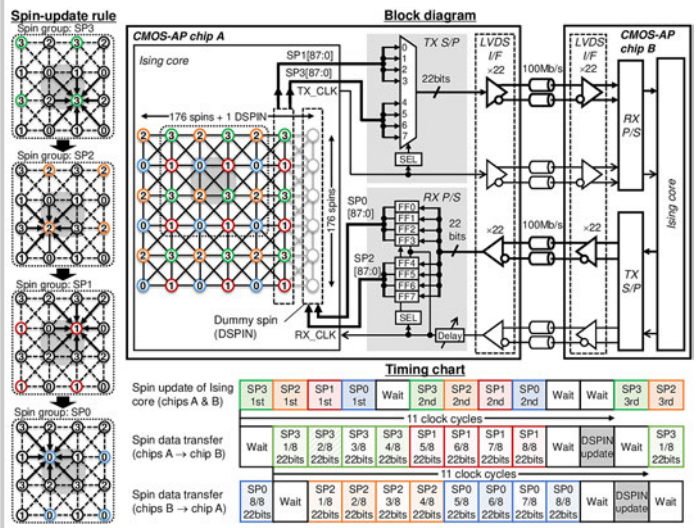


Figure 2.6.3: Spin update rule (left) and block diagram of inter-chip interface along with timing chart (right).

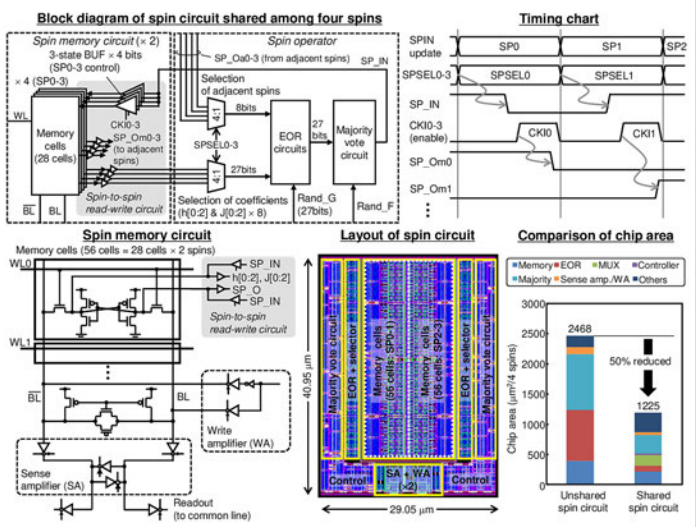


Figure 2.6.4: Block diagram of highly integrated spin circuit (top) and its layout (bottom).

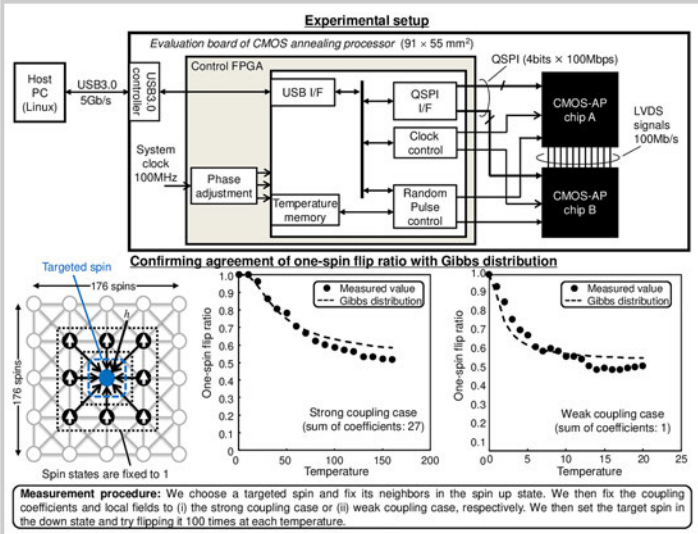


Figure 2.6.5: Experimental setup (top) and measured one-spin flip ratio (bottom).

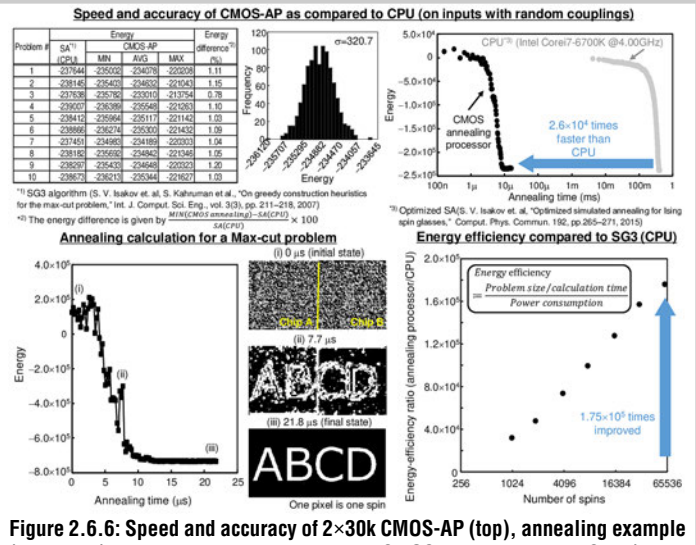


Figure 2.6.6: Speed and accuracy of 2×30k CMOS-AP (top), annealing example (lower left), and energy efficiency ratio of CMOS-AP compared to CPU (lower right).

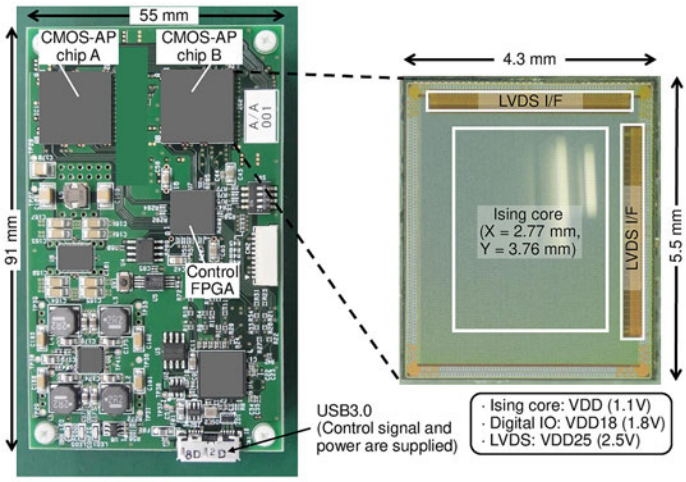


Figure 2.6.7: Chip micrograph with fabricated credit-card-sized 2×30k-spin annealing processor.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

2.7 A 28nm 600MHz Automotive Flash Microcontroller with Virtualization-Assisted Processor for Next-Generation Automotive Architecture Complying with ISO26262 ASIL-D

Sugako Otani, Norimasa Otsuki, Yasufumi Suzuki, Naoto Okumura, Shohei Maeda, Tomonori Yanagita, Takao Koike, Yasuhisa Shimazaki, Masao Ito, Minoru Uemura, Toshihiro Hattori, Tadaaki Yamauchi, Hiroyuki Kondo

Renesas Electronics, Kodaira, Japan

Automotive architecture has been rapidly progressing toward integration and centralization [1]. Integration of multiple electronic control units (ECUs) reduces vehicle weight by deleting the wire harness between ECUs and the cooperative control of multiple functions, resulting in energy savings. The power consumption of automotive MCUs must be single-digit Watts under thermal constraints, and many functions have to be squeezed into one MCU. Adoption of a 28nm process is necessary for speed, power and density, because a conventional 40nm embedded flash process is insufficient on these axes. Furthermore, the integration of ECUs also mixes software components with varying safety integrity levels in one MCU. Resource isolation is inevitable for functional safety to avoid latent faults. In the autonomous driving era, data size and performance requirements will increase owing to the number and complexity of sensors requiring fusion in real time. Fast networks within vehicles are vital. With the evolution of sensor fusion, information centralization, and control decentralization, automotive architectures have been progressing toward increased integration and centralization.

We developed a 28nm automotive MCU with 16MB of flash memory and four 600MHz CPUs with a lock-step mechanism, targeting a 5W power limit (Fig. 2.7.1). The MCU also realizes following integration requirements: the highest automotive safety integrity level (ASIL-D) and network performance improvements. Three key features of the MCU are: 1) a virtualization-assisted processor for functional safety, 2) a logic built-in self-test in the field (field-BIST) at standby resume for functional safety, and 3) a SGMII (serial gigabit media independent interface) for an in-vehicle network.

Figure 2.7.1 shows the overall architecture of the automotive control system. The system is divided by a firewall into two subsystems, the secure system and the application (non-secure) system. In the non-secure system, multiple integrated applications are executed. With the deployment of multicore in automotive applications, virtualization, originally used in server systems, is vital for handling increased complexity and the needed computation power. The hypervisor software realizes virtualization by emulating virtual hardware on physical hardware to execute several different virtual machines (VMs) without interference. However, virtualization in automotive-control applications has been limited by the overhead of hypervisors. To minimize the overhead, we developed a virtualization-assisted processor.

Virtualization overhead consists of VM context switching time and interrupt distribution processing time by the hypervisor software. Fig. 2.7.2 shows virtualization-assisted hardware in the processor, which has a register bank and high bandwidth between the CPU and RAM with save/restore instructions to reduce VM context switching overhead. The hardware also has an interrupt request (IRQ) affinity control interface, which specifies the VM services for a given interrupt. Therefore, a) when an IRQ to the active VM occurs, the interrupt handler on the active VM is activated directly. In this case, hypervisor software overhead is completely eliminated. Furthermore, b) the interrupt-suspending function can control VM switching according to both interrupt priority and VM priority. When a lower-priority interrupt to a background VM occurs, the IRQ is suspended until the background VM is activated. In this case, VMs do not interfere with the active VM. As a result, the execution time of VM context switching decreases by 72% (1.56μs to 0.44μs) and meets the requirement of 1μs. The overhead of the interrupt distribution processing time by the hypervisor is 0.0μs in these two cases. Interference among VMs is usually prohibited from the functional safety point of view. Once a fault is detected, the system needs to transition to a safe state immediately. Therefore, an interrupt request notifying a fault is assigned a high priority, even if this breaks the prohibition of interference among VMs. The area overhead of the hardware is less than 1% of the entire chip.

The random hardware failures per hour in ASIL-D must be less than 10^{-8}. To achieve this failure rate, field-BIST is necessary [2]. Field-BIST is executed during the standby-resume period, to avoid disturbing a CPU-processing period (Fig. 2.7.3). However, due to the current fluctuation (di/dt) limitation requirement in automotive applications, the large di/dt of field-BIST in standby-resume period is a problem, arising due to the shifting operations of scan chains. To solve this problem, we developed SR-BIST (standby-resume BIST) with a di/dt suppressor. The concept of SR-BIST is illustrated in Fig. 2.7.3. Every time the CPU resumes from standby mode, SR-BIST is executed prior to CPU processing. An on-chip oscillator clock is used as a test clock because of its rapid start-up. The di/dt at the beginning of each SR-BIST is reduced by gradually increasing the frequency of the scan clock by using an N/M divider. Experimental results show that di/dt is reduced from 6mA/μs to 1mA/μs, meeting the requirements.

Gigabit data throughput and low latency are required to support advanced driving assisted systems (ADAS) with multiple sensors and cameras. SGMII is desirable compared to parallel gigabit Ethernet (GbE), such as GMII/RGMII, due to superior electromagnetic compatibility. We integrated a SGMII interface whose PHY is 1.25Gbaud (=1Gb/s) LVDS into the MCU. The MCU consists of core and 5V transistors, because 5V I/O is mandatory for automotive applications. For reliability reasons tied to voltage surges, we use 5V transistors for the circuits directly connected with external pins. However, 5V transistors have narrower bandwidth than core transistors. To mitigate this problem, we developed the TX driver and the RX receiver shown in Fig. 2.7.4. The capacitor C_R in RX decreases the source impedance of the input pair at high frequency and enhances the AC gain at around 600MHz. The capacitor C_T in the TX conducts high-frequency current and compensates for attenuation caused by the 5V output driver at high frequency. Simulated waveforms of the TX drivers with and without C_T show that this simple technique reduces the rise/fall time by 50% and increases the amplitude by 22%. The measured TX eye pattern is shown in Fig. 2.7.4.

The CPU is a 32b RISC architecture processor with an 8-stage pipeline and is tightly coupled with both embedded flash memory and SRAM. In automotive application software, table-lookup occurs frequently. Due to the low latency of memories and the out-of-order CPU, the MCU running a vehicle control benchmark provides 5.6× the performance of [3]. A neural network (NN) is applied for digital signal processing of sensor data. The CPU employs a floating-point SIMD co-processor with four parallel arithmetic operations and memory access instructions. The SIMD manipulates data effectively by employing move and shuffle on data in the vector registers of 128b (4 words) × 32 entries. The benchmark results for the NN activation functions are shown in Fig. 2.7.5. The CPU with floating point SIMD provides 6.4-to-9.1× the performance of [3].

Measurement results for the MCU are shown in Fig. 2.7.6. The shmoo plots demonstrate 667MHz operation at 25°C. The 28nm MCU consumes 0.56W at 125°C, which is 45% less than the 40nm MCU.

The test chip was fabricated in a 28nm SG-MONOS e-Flash [4], high-κ metal gate CMOS, packaged in a BGA with 516 pins. Figure 2.7.7 shows the test chip micrograph and specifications. The automotive MCU can satisfy the requirements of ECU integration with ASIL-D with low power consumption of 0.52W, real-time virtualization via the virtualization-assisted processor, random hardware failures per hour of 10^{-8} by SR-BIST, and a GbE Interface by employing the SGMII standard.

References:
[1] The Wild West Of Automotive Semiconductor Engineering: http://semiengineering.com/the-wild-west-of-automotive/
[2] C. Takahashi, et al., "A 16nm FinFET Heterogeneous Nona-Core SoC Complying with ISO26262 ASIL-B: Achieving 10^{-7} Random Hardware Failures Per Hour Reliability," *ISSCC*, pp. 80-81, 2016.
[3] Renesas RH850/P1H-C Product Info: https://www.renesas.com/us/en/products/microcontrollers-microprocessors/rh850/rh850p1x/rh850p1hc.html
[4] Y. Taito, et al., "7.3 A 28nm Embedded SG-MONOS Flash Macro for Automotive Achieving 200MHz Read Operation and 2.0MB/S Write Throughput at Ti, of 170°C," *ISSCC*, pp. 132-133, 2015.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

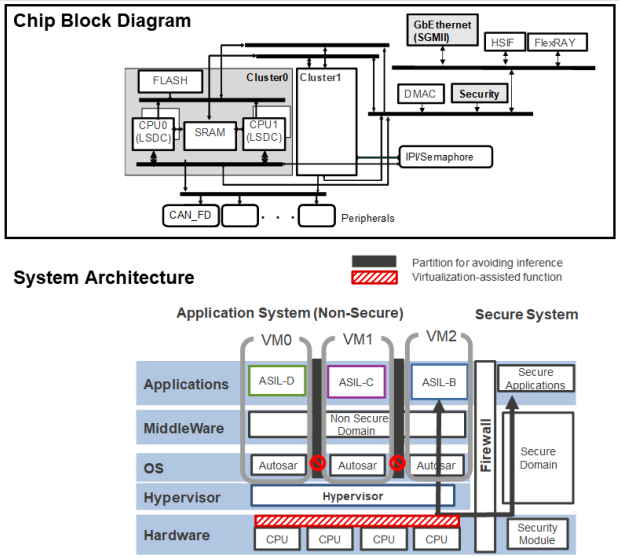
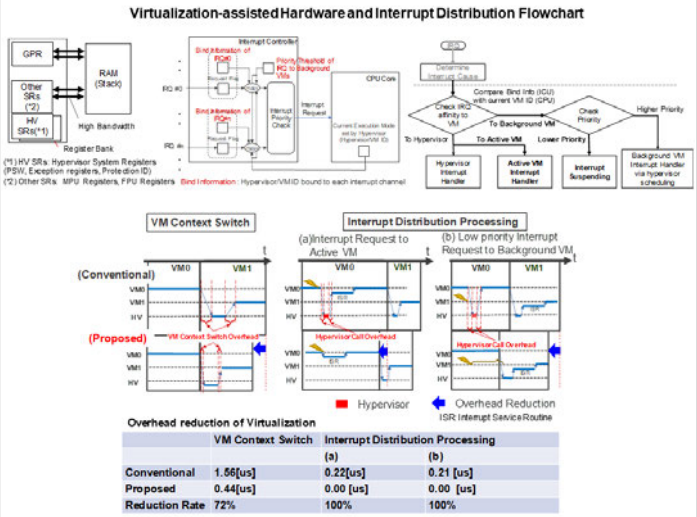


Figure 2.7.1: Overall architecture.



	VM Context Switch	Interrupt Distribution Processing	
		(a)	(b)
Conventional	1.56[us]	0.22[us]	0.21 [us]
Proposed	0.44[us]	0.00 [us]	0.00 [us]
Reduction Rate	72%	100%	100%

Figure 2.7.2: Virtualization-assisted function.

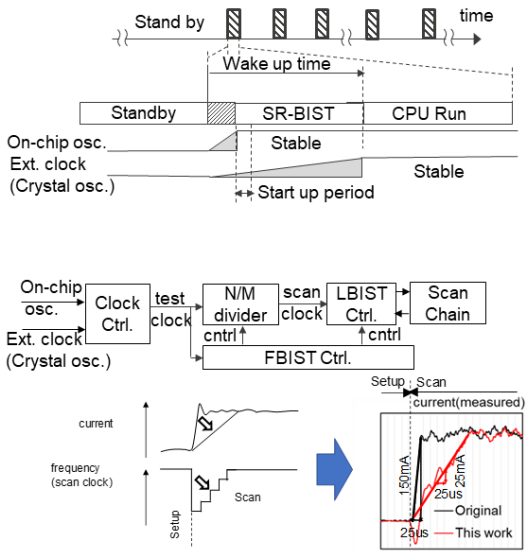


Figure 2.7.3: Standby-resume BIST of MCU for ASIL-D.

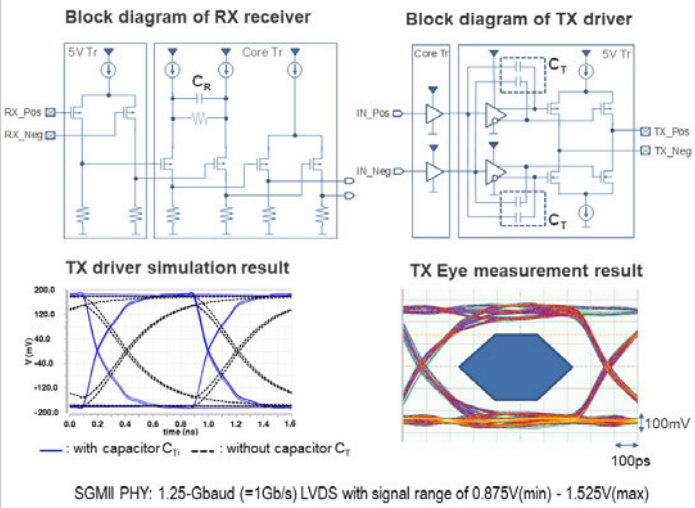
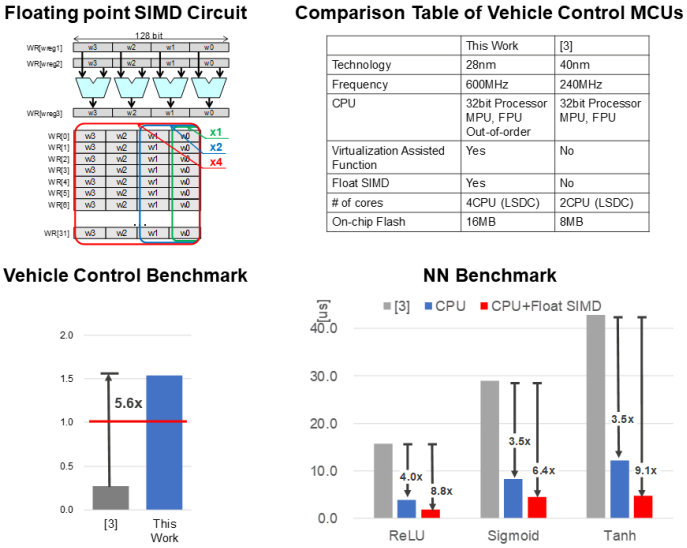


Figure 2.7.4: Serial gigabit ethernet interface (SGMII).



	This Work	[3]
Technology	28nm	40nm
Frequency	600MHz	240MHz
CPU	32bit Processor MPU, FPU Out-of-order	32bit Processor MPU, FPU
Virtualization Assisted Function	Yes	No
Float SIMD	Yes	No
# of cores	4CPU (LSDC)	2CPU (LSDC)
On-chip Flash	16MB	8MB

Figure 2.7.5: CPU with floating point SIMD.

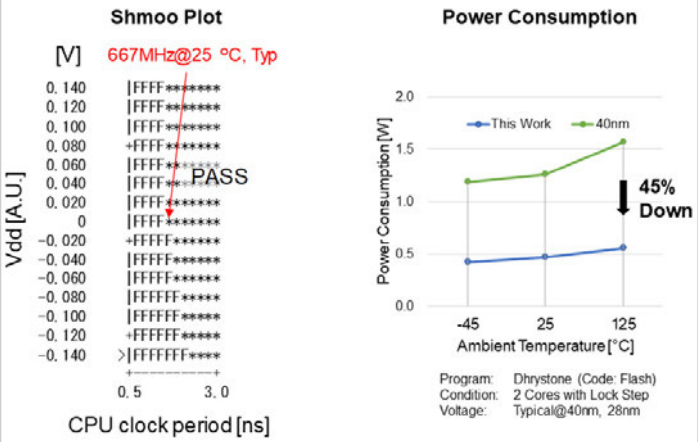
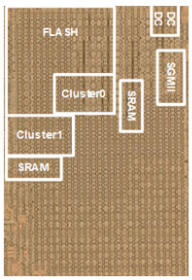


Figure 2.7.6: Measurement results.



Technology	28nm HKMG CMOS process w/ SG-MONOS embedded Flash
Clock Frequency	Processor: 600MHz max.
Chip Size	11.5mm x 8.2mm
Supply Voltage	3.0 .. 5.5 V with embedded DC/DC-ctrl
Operating Temperature	Tj= -40 .. +170°C
Power Consumption	521mW@25 °C, Typical
CPU	32bit Processor(Lock-Step) x 4 MPU, FPU Floating point SIMD Virtualization-assisted function
Memory	Code Flash: 16MB, SRAM: 3MB Local RAM : 256KB
CAN I/F	16 x RSCAN-FD
Ethernet	SGMII: 1bit, 1.25GHz, Serial
System, Security & Safety	[safety] Stress Monitor, Field-BIST, Over-The-Air, Error Collection Module [security] Hardware Security Module [system] Standby Modes, Low Power Sampler, 32ch DMA (+ 128 DTS), Clock Monitor, Temperature Sensor

Figure 2.7.7: Test chip micrograph and features.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 3 Overview: *Nyquist Rate ADCs*

DATA CONVERTERS SUBCOMMITTEE

Session Chair: ***Seng-Pan (Ben) U***
University of Macau, Macau, China

Associate Chair: ***Ahmed Ali***
Analog Devices, Greensboro, NC

Subcommittee Chair: Michael Flynn, University of Michigan at Ann Arbor, Ann Arbor MI

Extending the speed and resolution has been a major trend for Nyquist rate ADCs. This session presents Nyquist rate ADCs ranging in resolution from 11 to 16 bits and sample rates from 2MS/s to 5GS/s. Pipelined ADC architectures are primarily used to achieve these high resolutions, with a variety of hybrid approaches involving SAR and TDC stages used to enhance power efficiency. Residue amplification using open-loop Gm-R, dynamic amplifiers, and closed-loop ring-amplifier-based switched-capacitor circuits are employed for achieving high speed and low power. Calibration is widely used in many of these ADCs to enable high resolution with low power and area. Temperature stability is optimized to maintain calibrated performance over a wide temperature range.

1:30 PM

3.1 A 3.2GS/s 10 ENOB 61mW Ringamp ADC in 16nm with Background Monitoring of Distortion

B. Hershberg, imec, Leuven, Belgium

In Paper 3.1, imec describes a 13b, 3.2GS/s, 4× interleaved ring-amplifier-based pipelined ADC, which introduces a general method for background tracking of signal-to-distortion ratio using a single-comparator stochastic ADC. This 16nm ADC achieves 61.7dB SNDR with a 1.6GHz input while consuming 61mW.

2:00 PM

3.2 A 7.6mW 1GS/s 60dB SNDR Single-Channel SAR-Assisted Pipelined ADC with Temperature-Compensated Dynamic Gm-R-Based Amplifier

W. Jiang, University of Macau, Macau, China

In Paper 3.2, the University of Macau describes a single channel 12b, 1GS/s SAR-assisted pipelined ADC. A Gm-R-based residue amplifier with temperature-compensated bias is used. This 28nm ADC achieves >60dB SNDR at Nyquist while consuming 7.6mW.

2:30 PM

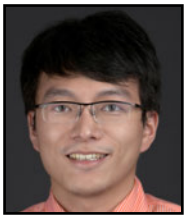
3.3 A 5GS/s 158.6mW 12b Passive-Sampling 8×-Interleaved Hybrid ADC with 9.4 ENOB and 160.5dB FoM_S in 28nm CMOS

A. Ramkaj, KU Leuven, Heverlee, Belgium

In Paper 3.3, KU Leuven describes a 5GS/s 12b passive-sampling 8×-interleaved hybrid ADC with a bandwidth in excess of 6GHz. This 28nm ADC achieves a 58.5dB SNDR at 2.4GHz while consuming 158.6mW.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

3:15 PM



3.4 A 0.01mm² 25μW 2MS/s 74dB-SNDR Continuous-Time Pipelined-SAR ADC with 120fF Input Capacitor

L. Shen, University of Texas, Austin, TX

In Paper 3.4, the University of Texas Austin describes a 2MS/s, 13b continuous-time pipelined-SAR ADC with a kT/C noise-free input stage. This 40nm ADC achieves a peak SNDR of 73.5dB while consuming 25.2uW.

3:45 PM

3.5 A 0.6V 13b 20MS/s Two-Step TDC-Assisted SAR ADC with PVT Tracking and Speed-Enhanced Techniques

M. Zhang, University of Macau, Macau, China

In Paper 3.5, the University of Macau presents a 20MS/s, 13b, two-step TDC-assisted SAR ADC with PVT tracking and speed enhancing techniques. This 65nm ADC achieves a Walden FoM of 1.4fJ/conversion-step using a 0.6V supply.

4:15 PM

3.6 A 6-to-600MS/s Fully Dynamic Ringamp Pipelined ADC with Asynchronous Event-Driven Clocking in 16nm

B. Hershberg, imec, Leuven, Belgium

In Paper 3.6, imec presents a single-channel 600MS/s, 11b, pipelined ADC with an asynchronous, event-driven approach to timing control that leads to fully dynamic power consumption. This 16nm ADC achieves 60dB SNDR and 78dB SFDR while consuming 6mW.

4:45 PM

3.7 A 10mW 16b 15MS/s Two-Step SAR ADC with 95dB DR Using Dual-Deadzone Ring-Amplifier

A. ElShater, Oregon State University, Corvallis, OR

In Paper 3.7, Oregon State University describes a 15MS/s, 16b, two-step SAR ADC using a dual-deadzone ring-amplifier. This 180μm ADC achieves 91dB SNDR while consuming 10mW.

3.1 A 3.2GS/s 10 ENOB 61mW Ringamp ADC in 16nm with Background Monitoring of Distortion

Benjamin Hershberg, Davide Dermit, Barend van Liempd,
Ewout Martens, Nereo Markulic, Jorge Lagos, Jan Craninckx

imec, Leuven, Belgium

Giga-sample ADCs targeting high performance communication applications such as direct-RF sampling all rely on some form of residue amplification to minimize the number of interleaved channels and meet demanding specifications. Despite architectural efforts to reduce the total number of amplifiers in the system, the challenges associated with designing them for high bandwidth and linearity has limited reported power efficiencies [1]. In this work, we show that ring amplification [2] can overcome this longstanding bottleneck. In an architecture using 36 ringamps, the 3.2GS/s ADC consuming 61.3mW has a Nyquist SNDR of 61.7dB, SFDR of 73.3dB, Walden FoM of 19.2fJ/conv-step, and Schreier FoM of 165.9dB. Furthermore, we demonstrate a general technique whereby the signal-to-distortion ratio (SDR) of any amplifier in the system can be independently monitored in the background with an analog hardware overhead of only one comparator.

The top-level system details are provided in Fig. 3.1.1. The ADC input is buffered by an AC-coupled pseudo-differential class-AB push-pull source follower. A clock buffer converts a differential sine into a single-ended square. The ADC core consists of four interleaved pipeline channels, each containing nine scaled 1.5b MDAC stages followed by a 1.5b+3b backend.

The stage 1 MDAC of the channel is shown in Fig. 3.1.2. Here, we introduce a passive-hold mode into the operation of the MDAC. In the illustrated switching scheme, during φ_Q the voltage at node X will be $-\alpha V_{IN}$, where α represents attenuation due to parasitics. The sub-ADC is attached to node X, and uses this held value for quantization during φ_Q. Because the charge sampled onto the MDAC capacitors (C_U) is used for both quantization and residue amplification, all skew and bandwidth related mismatch between the MDAC and sub-ADC is eliminated. This also eliminates the need for a dedicated sub-ADC sampling network, reducing the input buffer load. To minimize its parasitic influence on analog loop performance during φ_A, the sub-ADC comparator uses the source-shifted programmable threshold topology of [3] which gives low decision delay even with near-minimum sized input transistors.

The fully differential ring amplifier is illustrated in Fig. 3.1.3. Because it is only needed during φ_A, the structure is designed to support rapid power cycling. Stability is enforced with a CMOS resistor [4], which operates as a switch during power-down and a tunable bias control during normal operation. The effective resistance can be adjusted with the two charge-sharing DACs of Fig. 3.1.3. These set a trapped-charge bias voltage onto the gates of the CMOS resistor. The common-mode feedback (CMFB) of the ringamp consists of three feedback loops. Two of these loops operate on the trapped-charge present at floating node B1. This approach ensures correct biasing in the presence of kickback due to power cycling. Charge can only enter or exit node B1 through C_{SMALL}, and the first CMFB loop uses this slow injection path to regulate DC biasing. Per-cycle common-mode variations are suppressed by the fast second loop formed by C_{FB}. Transistors M_{CM1} and M_{CM2} form the third loop, and ensure stability by increasing phase margin.

To ensure optimal performance of the ringamps in the presence of PVT variation, a general technique for background measurement of SDR in switched capacitor feedback circuits is introduced in Fig. 3.1.4. It is based on the observation that the summing-node voltage V_X is equal to the input-referred error of V_{OUT}. A portion of this error is due to the finite gain of the amplifier (A_{OL}), and in [5] it is shown how repeated measurements of V_X and V_{OUT} can be used to estimate and correct for it. However, other well-known background calibration techniques can correct not only for finite gain error but also capacitor mismatch, so we are not directly concerned with the V_{OUT}/A_{OL} component of V_X in this work. Rather, we are interested in the remainder of V_X, which contains all other noise and distortion components of the feedback system – that is, SNDR. As described by the equations of Fig. 3.1.4, the estimated average open loop gain $\hat{A}_{OL}$ can be calculated with a linear regression, and after its removal from V_X, the SNDR can be estimated.

In a pipelined ADC, $D(V_{OUT})$ is already available in the pipeline's digital output. However, obtaining an accurate estimate of V_X requires additional hardware and is non-trivial due to the sensitivity of node X and the small amplitude of the signal – on the order of ~1mV. A direct approach using a high-resolution auxiliary ADC [5] is too complex and area-intensive to be practical for inclusion in each pipeline stage we wish to monitor.

An elegant alternative is to use a single-comparator stochastic ADC [6]. The procedure is summarized in Fig. 3.1.4. In an initial preparation step, the comparator is connected to an on-chip reference ladder, its threshold offset is nulled, and it is provided with several scaled reference values, e.g. $k \cdot V_{REF}/1024$, k = -3 to 3. The comparator output for each of these trials is used to determine its noise distribution in absolute terms (volts per sigma). In a second preparation step, the comparator is re-connected to node X and its threshold offset is again nulled while sampling V_X. Now the main measurement begins. As samples of V_X are quantized by the comparator, they are sorted with respect to $D(V_{OUT})$ into "bins". To reduce computational complexity, each bin processes a contiguous range of $D(V_{OUT})$ codes, rather than just one code per bin, without loss of final estimation accuracy. Sorting is simply a matter of demuxing with respect to the MSBs of $D(V_{OUT})$. Each bin computes the mean value of its binary input vector and converts this cumulative-distribution value to a corresponding probability-density "sigma" with the inverse error function. The output is then multiplied by the volts-per-sigma conversion factor acquired earlier and muxed with respect to $D(V_{OUT})$. The resulting output is an estimate of V_X that, due to the processing procedure, has been stripped of noise, but still contains a high accuracy estimate of all non-random errors. Consequently, the output of the estimator block becomes SDR rather than SNDR.

In Fig. 3.1.5 we show in measurement how this estimation technique is used to characterize the performance of the CH0, STG1 ringamp with respect to its digital bias control. For comparison, the THD of CH0, calculated from an FFT of its digital output, is overlaid. Considering that THD and SDR are not identical quantities, and that the THD also includes distortion from other stages in the channel, the correspondence between these two curves is very close. Also notable is the broad range of bias codes that give near-optimal performance, indicating that the ringamp is robust to variation.

The ADC is fabricated in a 16nm CMOS technology and occupies an active area of 0.194mm^2 (360µm × 540µm), excluding decoupling (Fig. 3.1.7). A single, fixed configuration of digital settings and stage gain correction coefficients (determined off-chip) are used in all measurements. Performance is summarized in Fig. 3.1.5 and Fig. 3.1.6. At 3.2GS/s, total power consumption is 61.3mW, with 11.2mW in the input buffer, 2.4mW in the clock buffer, and 11.9mW per channel. The references are 50mV and 800mV, allowing the ringamps to utilize 88% of the available 0.85V supply. In the input frequency sweep of Fig. 3.1.5, the dip in SFDR around 1GHz is dominated by HD3, and found to be packaging related. Compression near full-scale is due to input buffer HD3.

Compared to all ADCs above 1.7GS/s achieving SNDR > 51dB in [1], this work achieves the highest reported SNDR at Nyquist and advances the Walden and Schreier FoM state-of-the-art by 8× and 12dB respectively. This is made possible not only by the direct performance advantages of ring amplifiers, but also the indirect advantage of greater design freedom when amplification is relaxed as an architectural constraint.

References:
[1] B. Murmann, "ADC Performance Survey 1997-2018," [Online]. Available: http://web.stanford.edu/~murmann/adcsurvey.html.
[2] B. Hershberg, et al., "Ring Amplifiers for Switched Capacitor Circuits," *IEEE JSSC*, vol. 47, no. 12, Dec. 2012.
[3] E. Martens, et al., "Wide-tuning range programmable threshold comparator using capacitive source-voltage shifting," *IET Electronics Letters*, Sept. 2018.
[4] J. Lagos, et al., "A single-channel, 600Msps, 12bit, ringamp-based pipelined ADC in 28nm CMOS," *IEEE Symp. VLSI Circuits*, pp. 96-97, June 2017.
[5] A. Ali, et al., "A 16-bit 250-MS/s IF Sampling Pipelined ADC With Background Calibration," *IEEE JSSC*, vol. 45, no. 12, Dec. 2010.
[6] B. Verbruggen, et al., "A 60 dB SNDR 35 MS/s SAR ADC With Comparator-Noise-Based Stochastic Residue Estimation," *IEEE JSSC*, vol. 50, no. 9, Sept. 2015.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

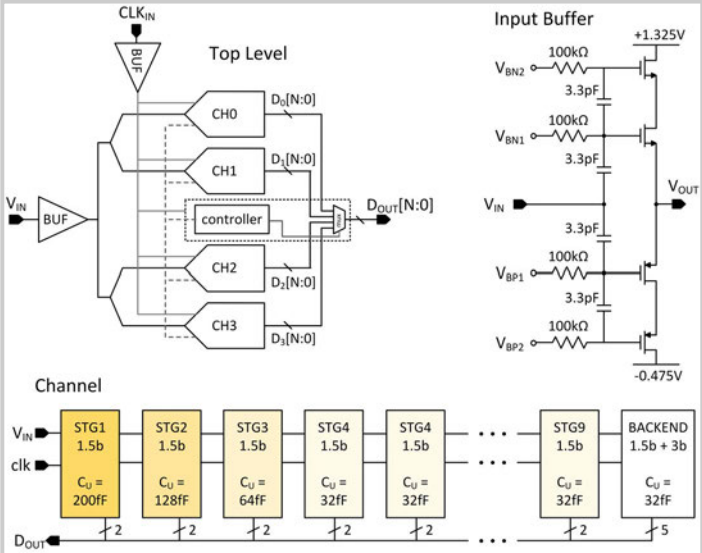


Figure 3.1.1: Top-level architecture.

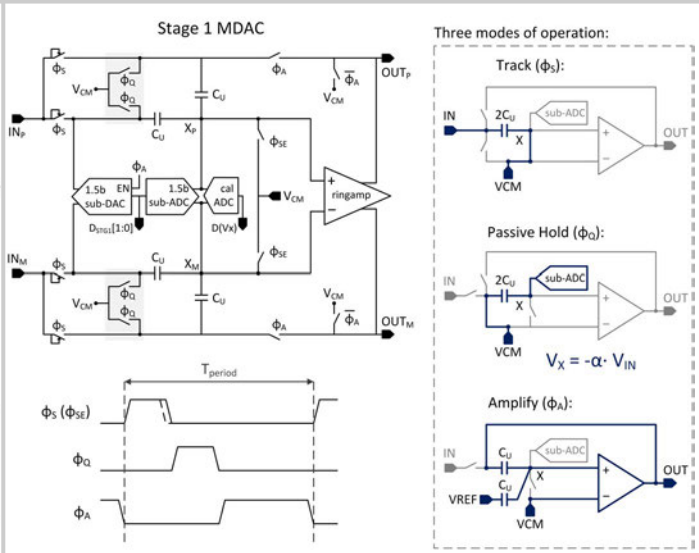


Figure 3.1.2: Merged MDAC and sub-ADC with passive hold mode.

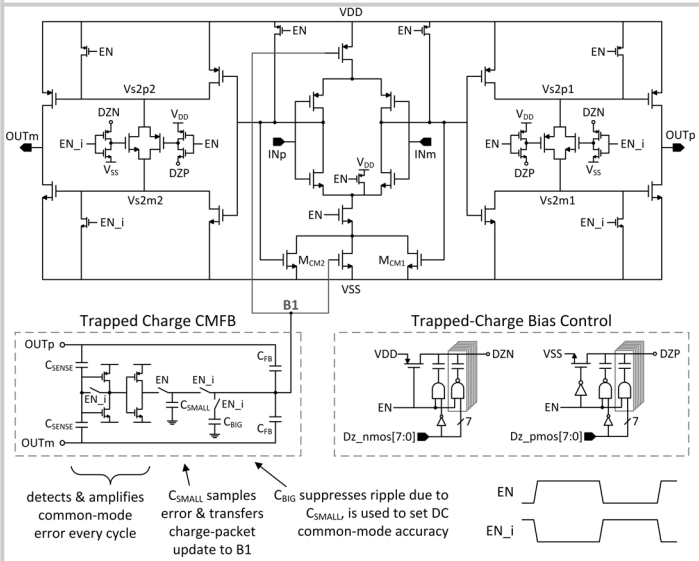


Figure 3.1.3: Fully differential ring amplifier.

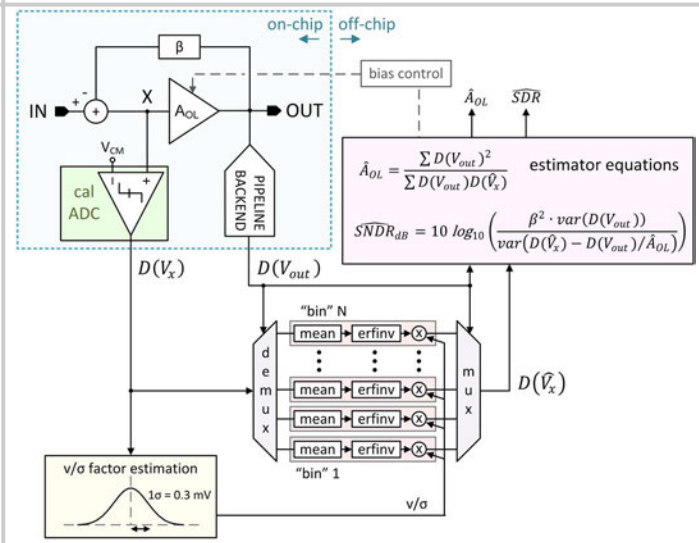


Figure 3.1.4: Method for background estimation of signal-to-distortion ratio (SDR).

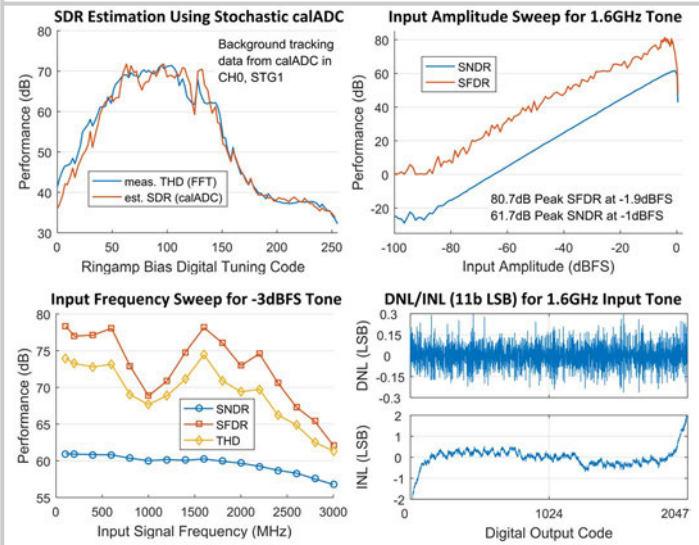
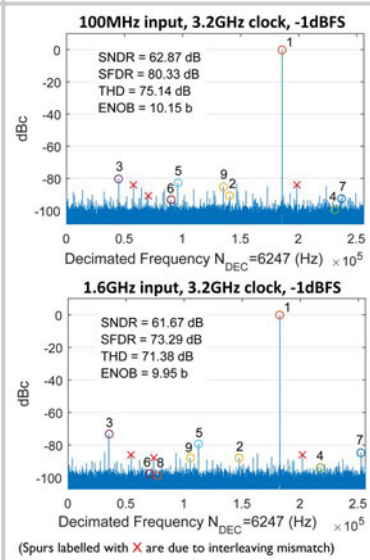


Figure 3.1.5: Demonstration of SDR measurement and key performance metrics.



Performance Summary

Technology	16nm CMOS	
Core Supply	0.85 V	
Buffer Supply	1.8 V	
Sampling rate	3.2 GS/s	
Interleave	4x	
Resolution	13 b	
Input Range	1.5 V pk-pk diff.	
Calibration	static 1st order stage gain	
Performance at -1dBFS:	*100 MHz input:*	*1.6 GHz input:*
ENOB	10.15 b	9.95 b
SNDR	62.9 dB	61.7 dB
SFDR	80.3 dB	73.3 dB
THD	75.1 dB	71.4 dB
Power	61.3 mW	
Walden FoM	19.3 fJ/c-step	
Schreier FoM	165.9 dB	
Active Area	0.194 mm^2	

Figure 3.1.6: Spectrums for LF/HF inputs and performance summary.

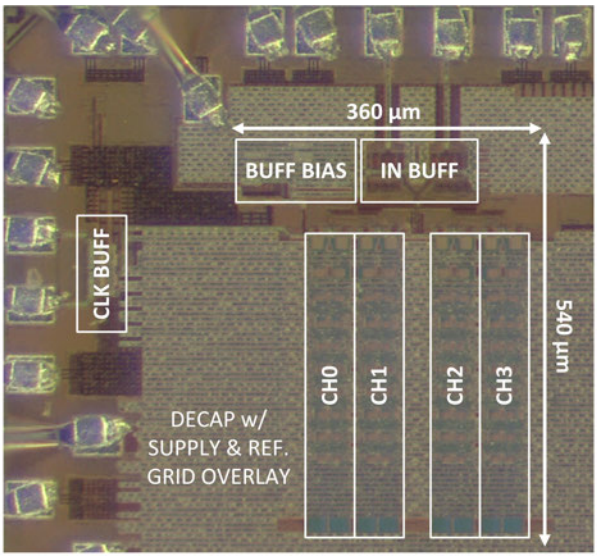


Figure 3.1.7: Chip micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

3.2 A 7.6mW 1GS/s 60dB SNDR Single-Channel SAR-Assisted Pipelined ADC with Temperature-Compensated Dynamic Gm-R-Based Amplifier

Wenning Jiang[1], Yan Zhu[1], Minglei Zhang[1], Chi-Hang Chan[1], Rui P. Martins[1,2]

[1]University of Macau, Macau, China
[2]Instituto Superior Tecnico/University of Lisboa, Lisbon, Portugal

Continuous technology scaling has allowed unceasing growth of the sampling rate of a single channel ADC in past decades. Such development not only helps reduce the number of channels in massively time interleaved ADCs, but also contributes to lower their overall jitter and input capacitance, thus enabling a further push on the ADC performance boundary. Being limited by metastability, the conventional SAR architecture is not suitable when both high resolution and speed are essential. While the pipeline SAR offers a higher speed alternative, it also keeps the low and dynamic power nature of the SAR architecture through adopting dynamic or integrating residue amplifiers (RAs) in recent works [1,2]. With an integrating characteristic, the amplification time can be short but in contrast, the linearity, extra reset time of the load, and PVT sensitivity pose significant design challenges. In this work, rather than adopting the integrating-type amplifier, we present a Gm-R based RA which has a complete-settled amplification characteristic, thus allowing us to compensate the gain variation over temperature easily with a tracking bias technique. Besides this, we use a two-stage amplification to alleviate the RA input parasitic capacitance that enables a small DAC size in all stages. The single channel prototype reaches 1GS/s with 60.02dB SNDR at a Nyquist input consuming 7.6mW from a 1V supply.

Figure 3.2.1 illustrates the overall architecture of our ADC. The first stage is a 4b asynchronous coarse SAR with code-transferred fine DAC while the 2nd and 3rd stages are asynchronous SARs. The 1st and 2nd RAs (RA1 & RA2) utilize different amplification schemes but they both share a complete-settled characteristic and temperature tracking bias circuit. During the sampling phase, both large DAC (L-DAC=540fF) and small DAC (S-DAC=60fF) in the 1st stage sample the input together. We use the S-DAC to quantize the first 4 MSBs in order to speed up the SAR conversion, where the outputs of the small SAR are fed to the L-DAC bit-by-bit for easy logic implementation. After the 4b residue is ready on the top-plate of L-DAC, the RA1 is enabled thus passing the 8× amplified residue to the next stage. The 2nd stage SAR resolves 4 more bits delivering its residue at DAC2 to the 3rd stage through the RA2 with 8× gain. The last stage SAR then determines the remaining 6 bits. One bit overlapping is allocated between each stage leading into a total 12b implementation. To improve the speed further, the last switching of DAC3 is single-ended to halve the DAC size. Both RAs are disabled when they are not in use to save power and their reset happens after the amplification. The offset of the comparator and RA, and inter-stage gain error due to process variations are corrected in the foreground off-chip.

In our design, the conversion time of the 3rd stage confines the overall sampling rate as it needs to resolve two more bits than the previous two stages. On the other hand, since this stage experiences a 64× gain from the input, its noise requirement is greatly relaxed; thereby, only 50fF capacitance is used, limited by the minimum unit capacitance (0.5fF) for matching. While the 2nd stage only experiences a gain of 8, the choice of capacitance in DAC2 considers mainly two aspects: 1) the settling requirement and limited amplification time of RA1; 2) non-linear capacitance and signal attenuation at the top-plate of DAC2. Small capacitance causes a large attenuation and nonlinearity, while enlarging DAC2 will lead to a longer amplification and conversion time with higher power in RA1 and the 2nd stage SAR, respectively, thus eventually tuning into a bottleneck of the design. In order to resolve the abovementioned constraints, we propose a two-stage Gm-R-based amplifier with reset configuration in Fig. 3.2.2. Each stage consists of a Gm cell with a resistive load with operations divided into amplification and reset mode by ϕ_{RA2}. Rather than allowing V_A to float, which causes a large parasitic variation, we reset it to VDD during ϕ_{RA2}=0 thus providing a complete current cut-off to the Gm-cell, thereby keeping its input capacitance constant. When ϕ_{RA2}=1, a current path is enabled from VDD to Gnd in the cell where the voltage difference at the input is converted into the current domain with a linearized input differential pair (M_1-M_4) [3] that can reach a current efficiency as high as 50% due to the complementary input. In the integrating-based amplifier, in order to obtain a high gain in a short time, the consumed instantaneous current must be large. While its gain depends on the integrating time, it is very sensitive under PVT variations. In contrast, we allow the amplifier to be completely settled. Its gain (1 stage) is independent of the enable time and DAC2&3 now do not need a dedicated reset phase, which greatly reduces the sensitivity of the gain variation and allows a simple temperature gain compensation technique (detailed next). Furthermore, we cascode a resistive load directly with the linearized Gm pair that optimizes energy efficiency of the RA for high bandwidth, with its BW determined mainly by R_{out} and C_{DAC}. In order to alleviate the attenuation caused by the input parasitic of RA2, the 8× gain from RA2 is realized by cascading two amplifiers. While the first amplifier provides a small gain with a small input pair, the remaining is compensated in the 2nd. Both amplifiers have a -3dB bandwidth of >5.7GHz that permits the amplification to be accomplished within 220ps in this design. Such a two-stage arrangement is only necessary in RA2 as L-DAC is relatively large for low noise. We select R_{out} to be 200Ω at the output of both RAs to balance the tradeoffs between the gain, bandwidth, and linearity.

Figure 3.2.3 depicts a simple temperature tracking gain compensation scheme applied in the RAs. We intentionally design the Gm-cell in both RAs to have a negative temperature coefficient (TC). While the resistance of R_{out} is relatively insensitive to temperature change, the gain of both RAs also experiences a negative TC characteristic (blue line). Therefore, we utilize a bias circuit with positive TC (red line) to compensate their gain drift under temperature variation. The polarity of the TC in the RAs and bias circuit can be manipulated through sizing of the PMOS and NMOS transistors. V_C is brought off-chip to cover the process variation. Figure 3.2.3 also illustrates the measured result, where the SNDR stays above 56.2dB within the temperature range of 0°C to 80°C without any further calibration from the nominal temperature.

The ADC fabricated in 28nm CMOS exhibits a full swing of 1.2V_{pp} occupying a core area of 0.0091 mm^2. Figure 3.2.4 illustrates the measured performance. The peak DNL and INL are +0.47/-0.39LSB and +1.87/-2.21LSB, respectively. Figure 3.2.4 also provides the FFT spectrum at low frequency and Nyquist input, respectively. The measured SFDR and SNDR at Nyquist input are 74.56dB and 60.02dB, respectively. Figure 3.2.5 plots the measured SFDR and SNDR of the ADC versus various sampling rates and input frequencies. The SNDR remains above 56dB even with the input frequency raised up to 1.2GHz at a 1GS/s sampling rate. Running at 1GS/s with a 1V supply, the power consumption is 7.6mW, resulting in a 9.28fJ/conv.-step Walden FoM and 168.2dB Schreier FoM. Figure 3.2.6 summarizes the ADC performance and compares it with state-of-the-art designs exhibiting similar speed and resolution. This work shows an outstanding energy efficiency and achieves the highest sampling rate among single channel 12b ADCs.

Acknowledgement:
This work was financially supported by the Macao Science & Technology Development Fund (FDCT) with Ref no: 117/2016/A3, Research grant from University Macau (MYRG2018-00113-AMSV) and NSFC Grant 61604180.

References:
[1] B. Vaz, et al., "A 13b 4GS/s digitally assisted dynamic 3-stage asynchronous pipelined-SAR ADC," *ISSCC*, pp. 276-277, Feb. 2017.
[2] H. Huang, et al., "A 12b 330MS/s pipelined-SAR ADC with PVT-stabilized dynamic amplifier achieving <1dB SNDR variation," *ISSCC*, pp. 472-473, Feb. 2017.
[3] R. G. Carvajal, et al., "The flipped voltage follower: a useful cell for low-voltage low-power circuit design," *IEEE TCAS-I*, vol. 52, no. 7, pp. 1276-1291, July 2005.
[4] K. Moon, et al., "A 9.1 ENOB 21.7fJ/conversion-step 10b 500MS/s single-channel pipelined SAR ADC with a current-mode fine ADC in 28nm CMOS," *IEEE Symp. VLSI Circuits*, pp. C94-C95, June 2017.
[5] D. Luu, et al., "A 12b 61dB SNDR 300MS/s SAR ADC with inverter-based preamplifier and common-mode-regulation DAC in 14nm CMOS FinFET," *IEEE Symp. VLSI Circuits*, pp. C276-C277, June 2017.
[6] L. Kull, et al., "28.5 A 10b 1.5GS/s pipelined-SAR ADC with background second-stage common-mode regulation and offset calibration in 14nm CMOS FinFET," *ISSCC*, pp. 474-475, Feb. 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

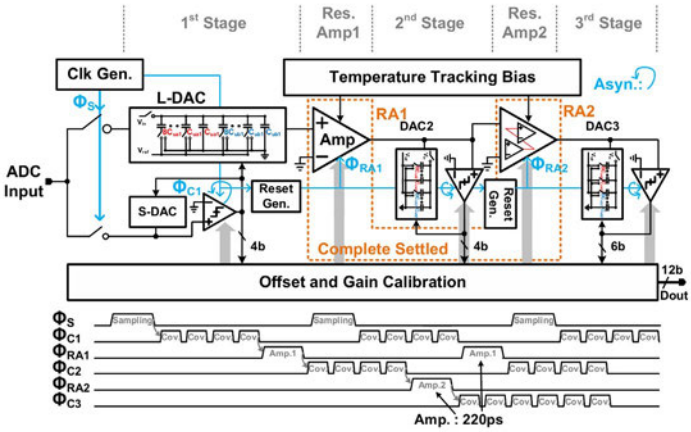


Figure 3.2.1: Block diagram of the SAR-assisted pipelined ADC.

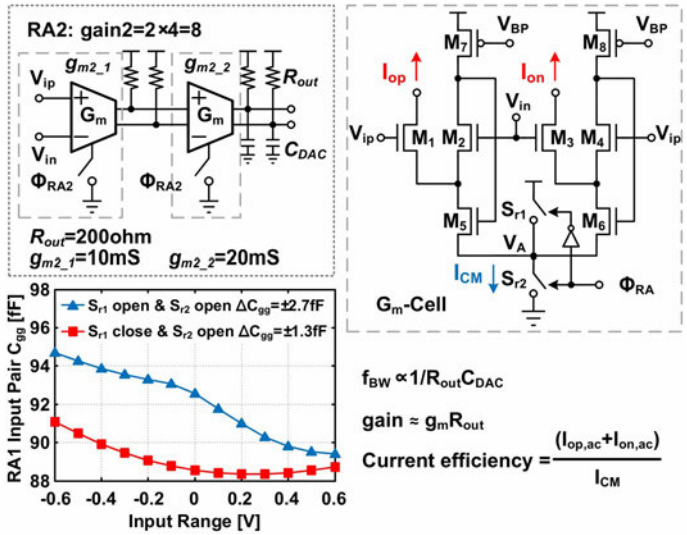


Figure 3.2.2: Two stage Gm-R based amplifier and input parasitic characteristic with and without V_A reset to VDD.

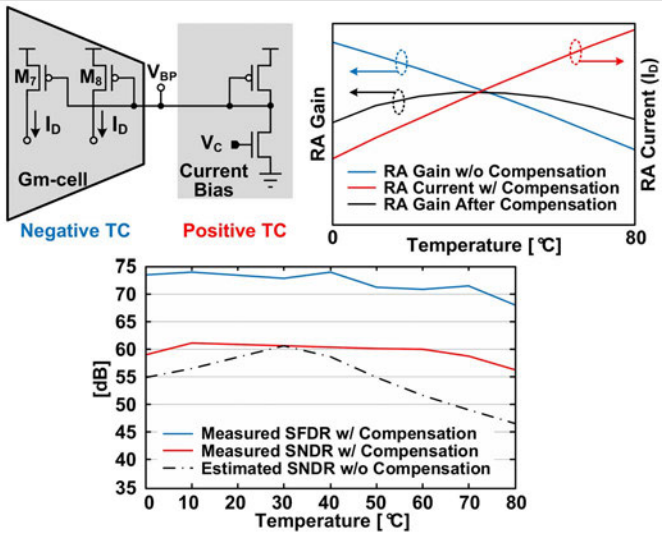


Figure 3.2.3: Temperature compensation concept in the RA with measurement results under temperature variation.

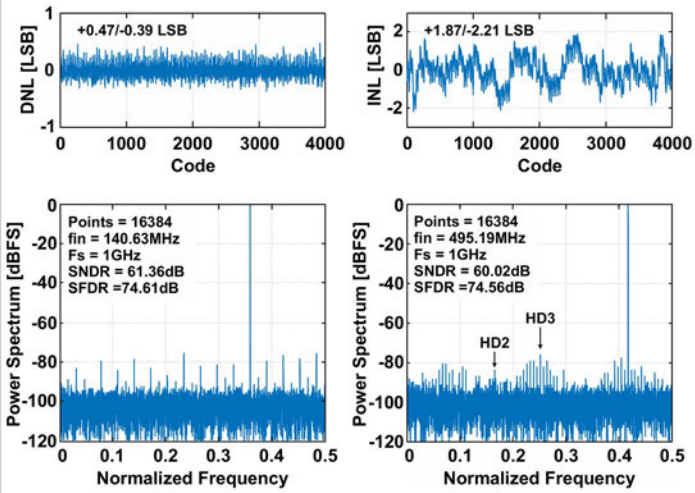


Figure 3.2.4: Measured DNL / INL and FFT spectrum (ADC out output decimated by 225×).

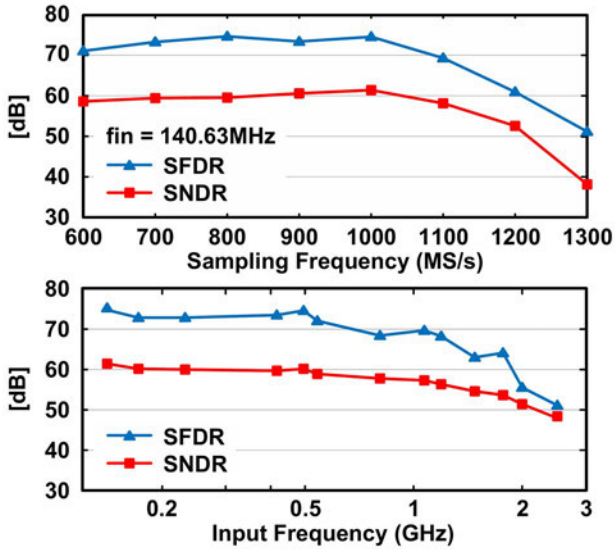


Figure 3.2.5: Measured sampling frequency and input frequency sweep.

	This work	[4]VLSI 2017 K.-J. Moon	[5]VLSI 2017 D. Luu	[6]ISSCC 2017 L. Kull	[2]ISSCC 2017 H. Huang
Architecture	Pipelined SAR	Pipelined SAR	SAR	Pipelined SAR	Pipelined SAR
Technology	28nm	28nm	14nm	14nm	65nm
Resolution [bits]	12	10	12	10	12
Sample Rate [MS/s]	1000	500	300	1500	330
Supply Voltage [V]	1	1	0.85	0.95	1.3
SFDR @Nyq. [dB]	74.56	69.2	78.5	58.39	75.8
SNDR @Nyq. [dB]	60.02	56.6	60.5	50.1	63.5
Power [mW]	7.6	6	3.3	6.92*	6.2
FoM_{Walden} @Nyq. [fJ/conv-step]	9.28	21.7	12.7	17.7*	15.4
$FoM_{Schreier}$ @Nyq. [dB]	168.2	162.8	167.1	160.5*	167.8
Area [mm^2]	0.0091	0.015	0.006	0.0016	0.08

* including the reference buffer

Figure 3.2.6: Performance summary and comparison.

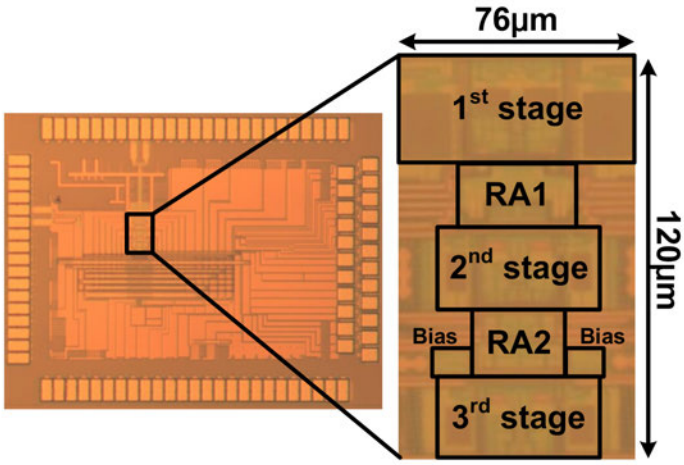


Figure 3.2.7: Die microphotograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

3.3 A 5GS/s 158.6mW 12b Passive-Sampling 8×-Interleaved Hybrid ADC with 9.4 ENOB and 160.5dB FoM_S in 28nm CMOS

Athanasios Ramkaj[1], Juan Carlos Pena Ramos[1], Yifan Lyu[1], Maarten Strackx[2], Marcel J. M. Pelgrom[1], Michiel Steyaert[1], Marian Verhelst[1], Filip Tavernier[1]

[1]KU Leuven, Heverlee, Belgium
[2]Nokia Bell Labs, Antwerpen, Belgium

Emerging 5G communication systems require ADCs to directly digitize wide bandwidth (BW) signals with high spectral purity at low power consumption. Current state-of-the-art solutions include mainly time-interleaved (TI) pipelined [1-4] or pipelined-SAR [5] architectures, enhanced by digital calibration. To ensure a sufficiently high input BW, all these designs employ a static front-end buffer. This buffer often dissipates more power than the ADC itself, significantly deteriorates the linearity and noise performance, and severely limits the available swing, unless over-voltage or multiple supplies are used [1-5].

This work aims to boost the input BW and dynamic performance, while minimizing the power of GHz-range high-resolution ADCs, with a 5GS/s passive-sampling, fully dynamic, TI 3-stage pipelined-SAR hybrid. Removing the front-end buffer offers at least 2× power reduction compared to [1-5], while an input BW larger than 6GHz and a Nyquist performance of 9.4ENOB and 160.5dB FoM_S are enabled by the combination of: (1) very low resistance/capacitance passive input network and proper on-chip termination; (2) on-chip clock division and distribution with negligible jitter; (3) optimized hybrid sub-ADC design to deliver maximum speed × resolution/power; and (4) on-chip co-designed sub-ADC/TI analog-digital calibration to enhance performance.

Figure 3.3.1 illustrates the ADC architecture. Eight 625MS/s channels are directly interleaved for moderate complexity and power-efficient sub-ADC operation, totaling a sample rate of 5GS/s. The final reconstructed outputs are stored in a 65k on-chip memory for performance evaluation. Sub-ADC impairments, as well as offset/gain errors, are digitally corrected with 23b to 27b coefficients. Timing skew between sub-ADCs is canceled in the analog domain by tuning their clock edges.

To achieve the highest possible input BW and sampling linearity, both resistance and capacitance in the critical path must be minimized. Utilizing a near full-scale input allows downsizing of the sampling capacitance to the noise limit (C_S=256fF), while bootstrapping the input switches based on [6] minimizes their on-resistance (~11Ω). Regarding the interconnect, a symmetrical intertwisted input/clock Y-tree is introduced, offering several benefits: (1) the BW mismatch and timing skew among the sub-ADCs are substantially reduced; (2) the input is shielded by the twisted clock such that their mutual differential coupling is cancelled, eliminating the need for ground isolation, and significantly enhancing the BW. Further, the Y-tree resistance is absorbed into the termination. The measured ADC transfer characteristic (including losses from board and bonding wires) with this passive front-end demonstrates a wide BW in excess of 6GHz (Fig. 3.3.1).

A clock jitter of less than 50fs is key to achieving more than 10b accuracy at high frequencies. Figure 3.3.2 shows the clock chain and its timing scheme. An external 5GHz differential signal (<40fs jitter) is terminated on-chip and turned into a master clock (MC), able to drive the Y-tree and sub-ADCs with 5ps clock edges. After division by eight, the output is aligned with MC and shifts along the sub-ADCs (SH7-SH0). Each local clock unit contains an align/shift flip-flop, logic gates to create the 12.5% duty-cycle pulses and a final align latch with a driver, offering the shortest path (one transistor and one inverter) from MC to the sampling clock (SAM). The total extracted jitter of 11fs, dominated by the CML circuit, is negligible compared to the jitter of the external source. To minimize timing skew between sub-ADCs, a segmented 6b-binary-3b-unary digital-to-delay tuner (DDT) is introduced in front of the driver. The segmentation combines fine tuning step with moderate resolution in a single stage, adding less jitter compared to multi-stage approaches [1]. The 9fs tuning step of this DDT aims for the desired timing spur level, and the tuning range of ±1.5ps covers the maximum anticipated skew.

The sub-ADC architecture is detailed in Fig. 3.3.3. Three dynamic SAR stages with two residue amplifiers (RA) in between are pipelined for maximum speed. The allocated 4-4-6 bits/stage with overrange total a 12b quantization at the targeted sample rate with minimum power and low design complexity. Bottom-plate sampling is preferred for SAR_1 for increased linearity, while top-plate sampling is chosen for $SAR_{2,3}$ to maximize speed. A V_{CM}-based switching is employed in all stages for its superior speed and power [6], while the supply rails are used as DAC references for minimum switch on-resistance. A single dynamic integrator RA is designed with a gain of 4× for relaxed linearity, and only a minor input-referred noise penalty. The gain is set by its output capacitance and the integration time is controlled by a detection circuit (Fig. 3.3.3) to set the required common-mode (0.5V) for the following stages. To accommodate the reduced output swing of the RA, the DAC capacitors of $SAR_{2,3}$ are split for half-scale and quarter-scale range by tying a fixed part (C_{H2}, C_{H3}) to V_{CM}. Despite the swing reduction, the LSB of each stage remains large enough to allow a high-speed and low-power comparator design. Asynchronous internal clocking is utilized for better timing allocation. Its timing sequence is shown in Fig. 3.3.3. As a result, a sub-ADC sample rate of 625MS/s is obtained with a single comparator/stage, avoiding the associated overhead of multiple comparators [1-5]. The digital calibration to correct DAC weights, RA gain and sub-ADC offset/gain is applied in the foreground, while the inter-stage overrange absorbs comparator and RA offset.

The prototype ADC, fabricated in 28nm bulk CMOS, occupies an active area of 1130μm×1600μm (Fig. 3.3.7). The bare die is placed in a cavity to minimize the length of the bonding wires. Each sub-ADC is laid out symmetrically and measures an area of 75μm×200μm. The blocks are carefully arranged to minimize speed-critical wiring.

Figure 3.3.4 illustrates the measured output spectra at 5GS/s for 600MHz (top) and 2.4GHz (bottom) input frequencies. The SFDR at 600MHz is 69.5dB, dominated by HD3, while at 2.4GHz it is 65.4dB, dominated by HD2, mainly due to the phase imbalance of the off-chip single-ended-to-differential balun. The remaining spurs after calibration are below -75dBFS. The SNDR is 61.3dB at 600MHz and remains 58.5dB at 2.4GHz, limited mainly by thermal noise.

Figure 3.3.5 plots the measured SFDR/SNDR vs. input frequency at 5GS/s (top-left) and vs. sample rate at 2.4GHz input frequency (top-right) for 5 different samples. The average peak SFDR is 73dB and remains above 60dB up to 4.9GHz input frequency (2nd Nyquist zone). The average SNDR at low frequencies is 62.4dB and drops to 53.2dB at 4.9GHz. SFDR/SNDR at 2.4GHz input frequency stay relatively flat up to 5GS/s after which they start dropping, mainly due to the RAs. SNDR vs. amplitude is shown (bottom-left) for 3 input frequencies. The degradation at large amplitudes is attributed to the input network harmonic distortion. Finally, the power consumption partitioning vs. sample rate at 1V supply is plotted (bottom-right). The total power consumption excludes only the output buffers.

Figure 3.3.6 compares the performance with recently published relevant ADCs. The proposed passive-sampling solution with negligible on-chip clock jitter and optimized hybrid sub-ADCs significantly advances the state-of-the-art in terms of power consumption (between 2× and 18×), both FoM_S/FoM_W and SNDR, with comparable or better BW and SFDR. It achieves at least 6.3dB better FoM_S among ADCs with sample rates above 3GS/s, as well as 2.5× higher sample rate than the closest competitor with the same (±0.5dB) FoM_S.

References:
[1] M. Straayer, et al., "A 4GS/s Time-Interleaved RF ADC in 65nm CMOS with 4GHz Input Bandwidth," *ISSCC*, pp. 464-465, Feb. 2016.
[2] J. Wu, et al., "A 4GS/s 13b Pipelined ADC with Capacitor and Amplifier Sharing in 16nm CMOS," *ISSCC*, pp. 466-467, Feb. 2016.
[3] A.M.A. Ali, et al., "A 14-bit 2.5GS/s and 5GS/s RF Sampling ADC with Background Calibration and Dither," *IEEE Symp. VLSI Circuits*, pp. 206-207, June 2016.
[4] S. Devarajan, et al., "A 12b 10GS/s Interleaved Pipeline ADC in 28nm CMOS Technology," *ISSCC*, pp. 288-289, Feb. 2017.
[5] B. Vaz, et al., "A 13b 5GS/s ADC with time-interleaved chopping calibration in 16nm FinFET," *IEEE Symp. VLSI Circuits*, pp. 99-100, June 2018.
[6] A.T. Ramkaj, et al., "A 1.25-GS/s 7-b SAR ADC With 36.4-dB SNDR at 5 GHz Using Switch-Bootstrapping, USPC DAC and Triple-Tail Comparator in 28-nm CMOS," *IEEE JSSC*, vol. 53, no. 7, pp. 1889-1901, July 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

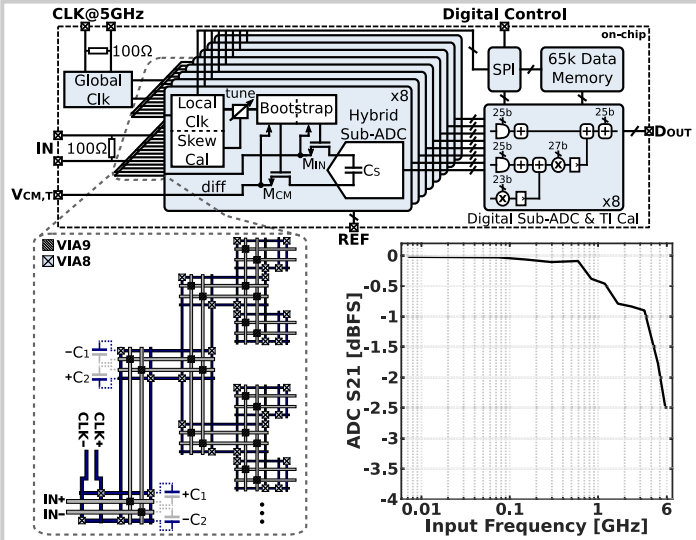


Figure 3.3.1: ADC block diagram with zoomed-in input/clock Y-tree (half shown) and measured transfer characteristic (BW>6GHz).

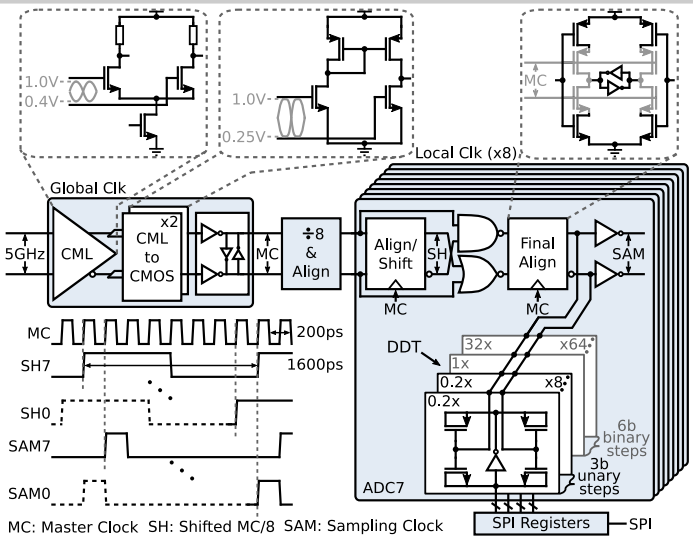


Figure 3.3.2: Total clock chain with timing sequence and digital-to-delay tuner (DDT) for timing skew calibration.

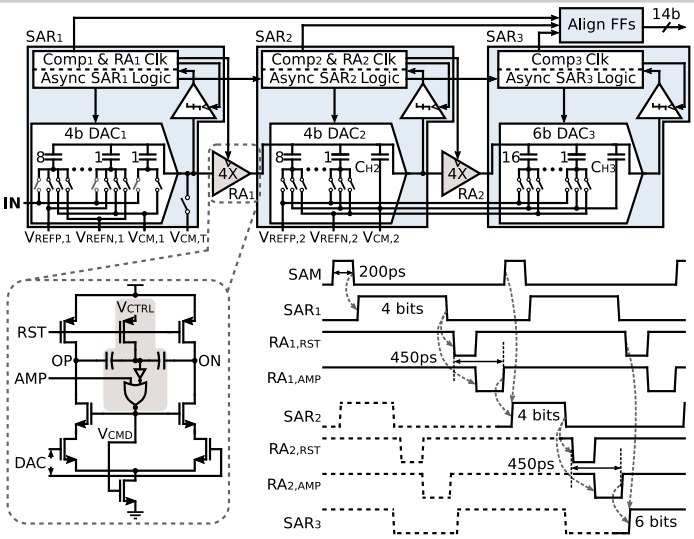


Figure 3.3.3: Sub-ADC schematic (single-ended shown) with its critical blocks and its asynchronous timing sequence.

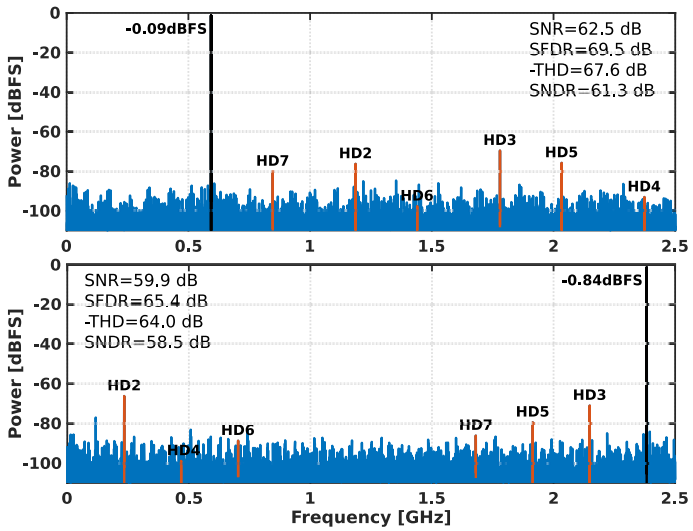


Figure 3.3.4: Measured FFT output spectra at 5GS/s with 600MHz (top) and 2.4GHz (bottom) input frequencies.

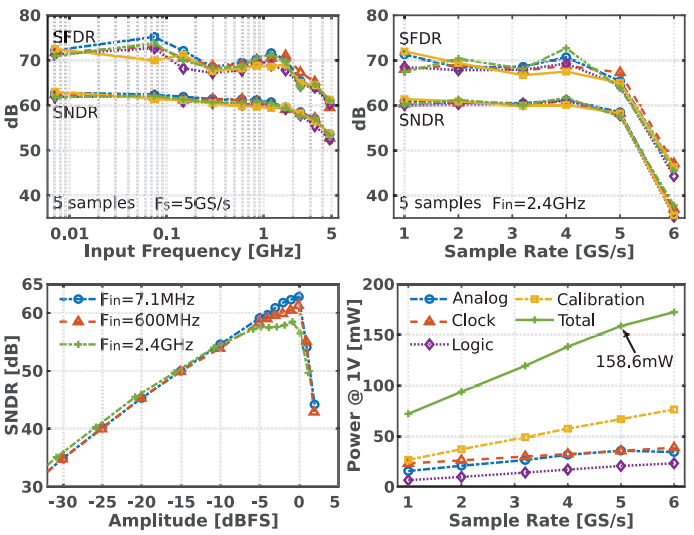
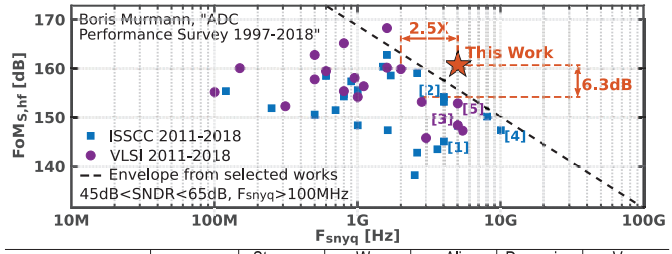


Figure 3.3.5: Measured performance vs. input frequency (top-left), sample rate (top-right), amplitude (bottom-left), and power vs. sample rate (bottom-right).



	This Work	Straayer ISSCC'16[1]	Wu ISSCC'16[2]	Ali VLSI'16[3]	Devarajan ISSCC'17[4]	Vaz VLSI'18[5]
Technology	**28nm**	65nm	16nm	28nm	28nm	16nm
Architecture	**TI-Pipe-SAR**	TI-Pipeline	TI-Pipeline	TI-Pipeline	TI-Pipeline	TI-Pipe-SAR
Sample Rate [GS/s]	**5.0**	4.0	4.0	5.0	10.0	5.0
Input Fin≤Nyq. [GHz]	**2.4**	1.8	1.9	2.0	4.0	2.4
Input Bandwidth [GHz]	**>6.0**	4.0	N.A.	5.0	7.4	N.A.
Resolution [bits]	**12**	N.A.	13	14	12	13
ENOB @ Fin [bits]	**9.4**	8.9	9.0	9.3	8.8	9.2
SNDR @ Fin [dB]	**58.5**	55.5	56.0	58.0	55.0	57.0
SFDR @ Fin [dB]	**65.4**	64.0	68.0	70.0	64.0	61.9
Power [mW]	**158.6**	2200.0	300.0	2300.0	2900.0	641.0
FoMw @Nyq. [fJ/conv]	**46.1**	1130.0	145.5	708.7	631.2	221.6
FoMs @Nyq. [dB]	**160.5**	145.1	154.2	148.4	147.4	152.9

Figure 3.3.6: Performance summary and comparison with the state-of-the-art.

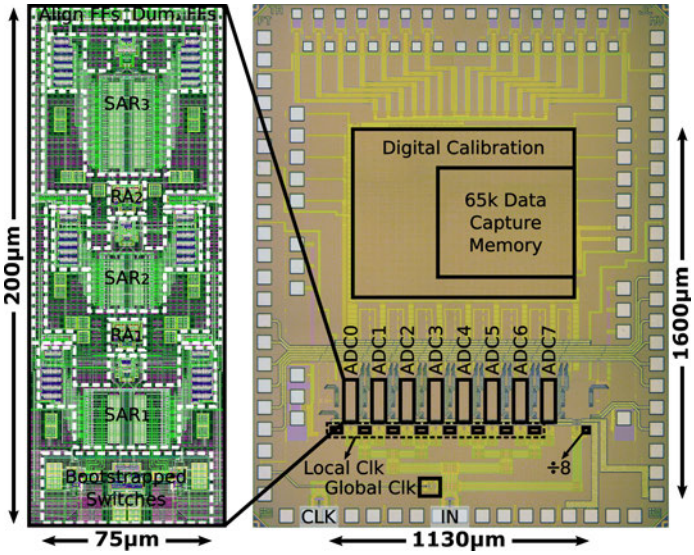


Figure 3.3.7: Die micrograph with a sub-ADC layout view.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

3.4 A 0.01mm² 25μW 2MS/s 74dB-SNDR Continuous-Time Pipelined-SAR ADC with 120fF Input Capacitor

Linxiao Shen[1], Yi Shen[1,2], Xiyuan Tang[1], Chen-Kai Hsu[1], Wei Shi[1], Shaolan Li[1], Wenda Zhao[1], Abhishek Mukherjee[1], Nan Sun[1]

[1]University of Texas, Austin, TX
[2]Xidian University, Xi'an, China

The kT/C noise poses a fundamental SNR limit for discrete-time (DT) ADCs with front-end sampling operations. To achieve a high resolution, the sampling capacitor needs to be sufficiently large (e.g., 3pF for 80dB SNR assuming $2V_{pp}$ swing) and it has to be quadrupled for every 1b increase in the resolution. A large sampling capacitor poses critical challenges for both the ADC input buffer and the reference buffer (Fig. 3.4.1). To meet the stringent linearity requirement of high-resolution ADCs, these buffers consume a significant amount of power, which can be comparable or even higher than the ADC itself. In addition, a large capacitor makes it difficult to design the ADC front-end S/H circuit, leading to considerable power/area costs and increased design complexity.

This paper presents a 13b pipelined-SAR ADC with a continuous-time (CT) input stage that is kT/C noise free (Fig. 3.4.1). Hence, the 1st-stage capacitor size is not bounded by the kT/C limit and can be significantly reduced. Unlike a classic pipelined-SAR ADC, the sampling operation is moved to the 2nd-stage. Hence, the kT/C noise from the 2nd-stage is suppressed by the 1st-stage gain, making possible the use of a small sampling capacitor in the 2nd-stage too. In the prototype ADC, the 1st-stage input capacitor is only 120fF (60fF single-ended), which is 25× smaller than 3pF (set by kT/C noise limit) in a classic DT pipelined-SAR ADC of the same resolution. This substantially smaller capacitor size reduces the chip area. Moreover, it greatly relaxes the performance requirement of the ADC input buffer and the reference buffer, leading to significant power saving on the system level.

In a pipelined ADC, the 1st-stage sub-ADC is fired a half clock cycle before the 1st-stage output is sampled, to leave sufficient time for the MDAC to settle. While it works well for a DT pipelined ADC with a held V_{in}, this timing mismatch poses a key challenge for a CT pipelined ADC. As its V_{in} moves continuously, the sub-ADC output is no longer an accurate estimate for V_{in} at the time of the 2nd-stage sampling, leading to a larger residue V_{res} and potential overflow. This timing misalignment problem can be addressed by intentionally delaying the V_{in} path using an LC lattice filter [1], but it is unsuitable for low-to-medium sampling rate applications, as the required long delay would result in large LC values and excessive chip area. Another solution is to add a prediction filter before the sub-ADC [2]. However, the OTA-based prediction filter consumes large power and its high-pass response makes it difficult to design the anti-aliasing filter. This work addresses the timing misalignment problem in a different way. Instead of using a flash sub-ADC whose comparators are triggered all at once, this work uses an asynchronous SAR that performs a continuous approximation of V_{in} (Fig. 3.4.2). Redundancy is embedded in the SAR DAC, so that it can track V_{in} even if a prior comparator decision is no longer correct as V_{in} moves. In addition, to limit the amount of V_{in} variation, dynamic logic is used to shorten the 10b SAR conversion time to be within 5ns. More importantly, the CT SAR performs the sub-ADC and sub-DAC operation simultaneously. Once the SAR finishes its conversion, V_{res} is readily available. Thus, it allows the dynamic amplifier to be triggered immediately, leading to a very short timing mismatch. Hence, no delay or prediction filter is needed in this work, leading to reduced chip area and power.

Figure 3.4.2 shows the single-ended version of the 1st-stage 10b CT SAR for simplicity. Unlike the conventional DT SAR, the input capacitor and SAR DAC are separated to allow simultaneous V_{in} tracking and conversion. A bridge capacitor C_B is used to minimize the total capacitor size while keeping the unit capacitor size of 1fF. Different from prior CT pipelined works that use input resistive coupling and an IDAC [1], [2], this work uses input capacitive coupling and a CDAC. It eliminates the noise from the input resistor and the IDAC. It also removes the IDAC static power. However, the tradeoff is that capacitive coupling blocks DC signals. A pseudo-resistor R_B has to be added to provide the DC bias for V_{res}. C_{in} and R_B form a high-pass filter that blocks low-frequency input. Nevertheless, the high-pass corner frequency can be made sufficiently low by increasing R_B. In the prototype ADC, the input BW is from 50Hz to 1MHz. The proposed capacitively coupled CT pipelined ADC is suitable for a wide range of applications where information does not reside at DC, such as audio, biological, and communication signals by setting the proper pass band frequencies.

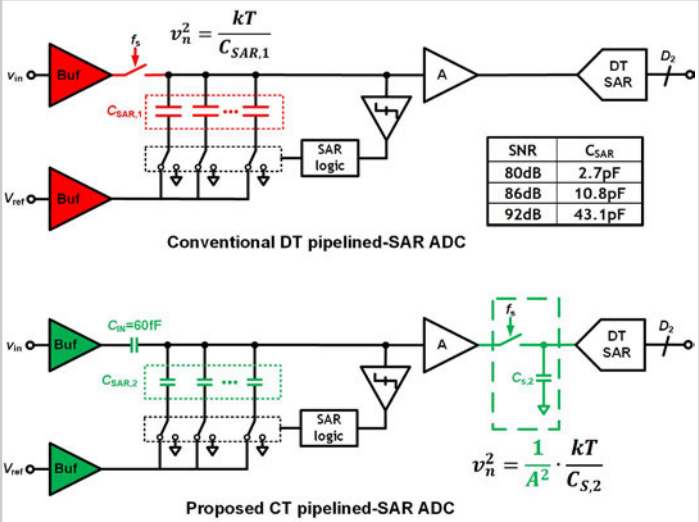
Figure 3.4.3 shows the overall schematic and timing diagram of the proposed ADC. At the rising edge of CLK_{sys}, the 1st-stage CT SAR is triggered. After its operation finishes, the dynamic amplifier (DA) starts to amplify V_{res}. The DA is self-timed and its output is sampled after a fixed amount of amplification time. Then, the DT SAR is triggered to complete the 2nd-stage conversion. With the 1b 1st-stage CT-SAR, V_{res} is sufficiently small, allowing the use of a DA to replace the static closed-loop amplifier to reduce power consumption. For a conventional DA, its output common-mode voltage is usually not well-controlled and sensitive to PVT variations, which limits the available output signal swing. To address this issue, an inverter-based floating DA is proposed (Fig. 3.4.3). During the reset phase, C_B is recharged to V_{dd} and ground. During the amplification phase, C_B works as a battery to provide the current needed for the DA as in [3]. Because C_B is a floating capacitor, it ensures that only differential-mode current can flow out of the DA; the output common-mode current is zero. As a result, the DA output common-mode voltage cannot move and is held at half V_{dd} in this design. To further boost the gain of the DA, a cross-coupled inverter is inserted inside the inverter [4]. The capacitor C_N is the integration cap and sets the DA input-referred noise. C_N is chosen to be 30fF to meet the 13b noise requirement.

The prototype ADC is fabricated in 40nm CMOS (die photo in Fig. 3.4.7). It consumes 25.2uW, where 7.1uW, 12.2uW, and 5.9uW are consumed by the 1st-stage CT SAR, the DA, and the 2nd-stage DT SAR, respectively. The 1st-stage CT SAR and the DA operates under 1.1V to enhance the speed, while the 2nd-stage DT SAR operates under 0.8V to save power. Figure 3.4.4 shows the measured spectra with 100kHz and 950kHz inputs. Foreground calibration is performed for the 1st-stage capacitor mismatch and the DA gain. Figure 3.4.5 shows the measured SNDR and SFDR across different input frequencies and amplitudes. It achieves a peak SNDR of 73.5dB, a peak SFDR of 87.8dB, and a dynamic range of 73.6dB. Figure 3.4.6 provides the performance summary and comparison table. Its input capacitor is orders of magnitude smaller than other works with similar SNDR, which is made possible by the kT/C noise-free CT front-end. Its chip area of 0.01mm² is also significantly smaller than others, due to the reduced capacitor size. The Walden and Schreier FoMs with Nyquist frequency input are 3.9fJ/conv-step and 177.8dB, respectively, and are in line with the state-of-the-art. Due to the substantially reduced input capacitor size, we also envision that significant power savings can be achieved for the ADC input and reference buffers.

Acknowledgement:
The authors would like to thank TSMC University Shuttle Program for chip fabrication.

References:
[1] H. Shibata, et al., "A 9-GS/s 1.125-GHz BW Oversampling Continuous-Time Pipeline ADC Achieving –164-dBFS/Hz NSD," *IEEE JSSC*, vol. 52, no. 12, pp. 3219–3234, Dec. 2017.
[2] D. Gubbins, et al., "Continuous-Time Input Pipeline ADCs," *IEEE JSSC*, vol. 45, no. 8, pp. 1456–1468, Aug. 2010.
[3] M. S. Akter, et al., "A Capacitively Degenerated 100-dB Linear 20–150 MS/s Dynamic Amplifier," *IEEE JSSC*, vol. 53, no. 4, pp. 1115–1126, Apr. 2018.
[4] S. Li, et al., "A 13-ENOB Second-Order Noise-Shaping SAR ADC Realizing Optimized NTF Zeros Using the Error-Feedback Structure," *ISSCC*, pp. 234–236, Feb. 2018.
[5] P. Harpe, et al., "An Oversampled 12/14b SAR ADC with Noise Reduction and Linearity Enhancements Achieving up to 79.1dB SNDR," *ISSCC*, pp. 194–195, Feb. 2014.
[6] M. Shim, et al., "Edge-Pursuit Comparator: An Energy-Scalable Oscillator Collapse-Based Comparator With Application in a 74.1 dB SNDR and 20 kS/s 15 b SAR ADC," *IEEE JSSC*, vol. 52, no. 4, pp. 1077–1090, Apr. 2017.
[7] Y. Lim and M. P. Flynn, "A 1mW 71.5dB SNDR 50MS/S 13b Fully Differential Ring-amplifier-based SAR-Assisted Pipeline ADC," *ISSCC*, pp. 458–459, Feb. 2015.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE



SNR	C_{SAR}
80dB	2.7pF
86dB	10.8pF
92dB	43.1pF

Figure 3.4.1: Comparison between the conventional DT pipelined-SAR ADC and the proposed CT pipelined-SAR ADC.

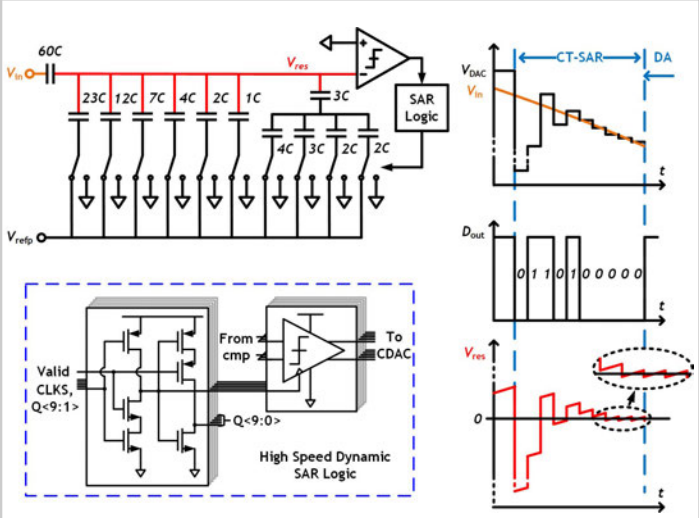


Figure 3.4.2: Schematic and behavior of the CT SAR stage with sub-radix-2 error correction CDAC and high-speed dynamic logic.

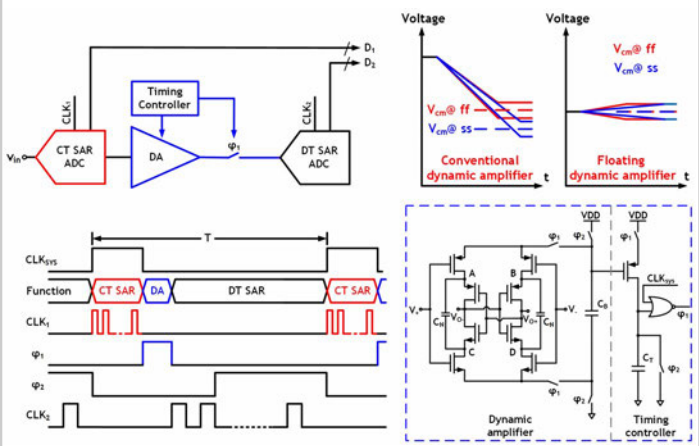


Figure 3.4.3: Block diagram (single-ended) and timing diagram of the proposed CT pipelined-SAR ADC, and the schematic of the floating dynamic amplifier.

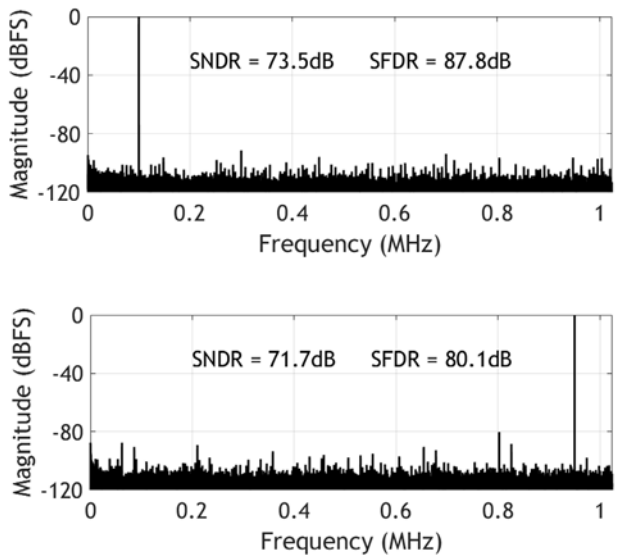


Figure 3.4.4: Measured output spectra with 100kHz and 950kHz inputs.

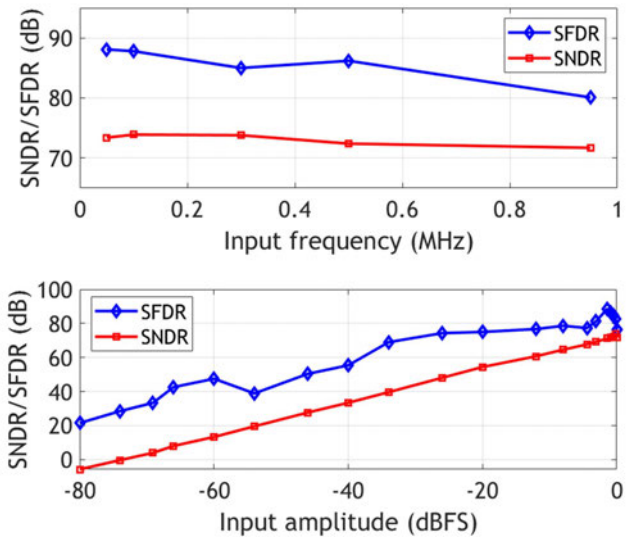


Figure 3.4.5: Measured SNDR and SFDR versus input frequency and amplitude.

	This work	ISSCC, 2017 [1]	JSSC, 2010 [2]	ISSCC, 2015 [5]	JSSC, 2017 [6]	ISSCC, 2015, [7]
Technology	**40nm LP**	28nm	180nm	40nm	40nm	65nm
Architecture	**CT pipe-SAR**	CT pipeline	CT pipeline	SAR	SAR	Pipe-SAR
1st-stage kT/C noise-free	✓	✓	✓	×	×	×
1st-stage input capacitance (pF)	0.06	Resistive	Resistive	9*	16	4
Area (mm^2)	0.01	5.1	1.9	0.18	0.32	0.054
Resolution (bits)	**13**	N/A	11	14	15	13
Sampling rate (MS/s)	**2**	9000	26	0.032	0.02	50
Input swing ($V_{pp,diff}$)	**2.5**	N/A	3.6	1.6	1.8	2.4
$SFDR_{NYQ}$ (dB)	**79.4**	73	67	78.5	95.1	84.6
$SNDR_{NYQ}$ (dB)	**71.7**	68	61.1	69.7	74.1	70.9
$ENOB_{NYQ}$ (bit)	**11.6**	11	9.9	11.3	12	11.5
Total power (uW)	**25.2**	2330000	26700	0.352	1.17	1000
FoM_w (fJ/conv-step)	**3.9**	495	1113	4.4	14.1	6.9
FoM_s (dB)	**177.8**	160	148	176.3	173.4	174.9

Figure 3.4.6: Performance summary and comparison with state-of-the-art ADCs.

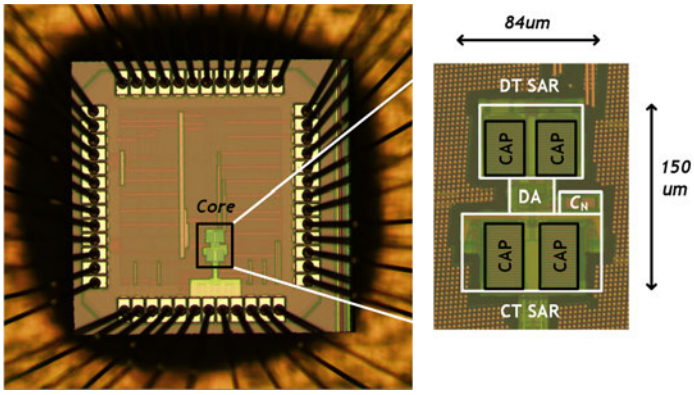


Figure 3.4.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

3.5 A 0.6V 13b 20MS/s Two-Step TDC-Assisted SAR ADC with PVT Tracking and Speed-Enhanced Techniques

Minglei Zhang[1], Chi-Hang Chan[1], Yan Zhu[1], Rui P. Martins[1,2]

[1]University of Macau, Macau, China
[2]Instituto Superior Tecnico/University of Lisboa, Lisbon, Portugal

The two-step SAR ADC is an energy-efficient architecture for high-resolution applications, which faces headroom challenges from the voltage-domain residue amplification under a low power supply. A TDC-assisted SAR ADC [1] uses a voltage-to-time converter (VTC) and a TDC as the back-end to quantize the residue voltage of the SAR ADC, which is attractive to the low power supply scenarios by moving the voltage domain quantization of a two-step SAR ADC into the time domain. However, the TDC-assisted SAR ADC encounters several design challenges. For example, it is sensitive to PVT variations due to partial time-domain operation, and its conversion speed is limited significantly by the long VTC latency when there is a small residue input voltage. This paper presents both PVT tracking and speed enhancement techniques for a 13b two-step TDC-assisted SAR ADC with 0.6V supply. The VTC and the back-end TDC are designed to have a common operation characteristic; thereby, their variations over PVT are inherently tracked without the need of any extra power or circuit overheads. The prototype ADC achieves less than 0.8dB SNDR drop across –50°C to 90°C and ±5% power supply variation at 20MS/s. The Walden FoM is 1.4fJ/conversion-step.

Figure 3.5.1 shows the block and timing diagram of the presented hybrid ADC. A 7b subranging SAR ADC adopts a detect-and-skip switching scheme modified from [2] for voltage-domain quantization. The switching scheme replaces the floated operation in [2] with a solid-connected V_{CM} to keep the common-mode voltage constant during the SAR conversion, while retaining the energy-efficient and high linearity features. A fully dynamic VTC incorporating discharging branches and threshold-crossing detectors (TCDs) converts the residue voltage of the SAR ADC into a time difference (T_P and T_N). Following this, a single-ended time generator produces the input (STA and STO) for the single-ended TDC and gives one sign bit result simultaneously. It also moves the reset instant of STA and STO to the end of the upcoming sampling period for more time margin to the TDC. Unlike conventional two-step TDCs that utilize the PVT-unfriendly time amplifier, the time difference is quantized directly by a 4b flash TDC cascaded with a 3.5b Vernier TDC herein. A 30MS/s maximum conversion rate (with >71dB@LF SNDR) is achieved under a power supply as low as 0.6V, which benefits from 3 design strategies in the time domain: 1) reconfigurable latency VTC; 2) two-step TDC for finer time resolution; 3) pushing part of the TDC conversion into the next sampling period.

The PVT-stabilized technique for the voltage-domain dynamic amplifier in [3] involves a static amplifier and complicated timing control logic. The developed PVT tracking technique eliminates such auxiliaries by tracking the VTC variations with the inherent back-end TDC. Figure 3.5.2 shows the circuit implementation and operation principle of the presented technique. The current source I_D of the VTC consists of one branch of 10 serial unit NMOS and 6 configurable branches of 20 serial unit NMOS. The delay cells in the flash TDC and Vernier TDC have the same topology but with different loading capacitance C_C. Moreover, the current drawn by the rising-edge-related PMOS βI_U is intentionally designed >4× larger than the NMOS counterpart I_U to alleviate the influence from the PMOS-related delay. Based on simulations, the factor β only changes ±3% across –50°C to 90°C and ±5% power supply variation, while the process influence is covered by the configurable I_D of the VTC. The unit NMOS of the current source I_D in the VTC has the same size as the rising-edge-related unit NMOS in the delay cell, and all biased by V_{DD} ($=V_B=V_{OS}=V_T$). Therefore, I_D and I_U share the same PVT response, which results in a PVT-insensitive full-scale (FS) code of the TDC. A low-noise TCD is designed to meet the 13b noise requirement with a dynamic pre-amp and C_1, C_2, while the latency of the VTC can be configured through V_{REF} to achieve a fast V-T conversion. The charge held on the fine SAR CDAC starts to be discharged through I_D at the rising edge of CK_T, and the time outputs T_P and T_N are generated once the residue voltages V_P and V_N cross V_{REF}, respectively. The TCD and the current source I_D are powered down after the time difference T_{RES} is generated for better energy efficiency.

The time residue generator (TRG) in Fig. 3.5.3 connects the flash TDC and Vernier TDC with inherent PVT robustness by removing a time amplifier. To shield the metastability from the time comparator in the flash TDC, the time residue (t_r+t_d) is generated in a complementary manner with a fixed offset t_d. Extra time offset t_d is also added to both STA and STO signal paths for more timing margins under low power supply. Figure 3.5.3 presents a time residue generated example with 1.38t_d input to the total TDC. The output code of the Vernier TDC should be inverted to match the residue-generated manner. Different from the TRG in [4] that consumes static power and suffers from non-linearity, a fully dynamic TRG with offset bit shifter is presented, as shown in Fig. 3.5.3. Nodes X and Y are pre-charged before the activation of the delay line in the flash TDC, and then the relevant discharging branches are enabled to generate the rising edges of the residue time difference (S and F). There are 12 delay stages in the 3.5b Vernier TDC, where 4 of them are used for time quantization and the remaining 8 for the offset measurement to the TRG. The offset caused by the TRG is inherently measured and suppressed at the digital back-end through counting the average output code of the Vernier TDC, and then subtracting the TDC output with the measured offset codes. The above operations do not require a known input signal property and have no external intervention, while no further linearity calibration to the TDC delay cells is applied.

The prototype ADC is fabricated in a 1P9M 65nm CMOS process with a core size of 0.053mm^2 as shown in Fig. 3.5.7. Only one-time foreground calibration to the VTC FS time difference and offset is processed, while all the other error sources are covered by the redundancies between stages and the intrinsic matching. A custom designed MOM capacitor array with a single-ended capacitance of 4pF is used to achieve a 13b-level INL of –1.03/+1.31 LSB in Fig. 3.5.4. From the measured 66536-point FFT spectrum in Fig. 3.5.4, the SNDR and SFDR at 20MS/s with a 0.49MHz input signal are 71.5dB and 91.9dB, respectively. The SNDR with Nyquist input is 71.0dB, and stays above 69.5dB with an 18MHz input (Fig. 3.5.4). The presented hybrid ADC exhibits a fully dynamic feature with a similar FoM across different conversion rates. The maximum conversion rate of the design is 30MS/s, while a large design room is reversed to guarantee that the ADC performance is not significantly affected by the sampling network and other non-idealities under PVT variations. The output data of the ADC is decimated by 5 to mitigate the ripple coupled inside the PAD ring. Figure 3.5.5 shows the measured SNDR variation versus temperature and power supply variations of three randomly selected chips, the maximum SNDR drop is less than 0.8dB and 0.5dB across –50°C to 90°C and ±5% power supply variations, respectively. The total power consumption at 20MS/s is 82μW. Figure 3.5.6 summarizes the performance of the ADC and compares it with other state-of-the-art hybrid ADCs. The prototype ADC is not only the only PVT-robust TDC-assisted SAR ADC to date, but it also achieves the best reported FoM of an ADC with both more than 70dB SNDR and 20MS/s conversion rate.

Acknowledgement:
This work was financially supported by the Macao Science & Technology Development Fund (FDCT) with Ref no: 0068/2018/A2, Research Grant from University of Macau (MYRG2018-00104-AMSV), and NSFC Grant 61604180.

References:
[1] Y. J. Chen, et. al., "A 2.02-5.16 fJ/conversion step 10 bit hybrid coarse-fine SAR ADC with time-domain quantizer in 90nm CMOS," *IEEE JSSC*, pp. 357-364, Feb. 2016.
[2] Y. Lim, M. P. Flynn, "A 1mW 71.5dB SNDR 50MS/s 13b fully differential ring-amplifier-based SAR-assisted pipeline ADC," *ISSCC*, pp. 458-459, Feb. 2015.
[3] H. Huang, et. al., "A 12b 330MS/s pipelined-SAR ADC with PVT-stabilized dynamic amplifier achieving <1dB SNDR variation," *ISSCC*, pp. 472-473, Feb. 2017.
[4] K. Kim, et. al., "A 7 bit, 3.75 ps resolution two-step time-to-digital converter in 65 nm CMOS using pulse-train time amplifier," *IEEE JSSC*, pp. 1009-1017, Apr. 2013.
[5] M. Ding, et al., "A 5.5fJ/conv-step 6.4MS/s 13b SAR ADC utilizing a redundancy-facilitated background error-detection-and-correction scheme," *ISSCC*, pp. 460-461, Feb. 2015.
[6] A. Sanyal, N. Sun, "A 18.5-fJ/step VCO-based 0-1 MASH ΣΔ ADC with digital background calibration," *IEEE Symp. VLSI Circuits*, pp. 26-27, June 2016.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

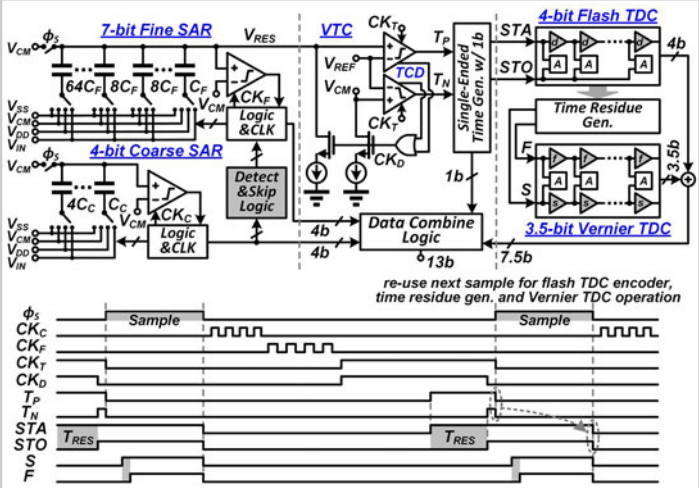


Figure 3.5.1: Block and timing diagram of the prototype 13b two-stage TDC-assisted SAR ADC with speed enhancement techniques (single-ended SAR ADC for simplicity).

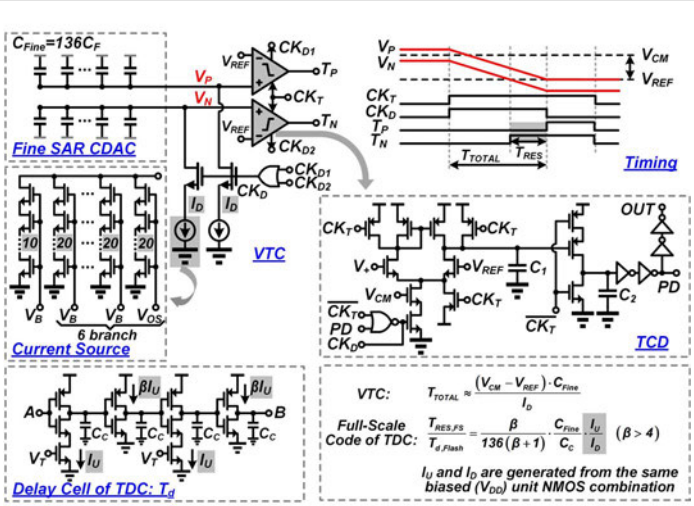


Figure 3.5.2: Circuit implementation and principle of the PVT tracking technique.

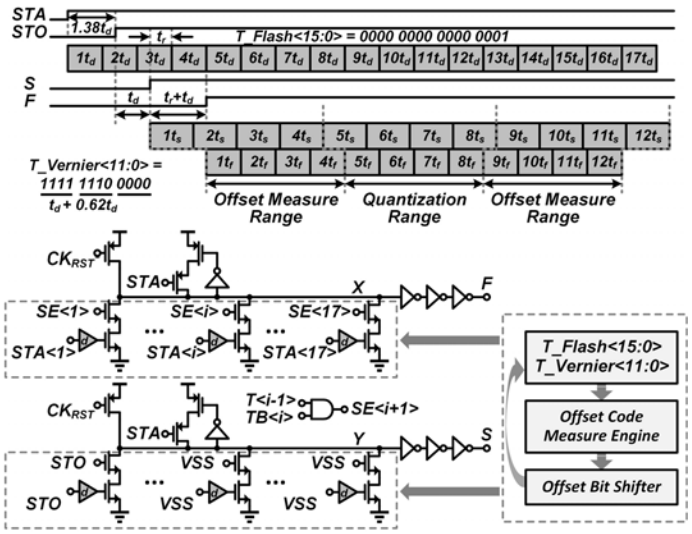


Figure 3.5.3: Fully dynamic time residue generation between the two stage TDC with offset bit shifter.

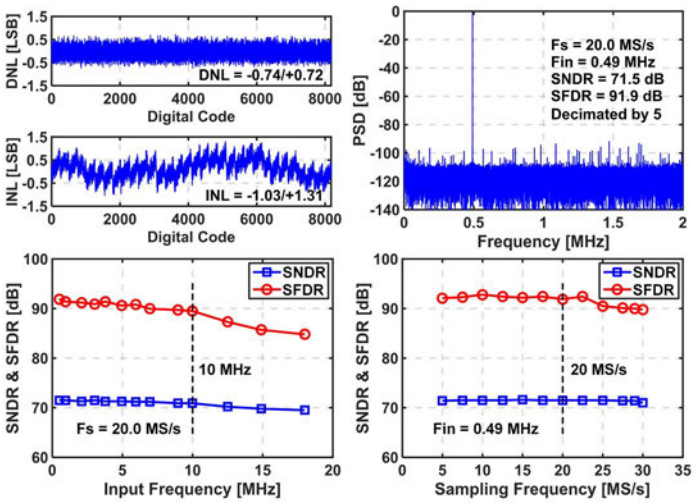


Figure 3.5.4: Measurement results (decimated by 5).

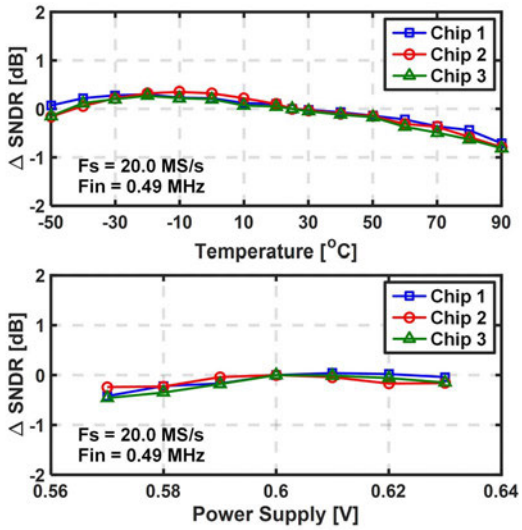


Figure 3.5.5: Measured SNDR variation versus temperature (–50°C to 90°C) and ±5% power supply variation with 20MS/s sampling frequency.

	This Work	[1] JSSC16	[2] ISSCC15	[5] ISSCC15	[6] VLSI16
Architecture	**SAR-TDC**	SAR-TDC	Pipe-SAR	SAR	SAR-VCO
Technology	**65nm**	90nm	65nm	40nm	40nm
Resolution [bits]	**13**	10	13	13	-
Active Area [mm²]	**0.053**	0.041	0.054	0.068	0.030
Power Supply [V]	**0.6**	0.6	1.2	1.0	1.1
Sample Rate (or 2×BW) [MS/s]	**20**	2	50	6.4	6
SNDR @ Nyq. [dB]	**71.0**	54.5	70.9	64.1	71.4
SFDR @ Nyq. [dB]	**89.5**	82.3	84.6	81.9	-
Total Power [μW]	**82**	4.6	1000	46	350
FoM_W @ Nyq. [fJ/conv.step]	**1.4**	4.0	7.0	5.5	18.5
FoM_S @ Nyq. [dB]	**181.9**	167.9	174.9	172.5	170.7

Figure 3.5.6: Performance summary and comparison with other hybrid ADCs.

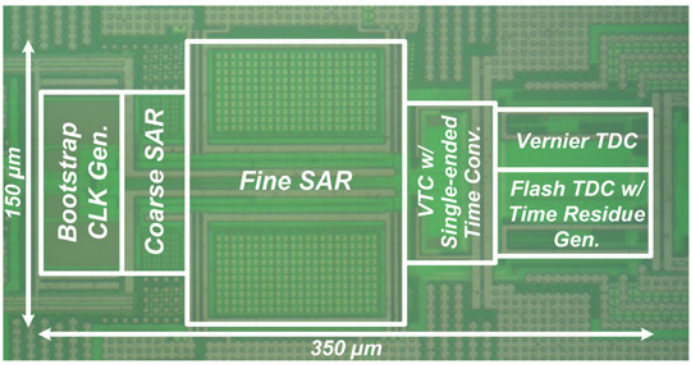


Figure 3.5.7: Die microphotograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

3.6 A 6-to-600MS/s Fully Dynamic Ringamp Pipelined ADC with Asynchronous Event-Driven Clocking in 16nm

Benjamin Hershberg, Barend van Liempd, Nereo Markulic, Jorge Lagos, Ewout Martens, Davide Dermit, Jan Craninckx

imec, Leuven, Belgium

At the upper end of achievable ADC operating speeds, clocking becomes a critical performance limiter. In "deep" pipelined ADCs that contain many stages, the clock tree constitutes a highly distributed network, with parasitics and mismatch creating skew between the different branches. Sufficient margin must be included in the timing generation such that all non-overlap and causal relationships are maintained. This leads to a difficult set of design tradeoffs in terms of power, speed, jitter, and reliability. Meanwhile, although residue amplifiers have traditionally dominated the power budget in deep pipelines, recent advances such as ring amplification have improved achievable efficiencies to the point that clocking is now the primary consumer in some cases [1,2].

It is already well known that asynchronous event-driven clocking techniques can provide speed and power advantages in SAR and pipelined-SAR ADCs. However, in deep pipelines this approach remains largely unexplored at the system level. In this paper, we propose a clocking paradigm where the conventional hierarchical clock tree is replaced by localized control units that interact with each other using event-driven triggers and handshake protocols. The result is a modular, correct-by-construction timing system that minimizes routing complexity and maximizes performance. The numerous advantages of the approach are combined with the efficiency of ring amplifiers to build a single-channel 11b, 60dB SNDR, 78dB SFDR pipelined ADC with fully dynamic power consumption that maintains better than 13fJ/conv-step Walden FoM from 6MS/s to 600MS/s.

Illustrated in Fig. 3.6.1, the single-channel ADC consists of seven 1.5b stages followed by a 1.5+3b backend. Each stage in the pipeline communicates with its direct neighbors through an event-driven protocol. This can be seen in the lower portion of Fig. 3.6.1, where a digital control unit in each stage sends and receives commands. The wires linking these control units together serve a dual-purpose: they provide both inter-stage communication and local line driving of circuits. This is more power and area efficient than hierarchical clocking where global distribution and local buffering are handled separately.

Figure 3.6.2 shows a simplified version of the event-driven timing that occurs during normal operation. It begins in the system reset state, where all stages are waiting to sample. The arrival of the *master clock* initiates a chain reaction. First, stage 1 samples the input and begins quantizing. When quantization is complete and stage 2 also indicates that it is ready, stage 1 amplifies the residue. After internal timing controls determine the residue is settled, *sample REQ* is sent to stage 2. Stage 2 then samples and returns an acknowledgement, allowing stage 1 to power down its ringamp and return to the *track* state. The inherent causality of this approach makes it correct-by-construction; safe and valid phase relationships are guaranteed.

Stage 2 now begins a similar exchange with stage 3, with a few notable differences. First, all stages except stage 1 use an early quantization technique that occurs in parallel with sampling. This removes the quantization phase from the critical timing path (explained later). Second, when stage 2 is ready to resume tracking, it will wait until the *begin* signal of stage 1 signal indicates that it is safe to reconnect. This prevents stage 2 from reconnecting to the output of stage 1 just as stage 1 samples, and avoids kickback related errors.

As illustrated in Fig. 3.6.1, the digital outputs of each stage propagate to the end of the pipeline for alignment and then shift back to stage 1 using the causally-related edges of *sample REQ*, where they can be re-synchronized with the master clock.

Event triggers can be generated with delay cells, derived directly from analog events, or a combination of both. For example, the sub-ADC comparators can self-detect completion, but also use a timeout delay to protect against metastability. Another example, as we introduce in Fig. 3.6.3, is a technique for sensing the end of a ringamp's slewing operation. Consider the binary value of the voltages at nodes A, B, C, and D. During slewing, the logic equation AB'CD' will never equal '1', because A=B and C=D. During settling, the dynamic formation of a dead-zone – in this case due to the CMOS resistor [2] – forces the four nodes apart such that A=C='1' and B=D='0' and the logic equation evaluates to '1'. Based on this principle, the simple detection circuit in Fig. 3.6.3 generates the *sub-ADC latch* signal of Figs. 3.6.1 and 3.6.2. This is used to implement early quantization [3], where the sub-ADC samples the residue of the previous stage early and quantizes it in parallel with continued residue settling. Our approach provides an optimal solution to the classic challenge of this technique, namely, when to sample. Sampling too soon may result in too much quantization error, and sampling too late will impede optimal settling due to disturbance of the amplifier output.

Triggering on analog events requires that the analog blocks themselves function correctly. This is not always the case during startup, when e.g. common-mode or differential saturation of the ringamp may occur. In such cases, slewing will never finish and the *sub-ADC latch* signal will never assert. The result is a deadlock. Fortunately, this can easily be detected (Fig. 3.6.2). Any deadlock will eventually propagate back to stage 1, and if a *master clock* edge arrives at stage 1 before it has begun sampling the input, we know that either: 1) the clock speed is too high for the given settings or 2) a deadlock has occurred. When this condition is detected, a "flush" command is issued that forces each stage into its *track* state. Analog blocks such as ringamp CMFB are designed so that repeatedly returning to this reset condition will result in correct bias convergence.

The internal processing speed of the pipeline is completely independent of the input clock. This has several important advantages. First, it automatically maximizes track time. Second, many critical leakage constraints are eliminated, significantly increasing the range of clock speeds that can be supported. For example, the MDAC switches can be sized for the maximum speed without concern of leakage corrupting the charge sampled onto the MDAC capacitors at low clock speeds. Third, reconfiguring the internal operation – even inserting or removing entire phases – is both trivial to implement and self-optimized for timing efficiency. Fourth, it eliminates the subtle but crucial tradeoff between input track time and sampling jitter in conventional clock trees that is particularly relevant at high speeds.

The ADC is fabricated in a 16nm technology, occupying an active area of 0.037mm^2 (110µm × 340µm), excluding decoupling (Fig. 3.6.7). It operates on a 0.85V supply with 50mV/800mV references. At 600MS/s with a Nyquist input it consumes 6.0mW (including clock buffer) and achieves 60.2dB SNDR, 78.3dB SFDR, 12.0fJ/conv-step FoM_W, and 167.2dB FoM_S. Measured performance is summarized in Figs. 3.6.4 and 3.6.5. All results use the same digital settings and stage gain coefficients (calculated off-chip). The fully dynamic, linear scaling of power with respect to clock frequency demonstrated in Fig. 3.6.4 makes this architecture an excellent candidate for reconfigurable applications such as multi-standard wireless. The supply sweep of Fig. 3.6.4 demonstrates the robustness of the ringamps to variation (to sweep down to 0.8V without saturating, the references are set to 50mV and 750mV for this test.)

A comparison with relevant state-of-the-art is provided in Fig. 3.6.6. This design achieves the best Walden FoM of any ADC above 410MS/s in [4], indicating that ringamp-based single-channel pipelined ADCs can offer a power-efficient alternative to multi-channel SAR architectures, without the overhead of interleave calibration. Furthermore, the versatility of event-driven control creates new possibilities for developing more efficient, reconfigurable, and robust solutions to classic challenges in pipelined ADCs.

References:
[1] J. Lagos, et. al., "A 1Gsps, 12-bit, single-channel pipelined ADC with dead-zone-degenerated ring amplifiers," *IEEE CICC*, May 2018.
(power breakdown by private communication)
[2] J. Lagos, et. al., "A single-channel, 600Msps, 12bit, ringamp-based pipelined ADC in 28nm CMOS," *IEEE Symp. VLSI Circuits*, pp. 96-97, June 2017.
[3] B.-M. Min, et. al., "A 69-mW 10-bit 80-MSample/s Pipelined CMOS ADC," *IEEE JSSC*, vol. 38, no. 12, pp. 2031-2039, Dec. 2003.
[4] B. Murmann, "ADC Performance Survey 1997-2018," [Online]. Available: http://web.stanford.edu/~murmann/adcsurvey.html.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

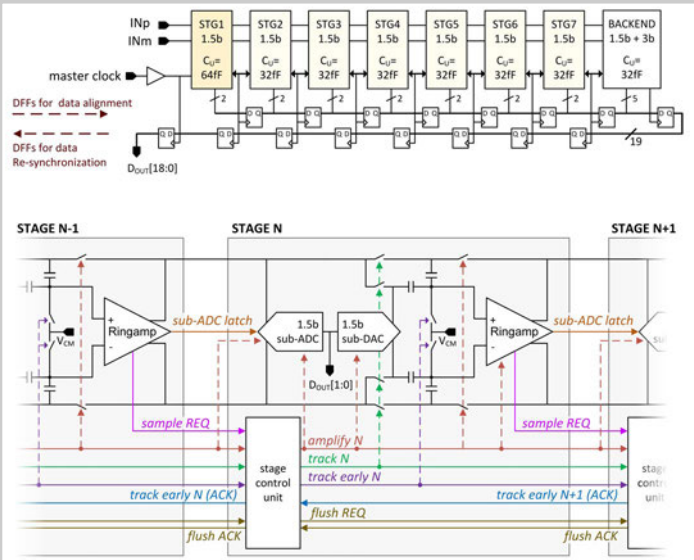


Figure 3.6.1: ADC top level and simplified MDAC with event-driven control.

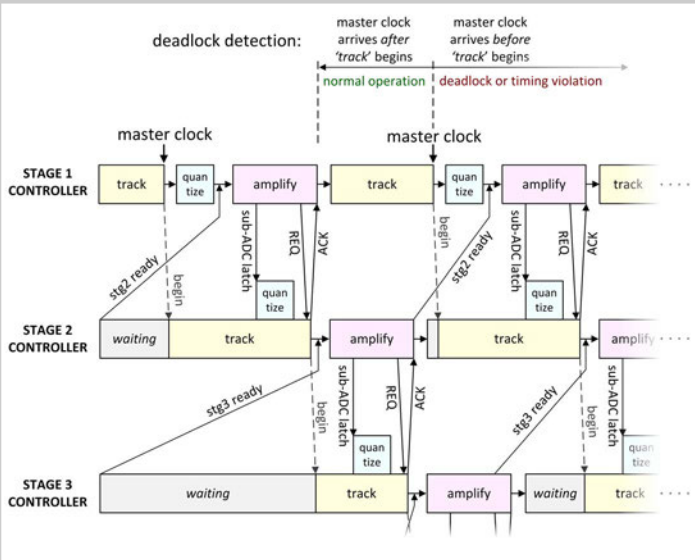


Figure 3.6.2: Timing diagram with simplified event-driven communication.

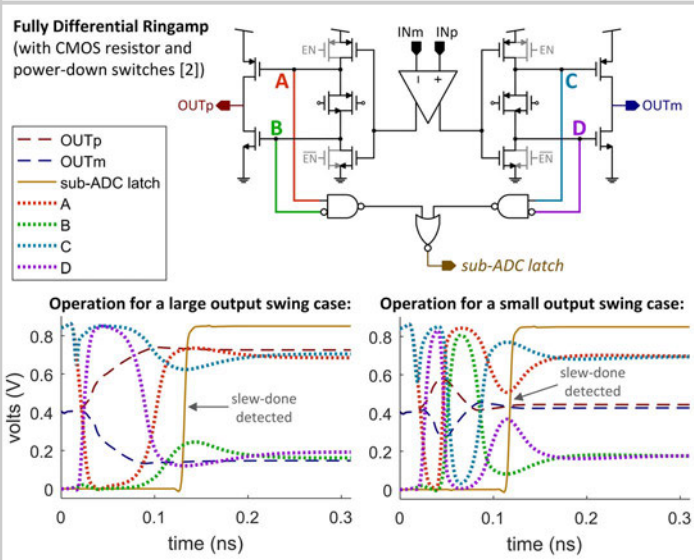


Figure 3.6.3: Simplified ringamp and two examples of slew-done detection.

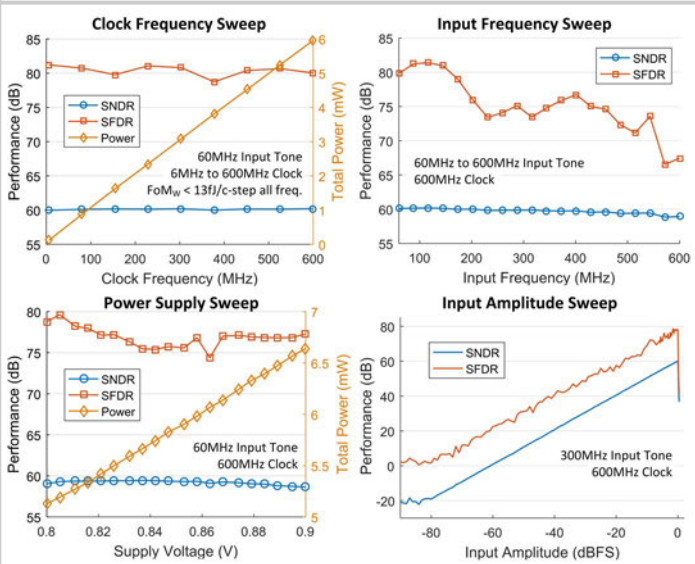


Figure 3.6.4: Characterization sweeps.

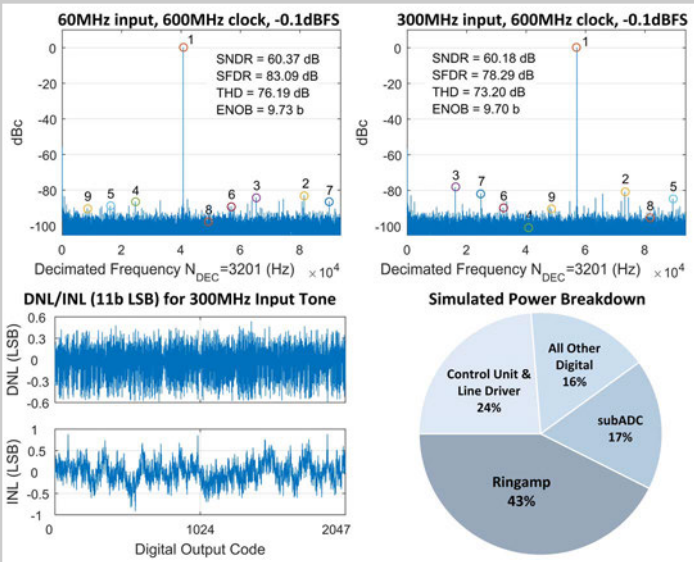


Figure 3.6.5: Measured dynamic and static performance.

	High Speed Ringamp SoTA			SoTA in [1] for SNDR > 56dB, Fs > 410MHz			
	This work	Lagos CICC 2018	Lagos VLSI 2017	Venca ISSCC 2016	Nam VLSI 2016	Lien VLSI 2016	Sung ISSCC 2015
Architecture	**Pipeline**	Pipeline	Pipeline	TI SAR ΔΣ	TI SAR ΔΣ	TI Pipe-SAR	TI FATI-SAR
Sampling rate (max) [MS/s]	**600**	1000	600	600	1600	800	1600
Number of channels	**1**	1	1	4	8	4	12
Per-channel rate [MS/s]	**600**	1000	600	150	200	200	133
Residue Amplifier	**Ringamp**	Ringamp	Ringamp	-	-	Opamp	-
Resolution [bit]	**11**	12	12	12	12	12	10
Technology [nm]	**16**	28	28	28	65	28	65
Supplies [V]	**0.85**	0.9	0.9	1.2, 1.5	1.1, 1.2	1	1.1
Input range [Vpp]	**1.5**	1.6	1.6	1.5	1.6	1.2	n.a.
ENOB LF input [bit]	**9.7**	9.2	9.4	9.8	11.6	10.0	9.2
ENOB Nyquist [bit]	**9.7**	9.1	9.1	9.3	10.6	9.8	9.0
SNDR LF input [dB]	**60.4**	57.1	58.1	60.7	71.8	62.2	57.2
SNDR Nyquist [dB]	**60.2**	56.6	56.3	58.0	65.3	60.8	56.1
SFDR LF input [dB]	**83.1**	74.6	67.5	77.5*	83.0	79.0*	68.0
SFDR Nyquist [dB]	**78.3**	73.1	69.2	66.0*	65.3	75.0	61.2
Power [mW]	**6.0**	24.8	14.2	26.5	37.7	14.6	17.3
FoM_{Walden} [fJ/c.step]	**12.0**	45.0	44.3	68.0	17.8	20.0	21.0
$FoM_{Schreier}$ [dB]	**167.2**	159.6	159.5	158.5	168.6	165.2	162.8
Core area [mm²]	**0.037**	0.540	0.620	0.042	0.900	0.093	0.360
Calibration	**gain**	gain	gain	offset, gain, mismatch	offset, gain, timing skew	offset, gain	offset

* estimated from figure

Figure 3.6.6: Comparison with relevant state-of-the-art.

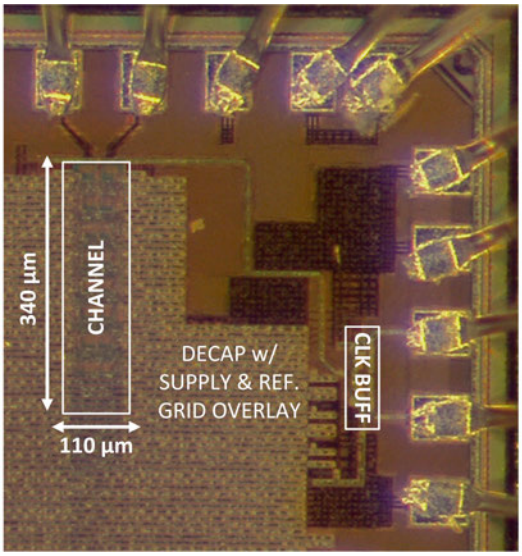


Figure 3.6.7: Chip Micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

3.7 A 10mW 16b 15MS/s Two-Step SAR ADC with 95dB DR Using Dual-Deadzone Ring-Amplifier

Ahmed ElShater[1], Calvin Yoji Lee[1], Praveen Kumar Venkatachala[1], Jason Muhlestein[1], Spencer Leuenberger[1], Kazuki Sobue[2], Koichi Hamashita[2], Un-Ku Moon[1]

[1]Oregon State University, Corvallis, OR
[2]Asahi Kasei Microdevices, Atsugi, Japan

The two-step SAR architecture has been a popular choice for power-efficient ADCs used in applications such as medical imaging. The simple and scalable architecture of the SAR ADC enables efficient multi-bit conversion per stage [1,2] however, the maximum performance of the two-step SAR is limited by the residue amplifier. The ring amplifier (*RAMP*) has enabled efficient two-step SAR ADCs [3]. Despite this, *RAMP*-based high-precision ADCs (>16b) remain unexplored. This paper presents a two-step SAR ADC utilizing a fully differential *RAMP* achieving 95dB DR. Two techniques that contribute to the energy-efficiency of the ADC are also presented: dynamic digital supply and DAC impedance-matching.

The two-step SAR ADC (Fig. 3.7.1) consists of a 9b 1st-stage (*STG1*) and 13b 2nd-stage (*STG2*) with an inter-stage gain of 64× from the residue amplifier (*RA*). Due to redundancy, the ADC yields a total of 18b quantization. *STG1* goes through four main phases. The input is sampled onto *STG1* DAC followed by *STG1* conversion. Once completed, the remaining residue is transferred and amplified onto *STG2* by the *RA* followed by *STG1* DAC reset. *STG2* is a conventional SAR with two phases: sampling and conversion. Both *STG1* and *STG2* SAR logic utilize asynchronous timing for bit decisions.

The table in Fig. 3.7.1 describes the partitioning of *STG1* DAC and operation in each phase. To minimize *STG1* conversion time, 4MSB decisions are resolved simultaneously with a 4b MSB-folding Flash ADC. A three-level switching reduces the capacitive loading on the reference while the MSB-folding architecture reduces input signal loading and power consumption compared to conventional 4b Flash. Redundancy between the Flash and SAR section of the *STG1* DAC account for Flash mismatch and offset errors.

The remaining bits of *STG1* DAC are resolved sequentially in a SAR fashion. Dither added during sampling randomizes the residue seen by the *RA* and *STG2* for improved DNL and SFDR performance. The dither reference is scaled (11/48*Vref) to ensure that all 15 levels of dither propagate to *STG2*. *STG1* SAR comparator offset is trimmed and redundancy employed between *STG1* and *STG2* ensures that the residual offset of the SAR1 comparator in addition to *STG1* SAR errors do not exceed the input dynamic range of *STG2*.

Figure 3.7.2 shows the architecture of the *STG2* ADC with three bridged sub-DACs. Redundancy is introduced between each sub-DAC allowing for shorter bit cycle durations for MSB decisions. The *STG2* input network provides an additional digital gain of 2× through the ratio between Cin2 and Cdac. Due to the large number of bits to be resolved, *STG2* design focuses mainly on minimizing the bit cycle duration. In addition to utilizing dynamic bit logic, dynamic digital supply and impedance-matching were employed to further speed up conversion time:.

Traditionally, variable bit cycle durations are implemented by digitally controlled delay cells in the asynchronous timing logic. In this work, variable durations are achieved by the gradual reduction of the digital supply (Fig. 3.7.2). The capacitor which is precharged to VDD during the sampling phase is disconnected at the beginning of conversion and acts as a charge reservoir for the digital supply. Charge is drawn with every bit cycle resulting in a gradual decay of voltage VDD_DYN. Hence, the bit cycle duration defined by the inverter-based delay line increases with the decaying VDD_DYN. Additionally, this technique mimics the power savings of operating the digital section at a lower supply without a dedicated lower supply rail and provides isolation between the analog and digital supply of *STG2*.

The impedance-matching technique achieves fast DAC settling at the comparator input without excessive switch sizing and power-hungry switch drivers. The comparator inputs settle almost instantly by ensuring an all-pass transfer function from the DAC references to the comparator input. By sizing the DAC switch resistances to have the same time constant for each capacitor weight, the transfer function from any reference to the comparator input, Vtop, becomes a frequency-independent ratio (Fig. 3.7.2). For a bridged-DAC structure like that of *STG2*, the bridged capacitors, Cb, require the insertion of equivalent series resistances, Rb, to maintain the impedance matching concept.

The *RA* plays an important role in the overall performance of a two-step SAR architecture by providing residue amplification which relaxes the noise requirements of the *STG2* comparator. However, since the noise and non-linearity of the *RA* directly affect the performance of the entire ADC, the *RA* limits the power efficiency of the structure. *RAMP* for residue amplification has been shown to be power efficient while exhibiting a fast settling response [3]. The settling response of *RAMP* (Fig. 3.7.3) can be roughly divided into three phases: slewing, stabilization and steady-state. During slewing, the first two stages of the *RAMP* quickly amplify the residue to near supply levels at the input of the *RAMP*'s output stage and the push-pull nature of the output stage provides large peak currents (I_{slew}) relative to the quiescent current (I_Q). The output voltage approaches the final settled value during the stabilization phase and in the steady-state phase, the output stage current returns to I_Q , set by Vbp and Vbn. Steady-state GBW is determined by the *gm* of the output stage ($gm3_{ss}$).

Current *RAMP* suffers from limitations when optimizing for peak I_{slew} and stability. For illustrative purposes, Fig. 3.7.3 shows the static V-I curves and effective large signal gm across input swing for two sizes: small and large. Each were biased to obtain the same $gm3_{ss}$. The stable gm region shown in the figure represents the region where amplifier dynamics would have good small signal phase margin. In practice, the bandwidth and large signal delay of the first two stages are influencers of these limits. For faster settling and higher I_{slew}, large devices would be preferred but maintaining gm_{ss} sets the output bias at smaller overdrive voltages. This causes a reduction in the stable region width and extends the stabilization phase. Additionally, smaller overdrive voltages make the amplifier more sensitive to the accuracy of the bias voltages.

The dual-deadzone output stage shown in Fig. 3.7.3 comprises of two parallel branches with low (*LVT)* and high (*HVT)* threshold voltages. The *HVT* branch is sized up to provide the I_{slew} current. By setting the bias voltage below the *HVT* threshold voltage, the *HVT* devices are switched off in steady state while the *LVT* branch defines the gm_{ss} and I_Q.. The additional degree of freedom represented in the *LVT* branch allows for large I_{slew} without compromising the stable region width as shown in Fig. 3.7.3. In addition, the bias sensitivity is reduced by operating the *LVT* devices with a large overdrive. Finally, the *LVT* branch is cascoded to obtain 110dB of DC gain – necessary for linearity requirements – without affecting I_{slew}.

The schematic of the fully differential *RAMP* is shown in Fig. 3.7.4. The local CMFB in *STG2* provides a lower CM impedance for stabilizing the global output CMFB. Auto-zeroing is performed between each stage sequentially during *STG1* conversion. Enable switches allow for duty-cycling *RAMP* power. The gm devices of the output stage *LVT* branches are digitally trimmable for three *LVTtrim* settings. Neutralization capacitors, *Cneut*, are placed across the output stage to avoid DC gain reduction due to *Cgd* of output stage devices. An on-chip amplifier settling measurement technique provides insight into the settling response of the residue amplifier for various bias and output stage sizes (Fig. 3.7.4).

The ADC is implemented in a 0.18µm CMOS process. Foreground bit weight calibration is performed. Figure 3.7.5 shows the static (INL) and dynamic performance of the ADC at 2MS/s and 15MS/s. Figure 3.7.6 summarizes this work and compares with other two-step SAR ADCs. This work achieves the highest figure-of-merit for a *RAMP* based ADC. A die photo is shown in Fig. 3.7.7.

References:
[1] H. Li, et al., "A Signal-Independent Background-Calibrating 20b 1MS/s SAR ADC with 0.3ppm INL," *ISSCC*, pp. 242-243, Feb. 2018.
[2] D. Hummerston, P. Hurrell, "An 18-bit 2MS/s pipelined SAR ADC utilizing a sampling distortion cancellation circuit with -107dB THD at 100kHz," *IEEE Symp. VLSI Circuits.* pp. 280-281, June 2017.
[3] Y. Lim, M.P. Flynn, "A Calibration-free 2.3mW 73.2dB SNDR 15b 100 MS/s Four-Stage Fully Differential Ring Amplifier Based SAR-Assisted Pipeline ADC," *IEEE Symp. VLSI Circuits*, pp. 98-99, June 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

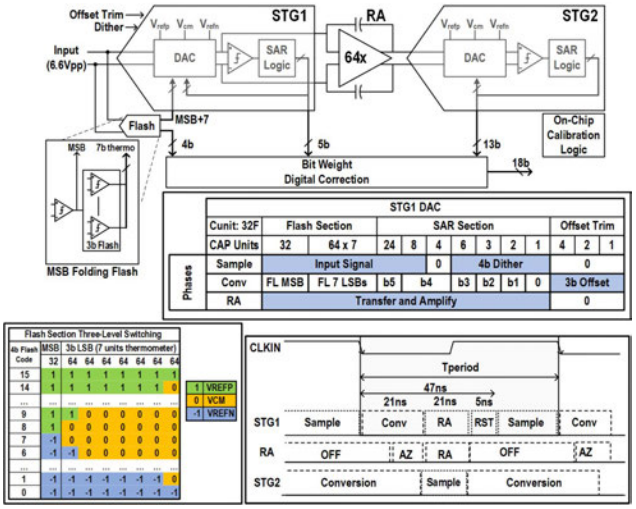


Figure 3.7.1: ADC block diagram and timing diagram with operation of three-level switching of Flash STG1 DAC section and partitioning/usage of STG1 capacitor DAC.

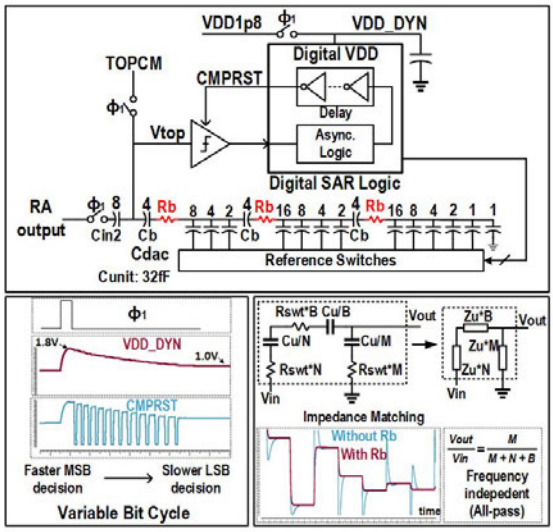


Figure 3.7.2: Block diagram of STG2 featuring dynamic digital supply operation and capacitor DAC impedance matching concept.

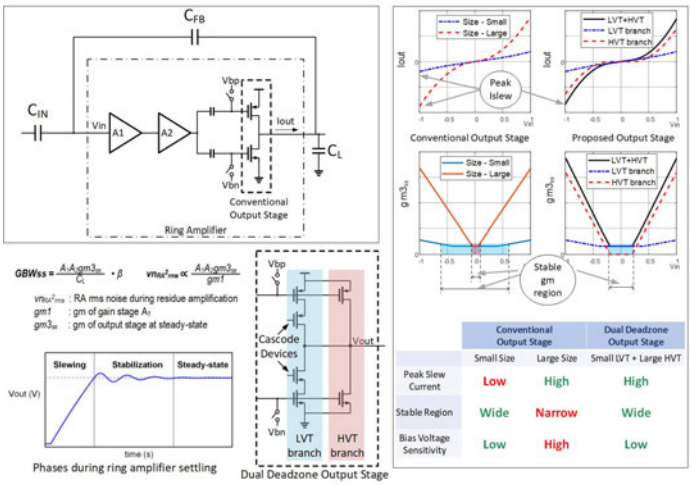


Figure 3.7.3: RAMP basic operation with conventional output stage design tradeoffs and dual deadzone output stage.

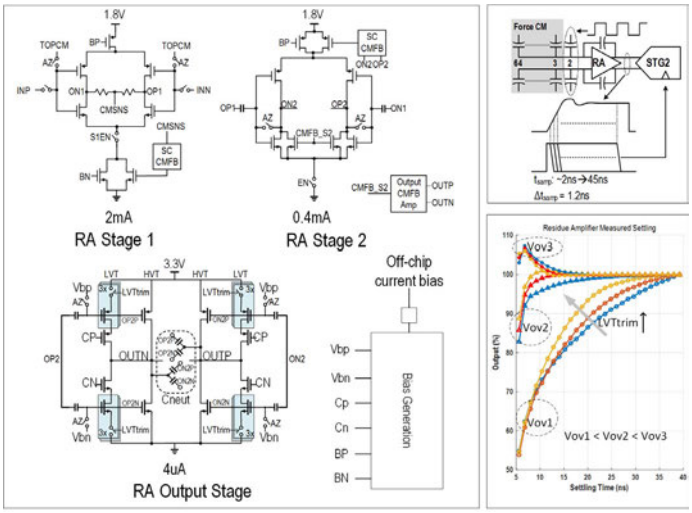


Figure 3.7.4: Simplified RAMP schematic with on-chip RAMP settling measurement functionality and measured waveforms.

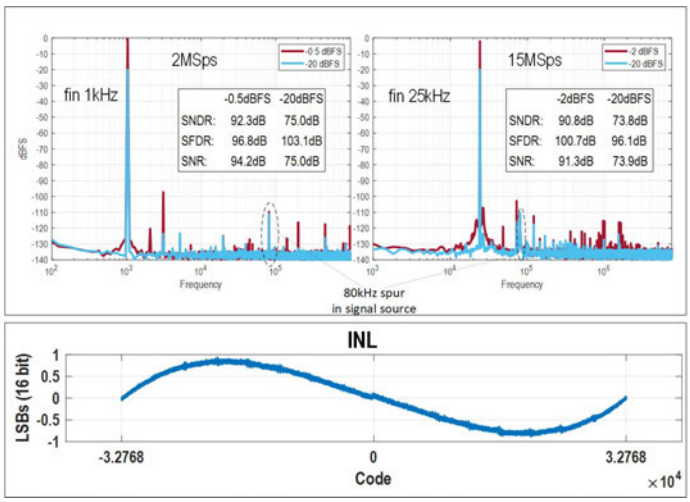


Figure 3.7.5: Measurements of dynamic performance at 2MS/s and 15MS/s and INL performance at 16 bit.

	This Work		Li	Hummerston	Ring Amplifier Based		Bannon	Hurrell
	2MS/s	15MS/s	[1] ISSCC 18	[2] VLSI 17	Lim [3] VLSI 17	Lim ISSCC 15	VLSI 14	ISSCC 10
Technology [nm]	180		180/500	180	40	65	180	250
Reference [V]	3.3		5.0	5.0	1.1	1.2	5.0	5.0
RA Gain	64		32	64**	64	32	64	42.66
Calibration	Foreground		Background	Trimmed**	None	None	Trimmed	Measured
Speed [MS/s]	2	15	1	2	100	50	5	12.5
SNDR [dB]	92.3	90.8†	101.5	99.2††	73.2	71.5	98.6	92.7
SFDR (THD) [dB]	96.8	100.7	132.0	-107.0	90.4	87.0	109.2	104/-100
SNR/DR [dB]	94.2/95.0	91.3†/93.9†	101.5††/102.7	100.0/100.5	73.3/-	71.9/-	99.0/100.2	93.0/93.0
ADC Power [mW]	1.61*	9.82*	12.9	11.35	2.3	1.0	30.52	105
Area [mm²]	1.82		4	3.87	0.068	0.054	5.74	5.74
FOM_{SNDR}	180.2	179.6†	177.4	178.6††	176.6	175.5	177.7	170.4
FOM_{SDR}	182.9	182.7†	178.6	179.9	–	–	179.3	170.7

* Not including digital output drivers.
** Information reported in table is based from previous design
† Actual performance is better by 1.1dB. Reported number is limited by ADC driver.
†† Information reported is calculated based on other performance numbers provided

Figure 3.7.6: Performance summary and comparison to state-of-the-art two-step SAR ADCs.

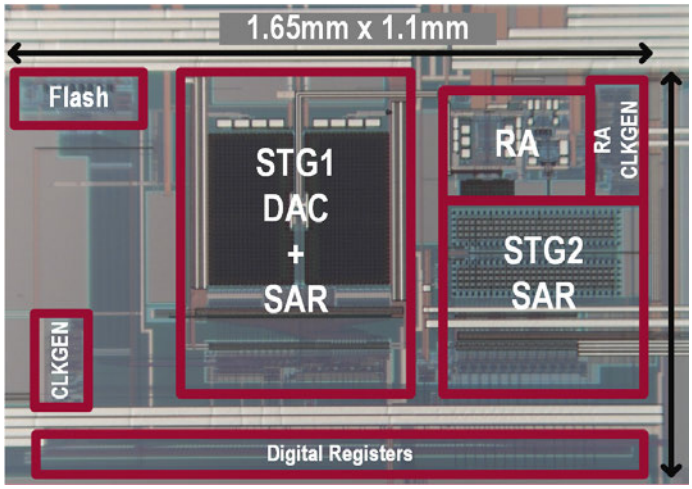


Figure 3.7.7: Die photo.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 4 Overview: *Power Amplifiers*

RF SUBCOMMITTEE

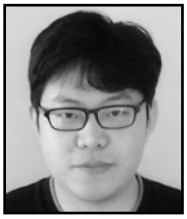
Session Chair: ***Krzysztof Dufrêne***
Intel, Linz, Austria

Associate Chair: ***Harish Krishnaswamy***
Columbia University, New York, NY

Subcommittee Chair: Piet Wambacq, imec, Leuven, Belgium

Simultaneously achieving high output power, high peak efficiency, and high efficiency under spectrally efficient modulation schemes has become an important topic of research as the industry prepares for the widespread deployment of 5G technologies. Doherty PAs are being extensively investigated at millimeter-wave frequencies to improve efficiency under back-off. There is a continuing trend at millimeter-waves to blend electronics with antenna design to achieve improved performance. Innovations in digital power amplifiers and broadband power amplifiers are also being seen for sub-6GHz RF applications as well as broadband wireline applications.

1:30 PM



4.1 A Watt-Level Phase-Interleaved Multi-Subharmonic Switching Digital Power Amplifier Achieving 31.4% Average Drain Efficiency

A. Zhang, University of Southern California, Los Angeles, CA

In Paper 4.1, the University of Southern California presents a phase-interleaved, multi-subharmonic switching digital power amplifier, providing inherent cancellation of the subharmonic components in the power back-off region. The prototype achieves a 30dBm peak power at 1.9GHz with 45.9/41.3/35.3/32.2/24.2% drain efficiency at 0/-3.5/-7/-9.5/-12 dB PBO.

2:00 PM

4.2 A Broadband Switched-Transformer Digital Power Amplifier for Deep Back-Off Efficiency Enhancement

L. Xiong, Fudan University, Shanghai, China

In Paper 4.2, Fudan University presents a broadband inverter-based class-D power amplifier with a switched-transformer technique to enhance deep back-off efficiency. Implemented in 40nm CMOS and powered from a 1.1V supply, the PA achieves 4 efficiency peaks in 1.3-to-3.5GHz range. It obtains 21.4dBm peak output power and 31.3/27.7/16.6/7.7% PAE at 0/6/12/18dB back-off at 1.5GHz.

2:15 PM

4.3 A Multiphase Interpolating Digital Power Amplifier for TX Beamforming in 65nm CMOS

Z. Bai, University of Utah, Salt Lake City, UT

In Paper 4.3, the University of Utah introduces a multiphase interpolating digital power amplifier for TX beamforming, leveraging switched capacitor PA techniques to simultaneously realize high linearity and power and system efficiency. Four transmitters are able to achieve peak P_{out} of 24.4dBm, while achieving <1deg phase resolution and <0.15dB gain error.

2:30 PM

4.4 A Highly Linear High-Power 802.11ac/ax WLAN SiGe HBT Power Amplifier Using a Compact 2nd-Harmonic-Shorting Four-Way Transformer and Integrated Thermal Sensors

I. Ju, Georgia Institute of Technology, Atlanta, GA

In Paper 4.4, the Georgia Institute of Technology and Skyworks Solutions propose a highly linear, high-power SiGe HBT power amplifier supporting 802.11ac/ax. A compact four-way output transformer with a 2nd-harmonic short is designed to obtain efficient power combining and linear operation without bulky harmonic rejection filters.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

2:45 PM

4.5 A 13.5dBm Fully Integrated 200-to-255GHz Power Amplifier with a 4-Way Power Combiner in SiGe:C BiCMOS

M. H. Eissa, IHP, Frankfurt (Oder), Germany

In Paper 4.5, IHP and Technische Universität Berlin present a fully integrated power amplifier at 240GHz. Four PA units are combined with a zero-degree power combiner, achieving P_{sat} of 13.5dBm and 3% drain efficiency. A small-signal gain of 15.5dB is measured across a 3dB 55GHz bandwidth between 200 and 255GHz.

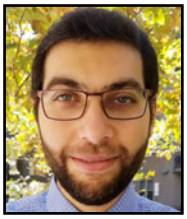
3:15 PM

4.6 A mm-Wave 3-Way Linear Doherty Radiator with Multi-Antenna Coupling and On-Antenna Current-Scaling Series Combiner for Deep Power Back-Off Efficiency Enhancement

H. T. Nguyen, Georgia Institute of Technology, Atlanta, GA

In Paper 4.6, the Georgia Institute of Technology presents a 3-way series-combining Doherty radiator exploiting dual-antenna coupling and on-antenna current-scaling series combiner for power back-off efficiency enhancement up to 9.6dB with high linearity. It achieves 21.2dBm P_{sat} and 21.8/19.5/18% PAE at 0/6/9dB PBO.

3:45 PM



4.7 A Compact DC-to-108GHz Stacked-SOI Distributed PA/Driver Using Multi-Drive Inter-Stack Coupling, Achieving 1.525THz GBW, 20.8dBm Peak P_{1dB}, and Over 100Gb/s in 64-QAM and PAM-4 Modulation

O. El-Aassar, University of California, San Diego, La Jolla, CA

In Paper 4.7, the University of California, San Diego, describes a compact stacked distributed power amplifier with multi-drive and inter-stack coupling techniques. The amplifier achieves a record 1.525THz gain-bandwidth product and 20.8dBm peak output P_{1dB}. By exploiting complementary nature of NMOS-PMOS to minimize AM-AM and AM-PM distortions, it achieves over 100Gb/s in both 64-QAM and PAM-4 modulations.

4:15 PM

4.8 A Highly Linear Super-Resolution Mixed-Signal Doherty Power Amplifier for High-Efficiency mm-Wave 5G Multi-Gb/s Communications

F. Wang, Georgia Institute of Technology, Atlanta, GA

In Paper 4.8, the Georgia Institute of Technology proposes a mixed-signal Doherty power-amplifier architecture. The power amplifier achieves 40.1% peak PAE with 23.3dBm P_{SAT} and 37.5/33.1% PAE at 5/6dB power back-off. It also achieves -24.5dB rms EVM with an average P_{out}/PAE of 15.6dBm/27.8% for a single-carrier 12Gb/s 64-QAM signal.

4:45 PM

4.9 A 60GHz CMOS Power Amplifier with Cascaded Asymmetric Distributed-Active-Transformer Achieving Watt-Level Peak Output Power with 20.8% PAE and Supporting 2Gsym/s 64-QAM Modulation

H. T. Nguyen, Georgia Institute of Technology, Atlanta, GA

In Paper 4.9, the Georgia Institute of Technology presents a cascaded asymmetric distributed active-transformer power amplifier combining 48 differential output ports of 24 PA units to deliver 29.1-to-30.1dBm peak output power with 18.5-to-20.8% peak PAE at 60GHz. The amplifier supports linear amplification of wideband modulations, demonstrating 2Gsym/s 16-QAM and 64-QAM signals at an average P_{out} of 22.3dBm and 20.9dBm, respectively.

4.1 A Watt-Level Phase-Interleaved Multi-Subharmonic Switching Digital Power Amplifier Achieving 31.4% Average Drain Efficiency

Aoyang Zhang, Mike Shuo-Wei Chen

University of Southern California, Los Angeles, CA

Modern wireless communication systems often utilize spectrum-efficient modulation schemes for higher data throughput, given the finite bandwidth. This type of modulation schemes, such as Orthogonal Frequency Division Multiplexing (OFDM), results in a high peak-to-average power ratio (PAPR) for the transmitted signal. Therefore, power amplifier efficiency in the power back-off (PBO) region has become an important design target. Meanwhile, obtaining high output power and high average efficiency still remains a key design challenge when developing an integrated CMOS PA. Recently, a subharmonic switching (SHS) digital PA architecture was reported in [1]. It toggles the PA cell at the subharmonic component of the carrier frequency (Fc) to achieve power back-off. The slower toggling rate reduces dynamic and conduction loss in the switching PA, resulting in better PBO efficiency. However, the SHS PA requires additional notch filtering of the subharmonic components in the matching network. Therefore, we propose a phase-interleaved architecture that combines three SHS PAs to increase output power (Watt-level) and inherently cancel the subharmonic components in the PBO mode, thereby alleviating the burden of the matching network. Moreover, multiple subharmonic components are utilized to create a greater number of efficiency peaks in the PBO region. This is referred to as a multi-SHS scheme. Lastly, a hybrid Class-G operation, in combination with the multi-SHS scheme, is used to further enhance average efficiency.

In the proof-of-concept prototype, a switched-capacitor (SC) PA [2] is employed as the baseline structure for digital-to-analog conversion. As shown in Fig. 4.1.1, the conventional SC PA utilizes the capacitor network to perform voltage division at the PA output, leading to power back-off. However, due to charge redistribution, the charge stored on the capacitor is lost, degrading PBO efficiency. The key idea of the SHS PA is to toggle the PA cells at a certain subharmonic frequency when the PA operates in the PBO region, e.g., Fc/3 (Fig. 4.1.1), where Fc is the carrier frequency, and to use the third harmonic component to carry the signal information, leading to PBO of -9.5dB in this particular case. Since all PA cells toggle together, dynamic charge loss in the switched capacitor bank is avoided. However, to remove the subharmonic components at the output spectrum, a notch response must be added to the matching network at subharmonic frequencies, thus incurring overhead.

The key concept of the proposed phase-interleaved architecture is to deploy an array of SHS PAs and balance the phases of their digital input waveforms in the PBO region, i.e. 120 degrees apart for the three interleaved PA channels. Accordingly, only the Fc component is aligned in phase, while the phases of the unwanted Fc/3 and 2Fc/3 components are inherently cancelled on the primary side of the transformer without additional loss (Fig. 4.1.1). Realistically, the effectiveness of the nullification depends on how well the phase and magnitude between the PA channels could be matched. This prototype is demonstrated to achieve more than 50dB subharmonic attenuation, given the extra design efforts in balancing it. It is noteworthy that, in the peak-output mode, the three PAs continue to operate at the carrier frequency (Fc), using the same input phases for power combining. In other words, phase-interleaved input is only applied when specific PA cells operate in subharmonic switching mode.

The key idea of multi-SHS is to use multiple subharmonic components for creating a greater number of efficiency peaks at PBO, in this case both Fc/3 and 2Fc/3 subharmonic components are used. Additionally, dual-power supplies (3VDD and 2VDD) are used in conjunction with the multi-SHS operation to create a total of five efficiency peaks located at various PBO points, i.e. 0dB, -3.5dB, -7dB, -9.5dB, and -13dB (Fig. 4.1.2). It is worth mentioning that phase balancing for the 2Fc/3 peak results in non-uniform toggling waveforms for each PA channel. By gating off every third cycle, the phases are balanced and hence the unwanted subharmonic components are rejected, which also leads to additional efficiency peaking at -7dB PBO. For the actual implementation, efficiency between peaks is further optimized by combining the different subharmonics and power supplies simultaneously in the individual PA cells.

When devising guidelines for the design, it is important to ensure that 1) the same number of PA cells are within each PA channel toggle for the same subharmonic component (either Fc/3 or 2Fc/3) in order to maintain equivalent transformer input impedance seen by each PA channel and 2) the input phases for the corresponding PA cell in each channel should always be balanced while operating in subharmonic mode. In that way, the PBO efficiency can be maximized while unwanted frequency components can be rejected.

The block diagram for the overall implementation of the PA is shown in Fig. 4.1.3. An external phase-modulated (PM) signal is used to generate phase-interleaved subharmonic components (Fc/3 and 2Fc/3), as well as the delayed PM signal via on-chip circuitry, including frequency dividers and delay lines. A high-speed multiplexer selects the correct PM signal for each PA cell, based on the corresponding amplitude code. The synthesized decoder converts the eight-bit amplitude code into a set that comprises amplitude, multi-SHS, and dual-power supply control words for each PA channel, based on the aforementioned switching rules for PBO efficiency enhancement. The output power of each PA channel can be tuned by making changes to the non-overlapping clock phase. A triple-stacked Class-D output driver is designed to enlarge the output voltage swing and lower the resultant impedance transformation ratio. Based on the hybrid operation, the pull-up path of the Class-D output driver selects either 3VDD (3.6V) or 2VDD (2.4V). This helps to reduce the power loss in the matching network and yield less efficiency degradation due to routing resistance.

A two-turn-transformer-based power combiner was designed to combine the power of the PA with subharmonic cancellation. A symmetric three-way power combiner layout was designed to balance the coupling factor for each PA channel, hence maintaining effective subharmonic cancellation. Parasitic coupling between the primary and secondly winding of the power combiner is minimized to reduce the ratio of direct subharmonic coupling to the PA output as it would have undermined the efficacy of subharmonic cancellation otherwise. The layout strategy used for the power combiner and the associated HFSS EM simulation results for insertion loss are shown in Fig. 4.1.3.

The proof-of-concept prototype is implemented in 65 nm CMOS as shown in Fig. 4.1.4. The PA delivers 30dBm peak output power with 45.9% drain efficiency at 1.9GHz and 28.7dBm peak output power with 42.4% drain efficiency at 2.2GHz. The proposed wideband three-way power combiner achieves a 3dB bandwidth of >500MHz. Figure 4.1.4 shows PA efficiency against different PBO levels, at both 1.9GHz and 2.2GHz. Five efficiency peaks are achieved in the PBO region after combining multi-SHS and hybrid Class-G operation. The last efficiency peak (24.2% drain efficiency) occurs at a -12dB PBO level (1dB PBO level offset due to peak power saturation), demonstrating 2× improvement compared with a normalized Class-B PA, which can benefit high PAPR modulation signal. Note that the multi-SHS scheme also helps to improve PBO efficiency between the five peaks. The AM-AM and AM-PM responses are measured for digital pre-distortion (DPD). The dynamic performance of the PA is evaluated using 5MHz 16-QAM OFDM signal. It achieves 22.8dBm average power with error vector magnitude (EVM) of -24.7dB and an adjacent channel power ratio (ACPR) of 35dBc. The average drain efficiency achieves 31.2% under real-time DPD operation using a 7.2dB PAPR modulated signal. The performance of our PA prototype is compared with that of prior-art CMOS PAs [3-8] (Fig. 4.1.6). The proposed architecture achieves superior PBO efficiency and compares favorably with the published Watt-level RF CMOS PAs in Fig. 4.1.6. Figure 4.1.7 shows a micrograph of the test chip.

Acknowledgements:
The authors would like to acknowledge partial support from NSF ECCS-SpecEES program.

References:
[1] A. Zhang and M. S.-W. Chen, "A Sub-harmonic Switching Power Amplifier with Hybrid Class-G Operation for Enhancing Power Back-off Efficiency," *IEEE Symp. VLSI Circuits*, pp. 213-214, June 2018.
[2] S.-M. Yoo, et al., "A Switched-Capacitor RF Power Amplifier," *IEEE JSSC*, vol. 46, no. 12, pp. 2977–2987, Dec. 2011.
[3] D. Cousinard et al., "A 0.23mm^2 Digital Power Amplifier with Hybrid Time/Amplitude Control Achieving 22.5dBm at 28% PAE for 802.11g," *ISSCC*, pp. 228–229, Feb. 2017.
[4] V. Vorapipat et al., "A Class-G Voltage-Mode Doherty Power Amplifier," *ISSCC*, pp. 46–47, Feb. 2017.
[5] S. Hu et al., "A Broadband CMOS Digital Power Amplifier with Hybrid Class-G Doherty Efficiency Enhancement," *ISSCC*, pp. 44–45, Feb 2015.
[6] P. Godoy et al., "A 2.4-GHz, 27-dBm Asymmetric Multilevel Outphasing Power Amplifier in 65-nm CMOS," *IEEE JSSC*, vol. 47, no. 10, pp. 2372–2384, Oct. 2012.
[7] K. Onizuka et al., "A 1.8GHz Linear CMOS Power Amplifier with Supply-Path Switching Scheme for WCDMA/LTE Applications," *ISSCC*, pp. 90–91, Feb. 2013.
[8] W. Tai et al. "A Transformer-Combined 31.5 dBm Outphasing Power Amplifier in 45 nm LP CMOS With Dynamic Power Control for Back-Off Power Efficiency Enhancement," *IEEE JSSC*, vol.47, no.7, pp.1646-1658, July 2012.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

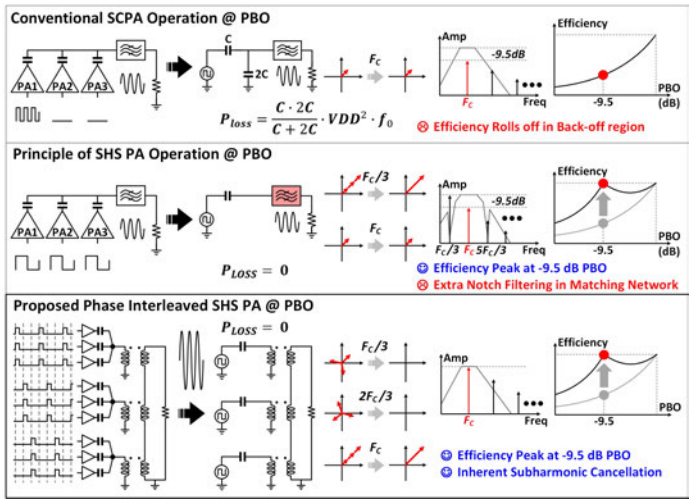


Figure 4.1.1: Proposed phase-interleaved architecture for the SHS PA in a PBO region.

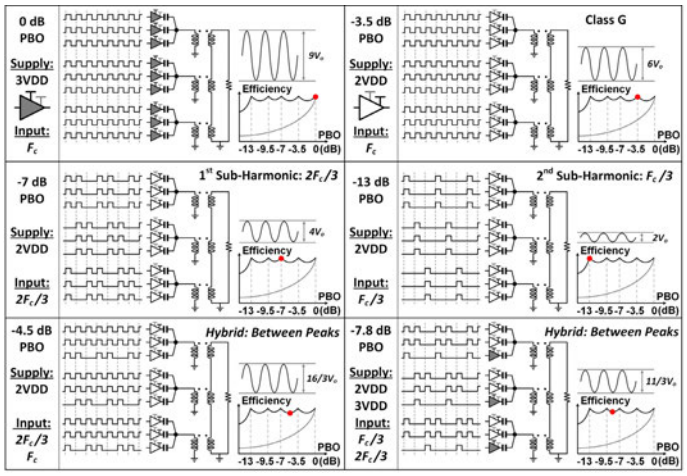


Figure 4.1.2: Hybrid operation modes for combining multi-SHS and dual power supplies.

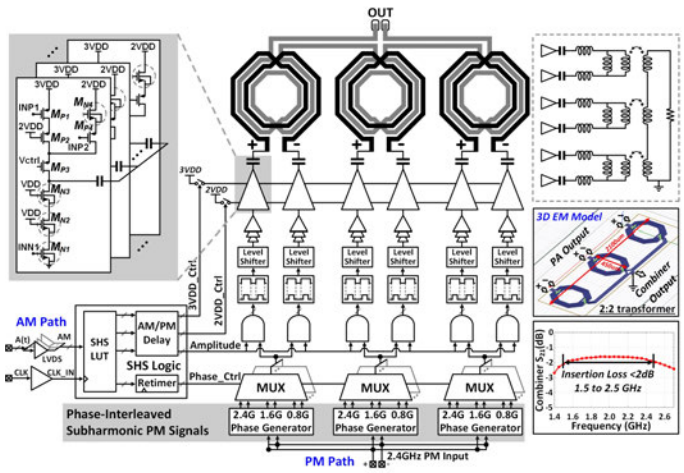


Figure 4.1.3: Simplified block diagram of the PA prototype. Here, $3V_{DD}$=3.6V and $2V_{DD}$=2.4V.

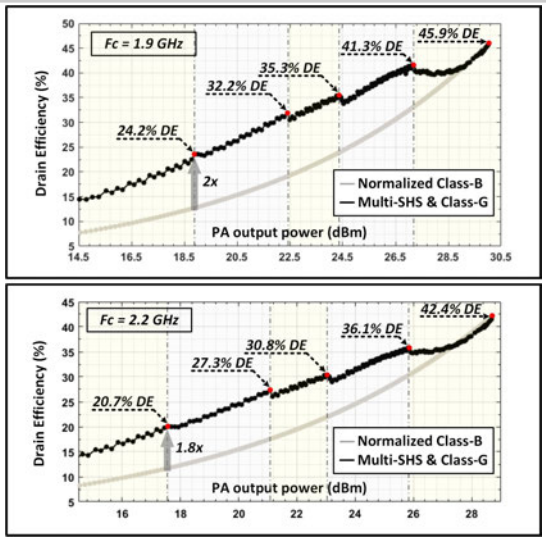


Figure 4.1.4: Measured drain efficiency using a CW signal at different PBO levels at 1.9GHz and 2.2GHz.

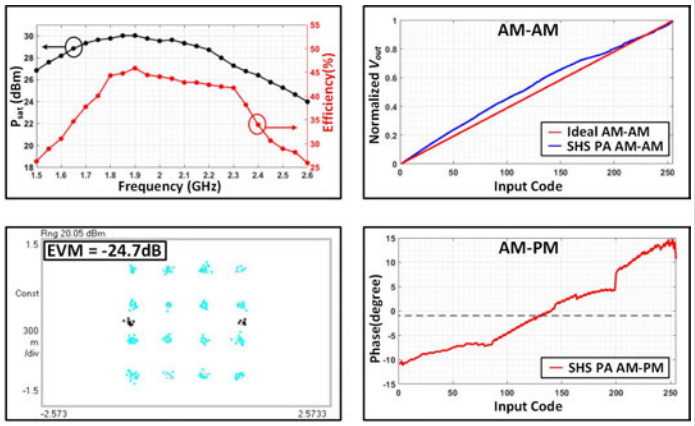


Figure 4.1.5: Measured EVM using 16-QAM OFDM signal with 7.2dB PAPR, AM-AM, AM-PM and peak static performance.

Specifications		This work		[3] ISSCC 2017	[4] ISSCC 2017	[5] ISSCC 2015	[6] ISSCC 2013	[7] JSSC 2012	[8] JSSC 2012
Process	[nm]	65		28	45	65	65	65	45
Frequency	[GHz]	1.9	2.2	2.45	3.5	3.71	1.8	2.4	2.4
Max P_{out}	[dBm]	30	28.7	28.2	25.3	26.7	27.2	27.7	31.5
Peak DE	[%]	45.9	42.4	39	30.4‡	40.2	30	45.1‡	27‡
-3.5 dB DE	[%]	41.3	36.1	33*	26‡, *	32.5*	24‡, *	39‡, *	22‡, *
-7 dB DE	[%]	35.3	30.8	23*	23‡, *	35*	19‡, *	37‡, *	17‡, *
-9.5 dB DE	[%]	32.2	27.3	18*	19‡, *	31.5*	14‡, *	24‡, *	14‡, *
-12 dB DE	[%]	24.2	20.7	14*	17.4‡	26.2	9‡, *	22‡, *	12‡, *
PAPR	dB	7.2		5.7	6.3	5.8	5.9	7.5	6.7
Average DE	[%]	31.4†		30	24‡	28.8	18‡	27.6‡	16‡
EVM	[dB]	-24.7†		-25	-35	-24	-22	-31.4	-25
Matching Network		On-Chip		On-Chip	On-Chip	On-Chip	On-Chip	Off-Chip	On-Chip
PBO Efficiency Enhancement Tech		Multi-SHS /Class-G		Polar/PWM	Doherty/ Class-G	Doherty/ Class-G	Supply Switching	AMO	Current Switching

*Estimated from reported figures
‡Reported Power Added Efficiency (PAE)
†Measured at center frequency 1.9 GHz

Figure 4.1.6: Comparison with prior-art CMOS PAs with PBO efficiency enhancement.

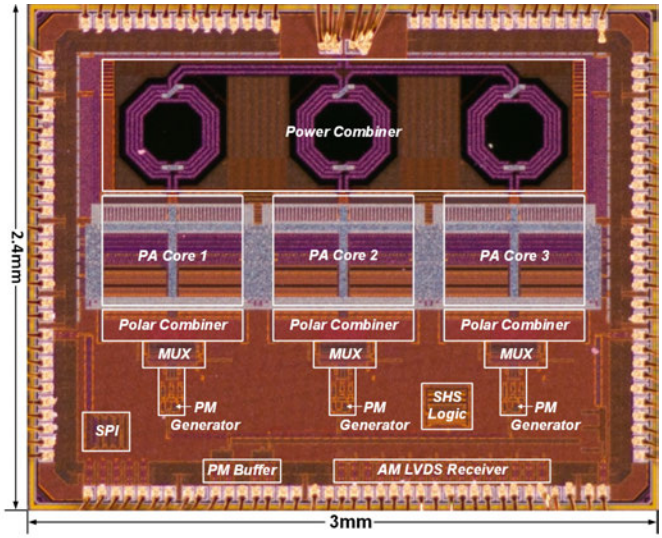


Figure 4.1.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

4.2 A Broadband Switched-Transformer Digital Power Amplifier for Deep Back-Off Efficiency Enhancement

Liang Xiong, Tong Li, Yun Yin, Hao Min, Na Yan, Hongtao Xu

Fudan University, Shanghai, China

Sophisticated OFDM modulation schemes with high spectrum efficiency and data throughput in modern wireless communication systems often result in a large peak-to-average power ratio (PAPR). Besides, wireless standards like LTE, WLAN, NB-IoT, etc., require wide transmission power range to accommodate various communication environments, and devices often function at low average output power. For better battery lifetime, it is critical to improve the power amplifier (PA) efficiency at deep power back-off (PBO) levels (e.g., 12/18dB or higher). Recently, several digital-style techniques have been employed to enhance PA PBO efficiency, such as dynamic power control [1], Class-G, and Doherty [2-4], as well as multi-level outphasing [5]. Class-G or Doherty techniques usually provide an efficiency peaking at 6dB PBO, and when combined together [2,3] or cascaded [6] they can further enhance the efficiency beyond 6dB PBO by introducing two efficiency peaks at 6/12dB PBOs. However, most of the Class-G Doherty PAs suffer from large area overhead with two power supply paths and glitches due to mode transitions. The dynamic power control or multi-level outphasing PA requires multiple phase modulators and amplitude-level transitions, which cause inherent discontinuities and degrade the linearity. In this work, a switched-transformer digital-PA technique is proposed for wide-range PBO efficiency enhancement. This topology does not require multiple power supplies and does not introduce AM/PM discontinuities. The PA achieves multiple efficiency peaks at 0/6/12/18dB PBOs and wide frequency coverage with a single-transformer footprint and only one supply voltage.

Figure 4.2.1 illustrates the operating principles of proposed switched-transformer PA. PA1/PA2/PA3/PA4 are four differential sub-PA pairs that are digitally controlled independently, and the transformer can be reconstructed to perform the function of power combining at different PBO levels. At 0dB PBO, all sub-PAs are fully switched on and the 5-coil transformer combiner performs an 8-way parallel power combining, where the four primary coils are driven by four differential sub-PA pairs to collect their output currents and provide the optimum single-ended load impedance Ropt. Besides, the transformer combiner is configured as shown in Fig. 4.2.1 to maintain all sub-PA pairs quasi-differential so as to reduce differential mismatches. With output power decreased, the unit cells of PA3 and PA4 are gradually switched off. At 6dB PBO, PA3/PA4 are fully turned off and the transformer combiner is reconfigured as a 4-way power combiner with SW3/SW4 switched on, where the single-ended impedance seen by PA1/PA2 reaches 2×Ropt and an efficiency peaking is achieved due to load modulation. With output power reduced continuously, PA2 is then switched off and the single-ended impedance seen by PA1 continues to increase until it reaches 4×Ropt, thus leading to another efficiency peaking at 12dB PBO. Furthermore, when PA1 goes from differential to single-ended, the single-ended impedance seen by PA1+ increases to 8×Ropt and an efficiency peaking appears at 18dB PBO. Therefore, deep back-off efficiency enhancement at 6/12/18dB PBOs are achieved with the proposed switched-transformer power-combing technique.

Figure 4.2.2 shows the detailed schematic of proposed switched-transformer digital PA, where the switched-capacitor PA topology is adopted for fine amplitude control due to its good linearity and efficiency [3]. In the design, the total resolution of 9b is employed and the 2 MSBs control four 7b sub-PAs. In each sub-PA, the hybrid unary-and-binary array is utilized to achieve a better resolution and reduce layout mismatch. Each sub-PA is divided into 4 groups controlled by AM1/2/3/4<6:5> and each group consists of 7 thermometer-coded cells and 2b LSB binary-coded cells. Here, PA2/PA3/PA4 are three identical sub-PAs while PA1 is slightly different from the others since it has the single-ended operation mode. A simple inverter-based Class-D topology is employed in the differential unit PA cell, which has two operation states in all the sub-PAs. During 0-to-12dB PBO, the unit cells of PA3/PA4 and PA2 are gradually transformed from the on-state to the high-impedance state in sequence, and the transformer combiner is reconstructed through the switches that are formed by three NMOS transistors between the differential outputs. Within 12-to-18dB PBO, the differential branches of PA1 are controlled independently. At 18dB PBO, the positive half of PA1 is fully at the on-state while the negative half is completely at the off-state with inverter outputs grounded, thus enhancing the deep back-off efficiency by reconstructing the transformer combiner and performing the load modulation.

Figure 4.2.3 presents the implementation of proposed switched-transformer-based matching network. The 8-way transformer combiner is realized within a compact single-transformer footprint. The 4 sub-PAs are located at the four corners of the transformer to achieve good differential matching and diagonal symmetry. The output currents from the 4 sub-PAs flow in parallel direction in the four primary coils, which achieves magnetic enhancement and increases the effective primary inductance, thus contributing to size reduction and lower loss. The passive matching network consists of sub-PA device parasitic capacitors, switched capacitors, a digitally reconstructive transformer, and 2 capacitors C_1, which transform the optimum load impedance for the 4 sub-PAs. It achieves broadband frequency coverage over 1.3 to 3.5GHz with less than 1dB output power variation. The total passive loss from the 4 sub-PAs to 50Ω load is less than 1.5dB in a wide frequency range. The balun function is also implemented on-chip with a single-ended output pad. In the design, a total of 300pF de-Q'ed decoupling capacitors are integrated to attenuate supply and ground fluctuations while further improving the linearity performance.

The proposed deep back-off efficiency-enhanced DPA has been fabricated in a 40nm CMOS process (Fig. 4.2.7), and occupies an area of 1.15×0.7mm^2 including all decoupling capacitors and I/O pads. The chip is powered by a single 1.1V supply, and the power consumption of all IO buffers, logic blocks, and PAs is included in the PAE calculation. Figure 4.2.4 shows the measured continuous-wave (CW) results, where the DPA achieves 21.4dBm output power with 31.3% peak PAE at 1.5GHz and 20.4dBm output power with 24.5% peak PAE at 3.5GHz. The 1dB bandwidth is over 90% spanning from 1.3 to 3.5GHz while maintaining good PAE performance. Thanks to the switched-transformer power-combining and load-modulation techniques, the expected back-off efficiency enhancement is realized, where the DPA achieves the PAE of 31.3%, 27.7%, 16.6%, and 7.7% for 0dB, 6dB, 12dB, and 18dB PBOs at 1.5GHz, respectively. Besides, our proposed DPA demonstrates the significant PAE improvements of 1.77×, 2.12×, and 1.97× at 6dB, 12dB, and 18dB PBOs, respectively, compared to the traditional Class-B PA.

In modulation tests, 1-D AM-AM and AM-PM memoryless look-up tables (LUTs) are used to linearize the DPA. For the 20MHz 64-QAM LTE signal with 6.2dB PAPR, the measured spectrum and constellation are shown in Fig. 4.2.5. With -32.5dB EVM, the DPA achieves the P_{avg} of 15.2dBm with average PAE of 25.3%, and the measured upper/lower ACLRs are -30.4dBc and -33.0dBc, respectively. Figure 4.2.6 summarizes the performance and makes a comparison with prior deep back-off efficiency-enhanced DPAs. By utilizing the switched-transformer power-combining technique, the proposed DPA achieves the effective deep back-off efficiency enhancement and broadband frequency coverage with the smallest die size, thus suitable for multiple communication standards especially with a wide dynamic power range.

Acknowledgements:
The authors would like to thank the State Key Laboratory of ASIC and System at Fudan University for measurement support.

References:
[1] W. Tai et al., "A Transformer-Combined 31.5 dBm Outphasing Power Amplifier in 45 nm LP CMOS with Dynamic Power Control for Back-off Power Efficiency Enhancement," *IEEE JSSC*, vol. 47, no. 7, pp. 1646-1658, July 2012.
[2] S. Hu et al., "A Broadband CMOS Digital Power Amplifier with Hybrid Class-G Doherty Efficiency Enhancement," *ISSCC*, pp. 44-45, Feb. 2015.
[3] V. Vorapipat et al., "A Class-G Voltage-Mode Doherty Power Amplifier," *ISSCC*, pp. 46-48, Feb. 2017.
[4] Y. Yin et al., "A Compact Dual-Band Digital Doherty Power Amplifier Using Parallel-Combining Transformer for Cellular NB-IoT Applications," *ISSCC*, pp. 408-410, Feb. 2018.
[5] M. Martelius et al., "A 30-dBm Class-D Power Amplifier with On/Off Logic for an Integrated Tri-Phasing Transmitter in 28-nm CMOS," *IEEE RFIC*, pp. 136-139, June 2018.
[6] P. Seddighrad et al., "Digitally-Scalable Transformer Combining Power Amplifier," US Patent 8,502,598, Aug. 2013.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

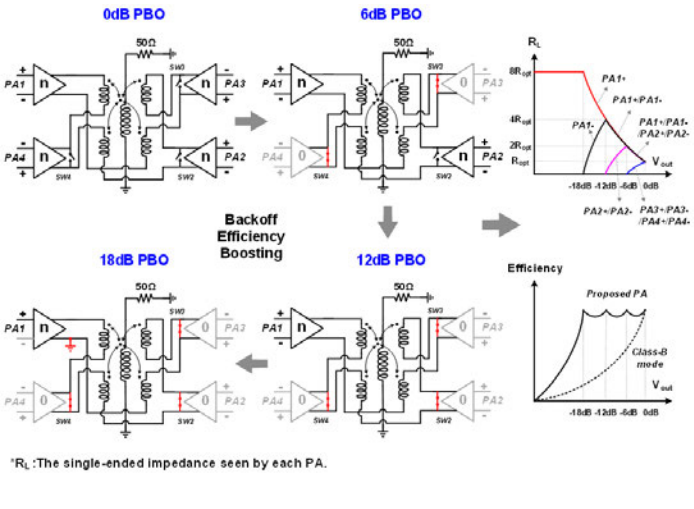


Figure 4.2.1: Operating principles of the proposed switched-transformer digital PA.

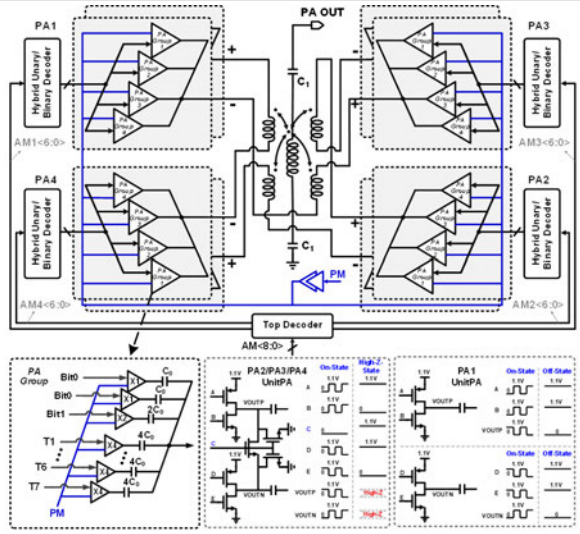


Figure 4.2.2: Detailed schematic of the proposed switched-transformer digital PA.

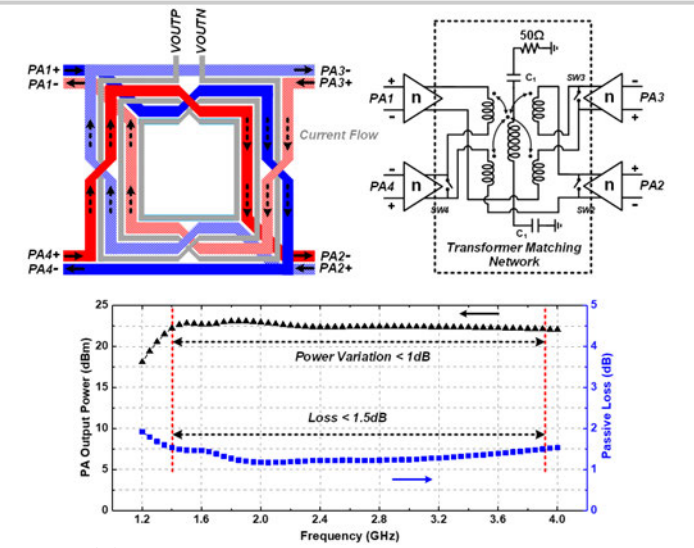


Figure 4.2.3: Implementation of the proposed switched-transformer-based matching network and its simulation results.

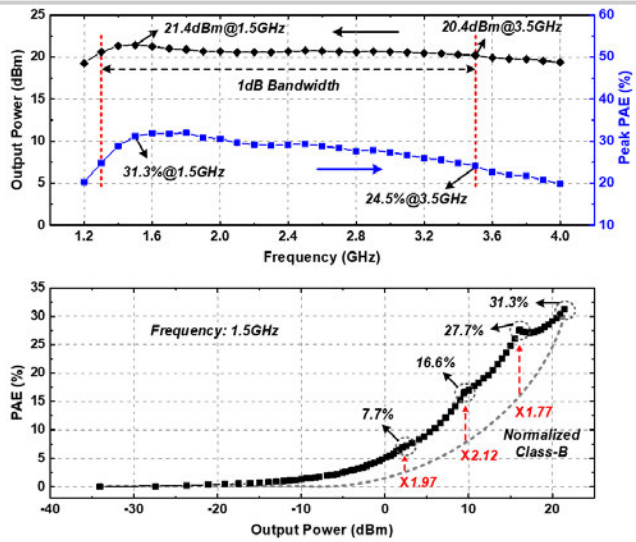


Figure 4.2.4: Measured CW results of output power, PAE (PAE=Pout/Pdc_total), and the frequency response.

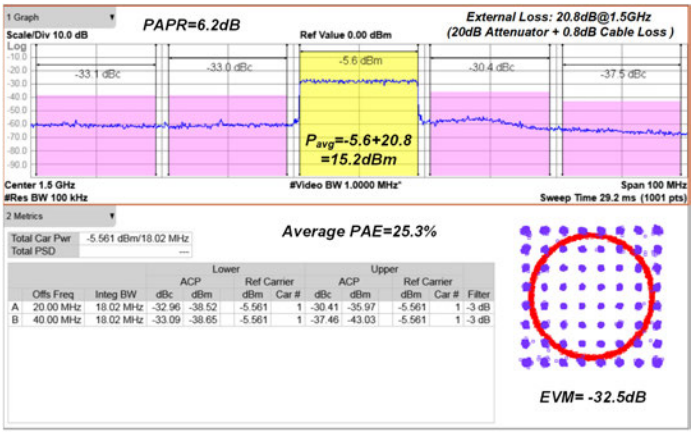


Figure 4.2.5: Measured DPA spectrum and constellation of 20MHz 64-QAM LTE signal.

	This work	ISSCC 2015 [2]*	ISSCC 2017 [3]*	RFIC 2018 [5]
Architecture	Switched transformer-based polar	Class-G+Doherty polar	Class-G+Doherty polar	Multi-level outphasing
Matching Newtork	On-chip (1 transformer)	On-chip (2 transformers)	On-chip (2 transformers)	Off-chip
PAE Improvement @Deep PBOs	6/12/18dB	6/12dB	6/12dB	2.5/6/12dB
Frequency (GHz)	1.5	3.71	3.5	1.7
1dB RF Bandwidth (GHz) (Frequency Coverage)	1.3-3.5 (91.7%)	3.1-4.7 (41.0%)	2.9-4.3 (38.9%)	NA
Peak Pout (dBm)	21.4	26.7	25.3	29.7
Peak PAE (%)	31.3	40.2 (DE)	30.4	34.7
PAE @6/12/18dB PBO (%)	27.7/16.6/7.7	37.0/26.2/NA (DE)	25.3/17.4/8†	20.5/11.5/3.5††
Modulation Signal	20MHz 64QAM LTE	1MS/s 16QAM	32 Carriers, 10MHz 256QAM	20MHz LTE
Average Pout (dBm)	15.2	20.8	19.0	23.1
Average PAE (%)	25.3	28.8 (DE)	24.0	15.3
EVM (dB)	-32.5	-24	-35.0	-21.9
Supply (V)	1.1	1.4/2.8	1.2/2.4	1.8/3.6
Chip Size (mm²)	0.8	3.2	1.2	1.31
Technology	40nm CMOS	65nm CMOS	45nm CMOS SOI	28nm CMOS

*Results measured with GSG probe. †Estimated from Fig. 6 in [3]. ††Estimated from Fig. 6 in [5].

Figure 4.2.6: Performance summary and comparison with prior works.

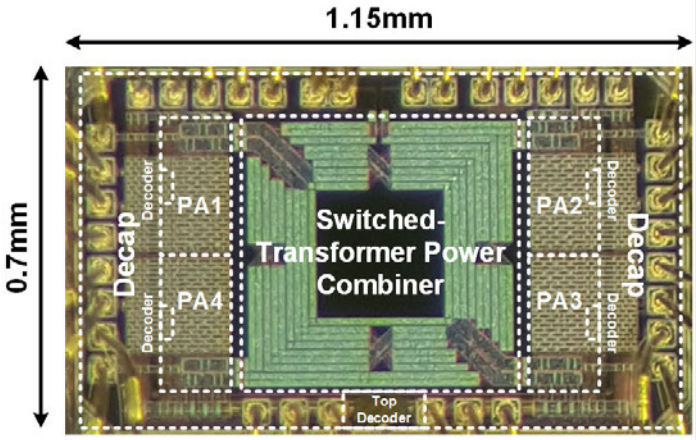


Figure 4.2.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

4.3 A Multiphase Interpolating Digital Power Amplifier for TX Beamforming in 65nm CMOS

Zhidong Bai, Wen Yuan, Ali Azam, Jeffrey Sean Walling

University of Utah, Salt Lake City, UT

Future 5G communications will heavily leverage beamforming and MIMO techniques, owing to the improvement in data transmission capacity. In a transmit (TX) beamformer, the phase and amplitude of an array of TXs can be arbitrarily adjusted to spatially steer a transmitted beam toward a user or multiple users. The amplitude and direction of the beam is adjusted by controlling the gain and phase of each TX in the array. Gain is typically adjusted with a variable-gain amplifier in the driver chain. In recent beamforming TXs, four primary means to control the output beam angle have been leveraged. Passive RF phase shifting is bulky and lossy, owing to high losses in the passive elements, and is difficult to design to provide phase control through 360°. LO phase shifting often uses a single passive phase shifter or a multiphase ring oscillator, but these devices typically provide lower phase resolution. Digitally controlled delay lines provide a wider bandwidth response but suffer from reduced phase resolution as frequency is increased and can consume high power [1]. Digital phase-shifting techniques have recently shown promise in polar systems [2,3], but the quadrature modulators, typically used for phase shifting, incur high loss if placed directly at the output stage.

We propose a multiphase controller to directly control a power amplifier (PA) allowing for high phase/gain resolution with reduced loss by performing a vector addition of the beam phase and amplitude. This has been used recently to provide amplitude and phase modulation [4]. In the proposed system (Fig. 4.3.1), an RF clock signal is input to an array of TXs to injection lock a multiphase ring oscillator (MRO) that outputs a set of basis phases (ϕ_1-ϕ_N) that can be weighted to provide amplitude/phase modulation data. Unlike [2], which only applies phase shifting, we combine both amplitude weighting and phase shifting in a single digital encoder block. A digital input code (B_{in}) drives a multiphase logic decoder (MLD) that selects the two adjacent phases to the desired output beam angle (ϕ_A, ϕ_B) and sends these two phases as the clock inputs to a multiphase switched-capacitor PA (MP-SCPA). With a control bit enabled, the amplitude- and phase-weighting for a desired beam output are calculated and stored. When the control bit is disabled, modulation is added to the stored beam output code.

The schematic for the proposed 4-channel beamforming TX is shown in Fig. 4.3.2. It is comprised of four identical MP-SCPAs. The SCPA is ideal for beamforming due to its high linearity, compact size, and the ability to output moderate-to-high output power [4,5]. Each MP-SCPA has a local MRO that is injection locked to a common global clock reference. The global clock reference is input to the chip via an LVDS clock RX. Care is taken to route the clock with equal delay to each ring oscillator to provide common time/phase basis. Beamforming and modulation are performed as follows. The local multiphase clock creates 16 equally spaced output phases that are input to a clock selection logic (CSL; e.g., a MUX tree). With *control*=1, the desired beam amplitude/phase is input to the chip via a two 9b control words; the CSL selects the two phases adjacent to the desired output phase; and this provides a coarse phase shift based on the two selected phases. The amplitude word controls the weighting of the basis phases so that after summation the beam phase and weighting are set. With the desired beam angle chosen, *control* is set to 0 and the beam angle is stored. The amplitude and phase modulation are then input as the 9b digital codewords. The original beam amplitude and phase are added to the modulation before being input to control the MP-SCPA. The decoder routes the two adjacent phases to the desired output (beam+modulation) to every cell in the MP-SCPA. The decoder output controls whether the MP-SCPA cell switches on ϕ_A, ϕ_B, or is held at ground. In this way, a vector addition on the capacitor array provides fine amplitude and phase modulation.

The MP-SCPA is ideal since the switched-capacitor array provides a linear output response. Polar and multiphase combinations on the array are native, by controlling the clock phase, and do not suffer the non-linearities that current-steering phase-shifting mechanisms can incur. Additionally, the multiphase technique does not suffer from 3dB output power reduction that quadrature combinations incur, allowing high efficiency combination at the output stage.

The 4-element TX array is fabricated in a 65nm CMOS technology. All decoders are designed in Verilog, synthesized, and automatically placed and routed. The combined area of all four TXs occupies 5mm^2, including pads and overhead circuitry (e.g., clock RX, clock distribution, pad drivers, and I/O routing), as shown in Fig. 4.3.7. The TX array is chip-on-board bonded to a PCB and an off-chip transformer balun converts the differential PA output signal to a single-ended output to drive an SMP jack. Individual TXs are characterized for their static and dynamic (modulated) characteristics.

The individual TXs show similar measured performance. The measurements include all losses including the off-chip balun and SMP connector. The output power and system efficiency (SE) are plotted in Fig. 4.3.3 and show peak P_{OUT}=24.4dBm and SE=24.2% at the center frequency of 1.75GHz. The output power characteristic varies linearly with code, and the efficiency characteristic shows a 6dB backoff SE≈14%. The modulation can be polar (injection clock is phase modulated) or multiphase (injection clock is not modulated). The individual SCPAs are tested with LTE signals with PAPR≈6.5dB in both a polar mode and multiphase mode, as shown in Fig. 4.3.4. The measured ACLR and EVM at the center frequency are <-30dBc and 3.27%-rms for polar mode and <-30dBc and 3.13%-rms for multiphase-mode without using digital pre-distortion (DPD). EVM and ACLR are maintained across the band without using DPD.

To verify the performance of the TXs as a function of the beamsteering angle, the phase error and output power vs. phase are measured for 4 TXs on a single die in Fig. 4.3.5. An off-line calibration is used to correct the static phase output and the phase error after calibration is ≈±1° across the 9b code range, with a phase error <0.32°-rms, across all four PAs. The output power varies by <1dB for outputs across the code range, with an output power error <0.15dB$_{rms}$, across all four PAs. Measured data from 4 PAs on a single die is used to synthesize a beam in Matlab for several pointing directions for a linear array with $\lambda/2$ spacing and compared to the ideal synthesized beam angle in Fig. 4.3.5. The synthesized beam from measured data matches well with the ideal synthesized beam across all output phases.

The performance of the beamforming DPA is compared with prior art in Fig. 4.3.6. The transmitter leverages the multiphase techniques presented in [4], as well as split-array techniques presented in [5]. Compared to [2], which is the most closely related work, the proposed work achieves similar phase and amplitude resolution, but at higher output power and without using DPD, while achieving better linearity (ACLR and EVM). Also, beam amplitude weighting is natively included in the proposed work. Due to the compact size of the SCPA, it also uses less chip area. Because of the relatively high output power and small area, this can be deployed in massive-MIMO systems that require a dense array with lower radiated power per element, without the requirement of an external PA.

References:

[1] L. Wang et al., "3-to-5GHz 4-Channel UWB Beamforming Transmitter with 1° Phase Resolution Through Calibrated Vernier Delay Line in 0.13µm CMOS," *ISSCC*, pp. 444–446, Feb. 2012.

[2] H. Qian et al., "A 3–7GHz 4-Element Digital Modulated Polar Phased-Array Transmitter with 0.35° Phase Resolution and 38.2% Peak System Efficiency," *IEEE CICC*, pp. 1–4, May 2017.

[3] B. Zheng et al., "A 0.19mm^2 128mW 0.8-1.2GHz 2-Beam 8-Element Digital Direct to RF Beamforming Transmitter in 40nm CMOS," *IEEE RFIC*, pp. 128–131, June 2018.

[4] W. Yuan and J. Walling, "A Multiphase Switched Capacitor Power Amplifier," *IEEE JSSC*, vol. 52, no. 5, pp. 1320–1330, May 2017.

[5] Z. Bai et al., "Split-Array, C-2C Switched-Capacitor Power Amplifiers," *IEEE JSSC*, vol. 53, no. 6, pp. 1666–1677, June 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

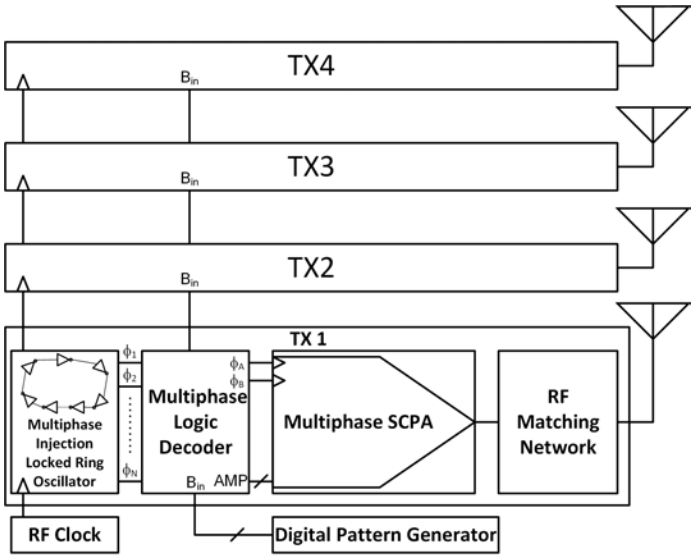


Figure 4.3.1: Multiphase beamforming TX concept.

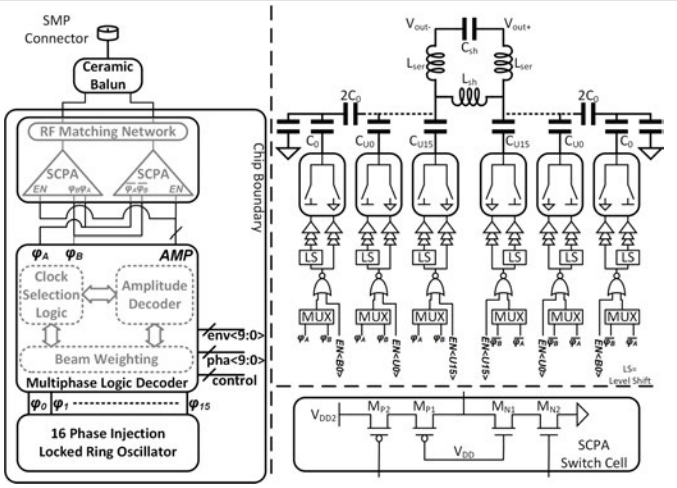


Figure 4.3.2: Schematic details of the 4-element beamforming transmitter. (left) Block-diagram schematic of the unit transmitter. (top right) Schematic of the pseudo-differential multiphase SCPA. (bottom right) Schematic of the cascoded unit switching cell.

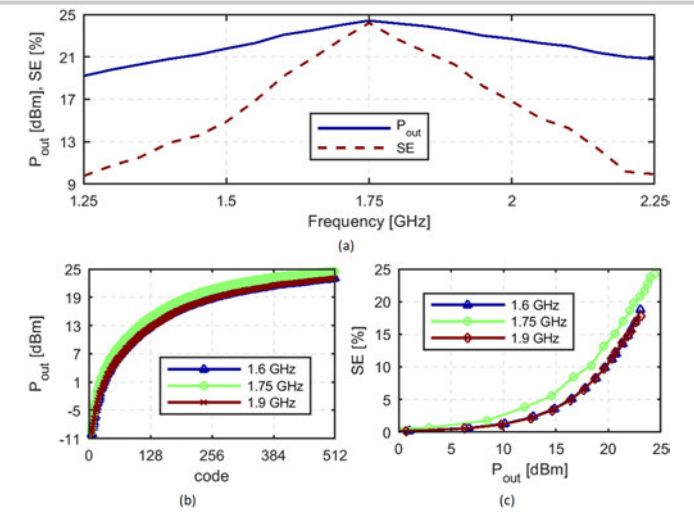


Figure 4.3.3: Single transmitter static measurements: (a) Measured output power and system efficiency (SE) vs. frequency, (b) output power vs amplitude code, and (c) SE vs. output power.

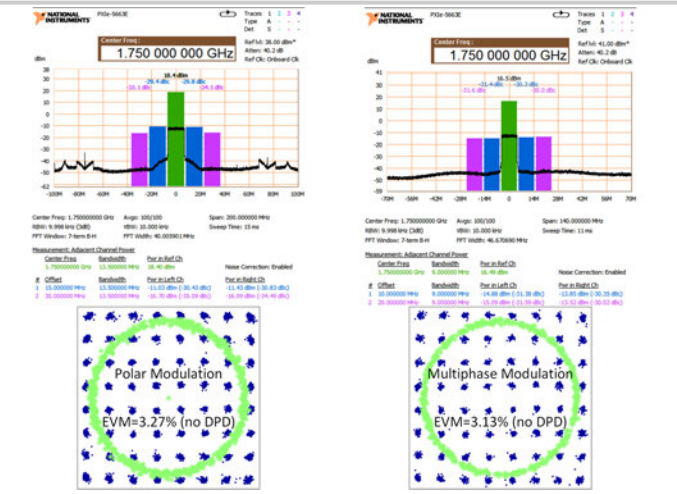


Figure 4.3.4: Measured ACLR (top left) and constellation (bottom left) for a 64-QAM, 15MHz OFDM signal (LTE) using polar modulation and ACLR (top right) and constellation (bottom right) for a 64-QAM, 10MHz OFDM signal (LTE) using multiphase modulation.

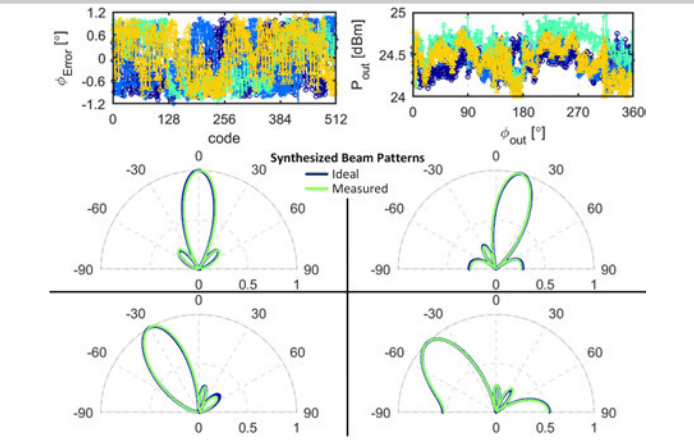


Figure 4.3.5: (top left) Measured beam steering phase error vs. steering code for four elements on same die (top right) Measured output power vs. beam steering phase for four elements on same die (bottom) Synthesized beam patterns using the measured output phase and power data for 4 elements from a single die compared to ideal synthesized beam phases (clockwise from top left: 0°, 15°, -30°, -45°).

	This Work	[2]	[3]	[1]	[4]
Beamforming Specifications					
Architecture	SCPA with multiphase phase-shifting	IDPA with Cartesian PM phase-shifting	DSM with Cartesian PM phase-shifting	UWB with Vernier delay line	MP-SCPA
Frequency (GHz)	1.45-2.15+	3-7+	0.8-1.2	3-5	1.4-2.2+
Phase Resolution (°)	0.7	0.35	11.25	1	N/A
RMS Phase Error (°)	0.32	0.2	0.8-1.2	N/A	N/A
RMS P_{out} Error (dB)	0.15	0.2	0.8-1.2	N/A	N/A
TX Specifications					
Peak P_{out} (dBm)	24.4#	21.8	N/A	N/A	26
Peak SE (%)	24.2#	38.2	N/A	7.5	24.9
Modulation	15 MHz, 64 QAM, OFDM	25 MHz 64-QAM	40 MHz 64-QAM	500 MHZ UWB	10 MHz, 64 QAM, OFDM
Average P_{out} (dBm)	18.4#	14.2	N/A	N/A	20.9
Average SE (%)	14#	22	NA	N/A	15.2
EVM (%-rms)	3.7	4.7	3.5/6.6	N/A	3.5
ACLR (dBc)	-30.5/-30.9	N/A	<-32	N/A	-30.3/-31.7
DPD	No	Yes	N/A	N/A	Yes
General Specifications					
CMOS Tech. (nm)	65	40	40	130	130
Supply Voltage (V)	1.4/2.8	1.1/1.2	1	1.2	1.5/3
Chip Area (mm^2)	5	8.6	0.19	7.2	3.8

+3-dB P_{sat} BW #Includes External Losses

Figure 4.3.6: Comparison to recent art.

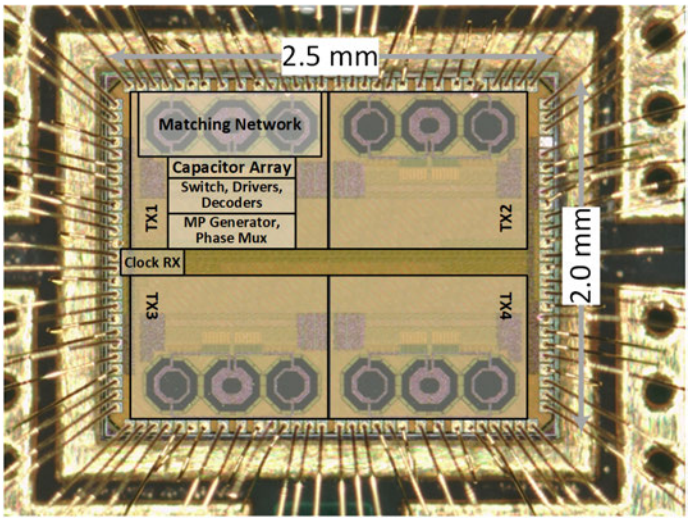


Figure 4.3.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

4.4 A Highly Linear High-Power 802.11ac/ax WLAN SiGe HBT Power Amplifier Using a Compact 2nd-Harmonic-Shorting Four-Way Transformer and Integrated Thermal Sensors

Inchan Ju[1], Mike McPartlin[2], Chun-Wen Paul Huang[2], Clifford DY Cheon[1], Mark Doherty[2], John D. Cressler[1]

[1]Georgia Institute of Technology, Atlanta, GA; [2]Skyworks Solutions, Andover, MA

Increasing demand on high-speed mobile access requires a spectrum-efficient wireless connectivity platform. With dense modulation and MIMO techniques, 802.11ac WLAN has been widely deployed as a viable solution. The emerging 802.11ax standard promises increased network capacity and higher data-rates. As 802.11ac/ax imposes stringent linearity requirements on power amplifiers (PAs), a high peak-to-average power ratio, operation at large power back-off, and reduced PAE are unavoidable. As a result, III-V processes [1] have dominated the development of high-output-power (P_{OUT}) WLAN PAs. To replace them with lower cost Si-based PAs, several linearization schemes [2,3] have been proposed, such as digital pre-distortion (DPD) and a spatial power combining, at the cost of increased complexity in the digital domain. In [4], a 2nd-harmonic short is used for linear Class-AB PAs, but an inductor on the supply lane occupies Si area and causes ohmic losses, degrading P_{OUT} and PAE. SiGe HBTs [5] are suitable for WLAN PAs due to their high power density, ruggedness, and CMOS compatibility. However, as current gain (β) drops with junction-temperature (T_J) rise, this distorts PA linearity and power gain.

To address these challenges, a highly linear, high-power 802.11ac/ax WLAN SiGe HBT PA is proposed. Both P_{OUT} and PAE are improved by using a compact 2nd-harmonic-shorted four-way output transformer. A dynamic bias circuit with an embedded thermal sensor is used to manage T_J of SiGe HBTs, preventing premature AM-AM compression.

Figure 4.4.1 shows the schematic of the 3-stage SiGe HBT PA. The emitter area ratio from one stage to the next remains about 4 for linear operation. Through-wafer vias (TWVs) are used as ground. A temperature-compensated dynamic bias circuit consists of a current source (I_{REF}) with a 3.3V CMOS control logic and a current mirror (Q_{1B}/Q_{2B}) with diodes (Q_{REF}/Q_{3B}). Only Q_{REF} ($S_1/S_2/S_3$) is thermally coupled to its corresponding SiGe HBT arrays. Enable (EN) is applied to the PA for dynamic operation. A compact input transformer balun is realized for equal power dividing with 180° phase shift, and an inter-stage four-way transformer [6] is designed to generate two differential signals for driving an output stage. A multi-section highpass matching network is inserted between the 2nd stage and the output stage to compensate for gain roll off. Neutralization capacitors (C_{neu}=0.73pF) are used at the output stage to mitigate the Miller effect and stabilize the PA. An output passive network is a parallel power-combining four-way transformer [7].

Details of the 2nd-harmonic-tuned four-way output transformer are revealed in Fig. 4.4.2. It is constructed using two primary windings and a secondary coil, all within a single footprint. Unique features of the back-end-of-line (BEOL) process provide a 4µm-thick aluminum layer (MA), a 3µm-thick copper layer (E1), and a continuous-run via (E2). The center and the innermost traces are a multi-layered stack connected through E2. With this structure, both broadside and lateral magnetic coupling among the primary windings and the secondary coil are obtained, leading to high magnetic coupling factor k=0.82, with a turn ratio n=1.9. Simulated passive efficiency of the transformer is 87.2% at 5.5GHz and maintains 86.0% from 4.5 to 6.5GHz. The multi-layered configuration has the two primary coils 4µm apart, resulting in a high primary-to-primary magnetic coupling factor k_P=0.53. Consequently, a large primary inductance modulation $(1+k_P)/(1-k_P)$=3.26 can be attained, depending on the mode of operation, enabling separate impedance matching at the fundamental and 2nd-harmonic frequencies. In an odd-mode, AC currents on the primary coils flow in the same direction, increasing the primary inductance L_P=254pH to $(1+k_P)L_P$=388pH, followed by an enhanced Q factor of 28.8 at 5.5GHz. The odd-mode equivalent circuit of the transformer for a differential pair consists of a leakage inductance $(1-k^2)(1+k_P)L_P$, a magnetizing inductance $k^2(1+k_P)L_P$, a PA parasitic capacitance C_P=1.9pF, and an equivalent load (C_L=0.33pF) shifted to the primary side. This 4th-order passive network transforms a 50Ω R_L to a 15Ω optimum impedance for a single-ended PA cell. In contrast, the even-mode primary inductance $(1-k_P)L_P$ is reduced to 119pH due to magnetic flux cancellation from the opposite current flow on the primary coils. As shown in the even-mode equivalent circuit, a C_{CT}=3.4pF at each center-tap (CT_1/CT_2) of the transformer synthesizes a highly capacitive pi-network, resulting in a low 2nd-harmonic impedance. Simulated odd-mode (Z_{f0}) and even-mode (Z_{2f0}) impedance plots show R_{f0} of 15.0±1.0Ω and R_{2f0} of 2.4±1.1Ω, which proves the proposed four-way transformer achieves simultaneous fundamental and 2nd-harmonic impedance matching with the efficient parallel power combining capability.

The electro-thermal behavior of the output stage is described in Fig. 4.4.3. The thermal sensor Q_{REF} is located between two SiGe HBT arrays Q_P/Q_N in a symmetrical fashion to sense and track their T_J. Other elements in the bias circuit are kept far away from Q_P/Q_N to minimize mutual thermal coupling. Due to complementary-to-absolute-temperature characteristic of Q_{REF}, the base voltage V_B decreases with increased power dissipation in Q_P/Q_N, which helps maintain their collector current and slow the β drop with increasing T_J. Simulations were performed using a layout-based thermal model provided by the process design kit. Average T_J of Q_P with thermally coupled Q_{REF} is lower by 12°C than thermally decoupled case at 24dBm. Since the bias circuit is operated in a closed loop, an insufficient bandwidth of the bias circuit can cause a static memory effect. Simulated normalized 1dB-gain bandwidth of the bias circuit is 300MHz, covering 3.75× of a 80MHz channel for fast tracking of an instantaneous PA-input-signal amplitude variation. In addition, the tight thermal coupling between Q_P/Q_N and Q_{REF} in the output stage reduces gain hysteresis in the PA dynamic operation, which indicates that the PA is less sensitive to a thermal memory effect compared to that without thermally coupling with Q_{REF}. Simulated P_{1dB}s are 32.0dBm and 31.2dBm with and without thermally coupled Q_{REF}, respectively. AM-PM conversion is further distorted if Q_P/Q_N are not thermally coupled to Q_{REF}. Thus, the optimal electro-thermal design would be of a key importance for highly linear and high power SiGe HBT PAs design.

A prototype was fabricated in GlobalFoundries SiGe BiCMOS 5PAX technology and the die size is 1.45×1.40mm^2 (Fig. 4.4.7). The BV_{CEO} and BV_{CBO} of the SiGe HBT are 6.1V and 17.7V, respectively. The existence of the avalanche diode Q_{3B} in the temperature-compensated dynamic bias circuit would extend collector-to-emitter voltage of a SiGe HBT beyond BV_{CEO}, making the PA robust under high-voltage-swing operation. Quiescent current is 225mA from a 5V supply. Figure 4.4.4 shows measured S-parameters and large-signal characteristics. Peak S_{21} is 32.2dB at 5.5GHz with <1dB gain variation from 4.9 to 5.9GHz. Wide bandwidth input matching and well-fitted output return loss are achieved. Measured output P_{1dB} and saturated P_{OUT} are 32.0±0.4dBm and 33.0±0.5dBm, respectively. Measured AM-PM conversion is within 2.5° at the output P_{1dB}. Very low 2nd/3rd-harmonic emissions (-47/-80dBm/MHz) indicate the excellent spectral purity of the PA without any harmonic rejection filters.

Dynamic EVM (DEVM) and PAE of the PA were plotted as shown in Fig. 4.4.5. For 802.11ac MCS9 VHT80 signal, the PA delivers 22.5/23.6/23.1dBm P_{OUT} (DEVM=-35dB) with 10.0/12.2/11.1% PAE at 5210/5530/5855MHz. P_{OUT} higher than 23.1dBm is observed under the various duty cycles at 5530MHz, indicating the PA is thermally robust. The PA also successfully supports 802.11ax standard (EVM=-38/40dB) at 21.4/19.4dBm. I/Q constellations and a spectral mask are depicted in Fig. 4.4.6. A table in Fig. 4.4.7 summarizes benchmarking against prior-art 802.11ac/ax WLAN PAs. The proposed SiGe HBT PA demonstrates the highest linear P_{OUT} and higher or similar PAE compared with Si-based 802.11ac/ax WLAN PA and is comparable to III-V counterparts in Fig. 4.4.7.

Acknowledgement:
We thank Skyworks Solutions Inc. for financial and technical support. We are also grateful to Integrand Software Inc. for providing the EMX® license.

References:
[1] Qorvo Inc. "TQP5525 Datasheet," 2016, Accessed on Aug. 1, 2017, <https://www.qorvo.com/products/d/da005626>
[2] Y. Chee et al., "A Digitally Assisted CMOS WiFi 802.11ac/11ax Front-End Module Achieving 12% PA Efficiency at 20dBm Output Power with 160MHz 256-QAM OFDM Signal," *ISSCC*, pp. 292-293, Feb. 2017.
[3] T.-M. Chen et al., "An 802.11ac Dual-Band Reconfigurable Transceiver Supporting up to Four VHT80 Spatial Streams with 116fs$_{rms}$-Jitter Frequency Synthesizer and Integrated LNA/PA Delivering 256QAM 19dBm per Stream Achieving 1.733Gb/s PHY Rate," *ISSCC*, pp. 126-127, Feb. 2017.
[4] W. Ye et al., "A 2-to-6GHz Class-AB Power Amplifier with 28.4% PAE in 65nm CMOS Supporting 256QAM," *ISSCC*, pp. 38-39, Feb. 2015.
[5] C.-W. Huang et al., "A Multimode 5-6 GHz SiGe BiCMOS PA Design Powers Emerging Wireless LAN Radio Standards," *IEEE BCTM,* pp. 38–41, Oct. 2017.
[6] T Cheung and J. Long, "A 21-26-GHz SiGe Bipolar Power Amplifier MMIC," *IEEE JSSC*, vol. 40, no. 12, pp. 2583-2597, Dec. 2005.
[7] J. Park et al., "A CMOS Wideband Current-Mode Digital Polar Power Amplifier With Built-In AM-PM Distortion Self-Compensation," *IEEE JSSC,* vol. 53, no. 2, pp. 340–356, Feb. 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

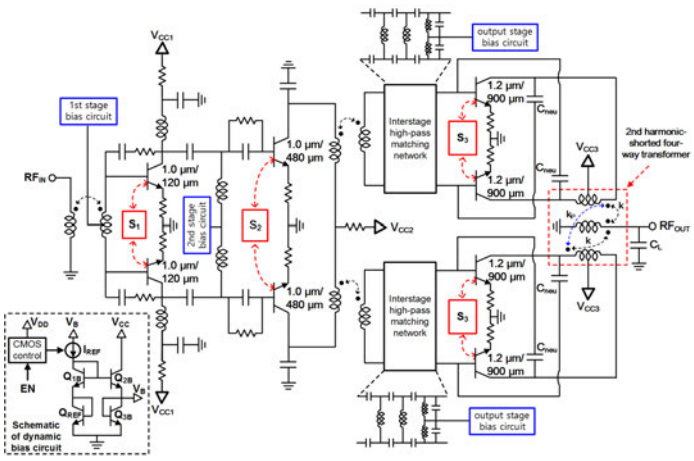


Figure 4.4.1: Schematic of the 3-stage SiGe HBT PA with a compact 2nd-harmonic-shorted four-way output transformer (highlighted as a red dotted box). A lower left inset is a schematic of a temperature-compensated dynamic bias circuit with an embedded thermal sensor (Q_{REF} is indicated as S_1, S_2, and S_3).

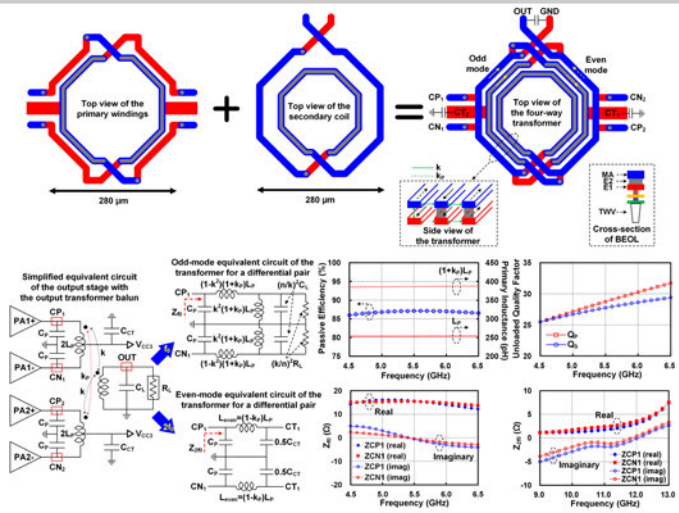


Figure 4.4.2: Implementation of the 2nd harmonic-shorted four-way output transformer balun, its odd- and even-mode equivalent circuits for a single differential pair, and 3D EM-based (by using EMX®) simulation results.

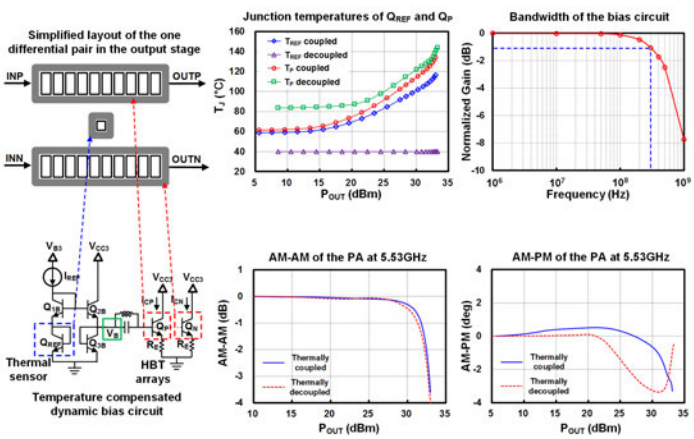


Figure 4.4.3: Electro-thermal design of the output stage and its simulation results; a junction temperature of Q_P and Q_{REF} over a P_{OUT} sweep, a bandwidth of the output-stage bias circuit, and AM-AM/AM-PM conversion of the PA with (solid) or without (dotted) thermal coupling among Q_P/Q_N and Q_{REF}.

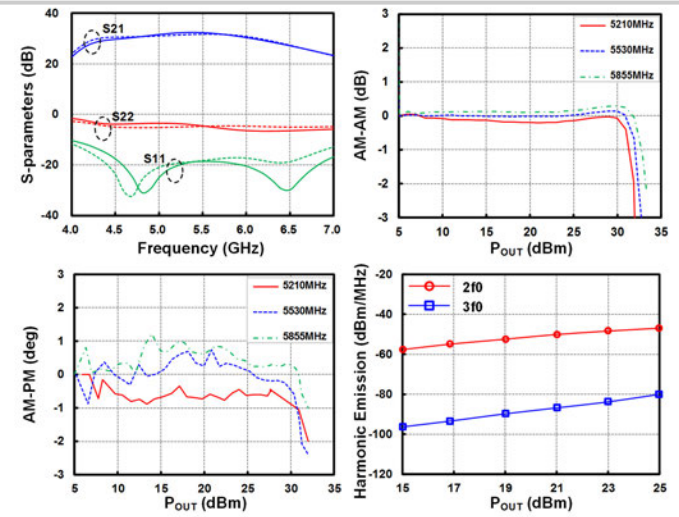


Figure 4.4.4: Measured S-parameters (solid line: measured; dotted: simulated), AM-AM (at 10% duty cycle), AM-PM (in CW mode), and 2nd/3rd-harmonic emissions at 5530MHz (802.11ac MCS9 VHT80 signal at 100% duty cycle).

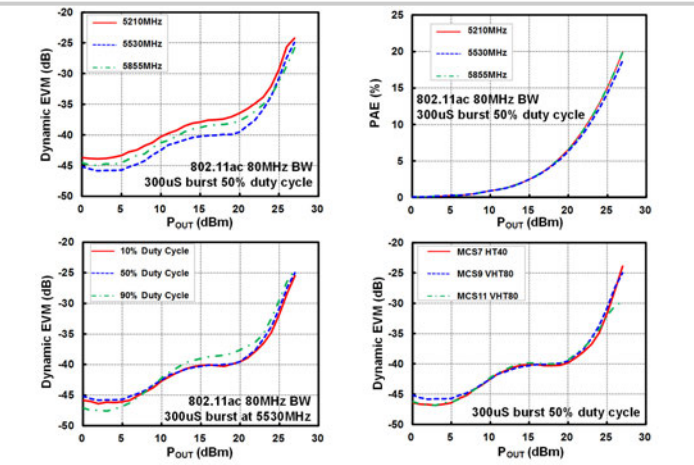


Figure 4.4.5: Measured DEVM (Upper left)/PAE (Upper right) vs P_{OUT} for an 802.11ac MCS9 VHT80 signal at 50% duty cycle, DEVM at various duty cycles at 5530MHz (Lower left), and DEVM under several modulations (Lower right) at 5510MHz (802.11n MCS7 HT40) and at 5530MHz (802.11ac MCS9 VHT80 and 802.11ax MCS11 VHT80).

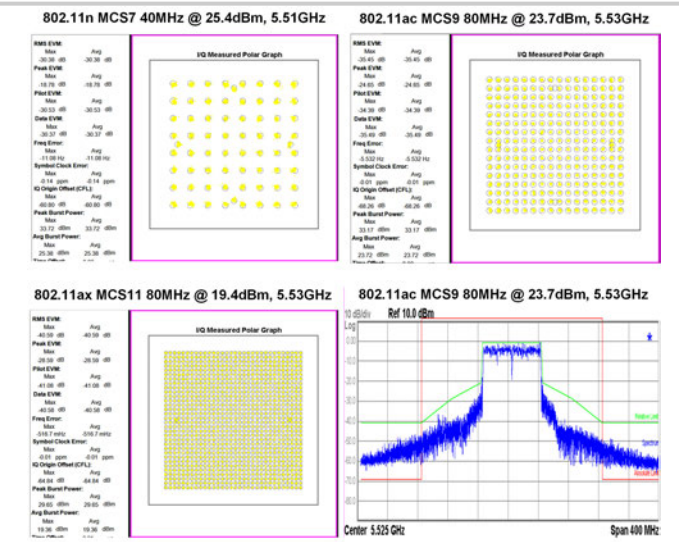
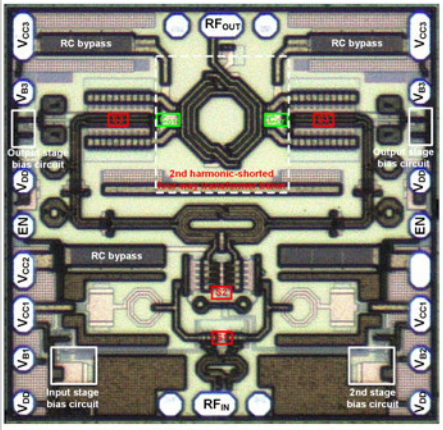


Figure 4.4.6: Measured I/Q constellations for various modulated signals and a spectral mask for 802.11ac MCS9 VHT80 test signal at 5530MHz.



		This work	[1]	[2]	[3]	[5]
Technology		SiGe BiCMOS	III-V HBT	55nm CMOS	40nm CMOS	SiGe BiCMOS
Supported WLAN Standards		11n/ac/ax	11n/ac	11n/ac/ax	11abgn/ac	11n/ac/ax
PA topology		Linear Class-AB	Linear Class-AB	Digitally assisted Doherty	Spatial power combining	Linear Class-AB
Frequency (MHz)		4900-5925	4900-5925	4900-5925	4900-5925	4900-5925
Supply (V)		5.0	5.0	3.3	-	5.0
Small Signal Gain (dB)		32.2	32.0	26.0	-	32.0
2nd/3rd harmonics (dBm/MHz)		-47/-80	-45/-45	-51/-51	-	-55*/-
P_{SAT} (dBm)		33	34	29	27	-
DPD requirement		No	No	Yes	Yes	No
802.11ac MCS9	P_{OUT} (dBm)	23.6	25.0	20.0	16.0/19.0	21.6
	PAE (%)	12.2	10.5	12.0**	-	10.2***
	DEVM (dB)	-35	-35	-35	-35	-35
	Ch. BW (MHz)	80	80	160	80/80+80	80
802.11ax MCS11	P_{OUT} (dBm)	21.4/19.4	-	18.0/15.0	-	20.3/19.8
	PAE (%)	8.5/6.1	-	10.0/7.0**	-	8.4/7.8***
	DEVM (dB)	-38/-40	-	-38/-40	-	-38/-40
	Ch. BW (MHz)	80	-	160	-	80
Chip size (mm^2)		2.03	-	1.42****	-	0.95

* estimated from S-parameters plot
** DPD power consumption is not included
*** calculated using I_{CC} plot
**** PA balun realized on 0.18μm SOI CMOS is not included

Figure 4.4.7: Die micrograph and a comparison table with prior-art 802.11ac/ax WLAN PAs.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

4.5 A 13.5dBm Fully Integrated 200-to-255GHz Power Amplifier with a 4-Way Power Combiner in SiGe:C BiCMOS

Mohamed Hussein Eissa[1], Dietmar Kissinger[1,2]

[1]IHP, Frankfurt (Oder), Germany; [2]Technische Universität Berlin, Berlin, Germany

In order to efficiently utilize the frequency band above 200GHz for radar and communication applications, enough transmitted output power is essential to overcome the elevated path loss. For silicon-based technologies, the available output power at this frequency range is limited. The enlarging of the transistor size and so the corresponding power handling leads to a reduced output impedance. This causes the impedance transformation ratio to increase allowing only for narrow-band designs. Power combining of multiple single power amplifiers (PAs) is a common approach to achieve high output power alleviating the need to enlarge the devices size [1]. In this paper, we present a fully integrated PA that combines the power of 4 differential PA units using a 4-way zero-degree combiner and a 4-way active splitter to feed the 4 PA units.

Different approaches are possible for power combining. Although Wilkinson power combiners guarantee certain isolation between the PA units, the expected achievable power-added efficiency (PAE) decreases due to the losses of the Wilkinson passive structure. Other approaches are to combine using transformers or T-Lines. In these cases, the combining structure is also utilized as the matching network, achieving better PAE compared to the Wilkinson power combiner. The T-line, or also called zero-degree power combiner, is more efficient compared to the transformer based combining as the targeted frequency increases. This is due to the more complex structure of the transformer-based combining network and the reduced mutual coupling for a few turns [2]. Both such combiners were designed and electromagnetically (EM) simulated. The 2-way T-line combiner simulation shows a loss of 0.4dB. This is a better performance than a 1.5dB loss for the transformer based 2-way combiner. In this work, we combine 4 differential PA units with a zero-degree power-combining structure, which is also the output matching of the single PA units.

In previous works, the power combiners were also utilized as the power distributing networks [1-3]. Instead, in this work a 2-cascaded 1-to-2 active power splitters are implemented. This enhances the overall PA bandwidth by optimizing the inter-stage matching between the stages so that the PA performance is not limited by the 4-way distribution network.

The schematic of the single PA unit, together with the circuit-core EM 3D view and its bias circuit, is shown in Fig. 4.5.1. The PA unit consists of a differential cascode amplifier. All transistors are biased at the maximum f_T bias conditions. The parasitic inductance at the bases of the cascode devices is minimized to guarantee unconditional stability of the PA unit. The bases of Q_3 and Q_4 are connected to guarantee a virtual ground and so to maximize the differential gain. The base capacitance (C_1) is still integrated to assure the bases are grounded across process and mismatch variations. The cascode base is connected to the bottom plate of C_1 to reduce the routing inductance, and then the upper plate is connected to the ground shield. S-parameters and load pulling simulations showed output impedance (Z_{out}) of (6-j42)Ω and the maximum saturated output power (P_{sat}) for Z^*_{opt} of (20-j32)Ω at 250GHz. Output matching is performed as a compromise between P_{sat} and S_{22}. For such impedance range, a short stub followed by a series transmission line is an optimal matching network.

Figure 4.5.2 shows the block diagram of the 4-way PA. To keep low impedance transformation ratios while combining, the impedance transformation is performed by TL_s and TL_1 before the first 2-to-1 combining and by TL_2 before the second 2-to-1 combining, as shown in the Smith chart in Fig. 4.5.2 at a chosen frequency of 250GHz. The 1-to-2 power splitters are also utilized as matching networks between the stages and are optimized for wideband response. The cross-overs within the zero-degree combiners and splitters were optimized to reduce the systematic mismatches by equalizing the electrical length of the differential signal as much as possible.

One bias circuit for each stage is implemented to avoid cross-coupling across the stages. For multi-stage designs, the supply and ground connections are critical to guarantee stability, especially since the k-factor simulated for the multi-stage amplifier is not sufficient to assure unconditional stability. For stability considerations, a small supply decoupling capacitor of 200fF is integrated in close proximity of each PA unit. Then larger capacitors, de-Q'ed with small resistors, are integrated as traces approach the DC-pads where supplies are combined.

Utilizing such power combining, power splitting, and PA-unit architectures, a fully integrated power amplifier was implemented in a 0.13µm SiGe:C BiCMOS technology with f_t\f_{max}=300\500GHz. The PA was characterized on wafer using two test structures. The first test structure is the standalone 4-way PA circuit, utilized to measure the 2-port S-parameters and the output 1dB compression point (OP1dB). While the second test structure is an integrated multiplier-by-8 chain driving the PA with enough power to measure P_{sat}.

In the first setup, waveguide probes (WR4.3), frequency extenders (RPG ZC260), and two DC probes for supply were used. Input and output Marchand baluns are integrated to allow for single-ended measurements. Measured loss of the Marchand balun and a pad is 1.5dB at each side.

Small-signal S-parameters in Fig. 4.5.3 show the maximum gain of 12.5dB across a 3dB bandwidth of 55GHz from 200 to 255GHz. This is shifted by approximately 4% from the simulation results mostly due to inaccuracies in the transistor models or the parasitic extractions. The input return loss (S_{11}) is centered on purpose at higher frequencies to enhance the gain flatness. As the design is meant for fully differential systems the S_{11} and S_{22} degradation due to the pad and balun are not compensated for. Stability k-factor is plotted showing values above 1 across the whole band. An important parameter for communication applications is the group-delay variation, which was measured to be only ±5ps across the whole band as plotted in Fig. 4.5.4, where the simulated and measured OP1dB are also plotted showing the maximum value of 9dBm. It becomes 10.5dBm after calibrating for the output balun and pad loss, which is equivalent to a power-added efficiency (PAE) at OP1dB of 1.47%.

Although the delivered power from the ZC260 was enough to measure the OP1dB, it was not enough to measure P_{sat}. Therefore, the second setup and a test structure were used where a 30GHz signal is delivered to the on-chip multiplier-by-8 chain, which then drives the PA with differential power of approximately 2dBm. The simulated and measured P_{sat} of the multiplier chain cascaded with the PA is shown in Fig. 4.5.5 across frequency, exhibiting a saturated output power of 12dBm. This becomes 13.5dBm, when omitting the losses of the output balun and the pad, leading to the drain efficiency (η_d) of 3%. Simulations results for 25°C and 40°C temperatures are plotted to highlight the effect of the elevated on-chip temperature on the performance.

Figure 4.5.6 summarizes the design parameters of the presented PA, together with the prior-art integrated PAs in silicon technologies above 200GHz. Very good performance can be observed when compared to the prior art. The maximum measured OP1dB of 9dBm is 5dB higher than the maximum achieved in designs reported in Fig. 4.5.6 with 7-times better PAE at OP1dB. The measured P_{sat} is approximately 2 times higher than the reported silicon-based power-combined PAs above 200GHz with 4 times higher drain efficiency [1]. All these results are achieved across a 55GHz of 3dB linear bandwidth, which is 1.8 times higher than the widest bandwidth reported in Fig. 4.5.6. The die micrograph of the integrated PA is shown in Fig. 4.5.7, where the different parts of the design are highlighted.

Acknowledgement:
This work was supported by the German Federal Ministry of Education and Research in the research project fast-spot (Project number 03ZZ0512A).

References:
[1] N. Sarmah et al., "A 200-225 GHz SiGe Power Amplifier with Peak P_{sat} of 9.6 dBm Using Wideband Power Combination," *ESSCIRC*, pp.193–196, Sept. 2016.
[2] W. Tai et al., "A 0.7W Fully Integrated 42GHz Power Amplifier with 10% PAE in 0.13µm SiGe BiCMOS," *ISSCC*, pp. 142-143, Feb. 2013.
[3] O. Momeni, "A 260GHz Amplifier with 9.2dB Gain and -3.9dBm Saturated Power in 65nm CMOS," *ISSCC*, pp. 140-141, Feb. 2013.
[4] N. Sarmah et al., "235-275 GHz (x16) Frequency Multiplier Chains with up to 0 dBm Peak Output Power and Low DC Power Consumption," *IEEE RFIC*, pp. 181–184, June 2014.
[5] M. Eissa et al., "216 - 256 GHz Fully Differential Frequency Multiplier-by-8 Chain with 0 dBm Output Power," *EuMIC*, pp. 201–204, Oct. 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

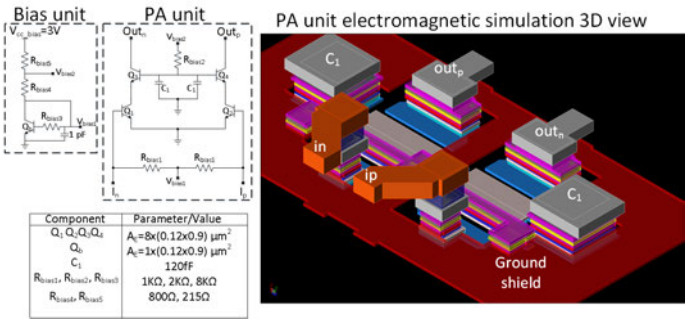


Figure 4.5.1: Schematic of the PA unit, the bias circuit, and the electromagnetic view for the transistor core.

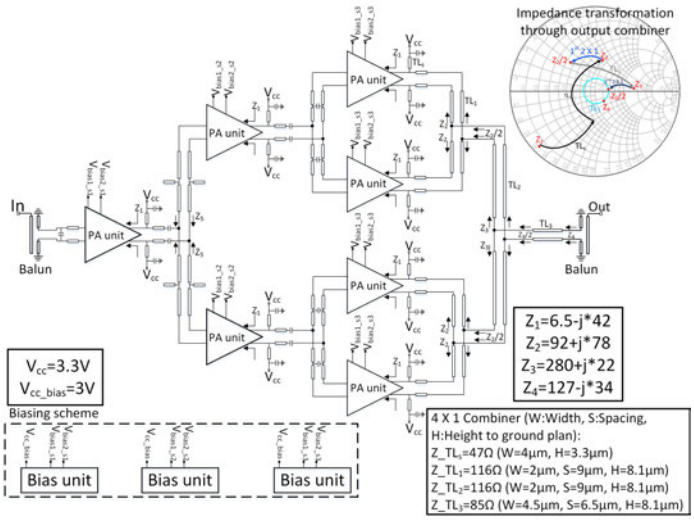


Figure 4.5.2: Block diagram of the 4-way power-combined PA.

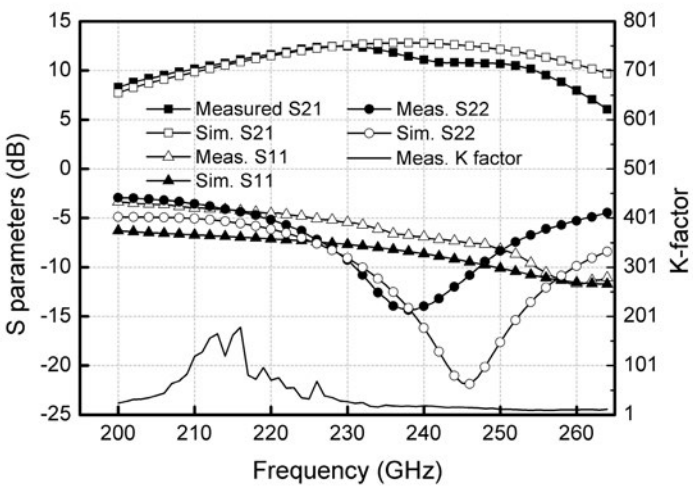


Figure 4.5.3: Simulated and measured S-parameters for the 4-way PA.

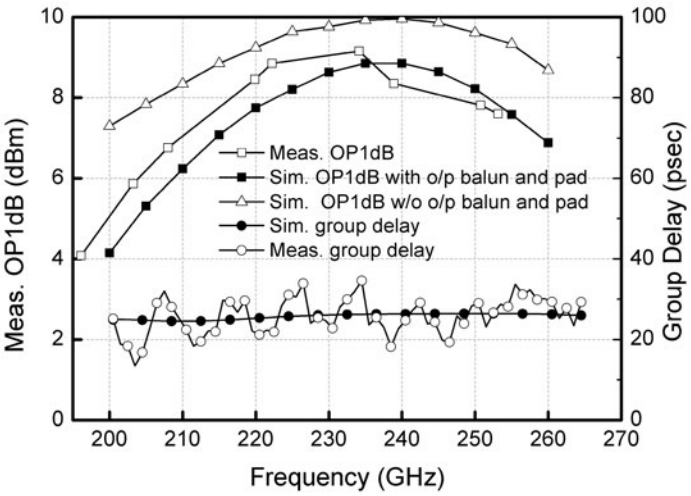


Figure 4.5.4: Simulated and measured output 1dB compression point and group delay variations across frequency.

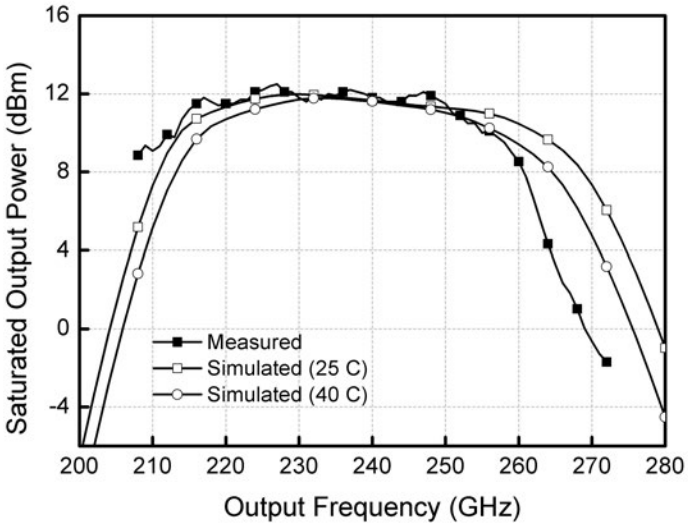


Figure 4.5.5: Simulated and measured saturated output power across frequency.

Reference	[1]	[3]	[4]	[5]	This work
Technology	130nm BiCMOS	65 nm CMOS	130nm BiCMOS	130nm BiCMOS	130nm BiCMOS
f_T/f_{max}	250/370	--	350/550	300/500	300/500
Power combinning Distributing structure	4-way zero-degree combiner Passive splitter	No	No	No	4-way zero-degree combiner Active splitter
-3dB BW (GHz)	200-220	251-263	228-258	208-258	200-255
Gain (dB)	25	9.2	10	9*	12.5* (15.5**)
Psat (dBm)	9.6	-3.9	5	0*	12* (13.5**)
OP1dB (dBm)	4	-5.5	--	-2.5	9* (10.5**)
DC Power (mW)	1824[1]	27.6	400	100	740
PAE at OP1dB (%)	0.14	0.89	--	0.5	1* (1.47**)
Max. Drain efficiency (%)	0.5	1.35	0.8[X]	1*	2.14* (3**)

*With no calibration for input and output baluns and pads.

** After calibration of input and output baluns and pads. (1.5dB for input and 1.5dB for output)

[1] Calculated from reported efficiency and saturate output power.

[X] Calculated from reported power consumption and saturated output power.

Figure 4.5.6: Performance comparison of silicon integrated PAs above 200 GHz.

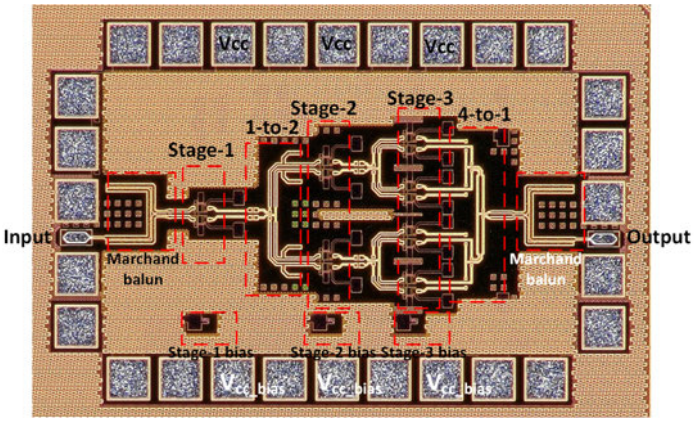


Figure 4.5.7: Die micrograph of the 4-way PA (Die area=0.83mm^2).

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

4.6 A mm-Wave 3-Way Linear Doherty Radiator with Multi-Antenna Coupling and On-Antenna Current-Scaling Series Combiner for Deep Power Back-Off Efficiency Enhancement

Huy Thong Nguyen, Sensen Li, Hua Wang

Georgia Institute of Technology, Atlanta, GA

There is a growing interest in exploring antenna-electronic co-designs that leverage antennas (including their low-loss metal structures, design versatility, multi-feed driving capabilities) as a new design paradigm to radically advance mm-wave front-end performance. In particular, enhancing the peak/back-off transmitting efficiency is essential for future mm-wave wireless communications, such as 5G, which will predominantly use complex modulations with large peak-to-average-power ratio (PAPR) and extra power back-off (PBO) for linearity and reliability. Consequently, mm-wave transmitters often operate at deep PBO (9 to 12dB) with poor average efficiency. Although mm-wave Doherty/Outphasing PAs have been widely explored, there is limited success in high-order active load modulation for deep-PBO efficiency enhancement with high PA linearity, mainly due to the complexity and prohibitive loss of on-chip passive networks. On the other hand, most existing antenna-electronic radiators only follow basic Class-AB/B efficiency roll-off with rapidly degraded PBO efficiency [1,2]. Some use active load modulation on a single multi-feed antenna, whose simple structures support only 6dB PBO efficiency enhancement and limited PBO extension [3].

In this paper, we present a general N-way series-combiner Doherty-PA architecture, and its 3-way design case is mapped to a 3-way linear Doherty radiator by exploiting dual-antenna coupling and an on-antenna current-scaling series combiner for simultaneous deep-PBO efficiency enhancement and high linearity. Performing Doherty combining/load-modulation on the antennas before radiation, the radiator also supports angle-independent Gb/s transmission over the entire antenna Field-of-View (FoV).

Figure 4.6.1 depicts the proposed general N-way Doherty-PA architecture based on current-scaling series power combiners and λ/4-line impedance inverters. To ease the implementation, we assume equal-weighting sub-PAs having the same max P_{out} but different onset points with the turn-on sequence of Main, Aux1, Aux2…, Aux(N-1) for N-way Doherty active load modulation. Although series combiners greatly facilitate voltage-limited silicon PA designs [4], their low-loss realizations at mm-wave are exceedingly challenging, let alone the realizations of specific scaling ratios. Note that mm-wave lumped transformers exhibit restricted scaling ratios, significant coil-to-coil capacitive coupling, limited resonant frequency, and substantial ohmic loss.

As an example, the schematic of a 3-way equal-weighting linear Doherty PA is shown in Fig. 4.6.1. It requires one λ/4 line and one series combiner with 2:1 current scaling that is electrically modeled as two transformers with 1:1 and 1:2 ratios. The Main PA connects to the 1:1 transformer, while the Auxiliary 1 (Aux1) and Auxiliary 2 (Aux2) PAs connect to the 1:2 transformer separated by a λ/4-line. The series transformers ensure $I_2 = 2I_1$, which accounts for three equal weighting sub-PAs that deliver the same RF current/power at 0dB PBO.

Inspired by recent antenna-electronics co-designs [1-3], we explore radiator-based solutions as near-field-coupled dual-slot antennas with offset feeds to synthesize a low-loss series combiner with a 2:1 current-scaling ratio. Figure 4.6.2 explains the electrical duality of a 2-feed slot antenna and two near-field-coupled slot antennas. The 2-feed slot antenna is equivalent to a 2-port parallel combiner [2], while the 2-feed near-field-coupled slots behaves as a 2-port series combiner [1]. Next, with offset feeds, the 2-feed slot antenna still guarantees a standing-wave voltage distribution at its resonance, resulting in voltage scaling ($V_1 = -kV_2$), where k is the standing-wave voltage ratio depending on the feed locations. By duality, the two near-field-coupled slot antennas with offset feeds exhibit current scaling with $I_1 = -I_2/k$, functioning as a series power combiner with a current-scaling ratio k. We then drive the port 1 of the antenna with the Main PA and connect port 2 to the Auxs PA in parallel with a λ/4-line and the Aux1 PA, achieving an equivalent circuit as the 3-way Doherty PA (Fig. 4.6.1).

When the Main/Aux1/Aux2 PAs follow the proper current relationship, the on-antenna network performs the desired 3-way Doherty active load modulation (Fig. 4.6.1). Assuming Zo is the optimum load-line impedance for one sub-PA at its max input drive, the Main PA load equals to 3Zo at a deep PBO and gradually decreases to Zo at 0dB PBO. After turning on, both Aux1/Aux2 PAs see a reducing load that equals Zo at 0dB PBO.

We use 3D EM simulations to verify the 3-way Doherty radiator behaviors and use far-field receiving dipole at boresight to check its AM-AM/AM-PM nonlinearity. The simulated [Y] matrix of the near-field coupled dual slot antennas with offset feeds validates its desired current scaling. The 3-way on-antenna Doherty load modulation is also verified. To evaluate the AM-AM/AM-PM nonlinearity only due to the 3-way on-antenna network, ideal Class-B PAs are used as the Main/Aux1/Aux2 PAs with different onset points. An excellent large-signal linearity is observed by far-field receiving dipoles with <0.2dB AM-AM and <2° AM-PM variation, verifying the use of the proposed design for practical wireless communication.

Figure 4.6.3 illustrates the schematic of the proposed 3-way Doherty radiator. The identical Main/Aux1/Aux2 PAs each comprise of a cascode PA at 2V VDD and a common-source driver at 1V V_{DD}. On-chip transformers are used as input/inter-stage matching and an output balun. To ensure the desired Main/Aux1/Aux2 PA output-current relationship, adaptive biasing circuits are used for the Aux1/Aux2 PAs to completely turn them off at deep PBO and sequentially turn them on at the desired power levels. The envelope-detector gate-biasing voltages are reconfigurable to further optimize Aux1/Aux2 PA turning-on operations. The input feed uses a balun, one 4.8dB coupler, and one 3dB coupler to generate three inputs with equal power and proper phase shifts for the 3-way Doherty radiator.

The 60GHz 3-way linear Doherty radiator is implemented in a 45nm CMOS SOI process, occupying 1.9×1.9mm^2 including pads and an on-chip antenna (Fig. 4.6.7). The chip is flip-chip packed to a Rogers PCB board to perform backside radiation. The gates of the envelope detector are biased at 0.27V/0.12V for Aux1/Aux2 PAs respectively for their desired onset sequence and the 3-way Doherty PA co-operation. The adaptive biasing settings are fixed for the CW and modulation measurements. The measured DC currents of the three PAs verify the adaptive biasing functionalities, showing a rapid turning-on of Aux1 PA and an even sharper turning-on of Aux2 PA (Fig. 4.6.4). The Doherty active load modulation and Main/Aux1/Aux2 PAs co-operation simultaneously enhances the PAE over 9dB PBO and achieves a linear PA gain. The output power of the 3-way Doherty radiator is calculated as the power delivered to the antenna network. At 57GHz, the saturated output power (P_{sat}) is 21.2dBm, which is based on the average of 21.5dBm P_{sat} derived from measured PA test structures and 21dBm P_{sat} extracted from the radiator EIRP and its antenna test structure measurements. The measured CW PAE closely resembles the ideal 3-way Doherty response (Fig. 4.6.1) with three PAE peaks of 21.8%/20.1%/19.3% at 0dB/4dB/8dB, showing 1×/ 1.42×/ 2.34× efficiency over a reference ideal Class-B PA. The PA OP1dB/P_{sat}/peak EIRP are 20.7dBm/21.2dBm/21.2dBm with high linearity. Similar deep back-off efficiency enhancement and large-signal linearity are achieved for 56/58GHz.

The 3-way Doherty radiator is then tested with modulations. Undistorted 64-QAM constellation is observed over the entire antenna E-plane/H-plane FoV. At 57GHz carrier frequency with 250Msym/s 64-QAM (1.5Gb/s), the average PAE curve over the average P_{out} also exhibits the "Doherty shape" with relatively flat response. Similar flat average PAE vs. PBO performance is achieved with 500Msym/s (Fig. 4.6.5). Both results validate the deep PBO PAE enhancement by the 3-way linear Doherty radiator. The radiator achieves 15.5dBm average P_{out} with 18.3% average PAE at 25.5dB SNR for 64-QAM 250Msym/s (1.5Gb/s) and 13.5dBm average P_{out} with 17.5% average PAE at 23.5dB SNR for 64-QAM 500Msym/s (3Gb/s). The modulation speed is limited by the radiation-based measurement setup, speed of adaptive biasing circuit, and antenna bandwidth.

Comparing with silicon 60GHz PAs and radiators reported in Fig. 4.6.6, the proposed 3-way linear Doherty radiator reports the first high-order Doherty active load modulation at high mm-wave and the highest PAE enhancement at deep PBO.

Acknowledgements:
We would like to thank GlobalFoundries for chip fabrication and members of Georgia Tech GEMS group for their technical supports.

References:

[1] B. Abiri and A. Hajimiri, "A 69-to-79GHz CMOS Multiport PA/Radiator With +35.7dBm CW EIRP And Integrated PLL," *ISSCC,* pp. 404–405, Feb. 2018.

[2] T. Chi et al., "A 60GHz On-Chip Linear Radiator with Single-Element 27.9dBm Psat and 33.1dBm Peak EIRP Using Multifeed Antenna for Direct On-Antenna Power Combining", *ISSCC,* pp. 296-297, Feb. 2017.

[3] H. Nguyen et al., "A 62-to-68GHz Linear 6Gb/s 64QAM CMOS Doherty Radiator with 27.5%/20.1% PAE at Peak/6dB-Back-Off Output Power Leveraging High-Efficiency Multi-Feed Antenna-Based Active Load Modulation," *ISSCC,* pp. 402-404, Feb. 2018.

[4] E. Kaymaksut et al., "Transformer-Based Doherty Power Amplifiers for mm-Wave Applications in 40-nm CMOS," *IEEE TMTT,* vol. 63, no. 4, pp. 1186-1192, Apr. 2015.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

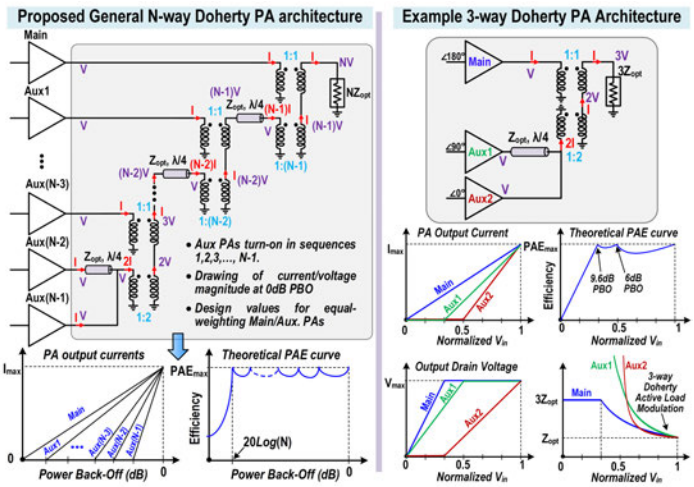


Figure 4.6.1: Proposed general N-way Doherty-PA architecture based on a series-combiner, design parameters, and theoretical operation for equal-weighting N-way Doherty PAs; and an example as a *N*=3 Doherty PA.

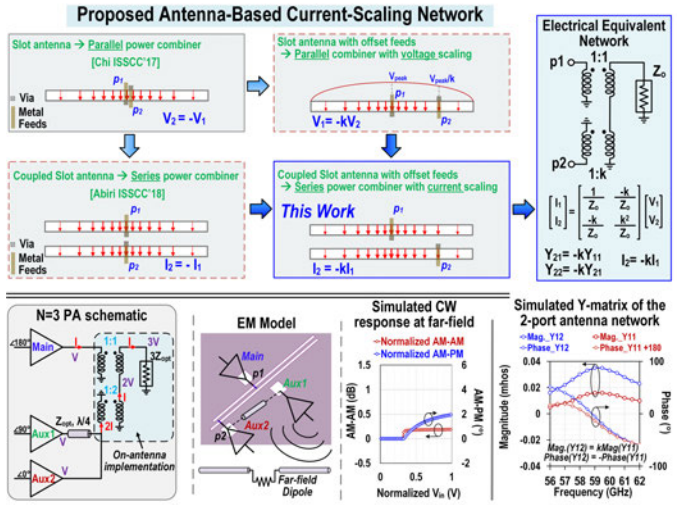


Figure 4.6.2: Electrical properties of 2-port radiation structures with center/offset-feed slot/near-field-coupled slots, equivalent electrical network, EM model, simulated Y-matrix, and CW responses at boresight.

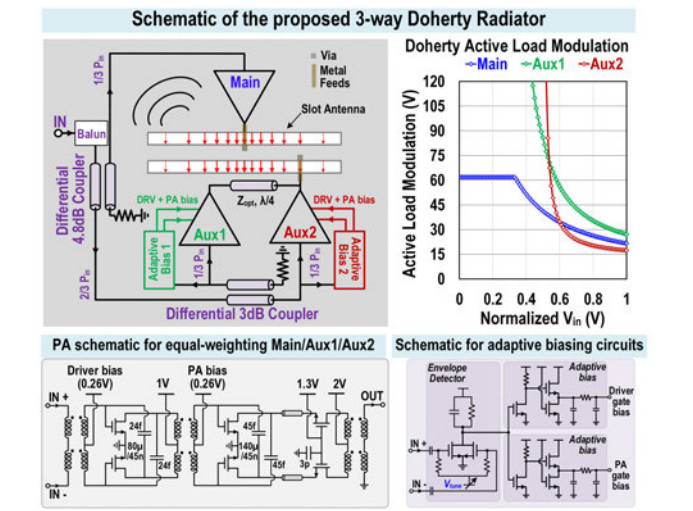


Figure 4.6.3: Top-level schematic of the proposed 3-way Doherty radiator, schematic for equal-weighting PAs, schematic for the adaptive biasing circuit, and simulated Doherty active load modulation.

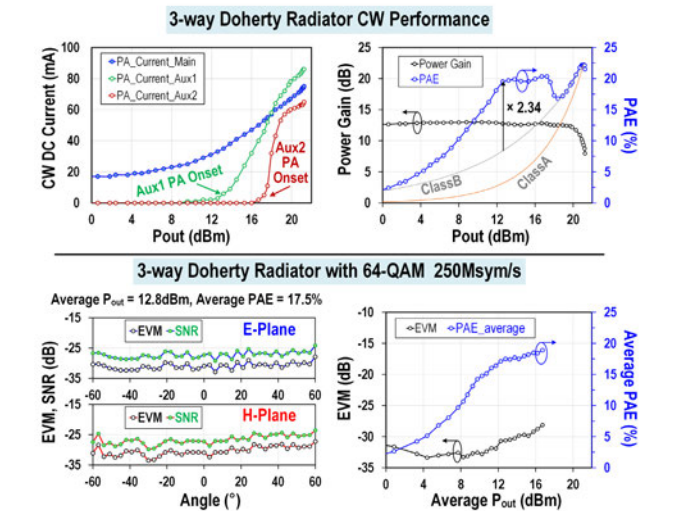


Figure 4.6.4: Measured large-signal CW performance at 57GHz, including DC current, power gain, PAE, and measured modulation performance with 64-QAM 250Msym/s at 57GHz at boresight and over the entire antenna Field-of-View.

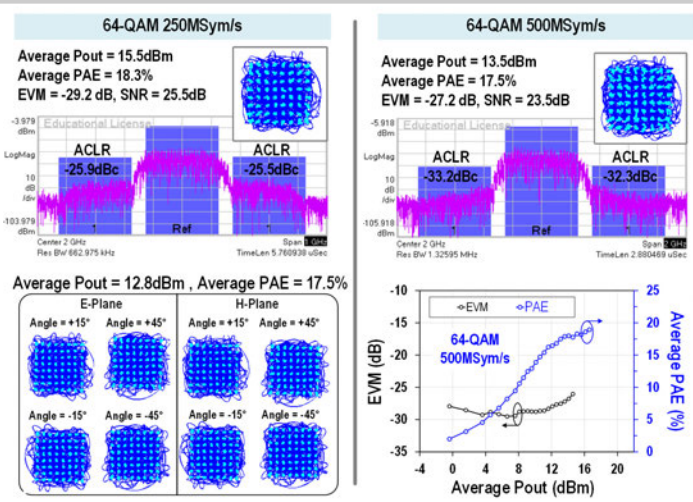


Figure 4.6.5: Demodulated constellations for 250Msym/s and 500Msym/s 64-QAM signals, and measured EVM and average PAE for 500Msym/s 64-QAM signals at the carrier frequency of 57GHz.

	This work	Greene, JSSC' 17	Kaymaksut TMTT'15	Nguyen, ISSCC'18	Chi, ISSCC'17	Khalaf, JSSC'16	Zhao, JSSC'12	Chappidi, JSSC'17
Architecture	3-way on-antenna Doherty	Doherty	Doherty	On-antenna Doherty	On-antenna combiner	Digital Polar Tx	Class AB	Asym. Combiner
Frequency (GHz)	57	62	72	65	60	60	62	55
P_{sat} (dBm)	21.2	17.5*	21	19.4	27.9	10.8	20.1	23.6
P_{1dB} (dBm)	20.7	17.1	19.2	19.2	25	7.4	19.7	19.9
Peak PAE	21.8%	23.7% *	13.6%	28.3%	23.4%	29.8%	18%	27.7%
PAE @ P_{1dB}	20.6%	23.7%	12.4%	27.5%	16.2%	15% *	18%	15.7%
PAE @ 6dB PBO	19.5%	13%	7%	20.1%	6%*	4.5% *	7%	7% *
PAE @ 9dB PBO	18%	7% *	4% *	14% *	3%	2% *	3%	3% *
Mod. Scheme Data Rate SNR (MER) Average P_{out} Average PAE	64-QAM 1.5Gb/s 25.5dB 15.5dBm 18.3%	64-QAM 3Gb/s 23.5dB 13.5dBm 17.3%	--	64-QAM 0.6Gb/s 25.6dB 15.9dBm 7.2%	64-QAM 6Gb/s 23dB 14.2dBm 20.2%	64-QAM 4.8Gb/s 25.4dB 19.3dBm 5.3%	16-QAM 6.7Gb/s 18.1dB 3.6dBm --	--	64-QAM 3Gb/s 21dB 14.8 dBm --
Area (mm²)	3.61*	0.6	0.19	3.23*	10.5*	<0.18	0.72	1.02
Technology	45nm CMOS SOI	130nm SiGe	40nm CMOS	45nm CMOS SOI	45nm CMOS SOI	40nm CMOS	130nm SiGe	130nm SiGe

* Estimated from reported figures
** Compared to an idealistic class-B PA with the same PAE at P1dB
* Including pads and on-chip antenna

Figure 4.6.6: Table of comparison to prior-art 60GHz PAs and radiators design.

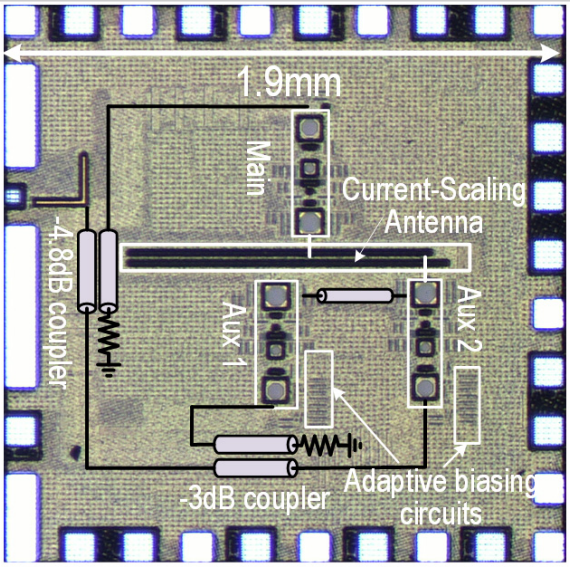


Figure 4.6.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

4.7 A Compact DC-to-108GHz Stacked-SOI Distributed PA/Driver Using Multi-Drive Inter-Stack Coupling, Achieving 1.525THz GBW, 20.8dBm Peak P_{1dB}, and Over 100Gb/s in 64-QAM and PAM-4 Modulation

Omar El-Aassar, Gabriel M. Rebeiz

University of California, San Diego, La Jolla, CA

Distributed amplifiers (DAs) are important circuit blocks for wireline optical links, high-resolution imaging systems, and millimeter-wave instrumentation. A clear trade-off exits in DAs between the gain and output power from one side and the bandwidth and chip area from the other. In this work, we propose a compact 23dB-gain CMOS distributed power amplifier (DPA) that operates from DC to 108GHz using several techniques to alleviate the output power-bandwidth trade-off. The DPA exploits the complementary nature of NMOS-PMOS to minimize AM-AM and AM-PM distortion and achieves over 100Gb/s in both 64-QAM and PAM4 modulation.

The proposed CMOS DPA consists of three cascaded sections (Fig. 4.7.1-top). An input 3-stage NMOS DA with supply-fed termination drives an intermediate inverted complementary stacked driver. The latter splits the signal in two paths, NMOS and PMOS, to feed a dual-input 5-stage triple-stacked output CMOS PA. The supply resistive biasing is only used in the first and intermediate stages and is essential to provide matching and flat gain from true DC. A 15dB-gain CMOS topology is used in the output stage to enable operation from DC with large voltage swing while no current flows in the output termination resistor under equal NMOS-PMOS drive. Without DC power dissipation in the termination resistor, the PA efficiency can be optimized. The presence of complementary paths (NMOS and PMOS), in both intermediate and output stages, is exploited to compensate for both AM-AM and AM-PM conversion by independent NMOS-PMOS biasing. The intermediate stage adopts an inverted complementary structure to provide a wideband real low input impedance (Z_{in}) (Fig. 4.7.1-bottom-left), where Z_{in} follows the lower impedance of both paths, controlled by their effective transconductance (Gm). The input capacitance variation is partially compensated by the opposite variation of NMOS and PMOS gate-to-source capacitance (Cgs) versus voltage swing. Stacking is applied inside the amplifier cells to break the trade-off between large required output power and lower transmission-line (TL) cut-off frequency due to larger-sized devices.

To arrive to the DA unit cell, a 3-stack inductor-peaked cell is first studied (Fig. 4.7.1-bottom-right). The peaking inductors should be placed in series to operate from DC at the expense of series resistive loss. The inductors resonate the intermediate-node parasitic capacitances and increase the effective Gm at high frequencies to help compensate for the input TL loss [1,2]. There remains, however, a part, i_{rc2}, of the current generated by the common-source device (M_{1CS}), i_{D1}, which is lost in the gate resistance (r_g) of the stacked common-gate device (M_{2CG}) through its Cgs. As frequency increases, the current at the top of the stack becomes lower than that generated by M_{1CS} (even if aligned), reducing the benefit of stacking. A multi-drive stacked cell [3] can alleviate this problem by forcing the common-gate device to inject an extra current compensating the current lost in r_g. However, the dissipated power of the common-gate drivers renders this technique non-efficient at lower frequencies, where less stack driving is needed. For a distributed structure, a stack multi-drive can be efficiently implemented by *tapping* a part of the voltage swing from the output of device M_n to drive the gate of the device M_{n+1} of the *following stage* (Fig. 4.7.1-bottom-right). The gate swing of M_{2CG} now slightly leads its original source voltage (V_{s2}), creating an extra current component (i_X) equal in amplitude and phase to i_{rc2} (in ideal operation) to align and equate i_{D1} and i_{S2}. Tapping from a previous stage also results in a leading voltage to compensate for the currents in the parasitic capacitance at V_{S2}, and thus the series inductor can be reduced for lower loss. The use of a small transformer (T1) for the feed-forward tapping ensures the technique is only applied at high frequencies. The coupling between the primary and secondary of T1 is adjusted by their separation to control the frequency bandwidth where the tapping is applied. T1 offers another advantage by having a small inductor at the gate of M_{2CG}. This inductor can create a negative real impedance component at V_{S2}, which is translated through the gate-to-drain capacitance (Cgd) of M_{1CS} to compensate for the input TL loss at high frequencies (Fig. 4.7.1-bottom-right). Similar to the emitter capacitive degeneration, used in most SiGe DAs, the inductor value should be kept small to ensure the real part of the TL input impedance remains positive over the entire frequency range.

The multi-drive inter-stack coupling DPA is shown in Fig. 4.7.2. Thin-oxide floating-body partially depleted SOI devices are used in all PA cells. The TLs, implemented in the upper 1.2μm-thick layer, are shaped as single-turn inductors to improve the inductance per unit length and keep the stages close for stack coupling, while the inter-stack coupling transformers are implemented in the upper two copper layers (both 1.2μm thick) with a patterned ground shield. RC lowpass filters are placed between the bias nodes to impede any common-mode instability due to routing. Biasing of the intermediate stage common-gates is used for pre-distortion and to control the AM-PM component (Fig. 4.7.2-top-left). Pads parasitic capacitance and routing are exploited to form small step-down impedance networks at both input and output ports.

The proposed CMOS DPA is implemented in a 45nm CMOS SOI process with a core area of 0.625×0.5mm^2 (Fig. 4.7.7). The measured and simulated small-signal parameters are presented in Fig. 4.7.3. The DA achieves a 23dB gain, DC-to-108GHz 3dB bandwidth, and a GBW of 1.525THz. The input and output matching are below -9dB up to 90GHz and below -6dB for the whole band. The group delay is within ±10ps up 90GHz (limited by the VNA calibration errors), and the amplifier is unconditionally stable (the dip in the k-factor at 75 to 120GHz can be explained by the degraded reverse isolation when using frequency extenders but remains higher than 2). The noise figure (NF) is measured up to 50GHz with a minimum of 6.9dB at 30GHz. The large-signal performance is shown in Fig. 4.7.4. The PA is measured up to 70GHz for two different settings: maximum P_{1dB} and minimum AM-PM conversion, by controlling the biasing of the intermediate and output stages. For the maximum P_{1dB} setting: the P_{1dB} and P_{sat} are higher than 17.4dBm and 18.4dBm up to 70GHz, respectively, and are within 1.2dB from 25 to 70GHz, demonstrating the efficacy of the AM-AM compensation using the pre-distortion knobs. Peak performance is obtained at 50GHz with P_{1dB}, P_{sat}, and peak PAE of 20.8dBm, 21.5dBm, and 13.4%, respectively (Fig. 4.7.4-top). At the minimum AM-PM conversion setting: the P_{1dB} and P_{sat} remain better than 16.8dBm and 18.3dBm, respectively, up to 70GHz. The AM-PM conversion at P_{1dB} is improved to +2.1°/-3.6° at DC to 70GHz, compared to +2.1°/-11.9° for the maximum P_{1dB} setting (Fig. 4.7.4-bottom-left). The power gain is reduced by ~1.9 dB while the PAE remains unchanged (Fig. 4.7.4-bottom-right).

The DPA is measured with 64-QAM signals from DC to 63GHz (Fig. 4.7.5-left). First, the carrier is centered around 14GHz and 102Gb/s (17GBaud) is demonstrated with an average P_{out} of +7.1dBm limited by the setup pre-DUT gain. Next, the frequency spectrum is divided into 5 setups and baud rates up to 5GBaud (30Gb/s) are tested for each setup, with the EVM kept below -25dB for all measurements. The CMOS DPA achieves an average P_{out} > 12.9dBm for the whole frequency range (except a dip at setup 4, Fig. 4.7.5 top-left). At 59GHz, the measured average P_{out} is +13.2dBm for 30Gb/s, with a PAE of 4.7%. The CMOS DPA is then measured with serial data PRBS9 PAM4 using a sampling-scope and a bit-error-rate tester (BERT) as a source with some pre-emphasis (Fig. 4.7.5-right). The DPA demonstrates open PAM4 eyes at 56, 100, and 112Gb/s with a de-embedded single-ended peak-to-peak swing of 5.2V, 4.5V, and 3.6V, respectively.

Compared with wide-band and distributed silicon-based PAs in Fig. 4.7.6, our CMOS DPA demonstrates the highest GBW and one of the highest P_{1dB} and P_{sat} up to 70GHz. The DPA also achieves the highest 64-QAM modulation bit-rate up to 59GHz, and single-ended output swing among PAs in Fig. 4.7.6. The efficiency is the highest among DAs [1-3], in a chip area comparable to non-distributed PAs.

Acknowledgements:
The authors would like to thank Analog Devices for their financial support, Integrand for EMX electromagnetic simulator, Keysight Technologies for providing PAM4 measurement equipment, and GlobalFoundries for chip fabrication.

References:
[1] P. V. Testa et al., "170 GHz SiGe-BiCMOS Loss-Compensated Distributed Amplifier," *IEEE JSSC,* vol. 50, no. 10, pp. 2228-2238, Oct. 2015.
[2] A. Jahanian and P. Heydari, "A CMOS Distributed Amplifier With Distributed Active Input Balun Using GBW and Linearity Enhancing Techniques," *IEEE TMTT,* vol. 60, no. 5, pp. 1331-1341, May 2012.
[3] A. Agah et al., "Multi-Drive Stacked-FET Power Amplifiers at 90 GHz in 45 nm SOI CMOS," *IEEE JSSC,* vol. 49, no. 5, pp. 1148-1157, May 2014.
[4] K. Fang et al., "Supply-Scaling for Efficiency Enhancement in Distributed Power Amplifiers," *IEEE JSSC,* vol. 51, no. 9, pp. 1994-2005, Sept. 2016.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

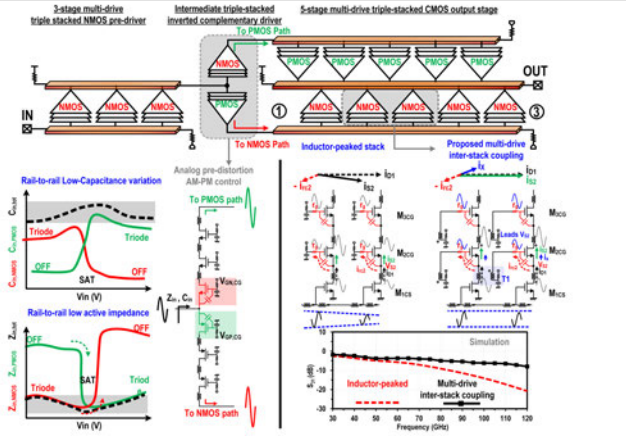


Figure 4.7.1: Architecture of the CMOS distributed amplifier with multi-drive inter-stack coupling (top). Intermediate stage inverted-complementary dual-path stacked-driver providing low input impedance and low capacitance variation across wide input swing (bottom-left). Schematic and current phasor diagram of the inductor-peaked unit cells compared to the proposed multi-drive inter-stack coupling (bottom-right).

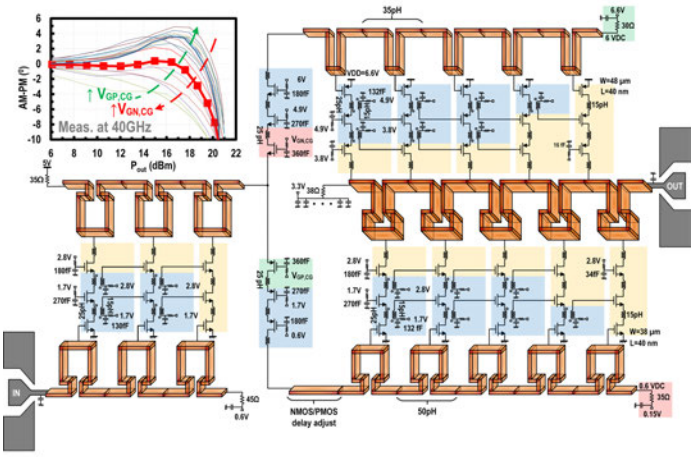


Figure 4.7.2: Schematic of the CMOS distributed amplifier and the measured AM-PM variation at 40GHz with intermediate stage pre-distortion.

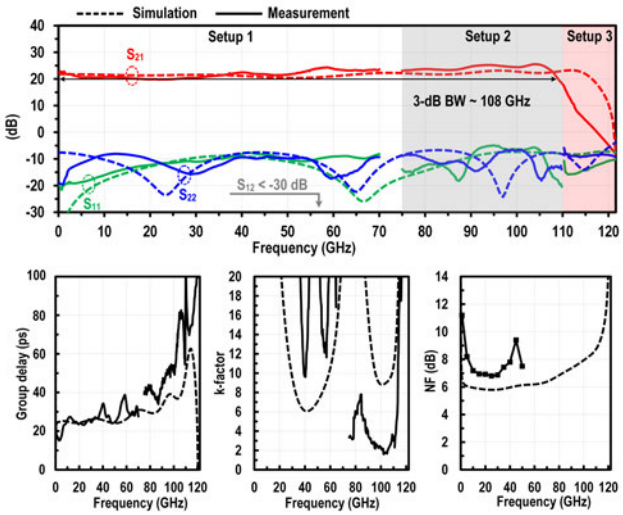


Figure 4.7.3: Small-signal performance versus frequency: measured and simulated S-parameters (top), group delay, stability factor (k), and noise figure (bottom).

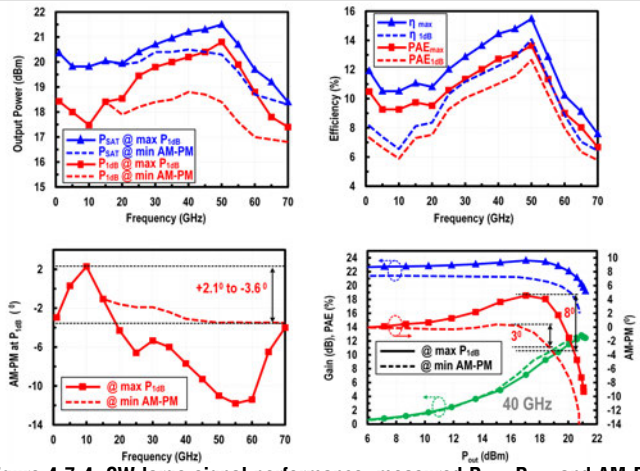


Figure 4.7.4: CW large-signal performance: measured P_{1dB}, P_{SAT}, and AM-PM across frequency for two settings: maximum P_{1dB}, and minimum AM-PM (left). Measured PAE_{P1dB}, PAE_{PSAT}, and drain-efficiencies η_{1dB}, η_{max} versus frequency (top-right). Measured gain, PAE, and AM-PM versus output power at 40GHz (bottom-right).

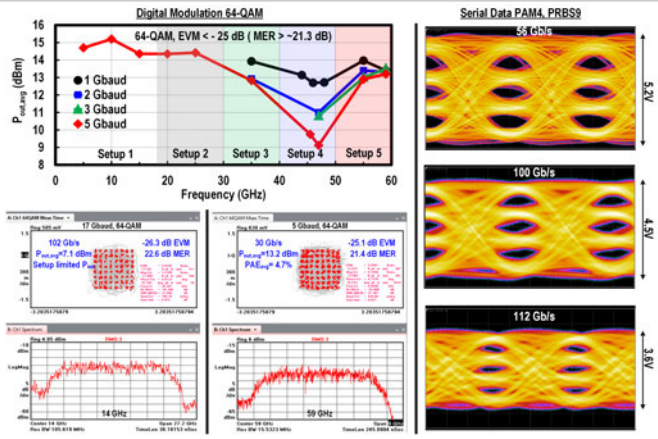


Figure 4.7.5: Modulation and time domain measurements: measured average output power for 64-QAM vs baud rate for an EVM < -25dB across frequency (top-left). Measured 64-QAM constellations: 102Gb/s at 14GHz, and 30Gb/s at 59 GHz (bottom-left). Measured PAM4 PRBS9 eye diagrams demonstrating 56, 100 and 112Gb/s with single-ended peak-to-peak voltage of 5.2V, 4.5V, and 3.6V, respectively (bottom-right).

Design	This work	T.W. Li et al., ISSCC'18	Vigilante et al., JSSC'18	Hu et al., ISSCC'17	Chappidi et al., JSSC'18	[1] JSSC'15	[2] TMTT'12	[4] JSSC'16
Technology process	45nm RFSOI	130nm SiGe	28nm	130nm SiGe	130nm SiGe	130nm SiGe	65nm	130nm SiGe
Frequency (GHz)	DC-108	19-29.5	29-57	23.3-39.7	30-55	DC-170	DC-65	14-105
Supply (V)	5 (pre-driver) / 6.6 (PA)	1.1	0.9	1.5	4, 1.6	3.6	1.3	4, 3.6, 3.2, 2.7
Gain (dB)	23	20	20.8	17*	23.4	10	22	12
GBW (GHz)	**1525**	110	307	116	370	537	818	362
[&]P_{1dB} (dBm)	**19.5 @ 25GHz 20.8 @ 50GHz 17.5 @ 70GHz**	15.2 @ 28.5GHz	13.4 @ 30GHz 11.1 @ 40GHz 10.9 @ 50GHz	15.2 @ 28GHz 15.5 @ 37GHz 15.4 @ 39GHz	N.R.	7.5* @ 50 GHz	10 @ 2GHz	14.9 @ 50GHz
[&]P_{sat} (dBm)	**20.5 @ 25GHz 21.5 @ 50GHz 18.4 @ 70GHz**	17 @ 28.5GHz	16.6 @ 30GHz 15.9 @ 40GHz 15.1 @ 50GHz	16.8 @ 28GHz 17.1 @ 37GHz 17 @ 39GHz	23.7 @ 40GHz	9.4* @ 50 GHz	12* @ 2GHz	17 @ 50GHz
PAE_{max}(%) / PAE_{P1dB}(%)	10.8/9.1 @ 25GHz 13.4/12.5 @ 50GHz 6.8/5.9 @ 70GHz	43.5/39.2	24.2/12.6 @ 30GHz 18.4/7.5 @ 40GHz 14.9/7.0 @ 50GHz	20.3/19.5 @ 28GHz 22.6/21.6 @ 37GHz 21.4/20.7 @ 39GHz	24.7* @ 30GHz 28.5 @ 40 GHz 23.4* @ 50 GHz	N.R.	N.R.	12.6/8.5* @ 50GHz
[&]AM-PM @P1dB (°)	+2.1 to-3.6 from DC to 70GHz	-3	0.45 @ 50 GHz	N.R.	N.R.	N.R.	N.R.	N.R.
Frequency (GHz) Modulation Bit-rate (Gb/s) #EVM (dB) Pout (dBm) PAE (%)	14 | 25 | 59 64-QAM 102 | 30 | 30 < -25 +7.1 | +14.3 | +13.2 2.6 | 5 | 4.7	28.5 64-QAM 18 -25 +9.8 18.4	34 64-QAM 6 -28.7 +5.9 2.3	39 64-QAM 3 -28.7 +9.3 17.2 (CE)	50 16-QAM 4 -21.75 +16.9 12.6*	N.R.	N.R.	N.R.
PAM4 bit-rate (Gb/s)/ Output swing (Vpp)	56/5.2 100/4.5 112/3.6	N.A.	N.A.	N.A.	N.A.	N.R.	N.R.	N.R.
Core Area (mm^2)	**0.31**	0.29	0.14	1.76^	0.96	0.38	0.93	1.51

N.R. = Not Reported
N.A. = Not Applicable
† Minimum value
& Best-setting at each frequency
* Estimated from plots
# Normalized to peak of the constellation
^ Including pads

Figure 4.7.6: Comparison with wideband millimeter-wave and distributed power amplifiers.

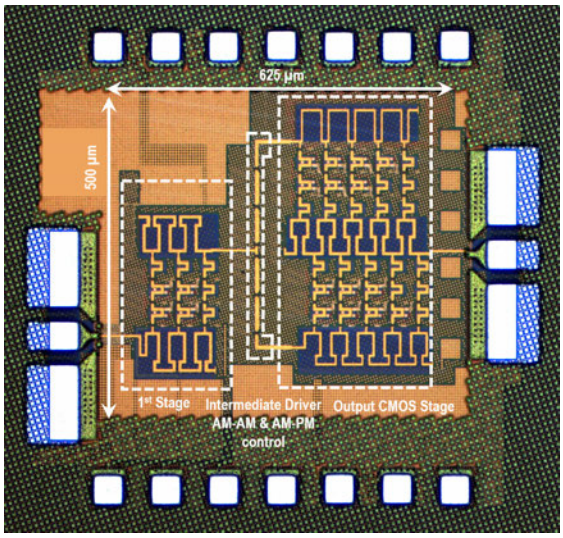


Figure 4.7.7: Die micrograph of the fabricated CMOS distributed amplifier chip in a 45nm SOI process.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

4.8 A Highly Linear Super-Resolution Mixed-Signal Doherty Power Amplifier for High-Efficiency mm-Wave 5G Multi-Gb/s Communications

Fei Wang, Tso-Wei Li, Hua Wang

Georgia Institute of Technology, Atlanta, GA

As the millimeter-wave (mm-wave) spectrum is believed to drive the next wireless communication revolution, there is a growing need for architectural research on advanced transmitter (TX) or power amplifier (PA) that can radically push the performance envelope of mm-wave power amplification to the next level [1]. This is especially true for future mm-wave 5G deployment that extensively employs high peak-to-average power ratio (PAPR) waveforms and poses demanding TXs/PAs requirements, including linearity, peak efficiency, power back-off (PBO) efficiency, and capability of supporting multi-Gb/s modulation [2-7].

Conventional analog mm-wave PAs usually suffer from steep PBO efficiency degradation and low average efficiency when amplifying high PAPR signals [2]. Although outphasing and Doherty PAs can enhance PBO efficiency, generating wideband outphasing signals often causes significant baseband overhead, while Doherty PAs with wideband modulations require delicate co-operations of the Main/Auxiliary (Aux) PAs [3,4]. Alternatively, digital PAs provide higher peak efficiency and better real-time controllability and reconfigurability than analog PAs. However, several fundamental limitations of digital PAs restrict their use in applications with high-quality multi-Gb/s modulations [5–7]. First, polar digital PAs suffer from large signal-path bandwidth expansion over the modulation bandwidth (i.e. 3× to 5× for the AM path and 5× to 7× for the PM path). Moreover, high oversampling requirement for out-of-band (OOB) image suppression greatly exacerbates the bandwidth challenges for Gb/s modulations. On the other hand, I/Q digital PAs often exhibit major I/Q cross-modulation and pulling, requiring extensive digital pre-distortion (DPD). Second, although RF digital PAs can use large effective number of bits (ENOBs) in AM controls, e.g., 13 bit [8], to ensure in-band/OOB linearity, it is challenging and power hungry to generate a large number of high-speed digital controls with precise synchronization for mm-wave digital PAs. Third, timing synchronization of AM/PM paths further limits the modulation quality/speed. In addition, most digital mm-wave PAs do not support PA efficiency enhancement better than the ideal Class-B PBO roll-off [7].

We propose a mixed-signal Doherty-PA (MSDPA) architecture that employs a hybrid use of analog and digital mm-wave PAs in the Main and Aux PA paths, respectively, which combines the advantages of the two yet overcomes their intrinsic limitations. The MSDPA consists of one analog PA as the Main PA and one digital PA (multiple binary-weighted sub-PAs in parallel) as the Aux PA (Fig. 4.8.1). The MSDPA operation is explained as follows. In contrast to conventional polar digital PAs that only take constant-envelope PM signals, the MSDPA is driven by the standard complex modulated signal with both phase/envelope modulations, avoiding signal-path bandwidth expansion. Simultaneously, based on the real-time AM envelope, different Aux sub-PA weightings are turned-on to extend the total PA compression point. Figure 4.8.1 compares the outputs of a 3b MSDPA, a 3b conventional digital PA, and a conventional analog PA. The analog PA often has a significant gain compression and thus large AM-AM distortion. The conventional digital PA only uses digitized and time-sampled interpolations of the desired envelopes, resulting in significant quantization errors, spectral images, and limited output dynamic range. In contrast, the MSDPA offers the following fundamental advantages. *(1) PA Dynamic Range and Quantization Error.* For small-input envelopes, MSDPA operates in its analog regime. Only the Main path analog PA is turned-on, and it operates identically as an analog mm-wave PA, supporting a small output signal (limited by analog PA noise) and a large dynamic range with no PA quantization error. *(2) PA PBO Efficiency.* For medium/large input envelopes, MSDPA operates in its mixed-signal regime. The Aux sub-PAs are dynamically turned-on, which modulates the Main PA load by Doherty operation to extend the PA output power with high PBO efficiency. The Aux PA digital bits also accurately control the Aux PA turning-on instances and guarantee optimum Doherty operation for more superior PBO efficiency than that of analog Doherty PAs. *(3) Quantization Error* is largely suppressed, since the MSDPA interpolates its output envelope using both envelope-varying input and real-time Aux PA digital controls, achieving "super resolution" over conventional digital PAs with the same ENOBs. *(4) Signal-Path Bandwidth Expansion* in conventional mm-wave digital-PA PM signals is also avoided by using the envelope-varying complex modulation input in the MSDPA.

Figure 4.8.2 shows the simulations of a 3b 28GHz MSDPA with extracted device parasitics in a 45nm SOI CMOS process. The blue curves stand for the MSDPA performance by gradually turning on the optimum Aux PA digital settings versus the input envelope, while the gray curves represent the MSDPA performance with various fixed Aux PA digital settings. The sequential turning-on points of the Aux sub-PAs are highlighted in red, optimized for low AM-AM gain ripple. Using only 3b controls, the MSDPA achieves <0.4dB AM-AM gain ripples and <5° AM-PM distortion (Fig. 4.8.2). With the optimum mixed-signal Doherty operation, it achieves 1.9× efficiency enhancement at 7dB PBO over a normalized Class-B PA (Fig. 4.82). The PA output spectra are simulated with a 1GSym/s 64-QAM modulation signal at 5× oversampling together with the MSDPA AM-AM/AM-PM nonlinearities. The MSDPA (using only 3b or 4b ENOB) substantially improves the close-in quantization noise floor and ACPR by 20dB compared with a conventional digital PA with the same ENOB, validating MSDPA "super-resolution" performance. Figure 4.8.2 also shows that the MSDPA (3b ENOB) suppresses the far-out OOB spectral images by 20dB compared to a 3b digital PA.

The proposed MSDPA is prototyped at 28GHz in a 45nm SOI CMOS process (Fig. 4.8.3). The Main PA stage is a cascode amplifier, while the Aux PA stage is a 3b binary-weighted cascode amplifier with source-switching to ensure reliability. The Main/Aux PA stages have the same total device size. A digital re-timing circuit synchronizes the three AM-based digital controls A<0>, A<1>, and A<2>, which are buffered by binary-weighted inverters to feed the Aux sub-PA source switching transistors. The chip occupies 2.87mm^2 with a core area of only 0.52mm^2 (Fig. 4.8.7).

Figure 4.8.4 shows the measured small-signal S-parameters from 20 to 40GHz. The S_{21} is 19.1dB at 27GHz with a 3dB bandwidth from 22.35 to 34.5GHz. The input matching is better than -10dB from 21.3 to 39.46GHz. The broadband gain flatness enables high-speed modulations. The large-signal continuous-wave (CW) results of different Aux PA 3b settings and the MSDPA performance are shown. The PA achieves 40.1% peak PAE with 23.3dBm P_{SAT} and 39.4% PAE at P_{1dB} of 22.4dBm at 27GHz. The PAE at 5dB/6dB is 37.5%/33.1% and 1.69×/1.68× over a normalized Class-B PA, all among the highest performance among reported 28GHz silicon PAs in [1]. The MSDPA mixed-signal gain ripple is <0.4dB.

Figure 4.8.5 shows modulation tests using a single-carrier 64-QAM signal (PAPR=6.5dB) with no DPD. The digital controls are generated in MATLAB based on the time-sampled envelope of the input modulated signal and measured CW performance. For a 12Gb/s (2GSym/s) 64-QAM signal, the MSDPA achieves -24.5dB rms EVM in its mixed-signal Doherty mode with an average P_{out}/PAE of +15.6dBm/27.8%. The digital circuits run at 8GSamples/s and 9mW, leading to 26.2% total average system efficiency.

The proposed MSDPA architecture advances the linearity-efficiency (peak/PBO)-modulation performance envelope of mm-wave PAs reported in Fig. 4.8.6.

Acknowledgements:
We thank GlobalFoundries for chip fabrication. This work was supported in part by Qorvo and Texas Instruments.

References:
[1] Hua Wang et al., "Power Amplifiers Performance Survey 2000-Present," Accessed on Sept. 10, 2018 [Online].
Available: https://gems.ece.gatech.edu/PA_survey.html
[2] S. Shakib et al., "A Wideband 28GHz Power Amplifier Supporting 8×100MHz Carrier Aggregation for 5G in 40nm CMOS," *ISSCC*, pp. 44-45, Feb. 2017.
[3] B. Rabet and J. Buckwalter, "A High-Efficiency 28GHz Outphasing PA with 23dBm Output Power Using a Triaxial Balun Combiner," *ISSCC*, pp. 174-176, Feb. 2018.
[4] S. Hu et al., "A 28GHz/37GHz/39GHz Multiband Linear Doherty Power Amplifier for 5G Massive MIMO Applications," *ISSCC*, pp. 32-33, Feb. 2017.
[5] K. Khalaf et al., "Digitally Modulated CMOS Polar Transmitters for Highly-Efficient mm-Wave Wireless Communication," *IEEE JSSC*, vol. 51, no. 7, pp. 1579-1592, July 2016.
[6] K. Dasgupta et al., "A 25 GB/s 60GHz Digital Power Amplifier in 28nm CMOS," *ESSCIRC*, pp. 207-210, Sept. 2017.
[7] C. Chappidi et al., "A Digital mm-Wave PA Architecture with Simultaneous Frequency and Back-off Reconfigurability," *IEEE RFIC*, pp. 328-331, June 2017.
[8] M. Alavi, *et al.*, "A Wideband 2×13-bit All-Digital I/Q RF-DAC," *IEEE TMTT*, vol. 62, no. 4, pp. 732-752, April 2014.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

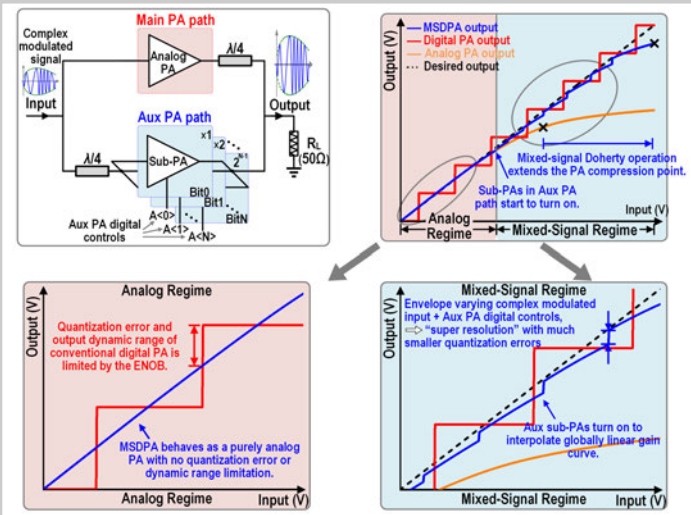


Figure 4.8.1: The conceptual diagram of the proposed mixed-signal Doherty PA (MSDPA), its operating principles, and comparison between a 3b MSDPA, 3b conventional digital PA, and conventional analog PA.

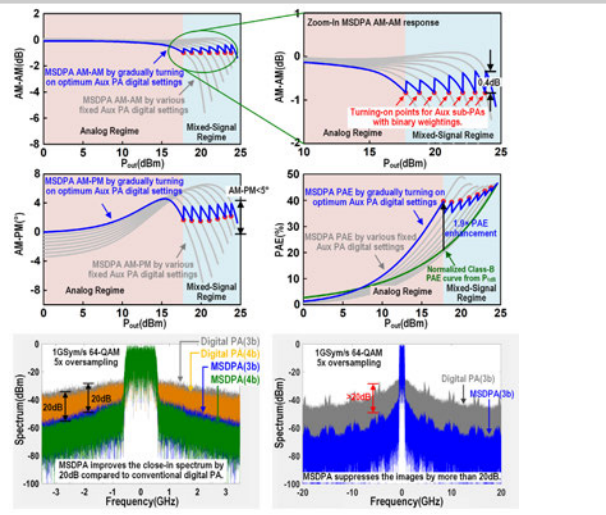


Figure 4.8.2: The continuous-wave (CW) simulation results of a 3b MSDPA using extracted transistors in a 45nm SOI CMOS process and spectrum simulation results comparison among the MSDPAs and digital PAs.

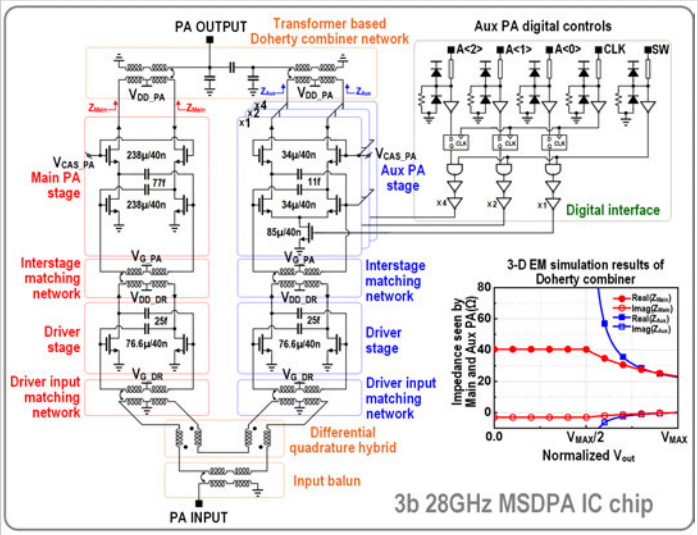


Figure 4.8.3: The top schematic of a proof-of-concept 3b MSDPA at 28GHz implemented in a 45nm SOI CMOS process.

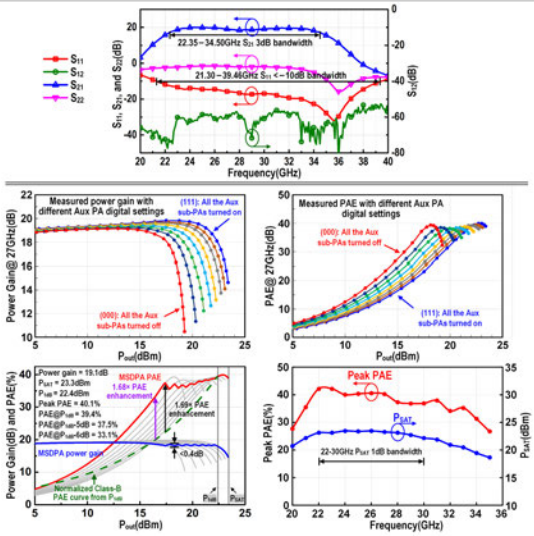


Figure 4.8.4: The small-signal S-parameter measurement results and the large-signal continuous-wave (CW) measurement results.

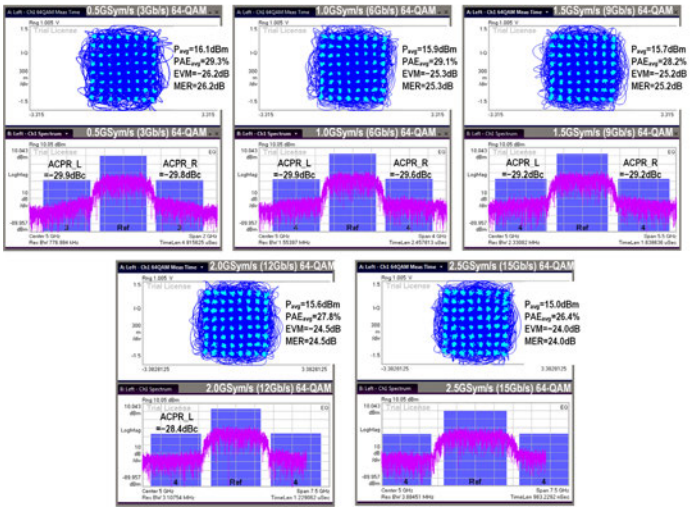


Figure 4.8.5: The modulation measurement results of the MSDPA.

Comparison with recently reported mm-wave digital and analog PAs in silicon

	This work		Chappidi, JSSC 18'	Khalaf JSSC 16'	Dasgupta ESSCIRC 17'	Hu, ISSCC 17'	Indirayanti, RFIC 17'	Rabet, ISSCC 18'	Shakib, ISSCC 17'
Technology	**45nm SOI CMOS**		0.13μm SiGe	40nm CMOS	28nm CMOS	0.13μm SiGe	28nm CMOS	0.13μm SiGe	40nm CMOS
Architecture	**3b Mixed-Signal Doherty PA**		4b Digital Asymmetrical PA	4b Digital Polar PA/TX	6b Class-E Digital PA	Multiband Analog Doherty PA	Analog Doherty PA	Triaxial Balun Outphasing PA	Analog PA
Supply (V)	**2.0**		4.0	0.9	1.1	1.5	1.0	4	1.1
Freq. (GHz)	**27**		40	60	60	28	32	28	27
Gain (dB)	**19.1**		23.4	24.3	9.3	18.2	22	14.0	22.4
P_{SAT} (dBm)	**23.3**		23.7	10.8	7.4	16.8	19.8	23	15.1
P_{1dB} (dBm)	**22.4**		N.A.	7.4	N.A.	15.2	16	N.A.	13.7
Peak PAE	**40.1%**		28.5%	29.8%	17.7%	20.3%	21%	41.4***	33.7%
PAE@P_{1dB}	**39.4%**		N.A.	15%†	N.A.	19.5%	12.8%	N.A.	31.1%
PAE@PBO	**33.1%@ 6dB PBO**		14.8%@ 6dB PBO	5.6%@ 6dB PBO	6.9%@ 6dB PBO	13.9%@ 5.9dB PBO	4%@ 6dB PBO†	34.7%@6dB PBO***	15.1%@ 5dB PBO
Modulation scheme	**64-QAM**		16-QAM	16-QAM	64-QAM	64-QAM	64-QAM	64-QAM OFDM	64-QAM OFDM
Data rate (Gb/s)	**6**	**15**	4	6.7	24	6	15	0.48	4.8
EVM(dB)	**-25.3**	**-24.0**	-18.2	-16.5/-18.1†	-26.6	-26.6	-25.0	-30.5†	-25
ACPR (dBc)	**-29.6**	/	-25*	N.A.	N.A.	-25.4	-20*	-33†	-29.4
P_{avg} (dBm)	**15.9**	**15.0**	16.4	7.2/3.6†	0.6	7.2	11.7	14.3	6.7
PAE_{avg}	**29.1%**	**26.4%**	19.9% CE**	19.8%/11.6%†	5.8%	14.4% CE**	5.75%	25.3%***	11%
Area (mm^2)	**2.87(0.52‡)**		1.85(0.96‡)	2.38(0.18‡)	0.1‡	1.76	1.79(0.59‡)	0.56	0.225‡

*graphically estimated **last stage collector efficiency ***1-stage PA †with digital predistortion ‡core area

Figure 4.8.6: Comparison table with recently reported silicon-based mm-wave digital PAs and analog PAs with PBO efficiency enhancement techniques.

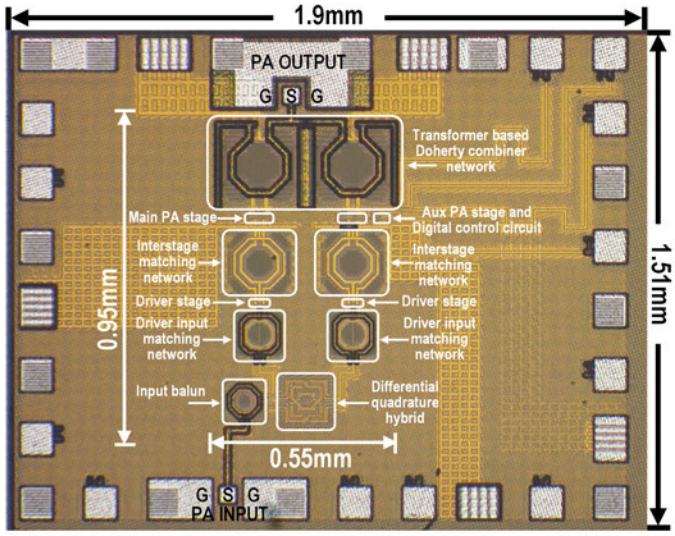


Figure 4.8.7: Die micrograph (1.9mm×1.51mm with a core area of only 0.55mm×0.95mm).

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

4.9 A 60GHz CMOS Power Amplifier with Cascaded Asymmetric Distributed-Active-Transformer Achieving Watt-Level Peak Output Power with 20.8% PAE and Supporting 2Gsym/s 64-QAM Modulation

Huy Thong Nguyen, Doohwan Jung, Hua Wang

Georgia Institute of Technology, Atlanta, GA

Since its invention in early 2000s, Distributed Active Transformer (DAT) has been a popular power-combiner technique for high-power high-efficiency Power Amplifiers (PAs) in voltage-limited Silicon processes at GHz frequencies [1,2]. However, successful DAT implementations at high mm-wave frequencies still remain elusive. Substantial inter-winding capacitance causes asymmetry in impedance transformation among the DAT sections, resulting in mismatched and non-optimum loads for high-frequency mm-wave DAT PAs, including those using transformer-based DAT combiners [3]. Moreover, the DAT physical size at mm-wave frequencies is comparable to a mm-wave signal wavelength, which leads to significant physical phase delays between DAT sections. As a result, the combined vectoral signal at the DAT output is substantially mismatched, causing appreciable degradation in the DAT output power and efficiency [4].

In this work, we present a cascaded asymmetric DAT PA that allows all DAT PA sections to achieve symmetric and optimum load impedance as well as in-phase high-efficiency power combining even with large inter-winding capacitance and a wavelength-comparable DAT combiner. A proof-of-concept design is implemented in a CMOS 45nm CMOS SOI process together with an 8:1 zero-degree combiner, which efficiently combines 48 differential output ports of 24 unit-PAs into one single-ended 50Ω port and delivers 29.1-to-30.1dBm peak output power with 18.5-to-20.8% peak PAE at 60GHz. This watt-level 60GHz CMOS PA also supports linear amplification of wideband modulations, and 2Gsym/s 16-QAM and 64-QAM signals are demonstrated at an average P_{out} of 22.3dBm and 20.9dBm, respectively.

Figure 4.9.1 depicts the general topology of the proposed cascaded asymmetric DAT combiner. It consists of *N*-unit DAT sections, each of which is loaded by an output dual coupled-line combiner and is fed by a phase-shifted input signal. In contrast to conventional symmetric DAT, each DAT section has a different coupler even-/odd-mode impedance and input phase shift, both of which are determined by the DAT section location. The cascaded asymmetric DAT resolves the issues of the conventional DAT as follows. First, as the DAT size is comparable to the wavelength, signals from different DAT PA sections experience various phase delays to travel to output. Proper input phase shifting is needed to align the signals from multiple DAT sections at the total DAT PA output, which results in maximum vectoral addition and high overall DAT combing efficiency. Second, the coupled-line combiner naturally absorbs the inevitable inter-winding capacitance at mm-wave. It also provides inherent differential-to-single-ended combining, obviating the need for additional baluns and ensuring balanced differential load impedance and low loss. Third, due to output combining and inter-winding capacitance, a conventional practical DAT combiner experiences significant active load modulation on each section, resulting in their further load mismatches. To ensure a symmetric and optimum load, each section in our proposed asymmetric DAT employs its output dual coupled-line combiner with a judiciously designed coupling factor and even/odd impedances. In this paper, the theoretical analysis of the proposed asymmetric DAT PA is presented and further verified by EM/device simulations, enabling equation-driven design and demonstration of the 1W 60GHz DAT PA in a 45nm CMOS SOI process.

The design theory of the asymmetric DAT PA is analyzed next (Fig. 4.9.2). First, based on the [Y] matrix of an ideal coupled-line, the [Y] matrix of a unit DAT PA section with its dual coupled-line combiner is derived, considering its differential PA driving and single-ended interaction with adjacent DAT sections. Next, to obtain the load impedance of each DAT PA section, we connect *N* DAT PA sections in cascade and apply the termination conditions, i.e. AC ground on the left-most port, inter-section interaction, and antenna load on the right-most port. To ease practical implementations, we assume all unit PAs are identical and see the same optimum load-line impedance, so that the coupled-line parameters and input phase shifts are solved.

For further comparison, a conventional symmetric DAT at 60GHz with in-phase driving leads to asymmetric and non-optimum PA loads (Fig. 4.9.2). This impedance mismatch further deteriorates when the number of DAT sections (*N*) increases. For *N*=3, the load seen by PA 1 is substantially far away from the load seen by PA 3.

In contrast, with the cascaded asymmetric DAT architecture, the loads of all DAT sections are identical in theory and matched to the optimum load-line impedance at 60GHz (Fig. 4.9.2). Using 3D EM simulation to model asymmetric couplers predicted by the design equations, even including the loss and other non-idealities, the resulting loads for all DAT sections (*N*=3) exhibit close symmetry and balanced differential impedance. Therefore, the design theory of the proposed cascaded asymmetric DAT PA predicts the PA behavior well in practice and enables its "equation-driven" design at high mm-wave frequency.

Different from the conventional DAT with a folded-transformer layout, the proposed structure embraces a straight-line configuration, which, for *N*=3, combines the power from 6 differential output ports of 3 DAT sections to a single-ended port (Fig. 4.9.3). This convenient physical layout allows us to additionally incorporate the cascaded DAT with parallel combiners for further output-power boosting. For example, we employ an 8-to-1 zero-degree combiner to combine 8 cascaded asymmetric DAT outputs. The simulated insertion loss of the 3-section DAT output combiner is 0.9dB, while the zero-degree combiner adds an additional 0.75dB (Fig. 4.9.3). Overall, the proposed asymmetric DAT PA combines the power from 48 differential ports (24 unit PAs) to a single-ended 50Ω load to achieve Watt-level output power.

Figure 4.9.3 depicts the top-level schematic of the proposed DAT PA. The unit PA consists of a common-source driver stage and a cascode PA stage. Transformers are used for input and inter-stage matching. Extra T-lines improve the current gain of the cascode PA, and neutralization capacitors enhance reverse isolation. The asymmetric DAT PA sections demand similar input power with desirable phase shifts. Various techniques can achieve this input power-splitting and phase-shifting. In this work, we employ coupler-based equivalent T-junctions for power splitting and delay T-lines for phase shifting to achieve the 1-to-6 input feedline for the DAT PA.

As a proof-of-concept, the proposed cascaded asymmetric DAT PA is implemented in a 45nm CMOS SOI process and occupies a 3×2.2mm^2 active area. The nominal supply/bias voltage is 2V for the PA stage, 1V for the driver, and 1.3V for the PA cascode gate bias. As shown in Fig. 4.9.4, for the high-gain mode (driver and PA gate bias =0.42V), the PA demonstrates a small-signal gain of 25dB, a 3dB S21 bandwidth from 53 to 65GHz, and a broadband -10dB input matching from 49 to 63GHz. The large-signal continuous-wave (CW) test shows 29.1dBm P_{sat} with 18.5% peak PAE. The -1dB P_{sat} bandwidth is 55 to 64GHz. With a slightly elevated PA/driver supply to 2.2V/1.1V, the PA achieves 30.1dBm P_{sat} with 20.8% peak PAE. For the high-linearity mode (driver and PA gate bias =0.27V), the CW test demonstrates 28.5dBm P_{sat} with 18.9% peak PAE and 26.5dBm P_{1dB} with 15.5% PAE at P_{1dB}.

The PA is tested with 16-QAM and 64-QAM modulated signals at a high-gain mode. Without any digital pre-distortion, the PA achieves -23.6dB EVM with 22.5dBm average P_{out} for an 8Gb/s (2Gym/s) 16-QAM signal, and -27.1dB EVM with 20.9dBm average P_{out} for a 12Gb/s (2Gsym/s) 64-QAM signal.

The series connection of the proposed cascaded asymmetric DAT allows the PA to access a broad physical area to collect power. As a result, the PA demonstrates the highest output power and highest average modulation power compared to all reported silicon PAs at high mm-wave in [3-9] (Fig. 4.9.6).

Acknowledgements:
We would like to thank GlobalFoundries for chip fabrication and members of Georgia Tech GEMS group for their technical supports.

References:
[1] I. Aoki et al., "Fully Integrated CMOS Power Amplifier Design Using the Distributed Active-Transformer Architecture" *IEEE JSSC*, vol. 37, no. 3, pp. 371-383, Mar. 2002.
[2] I. Aoki et al., "Distributed Active Transformer-A New Power-Combining and Impedance-Transformation Technique", *IEEE TMTT*, vol. 50, no. 1, pp. 316-331, Jan. 2002.
[3] U. R. Pfeiffer and D. Goren, "A 23-dBm 60-GHz Distributed Active Transformer in a Silicon Process Technology," *IEEE TMTT*, vol. 55, no. 5, pp. 857-865, May 2007.
[4] K. Wang et al., "A 1V 19.3dBm 79GHz Power Amplifier in 65nm CMOS," *ISSCC*, pp. 260-261, Feb. 2012.
[5] R. Bhat et al., "Large-Scale Power Combining and Mixed-Signal Linearizing Architectures for Watt-Class mmWave CMOS Power Amplifiers," *IEEE TMTT*, vol. 63, no. 2, pp. 703–718, Feb. 2015.
[6] K. Datta and H. Hashemi, "Watt-level mm-Wave Power Amplification with Dynamic Load Modulation in a SiGe HBT Digital Power Amplifier," *IEEE JSSC*, vol. 52, no. 2, pp. 371–388, Feb. 2017.
[7] T. Chi et al., "A 60GHz On-Chip Linear Radiator with Single-Element 27.9dBm Psat and 33.1dBm Peak EIRP Using Multifeed Antenna for Direct On-Antenna Power Combining", *ISSCC*, pp. 296-297, Feb. 2017.
[8] H. Park et al., "Millimeter-Wave Series Power Combining Using Sub-Quarter-Wavelength Baluns," *IEEE JSSC*, vol. 49, no. 10, pp. 2089-2102, Oct. 2014.
[9] H. Wang et al., "Power Amplifiers Performance Survey 2000-present," Accessed on Sept. 10, 2018, [Online].
Available: https://gems.ece.gatech.edu/PA_survey.html

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

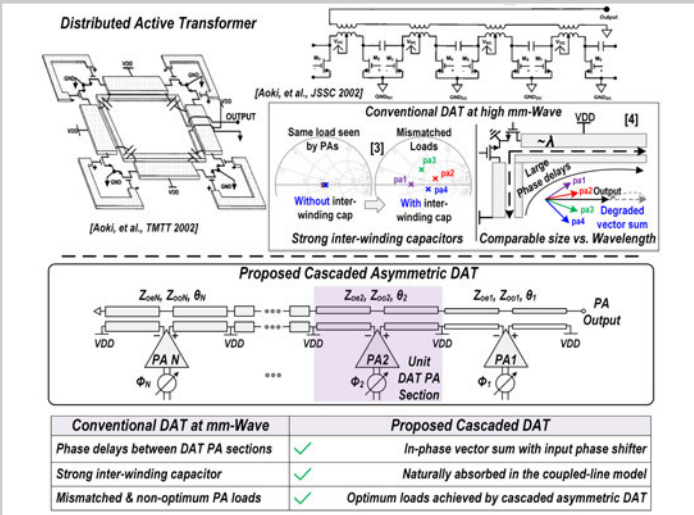


Figure 4.9.1: Conventional symmetric distributed active transformer at high mm-wave frequencies, proposed cascaded asymmetric DAT, and the comparison of design concepts.

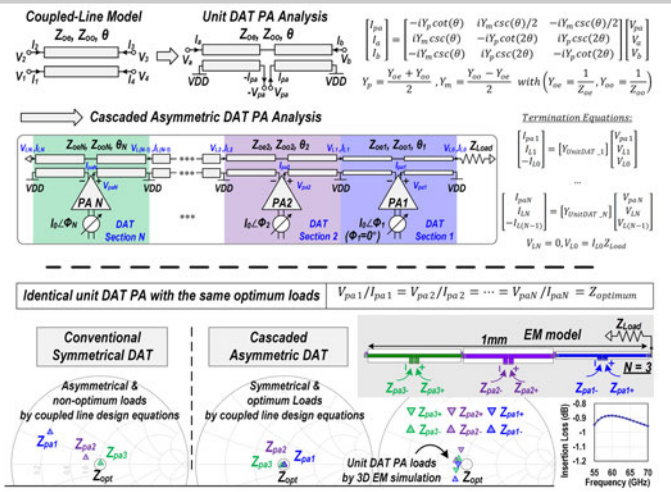


Figure 4.9.2: Equation-driven design of the proposed cascaded asymmetric DAT PA using the coupled-line model, calculated and 3D EM simulated load impedance for a conventional symmetric DAT PA and the proposed asymmetric DAT PA with 3 DAT sections in 45nm CMOS SOI at 60GHz.

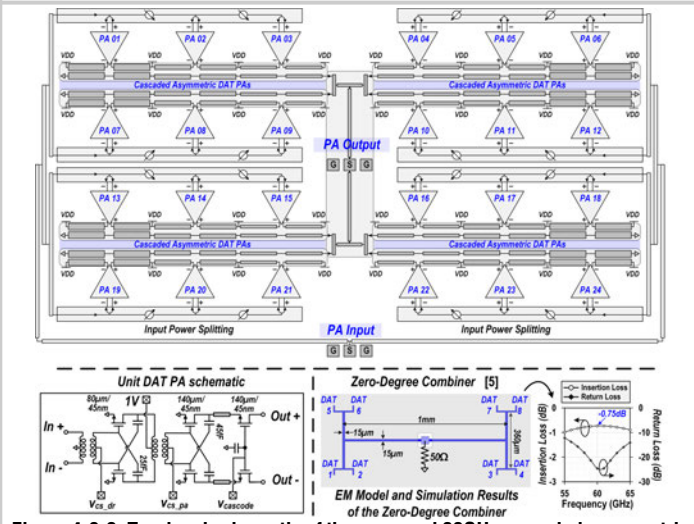


Figure 4.9.3: Top-level schematic of the proposed 60GHz cascaded asymmetric DAT PA with 24 unit DAT PAs, a unit DAT PA schematic, and an 8-to-1 zero-degree PA output combiner.

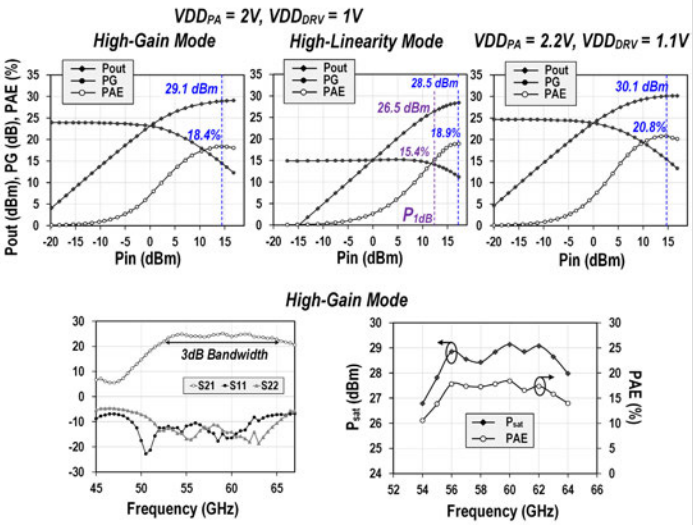


Figure 4.9.4: PA small-signal and large-signal CW performances at 60GHz.

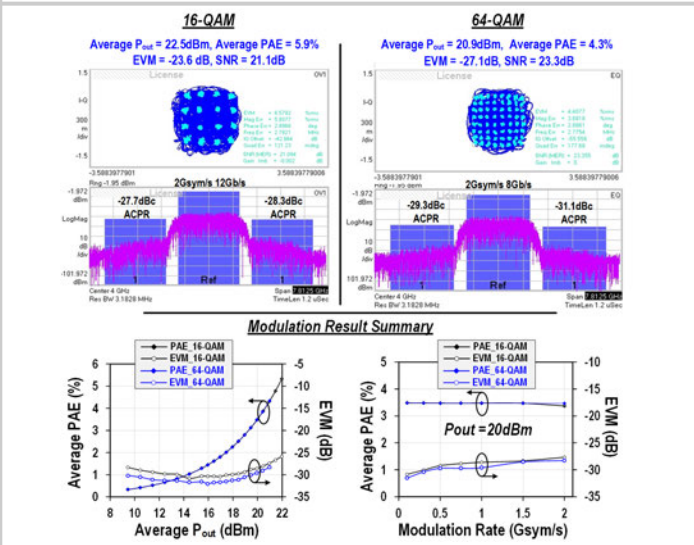
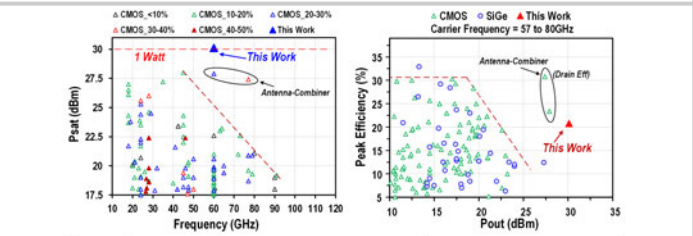


Figure 4.9.5: Dynamic performances of the proposed cascaded asymmetric DAT PA at high-gain mode with 16-QAM and 64-QAM modulated signals.



	This work			mm-wave high-power PAs					60 to 80GHz PAs				
				Datta, JSSC'17	Bhat, TMTT'15	Chi, ISSCC'17	Lin, TMTT'16	Park, JSSC'14	Pfeiffer, TMTT'07	Komijani, JSSC'06	Hsiao, TMTT'13	Zhao, JSSC'12	Khalaf, JSSC'16
Architecture	Cascaded Asymmetric DAT			Stacked PA	Stacked PA	Antenna combiner	Class AB	Triaxial Balun	DAT	Class AB	32-way combiner	4-way 3-stage	Digital Polar Tx
Frequency (GHz)	60			46	42.5	60	76	81	60	77	64	62	60
V_{DD} (V)	2	2	2.2	5	4.8	2	1.8	2.75	4	1.8	1.2	1.8	0.9
Gain (dB)	14.9	23.9	24.7	13	19.4	17.4	19.3	17.5	20	17	16.3	20.6	24.3
P_{sat} (dBm)	28.5	29.1	30.1	28.9	27.2	27.9	27.3	26.7	23	17.5	23.2	20.1	10.8
P_{1dB} (dBm)	26.5	24.7	25.4	–	21*	25	22.3	–	21*	14.5	19.6	19.7	7.4
Peak PAE	18.9%	18.4%	20.8%	18.4%	10.7%	23.4%	12.4%	23.4%	6.3%	12.8%	10%	18%	29.8%
Modulation Data Rate EVM Avg Pout	64-QAM 12Gb/s -27.1dB 20.9dBm		16-QAM 8Gb/s -23.6dB 22.5dBm	–	–	64-QAM 4.8Gb/s -25.4dB 19.3dBm	–	–	–	–	–	–	16-QAM 6.7Gb/s -18.1dB 3.6dBm
Active Area (mm²)	6.6			13.7	4.16	10.5**	6.48	1.06	3.42	0.6	2.04	0.72	<0.18
Technology	45nm CMOS SOI			0.13μm SiGe	45nm CMOS SOI	45nm CMOS SOI	90nm SiGe	0.25μm InP HBT	0.13μm SiGe	0.12μm SiGe	65nm CMOS	0.13μm SiGe	40nm CMOS

* Estimated from reported figures
** Including pads and on-chip antenna

Figure 4.9.6: Comparison to the prior-art mm-wave high-power PA designs and 60-to-80GHz PA designs.

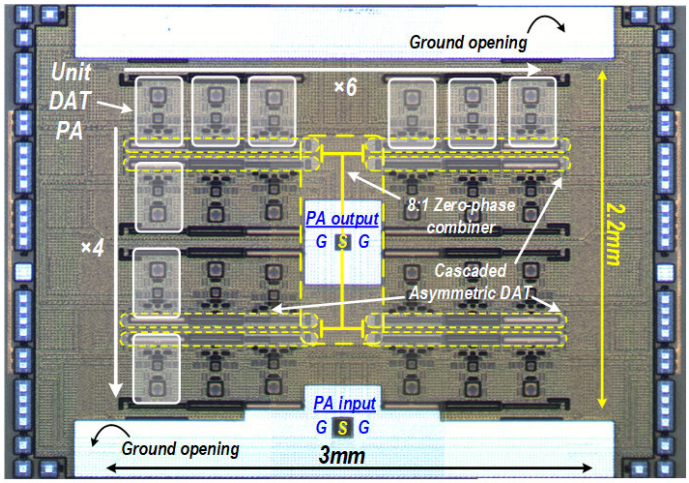


Figure 4.9.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 5 Overview: *Image Sensors*

IMMD SUBCOMMITTEE

Session Chair: ***Kazuko Nishimura***
Panasonic Corporation, Moriguchi, Japan

Associate Chair: ***Jun Deguchi***
Toshiba Memory Corporation, Kawasaki, Japan

Subcommittee Chair: Chris Van Hoof, IMEC, Leuven, Belgium

This session presents advances in image sensors covering 3D-stacked BSI, global shutter, motion/object detection, novel-column ADC architecture, coded exposure, data-compressive imaging, vertical APD, high dynamic range, SPAD, LiDAR and THz imaging. The first paper by SmartSens presents a BSI global shutter with 99% shutter efficiency. This is followed by the University of Michigan who present a low-power image sensor with energy efficient SAR ADCs for IoT applications. Stanford University presents a 127pJ/pixel image sensor for HOG-based object detection. A VGA CMOS image sensor with time-stretched single-slope ADCs is presented by Yonsei University, while the University of Toronto presents a 2-tap coded-exposure image sensor with a tap contrast ratio of 99% at 180fps. Panasonic presents a CMOS image sensor with vertical avalanche photodiodes (VAPDs) with "relaxation-quenching". The University of Edinburgh presents a stacked SPAD array for LiDAR applications. Finally, Hokkaido University presents a THz image sensor with pixel-parallel ADCs.

1:30 PM

5.1 A Stacked Global-Shutter CMOS Imager with SC-Type Hybrid-GS Pixel and Self-Knee Point Calibration Single-Frame HDR and On-Chip Binarization Algorithm for Smart Vision Applications

C. Xu, SmartSens Technology, San Jose, CA

In Paper 5.1, SmartSens presents a global shutter BSI 3D-stacked 1.3Mpixel 1/4" CMOS image sensor. High dynamic range of 110dB and low dark current are achieved at 99% shutter efficiency operating at 120fps consuming 205mW.

2:00 PM

5.2 Energy-Efficient Low-Noise CMOS Image Sensor with Capacitor Array-Assisted Charge-Injection SAR ADC for Motion-Triggered Low-Power IoT Applications

K. D. Choo, University of Michigan, Ann Arbor, MI

In Paper 5.2, the University of Michigan presents a low-power image sensor for IoT applications using a 65nm CIS process to implement 1.5μm 4T pixels along with a column-parallel capacitor array-assisted charge-injection SAR ADC. The energy efficient SAR ADC architecture achieves a noise floor of 226μV_{rms} or 106μV_{rms} using multiple sampling, and facilitates power-efficient motion detection.

2:30 PM

5.3 A Data-Compressive 1.5b/2.75b Log-Gradient QVGA Image Sensor with Multi-Scale Readout for Always-On Object Detection

C. Young, Stanford University, Stanford, CA

In Paper 5.3, Stanford University presents an image sensor with 127pJ/pixel for HOG-based object detection. Captured images are compressed to 1.5b and 2.75b logarithmic gradients to reduce backend data by 25× compared to an 8b image. Despite the low-resolution logarithmic gradients, the robustness to illumination is comparable to images with 9b linear gradients.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

3:15 PM

5.4 A 76mW 500fps VGA CMOS Image Sensor with Time-Stretched Single-Slope ADCs Achieving 1.95e⁻ Random Noise

I. Park, Yonsei University, Seoul, Korea

In Paper 5.4, Yonsei University presents a VGA CMOS image sensor with a frame rate of 500fps, a random noise level of $1.95e^{-}_{rms}$ and power consumption of 76mW using 110nm 1P4M CMOS process. A low figure of merit of $0.96e^{-}\cdot nJ$ is achieved with 8× analog gain and time-stretched single-slope ADCs. A horizontal conversion time of only 4μs is facilitated using a low-rate 100MHz counter clock.

3:30 PM

5.5 Dual-Tap Pipelined-Code-Memory Coded-Exposure-Pixel CMOS Image Sensor for Multi-Exposure Single-Frame Computational Imaging

R. Gulve, University of Toronto, Toronto, Canada

In Paper 5.5, the University of Toronto presents a 2-tap coded-exposure-pixel image sensor for computational imaging application with a tap contrast ratio of 99% at 180sub-fps. A small pixel size of 11.2μm is realized with a fill factor of 45.3%. The sensor is demonstrated in single-frame structured-light and photometric-stereo applications.

3:45 PM

5.6 A 400×400-Pixel 6μm-Pitch Vertical Avalanche Photodiodes CMOS Image Sensor Based on 150ps-Fast Capacitive Relaxation Quenching in Geiger Mode for Synthesis of Arbitrary Gain Images

Y. Hirose, Panasonic, Nagaokakyo, Japan

In Paper 5.6, Panasonic presents a VAPD CMOS image sensor. A new method of "relaxation-quenching" enables operation with variable gain from unity to 80000 with fast and clean (after-pulse-suppressed) quenching. Synthesis of gradation and direct ToF binary images are demonstrated under 100Klux sunlight.

4:15 PM

5.7 A 256×256 40nm/90nm CMOS 3D-Stacked 120dB-Dynamic-Range Reconfigurable Time-Resolved SPAD Imager

R. K. Henderson, University of Edinburgh, Edinburgh, United Kingdom

In Paper 5.7, the University of Edinburgh presents a 256×256 array of stacked SPADs at 9.2μm pitch for high-dynamic-range photon counting with multiple reconfigurable modes useful for ToF imaging. Using a 40nm/90nm 3D-stacked BSI process, 4096 14b TDCs with time resolution of 35ps are integrated into the bottom-tier CMOS to facilitate the processing of the SPAD array.

4:45 PM

5.8 A 32×32-Pixel 0.9THz Imager with Pixel-Parallel 12b VCO-Based ADC in 0.18μm CMOS

S. Yokoyama, Hokkaido University, Sapporo, Japan

In Paper 5.8, Hokkaido University presents a 32×32 CMOS THz image sensor using 0.18μm CMOS for low cost. A responsivity of 218kV/W at 0.93THz, NEP of 91pW/μHz at 31Hz, and 400fps image capturing with digital CDS are achieved by envelope detectors with a square-law-based amplifier and pixel-parallel VCO-based ADCs.

5.1 A Stacked Global-Shutter CMOS Imager with SC-Type Hybrid-GS Pixel and Self-Knee Point Calibration Single-Frame HDR and On-Chip Binarization Algorithm for Smart Vision Applications

Chen Xu[1], Yaowu Mo[1], Guanjing Ren[2], Weijian Ma[2], Xin Wang[1], Wenjie Shi[2], Jinjian Hou[2], Ke Shao[2], Haojie Wang[2], Pengge Xiao[2], Zexu Shao[2], Xiao Xie[1], Xiaoyong Wang[1], Chris Yiu[1]

[1]SmartSens Technology, San Jose, CA
[2]SmartSens Technology, Shanghai, China

Request for smart vision related applications, such as face identification, VR/AR, gesture recognition, 3D imaging, and artificial intelligence (AI), has driven demand for high-performance global-shutter (GS) sensors. Most commercially available GS sensors use a charge-domain storage gate implementation, which suffers from serious light leakage and leads to lower shutter efficiency. This situation worsens when using a BSI fabrication process [1]. In addition, the traditional frame-based or line-based HDR method utilizing multiple exposures adds motion artifact to fast-moving objects, which defeats the purpose of having a global shutter. Moreover, some smart vision applications such as QR 2D barcode scanners and 3D facial recognition with structured light method need image sensors to "read" a certain pattern and "understand" the information within. However, image sensors usually capture a full image that needs to be further transferred to and processed by a companion SoC. Higher resolution and increased complexity of the target pattern pose a growing challenge to transfer and process the entire image at real time, also the required high power consumption lowers handheld device's battery life.

To address these issues, we present a GS image sensor fabricated in TSMC 40nm/65nm stacked backside-illumination (BSI) CMOS process. Figure 5.1.1 is a schematic diagram illustrating the switched-capacitor-type hybrid GS pixel with selectable rolling shutter and global shutter readouts and dynamic range enhancement through self-calibrating knee point HDR. The operation of this hybrid GS is similar to the voltage-domain GS as depicted in [2], except the drain of SF1 is connected a controlled signal *Vrsf* and there is no bias transistor to the first source follower SF1. The advantages of this configuration are that power consumption is significantly lower during the GS operation and image quality is improved. Since the true CDS operation is performed by incorporating capacitors and switches, a good PRNU of 0.6% at 50% of full-well and low random noise performance of 3.5e- rms can be achieved. Furthermore, reset and signal are stored at the storage capacitors *Csig* and *Crst* after the GS operation. Since it is a voltage value instead of charges stored in the GS operation, no light leakage during readout occurs. Thus a high shutter efficiency of 20000:1 can be achieved even with a BSI process. There is also no excess dark-current leakage path in the pixel compared to the traditional charge-domain GS pixel leading to a good dark-current performance of 200 e-/s @ 80°C.

A high-density MIM process is used to form the storage capacitors *Csig* and *Crst* and help to reduce the random noise. With BSI technology, a MIM capacitor is located at the backside, thus a higher fill factor of 80% and a better sensitivity of 5000mV/lux·s can be achieved compared to a charge-storage-type GS with the same pixel size. Furthermore, beside the native GS mode, the pixel also supports a selectable rolling-shutter mode. It may readout an image signal in a rolling-shutter mode through *Rs_sel* as *rs_Pix* or in a GS mode through *Gs_sel* as *gs_Pix* for flexibility.

As shown in Fig. 5.1.1, one of the key pixel designs is to support self-calibrating knee point HDR denoted by *Vrab Mode_sel*. *Vrab Mode_sel* provides a bimodal selection of either connecting bias *Vhi* or *Vlo* to the drain of reset transistor *Rst*. *Vrab Mode_sel* provides either connecting *Vhi* or *Vlo* to the drain of the reset transistor, resulting in reduction of any transfer transistor threshold voltage variation impact on resulting imaging array fixed pattern noise (FPN) in the HDR mode.

The chip architecture and readout circuitry block diagram are shown in Fig. 5.1.2. The chip uses a stacked technology and comprises a sensor die and an ASIC die. Hybrid bond points are used between the sensor and ASIC die for interconnections. The readout architecture utilizes a column-parallel single-slope ADC architecture to reduce power and improve noise performance, similar to [3]. The comparator is a differential input and single-ended output 3-stage configuration with programmable gain control at the input of the comparator. A programmable noise reduction block is inserted between 1st and 2nd stage of the comparator for better noise and speed optimization. In the counter/memory readout block, it uses a 13b counter for fast counting speed up to 2G counts/s and a multiple channel DFFR sequential memory readout configuration to reduce the noise interference to other analog blocks, thus achieves a better noise performance.

The principle of the single frame HDR is explained in Fig. 5.1.3. During the integration of the photodiode, the TX gate signal is controlled by a pulse signal as shown in Fig. 5.1.3. The dynamic range of the photodiode is extended since the excess charge will be drained away by lowering the barrier of TX transistor and thus a nonlinear photo response can be achieved. Since this single-frame HDR function is performed simultaneously in each pixel across the entire array, there is no motion artifact for fast-moving object capturing in HDR mode.

The origin of the FPN issue at the knee point is also illustrated in Fig. 5.1.3. The bottom curves show the upper line segments for the three pixels to be displaced from one another. This displacement is due to the variation in the respective transfer transistor gate electrode threshold voltages for the three pixels illustrated. To reduce FPN at the knee point, a self knee point calibration (SKPC) method is designed. The control timing is shown in Fig. 5.1.4. The barrier voltage level information is readout row by row through RS mode and then a dedicated recover algorithm is applied to the original GS HDR image data and knee point data. It normalizes the photo transfer curves into one unified response curve (dash line) as shown in Fig. 5.1.3. The accuracy of the result at the knee point with in-pixel SKPC on and off is also shown in Fig. 5.1.4. The standard deviation of the image before and after the calibration is 5.8 and 1.6 rms respectively, indicating the pixel FPN at the knee point can be effectively suppressed. Since the SKPC operation needs one row time to calibrate the knee point and readout, so the frame rate is reduced when SKPC is enabled. The sensor can achieve a frame rate of 120fps without SKPC and 60fps with SKPC.

A digital algorithm at the forefront of the pixel array improves image processing efficiency and saves power consumption of the entire system, since only necessary information is readout instead of the full image [4]. The stacked technology enables advanced digital circuitry with more image processing power. In this design, we use a 65nm digital process node for the ASIC die. In this ASIC die, beside the readout circuitry, voltage reference, interface circuit and digital core block for the basic sensor function, a pre-processing digital block performing an advanced binarization algorithm is implemented. The final sensor output is a 1b binary data and contains all the information necessary for structure light 3D and QR 2D barcode applications, instead of 10/12b full-resolution RAW imaging data. As a result, recognition efficiency can be effectively improved and power consumption of the system is greatly reduced.

The sensor is fabricated in a TSMC 40nm/65nm stacked BSI CMOS process. The performance summary and key parameter comparison with existing literature are included in Fig. 5.1.5. A structured light 3D demo system is setup to demonstrate the sensor function and performance. The image with a structured light pattern emitter is captured and shown in raw format, after on-chip binarization algorithm format and reconstructed 3D depth map format, as shown in Fig. 5.1.6. Top images are without HDR and bottom are with single frame HDR for comparison. The result shows a clear advantage for extending the usable range of light condition with the specific applications. The micrograph of the sensor is shown in Fig. 5.1.7 for both sensor and ASIC die. As a conclusion, the global shutter sensor presented in this paper can address many existing issues and is suitable for smart vision applications, i.e., face identification, machine vision, facial recognition, 3D imaging and AI, etc.

Acknowledgements:
Thanks to Taiwan Semiconductor Manufacturing Co. for fabricating the chip and provide the technical support for the process development of HD MIM technology.

References:
[1] K. Yasutomi, et al., "A 2.7e- Temporal Noise 99.7% Shutter Efficiency 92dB Dynamic Range CMOS Image Sensor with Dual Global Shutter Pixels," *ISSCC Dig. Tech. Papers*, pp. 398-399, Feb. 2010.
[2] G. Meynants, et al., "Backside Illuminated Global Shutter CMOS Image Sensors," *Int. Image Sensor Workshop*, pp. 305-308, June 2011.
[3] Y. Nitta, et al., "High-Speed Digital Double Sampling with Analog CDS on Column Parallel ADC Architecture for Low-Noise Active Pixel Sensor," *ISSCC Dig. Tech. Papers*, pp. 2024-2031, Feb. 2006.
[4] T. Yamazaki, et al., "A 1ms High-Speed Vision Chip with 3D-Stacked 140GOPS Column-Parallel PEs for Spatio-Temporal Image Processing," *ISSCC Dig. Tech. Papers*, pp. 82-83, Feb. 2017.
[5] T. Geurts, et al., "A 25 Mpixel, 80fps, CMOS Imager with an In-Pixel-CDS Global Shutter Pixel," *Int. Image Sensor Workshop*, pp. 7.02, June 2015.
[6] L. Stark, et al., "Back-Illuminated Voltage-Domain Global Shutter CMOS Image Sensor with 3.75μm Pixels and Dual In-Pixel Storage Nodes," *IEEE Symp. VLSI Tech.*, pp. 1-2, June 2016
[7] Y. Oike, et al., "8.3 M-Pixel 480-fps Global-Shutter CMOS Image Sensor with Gain-Adaptive Column ADCs and Chip-on-Chip Stacked Integration," *IEEE JSSC*, vol. 52, no. 4, pp. 985–993, Apr. 2017.
[8] M. Sakakibara, et al., "A Back-Illuminated Global-Shutter CMOS Image Sensor with Pixel-Parallel 14b Subthreshold ADC," *ISSCC Dig. Tech. Papers*, pp. 80-81, Feb. 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

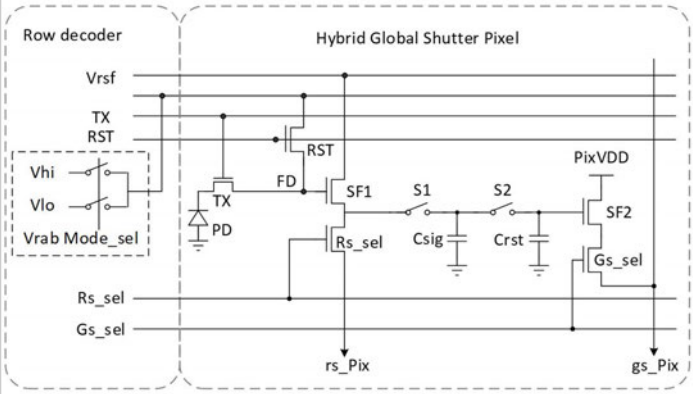


Figure 5.1.1: Schematic of the switched-capacitor-type hybrid global-shutter pixel.

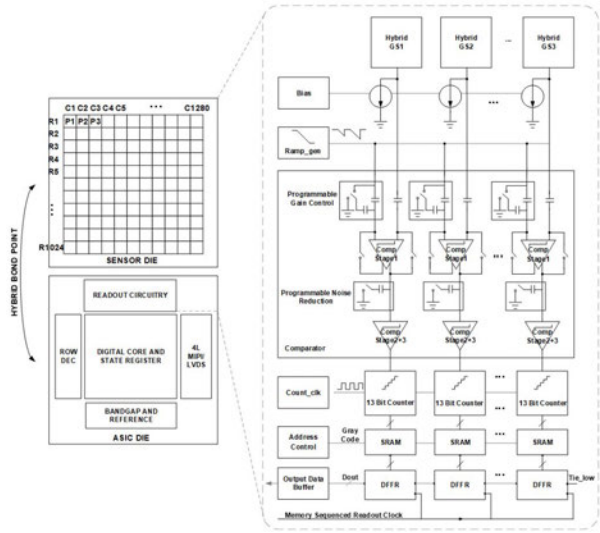


Figure 5.1.2: Block diagram of the chip architecture and column-readout circuitry.

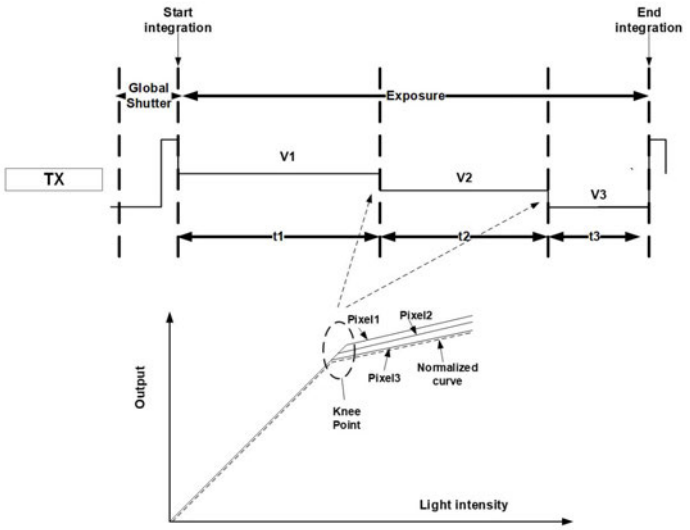


Figure 5.1.3: Control timing used to achieve extended dynamic range of the pixel, corresponding HDR photo transfer curve, and knee point FPN issue.

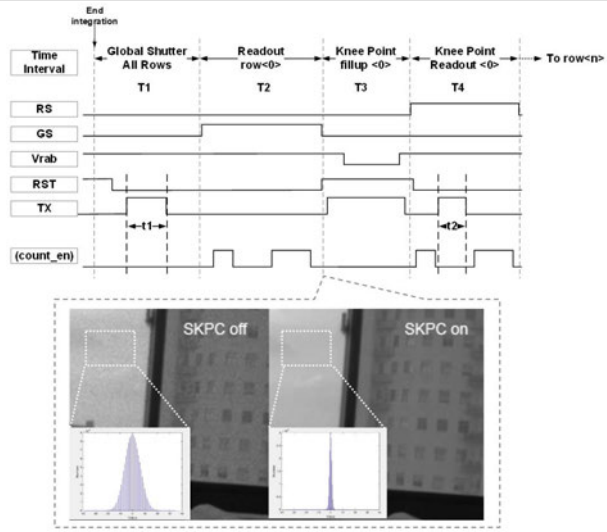


Figure 5.1.4: Control timing to achieve a self-knee point calibration (SKPC) during the readout period and results comparison.

Parameters	Value
Pixel size	2.7um x 2.7um stacked BSI
Resolution	1080(H) x 1280(V)
Optical Size	1/4"
HDR	Single Frame HDR with SKPC
Frame Rate	1080x1280 @ 120fps 10bit
Output	RAW @ 4L MIPI/LVDS
SNRmax	40dB
Dynamic Range	110dB with HDR
Sensitivity	5000mV/lux*s
PRNU	0.6% at 1/2 of fullwell
Package	CSP, RW
Power	205mW @1080x1280 120fps with on-chip binarization algorithm
550nm-QE	95%
NIR-QE	58% @ 850nm, 36% @ 940nm
CRA	25°
DC	200 e-/s @ 80°C
FWC	10K e-
RN	3.5 e-
Shutter Efficiency	20000:1

	This work	2015 ISW [5]	2016 VLSI [6]	2017 JSSC [7]	2018 ISSCC [8]
Pixel Size	2.7um stacked BSI	4.5um FSI	3.75um BSI	5.86um FSI	6.9um stacked BSI
Frame Rate	120fps	80fps	50fps	480fps	660fps
Dynamic Range	110dB	-	102dB	76.3dB	70.2dB
Random Noise	3.5 e-	12 e-	8.5 e-	4.6 e-	5.15 e-
Shutter Efficiency	20000:1	5000:1	3000:1	96000:1	5600:1
FWC	10k e-	10k e-	8.1k e-	30.5k e-	16.6k e-
Sensitivity	5000 mV/lux·s	-	-	~2500 mV/lux·s	3690 mV/lux·s
HDR	Yes	No	Yes	No	No
FoM = (Power*Noise)/(Res*FPS) (e-.nJ)	4.33	22.89	-	6.04	3.98
GS Type	Voltage Domain	Voltage Domain	Voltage Domain	Charge Domain	Charge Domain

Figure 5.1.5: Performance summary table and comparison of key parameters.

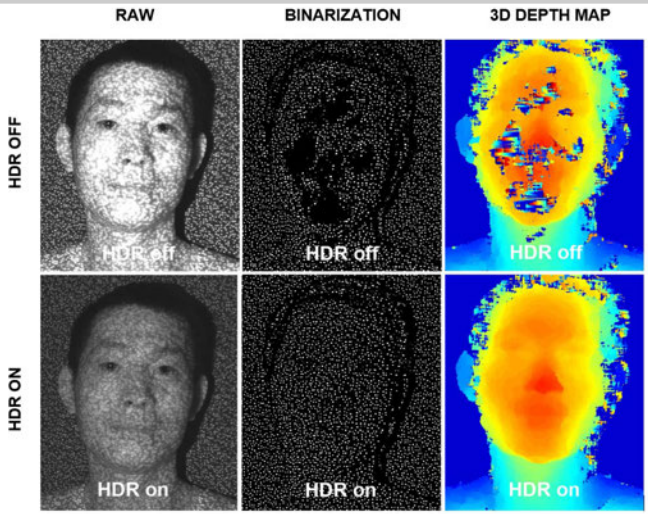


Figure 5.1.6: A structured-light facial image shown in raw format, after on-chip binarization algorithm format and reconstructed 3D depth map format. Top is without HDR and bottom is with single frame HDR for comparison.

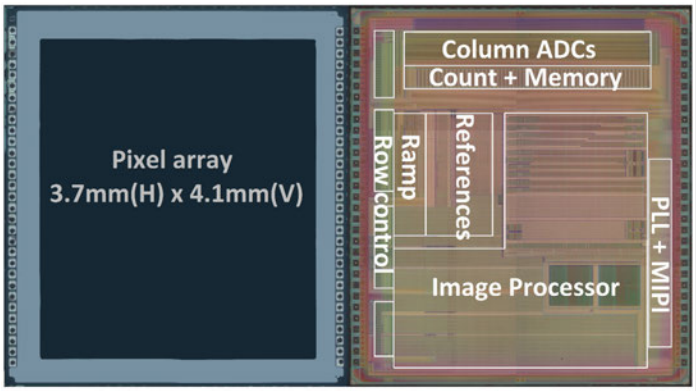


Figure 5.1.7: Chip micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

5.2 Energy-Efficient Low-Noise CMOS Image Sensor with Capacitor Array-Assisted Charge-Injection SAR ADC for Motion-Triggered Low-Power IoT Applications

Kyojin D. Choo, Li Xu, Yejoong Kim, Ji-Hwan Seol, Xiao Wu, Dennis Sylvester, David Blaauw

University of Michigan, Ann Arbor, MI

As IoT is increasingly integrated into our everyday life, the demand for different sensor modalities, especially imaging, is rising. IoT imagers are often compact and have small form-factor batteries and thus must be designed with both low power (improving battery life) and high image quality (maximizing utility). Previously reported sensors [1,2], along with this work, adopt motion-detection (MD) triggering of full-array capture where MD is performed on a heavily subsampled frame to enable continuous low-power operation. MD limits energy-hungry full-array captures to cases where activity is detected. To further reduce power consumption, the full-frame capture energy itself needs to be addressed, which is typically dominated by the ADC.

Conventional imagers use a single-slope ADC with 4T pixel structure [1], which provides high fidelity but is inherently energy inefficient. This is particularly the case at higher resolution due to its exponential energy scaling with bit-width or when used in conjunction with sub-sampling. SAR-based ADCs have superior energy efficiency and are used extensively in other application areas. However, to date SAR ADCs have been rarely used in imagers due to their large size and limited accuracy. To achieve both high quality and low power, we propose an energy-efficient, low-noise capacitor array-assisted charge-injection SAR ADC (c-ciSAR) structure. The structure merges an area-efficient charge-injection cell (ci-cell) [3] with a small capacitor array structure to extend its performance to 10b (Fig. 5.2.1). The merged c-ci structure achieves significantly higher ADC energy efficiency compared to a single-slope ADC, with a measured ADC FoM of 14.4μV·nJ (full capture mode) and 90.4μV·nJ (MD mode).

In addition, for MD, we propose an *n*-way side-step mapping of *n* motion pixel rows that re-uses the in-column hardware of the full-adder (FA) in the SAR logic (Fig. 5.2.1). This allows MD without transferring the motion pixel values to external storage for motion frame differencing. Unlike prior in-pixel analog MD [2], which degraded pixel performance by -25dB compared to typical 4T pixels, the MD approach maintains an un-modified conventional 4T pixel structure for high image quality. The complete image sensor is implemented in a 65nm CIS process with VGA active resolution and achieves 70.1μJ/full-array or 167.5pJ/frame/pixel, which is 2.4× lower than recent prior art [1], and consumes 1.7μJ/MD-frame (64.7× lower) in continuous MD mode.

The MD pixels are subsampled every n^{th} column and i^{th} row. For row *r*, an MD pixel at coordinate ($r \bullet i$, $c \bullet n$), where $c = \{0, 1, \ldots\}$, is re-routed sideways to column ($c \bullet n + r$) for data retention and in-column MD computation. As readout progresses vertically across subsampled MD rows, the mapped ADCs shift horizontally, and all MD pixels are mapped to column ADC. Both the storage of the MD pixel values and the differencing of values between MD frames reuse existing hardware of the c-ciSAR, thereby incurring minimal area overhead. The subsample stride *(n, i)* and resulting MD frame size *(h, v)* are configurable with the constraint that $h \bullet v < k$, where *k* is the total number of full active-array columns (640 for VGA). Skipped pixels can be enabled and binned together to minimize blind spots using three methods: floating diffusion (FD) binning on 2 shared Gr-Gb pixels, source-follower (SF) binning across horizontal/vertical direction, and digital binning in the column logic. For our evaluation case, $n = 20$ and $i = 24$, resulting in an MD frame resolution of 20×32 pixels.

The c-ciSAR ADC structure uses a c-ciDAC to create a local reference voltage, which is compared to the pixel voltage (Fig. 5.2.2). The comparator uses a StrongARM structure supplemented with an NMOS-input 5T-OTA placed as a pre-amplifier. Auto-zeroing (DC decoupling) capacitors are applied at both input branches of the pre-amplifier. Three capacitors (8×, 1×, 1×) on the DAC side provide coarse/fine pathways that transfer and sum charge-injection-based voltage changes at PDC (*pull-down coarse)*, PDF (*pull-down fine)* and PUF (*pull-up fine)* at *i_DAC*.

For each ci-cell (Fig. 5.2.2, right), the *Inject bar* signal pulses low to transfer constant charge from output (*Out*) and induces a repeatable change in its voltage. During each injection, the voltage on node *i_cap* rises from its reset value (*Gnd*) to the voltage controlled by node *Bias*, whose value sets the charge-injection amount. The cascode configuration of *M2* and the direction steering switches (*M1* and *M1d*) greatly improve the output impedance of the ci-cell, enabling 10b linearity. We pre-charge the output capacitor (C_{inj}) to a sufficiently high voltage (3.4V in our chip) during auto-zero phase to ensure that *M1* and *M2* operate in saturation mode throughout the ADC sequence. The dummy injection path makes the current draw from the power supply decision-independent, which ensures a constant injection charge even for high impedance power supply regulation. This enables low power overhead supply regulation, which is often needed in battery-operated devices.

Ci-cells are replicated 8 times per path to provide 3b of resolution and are re-activated multiple times to further refine resolution. Each set has a 7b calibration capacitor C_{cal}. There are only two types of ci-cells (up, down), reducing design complexity. Also, the coarse/fine capacitor array reduces the number of ci-cell activations by 4.3× for 10b ADC operation, enabling compact and energy-efficient DAC design.

Figure 5.2.3 shows the c-ciSAR ADC operation for one row. We perform dual-CDS, both in the analog domain. The first is performed during auto-zeroing. The second is performed by a non-binary search with fine ci-cells that match the DAC output (*i_DAC*) voltage to the *i_pixel* voltage. This is followed by coarse and fine searches to quantize the pixel signal value. Unlike a conventional capacitive DAC, the c-ciDAC does not require the fine CDS search code to be remembered and subtracted from the final signal code, simplifying ADC operation.

The signal value quantization first uses a coarse linear search (4b) instead of a binary search, which offers three advantages: 1) smaller and regular changes in DAC voltage allow for the use of a slow, low-power pre-amplifier; 2) the pull-up coarse (PUC) path can be omitted; and 3) noisy PDC operation is avoided at darker light levels where its higher DAC noise (due to its high gain) impacts image quality more significantly (Fig. 5.2.3). This is unlike a binary search, which involves noisy PDC/PUC operations regardless of light level. To further this improvement, the DAC value is shifted lower with PDF operation before entering the coarse linear search to avoid any PDC operations for the dark level.

Next, we perform a 7b fine binary search with a 1b redundant range where the DAC voltage is again offset to the center of the coarse search step to maximize overlap. Unlike a conventional capacitive SAR ADC, a non-binary search can be easily programmed since the sequence of operations is not restricted by the hardware configuration. Another advantage of this flexibility is the option to perform multiple sampling on the fine LSB to reduce noise. We implemented two types of multiple sampling (Fig. 5.2.3): 1) repeat the LSB decision cycle for MS1 times, and/or 2) fire the comparator decisions MS2 times within a cycle and perform majority voting. When the noise level is comparable to the LSB size, using the first approach is both effective and energy efficient.

Figure 5.2.4 shows the MD mode procedure and column logic, which has four unit operations: 1) arbitrary number addition/subtraction, 2) inversion, 3) absolute value, and 4) storage. Using these operations, we store the MD pixels in-column and detect those with significant changes in their code. The controller gathers and accumulates the MD flags (sign-bit of difference result) along with the frame values to determine whether to proceed to full capture and to adjust exposure value. The sensor is designed primarily with thick-oxide transistors to enable a low-power standby mode; column logic circuits use standard thin-oxide devices but are power gated.

Figure 5.2.5 shows captured images from full-array and MD operation, while Fig. 5.2.6 provides a comparison with other recent works with MD capability or SAR ADCs. The energy efficiency of the complete sensor is 37.9μV·nJ (ADC FoM). However, this number is dominated by the fully programmable controller (for flexible testing purposes); a simple hardwired state machine could be used in a more lightweight implementation. Excluding the controller energy consumption and evaluating only the readout energy (pixel, ADC and stream-out), E/frame/pix is 63.6pJ, and the ADC FoM is 14.4μV·nJ. The E/frame/pix in MD mode is 230.5× smaller than that of a single-slope sensor [1]. Also, the sensor can operate with its on-chip PMU, which generates internal voltages from a single external 2.5V battery, enabling integration into an IoT sensor. Figure 5.2.7 provides the chip photograph; the die area is 16.7mm^2.

References:

[1] O. Kumagai, et al., "A 1/4-inch 3.9Mpixel Low-Power Event-Driven Back-Illuminated Stacked CMOS Image Sensor," *ISSCC Dig. Tech. Papers*, pp. 86-87, Feb. 2018.

[2] G. Kim, et al., "A 467nW CMOS Visual Motion Sensor with Temporal Averaging and Pixel Aggregation," *ISSCC Dig. Tech. Papers*, pp. 480-481, Feb. 2013.

[3] K. Choo, et al., "Area-Efficient 1GS/s 6b SAR ADC with Charge-Injection-Cell-Based DAC," *ISSCC Dig. Tech. Papers,* pp. 460-461, Feb. 2016.

[4] S. Ji, et al., "A 220pJ/Pixel/Frame CMOS Image Sensor with Partial Settling Readout Architecture," *IEEE Symp. VLSI Circuits*, pp. 226-227, June 2016.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

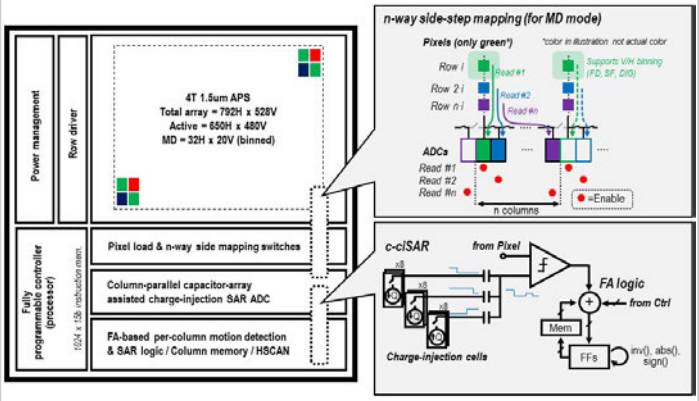


Figure 5.2.1: Overall sensor organization. n-way side-step mapping is used to enable low-power motion detection (MD) operation. Capacitor array-assisted charge-injection SAR ADC (c-ciSAR) is proposed to improve energy efficiency of the overall sensor.

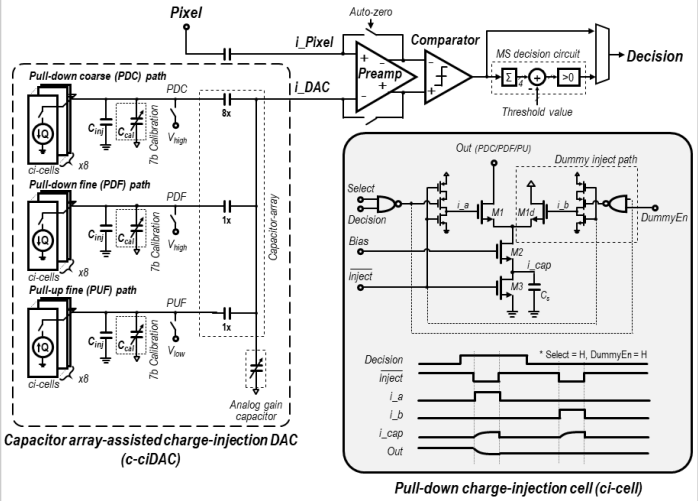


Figure 5.2.2: Detailed circuit diagram of capacitor array-assisted charge-injection SAR ADC (c-ciSAR). Capacitor array complements the charge-injection cell (ci-cell) to enable compact and energy efficient 10b operation.

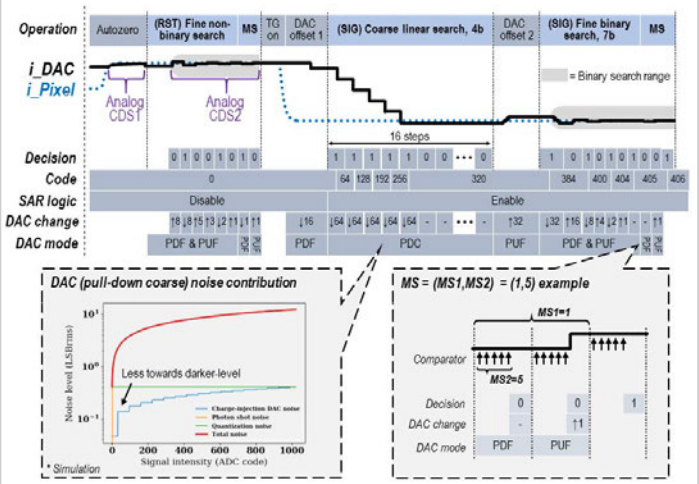


Figure 5.2.3: c-ciSAR row-conversion sequence. Programmability of c-ciSAR allows the ADC to perform not only binary search but also non-binary or linear search. Using appropriate search methods, the DAC noise contribution is minimized at dark level. Multiple-sampling mode is implemented.

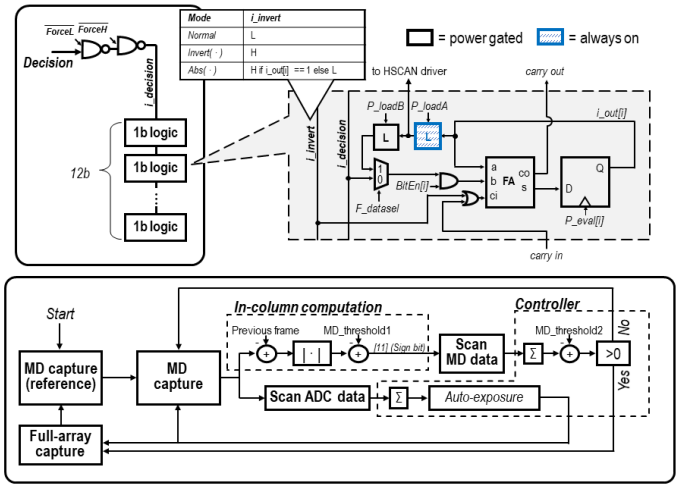


Figure 5.2.4: Circuits and algorithms for in-column motion evaluation. FA-based SAR logic is reused for in-column motion evaluation. Motion flags generated in-column are aggregated by the controller and used in auto-exposure adjustment and to trigger full-array capture.

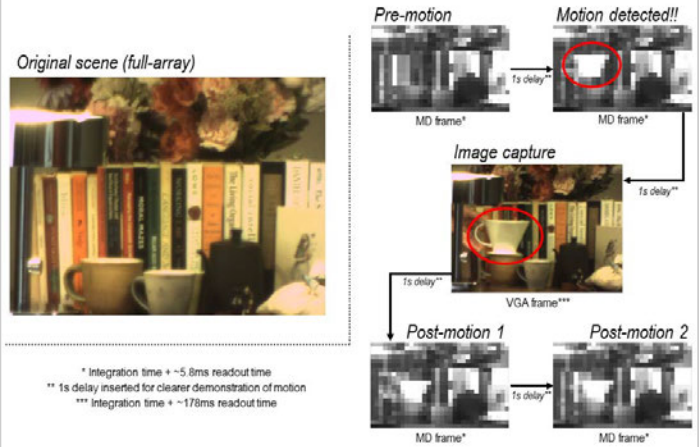
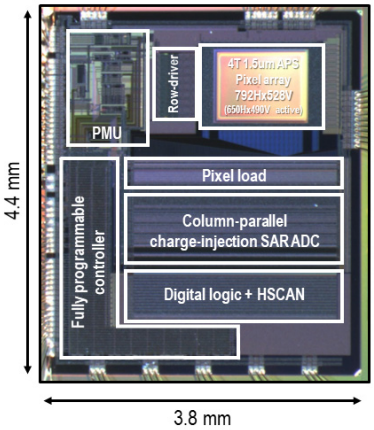


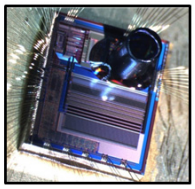
Figure 5.2.5: Captured images from full-array mode and motion detection (MD) mode. The sequence of images (right) shows the sensor in MD mode successfully triggering a full-array capture when motion (red circle) is detected.

		This work			Kumagai ISSCC 2018 [1]	Kim ISSCC 2013 [2]	Ji VLSI 2016 [4]
Technology		65 nm CIS			90 nm CIS + 40 nm	130 nm	180nm
Pixel structure		4T 1.5 um			4T 1.5 um	3T 6.4 um	3T
Full array resolution		792 x 528 = 418.1 K (Active=VGA)			2560 x 1536 = 3.93 M	128 x 128 = 16.3 K	320 x 240 = 76.8 K
MD mode support (& MD resolution)		O (32 x 20 = 640)			O (16 x 5 = 80)	O (48 x 16 = 778)	X
Readout circuit	ADC type	Capacitor-array assisted charge-injection SAR ADC			Single-slope ADC	Single-slope ADC	SAR ADC
	Resolution	10 b			10 b	8 b	10 b
	Full signal range	500 mV			435 mV	-	-
	ADC noise [uVrms]	Condition	Meas.	Q. adj*4	Cond. / Meas / Q. adj		
		no MS, AG*1≈7.7 dB	226*3	211	AG=18 dB / 100.4 / 98	-	450 (w/ pixel noise)
		MS=(4,1)*2, AG=7.7dB	180	161			
		MS=(1,5), AG=7.7 dB, I_{amp}↑	136	108			
		MS=(2,15), AG=7.7 dB, I_{amp}↑	106	68			
Full capture mode	Dynamic range	64.3 dB			67 dB	38.5 dB	-
	Max frame-rate	5.6 fps			60 fps	19 fps	130 fps
	@ dark condition, no MS, PMU off	Readout only (ADC + pixel)	Sensor (w/ controller)		Sensor	Sensor	Readout only (ADC + pixel)
	Energy / frame [uJ]	**26.6**	70.1		1583	1.5	16.9
	Energy / frame / pix [pJ]	**63.6**	167.5		400	610	220
	ADC FoM [uV·nJ]*5	**14.4**	37.9		40.1	-	99 (w/ pixel noise)
	ADC FoM2 [uV·pJ/DRU]*6	**15.7**	41.4		73.5	-	-
MD mode	Res. & max. frame-rate	10 b, 170 fps			8 b, 10 fps	3 levels, 30 fps	-
	Energy / frame [uJ]	**0.26**	1.7		110	0.036	-
	Energy / frame / pix [pJ]	**400**	2650		1375000	46.3	-
	ADC FoM [uV·nJ]*5	**90.4**	598.9 (x1)		138050 (x230.5)	-	-

*1 AG = Analog Gain
*2 Multiple-sampling (MS) notation : MS=(MS1, MS2)
*3 This number is used for the rest of the table
*4 Quantization noise adjust : sqrt(ADCnoise^2-2/12·LSB^2)
*5 ADC FoM : ADCnoise [uV] * (Energy / frame / total pix) [nJ]
*6 ADC FoM2 : ADC FoM [uV·nJ] *1000 / DRU

Figure 5.2.6: Comparison table.





Technology	TPSCo CIS 65nm 1P3M
Die Size	3800 um (H) x 4400 um (V)
Pixel pitch	1.5um
APS array	792H x 528V (Active : 650H x 490V > VGA)
# of ADC ch.	792
ADC pitch	3.1 um

Figure 5.2.7: Chip photograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

5.3 A Data-Compressive 1.5b/2.75b Log-Gradient QVGA Image Sensor with Multi-Scale Readout for Always-On Object Detection

Christopher Young[1], Alex Omid-Zohoor[1,2], Pedram Lajevardi[3], Boris Murmann[1]

[1]Stanford University, Stanford, CA
[2]*now with K-Motion Interactive, San Francisco, CA
[3]Robert Bosch, Sunnyvale, CA

Histograms of Oriented Gradients (HOG) are attractive features for object detection in embedded vision applications, as they provide a good trade-off between complexity and detection accuracy. A custom 8b CMOS imager that computes these features on-chip consumes only 52pJ/pixel [1]. However, a complete system also requires a backend detection algorithm, which consumes 940pJ/pixel in an optimized implementation [2]. As shown in the system study of [3], this imbalance mostly stems from the large amount of data seen by the detector. To remedy this issue, the work of [3] studies a feature-extraction approach that aggressively log-quantizes the data, thereby eliminating unnecessary illumination-related bits from the histograms. The custom log-gradient image sensor described in this paper demonstrates this concept in CMOS. It consumes 127pJ/pixel and offers two log-gradient modes (1.5b and 2.75b) along with multi-scale readout. With several algorithmic enhancements enabled by the log gradients as described in [3], the resulting HOG feature compression ratios are 25× (1.5b) and 9.5× (2.75b) when compared to a standard 8b image (without image pyramid). Using 1.5b log gradients, [3] conservatively estimates a 3.3× reduction in backend detection energy for a deformable parts model (DPM) based detector.

Figure 5.3.1 motivates the log-gradient approach with images from a relatively large dynamic range (DR) scene. Raw 8b images are used to calculate 9b linear gradients for comparison against 2.75b (7 level) log gradients. The linear images and log gradients are both taken with the reported imager. To capture all the linear gradient information, high bit-depth is needed due to the varying illumination across the image. This can be seen when the DR requirement is stressed further by placing a 2-stop neutral density filter in front of the imager reducing the exposure by 4×. The linear gradient's magnitude is reduced since the gradients scale with the illumination profile (α), which requires enough gradient quantization levels to keep the values in range. However, if the gradient is taken in the log domain, α cancels and fewer bits are needed to retain the gradient. The example of Fig. 5.3.1 shows that 2.75b log-gradients are robust to the 4× illumination stress test and essentially strip the unnecessary illumination information from the image.

Figure 5.3.2 shows a block diagram of our QVGA (320×240) imager system, which employs standard 3.3V, 5μm, 4T pixels with pinned-photodiodes. Column-parallel, inverter-based amplifiers perform correlated double sampling (CDS) and function as a three-row cyclic buffer for reading the pixel values multiple times. Each pixel is required in the calculation of four gradients: two horizontal, and two vertical. An extra row of storage is also included in the analog buffers for pixel-binning. To compute the log-gradients, this work introduces column-parallel ratio-to-digital converters (RDCs). Rather than directly computing the log of each pixel, then taking the horizontal and vertical gradients, and finally quantizing, the RDCs digitize the ratio of neighboring pixels in the vertical and horizontal directions directly. In addition to its own column, each RDC is connected to the row buffer output on its left and right via a MUX to compute a centered gradient (i.e., [-1 0 1]). An alternative way to realize the log gradient functionality is to employ logarithmic pixels. However, this comes at the cost of high fixed-pattern noise (FPN) or extra in-pixel circuitry that reduces the fill-factor. Because the gradients are sensitive to differential noise and offset between adjacent pixels, we preferred linear pixels with true CDS. The generated log-gradients are sent off chip for HOG feature generation and object detection by the DPM described in [3] for end-to-end system evaluation. As an additional testing aid and to produce standard images, the chip contains a global buffer for digitizing the analog pixel values with an external ADC.

Figure 5.3.3 shows the column circuit of the analog row buffer, which contains an inverter-based amplifier whose bias point is set by a floating current source [4]. The timing is similar to a traditional readout with CDS. A column's pixel signal value is subtracted from its reset value and the resultant charge is pushed on one of four 50fF hold capacitors (C_H) using selection switches $s_{1...4}$. The stored value is accessed via the corresponding selection switch and by operating the circuit in a flip-around configuration. While only three hold capacitors are required for computing the gradients, a fourth capacitor allows for pixel-binning that averages pixel blocks for multi-scale readout (see image pyramid in Fig. 5.3.3). The imager uses two register rows that are programmed with a bit-pattern defining how the pixels should be averaged. Figure 5.3.3 shows a bit pattern for 2×2 spatial binning. The top register row is static and controls the pixel row averages across the columns in a bin with signals *ravg*. The bottom row holds the same bit pattern but is shifted once per row to assert the alternate-cap (*acp*) signals to indicate which columns use the extra hold capacitors. For an arbitrary sized *n* by *n* bin, after *n* row reads, each of the *n* columns holds one of the *n* row averages. These columns then average the *n* values together to yield the final bin value.

Figure 5.3.4 describes the column RDC, which is programmable to 1.5b or 2.75b. Two pixel values are sampled sequentially from the row buffers onto separate capacitive attenuators (ratio cells). The RDC performs a binary search for the ratio, similar to a successive-approximation ADC. The switching diagram in Fig. 5.3.4 shows the 1.5b case. The MSB is calculated by comparing the two sampled voltages. The larger voltage is then attenuated by a factor of two and another comparison yields the LSB. While the two comparisons provide 2b, the log-gradients use only three bins, thus producing 1.5b. The measured 1.5b RDC transfer characteristic is plotted in Fig. 5.3.4 and should ideally show parallel lines on a log scale. The curves deviate at low levels due to nonidealities, which along with random noise sets the DR of the readout. A similar procedure is followed in the 2.75b mode (decision levels shown in Fig. 5.3.4) but requires a third comparison.

To evaluate the system, we performed an object-detection task on cars using the 2130-image validation set from [3] and compared the log vs. linear gradient performance (see Fig. 5). An LED monitor displayed each scene and we took images at two different exposures (base exposure and 4× lower to simulate poor exposure metering). The base exposure is fixed and chosen by maximizing pixel coverage in histograms of several database images. At base exposure, the average precision (AP) is comparable between 9b/6b linear gradients and 2.75b log gradients. With 4× less light, the AP is roughly constant in the 9b linear and 2.75b log gradient cases. However, the AP drops for the 6b linear gradients, and the 3b linear gradients completely fail as there is not enough DR. The sample images in Fig. 5.3.5 show that 1.5b log gradients have less information than 2.75b log gradients resulting in the lowest base AP. However, the 1.5b log gradient AP increases slightly at lower light levels, possibly due to more pixel values residing in the full-scale range of the imager. Despite its lower AP, the 1.5b option may be attractive in applications where high accuracy is not required (i.e., a wakeup mechanism) and maximum data compression is preferred.

Figure 5.3.6 compares the specs of our imager chip (see 4mm^2 die in Fig. 5.3.7) to similar work. The design of [1] achieves lower energy mainly due to a custom 1.3V pixel implementation. However, its detection accuracy is evaluated without using the on-chip imager on a more simple single classification task. The work of [5] uses log pixels with four photodiodes to calculate horizontal/vertical ratios locally at a very high frame rate but exhibits higher energy. The imager in [6] contains a complete face-detection algorithm but utilizes less-robust Haar-like features at a low frame-rate (1fps) and only has three readout scales.

Acknowledgements:
We thank Bosch RTC for partial funding and design support. This work was funded in part by Systems on Nanoscale Information fabriCs (SONIC), one of the six STARnet Centers, sponsored by MARCO and DARPA.

References:
[1] J. Choi, et al., "A 3.4μW CMOS Image Sensor with Embedded Feature-Extraction Algorithm for Motion-Triggered Object-of-Interest Imaging," *ISSCC Dig. Tech. Papers*, pp. 478-479, Feb. 2013.
[2] A. Suleiman, et al., "A 58.6mW Real-Time Programmable Object Detector with Multi-Scale Multi-Object Support Using Deformable Parts Model on 1920×1080 Video at 30fps," *IEEE Symp. VLSI Circuits*, 2 pages, June 2016.
[3] A. Omid-Zohoor, et al., "Toward Always-On Mobile Object Detection: Energy Versus Performance Tradeoffs for Embedded HOG Feature Extraction," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 28, no. 5, pp. 1102-1115, May 2018.
[4] B. Gönen, et al., "A 1.65mW 0.16mm^2 Dynamic Zoom-ADC with 107.5dB DR in 20kHz BW," *ISSCC Dig. Tech. Papers*, pp. 282-283, Feb. 2016.
[5] N. Katic, et al., "A Relative Imaging CMOS Image Sensor for High Dynamic Range and High Frame-Rate Machine Vision Imaging Applications," *IEEE Sensors Journal*, vol. 15, no. 7, pp. 4121-4129, July 2015.
[6] K. Bong, et al., "14.6 A 0.62mW Ultra-Low-Power Convolutional-Neural-Network Face-Recognition Processor and a CIS Integrated with Always-On Haar-Like Face Detector," *ISSCC Dig. Tech. Papers*, pp. 248-249, Feb. 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

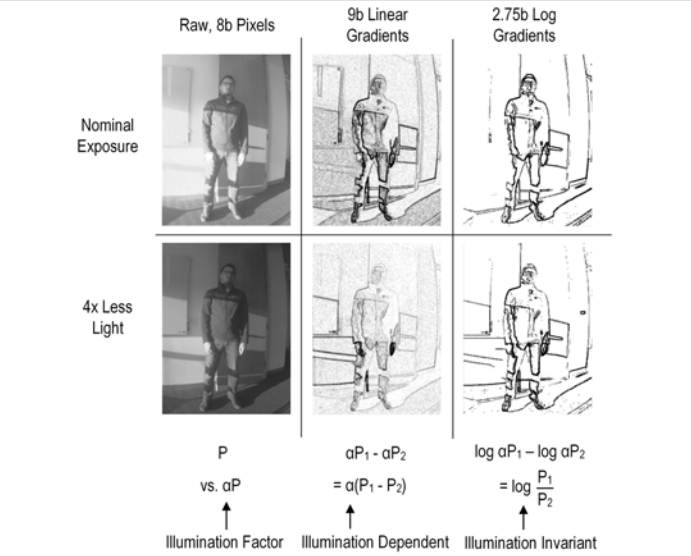


Figure 5.3.1: Linear gradients vs. log gradients under illumination change.

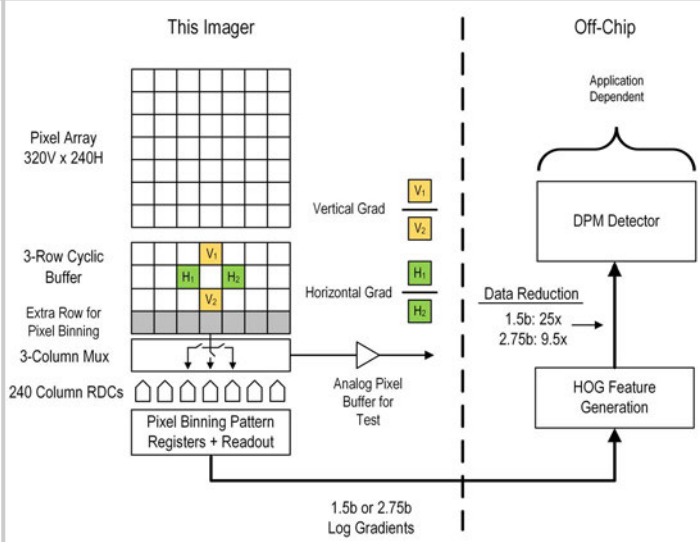


Figure 5.3.2: System architecture.

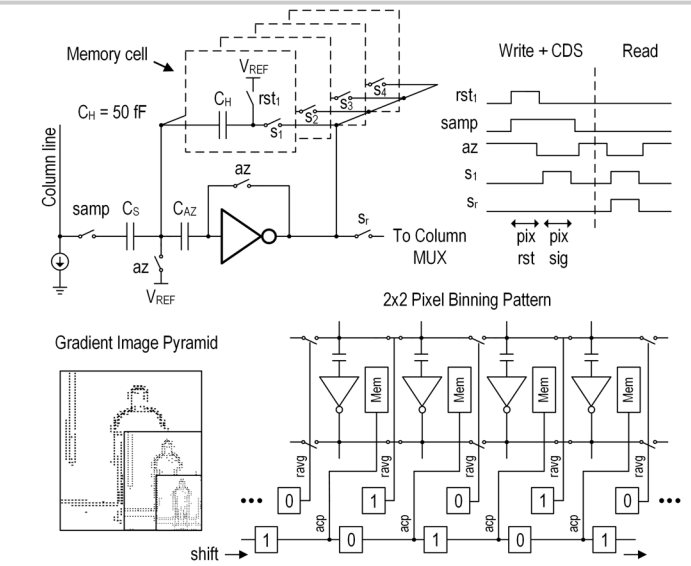


Figure 5.3.3: Inverter-based analog row buffer with pixel binning.

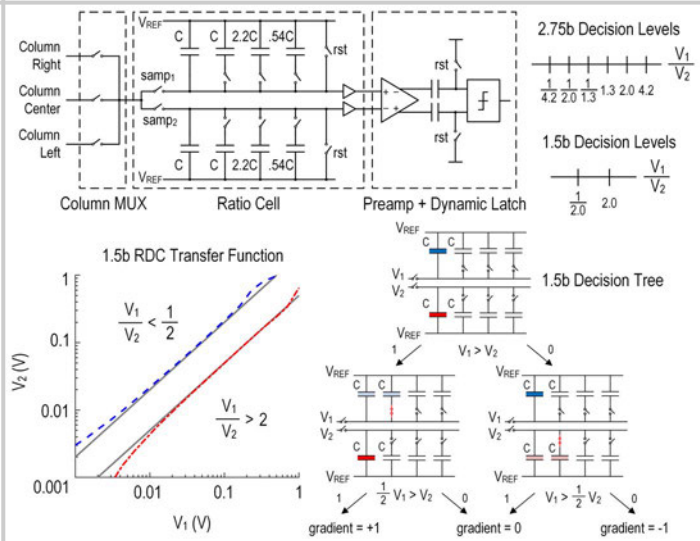


Figure 5.3.4: Ratio-to-Digital Converter (RDC) and measured 1.5b transfer characteristic.

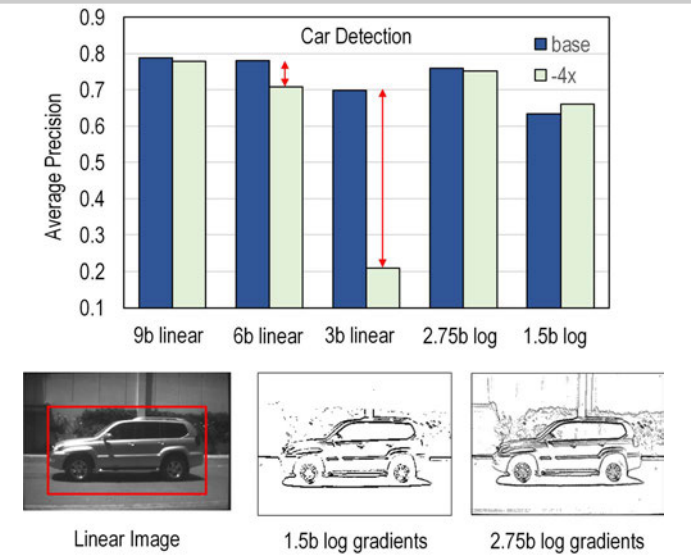


Figure 5.3.5: Object detection evaluation.

	This Work	[1]	[5]	[6]
Technology	0.13μm 1P4M	0.18μm 1P4M	0.18μm	65nm 1P8M
Resolution	320x240	256x256	32x32	320x240
Pixel Size	5μm x 5μm	5.9μm x 5.9μm	31μm x 26μm	7μm x 7μm
Feature Type	log-gradients	linear HOGs	relative ratios between pixels	linear Haar-like w/ face-detector
Frame Rate	207fps max	15fps - reported	9756fps nom. 24000fps max	1fps - reported
Dynamic Range	59.3dB[1]	54.8dB	43dB[2]	-
Power Consumption	292μW @ 30fps	51μW @ 15fps	4mW @ 9765fps	24-96μW @ 1fps
Energy per Pixel	127pJ/pixel	52pJ/pixel	404pJ/pixel	312 - 1250pJ/pixel
Multi-Scale	Yes - arbitrary square bins	No	No	Yes - three scales

1. At output of row buffer, without RDC
2. Pixel-to-pixel dynamic range

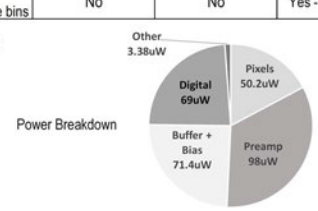


Figure 5.3.6: Comparison table.

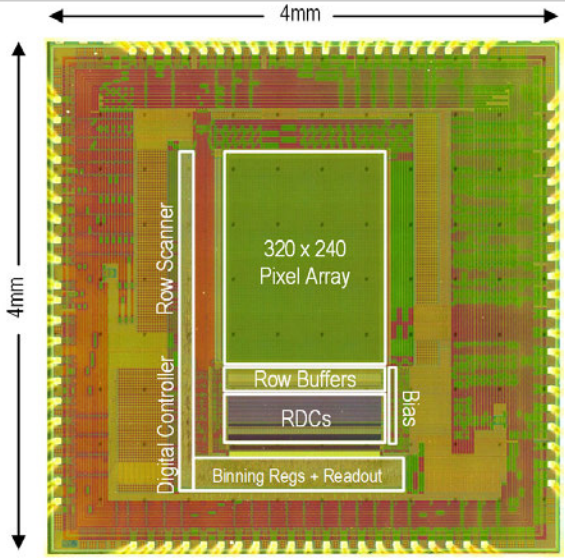


Figure 5.3.7: Die photo.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

5.4 A 76mW 500fps VGA CMOS Image Sensor with Time-Stretched Single-Slope ADCs Achieving 1.95e$^-$ Random Noise

Injun Park[1], Chanmin Park[1], Jimin Cheon[2], Youngcheol Chae[1]

[1]Yonsei University, Seoul, Korea
[2]Kumoh National Institute of Technology, Gyeongbuk, Korea

The demand for high-frame-rate CMOS image sensors is steadily increasing. Column-parallel single-slope (SS) ADCs are widely used in CMOS image sensors, because they can be implemented with small area, low noise, and high energy efficiency. To achieve high frame rate and low noise simultaneously, several techniques using SS ADCs, such as parallel multiple sampling [1,2], dual-gain slopes [3], and dual-gain amplifiers [4], have been investigated. However, since the clock frequency of the SS ADC is already in the GHz range, it is very challenging to maintain energy efficiency as the frame rate increases further.

This paper presents a time-stretched (TS) SS ADC architecture for a high-frame-rate CMOS image sensor. In this architecture, a column-parallel time stretcher is proposed, where the time residue of the SS ADC is expanded to 16 times and the conversion cycle of the SS ADC is significantly reduced (only 80 cycles for 10b resolution). This makes it possible to use a lower clock frequency for column counters, resulting in a significant power reduction. Since the TS SS ADC uses a single ramp generator as a conventional SS ADC, it can easily maintain the noise level and analog gain functionality of the SS ADC. A prototype VGA sensor is fabricated in a 110nm 1P4M CMOS process. The sensor achieves a horizontal (1-H) conversion time of 4μs with a 100MHz counter clock and a random noise (RN) level of 1.95e$^-_{rms}$ at 500fps, while consuming 76mW. This corresponds to an energy efficiency FoM of 0.96e$^-$×nJ.

Figure 5.4.1 shows the block and timing diagrams of the column-parallel TS SS ADC. The column circuit consists of a comparator, a time stretcher, and MSB/LSB counters. The 1-H conversion time is set to 4μs, allowing 500fps for VGA pixel array. Before the conversion, the comparator offset is mitigated with auto-zeroing ϕ_{AZ}. The reset signal of the pixel is applied to the comparator, and the reset conversion is carried within a 0.6μs conversion time. When the V_{ramp} is equal to the output voltage of the pixel source follower V_{SF}, the comparator output ϕ_{comp} triggers a 7b MSB counter operating at a 100MHz clock frequency. There is still a time residue T_{res} of the SS ADC between the negative edge of ϕ_{comp} and the next falling edge of the counter clock ϕ_{CLK}. The time stretcher expands the T_{res} by a factor of 16 and provides a counter clock ϕ_{TS} into a 4b LSB counter. The LSB counter also operates at the same clock frequency as ϕ_{CLK}. For a 10b conversion, the TS SS ADC requires 80 cycles that are calculated as the sum of 64 (MSB) and 16 (LSB) cycles. To avoid out-of-range errors, the MSB counter employs a 1b redundancy, which relaxes the requirements on accuracy of the time stretcher. After reset conversion, each bit of the counter output is inverted. This bit-inversion (BI) provides the negative value of the reset signal ($-D_{rst}$) and is used as the initial value of the next conversion. After the charge transfer in the pixel, the signal level of V_{SF} (voltage swing = 1V) is then readout by using the TS SS ADC within a conversion time of 0.8μs. The final result in the MSB/LSB counters holds the difference of two signals (D_{sig} - D_{rst}) that realizes digital CDS operation. After the conversion, the counter output of each column is transferred simultaneously to SRAM and readout by sense amplifiers during the conversion of the next row.

Figure 5.4.2 describes the operation of the time stretcher. By comparing the comparator output ϕ_{comp} with the falling edge of ϕ_{CLK} signal, two control signals (ϕ_{RES} and ϕ_{STR}) are generated. The time stretcher consists of a current source I_B, two capacitors (C_{TS} and $16C_{TS}$), and an offset-cancelled dynamic comparator. Before the conversion, the output voltages (V_F and V_S) on two capacitors are reset. At the same time, the bias voltage of the comparator's preamplifier is stored on the offset-storage capacitor C_C. During the residue-sampling phase via ϕ_{RES}, I_B charges C_{TS} (0.2pF), and the output voltage V_F ramps up linearly, while the V_S remains zero. Considering the errors of the comparator, the maximum V_F swing is selected to be 0.8V, which corresponds to an LSB step of 50mV for a time resolution of 625ps. At the start of the time-stretching phase, I_B switches to $16C_{TS}$ and then V_F is set to $T_{RES}\, I_B / C_{TS}$. The V_S starts to ramp up at a 16× slower slew rate, and the dynamic comparator generates an output clock ϕ_{TS} for the LSB counter until V_S reaches the V_F value. Therefore, the operation period of ϕ_{TS} represents $16T_{RES}$ and its cycles determine the value of LSBs.

Since the I_B is used for both sampling and stretching phases, the gain of the time stretcher is ideally determined only by the ratio of two capacitors. In practice, the gain error is also caused by a delay mismatch between ϕ_{RES} and ϕ_{STR}, and a residual offset of the preamplifier, resulting in a time offset ΔT and its spread. This gain error is eliminated by applying the digital CDS to the LSB counter. A kick-back noise of the latched comparator also causes a gain error, but it is mitigated via the AC coupling by the capacitor C_C before the comparator. Therefore, unlike conventional time amplifiers, the time stretcher does not require extra calibration or trimming.

A simplified block diagram of the MSB/LSB counters is shown in Fig. 5.4.3 (top). There is no interaction between MSB and LSB counters, thus allowing their independent operation. Each counter cell includes a toggle flip-flop (TFF) and a BI block. For counter operation, the BI block passes the bits from LSB to MSB cell. For CDS operation, it generates a toggle signal regardless of LSB state and inverts the state of each counter cell. To implement a 1b redundancy, the MSB counter employs a double-data-rate (DDR) cell, which stores a state of the ϕ_{CLK} when ϕ_{COMP} is triggered, thus resulting in a 1b extension (M_0) of MSB counter. Due to the exclusion of the clock switching, the power consumption of the DDR cell is negligible. After CDS operation, unlike the value of MSB counter (e.g., 16), the value of the LSB counter can be either negative or positive (e.g., -1), as shown in Fig. 5.4.3 (bottom). This prohibits a direct bit-combination between the results of two counters. However, the problem is solved by adding a sign flag bit (L_4) in the LSB counter. After the conversion including the CDS operation, the ADC output can be obtained by summing the MSB and LSB outputs (e.g., 15=16-1). This summing operation can be processed at the output of the sense amplifier, and thus it does not increase the column level complexity.

The chip occupies an area of 5.5×4.5mm^2 (Fig. 5.4.7). It consumes 76mW from a 3.3/1.5V (analog/digital) supply. The pixel array consists of 640×480 pixels with 4μm pixel pitch and is implemented by a standard 4T pixel. A dark RN is measured with respect to the analog gain of the ramp generator (Fig. 5.4.4). It achieves an RN of 2.85e$^-_{rms}$ with a pixel conversion gain of 82μV/e$^-$. This RN is further improved to 1.95e$^-_{rms}$ with an 8× analog gain. The gain of the time stretcher is also measured for all columns. It exhibits an average gain of 16.12 and a gain variation between 15.84 and 16.34, corresponding to only ±0.2LSB error without any missing code. It achieves a vertical FPN (VFPN) of 0.71e$^-_{rms}$ and 0.18e$^-_{rms}$ for the analog gain of 1× and 8×, respectively. Fig. 5.4.5 shows a sample image taken by the prototype chip at 500fps.

Figure 5.4.6 shows the performance summary and comparison with other state-of-the-art works [1,4-6]. Although the pixel rate of this work is lower than the others, this work achieves one of the fastest 1-H times among all SS ADCs even with lower clock frequency. In terms of energy efficiency, this work achieves a state-of-the-art FoMs of 1.41 and 0.96e$^-$×nJ with 1× and 8× analog gain, respectively.

Acknowledgments:
This work was supported by Samsung Research Funding & Incubation Center of Samsung Electronics under Project Number SRFC-IT1701-08

References:
[1] A. Suzuki, et al., "A 1/1.7-inch 20Mpixel Back-Illuminated Stacked CMOS Image Sensor for New Imaging Applications," *ISSCC Dig. Tech. Papers*, pp. 110-111, Feb. 2015.
[2] T. Haruta, et al., "A 1/2.3inch 20Mpixel 3-Layer Stacked CMOS Image Sensor with DRAM," *ISSCC Dig. Tech. Papers*, pp. 76-77, Feb. 2017.
[3] Y. Oike, et al., "8.3M-Pixel 480-fps Global-Shutter CMOS Image Sensor with Gain-Adaptive Column ADCs and Chip-on-Chip Stacked Integration," *IEEE JSSC*, vol. 52, no. 4, pp. 985-993, Apr. 2017.
[4] H. Totsuka, et al., "An APS-H-Size 250Mpixel CMOS Image Sensor Using Column Single-Slope ADCs with Dual-Gain Amplifiers," *ISSCC Dig. Tech. Papers*, pp.116-117, Feb. 2016.
[5] T. Toyama, et al., "A 17.7Mpixel 120fps CMOS Image Sensor with 34.8Gb/s Readout," *ISSCC Dig. Tech. Papers*, pp. 420-421, Feb. 2011.
[6] T. Arai, et al., "A 1.1μm 33Mpixel 240fps 3D-Stacked CMOS Image Sensor with 3-Stage Cyclic-Based Analog-to-Digital Converters," *ISSCC Dig. Tech. Papers*, pp.126-127, Feb. 2016.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

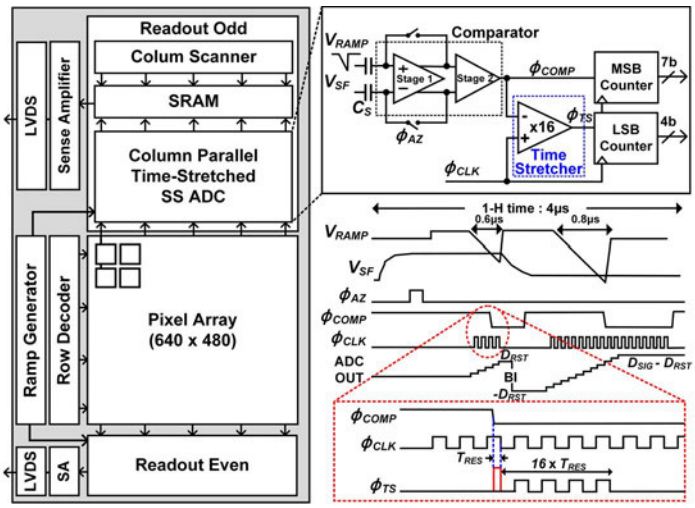


Figure 5.4.1: Block diagram and timing diagram of the CMOS image sensor.

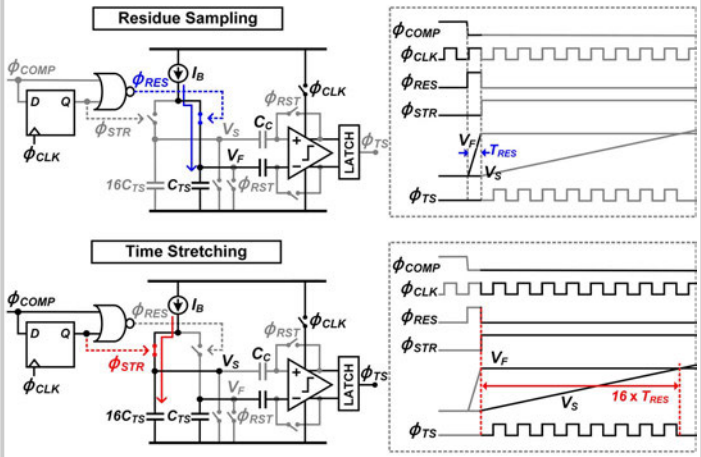


Figure 5.4.2: Block diagram and operation principle of the time stretcher.

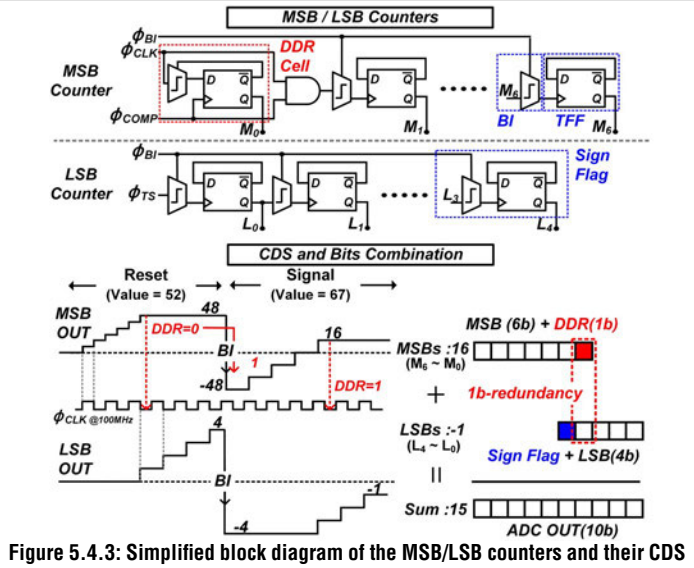


Figure 5.4.3: Simplified block diagram of the MSB/LSB counters and their CDS operation.

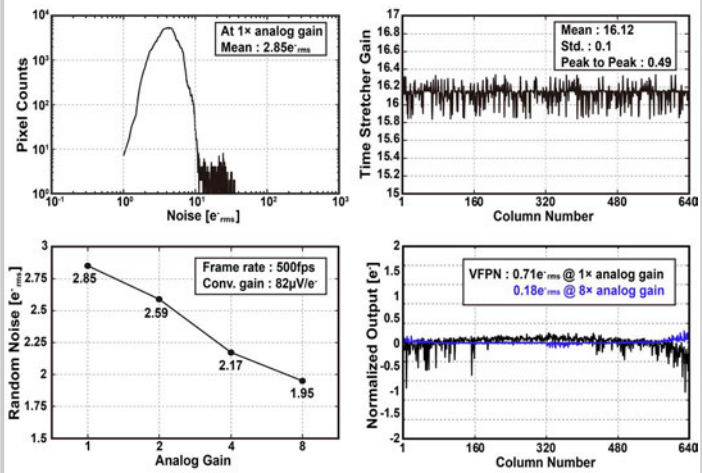


Figure 5.4.4: Measurement results.

Figure 5.4.5: Sample image at 500fps.

Publication	This Work			ISSCC16 Totsuka[4]	ISSCC15 Suzuki[1]	ISSCC11 Toyama[5]	ISSCC16 Arai[6]	
Process	110nm 1P4M			130nm 1P4M	90nm 1P4M / 65nm 1P7M	90nm 1P4M	45nm 1P4M / 65nm 1P5M	
ADC technique	Time-Stretched SS			Dual-Gain SS	Double SS	SS	Cyclic-SAR	
Pixel rate† [MHz]	154			1232	620	2123	7960	
1-H time [μs]	4			15.9**	8.5**	3.9**	0.92	
A/D clock freq. [MHz]	100			810	936	2376	-	
ADC resolution [bit]	10			12	12 / 10	12	12	
Conversion gain [μV/e-]	82			91	76.6	-	92	
Power [mW]	76			1970	532	3000	3000	
Analog gain	1	4	8	4*	27dB	21dB	1	4
Random noise [e-rms]	2.85	2.17	1.95	3.5	1.3	2.75	4.5	3.6
FoM†† [e--nJ]	1.41	1.07	0.96	5.59	1.11	3.88	1.7**	1.36

† Pixel rate = (Frame rate × Total pixel #)
†† FoM = (Power × Noise) / (Frame rate × Total pixel #)
* Adaptive gain = 1, 4
** Calculated

Figure 5.4.6: Performance summary and comparison with the state of the art.

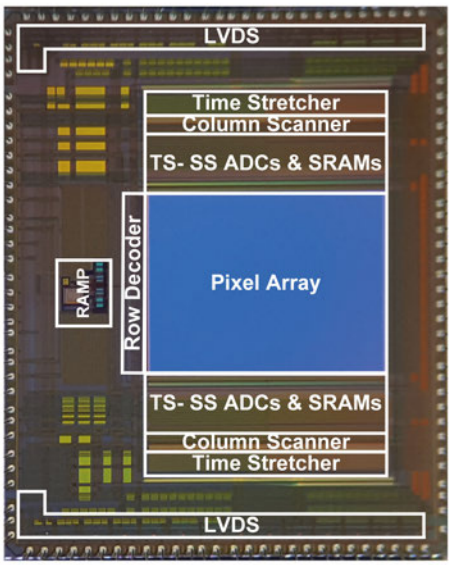


Figure 5.4.7: Chip micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

5.5 Dual-Tap Pipelined-Code-Memory Coded-Exposure-Pixel CMOS Image Sensor for Multi-Exposure Single-Frame Computational Imaging

Navid Sarhangnejad[1], Nikola Katic[2], Zhengfan Xia[1], Mian Wei[1], Nikita Gusev[1], Gairik Dutta[1], Rahul Gulve[1], Harel Haim[1], Manuel Moreno Garcia[3], David Stoppa[4], Kiriakos N. Kutulakos[1], Roman Genov[1]

[1]University of Toronto, Toronto, Canada
[2]Synopsys, Toronto, Canada
[3]Fondazione Bruno Kessler, Trento, Italy; [4]ams AG, Ruschlikon, Switzerland

Modern computational photography applications such as 3D sensing, gesture analysis, and robotic navigation drive the growing need for programmability, or coding, of the camera exposure at the individual-pixel level. Unlike conventional cameras, which record all light incident onto a pixel, the emerging class of coded-exposure-pixel (CEP) cameras can be programmed to selectively detect only *some* of that light [1] or, better, sort *all* of the light [2,3], depending on the pixel code. In conjunction with a concurrently coded illumination, this enables a wide range of new coded multi-exposure single-readout-frame imaging capabilities at video rates. This work demonstrates such an image sensor where multiple pixel-wise-coded exposures, or subframes, are accumulated during one video frame.

Over the past decade, CEP cameras were initially implemented using bulky, distortive, slow and expensive components like digital micromirror devices (DMD) together with off-the-shelf image sensors. In the last 2-3 years, the initial prototypes of fully integrated CEP image sensors have emerged, including a one-tap imager for compressive sensing [1]. One-tap imagers lose light when the tap is off and cannot sort photo-generated charge. Low-resolution (80×60 or less) two-tap light-sorting image sensors for primal-dual coded imaging [2] and motion-deblurring [3] have recently been reported. By using two taps, all incident light is utilized regardless of applied codes. These imagers have large pixels (>12μm pitch) and low fill factors (<34%). None of them use a pinned photodiode-based (PPD) image sensor CMOS technology and thus suffer from low image quality and low tap contrast (defined as the transferred minus residual charge divided by their sum), and as a result, are prone to degraded photo-generated charge sorting. Finally, the essential feature of global-shutter-style coded-subframe exposure for non-static codes is currently not available.

This work demonstrates a next-generation CEP image sensor that addresses all of the aforementioned limitations, and validates it in two novel single-frame computational imaging applications as illustrated in Fig. 5.5.1 (top). Figure 5.5.1 (top, left) depicts the single-frame structured light imaging application where: (1) 4 spatial sinusoidal patterns separated by a 90 degree phase shift are cyclically projected onto the scene; (2) synchronously with the projected patterns, 4 code matrices with time-multiplexed Bayer-like mosaic pattern, as shown, are cyclically submitted to the camera; and (3) the sorted photo-generated charges are accumulated across every 4 subframes and read out as two images per frame. Off-chip image processing reconstructs all lighting conditions at full spatial resolution. Finally, disparity and albedo maps are computed. Figure 5.5.1 (top, right) depicts the single-frame photometric stereo application. In this example, instead of sinusoidal patterns, 4 LED light sources are turned on, one at a time, during the 4 respective subframes within one frame. Following a procedure similar to that in structured light imaging, maps of surface normals and albedo are computed.

To enable such applications, a sensor with a general operating principle described in the flow chart in Fig. 5.5.1 (bottom, left) is needed. The photo-generated charge is stored on taps 1 or 2 for codes 0 and 1, respectively, during each coded subframe, and the results are accumulated over N subframes within one video frame. We introduce the code-memory pixel (CMP) architecture that operates following the aforementioned principle. To perform global coded-subframe exposure it requires an in-pixel dual digital code memory (Fig. 5.5.1 bottom, right). The code memory is pipelined: a code value is pre-loaded row-wise into each pixel during the previous subframe and is applied at the beginning of the new subframe. Photo-generated charge is collected based on the current code while the next code is being pre-loaded into the pixel.

Figure 5.5.2 (left) depicts the CMP architecture and the relevant timing diagrams. The CMP includes two D-latches, one controlled by ROW_LOAD to pre-load the code patterns row-by-row, and the other controlled by GLOB_LOAD to activate this pattern globally. Based on the code in each pixel, one of the two transfer gates, TG1 or TG2, connects the pinned-photodiode (PPD) to the corresponding floating diffusion node, FD1 or FD2, respectively. The pattern in each pixel is gated with EXPOSURE signal to stop any charge transfer during the read-out phase. EXPOSURE is also kept low during the global code updates to ensure that signal and supply glitches caused by digital switching in the array do not affect the tap contrast. Figure 5.5.2 (right) depicts the pixel layout floorplan and its cross-sectional electrostatic potential diagrams. A fill-factor of 45.3% is achieved with 27% of the area occupied by the latches and logic gates. The CMP operates at the maximum of 180 subframes/s with a high tap contrast of 99%.

Figure 5.5.3 depicts the 280×176-pixel sensor system architecture. Unlike conventional image sensors, during each subframe the CEP sensor receives pixel codes to control all pixels' exposure individually. The sensor has 18 digital input channels for streaming the codes row by row, 18 code-deserializers with logic for arranging the codes, and vertically organized code loading control circuits to ship the codes to their respective rows. Additional column-sharing and row-sharing logic allows for faster streaming of repeated patterns, to save power and time needed for shipping the same pattern from an off-chip DRAM. The column-parallel outputs are amplified by switched-capacitor programmable-gain amplifiers (PGAs), which can apply different gains to each of the two taps. The output is then serialized and buffered over 6 analog output channels.

Figure 5.5.4 depicts raw outputs of the sensor for four different camera configurations for the same scene, under uniform lighting conditions. The first two rows show images where only one subframe is recorded per frame. The image is collected on taps 1 and 2 for black and white codes, respectively. In the 3rd row, there are two subframes, with tiled 2×2-pixel codes, as shown, sent to the pixel array. The projector projects all-white during the first subframe and all-black during the second one. The magnified inset for tap 1 shows that the image is black for 1 out of 4 pixels. The last row depicts the same conditions, but the projector projects all-white for both subframes. The magnified inset shows that the alternating rows are brighter because they have collected light for two subframes, while the darker rows have collected light for one subframe only.

Figure 5.5.5 (top) depicts the albedo and disparity map reconstruction pipeline for the single-frame structured light 3D imaging setup where 4 subframes are used. As discussed above for Fig. 5.5.1 (top, left), at the end of a frame, each of the 2 tap images capture the visual scene sampled 4 times and coded by a four-pixel time-multiplexed Bayer-like mosaic pattern, corresponding to 4 90-degree-shifted sinusoidal illumination patterns. The Bayer-coded raw images from taps 1 and 2 (Fig. 5.5.5 top, left) are separated out into 4 images for each tap, each containing different structured light sine phase information (Fig. 5.5.5 top, middle). These images are then upsampled and processed to obtain 4 full-resolution images from which the output disparity and albedo maps are computed. Figure 5.5.5 (bottom) shows the results from a photometric stereo experiment. The top row depicts the surface normal maps of three different scenes captured at 20fps. The images are color-coded with the orientation of the surface normal at each pixel. The bottom row shows the albedo maps containing the intensity images of the same scenes.

Figure 5.5.6 compares the design with state-of-the-art multi-tap and other computational image sensors. Figure 5.5.7 depicts the micrograph of the chip.

Acknowledgement:
The authors gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada under the RGPIN, RTI and SGP programs, of DARPA under the REVEAL program, and of CMC Microsystems. We also wish to thank Hui Feng Ke, Hui Di Wang and Gilead Wolf Posluns for help on test system development.

References:
[1] J. Zhang, et al.,"Compact all-CMOS Spatiotemporal Compressive Sensing Video Camera with Pixel-wise Coded Exposure," *Optics Express*, vol. 24, no. 8, pp. 9013-9024, 2016.
[2] N. Sarhangnejad, et al.,"CMOS Image Sensor Architecture for Primal-Dual Coding," *Int. Image Sensor Workshop*, 2017.
[3] Y. Luo, et al.,"Exposure-Programmable CMOS Pixel With Selective Charge Storage and Code Memory for Computational Imaging," *TCAS-I*, vol. 65, no. 5, pp. 1555-1566, 2018.
[4] F. Mochizuki, et al.,"Single-Shot 200Mfps 5×3-Aperture Compressive CMOS Imager," *ISSCC Dig. Tech. Papers*, Feb. 2015.
[5] G. Wan, et al.,"CMOS Image Sensors With Multi-Bucket Pixels for Computational Photography," *JSSC*, vol. 47, no. 4, pp. 1031-1042, 2012.
[6] M. Seo, et al., "A 10.8ps-Time-Resolution 256×512 Image Sensor with 2-Tap True-CDS Lock-In Pixels for Fluorescence Lifetime Imaging," *ISSCC Dig. Tech. Papers*, 2015.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

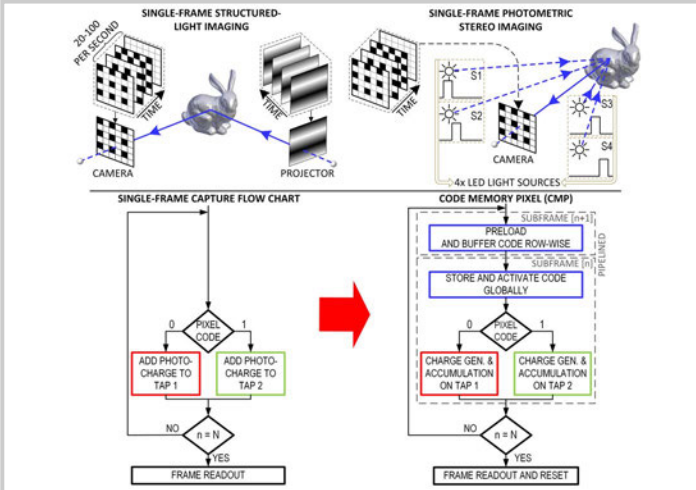


Figure 5.5.1: System configuration of single-frame structured light 3D imaging (top-left) and single-frame photometric stereo imaging (top-right). A generic flowchart of a dual-tap CEP sensor (bottom-left) and the flow chart for the CMP pixel architecture (bottom-right) are shown.

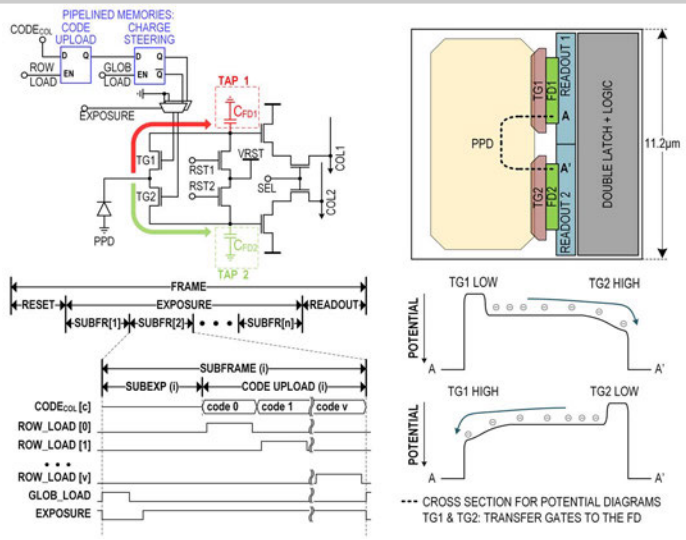


Figure 5.5.2: Schematic and layout of the code-memory pixel (CMP) with the associated timing and potential diagrams.

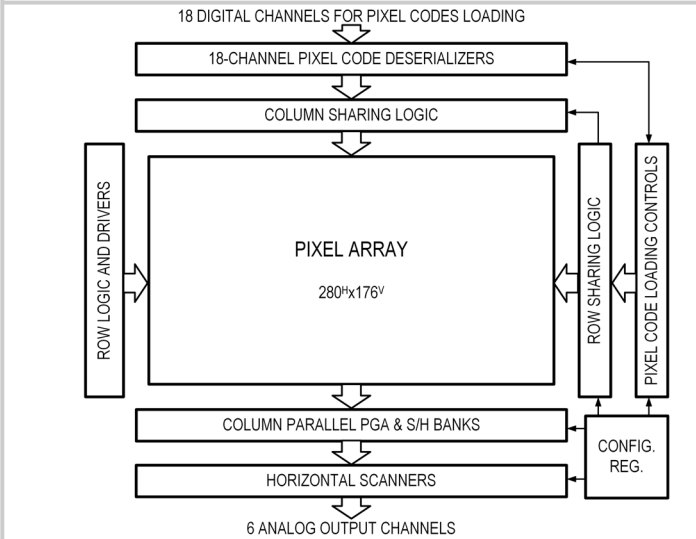


Figure 5.5.3: Image sensor top-level architecture with the code loading circuits and readout.

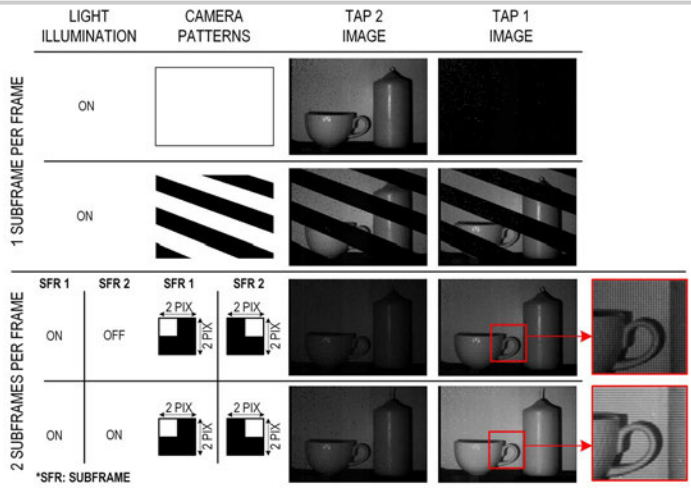


Figure 5.5.4: Raw output images of the sensor for two uniform-lighting cases: one (top) and two (bottom) subframes per frame. The insets on the right illustrate the difference between the two frames, with the same two sensor patterns but different illumination.

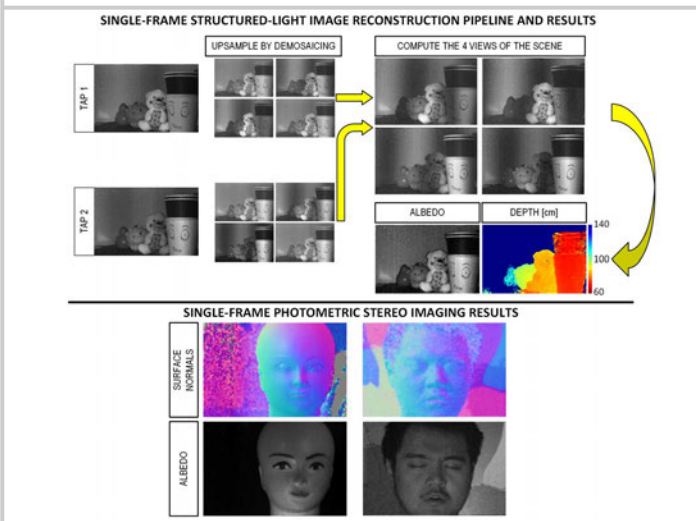


Figure 5.5.5: The albedo and disparity map reconstruction pipeline of single-frame structured light imaging (top); and examples of color-coded surface normal and albedo maps of the photometric stereo imaging (bottom).

		THIS WORK	[3]	[2]	[1]	[4]	[5]	[6]
	CODED-EXPOSURE	PER-PIXEL (*i.e.* PIXELWISE SPATIAL AND TEMPORAL CODING)				PER 1/15 OF ARRAY	PER FULL ARRAY (*i.e.* TEMPORAL CODE ONLY)	
PIXEL	TECHNOLOGY [nm]	**110 CIS**	130 CMOS	350 CMOS	180 CIS	**110 CIS**	130 CIS	**110 CIS**
PIXEL	PINNED PHOTODIODE	**YES**	NO (NW/P)	NO (PG)	**YES**	**YES**	**YES**	**YES**
PIXEL	PIXEL PITCH [μm]	**11.2**	12.1x12.2	25	10	11.2X5.6	**5**	11.2X5.6
PIXEL	FILL FACTOR [%]	45.3	33.2	20.5	**52**	–	42	–
PIXEL	NUMBER OF TAPS	**2**	**2**	**2**	1 (PHOTONS ARE LOST)		**2**	**2**
PIXEL	TAP CONTRAST	**99 @ 180afps** [1]	–	LOW [2]	N/A	N/A	–	94 [3]
ARCHITECTURE	PIXEL COUNT [HxV]	244 x 162	10 x 10	80 x 60	127 x 90	64 x 108	**640 x 576**	256 x 512
ARCHITECTURE	FRAME RATE [fps]	25	60	25	**100**	32	N/A	12
ARCHITECTURE	POWER [mW]	34.4	**1.23**	7	1.3	1620	N/A	540
ARCHITECTURE	POWER FoM [nJ] [4]	34	205	58	**1.14**	7324	N/A	343
SYSTEM	CODE MEMORY	IN-PIXEL (2 LATCHES)	IN-PIXEL (DRAM)	IN-PIXEL (2 LATCHES)	IN-PIXEL (SRAM)	OFF-PIXEL	N/A	N/A
SYSTEM	SUBFRAME RATE [afps [1]]	180	**300000**	–	100	N/A		
SYSTEM	PIXEL-CODE RATE [MHz]	7.1	**30**	–	0.11			
SYSTEM	ARBITRARY CODE / ROI	**YES/YES**	YES/–	YES/–	NO/–			
SYSTEM	SUBFRAME SHUTTER	**GLOBAL**	ROLLING	**GLOBAL**	ROLLING			
SYSTEM	APPLICATIONS	1.One-shot structured-light, 2.One-shot photometric stereo, 3.Other: compressive sensing, etc.	Deblurring	Transport-Aware imaging	Spatiotemporal compressive sensing	Ultra high speed w/ compressive sampling	High-dynamic range	Fluorescence lifetime imaging

1: subframes per second
2: Limitations due to the type of the photo-detector (photogate-based)
3: Also known as Extinction Ratio
4: $FoM = Power/((Number\ of\ Pixels) \times (Frame\ rate))$
N/A: Not Applicable
–: Not Available

Figure 5.5.6: Comparison table.

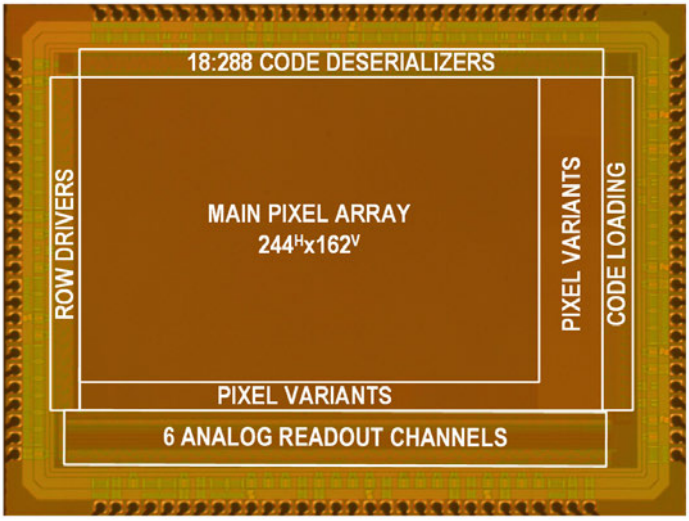


Figure 5.5.7: Chip micrograph (the dimensions are 4mm × 3mm).

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

5.6 A 400×400-Pixel 6μm-Pitch Vertical Avalanche Photodiodes CMOS Image Sensor Based on 150ps-Fast Capacitive Relaxation Quenching in Geiger Mode for Synthesis of Arbitrary Gain Images

Yutaka Hirose, Shinzo Koyama, Toru Okino, Akito Inoue, Shigeru Saito, Yugo Nose, Motonori Ishii, Seiji Yamahira, Shigetaka Kasuga, Mitsuyoshi Mori, Tatsuya Kabe, Kentaro Nakanishi, Manabu Usuda, Akihiro Odagawa, Tsuyoshi Tanaka

Panasonic, Nagaokakyo, Japan

The intensive development of Single-photon avalanche photodiode (SPAD) based CMOS image sensors (CIS) continues, with rapid progress [1-6]. Yet, due to unestablished quenching operation [5,6], realization of SPADs onto a CIS alongside conventional pixel circuitry has been a fundamental challenge.

In this paper, we present a CIS based on a vertically flowing carrier avalanche photodiode (VAPD) with a "relaxation-quenching (RQ)" method in which the cathode of VAPD and its parasitic capacitances including floating diffusion (FD) are left "floating" during the Geiger-mode (GM) operation. The technique utilizes a negative-feedback effect of avalanche-generated charge accumulation on the floating node to reduce the ionization field of VAPD. RQ is intrinsically fast due to small capacitances of the floating node, and a no-overflow pixel is possible. The technique was originally employed in a photon-counting imager for a pulse latch [7]. However, the fact that this action occurs with any amount of charge can further be exploited to implement an imager with arbitrary charge gain, i.e., from low gain (gradation) to linear avalanche and to high gain (Geiger mode) with "guaranteed fast quenching". We demonstrate such imaging by synthesis of gradation and ToF range images taken under a 100Klux sunlight environment.

The concept of RQ is visualized at the top of Fig. 5.6.1. Initially (left), a high voltage is applied to the p-n junction of VAPD. After detecting a photon, VAPD goes into the GM mode and a large amount of charge builds up and accumulates in the series capacitance. This charge, in turn, reduces the internal field until VAPD "relaxes" to a non-breakdown regime and avalanching is automatically quenched (right) with a corresponding voltage drop (V_Q). In Fig. 5.6.1, the RQ process is simulated by numerically solving a set of nonlinear difference equations. Amounts of accumulated charge on the capacitance (middle) and of avalanche generated charge (bottom) in the depletion region are plotted as functions of time with substrate bias as a parameter. For $|V_{Bias}|>28.5V$, the process is estimated to complete within 150ps, much faster than the period of the resistive quenching [2]. Since there is no rebuild phase until the next reset, dead time and after-pulse are irrelevant within a frame. Thus, one can use the stored charge as image signal without further disturbance.

The pixel structure in the vertical direction is drawn in Fig. 5.6.2. In a p-type substrate, a p-n junction with a high built-in field is formed within the photodiode (PD) area, giving a 70% fill factor. Potential profiles of each cross section are also plotted. A pixel circuit is formed in a triple well, and its diagram is drawn on the right. For gain tuning, the drain voltage of the reset transistor (V_{RSD}) is swept between 3 and 4.5V with fixed substrate bias. By setting the maximum amount of charge to be stored in the capacitance to the saturation charge, V_Q can be calculated as shown at the right bottom. We note that pixel overflow is prevented by setting the isolation barrier between PD and the pixel circuit region higher than $|V_Q|$.

In Fig. 5.6.3, a block diagram of the CIS is shown. The substrate voltage is externally controlled to be 0V for the normal imaging mode and -21.5V for the onset of the avalanche mode. The reset transistor gate control is switchable between rolling- and a global-shutter/ToF modes. The readout path consists of the gradation image readout circuit for low-gain modes (<10000) and of the binary image readout circuit for GM (>10000). The former circuit comprises a column amplifier (COLAMP), a correlated double sampling (CDS) circuit, and horizontal shift registers (HSR). The GM binary signal is directly fed into the same HSR operated both as a binarizing and as a data-transferring circuit. The specification of the fabricated CIS and comparison to reported references is summarized in Fig. 5.6.6. A chip photograph is shown in Fig. 5.6.7.

By time-gating repetitive photon pulses (0.6photon/pixel), the RQ signals of the source-follower (SF) outputs are monitored as a function of bias voltage. The oscilloscope waveforms are shown on the left of Fig. 5.6.4. Initially, the VAPD cathode node and FD are reset (green line) to HIGH level. Exposure starts after the TRN (pink line) is turned on. On arrival of a photon pulse (blue line), the SF output (yellow) immediately drops stepwise by 0.7V for $|V_{sub}|=25.5V$ and by 1V for $|V_{sub}|=26V$. The fall time is measured to be 120ns, which is limited by the resistance of the SF load. The device internal fall time of RQ is considered to be 150ps as explained in Fig. 5.6.2. Since the signal drop occurs only once stepwise without any disturbance, there is no after-pulse effect within the frame. Histograms corresponding to the three conditions are plotted in the middle column. The fact that the highest signal level does not exceed the SF saturation level (1.3V), plotted as a red dotted line, demonstrates successful quenching without overflow due to after-pulses. When $|V_{sub}|\geqq 25.5V$, the histogram clearly shows binary feature, and therefore, the taken images on the right are binarized by the HSR input threshold voltage. The most significant dark pulses are found to occur during the reset phase of the rolling mode, probably due to the optical cross-talk effect.

Finally, we synthesized gradation and ToF images taken in a daytime under a 100Klux sunlight environment (Fig. 5.6.5). The left top is the gradation image taken with normal operation. The left bottom is the direct ToF image taken with the binary-mode circuit. The full range (0 to 270m) is divided into 36 subranges giving a range resolution of 7.5m. In each subrange, 3800 pulses with 1200W peak from a 940nm wavelength laser are time-gated in a pulse per frame manner, and are counted and separated from the background by post-processing. The subrange images are mapped to a color scale (varying from orange to blue). Although the reflectivity is uncertain, the furthest building walls at ~250m ahead are correctly measured to be in the deepest subrange (blue). A synthesized image of the two images is shown on the right. The effect that the range image becomes to have additional texture and luster features is clearly seen and is expected to be useful for object recognition.

In summary, we have developed a VAPD based CIS with fast relaxation quenching of the Geiger mode. This enables imaging with arbitrary gain with the same pixel circuit and to attach features of low-gain images to high-gain (binary) images. This is expected to be useful not only for ToF systems, but also for surveillance cameras, robotics, and AI applications.

Acknowledgement:
The authors are grateful to Dr. E. Fujii for his continuous support and encouragement.

References:
[1] L. Gasparini, et al., "A 32×32-Pixel Time-Resolved Single-Photon Image Sensor with 44.64μm Pitch and 19.48% Fill-Factor with On-Chip Row/Frame Skipping Features Reaching 800kHz Observation Rate for Quantum Physics Applications," *ISSCC Dig. Tech. Papers*, pp. 98-99, Feb. 2018.
[2] N.A.W. Dutton, et al., "A SPAD-based QVGA Image Sensor for Single Photon Counting and Quanta Imaging," *IEEE Trans. Electron. Dev.*, vol. 63, no. 1, pp. 189-196, Jan. 2016.
[3] A.R. Ximenes, et al., "A 256×256 45/65nm 3D-Stacked SPAD-Based Direct TOF Image Sensor for LiDAR Applications with Optical Polar Modulation for up to 18.6dB Interference Suppression," *ISSCC Dig. Tech. Papers*, pp. 96-97, Feb. 2018.
[4] M. Perenzoni, et al., "A 64×64-pixel Digital Silicon Photomultiplier Direct ToF Sensor with 100Mphotons/s/pixel Background Rejection and Imaging/Altimeter Mode with 0.14% Precision up to 6km for Spacecraft Navigation and Landing," *ISSCC Dig. Tech. Papers*, pp. 118-119, Feb. 2016.
[5] S. Koyama, et al., "A 220 m-Range Direct Time-of-Flight 688×384 CMOS Image Sensor with Sub-Photon Signal Extraction (SPSE) Pixels Using Vertical Avalanche Photo-Diodes and 6 kHz Light Pulse Counters," *Symp. VLSI Circuits*, pp. 360-361, 2018.
[6] M. Mori, et al., "A 1280×720 Single Photon Detecting Image Sensor with 100 dB Dynamic Range Using A Sensitivity Boosting Technique," *ISSCC Dig. Tech. Papers*, pp. 120-121, Feb. 2016.
[7] M. Perenzoni, et al., "A 160×120 Pixel Analog-Counting Single-Photon Imager With Time-Gating and Self-Referenced Column-Parallel A/D Conversion for Fluorescence Lifetime Imaging," *IEEE JSSC*, vol. 51, no. 1, pp. 155-167, Jan. 2016.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

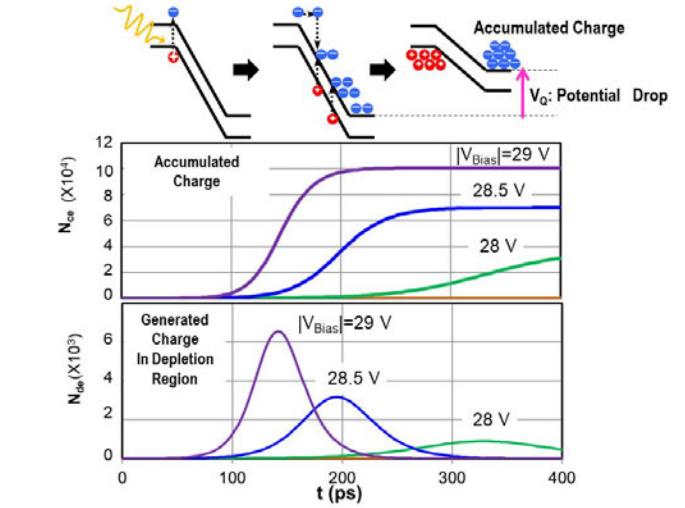


Figure 5.6.1: Band diagrams (top), calculated charge accumulated on the floating capacitor (middle), and avalanche charge (bottom) during relaxation quenching.

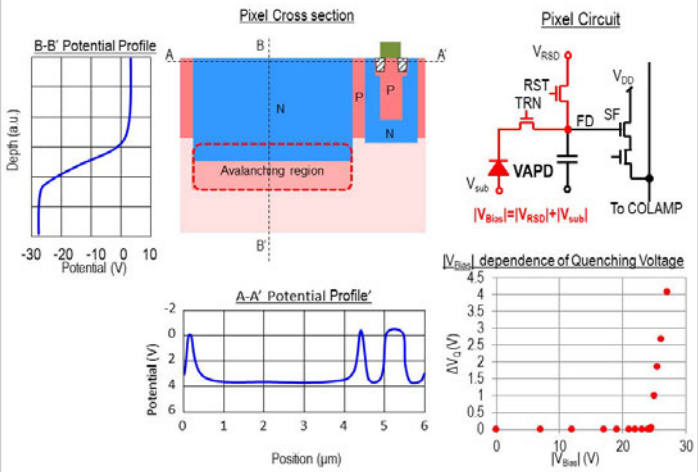


Figure 5.6.2: Schematics of pixel cross section, potential profiles, pixel circuit, and bias dependence of quenching voltage drop (V_Q).

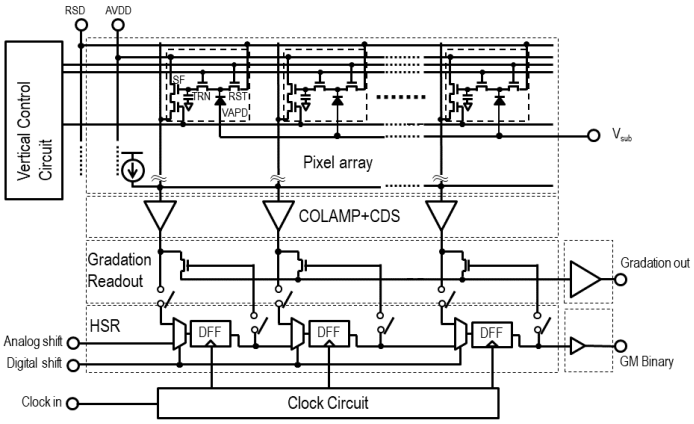


Figure 5.6.3: Block diagram of VAPD CMOS image sensor.

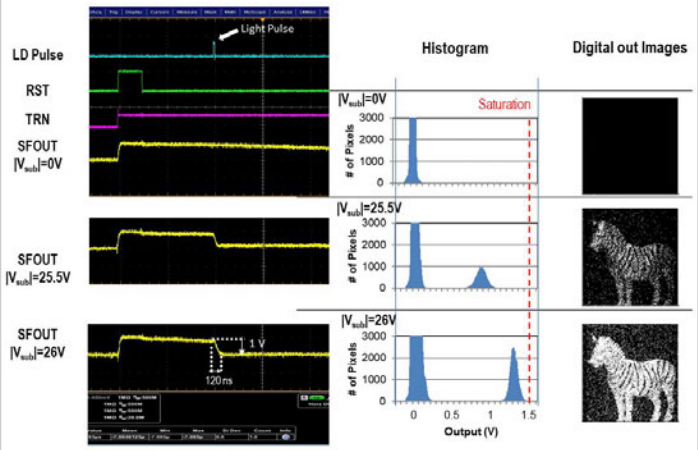


Figure 5.6.4: Oscilloscope waveforms (left), histograms (middle), and binary mode images (right) under low light level conditions taken with $|V_{sub}|$=0V, 25.5V, and 26V.

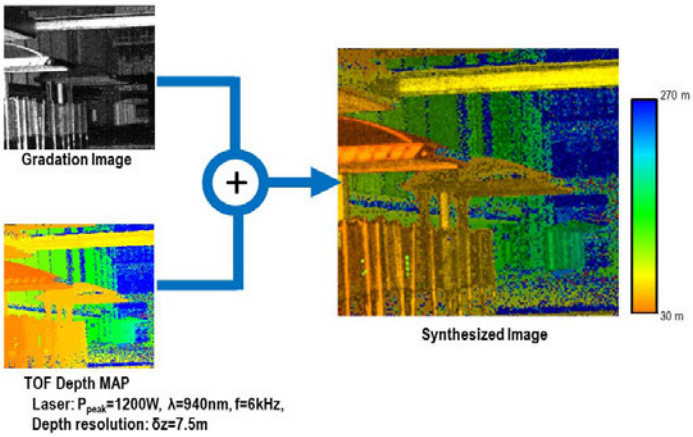


Figure 5.6.5: Gradation and daytime (100000lux) depth map images synthesis.

	This Work	[1]	[2]	[3]	[4]	[5]	[6]
CMOS Technology	65nm	150nm	130nm	45nm/65nm	150nm	110nm	110nm
Pixel Size	6μm	44.64μm	8μm	19.88μm	60μm	11.2μm	3.8μm
Array size	400x400	32x32	320x240	256x256	64x64	688x384	1280x720
Physical Signal	Photo-Charge	Current-Pulse	Current-Pulse	Current-Pulse	Current-Pulse	Photo-Charge	Photo-Charge
Quenching Type	Floating Capacitance	Series Resistance	Series Resistance	Series Resistance	Series Resistance	N.A.	N.A.
Fill Factor	70%	19.48%	26.8%	50.6%	26.5%	N.A.	N.A.
Operation Voltage	-3.3V~-29V	N.A.	N.A.	N.A.	21.8V	-76 V	-21.3V~ -23.3V
DCR(@RT)	100cps	600cps	47cps	6700cps	6800cps	N.A.	0.1cps
Imaging Modes	• Gradation • Binary	Pulse-Counts	Pulse-Counts	TDC*-Counts	TDC*-Counts	• Gradation • Pulse-Counts • Binary	• Gradation • Binary
Applications	• TOF • Surveillance • Scientific	• Photon-Counting • Scientific	Photon-Counting	TOF	TOF	TOF	• Surveillance • Scientific
Frame Rate	60fps	250kHz	16kfps	N.A.	7.68fps	60fps	15fps

*Time-to-Digital-Converter

Figure 5.6.6: Chip performance summary and comparison with state-of-the-art APD-based CISs.

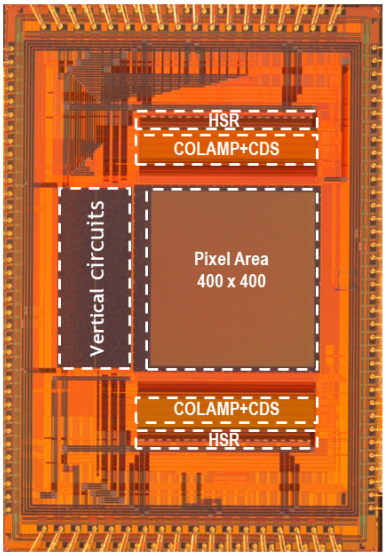


Figure 5.6.7: Chip photograph of the VAPD CMOS image sensor.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

5.7 A 256×256 40nm/90nm CMOS 3D-Stacked 120dB-Dynamic-Range Reconfigurable Time-Resolved SPAD Imager

Robert K. Henderson[1], Nick Johnston[1], Sam W. Hutchings[1], Istvan Gyongy[1], Tarek Al Abbas[1], Neale Dutton[2], Max Tyler[3], Susan Chan[3], Jonathan Leach[3]

[1]University of Edinburgh, Edinburgh, United Kingdom
[2]STMicroelectronics, Edinburgh, United Kingdom
[3]Heriot-Watt University, Edinburgh, United Kingdom

Light Detection and Ranging (LIDAR) applications pose extremely challenging dynamic range (DR) requirements on optical time-of-flight (ToF) receivers due to laser returns affected by the inverse square law over 2-3 decades of distance, diverse target reflectivity, and high solar background [1]. Integrated CMOS SPADs have a native DR exceeding 140dB, typically extending from the noise floor of few cps to 100's Mcps peak rate. To deliver this DR to downstream DSP, large SPAD time-resolved imaging arrays must count and time billions of single photon events per second demanding massively parallel on-chip pixel processing to achieve practical I/O power consumption and data rates. Hybrid Cu-Cu bonding offers a mass-manufacturable platform to implement these sensors by providing high-fill-factor SPADs optimised for NIR stacked on dense nanoscale digital processors [2]. Stacked sensor architectures involving pixel-level histogramming, on-chip peak detection and TDC/processor resource sharing are now being investigated [3-5].

This paper presents a large imaging array of compact, reconfigurable SPAD time-resolved pixels in a 3D stacked 40nm/90nm CMOS technology. Figure 5.7.1a shows the chip organisation, 8 I/O pads generate a modest 31.6Mbps of pixel data at the targeted 30fps (a maximum of 800Mbps at 760fps) via either a rolling or global shutter readout scheme. A uniform top-tier BSI matrix of 256×256 pwell/deep-nwell SPADs at 9.2µm pitch with 51% fill-factor and median dark count rate of 20cps at 1.5V excess bias. In the bottom-tier a 64×64 matrix of modular photon processors at 38.4µm pitch is integrated in 40nm CMOS technology with ~40M transistors. Figure 5.7.1b shows an interface circuit that is agnostic to the SPAD polarity for compatibility to future detector generations. External global *Vqp* and *Vqn* voltages and the *Invert* signal allow selection of either NMOS or PMOS quench transistors and time gating. An enable SRAM allows masking of noisy detectors. Two time-gated paths are implemented for photon counting or photon timing as *SPAD(k)* and *Start(k)* respectively. For photon counting, under control of *ToggleEn* signal, the *Start(k)* signal can encode photons on both rising and falling edges or transition once to report the first photon in an exposure time. For photon timing, it is possible to count STOP pulses by re-using the same path to generate macro-time stamps.

The 16 *Start(k)* toggling outputs are mixed via a XOR tree to a single sequence *PhotonEdges* that is passed to the multi-event TDC (METDC). In addition, *PhotonEdges* is processed by a multi-photon trigger circuit [5], which continually counts photons from the first rising edge received in a laser cycle. A threshold of 1, 2, 4, 8 photons occurring within a 1-to-10ns delay time (adjustable by current starving voltage *Vndelay*) generates an output trigger to start the gated ring oscillator (GRO). This circuit saves power by only activating the GRO if there is a highly correlated burst of photons likely to belong to a laser return while suppressing uncorrelated background light. The signal *Qtrig* enables a qualified trigger as the delayed first photon in the burst if the threshold has been met. In addition, a toggling trigger sequence can be generated on multiple bursts for use in the METDC. The METDC is based on the architecture presented in [6] operating with a shift-register delay chain to allow adjustable time bin resolution. This architecture allows time digitisation of up to 16 photons per laser cycle, or 1 photon per SPAD per laser cycle. Time bin resolution is set over a range 0.56ns to 560ns by *TDCclk* by selection of a divided phase of the GRO (Fig. 5.7.2b). This clock starts only on the first trigger (single or multi-photon) for power saving. The 14b GRO TDC is adjustable by the *Vddro* voltage over a full-scale range of 560ns to 1.6µs for a 35-to-100ps LSB resolution, matched to common automotive LIDAR ranges. Alternatively, for longer distance range *TDCclk* can connected to a lower frequency external clock *ExtClk*.

Finally, the 16×14b ripple counter array and mode multiplexer allow the sensor to operate in three main imaging modalities: (1) global-shutter (GS) single photon counting (SPC) for gated intensity imaging, and time-resolved imaging via (2) triggered timestamping or (3) multi-event coarse histogramming.

For SPC mode, the 16 counters receive inputs from the 16 *SPAD(k)* signals capturing full 256×256 image resolution with 16k photon full-well capacity (84dB DR). GS-HDR modes are realised by chaining the counters into 8×28b whilst binning groups of 4 SPADs allowing 256M photons per SPAD to be counted in an exposure period with a 64×264 spatial resolution. Figure 5.7.3a-d shows a HDR scene taken in 1k and 100 klux illumination at 30fps showing a 14b 1klux image, rollover of the 14b counters at 100klux and a tone-mapped unsaturated 100klux 28b image. The GS capture is parasitic light insensitive as the BSI sensor is all-digital. A DR of 120dB is seen in the photon transfer curve in Fig. 5.7.3c with peak count rates of 200Mcps/SPAD or 13Tcps for the whole array. We now achieve full native SPAD dynamic range is available in a video-rate image sensor.

For the two time-resolved modes, the 16 METDC bin outputs *TDC(k)* are directed to the counters, capturing 262k triggers (single or multiphoton) per histogram before saturation. The mode multiplexer can also be configured to an HDR METDC mode by chaining pairs of 14b counters to 28b with a reduced number of time bins. Sparing activation of the 14b GRO TDC through the multiphoton trigger is economic in both power consumption and I/O bandwidth, as these resources are not expended on background photon detections. Figure 5.7.4a shows histograms generated with 1 and 2 photon trigger levels showing considerable background rejection and sharper peak. Figure 5.7.4c shows a TDC nonlinearity of ±10cm over a 50m range. These data were obtained with a 671nm Picoquant pulsed laser coupled into a 100µm multimode fiber and passed through a 15mm lens to flood illuminate the scene. The pulse duration of the laser was ~100ps, the repetition rate was 1.9MHz, and after the fibre and imaging lens, the laser power was measured to be 1.8mW. Figure 5.7.4d shows the impulse response function of 240ps FWHM. Figure 5.7.5a and b detail two-step timings to allow the coarse METDC mode to be applied to generate 3D video images at >50m distances. The first timing approach operates the 16 bins over the full distance range with a *ExtClk* synchronised to 16 times the laser rate set by the round trip time. A first exposure is taken to determine the four MSBs of the target range. In a second exposure, the laser rate is increased by 16× and the in-pixel GRO *TDCclk* generates the METDC clock. This determines the next four LSBs of the target range but aliases the background noise. The second timing approach to achieve the same disambiguation is to apply the multiphoton triggered mode with a high threshold to timestamp the laser return. In this case, power consumption and jitter of the GRO can be reduced by operating with a STOP clock at 16× laser repetition rate and using the macro-time stamp feature to count STOP cycles. Figure 5.7.5c shows distance accuracy improved over that expected from the 560ps bin size by spreading the pulse energy over bins and performing centroiding around the peak. In this operating condition, the sensor consumes 77.6mW including SPAD, GRO and digital core, considerably lower than TDC array approaches. Figure 5.7.6 shows a table of comparison of this sensor.

Acknowledgements:
This work was supported by the Engineering and Physical Sciences Research Council through the Quantum Hub in Quantum Enhanced Imaging EP/M01326X/1. We are grateful to STMicroelectronics and the ENIAC-POLIS project for chip fabrication. The EPSRC CDT in Intelligent Sensing and Measurement, Grant Number EP/L016753/1 supported Sam Hutchings' work.

References:
[1] C. Niclass, et al., "A 100-m Range 10-Frame/s 340 x 96-Pixel Time-of-Flight Depth Sensor in 0.18-µm CMOS," *IEEE JSSC*, vol. 48, no. 2, pp. 559-572, Feb. 2013.
[2] T. A. Abbas, et al., "Backside Illuminated SPAD Image Sensor with 7.83µm Pitch in 3D-Stacked CMOS Technology," *IEDM*, pp. 8.1.1-8.1.4, 2016.
[3] S. Lindner, et al., "A 252 × 144 SPAD Pixel FLASH LiDAR with 1728 Dual-clock 48.8 ps TDCs, Integrated Histogramming and 14.9-to-1 Compression in 180nm CMOS Technology", *IEEE Symp. VLSI Circuits*, pp. 69-70, 2018.
[4] A. Ximenes, et al., "A 256×256 45/65nm 3D-Stacked SPAD-Based Direct TOF Image Sensor for LiDAR Applications with Optical Polar Modulation for up to 18.6dB Interference Suppression", *ISSCC Dig. Tech. Papers*, pp. 96-98, Feb. 2018.
[5] M. Perenzoni, et al., "A 64 x 64-Pixels Digital Silicon Photomultiplier Direct TOF Sensor with 100-MPhotons/s/pixel Background Rejection and Imaging/Altimeter Mode with 0.14% Precision Up To 6 km for Spacecraft Navigation and Landing," *IEEE JSSC*, vol. 52, no. 1, pp. 151-160, Jan. 2017.
[6] N.A.W. Dutton, et al., "11.5 A Time-Correlated Single-Photon-Counting Sensor with 14GS/S Histogramming Time-to-Digital Converter," *ISSCC Dig. Tech. Papers*, Feb. 2015.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

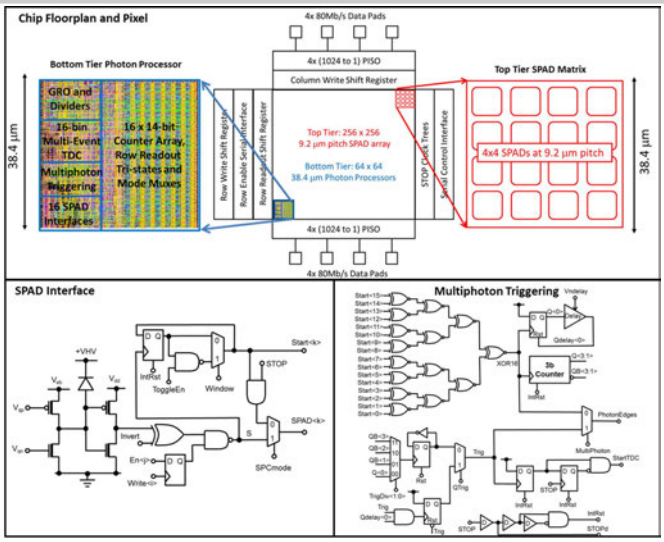


Figure 5.7.1: (a) Chip floorplan and pixel organisation, (b) SPAD interface, and (c) multiphoton triggering circuit.

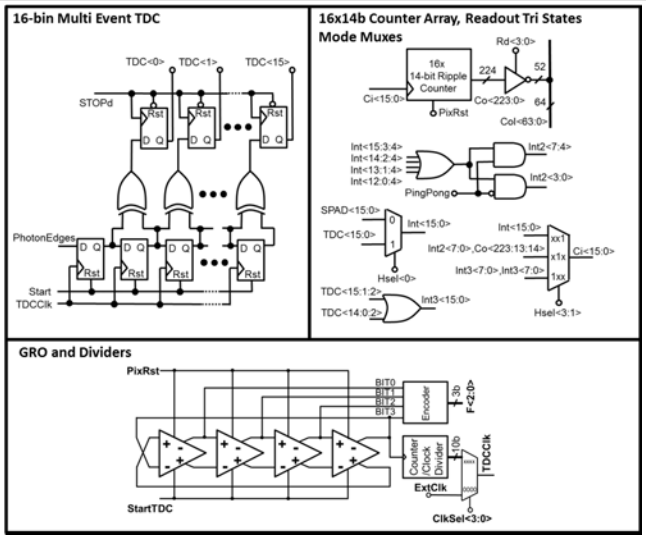


Figure 5.7.2: (a) 16-bin multi-event TDC; (b) 16×14b counter array, readout tri-states and mode mux; (c) GRO and dividers.

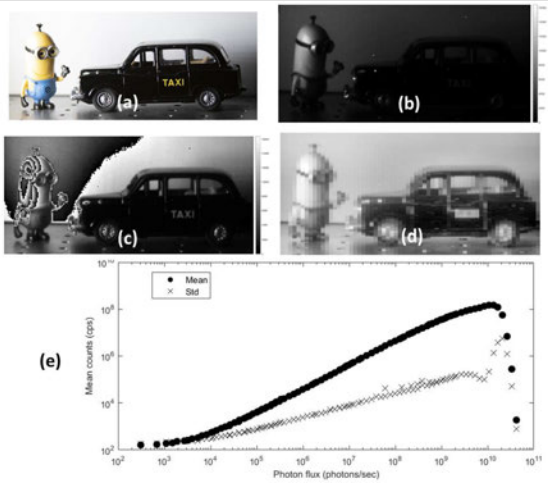


Figure 5.7.3: SPC mode: (a) side-illuminated scene with Nikon camera; (b) 1klux illumination 14b image; (c) 100klux illumination 14b code roll-over image; (d) 100klux illumination 28b image with 4× pixel binning; and (e) photon transfer curve showing 120dB DR in 28b mode.

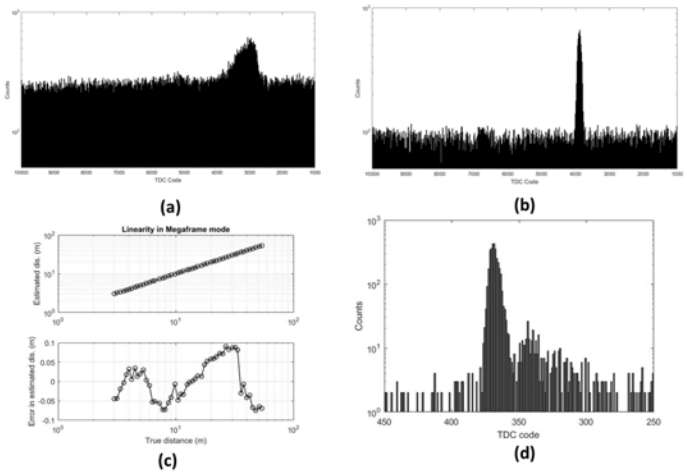


Figure 5.7.4: Multiphoton triggering mode: (a) reflected laser pulse in 1klux ambient; (b) with multiphoton triggering enabled with 2 threshold; (c) linearity sweep over 50m; (d) impulse response function.

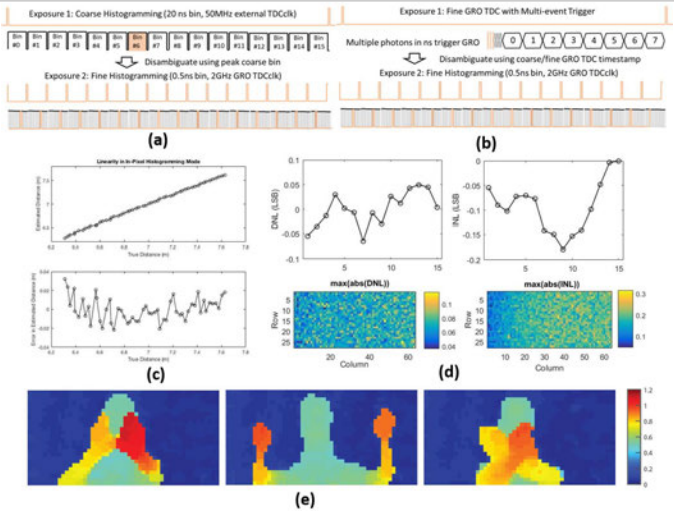


Figure 5.7.5: In-pixel histogramming mode: (a),(b) two-step timing schemes for LIDAR; (c) distance sweep over 1.6m; (d) histogram bin uniformity over full array; (e) 3 images from a 30fps movie taken at 50m.

Parameters	Units	This Work	[3]	[4]	[5]	[1]
Sensor						
CMOS Technology	-	**40nm/90nm**	180nm	45nm/65nm	150nm	180nm
Pixel Array	-	**256x256/64x64**	252x144	16x8	64x64	16x1/32x1
Pixel Pitch	µm	**9.2/38.4**	28.5	19.8	60	21
Fill-factor	%	**51**	28	31.3/50.6	26.5	70
Median DCR@Vex	cps	**20@1.5V**	195 @5V	5.3k@2.5V	6.8k@3V	2.65k@NA
TDC depth	bits	**14/4**	12	14	16/15	12
TDC resolution	ps	**35/560**	48.8	60-320	250-20000	208
TDC area	µm²	**130/150**	4200	550	NA	NA
TDC number	-	**4096**	1728	1	4096	64
TDC linearity DNL/INL	LSB	**+0.05/-0.05** **+0.1/-0.08**	+0.6/-0.48 +0.89/-1.67	+0.8/-0.7 +3.4/-0.8	+1.2/-1 +4.8/-3.2	+0.15/-0.17 +0.32/-0.56
LIDAR Measurement						
Image resolution	-	**64x64**	252x144	256x256	64x64	202x96
Laser projection	-	**Flash**	Flash	Scanning	Flash	Scanning
Wavelength	nm	**671**	637	532	470	870
Repetition rate	MHz	**1.9**	40	1	NA	0.133
Mean illumination power	mW	**1.8**	2	6	NA	21
PDP@ illumination wavelength@Vex	%	**23@3V**	33.7@5V	21@2.5V	20@3V	NA
Max distance	m	**50**	50	150-430	367-5862	128
Imaging range	m	**50**	0.7	4.5	NA	100
FOV	deg.	**1.2x1.2**	40x20	NA	NA	
Accuracy	m(%)	**0.17(0.34)**	0.088(0.17)	0.07(0.3)	1.5(0.37)	11(0.11)
Background light	-	**1klux**	dark	NA	100M/pix/s	70 klux
Target reflectivity	-	**white**	white	white	NA	9%
Power consumption	mW	**77.6**	2540	NA	93.5	530

Figure 5.7.6: Table of comparison.

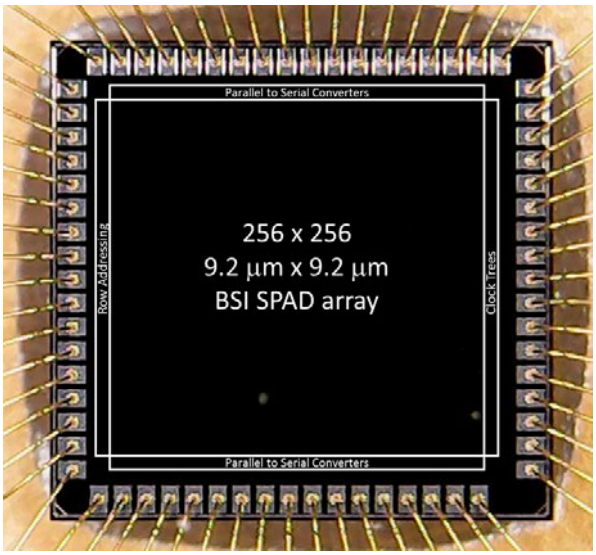


Figure 5.7.7: Chip micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

5.8 A 32×32-Pixel 0.9THz Imager with Pixel-Parallel 12b VCO-Based ADC in 0.18μm CMOS

Sayuri Yokoyama[1], Masayuki Ikebe[1], Yuri Kanazawa[1], Takahiro Ikegami[1], Prasoon Ambalathankandy[1], Shota Hiramatsu[1], Eiichi Sano[1], Yuma Takida[2], HIroaki Minamide[2]

[1]Hokkaido University, Sapporo, Japan
[2]RIKEN, Sendai, Japan

There are many compelling characteristics of signals in the terahertz band (100GHz to 10THz) located between the millimeter wave band and the infrared band. In particular, terahertz waves have higher spatial resolution than mm-waves. Moreover, they can transmit through various substances such as plastics, fibers and paper, and can detect hazardous substances. Because of these features, interest in terahertz applications such as security screening is rising. However, there is a paucity of low-cost terahertz detectors. The Si-CMOS process technology is low-cost and highly integratable with readout electronics and on-chip signal processors. A key consideration for many of the terahertz-detector technologies is the need for additional process steps to make them compatible with CMOS technologies [1]. Recent antenna-type pixel detectors have shown success in high-speed operation and do not require extra process steps [2]; however, power consumption of each pixel and the sequential read-out architecture offsets the speed advantage.

Figure 5.8.1 illustrates a block diagram of our imager. It consists of 32×32 terahertz-wave detectors, 4 blocks of 16 ADC clusters with the decoder and four 12b output channels. The imager is partitioned in to 4 sub-imagers with 16×16 detectors and one 16-ADC cluster block, where each sub-imager works independently. One ADC cluster has 13μm width and 16 column ADCs are placed at the bottom of the 215μm pitch of 16 rows×1 column detector. Each ADC keeps a connection to each detector in parallel and has two data buffers for global-shutter operation and ON/OFF keying (OOK) modulated terahertz imaging. The decoder supporting 16 ADC clusters selects 512 data buffers for ON/OFF states of the sub imager with 16×16 detectors. The 4-output channel relaxes the operation frequency requirement for a 24kb readout, which is related to one frame of data. Running the decoder at 512kHz, we achieve 1kfps for OOK-modulated signals (that is, 2kfps for ON/OFF states). Since our imager employs a pixel-parallel architecture, this solution can open new vistas for terahertz imaging based on global shutter, such as multi-frame super resolution, multimodal imaging, and so on.

Figure 5.8.2 shows the circuit composition of the detector circuit. The detector consists of a patch-type on-chip antenna and cascode amplifier with a subthreshold-biased operational amplifier (subVth-OP amp). The terahertz wave detection is based on square-law nonlinearity of the nMOSFET in the cascode amplifier having usual lowpass characteristics. A GND-biased nMOS current source works at sub-Vth operation with nA-order leak current. The OPamp slowly responds to only DC components of the signal. In this way, the cascode amp has low-pass filtering characteristics and a negative feedback loop while the subVth-OP amp operates like a high-pass filter for DC-cut operation. Thus, the detector operates as a square-law envelope detector for terahertz carrier frequencies that exceed the cutoff frequency of the MOSFET with DC servo to stabilize the DC level of the OOK modulated output voltage

The output of the pixel circuit is connected to a voltage-controlled-oscillator (VCO) based ADC. Basically, VCO-based ADCs work continuously and integrate the analog values during their A/D conversion period. Thus, a prefilter is not required, and analog value estimation with averaging can be performed. Figure 5.8.3 shows the block diagram of our ADC configuration. It consists of two ramp-wave generators, OP amps, a SR-FF, and two 12b counters with data buffer. When the ramp signal rises over the threshold voltage Vctrl, a reset signal is generated for another ramp generator, resetting it through the delay inverters and the RS-FF. Complementary resetting of two generators leads to clock pulse generation with 50% duty cycle. The slope of the ramp signal changes along with the detector output voltage. This value also controls the clock frequency as a VCO. Using a lock-in ADC scheme with OOK, we can separately estimate "THz-ON" or "THz-OFF" states and convert them using two counters. When the obtained data is stored to buffers, these conversion and background readouts are done simultaneously.

In Fig. 5.8.4, our imager acquires one frame by integrating sub frames. For example, when a 100kHz OOK-modulated terahertz pulse comes to the sensor, the detector output Vpix changes along with the terahertz wave intensity, and the VCO converts the Vpix value to frequency VCO_OUT. Using lock-in signals ϕTHz_ON/OFF, the ON/OFF states are separated and accumulated in each of the counters. After 100-time integration, the data is stored to the data buffers. The imager outputs the ON/OFF state sub-frame data as a 2kfps image. While in background readout, the next frame's A/D conversion is performed. Subtracting between ON/OFF state frames results in digital CDS.

For image capture an injection-seeded terahertz-wave parametric generator (is-TPG) was used as a frequency-tunable terahertz-wave source [3]. By changing the noncollinear phase-matching condition in MgO:LiNbO3 crystal, the terahertz-wave frequency was continuously tuned from 0.8 to 1.1THz with a linewidth of approximately 4GHz. Figure 5.8.5 shows image capturing using our imager. The frame rate is restricted by an OOK modulation frequency of 200Hz for the terahertz source. Overall, images are output at a rate of 400fps. The terahertz beam has a Gaussian-like form, average power is 18.8μW@0.93THz, and the beam spot is displayed on the upper side of the image. Although pixel variation is observed for THz_ON state image, external digital CDS cancels the variation. We prepared a 0.08mm-width narrow wire and 0.25mm-diameter aperture for terahertz beam spot control. When the narrow wire is placed on the sensor across the 1mm thick mold, we confirmed real-time splitting of the beam spot. Also, real-time erosion of the beam spot was observed while putting the metal aperture on the sensor. On the shifting beam spot, real-time capturing of the shift, to the right direction is performed and its position displayed in the image sequence.

Figure 5.8.6 compares our work with recently reported antenna-type CMOS terahertz imager [2,4-7]. Almost all imagers use a MOSFET rectifier in 65nm or 130nm CMOS. They do not support A/D conversion and global shutter. Our imager, on the other hand, applies a pixel-parallel ADC architecture using column ADCs and captures 2× images for digital CDS while keeping high responsivity over 200kV/W, low NEP, and low power consumption of pixel circuits, in a 180nm CMOS process. The pixel-parallel architecture achieves higher framerate, and global shutter is known to reduce image distortion while capturing movies. This combination ensures the imager is useful for applications like multi-frame super resolution.

Measured DNL with 12b resolution is +1.85/-1.72LSB and INL is +45.58/-23.98LSB at 2kS/s. Although the INL has some distortion, easy compensation of INL can be applied, because the VCO-based ADC has monotonically increasing characteristics. The analog part, which consists of the VCO, consumes 116μW. However, we can reduce power to 38μW by adjusting the bias setting of the amplifiers. We also confirm a responsivity of 218kV/W@0.93THz in the 78.5×78.5μm^2 antenna area and 34.6kV/W@0.93THz in the 215×215μm^2 detector area from the given Gaussian-form beam profile. The former responsivity is effective for interpolation-based image reconstruction, because the antenna area is considered to sub sampling of whole image. Measured NEP was 91pW/Hz$^{0.5}$@31Hz, and 13.7pW/Hz$^{0.5}$@100kHz was obtained as estimated NEP.

Figure 5.8.7 shows the imager microphotograph and detector geometry. To summarize, our imager is aimed towards low-cost terahertz imaging with three main features: 1) global shutter and digital CDS for OOK-modulated signal by pixel-parallel VCO-based ADCs, 2) >200kV/W responsivity with low power using a MOSFET amplifier-based detector, 3) achieving real-time terahertz imaging using low-cost 180nm CMOS.

Acknowledgements:
This work was partially supported by the Ministry of Internal Affairs and Communications of Japan / Strategic Information and Communications R&D Promotion Programme (MIC/SCOPE) #151301001 and the VLSI Design and Education Center (VDEC), University of Tokyo in collaboration with Cadence Design System, Inc. and Keysight Technologies Japan, Ltd.

References:
[1] S. Eminoglu, et al., "A Low-Cost 128×128 Uncooled Infrared Detector Array in CMOS Process," *J. Microelectromechanical Syst.*, vol. 17, no. 1, pp. 20-30, Feb. 2008.
[2] B. Assim, et al., "A Low-Noise CMOS THz Imager Based on Source Modulation and an In-Pixel High-Q Passive Switched-Capacitor n-Path Filter," *Sensors*, vol. 16, no. 3, p. 325, Mar. 2016.
[3] S. Hayashi, et al., "Ultrabright Continuously Tunable Terahertz-Wave Generation at Room Temperature," *Scientific Reports*, vol. 4, article #5045, June 2014.
[4] A. H. Richard, et al., "A 1 k-pixel Video Camera for 0.7–1.1 Terahertz Imaging Applications in 65-nm CMOS," *JSSC*, vol. 47, no. 12, pp. 2999-3012, Dec. 2012.
[5] K. D. Yeon, et al., "Design and Demonstration of 820-GHz Array using Diode-Connected NMOS Transistors in 130-nm CMOS for Active Imaging," *IEEE Trans. Terahertz Science and Technology*, vol. 6, no. 2, pp. 306-317, Mar. 2016.
[6] S. Kiryong, et al., "A CMOS 300-GHz 7 by 7 detector array for THz imaging," *RFIT*, Sept. 2017.
[7] R. Han, et al., "Active Terahertz Imaging using Schottky Diode in CMOS: Array and 860-GHz Pixel," *JSSC*, vol. 48, no. 10, pp. 2296-2308, Oct. 2013.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

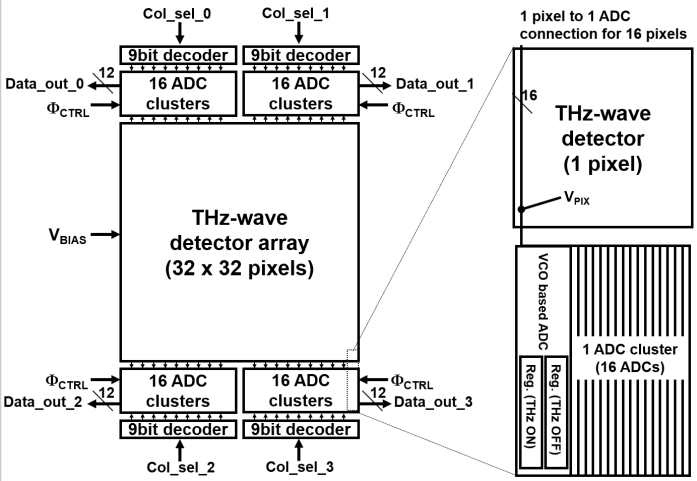


Figure 5.8.1: Block diagram of the terahertz Imager.

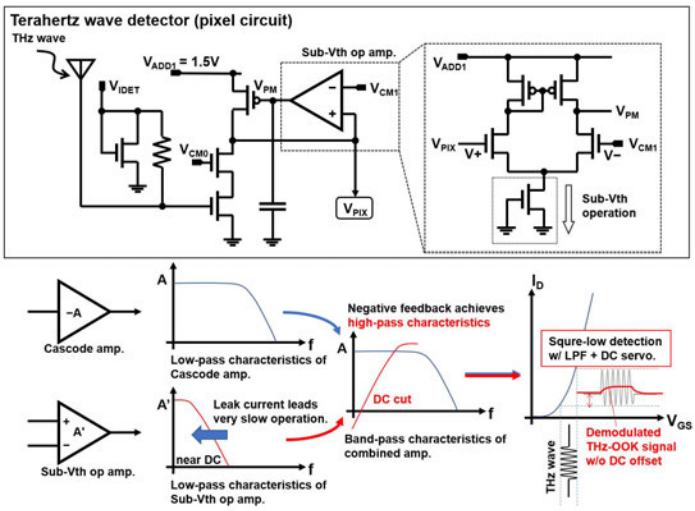


Figure 5.8.2: Detector-circuit schematic and terahertz-wave detecting scheme.

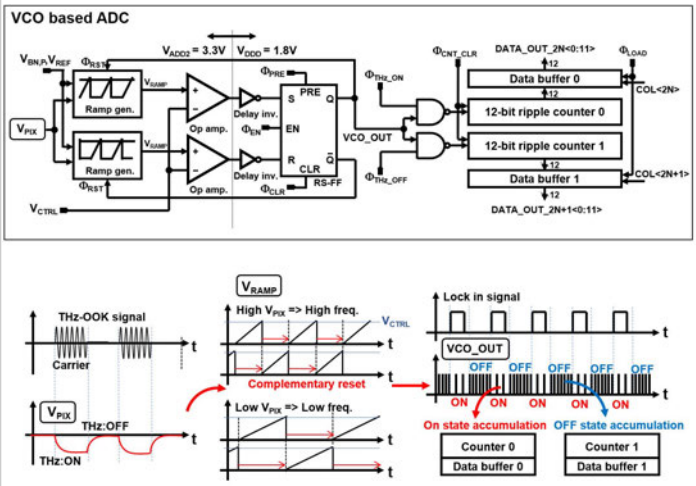


Figure 5.8.3: Simplified block diagram of pixel-parallel ADC and its function.

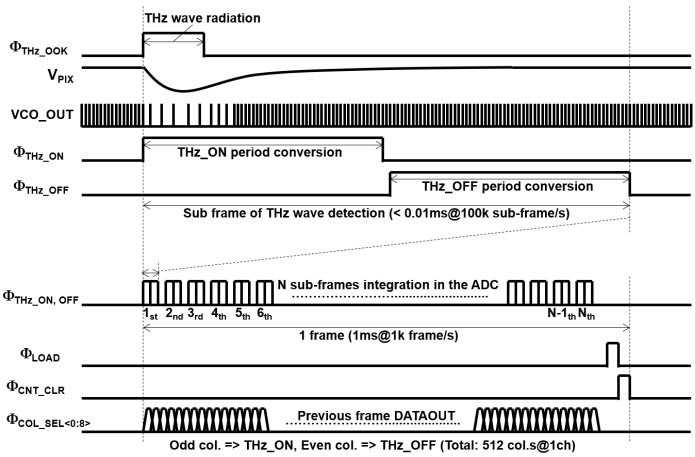


Figure 5.8.4: One-frame timing diagram of pixel-parallel ADC operation.

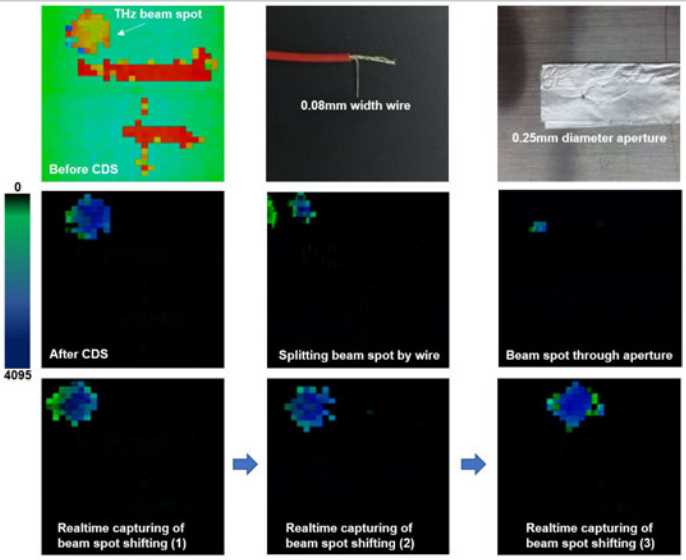


Figure 5.8.5: Captured images at 0.93THz.

	Sensors2016 [2]	JSSC2012 [4]	IEEE THz Sci. 2016 [5]
Process	130nm CMOS	65nm CMOS	130nm CMOS
Array	31 x 31	32 x 32	8 x 8
Pixel pitch	240μm x 240μm	80μm x 80μm	159μm x 159μm
Power consumption	174μW/pixel	2.5μW/pixel	150μW/pixel
Antenna	Bow tie	Ring	Patch
Detection Device	MOSFET rectifier	MOSFET rectifier	MOSFET rectifier
Responsivity	300kV/W@270GHz	115kV/W@860GHz	3.46kV/W@820GHz
On-chip gain	58dB	50dB	13.5dB
ADC	External	External	External
DNL, INL	-	-	-
Shutter	Focal plane	Focal plane	Focal plane
NEP	18.7pW/Hz$^{0.5}$@156kHz	100pW/Hz$^{0.5}$@5kHz	21.2pW/Hz$^{0.5}$@1MHz
	RFIT2017 [6]	JSSC2013 [7]	This work
	65nm CMOS	130nm CMOS	180nm CMOS
	7 x 7	4 x 4	32 x 32
	500μm x 500μm	500μm x 500μm	215μm x 215μm
	-	-	4.5μW/pixel
	Differential patch	Patch	Patch
	MOSFET amplifier	SBD rectifier	MOSFET amplifier
	1.2kV/W@303GHz	336V/W@860GHz	218kV/W@930GHz
	-	-	45dB
	External	External	12b pixel parallel
	-	-	+1.85/-1.72, +45.58/-23.98
	Focal plane	Focal plane	Global
	20pW/Hz$^{0.5}$	29pW/Hz$^{0.5}$@1MHz	91pW/Hz$^{0.5}$ @31Hz
			*13.7pW/Hz$^{0.5}$ @100kHz

*Estimation from noise spectrum of detector

Figure 5.8.6: Performance comparison.

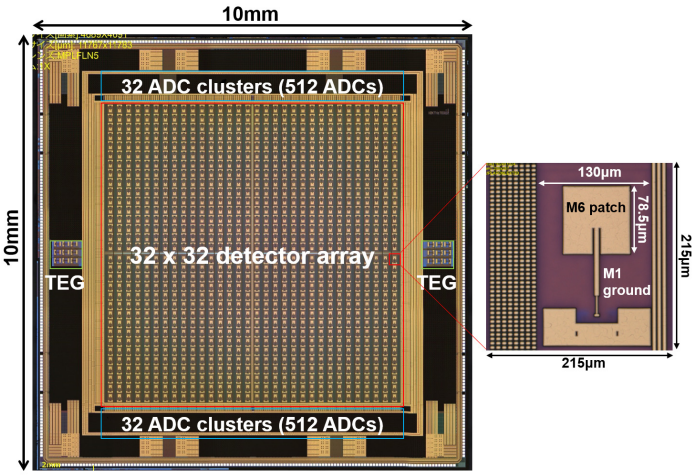


Figure 5.8.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 6 Overview: *Ultra-High-Speed Wireline*

WIRELINE SUBCOMMITTEE

Session Chair: ***Tony Chan Carusone***
University of Toronto, Toronto, Canada

Associate Chair: ***Takayuki Shibasaki***
Fujitsu Laboratories, Kawasaki, Japan

Subcommittee Chair: Frank O'Mahony, Intel, Hillsboro, OR

Digital traffic is aggregated within and between data centers through ultra-high-speed wireline links, which are poised to reach 100Gb/s line rates as the bandwidth per port in data routers keeps increasing. Thus, four papers in this session describe circuits at or above 100Gb/s. Wireline transceiver research is at the forefront of solid-state circuits technology, integrating high-performance analog and digital in the most advanced nanoscale CMOS. For instance, three of the papers in this session are implemented in 7nm FinFET technology. Increasingly complex signal processing and communication techniques are being brought to bear on these problems, with four of the papers relying on digital signal processing. However, practical constraints on the circuits' power consumption and area remain, so that advances in analog/mixed-signal transceiver circuits are the focus of the four remaining papers.

1:30 PM

6.1 A 100Gb/s 1.1pJ/b PAM-4 RX with Dual-Mode 1-Tap PAM-4 / 3-Tap NRZ Speculative DFE in 14nm CMOS FinFET

A. Cevrero, IBM Research, Ruschlikon, Switzerland

In Paper 6.1, IBM Research describes mixed-signal PAM-4 clock-and-data recovery at 100Gb/s. The analog front-end shapes the channel response to permit an efficient implementation of a speculative DFE consuming only 1.1pJ/bit in 14nm FinFET technology.

2:00 PM

6.2 A 60Gb/s PAM-4 ADC-DSP-Based Transceiver in 7nm CMOS with SNR-Based Adaptive Power Scaling Achieving 6.9pJ/b at 32dB Loss

M-A. LaCroix, Huawei Technologies, Ottawa, Canada

In Paper 6.2, Huawei Technologies reports a 60Gb/s ADC-DSP based transceiver in 7nm FinFET technology that automatically and adaptively minimizes its power consumption based upon the channel loss and crosstalk. Over 30% power savings is achieved when the link loss varies from 38dB to 18dB.

2:30 PM

6.3 A Sub-250mW 1-to-56Gb/s Continuous-Range PAM-4 42.5dB IL ADC/DAC-Based Transceiver in 7nm FinFET

M. Pisati, eSilicon Italy, Pavia, Italy

In Paper 6.3, eSilicon presents a PAM-4 transceiver that incorporates a DAC, ADC, and DSP in 7nm FinFET technology. The transceiver consumes less than 250mW and compensates up to 42dB loss at 56Gb/s.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

3:15 PM

6.4 A 180mW 56Gb/s DSP-Based Transceiver for High-Density IOs in Data Center Switches in 7nm FinFET Technology

T. Ali, MediaTek, Irvine, CA

In Paper 6.4, MediaTek describes a 56Gb/s transceiver employing low-power techniques for their 7b DAC and ADC-based front ends. The analog power consumption is 180mW/lane in 7nm FinFET technology.

3:45 PM

6.5 A 400Gb/s Transceiver for PAM-4 Optical Direct-Detect Application in 16nm FinFET

C. Loi, Inphi Corporation, Singapore

In Paper 6.5, Inphi reports on a 400Gb/s 4-lane transceiver for optical communication in 16nm FinFET technology. The quad transceiver occupies 6.16mm^2 while consuming a total of 900mW.

4:15 PM

6.6 A 128Gb/s 1.3pJ/b PAM-4 Transmitter with Reconfigurable 3-Tap FFE in 14nm CMOS

Z. Toprak-Deniz, IBM T. J. Watson Research Center, Yorktown Heights, NY

In Paper 6.6, IBM Research presents a transmitter with reconfigurable 3-tap FFE operating up to 128Gb/s. A 1V_{ppd} output swing is achieved with 1.3pJ/b power consumption in 14nm FinFET technology.

4:45 PM

6.7 A 112Gb/s PAM-4 Voltage-Mode Transmitter with 4-Tap Two-Step FFE and Automatic Phase Alignment Techniques in 40nm CMOS

P-J. Peng, Yuan Ze University, Taoyuan City, Taiwan

In Paper 6.7, Yuan Ze University describes a PAM-4 voltage-mode transmitter incorporating automatic phase alignment in a 4:1 multiplexer. A maximum data rate of 112Gb/s is achieved in 40nm CMOS technology.

5:00 PM

6.8 A 36Gb/s Adaptive Baud-Rate CDR with CTLE and 1-Tap DFE in 28nm CMOS

D. Yoo, University of Toronto, Toronto, Canada

In Paper 6.8, the University of Toronto incorporates automatic equalization and timing adaptation into a mixed-signal receiver that uses the same latches for both baud-rate clock recovery and a 1-tap DFE. The 36Gb/s NRZ receiver consumes 106mW for an 18dB-loss channel in 28nm CMOS technology.

6.1 A 100Gb/s 1.1pJ/b PAM-4 RX with Dual-Mode 1-Tap PAM-4 / 3-Tap NRZ Speculative DFE in 14nm CMOS FinFET

Alessandro Cevrero, Ilter Ozkaya, Pier Andrea Francese, Matthias Brandli, Christian Menolfi, Thomas Morf, Marcel Kossel, Lukas Kull, Danny Luu, Martino Dazzi, Thomas Toifl

IBM Research, Ruschlikon, Switzerland

The growing demand for higher data rates in wireline communications has led to emerging standards in the 100Gb/s+ range using PAM-4 signaling. ADC-based receivers have demonstrated robust operation over channels with high losses (>20dB) [1], but their power consumption (>500mW/lane excluding DSP) is prohibitive for applications requiring large port counts in short-reach chip-to-chip and chip-to-module links (such as OIF-CEI-112G-VSR/XSR/USR). This work demonstrates a dual PAM-4/NRZ RX implemented in 14nm FinFET and measured up to 100Gb/s in PAM-4 mode achieving pre-FEC BER<10^{-12} across a 19.2dB-loss channel with low power. To achieve low BER with >16dB loss channel the RX uses a CTLE combined with a 1-tap speculative DFE. To minimize power consumption, the number of slicing levels to resolve the 1-tap PAM-4 DFE speculation is reduced from 12 to 8 by shaping the channel to a 1+0.5D response (h0+0.5*h0) with CTLE and TXFFE. With a 1+0.5D channel, 4 out of 12 speculative decisions are overlapped. Moreover, comparators used for DFE speculation can be shared for phase detection in a baud-rate CDR scheme, extending the concept proposed in [2] to PAM-4. In NRZ mode, the RX features a fully speculative 3-tap DFE to equalize high loss channels (>35dB) at 56Gb/s. The RX incorporates the following key advances: (1) a trans-admittance trans-impedance (TAS-TIS) CTLE circuit resulting in a low-power and compact analog-front-end (AFE), (2) a 1-tap PAM-4 speculative DFE based on a 1+0.5D response, and (3) a CMOS quadrature-DLL (QDLL) generating quadrature clock phases resulting in a low jitter/power clock path.

Figure 6.1.1 shows the block diagram of the architecture. The RX input signal goes to the 50Ω termination front-end, which includes asymmetric T-COILs to absorb the ESD capacitance achieving >30GHz bandwidth with minimal group delay variation. The AFE consists of a two-stage TAS-TIS CTLE providing both high frequency peaking and long-tail-equalization (LTE). The CTLE output is amplified by the VGA and sampled by the 4-way time-interleaved comparator arrays, whose thresholds can be adjusted individually by resistive-ladder-based voltage-DACs with 9b resolution [3]. Each interleaved array consists of 8 slicers for 1-tap DFE speculation in PAM-4 mode or 3-tap DFE speculation in NRZ mode, as well as baud-rate phase detection. The comparator is a two-stage self-timed design with a single clock phase to minimize clock load. Speculative samples D_{spec} are then deserialized by a 4:32 DMUX and processed by the RX DSP. In PAM-4 mode, the DSP resolves the 1-tap speculation and generates early/late (E/L) signals processed by the CDR logic. Alternatively, the DSP processes NRZ data to resolve the 3-tap DFE and generates E/L information. For both PAM-4/NRZ, the E/L logic consists of a pattern filter that selects the correct speculative samples based on valid data transitions. The CDR logic outputs UP/DN signals that drive the phase-rotator (PR) register. The clock path starts with a 28GHz CML divider that generates I/Q inputs for a 128-step octagonal PR [4]. The PR output is converted to rail-to-tail CMOS levels to drive the QDLL, which creates quadrature clock phases. The QDLL implements a feedback loop that regulates a delay line to match 1UI (Fig. 6.1.1 bottom left). Buffered quadrature clocks drive the comparators directly. For minimum I/Q mismatch the quadrature error is sensed at the buffer output. The proposed I/Q generator topology results in a short clock path, as it is integrated in the fanout of the clock distribution, thus minimizing RMS jitter. The QDLL consumes 14mW from a 0.9V supply.

The circuit schematic of the AFE is shown in Fig. 6.1.2. The TAS-TIS CTLE topology was selected due to its low output impedance. High-frequency boost and LTE are implemented by two transconductor stages, Gm_{HF} and Gm_{LF} whose outputs are combined to generate the signal currents $IOUT_{p,n}$ that feed the TIS. The main CTLE Gm_{HF} consists of two source degenerated differential pairs $P_{1,2}$ whose output currents are superimposed implementing two independent zero-pole pairs. A programmable peaking resistor R_p provides up to 4dB peaking in addition to source degeneration. Up to 14dB AC boost is observed in simulation. Gm_{LF} implements a closely spaced zero-pole pair around 1GHz. Through a current subtraction, the LTE maintains 0 DC gain, whereas at high frequency only the current from differential pair P_3 flows into the TIS, due to the current mirror pole controlled by C_{LF}, which boosts the AC gain. A TAS-TIS VGA amplifies the CTLE signal to 600mV_{ppd}. An inductor in the TIS feedback path is optimally sized to 350pH to drive the 100fF load (32 slicers+wiring) with >20GHz bandwidth. The AFE output common mode can be adjusted by injecting a DC current across the VGA feedback path. Measured AFE noise (including slicer) is 4.5mV_{rms} with the CTLE configured to equalize a 37dB-loss channel.

Figure 6.1.3 presents the speculative DFE concept and phase detection scheme for a 1+0.5D channel. Assuming that the PAM-4 transmitted levels are {-3, -1,+1,+3}, there are 10 distinct levels seen by the RX with a 1+0.5D response. As 4 speculative slicing thresholds are overlapped, only 8 comparators are required to resolve the DFE. Moreover, the signal zero crossing corresponds to the DFE slicing levels so that they are re-used for phase detection as can be seen in the measured PAM-4 eye diagram. Fig. 6.1.3 (bottom) shows the block diagram of the PAM-4 DFE look-ahead logic running at 1/32 rate. After thermometer-to-binary conversion, it performs a selection among pre-calculated look-ahead signals. The gates in the critical path of the final data selection are six 4-to-1 muxes, which are easily accommodated in a single cycle. The 3-tap DFE selection logic uses a similar tree structure. For NRZ operation, 8 slicing levels implement a fully speculative DFE. In this case, to reuse DFE levels for CDR we require that h1=0.5*h0, as in [2]. However, with a 3-tap DFE the pattern filter must select the correct speculative decision based on previous bit history.

The RX was tested with a PRBS15 sequence generated by a 14nm 112Gb/s TX with an 8-tap FFE [6]. The TX, which provides 900mV_{ppd} DC swing, is wire-bonded to a custom PCB. RX dies were tested on a probe station accessing pads via RF probes. Two strip-line channels (Fig. 6.1.4) of different lengths (7.25 and 15.5 inches) were used to investigate equalization performance. The eye measurements and BER tests were performed with the on-die scope and PRBS checker. Bathtub plots have been measured by stepping the internal PR with a resolution of 32 steps/UI. Figure 6.1.4 displays PAM-4 silicon measurements. At 80Gb/s, the 7.25-inch channel can be fully equalized by the 8-tap TX-FFE (7.2dB de-emphasis) followed by the CTLE with 12.5% horizontal margin at BER<10^{-12}. Channel loss is 14.8dB at 20GHz. At 100Gb/s, the 1-tap DFE is essential to ensure error-free operation with 9.4% horizontal margin in the bathtub across a 19.2dB-loss channel. The RX also achieves BER<10^{-12} with a 26dB-loss channel at 80Gb/s. For the PAM-4 measurements, 8 FFE taps were optimized by recording a full PRBS7 sequence with the RX configured as time-correlated ADC and calculating the optimal FFE taps to match a 1+0.5D response. For 56Gb/s NRZ measurements, only 2 FFE taps were sufficient to equalize a 37dB loss channel with the CTLE and 3-tap DFE. The CTLE is optimized to cancel ISI the signal zero crossing at the expense of leaving residual ISI in the eye center as can be seen in the eye diagram of Fig. 6.1.5. However, speculative DFE eyes are fully open as demonstrated by the bathtub plot showing 43.7% horizontal margin at BER<10^{-12}. Figure 6.1.6 presents the RX performance summary and comparison with previous work. To the best of our knowledge, the RX achieves the highest data-rate among analog PAM-4 receivers while being lower power than ADC-based designs in the same speed range. Measured power at 100Gb/s is 111.4mW or 1.1pJ/b. The link power efficiency including the 14nm 8b DAC and 8-tap FFE is 3.7pJ/b.

References:

[1] J. Hudner, et al., "A 112Gb/s PAM4 wireline receiver using a 64-way time-interleaved SAR ADC in 16nm FinFET", *IEEE Symp. VLSI Circuits*, pp. 47-48, June 2018.

[2] T. Shibasaki, et al., "A 56Gb/s NRZ electrical 247mW/lane serial link transceiver in 28nm CMOS", *ISSCC*, pp. 64-65, Feb. 2016.

[3] P.A. Francese, et al., "A 50Gb/s 1.6pJ/b RX data-path with quarter-rate 3-tap speculative DFE", *IEEE Symp. VLSI Circuits*, pp. 267-268, June 2018.

[4] A. Cevrero, et al.,"A 60Gb/s 1.9pJ/b NRZ optical receiver with low latency digital CDR in 14nm CMOS FinFET", *IEEE Symp. VLSI Circuits*, pp. C320-C321, June 2017.

[5] E. Depaoli, et al., "A 4.9pJ/b 16-64Gb/s PAM-4 VSR transceiver in 28nm FDSOI CMOS", *ISSCC*, pp. 104-106, Feb. 2018.

[6] C. Menolfi, et al.," A 112Gb/S 2.6pJ/b 8-Tap FFE PAM-4 SST TX in 14nm CMOS", *ISSCC*, pp. 112-114, Feb. 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

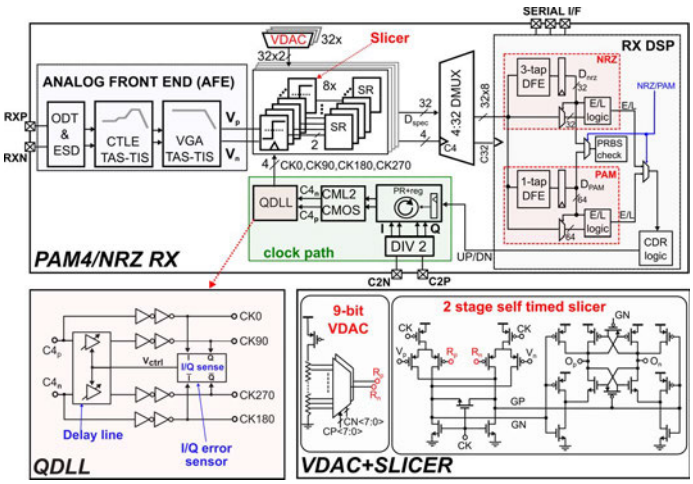


Figure 6.1.1: Receiver block diagram (top) and circuit details of the QDLL (bottom left) and slicer (bottom right).

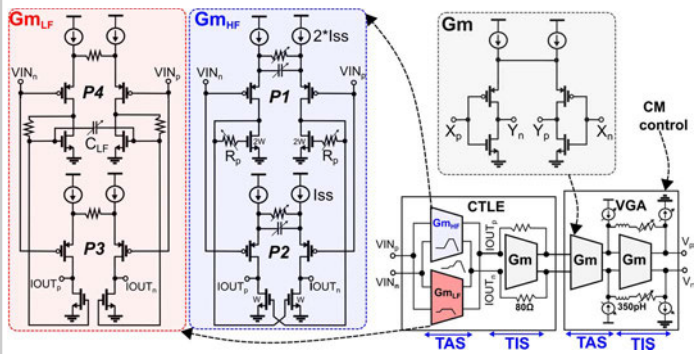


Figure 6.1.2: Analog-front-end block diagram (bottom right), circuit details of CTLE and VGA.

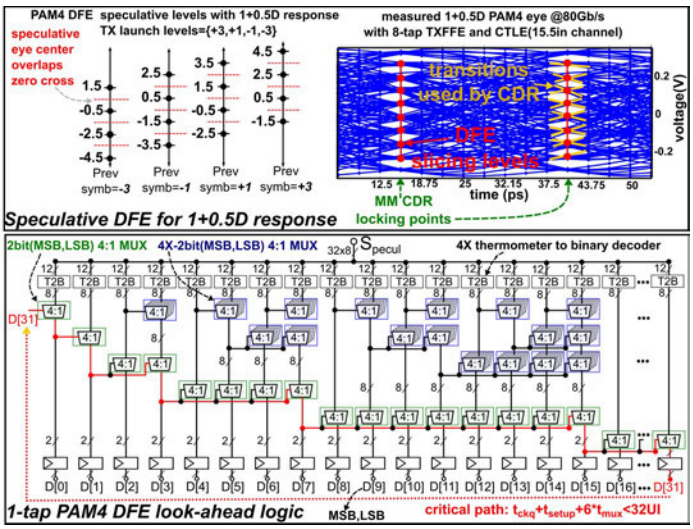


Figure 6.1.3: DFE concept with 1+0.5D response (top), 1-tap PAM-4 DFE look-ahead logic circuit diagram.

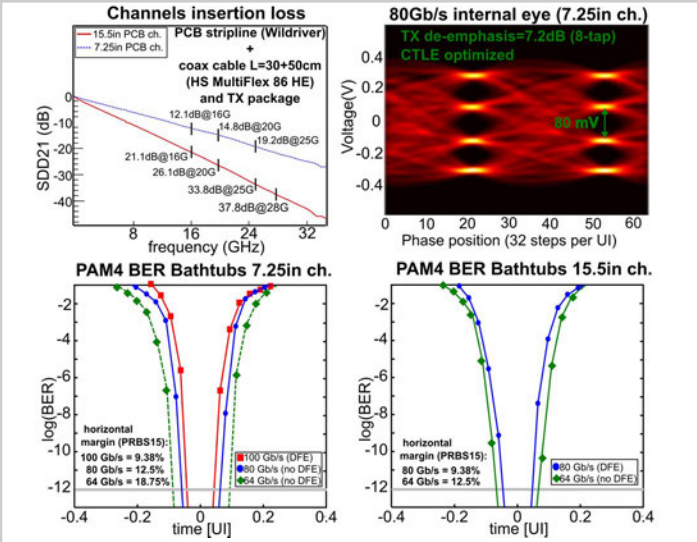


Figure 6.1.4: Channel insertion losses, measured PAM-4 eye-diagram captured with on-die scope and measured bathtub curves.

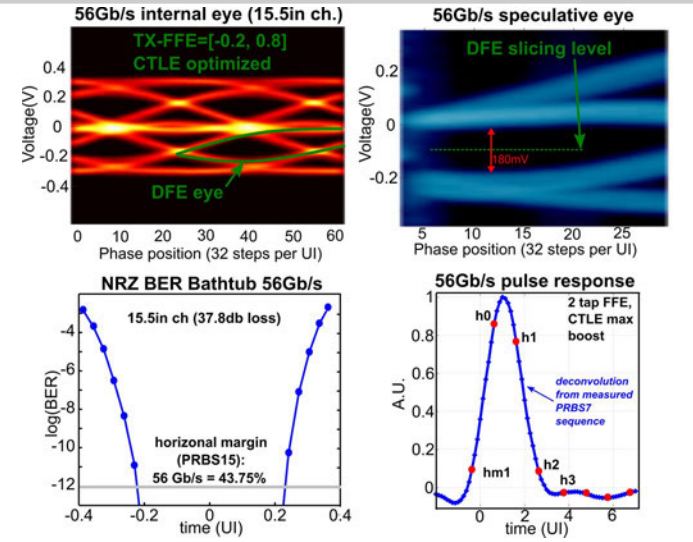


Figure 6.1.5: NRZ eye diagrams (top), BER-bathtub curve and measured pulse response (bottom).

Measured RX power@ 56Gbaud	Analog Front-end CML Divider+PR biasing (1V)	32X Slicers CML2CMOS, QDLL 64X VDACs (0.9V)	Aligner 8x DMUX 4:32 RX DSP (0.8V)
111.4mW	41.63mW	27.80mW	41.96mW

	[2]	[3]	[5]	[1]	This work	
Technology	28nm CMOS	14nm FinFET	28nm FDSOI	16nm FinFET	14nm FinFET	
Modulation	NRZ	NRZ	PAM4	PAM4	NRZ	PAM
Data-rate [Gb/s]	56	50	64	112	56	100
Loss at Nyq. [dB]	18.4	32	16.8	20	37.8	19.2
Link Equalization	2-tap TXFFE CTLE 1-tap DFE	2-tap TXFFE CTLE 3-tap DFE	4-tap TXFFE CTLE	3 tap TXFFE CTLE, ADC based 32-tap RXFFE 1-tap DFE	2-tap*** TXFFE CTLE, 3-tap DFE	8-tap TXFFE CTLE, 1-tap DFE (1+0.5D)
Pre FEC BER	<1E-12	<1E-12	<1E-12	2E-5	<1E-12	
PRBS	15	15	Q13	31	15****	
CDR	Baud rate(1+0.5D)	No CDR	Edge based	Baud rate	Baud rate(1+0.5D)*****	
Area [mm2]	1.40*	0.019	0.32	0.67**	0.053	
efficiency [pJ/bit]	2.53	1.6	2.8	5.26**	2	1.1

*2-lane transceiver area reported
**DSP Power/area not included
***pre and main cursor only configured to stress CTLE/DFE performance
****PRBS15 pattern used since this is the longest sequence fitting in the TX on chip-SRAM
*****in NRZ mode CDR locks to a signal with 3 postcursor taps provided that h(1)=0.5*h(0) to reuse speculative levels

Figure 6.1.6: Receiver power breakdown at 56Gb/s NRZ and comparison with previously reported work.

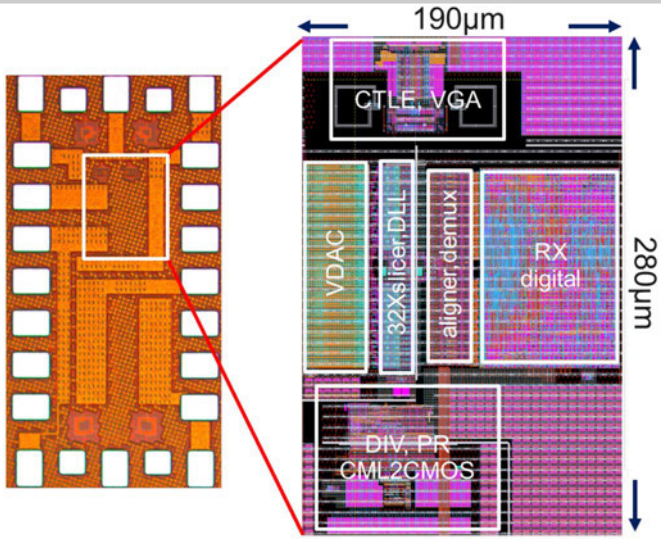


Figure 6.1.7: RX micrograph and layout details in 14nm CMOS FinFET technology.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

6.2 A 60Gb/s PAM-4 ADC-DSP Transceiver in 7nm CMOS with SNR-Based Adaptive Power Scaling Achieving 6.9pJ/b at 32dB Loss

Marc-Andre LaCroix, Henry Wong, Yun Hua Liu, Huong Ho, Semyon Lebedev, Petar Krotnev, Dorin Alexandru Nicolescu, Dmitry Petrov, Carlos Carvalho, Stephen Alie, Euhan Chong, Faisal Ahmed Musa, Davide Tonietto

Huawei Technologies, Ottawa, ON, Canada

With the introduction of PAM-4 signaling at 56Gb/s and the decreased benefits of CMOS scaling for high-speed mixed-signal designs, SerDes designers and system architects are faced with severe performance versus power budget constraints. Power management and energy efficiency have become the main drivers for system design. However, industry standards such as EEE have failed to keep up with efficiency demands. In this context the choice between a so-called analog mixed signal (AMS) SerDes architecture vs. an ADC-DSP-based one has been debated at length. AMS provides significantly lower maximum power [2,4] while ADC-DSP provides higher link margin [1] thus avoiding expensive and power hungry repeater ICs that largely negate the power advantage of AMS SerDes in a system. AMS provides an easier and cheaper approach to implement multi-tap DFEs [3] compared to DSP where it is typically very expensive to implement more than a 1-tap DFE. This paper will show an ADC-DSP SerDes transceiver with a 2-tap DFE is capable of operating error-free over a 38dB link yet having an overall power budget similar to AMS. The same basic SerDes architecture is implemented (Fig. 6.2.1) with minor differences in 16nm and 7nm FinFET, however, power scaling is incorporated into the 7nm version only.

Clocks are generated from a 156.125MHz reference by two identical and independent LC-PLLs, each one continuously tunable from 2.5-15GHz to support independent data rates from 8-60Gb/s. A Nyquist-rate clock from each PLL is distributed to each lane via single ended CMOS repeaters. Using the available clock distribution power budget on a single ended rather than differential buffer chain allows more power to be spent on each buffer thus achieving the highest drive capability and fastest raise/fall times possible in the technology. A shielding structure, shown in Fig. 6.2.2, surrounds each conductor achieving better than 80dB isolation in the frequency range of interest, while minimizing the overall parasitic capacitance. Individual on-die power supply regulators for the buffers and in-line passive DCD filtering reduces jitter accumulation along the chain. The jitter measured between 7MHz and 10GHz at the TX output of the closest and furthest lanes to the PLL is $160fs_{rms}$.

The dual-mode PAM-2/PAM-4 transmitter operates up to 60Gb/s using a differential voltage-mode transmitter with programmable 3-tap FIR equalizer. Both 16nm and 7nm versions were tested, the main difference being the addition of an asymmetric T-coil in 7nm. The achieved efficiency of 1.3pJ/b (in 7nm), is better than those reported in [2,6]. Output eye-diagrams are shown in Fig. 6.2.3. The 7nm version can generate amplitudes up to $1.2V_{ppd}$ across 100Ω. The addition of the T-coil for the 7nm version has yielded 17% improvement in eye openings.

The receiver is shown in Fig. 6.2.4. The AFE includes the termination, a passive attenuator, and an inductor-less CTLE. The CTLE consists of either 1-stage (7nm) or 2-stages (16nm). Each stage is comprised of a trans-admittance (TAA) and trans-impedance amplifier (TIA). To maximize the available high-frequency boost, the signal is applied to the source-degenerated pair (M1-M2) and coupled to M3-M4 via a programmable high-pass filter. Each CTLE generates 20dB boost at Nyquist while also providing 10dB of DC gain range. DC gain is controlled via resistors Rf_1 and Rf_2 to minimize gain-boost interaction and simplify adaptation. The 7nm version consumes 22mA, which is superior to recently published 50+Gb/s CTLEs [3]. The AFE is followed by 4 interleaved T&Hs clocked at 7.5GHz. The flip-around T&Hs acquire the input signal over a time span of approximately 1UI leaving 3UI to transfer the sample to one of the eight subsequent asynchronous SAR sub-ADCs per cycle. A single global timer maintains the ASAR connection sequence. The sub-ADC resolution can be adjusted from 2b to 7b and is powered from a dynamically controlled power supply.

DSP performs channel equalization and clock-and-data recovery. Equalization is achieved with a 15-tap FFE and a 2-tap DFE. Two DFE taps enable superior crosstalk immunity over a wide-range of insertion losses compared to previously reported ADC-DSP receivers with only 1 tap [1,5].

Previous work has shown coarse and partial power scaling [1]. In this work we demonstrate progressive scaling against link IL achieving ~10mW per dB of IL and netting 40% reduction between 38 and 18dB when operating at 56Gb/s. Our method dynamically controls and balances system wide parameters including: ADC resolution, DSP tap count, size of the CDR decision vector, and on-die LDO regulator voltages. Furthermore, AVS (adaptive voltage scaling) control of the digital, and analog power supplies provides further power optimization globally across all lanes. The system is depicted in Fig. 6.2.1.

As a first step, global AVS is performed using feedback from several PVT sensors embedded in each SerDes lane. One such sensor measures the average conversion time (ACT) for each asynchronous SAR sub-ADC. The method is depicted in Fig. 6.2.4. ACT is not dependent upon a clock cycle and varies with PVT, thus this metric provides direct visibility of local conditions. Along with other metrics ACT is used to inform upon optimal power supply conditions. Given the large local on-die variations of advanced FinFET technologies, AVS is conservatively driven by the worst case sensor reading from a SerDes lane in a given ASIC.

Adaptation of local parameters is achieved automatically for each lane. The raw SNR is used as figure of merit since raw BER is not readily available in a conventional SerDes. Our solution was to create an estimated SNR figure of merit where LMS error or slicer error represents the noise and the adapted main cursor represents the signal. The SNR and other related metrics are calculated within the SU and are provided to the performance controller via its interface.

In the first phase of the adaptation, RX and TX equalization parameters are optimized to obtain the lowest possible raw BER for the given link. This is obtained using a combination of LMS for the DSP, heuristic SNR-driven search for CTLE and standard link training for TX. Once this is completed, a power optimization phase begins whereby a wide variety of parameters are evaluated for their impact on link performance. Over the course of this process raw BER remains between 10^{-5} and 10^{-6}. This target provides approximately 1 order of magnitude margin at the KP4 FEC input ensuring 10^{-15} at the output. The optimization is autonomous and performed entirely on-chip.

The 7nm design is compared with a previous version in 16nm without power scaling. Both testchips include either a 4-lane or 8-lane (7nm) SerDes macro, and 4-lane KP4 FEC with SU. Figure 6.2.5 shows the maximum tolerable ICN versus loss for a raw BER ≤1E-05 when operating at 56Gb/s with full power, plus individual data points confirming transceiver performance at the maximum data rate, 60Gb/s. In order to test power scaling potential we assumed a system with many channels whose total insertion loss including the package varies from 18dB to 38dB as per the distribution in Fig. 6.2.5. An average integrated crosstalk noise (ICN) of 2mV is applied for all loss cases. At ≥38dB the 16nm and 7nm versions consume 625mW and 457mW, respectively. Power scaling of the 7nm version is demonstrated against the assumed channel distribution. Results show a reduction in average power consumption of ~20% yielding an effective per channel power of 364mW. The reduction is equivalent to the difference between our 16nm and 7nm versions at full power. This work is compared against other recent 50-64Gb/s PAM-4 transceivers in Fig. 6.2.6.

Acknowledgement:
The authors thank the HiLink physical design and application teams for their contributions.

References:
[1] P. Upadhyaya, et al., "A fully adaptive 19-to-56Gb/s PAM-4 wireline transceiver with a configurable ADC in 16nm FinFET," *ISSCC*, pp. 108-110, Feb. 2018.
[2] E. Depaoli, et al., "A 4.9pJ/b 16-to- 64Gb/s PAM-4 VSR Transceiver in 28nm FDSOI CMOS," *ISSCC*, pp. 112–113, Feb. 2018.
[3] P. Francese, et al., "A 50Gb/s 1.6pJ/b RX Data-Path with Quarter-Rate 3-tap Speculative DFE," *IEEE Symp. VLSI Circuits*, pp. 267–268, Oct. 2018.
[4] J. Im, et al., "A 40-to-56Gb / s PAM-4 Receiver with 10-Tap Direct Decision-Feedback Equalization in 16nm FinFET," *ISSCC*, 2017, pp. 114–115.
[5] K. Gopalakrishnan, et al., "A 40/50/100Gb/s PAM-4 Ethernet Transceiver in 28nm CMOS," *ISSCC*, 2016, pp. 62–63.
[6] L. Wang, et al., "A 64Gb/s PAM-4 Transceiver Utilizing an Adaptive Threshold ADC in 16nm FinFET," *ISSCC*, 2018, pp. 110–111.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

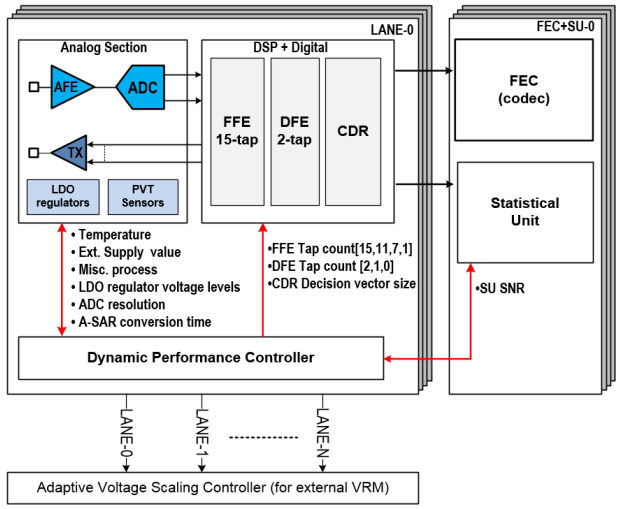


Figure 6.2.1: 4-Lane macro featuring adaptive power scaling based on per-lane BER. Each lane is connected to a FEC and statistical unit. The entire system is implemented on chip.

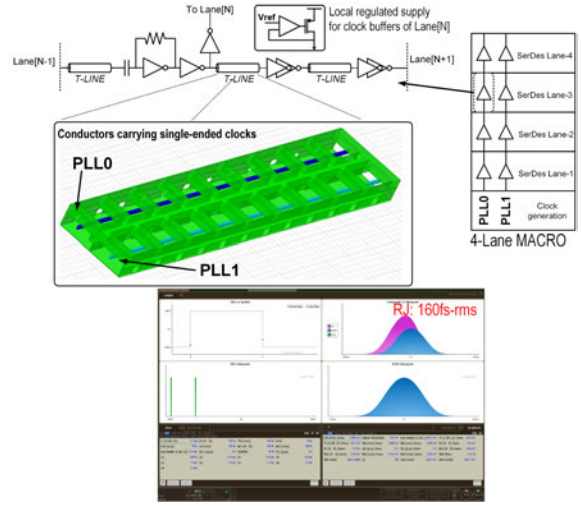


Figure 6.2.2: Single-ended clock distribution of Nyquist rate clock from PLL0 (PLL1 method is identical), shielding structure to minimize interference between the clocks, jitter decomposition of 28.125Gb/s 1010 pattern from TX.

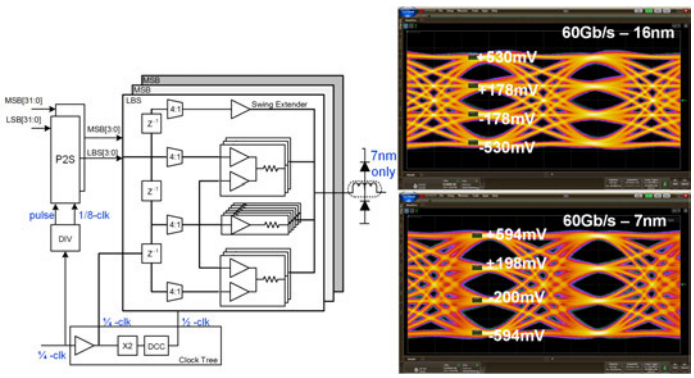


Figure 6.2.3: 60Gb/s PAM-4 TX: block diagram, and de-embedded PAM-4 PRBS13Q eyes for 16nm and 7nm versions. 7nm version achieves 1.2V_{ppd}.

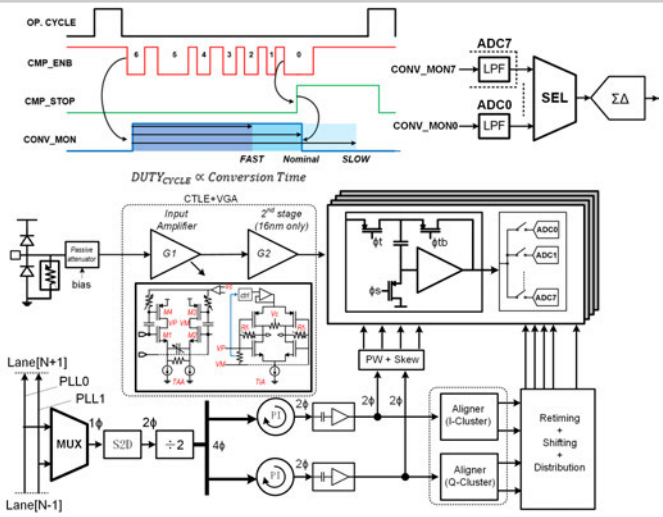


Figure 6.2.4: PAM-4 ADC based RX: Block diagram of analog, including depiction of system and method for asynchronous SAR sub-ADC conversion cycle measurement.

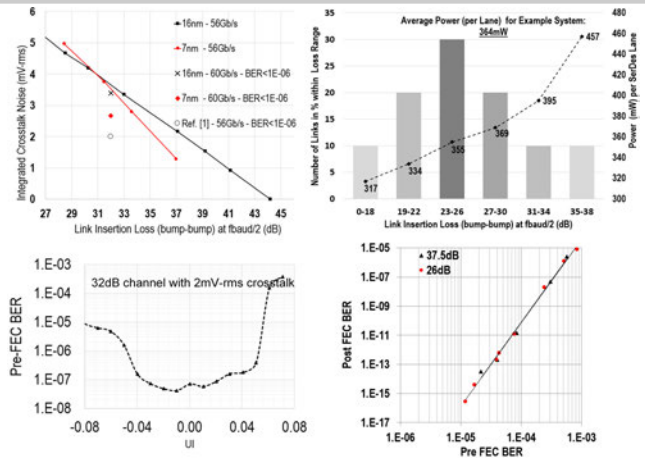


Figure 6.2.5: Maximum ICN tolerance at 56Gb/s (ICN referred to the bump) versus link insertion loss while achieving Pre-FEC BER < 1E-05, power scaling versus insertion loss across link distribution for Pre-FEC BER < 1E-05 with crosstalk of 2mV_{rms}, RX BER across a range of sampling phases, post-FEC versus pre-FEC BER for 37.5dB and 26dB loss channels.

	[Xilinx ISSCC 2018]	[Wang ISSCC 2018]*	This Work		
Technology	16nm	16nm	16nm	7nm	
Data rate [Gbps]	56	64	56	56	60
Area [mm²] for single lane -With clock gen. (for 4-lane) -Transceiver only	2.20	0.25*	1.048 0.807	0.840 0.615	
Supply Voltages [V]	0.85, 0.9, 1.2, 1.8	0.9, 1.2	0.8, 0.9, 1.2	0.75, 0.8, 1.2	
Max ICN (BER < 1E-06) @32dB	2	–	3.39	2.7	2.7
Rx CTLE boost [dB], # stages	–, 3	6, 1	20, 2	20, 1	
Rx ADC res. [b]	7	6	7	7	
Rx DSP # taps FFE / DFE	14 / 1	–, –	15 / 2	[15,11,7,1] / [2,1,0]	
Max. Swing [V_{pp-dif}]	1.0	1.1	1.1	1.2	
Tx RJ [fs_{rms}]	180	162	160	160	
Tx RLM	0.98	0.99	0.99	0.99	
Power/Lane@32dB [mW] Analog / DSP	545 325 / 220	374 / -	625 400 / 225	395 290 / 104	414 303 / 111
Efficiency/Lane [pJ/bit] @32dB	9.7	5.8* @ 29dB	11	7.0	6.9

*Excludes DSP

Channel Loss @ 14GHz	Total Lane Power [mW] / Efficiency [pJ/bit]	Total Lane Power [mW] / Efficiency [pJ/bit] excluding clocking LDOs
>38dB	457 / 8.1	407 / 7.2
32dB	395 / 7.0	345 / 6.1
24dB	355 / 6.3	305 / 5.4
18dB	317 / 5.6	267 / 4.7

Figure 6.2.6: Comparison of 4-PAM DSP-based 50 to 64Gb/s wireline transceivers. Summary of 56Gb/s power consumption and efficiency for various channels.

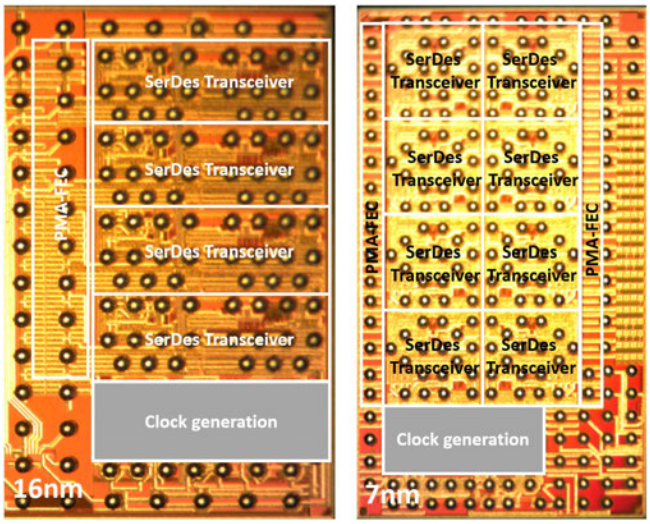


Figure 6.2.7: Microphotographs of 16nm 4-lane, and 7nm 8-lane SerDes transceiver macros.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

6.3 A Sub-250mW 1-to-56Gb/s Continuous-Range PAM-4 42.5dB IL ADC/DAC-Based Transceiver in 7nm FinFET

Matteo Pisati, Fernando De Bernardinis, Paolo Pascale, Claudio Nani, Marco Sosio, Enrico Pozzati, Nicola Ghittori, Federico Magni, Marco Garampazzi, Giacomino Bollati, Antonio Milani, Alberto Minuti, Fabio Giunco, Paola Uggetti, Ivan Fabiano, Nicola Codega, Alessandro Bosi, Nicola Carta, Demetrio Pellicone, Giorgio Spelgatti, Massimo Cutrupi, Andrea Rossini, Roberto Massolini, Giovanni Cesura, Ivan Bietti

eSilicon Italy, Pavia, Italy

The enormous availability of data on the internet and the increased computational power available in the cloud require a constant increase in the bandwidth needed to transport data across the network. In order to mitigate the cost of infrastructure, the industry has proposed standards for 56Gb/s PAM-4 interfaces that can support legacy channels in conjunction with forward error correction (FEC) techniques. The reduced signal to noise ratio (SNR) available with PAM-4 modulation vs. NRZ requires sophisticated signal processing, making DSP-based solutions very attractive in terms of performance but, so far, at the expense of extra power vs. classical analog solutions. This paper demonstrates that this trade-off can be broken and presents a compact ADC/DAC DSP-based long-reach transceiver in 7nm FinFET technology that operates seamlessly from 3.5-to-60Gb/s in PAM-4 (from 1-to-30Gb/s in NRZ mode) and consumes less than 250mW at 56Gb/s. The presented IP extends the deployment of DSP-based transceivers from repeater/re-timer applications to big ASICs with a very large number of ports.

The overall transceiver architecture is shown in Fig. 6.3.1. The receiver (RX) front-end consists of the on-die-termination (ODT) and a 2-stage continuous-time linear equalizer (CTLE). The CTLE drives a 40-way time-interleaved (TI) SAR-ADC. The receiver DSP consists of a 17-tap FFE working concurrently with a 1-tap speculative DFE and a reflection-canceling FFE, which implements 4 individually roaming taps. Clock recovery is achieved on a dedicated low latency data-path consisting of a 5-tap FFE, slicer, TED and loop filter driving a dedicated LC-DCO. Additional blocks perform the TI ADC gain and sampling time background calibrations, and the ADC offset correction. Receiver adaptation is based on hardware LMS loops and firmware control and optimization through an on-board MCU (not shown). The transmit section consists of the PMA functions, a variety of pattern generators, a 5-tap FIR section and a terminated DAC as the analog transmitter. The overall architecture has been implemented with extensive analog/digital co-design at the architecture and circuit levels that fully exploits the benefits of technology scaling in order to minimize power consumption and area while delivering high levels of performance.

In the RX analog section, a T-coil structure, L1 + L2 in Fig. 6.3.2, is resonating with the ESD parasitic capacitance and a peaking inductor L3, in series with the 50Ω termination, extends the RX bandwidth. Each CTLE stage is made of an AC coupler and a g_M stage followed by a trans-impedance amplifier (TIA) [1]. The AC coupler of the first stage embeds the low-frequency compensation (LFC) [2] equalizer. The g_M stage consists of a PN RC-degenerated transconductor feeding the TIA to obtain the optimal gain and noise performance. The value of the degeneration capacitor is programmable to provide boost and slope control on the transfer function. The gain of the CTLE is controlled by programming the current in the g_M stage, with the target of providing a full swing signal ($500mV_{ppd}$) to the ADC input. The TIA consists of a replica-bias regulated inverter with feedback resistor. Such a closed-loop topology is used to keep the linearity of the chain above 35dB (required by the PAM-4 signal), even with the relatively high output swing required by the ADC. In the CTLE first stage, the TIA feedback also includes series inductive peaking (L_F) to increase the bandwidth of the stage. The CTLE and ADC stages are connected through an AC coupler stage that also acts as the insertion point for the baseline wander (BLW) DAC.

The 7b 40-way time-interleaved ADC uses a hierarchical architecture with 8 front-end THs running at 3.75GS/s. Each TH is connected to an array of five 750MS/s non-binary charge-redistribution DAC-based asynchronous SAR ADCs [3] sharing a common reference buffer (Fig. 6.3.2). The TH consists of a top plate sampling stage connected to a super-follower-based buffer for distributing the sampled signal to the SAR array. Starting from the half-rate input clock, an ADC clocking block generates the 8 TH clocks at 8UI and the 40 SAR clocks at 40UI. The TH sampling phases are calibrated through analog current-controlled delay lines, driven by the static timing calibration (STC) block in the DSP, to reduce the relative skews below 130fs. Offset mismatches stemming mainly from the SAR comparators are corrected in the DSP while gain mismatches can be calibrated in analog on a per-TH base acting on the reference buffer.

A single LC-tank DCO provides the low-jitter clock to the ADC. A custom "candy shaped" (Fig. 6.3.7) inductor has been designed to minimize magnetic pulling-crosstalk. The VCO covers the 7.5G-to-15GHz tuning range by switching MoM capacitor banks. The CDR loop is closed through the digital DTL acting on a 7b vtune-DAC that controls the tuning voltage of the varactors. In the test chip described in this paper, the TX-PLL is not present, although a placeholder is put in the die area. However, as shown in Fig. 6.3.1, the recovered clock from the CDR has been routed to the transmitter (TX) block that can re-transmit the received data in a reference-less fashion.

The TX architecture is shown in Fig. 6.3.3. The TX is a 7b 1-to-30GS/s digital-to-analog converter. PAM-4/NRZ reconfigurability and equalization are obtained using a digital 5-tap FIR filter that directly provides 7b data samples to the TX-DAC. A half-rate SST driver with series analog impedance control has been chosen due to intrinsic high bandwidth and low power consumption [4]. The output swing with this implementation is $0.9V_{ppd}$. The 2:1 serializer stage is embedded in the driver section to minimize the delay between the final data re-timing and the output nodes. The T-coil is used to improve termination matching. Duty-cycle correction is implemented through a 6b DAC injecting current into a dedicated clock buffering stage in front of the TX main clock-tree to adjust impairments in the clock distribution.

The prototype is fabricated in TSMC 7nm FinFET CMOS. The transceiver area including DSP is 390µm × 1206µm (Fig. 6.3.7). The CTLE measured 14GHz boost and gain are 14dB and 19.9dB respectively as shown in Fig. 6.3.4. The same figure reports the performance of the ADC and CTLE as a function of input frequency after calibration of TI artifacts and an ADC output spectrum with a -1dBFs 13.4GHz signal. Since these measurements have been done without a CDR loop, close in phase noise within 2MHz offset has been considered as input signal power. After this effect is de-embedded, Nyquist SNDR/-THD are 32dB/39.8dB while offset and gain/skew spur powers are -40dBc and -46dBc respectively.

Link measurements are shown in Fig. 6.3.5 for 2 different channels with measured total insertion loss (IL) of 30dB and 42.5dB (bump to bump) at 14GHz respectively. In performing this measurement, a reference TX is used to output a 56Gb/s PRBS31 signal to the channel input. With the optimum setting for the CTLE, a pre-FEC BER better than 1×10^{-7} is measured on the longer channel, while the LMS in the DSP continuously adapts the FFE and DFE coefficients. The bathtub plot is showing that the communication over the 30dB channel is error-free. Fig. 6.3.5 also shows the internal RX eye after digital equalization (for the 40" channel case) and the jitter tolerance results in NRZ mode for 2 different channel lengths, showing the large CDR bandwidth that can be obtained. Fig. 6.3.3 shows a 60Gb/s transmitted PAM-4 eye diagram after FIR equalization.

The overall power consumption is 234mW at 56Gb/s, leading to 4.17pJ/b efficiency. The analog RX draws 91.5mW from a 0.9V supply while the TX, including the recovered clock distribution, draws 44mW. The digital part consumes only 98.5mW from a 0.75V supply with 17-tap FFE, 1-tap DFE, DTL, CDR, and LMS enabled. In the production device, a TX-PLL based on the VCO described above will be added to clock the TX data with low jitter. The projected overall power consumption of this complete solution is 244mW. The performance of the complete lane is compared against recently published solutions in Fig. 6.3.6 and achieves the best power efficiency and the best FoM.

References:
[1] H. Kimura, et al., "28Gb/s 560mW Multi-Standard SerDes with Single-Stage Analog Front-End and 14-Taps Decision-Feedback Equalizer in 28nm CMOS", *ISSCC*, pp. 38-39, Feb. 2014.
[2] S. Parikh, et al., "A 32Gb/s Wireline Receiver with a Low Frequency Equalizer, CTLE and 2-Tap DFE in 28nm CMOS," *ISSCC*, pp. 28-29, Feb. 2013.
[3] P. Upadhyaya, et al., "A Fully Adaptive 19-to-56Gb/s" PAM-4 Wireline Transceiver with a Configurable ADC in 16nm FinFET", *ISSCC*, pp. 108-109, Feb. 2018.
[4] C. Menolfi, et al., "A 112Gb/s 2.6pJ/b 8-Tap FFE PAM-4 SST TX in 14nm CMOS", *ISSCC*, pp. 104-105, Feb. 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

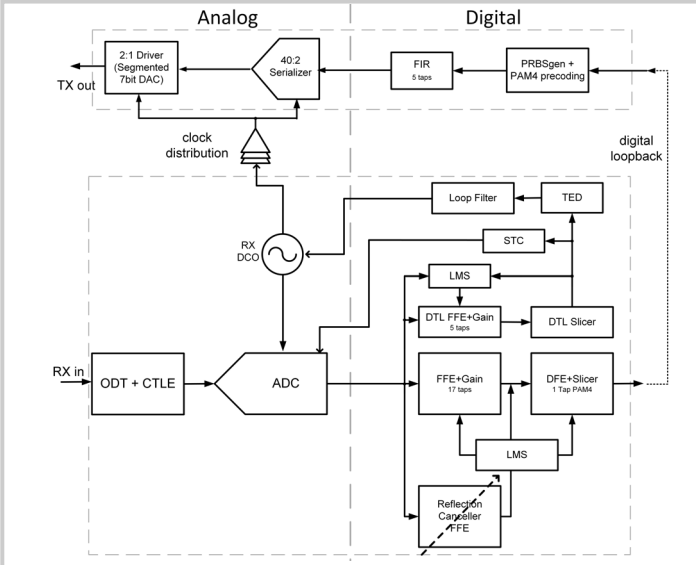


Figure 6.3.1: Transceiver architecture.

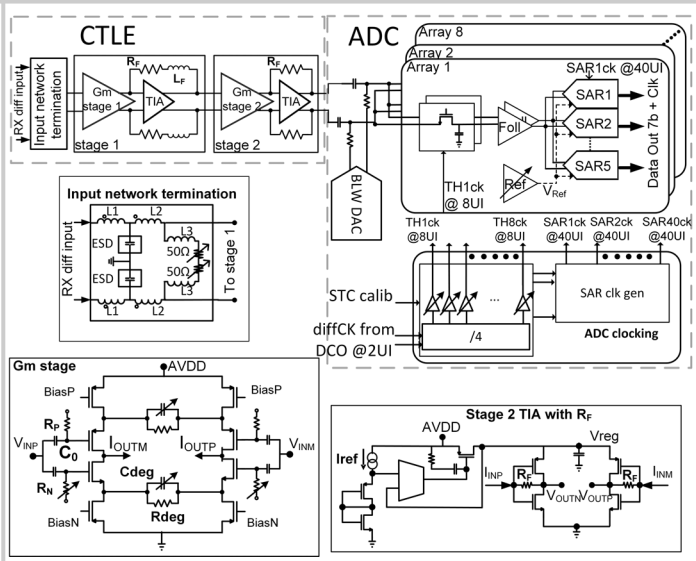


Figure 6.3.2: RX analog-front-end and ADC block diagrams.

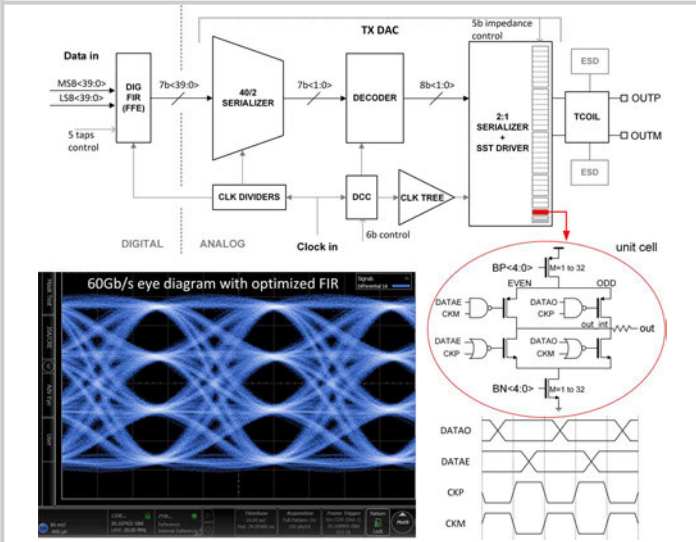


Figure 6.3.3: TX block diagram with the driver stage embedding the last 2-to-1 serializer stage, and related eye diagram.

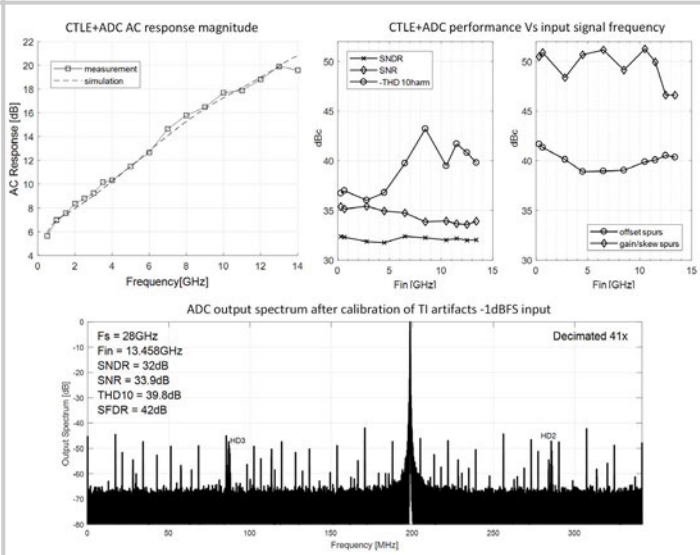


Figure 6.3.4: RX measurements.

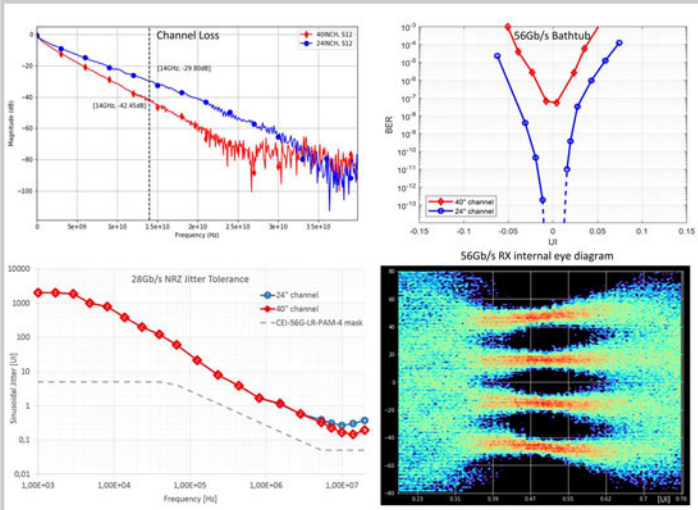


Figure 6.3.5: Channel response showing the IL of the 2 analyzed links, and the corresponding RX bathtub and jitter tolerance plots; internal RX eye diagram after DSP equalization for the 40" channel.

	This work	JSSC 2017 Frans	ISSCC 2018 Upadhyaya [3]	ISSCC 2018 Wang	ISSCC 2018 Depaoli
Technology	**7nm FinFET**	16nm FinFET	16nm FinFET	16nm FinFET	28nm FDSOI
Power Supply [V]	**0.9 / 0.75**	1.2 / 0.9	1.8/1.2/0.9/0.85	1.2 / 0.9	1
Data rate [Gb/s]	**1 – 60**	9.5 - 56	9.5 - 56	64.375	64
IL @ Nyquist [dB]	**42.5 @ 56Gb/s**	31 @ 56Gb/s	32 @ 56Gb/s	29.5 @ 64.375Gb/s	16.8 @ 64Gb/s
Area / lane [mm²]	**0.468**	1.4	2	0.25	0.44
Power / lane @56Gb/s	**234mW (a) current implementation** **244mW (b) projected complete implementation**	550 (DSP power not included)	545	373.6	315
Power efficiency [pJ/bit]	**4.1 (a) / 4.36 (b)**	9.8	9.7	5.8	4.9
FoM [pJ/bit/dB]	**0.098 (a) / 0.102 (b)**	0.29	0.3	0.19	0.29

Figure 6.3.6: Table of comparison with state of the art.

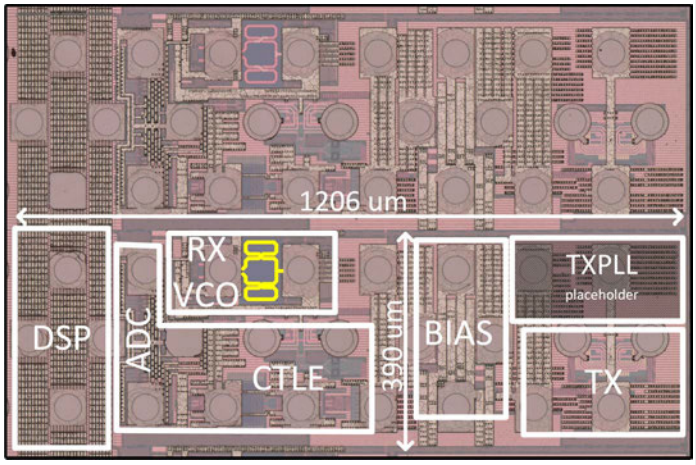


Figure 6.3.7: Die photo (2 lanes).

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

6.4 A 180mW 56Gb/s DSP-Based Transceiver for High-Density IOs in Data Center Switches in 7nm FinFET Technology

Tamer Ali[1], Ramy Yousry*[1], Henry Park*[1], Ehung Chen*[1], Po-Shuan Weng[2], Yi-Chieh Huang[2], Chun-Cheng Liu[2], Chien-Hua Wu[2], Shih-Hao Huang[2], Chungshi Lin[2], Ke-Chung Wu[2], Kun-Hung Tsai[2], Kai-Wen Tan[2], Ahmed ElShater[1], Kuang-Ren Chen[2], Wei-Hao Tsai[2], Huan-Sheng Chen[2], Weiyu Leng[1], Mazen Soliman[1]

[1]MediaTek, Irvine, CA
[2]MediaTek, Hsinchu, Taiwan
*Equally-Credited Authors (ECAs)

A soaring amount of data transfer has been witnessed in recent years. By 2020, 50 billion connected devices are expected, which will generate more than 2 zettabytes of data traffic annually. Given the constraints in power & space, this explosive growth puts high stress on network infrastructure, which demands low power, high BW, and area efficient transceivers. To improve BW efficiency, modern transceivers use PAM-4 instead of NRZ to double the throughput in the same BW. However, PAM-4 introduces substantial ISI, reduces peak-to-average-ratio, and imposes non-linearity constraints compared to NRZ modulation. Scaling in CMOS technology assisted the rise of DSP-based transceivers and digital equalization schemes to compensate PAM-4 non-idealities and achieve higher SNR at the receiver output [1-3].

This work presents a 56Gb/s DSP-based transceiver built in state-of-the-art 7nm FinFET technology that targets high-density IO switch chips (256+ lanes). The paper focuses on the analog side of the transceiver macro. Each macro has one PLL and bias circuits, 4 TX and RX lanes, and global clock distribution. Figure 6.4.1 shows the block diagram of the RX. At the input of the RX, a wide bandwidth resistive network provides termination and attenuation up to 14dB to prevent signal clipping at the RX under large input swing. A continuous-time linear equalizer (CTLE) provides up to 8dB of peaking to relax the required ADC resolution. It also improves pulse shape for robust timing recovery. A single-stage VGA amplifies the CTLE output by up to 8dB to maximize SNR at the ADC input. The ADC is a 32-way time interleaved ADC, with a 4-way T/H sampler to reduce VGA loading and simplify clocking. The T/H comprises a PMOS top-plate sampling switch, with 25% and 75% tracking & holding duty cycles, respectively. Switch size is optimized for BW and leakage trade-off, without need for bootstrapping. Sampled data is buffered and fed to a sub-ADC. RX front-end, T/H and buffers are all powered by a 1V supply.

Each sub-ADC has 8 asynchronous 7b SARs, and a token generator that generates an 875MHz token pulse for each SAR. Figure 6.4.2 shows the block diagram of the 7b asynchronous SAR ADC powered from a 0.8V supply. One redundant bit is added to improve the ADC immunity to the noisy supply. The first 5 conversion cycles incorporate a differential split-CDAC to stabilize comparator common-mode (CM) voltage, offset, noise, and SAR speed. The last 3 cycles, with minor CM impact, implement monotonic CDAC switching for simpler logic and smaller area. Figure 6.4.2 also shows the SAR double tail comparator. Because comparator speed and noise vary with a wide range across PVT, the size of tail devices is controlled by a digitally assisted loop to keep noise and speed within the design target. For instance, in a slow corner, the SAR may not produce the *conversion_done* signal within the desired period. The digital loop senses the absence of *conversion_done* and enlarges the tail device size to speed-up the comparator and restore timing margin, while achieving best noise performance. In a fast corner, the digital loop reduces tail device size to slow down the comparator to allow just enough timing margin for SAR conversion, while minimizing comparator noise. Offset calibration is performed by an auxiliary differential pair which tracks main differential pair offset variations over VT. Gain calibration is achieved by a variable cap C_P, which changes the V_{REF} value seen at the CDAC.

The TX is a 7b SST DAC driver (Fig. 6.4.3) to alleviate the complexity of building PAM-4 analog pre-emphasis taps. It also gives flexibility to implement any number of digital pre-emphasis or sliding taps to correct package reflection. The DAC driver is powered directly from a 1V supply, while the rest of the TX data path uses a 0.8V supply. The pre-driver uses 0.15V and 0.85V LDOs to interface between the 1V driver stage and prior 0.8V stages. This supply scheme protects the driver devices from exceeding the maximum tolerable voltage limit in 7nm (0.96V). TXP and TXN data streams are clocked separately, which allows correction of channel skew between the TXP/TXN outputs. As shown in Fig. 6.4.3, delay between CLKP and CLKN is programmed at startup with a VCO-based TDC to account for any on-die or channel skew.

High-quality clock distribution and generation is key in PAM-4 transceivers. In the TX, PJ and RJ have a tight budget, while in the RX, timing mismatches between clock phases and jitter dictate the SNR performance at frequencies close to Nyquist. A tuned standing-wave clock distribution structure is implemented to distribute the clock to 4 TX and RX lanes as shown in Fig. 6.4.4(a). Two shunt inductors are placed close to middle section of the structure to set boundary conditions for the transmission line (TL) and tune nearly equal amplitudes at all drop points. Because of its tuned nature, only one clock buffer is needed to drive the TL, which is crucial for low power and low jitter accumulation compared to traditional clock repeaters. CM termination is necessary to reject CM ripple and reduce single-ended clock PJ. This is achieved by adding a capacitor in each inductor center tap. The receiving buffers, shown in Fig. 6.4.4(a), have cross-coupled miller capacitors to increase buffer input impedance and achieve larger swing.

And injection-locked phase rotator (INJPR) is adopted in the RX clock path instead of a traditional PI [4] due its inherent advantages (Fig. 6.4.4(b)). Differential-to-quadrature phase generation occurs inside the INJPR, which saves area and power of the global clock routing with smaller I-Q phase mismatches. The circuit generates 4 divide-by-2 clocks. It also reduces input static clock mismatches, and filters jitter beyond the INJPR BW. A 32-phase VCO with 8-fractional phase injection enables 8b phase control. 7GHz quadrature outputs from the INPR are then applied to AC-coupled buffers with built-in skew adjustment circuit. I_{SKEW} is generated by tracking the INJPR VCO swing, and used as the reference current in the duty-cycle control circuit. The biased duty cycle is then converted into skew adjustment by a NAND operation inside the T/H clock gen to produce accurate quadrature clocks with 75% duty cycle (Fig. 6.4.1). Building skew adjustment inside the AC-coupled buffer alleviates the need for additional buffers with slower edges, which deteriorates random jitter and PSRR.

The transceiver includes all-digital calibration loops, baud-rate timing recovery and the DSP engine. The DSP gears 5-tap TX pre-emphasis, 12-16 tap programmable RX FFE, 1-tap DFE, sliding-tap DFE, and 1-tap MLSD. It is packaged in a large ASIC package with ~3dB Nyquist loss in round trip TX and RX. Figure 6.4.5 shows the RX front-end and ADC performance, which achieves 36.1dB SNDR (5.71 ENOB) at DC, and 30.3dB (4.74 ENOB) at Nyquist. Performance at Nyquist is strictly limited by jitter performance, as shown in Fig. 6.4.5(c), the SNR plot tracks 225fs PLL jitter from 10GHz to Nyquist. The RX offset and gain mismatch are calibrated to ~9b performance, while skew is calibrated to ~6b, thanks to well-matched analog layout and background calibration loops. Figure 6.4.6 shows the TX eye diagram at 56Gb/s which achieves 37dB SNDR and 98.5% RLM. The table in Fig. 6.4.6(b) compares transceiver performance with recently published DSP-based PAM-4 designs. The transceiver is able to achieve a BER of $1E^{-5}$ with a 40dB channel loss, that is 5dB more loss than [2]. For a 35dB channel loss the transceiver is able to achieve a BER of $3E^{-9}$, almost 2 orders of magnitude lower than [2]. All link measurements are performed with near-end and far-end cross talk aggressors. The analog transceiver occupies an area of $0.32mm^2$/lane and consumes 180mW/lane, with PLL, biasing and clock distribution are amortized per lane. Quad-lane transceiver die photo is shown in Fig. 6.4.7.

References:

[1] P. Upadhyaya, et al., "A fully adaptive 19-to-56Gb/s PAM-4 wireline transceiver with a configureable ADC in 16nm FinFET", *ISSCC*, pp. 108-110, Feb. 2018
[2] K. Gopalakrishnan, et al., "A 40/50/100Gb/s PAM-4 Ethernet transceiver in 28nm CMOS", *ISSCC*, pp. 62-63, Feb. 2016.
[3] D. Cui, et al., "A 320mW 32Gb/s 8b ADC-based PAM-4 analog front-end with programmable gain control and analog peaking in 28nm CMOS", *ISSCC*, pp. 58-59, Feb. 2016.
[4] Y-C Huang, et al., "An 8-Bit Injection-locked Phase Rotator with Dynamic Multiphase Injection for 28/56/112Gbps Serdes Application", *ISSCC*, Feb. 2019.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

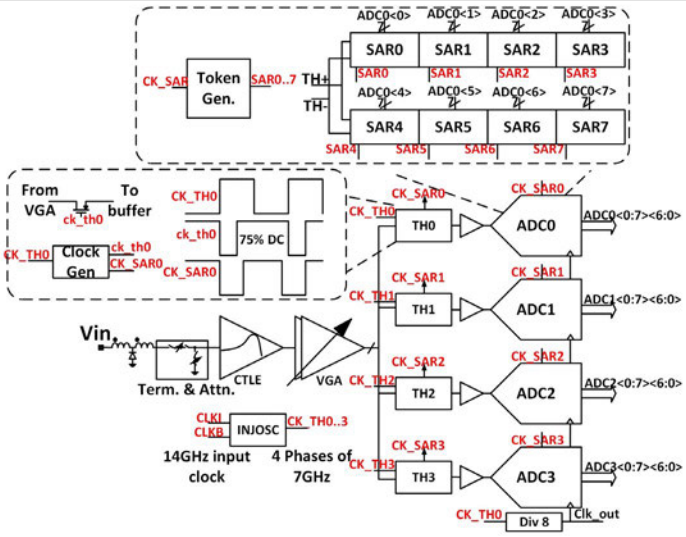


Figure 6.4.1: RX block diagram, inset showing sub-ADC and token gen for generating SAR clocks.

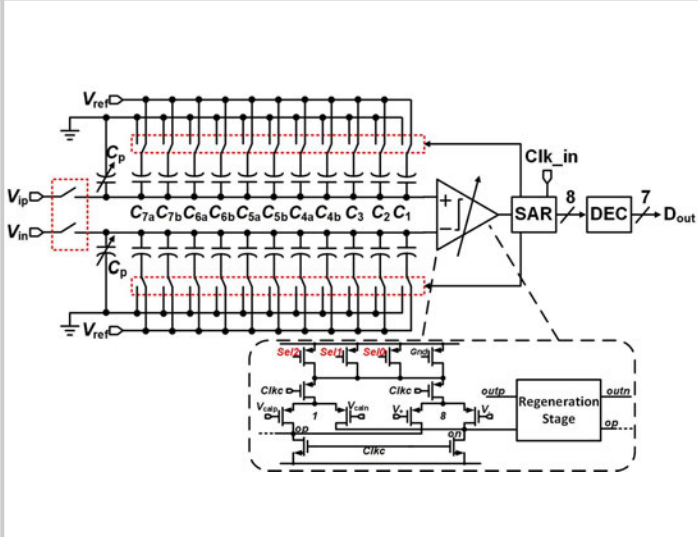


Figure 6.4.2: SAR ADC block diagram, inset shows double tail comparator.

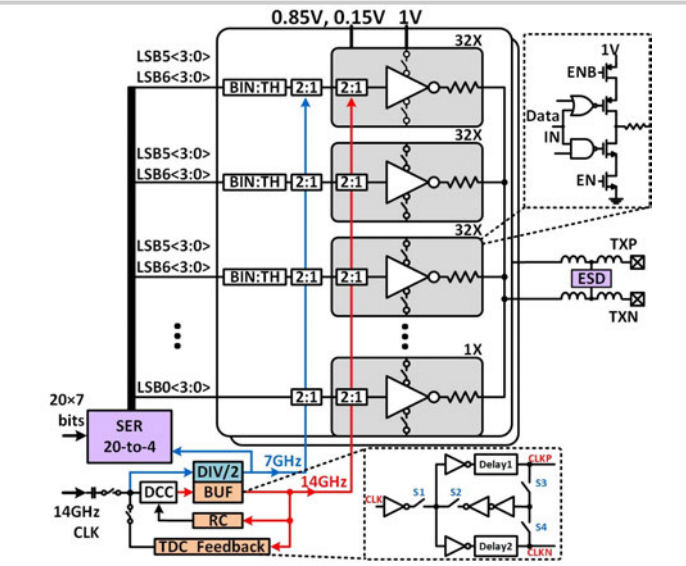


Figure 6.4.3: TX block diagram, showing skew calibration.

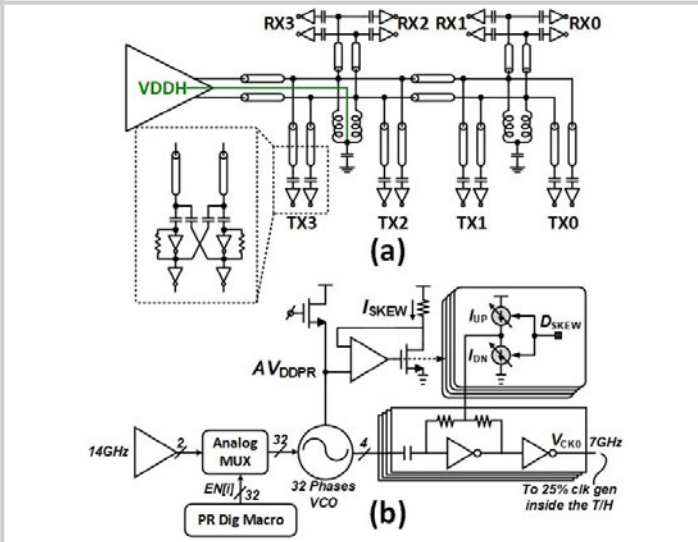


Figure 6.4.4: (a) Clock distribution channel, (b) injection-locked phase rotator and skew adjustment circuit.

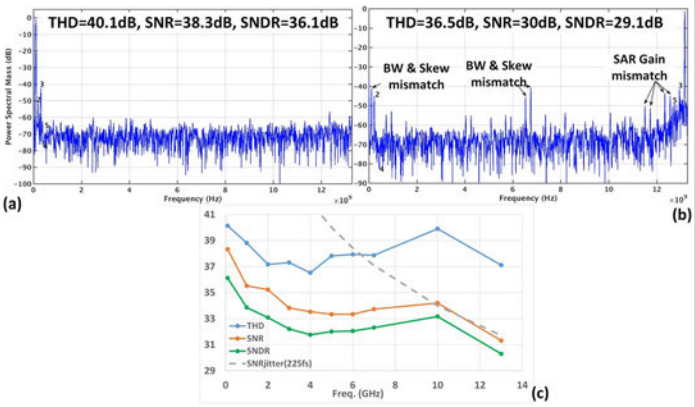
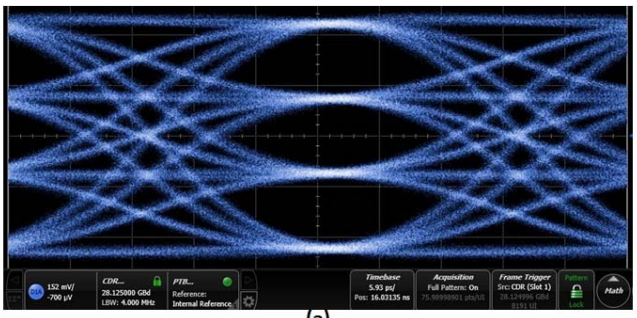


Figure 6.4.5: RX performance plots, (a) low-freq., (b) Nyquist, (c) THD and SNR vs. frequency.



(a)

	[1]	[2]	[3]	This Work
Technology	16nm	28nm	28nm	7nm
Data Rate (Gb/s) Channel loss at Nyquist (dB)	56 32 @ BER<1e-6	56 35 @ BER<2e-7	32 32 @BER N/A	56 >35 @ BER <3e-9 40 @ BER <1e-5
Analog Power (mW/Lane) Analog Area (mm^2/lane)	325 1.2	165 (RX ADC only) 3.2	320 (RX only) 0.89 (RX only)	180 0.32
TX Architecture	CML 3-tap FIR	CML 3-Tap FIR	N/A	7-bit DAC 5-tap FIR
RX Architecture	Analog: 7-bit ADC/CTLE DSP:14-FFE/1-DFE	Analog: 7-bit ADC/CTLE DSP: N/A	Analog:8-bit ADC/CTLE DSP:N/A	Analog:7-bit ADC/CTLE DSP:12-16-FFE/1-DFE/SDFE/MLSD

(b)

Figure 6.4.6: (a) TX eye diagram at 56.25Gb/s, (b) comparison table.

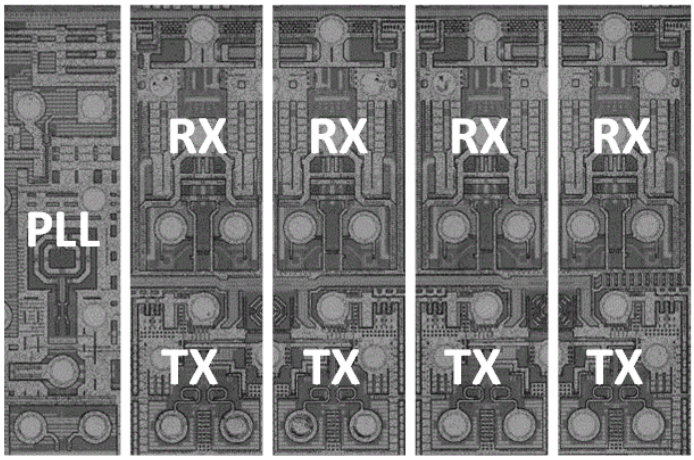


Figure 6.4.7: Transceiver photograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

6.5 A 400Gb/s Transceiver for PAM-4 Optical Direct-Detect Application in 16nm FinFET

C. Loi[1], A. Mellati[2], A. Tan[3], A. Farhoodfar[3], A. Tiruvur[3], B. Helal[3], B. Killips[4], F. Rad[3], J. Riani[3], J. Pernillo[3], J. Sun[1], J. Wong[3], K. Abdelhalim[2], K. Gopalakrishnan[3], K. Kim[2], L. Tse[3], M. Davoodi[2], M. Le[2], M. Zhang[3], M. Talegaonkar[2], P. Prabha[2], R. Mohanavelu[3], S. Chong[1], S. Forey[4], S. Netto[3], S. Bhoja[3], W. Liew[1], Y. Duan[3], Y. Liao[3]

[1]Inphi Corporation, Singapore, Singapore
[2]Inphi Corporation, Irvine, CA
[3]Inphi Corporation, Santa Clara, CA
[4]Inphi Corporation, Northants, United Kingdom

PAM-4 has become the mainstream modulation format for short-reach (<2km) optical interconnects at data rates of 50Gb/s and above [1,2]. For more efficient utilization of optical bandwidth in large-scale data centers, data rates of 100Gb/s per wavelength are required. This data rate imposes aggressive requirements on transceiver bandwidth (BW) and jitter. In addition, due to thermal constraints, transceivers must consume low power to help meet optical module power budgets specified by the multisource agreements (MSA) such as QSFP-DD and OSFP. This paper presents a low-power and low-jitter 4×100Gb/s transceiver for PAM-4 optical interconnects.

An overview of the transceiver in a typical optical application is shown in Fig. 6.5.1. The baud rate per channel is 53GBaud/s due to FEC overhead. Each transmitter (TX) uses a 53GS/s DAC. An FIR filter pre-compensates for the channel bandwidth limitation. Digital pre-distortion adjusts PAM-4 inner levels for optimum SNR performance. The receiver (RX) is a 53GS/s self-calibrated time-interleaved ADC. The sampled signal is equalized by the DSP using FFE and a reflection canceller. The DSP employs different detection schemes, including nonlinear DFE and MLSD, to provide optimal tradeoff between performance and power consumption under different channel conditions. A Mueller-Muller (MM) CDR is used for timing recovery.

The RX input network is critical to maintain good signal integrity in the interface of the optical sub-assemblies. Figure 6.5.2 shows the RX analog front-end using a T-coil input network to optimize return loss while achieving a maximally flat response. An additional inductor (L_2) is inserted between the ESD circuit and variable gain amplifier (VGA) to further extend BW. The DSP adjusts VGA gain (-2 to 2dB) to compensate for external TIA output amplitude variation and optimize the dynamic range at the ADC input. The VGA adopts an NMOS gm/gm architecture to minimize gain variations over PVT. Its common-mode output is set by controlling the gate bias of $M_{3,4}$ through a common mode feedback loop using an OTA with charge pump. Vpump is set higher than the supply to optimize for maximum swing and linearity. To drive the large input capacitance of the 53GS/s ADC, the VGA uses large NMOS transistor loads ($M_{3,4}$). The inductive nature looking into the source of the NMOS load pair ($M_{3,4}$) provides partial equalization of the capacitive output load.

The 53GS/s ADC consists of 64 7b SAR-ADCs as shown in Fig. 6.5.3. It employs a 2-rank hierarchical time-interleaved sampling architecture to reduce the VGA load and the number of jitter-critical clock signals [3], [4]. Rank-1 consists of 16 time-interleaved TAHs that individually operate at 3.3GHz. To further reduce the VGA load, 25% duty cycle clocks are used to drive the TAH, so that only 4 of them are closed at any given time. Rank-2 consists of 64 sample-and-hold switches (SAH) embedded in the SAR-ADCs. 16 source-followers are used to isolate the large input capacitance of the SAR-ADCs from the Rank-1 TAH switches. Together, the TAH and SAH switches function as 16× 1-4 analog demultiplexers enabling a SAR conversion rate of 828MS/s. As shown in the timing diagram, these SAH switches are closed only during the hold time of Rank-1 TAH switches, which leaves 1ns hold time for the 7b SAR conversion. In a synchronous SAR-ADC, excessive comparator delay caused by meta-stability events may lead to incomplete bit-conversion of MSBs. To alleviate this issue, an asynchronous SAR architecture is used in this work. As the probability of meta-stability events decays exponentially with time, the MSBs are designed to settle with significantly longer time to resolve their meta-stability compared to the LSBs. Therefore, the asynchronous SAR improves the error rates of more significant bits at the cost of losing less significant bits, thus effectively converting large-amplitude errors into smaller amplitude. This SAR ADC design has a meta-stability error-rate below the noise floor, corresponding to 1e-15 BER. Digital calibration algorithm similar to [5] is used to correct inter-channel gain, offset, and sampling phase errors.

Figure 6.5.4 shows the TX architecture. It consists of a 7b main DAC (2 MSB thermometer bits and 5 binary bits) and a 2b predistortion-DAC to improve linearity. The DAC driver provides $1.1V_{ppd}$ swing from a 1.2V supply. The CML differential pairs are fully switched using CMOS inverters. An on-chip T-coil improves both the BW and return loss by shielding the DAC parasitics and ESD diodes. Predistortion is used to reduce DAC compression and ensure that the ratio-of-level-mismatch (RLM) is always >0.95. A bleeder raises the output common-mode and a cascode current DAC improves the linearity of the TX. The pre-driver adopts a resistor feedback inverter topology for a sub-UI equalizer to improve BW and ISI [6]. The serializer adopts a 2:1 MUX at the final stage. A duty cycle correction circuit corrects the 26.5GHz clock driving the 2:1 MUX. A half-rate clock architecture is chosen over a quarter rate clock as it simplifies the clock path to achieve optimum jitter performance. Separate voltage regulators are used to isolate the self-generated noise from data path to the clock path for better DJ performance.

At high clock frequencies, clock distribution over long distances can significantly degrade jitter performance due to jitter amplification. To reduce the high frequency clock routing path length, one TX PLL is shared between two TXs while each RX utilizes independent RX PLL/CDR as shown in Fig. 6.5.1. A wide-bandwidth PLL architecture (Fig. 6.5.5) is chosen to mitigate the effect of VCO pulling. The PLL consists of an XOR phase/frequency detector (XOR-PFD) [7] followed by a programmable-offset charge pump (CP), a conventional proportional-integral analog loop filter, a VCO and a multi-modulus divider (MMD). The phase offset between feedback clock and reference clock can be set by choosing an appropriate ratio between UP current, I_{CP}, and down offset current, αI_{CP}, in the charge pump. The combination of XOR-PFD and CP mitigates $\Delta\Sigma$ noise folding due to CP nonlinearity while minimizing the noise contribution from CP current. To ensure robust operation at 26.5GHz, without a pre-scaler, the MMD utilizes both dynamic and static ÷2/3 cells with NAND/NOR logic. A CMOS voltage-biased VCO with two compact tail inductors for reduced flicker noise up-conversion is implemented to cover 15% tuning range.

The chip is fabricated in a standard 16nm FinFET process. Figure 6.5.2 shows that the measured RX analog front-end bandwidth is >32GHz over temperature and the measured differential input return loss (Sdd11) is < -10dB at 26.5GHz. The measured receiver SNDR is 34.5dB at 1GHz, 34dB at 10GHz and 31dB at 26.5GHz, as shown in Fig. 6.5.3. The measured SNDR of the TX at 53GBaud/s is 34.5dB with a RLM=0.95. In NRZ mode, the measured TX ISI DDJ is 2.3ps. The PLL achieves an integrated jitter (10kHz-100MHz) of $197fs_{rms}$ and $207fs_{rms}$ in integer-N mode and fractional-N mode, respectively, with an in-band fractional spur of -69.6dBc at 1MHz offset (Fig. 6.5.5). Figure 6.5.6 shows eye diagrams at different locations in an optical link. The RX jitter tolerance (JTol) performance over an electrical channel with 10dB loss at Nyquist, meets JTol mask at 1e-6 BER, with additional 0.05UI RJ added. Each RX/TX consumes 900mW (8.5pJ/b) with clocking power amortized across the lanes. The total area of the quad RX/TX is $6.16mm^2$.

Acknowledgments:
The authors would like to thank the entire Inphi CE team, especially G. Thomas, G. Mallareddy, A. Bhargav, M. Ren and J. Jiang for the valuable lab support.

References:
[1] P. Upadhyaya, et al., "A Fully Adaptive 19-to-56Gb/s PAM-4 Wireline Transceiver with a Configurable ADC in 16nm FinFET," *ISSCC*, pp. 108-109, Feb. 2018.
[2] L. Wang, et al., "A 64Gb/s PAM-4 Transceiver Utilizing an Adaptive Threshold ADC in 16nm FinFET," *ISSCC*, pp. 110–112, Feb. 2018.
[3] L. Kull, et al., "A 90GS/s 8b 667mW 64x interleaved SAR ADC in 32nm digital SOI CMOS", *ISSCC*, pp. 378-379, Feb. 2014.
[4] Y. Duan, et. al., "A 6b 46GS/s ADC with >23GHz BW and sparkle-code error correction", *IEEE Symposium on VLSIC*, pp. C162-C163, June 2015.
[5] M.Q. Le, et al., "A Background Calibrated 28GS/s 8b Interleaved SAR ADC in 28nm CMOS," *IEEE CICC*, pp. 1-4, Apr. 2017.
[6] M. Erett, et al., "A 126mW 56Gb/s NRZ Wireline Transceiver for Synchronous Short-Reach Applications in 16nm FinFET," *ISSCC*, pp. 102-103, Feb. 2018.
[7] M. Perrott, et al., "A 27 mW CMOS fractional-N synthesizer/modulator IC," *ISSCC*, pp. 366-367, Feb. 1997.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

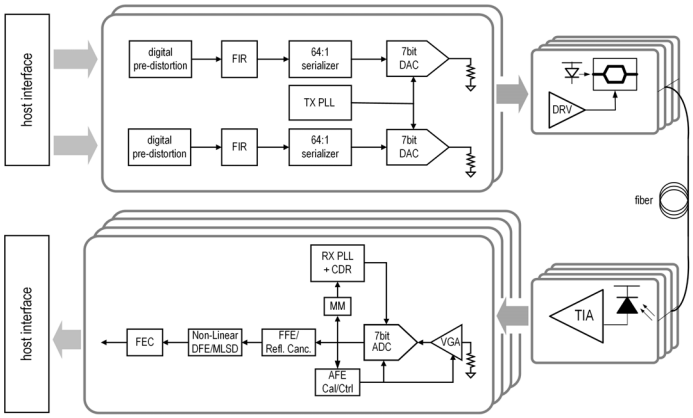


Figure 6.5.1: Block diagram of the 400Gb/s PAM-4 DSP-based transceiver in an optical direct-detect application with external modulator drivers and TIAs.

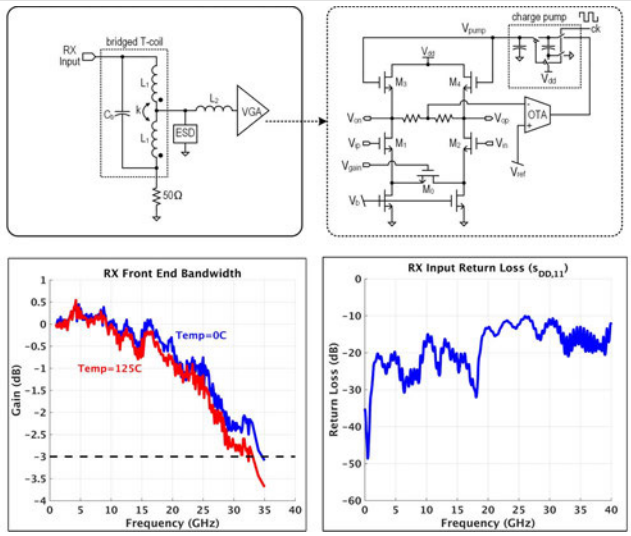


Figure 6.5.2: RX termination and variable gain amplifier (VGA) schematic. Gain transfer function measured at TAH output. RX input differential return loss.

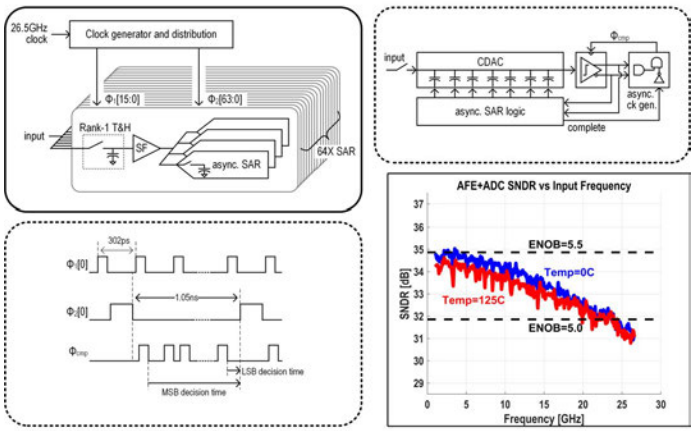


Figure 6.5.3: ADC architecture consists of 64 SARs. Block diagram of the asynchronous SAR. Timing diagram of TAH and SAH, and AFE+ADC SNDR measurement.

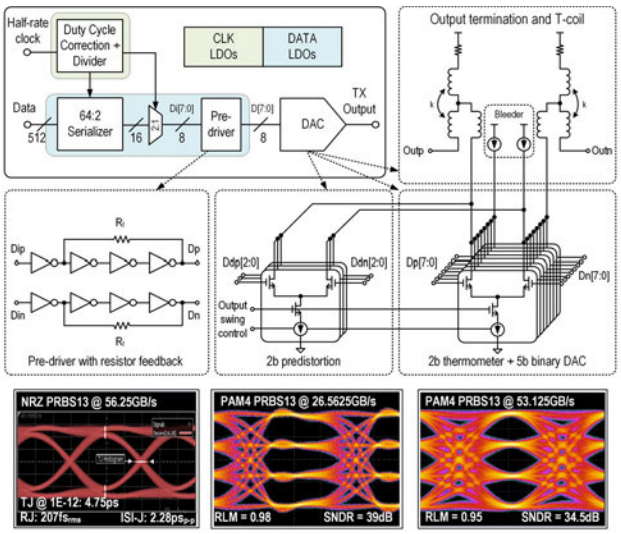


Figure 6.5.4: TX architecture showing different power domains. Pre-driver and output DAC stage schematic. Eye diagram of NRZ/PAM-4 at different baud rates.

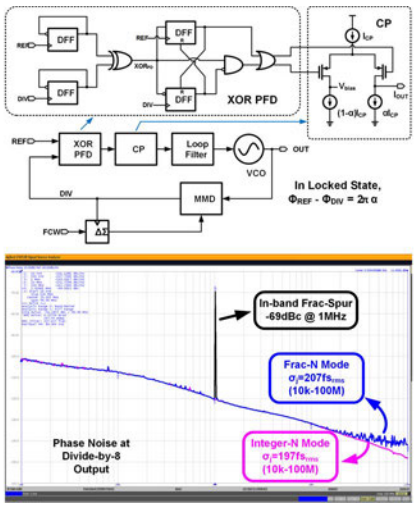
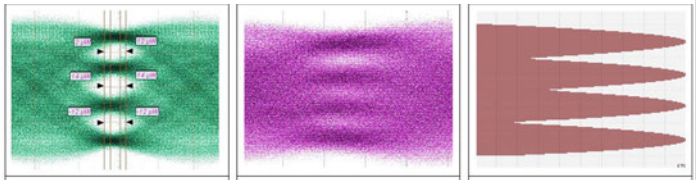


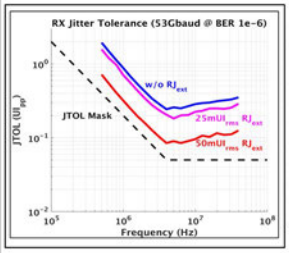
Figure 6.5.5: PLL architecture and XOR PFD schematic. Phase noise measurement shows Frac-N spurs and integrated RJ in integer-N and Frac-N mode.



Electro-Absorption Modulated Laser (EML) optical output.

Electrical eye at TIA output. Optical modulation amplitude (OMA) at optical receiver = -6dBm.

Histogram of recovered PAM4 data (pre-FEC)



	[1]	[2]	This work
Technology	16nm FinFET	16nm FinFET	16nm FinFET
Power Supply (V)	0.85/0.9/1.2/1.8	0.9/1.2	0.75/1.1/1.2
Architecture	ADC-based DSP	ADC-based DSP	ADC-based DSP
ADC resolution	7b	6b	7b
PAM4 Data Rate (Gb/s)	19 - 56	64.375	53/106
Max TX Swing (V)	1	1	1.1
TX RLM	-	0.99	0.95
TX RJ (fsrms)	180	162	197
Energy/bit (pJ/b)	9.5*	6.2p**	8.5
Area/lane (mm^2)	2.20	0.2525***	1.54

* at 32dB IL channel ** at 29.2dB IL channel (excluding CDR, PLL and equalizer)
*** TX/RX active area only

Figure 6.5.6: Eye diagrams / signal histogram at different optical link locations. Jitter tolerance over electrical channel (10dB loss at Nyquist) with RJ added.

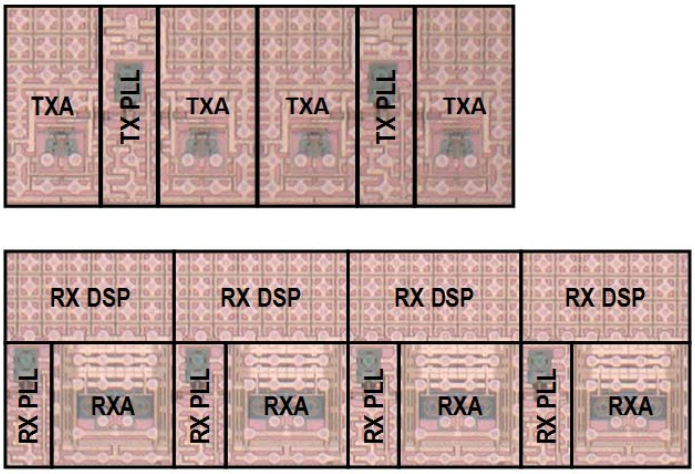


Figure 6.5.7: Die photo of TX dual (two transmit channels sharing single PLL) and four RX(s) in a 4×100Gb/s optical direct detect DSP chip in 16nm FinFET.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

6.6 A 128Gb/s 1.3pJ/b PAM-4 Transmitter with Reconfigurable 3-Tap FFE in 14nm CMOS

Zeynep Toprak-Deniz, Jonathan E. Proesel, John F. Bulzacchelli, Herschel A. Ainspan, Timothy O. Dickson, Michael P. Beakes, Mounir Meghelli

IBM T. J. Watson Research Center, Yorktown Heights, NY

The ever-increasing demand for higher bandwidth continues to fuel the need for faster and more power-efficient IOs, with the next generation high-speed serial links expected to reach data rates higher than 112Gb/s using PAM-4 signaling [1-3]. While PAM-4 spectral efficiency is better than that of NRZ, it is less tolerant of residual ISI and noise. As a consequence, a driver with high bandwidth and large output amplitude is required. This paper presents a 64Gbaud PAM-4 TX with a fully reconfigurable 3-tap FFE, which achieves a power efficiency of 1.3pJ/b in PAM-4 mode and 2.7pJ/b in NRZ mode for a differential output swing of $1V_{ppd}$. A feature of the FFE construction is the use of fully re-assignable FFE segments among the 3 taps, which allows a reduced number of segments for lower capacitance and higher driver bandwidth. To minimize power consumption, a quarter-rate clocking architecture is adopted with a tailless 4:1 multiplexer, which also acts as a pre-driver to a tailless CML output driver.

The quarter-rate TX architecture implementing the fully reconfigurable 3-tap baud-spaced FFE is shown in Fig. 6.6.1. The data path includes a pattern generator and a 64:4 MUX with optional Gray encoding followed by three equally weighted driver slices. Thermometer-encoded quarter-rate data (TA/TB/TC) is routed to the 3 driver slices in PAM-4 mode, while all 3 slices receive the same data in NRZ mode. Driver slices incorporate a tap delay generator to retime the incoming data and provide 1-UI-staggered quarter-rate data (D0-D3), which are fed to 12 identical segments. Key circuits of the segment are the quarter-rate (C4) clock phase selector, 1-UI-wide pulse generator, 4:1 MUX/pre-driver, and tailless CML driver (only half circuits are shown for simplicity). Within each segment, the 4-UI-wide data bits (D0-D3) are carved into 1-UI-wide pulses (D0'-D3') by gating (AND) them with appropriate phases of the quarter-rate clock; these 1-UI-wide pulses are then converted to full-rate data by the 4:1 MUX/pre-driver. By changing the timing of where the 1-UI-wide data is carved (by means of clock phase selection [4]), segments are reassigned to different taps to modify the FFE coefficients. While in principle there are four possible positions for the 1-UI-wide pulse carving (which would make a 4-tap FFE feasible without extra shift registers), in this 3-tap FFE implementation, restricting usage to only three of them allows timing constraints to be satisfied without complex delay compensation circuitry. All segments remain active with different tap configurations, maintaining constant peak power and efficient use of transistor sizing without excess parasitics. In addition, the current of each segment can be adjusted with the V_{SEG} gate voltage in the driver, which is generated by an op-amp-based replica bias circuit. Due to the increased headroom, the tailless CML approach for driver [5] and pre-driver [6] enables smaller FET sizing resulting in reduced parasitic capacitance and enhanced speed. In the 4:1 MUX/pre-driver, the PFET gate voltage V_{CTRL} is biased to maintain optimum performance over PVT. Current saturation of the PFET with large V_{DS} reduces the static power dissipation compared to a pullup resistor for a given settling time. Choosing 12 segments per slice permits efficient reallocation of active driver width without excessive wiring overhead while avoiding minimum device size limitations. The pre-driver and driver are sized to generate 20mA of drive current with the lowest parasitics while obeying EM current limitations. To minimize bandwidth and return loss degradation when an ESD device (1kV HBM, 250V CDM, 30V MM) is added, an optimized T-coil is inserted between driver and pad.

A half-rate (C2) external clock is divided by a CML I/Q divider to produce C4 clocks. After conversion to CMOS logic levels, the $C4_I$ clock is divided to generate C8, C16 and C32 clocks. For a high-speed quarter-rate TX, timing accuracy of the quarter-rate clocks is critical. Quadrature error sensing (QES) and duty cycle error sensing (DCS) are based on measurement of average voltages generated by an XOR-based mixer and LPF for QES and only an LPF for DCS. Positive or negative current injection is used to perform duty-cycle correction (DCC) in both quarter-rate clock paths $C4_I$ and $C4_Q$. Sub-100fs DCC resolution is achieved using 6b IDACs. Quadrature-error correction (QEC) is done by adding a differential voltage offset to the half-rate input clock.

Figure 6.6.2 shows the timing diagrams for FFE tap generation. The top timing diagram shows the carving of the main cursor when the FFE is configured for 1 pre- and 1 post-cursor. As an example, 1-UI carving for D0' is initiated by $C4_{IB}$ and terminated by $C4_Q$. A segment is reconfigured as a pre- or post-cursor by changing the C4 clock phase selection so that the carving is shifted left or right (dashed arrows). The bottom timing diagram shows main cursor carving for FFE configuration with 2 pre-cursors. Moving the main cursor carving from the 3rd to the 4th UI of the quarter-rate data D0-D3 by C4 clock phase selection improves timing margins for generating the 2nd pre-cursor. As an example, 1-UI carving for D0' is now initiated by $C4_{QB}$ and terminated by $C4_{IB}$. As shown in Fig. 6.6.3, the C4 clock phase selector is implemented as a 4:4 MUX with transmission-gate-based MUXes and tri-state inverters for minimal power and area cost. Any of the 4 incoming C4 phases can be assigned to a given output ($C4_{IS/ISB}$, $C4_{QS/QSB}$) with $SEL_{<3:0>}$ and EN_{FFE}.

The TX is built in 14nm CMOS FinFET technology, and testing was performed by wafer probing. Two test chip variants were created, with and without ESD and T-coil, to observe the effects on driver performance. Figure 6.6.4 shows the measured PRBS31 NRZ/PAM-4 eyes at 64Gbaud/s with 1 pre- and 1 post-cursor FFE setting for both test chips. At a nominal supply of 0.95/1.2V V_{DD}/V_{DDRV}, the TX core including 64:4 serializer, clock generation and three slices consumes 170mW (1.3pJ/b) with 128Gb/s PAM-4 signaling. Peak-to-peak differential launch amplitude for all these measurements is 1V. The relative level mismatch (RLM) is >98.6% after optimizing segment assignments to compensate for random mismatch. The total losses at half-baud frequency are 6.0dB without ESD and 7.1dB with ESD. (Fig. 6.6.5 shows the impulse responses and insertion losses (S_{21}) derived from measured step responses for both test sites). The loss difference with and without ESD is only 1.1dB at half-baud frequency but noticeably higher at most other frequencies, so the difference in ISI is significant. Approximately 5dB of this loss is due to connectors, cables, and the 60GHz sampling scope head. 128Gb/s performance is significantly degraded by the ESD despite the use of a compensation T-coil but could be improved by adopting a lower capacitance ESD structure at a reduced level of protection. After equalization, the total peak-to-peak jitter is less than 4ps for 64Gb/s NRZ eyes. The test chip was also measured with 112Gb/s PAM-4 signaling at a $0.6V_{ppd}$ launch amplitude and reduced supply of 0.85/0.9V V_{DD}/V_{DDRV} as shown in Fig. 6.6.5 (bottom portion). Under these conditions, the power efficiency is 1.0pJ/b and the measured RLM is 99%. Figure 6.6.6 summarizes the TX performance. Compared with recently reported PAM-4 transmitters operating above 64Gb/s, this work achieves the highest data rate as well as the best energy efficiency. Figure 6.6.7 shows the micrograph of the test chip with ESD. The active area of this version is 165μm×294μm.

Acknowledgments:
This project has been supported and partially funded by the Lawrence Livermore National Laboratory on behalf of the U.S. Department of Energy, under Lawrence Livermore National Laboratory subcontract no. B621073. The authors acknowledge support from GlobalFoundries for chip fabrication.

References:
[1] J. Kim, et al., "A 112Gb/s PAM-4 Transmitter with 3-Tap FFE in 10nm CMOS," *ISSCC*, pp. 102-103, Feb. 2018.
[2] K. Tan, et al., "A 112-Gb/s PAM4 Transmitter in 16nm FinFET," *IEEE Symp. VLSI Circuits*, pp. 45-46, June 2018.
[3] C. Menolfi, et al., "A 112Gb/s 2.6pJ/b 8-Tap FFE PAM-4 SST TX in 14nm CMOS," *ISSCC*, pp. 104-105, Feb. 2018.
[4] T. Dickson, et al., "A 1.8pJ/b 56Gb/s PAM-4 transmitter with fractionally spaced FFE in 14nm CMOS," *ISSCC*, pp. 118-119, Feb. 2017.
[5] G. Steffan, et al., "A 64Gb/s PAM-4 transmitter with 4-tap FFE and 2.26pJ/b energy efficiency in 28nm CMOS FDSOI," *ISSCC*, pp. 116-117, Feb. 2017.
[6] Y. Frans, et al., "A 40-to-64Gb/s NRZ transmitter with supply-regulated front-end in 16nm FinFET," *ISSCC*, pp. 68-69, Feb. 2016.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

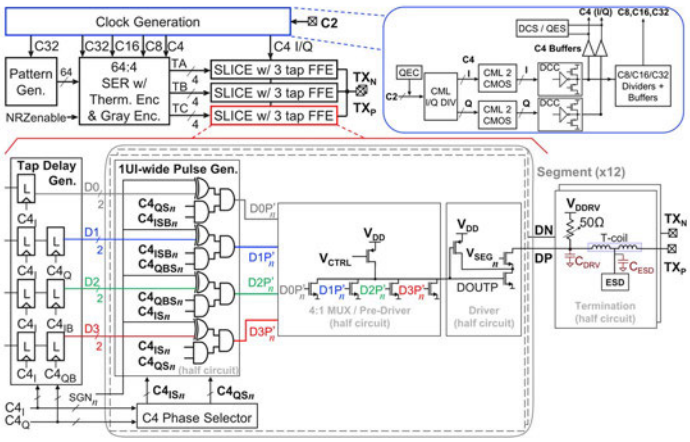


Figure 6.6.1: Block diagram of 64GBaud NRZ / PAM-4 TX.

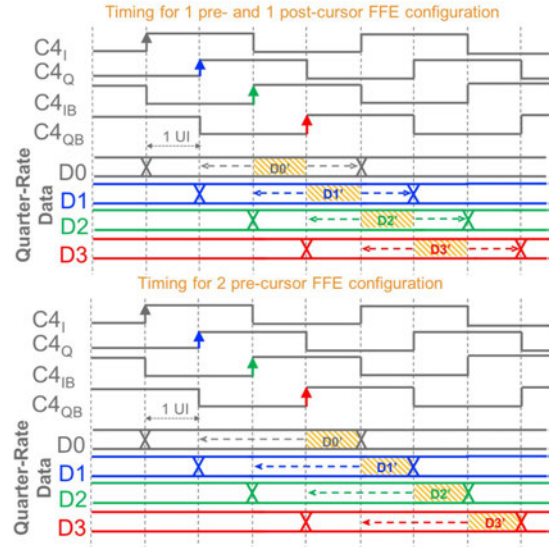


Figure 6.6.2: Timing diagrams for FFE tap generation.

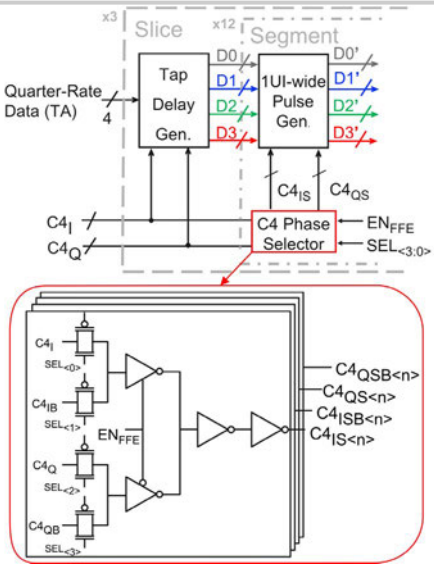


Figure 6.6.3: C4 phase selector circuit.

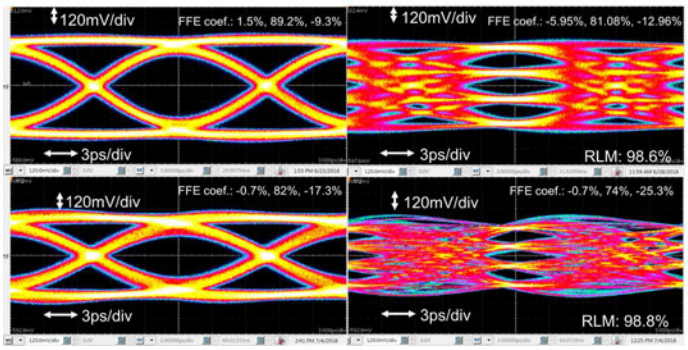


Figure 6.6.4: Measured PRBS31 eye diagrams w/ FFE at 64Gbaud/s. Upper row: no ESD, lower row: with ESD test sites. Peak-to-peak differential launch amplitude for all these measurements is 1V.

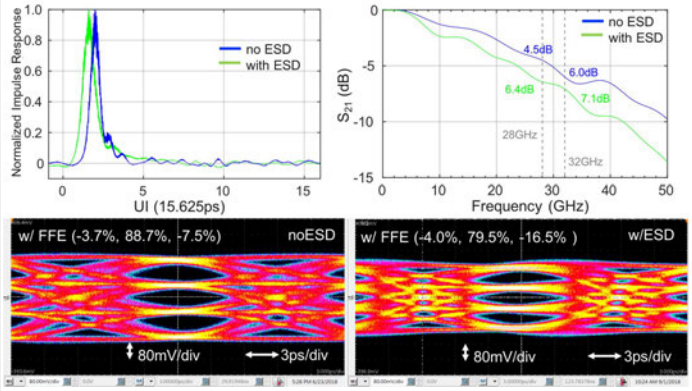


Figure 6.6.5: Top: impulse responses and insertion losses (S21) derived from measured step responses. Bottom: measured 112Gb/s PAM-4 PRBS31 data eyes at 600mV$_{ppd}$ launch amplitude with 1.0pJ/b power efficiency (Left: no ESD, right: with ESD test sites).

	[1]		[2]	[3]	[5]	[6]	This Work		
Technology	10nm FinFET		16nm FinFET	14nm FinFET	28nm FDSOI	16nm FinFET	14nm FinFET		
Architecture	Quarter-rate		Quarter-rate	Quarter-rate	Quarter-rate	Quarter-rate	Quarter-rate		
Modulation	NRZ	PAM-4	PAM-4	PAM-4	PAM-4	NRZ	NRZ	PAM-4	PAM-4
Data Rate [Gb/s]	56	112	112	112	64	64	64	128	112
Clock Source	On-chip PLL		On-chip PLL	External	On-chip PLL	On-chip PLL	External		
FFE	3-tap		4-tap	8-tap	4-tap	3-tap	3-tap		
Driver Type	CML		SST	SST	CML*	CML	CML*		
Output Swing [V_{ppd}]	0.75		1	0.92	1.2	0.8	1	1	0.6**
Supply [V]	1.0		0.9 / 1.2 / 1.8	0.95	1.0 / 1.2	0.8 / 1.1	0.95 / 1.2		0.85 / 0.9
RLM (%)	NA	98.5	96.3	NA	94	NA	NA	98.8	99.1
TX (only) Area [mm^2]	0.03		0.38	0.095	NA	0.32	0.048		
Power Efficiency [pJ/b] (excluding PLL)	3.44	1.72	3.08	2.6	2.26	3.51	2.7	1.3	1.0

*Tailless CML driver.
**Low power setting.

Figure 6.6.6: TX performance comparison.

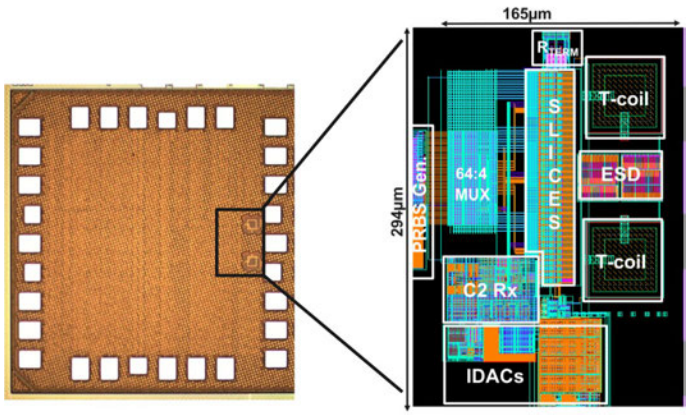


Figure 6.6.7: Die micrograph and layout details for the test site with ESD.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

6.7 A 112Gb/s PAM-4 Voltage-Mode Transmitter with 4-Tap Two-Step FFE and Automatic Phase Alignment Techniques in 40nm CMOS

Pen-Jui Peng[1], Yan-Ting Chen[1], Sheng-Tsung Lai[1], Chao-Hsuan Chen[1], Hsiang-En Huang[1], Ted Shih[2]

[1]Yuan Ze University, Taoyuan City, Taiwan
[2]Teletrx, Taipei, Taiwan

The continuous development of wireline communication encourages transmitters to operate at higher speeds. The applications of 400GbE also push the transmitter to be designed at 112Gb/s for a single lane [1-2]. However, the use of advanced processes (<16nm) hardly reduces the costs. This paper presents a 112Gb/s PAM-4 voltage-mode transmitter fabricated in 40nm CMOS by using the proposed two-step FFE and the automatic phase alignment techniques, improving the output bandwidth as well as the power dissipation. It delivers high-quality eye diagrams under 5.5dB loss at 28GHz with 3.89pJ/b efficiency.

The PAM-4 transmitters with voltage-mode drivers have been used in recent years due to their low power and high linearity characteristics [2-5]. Compared to NRZ signaling, the PAM-4 signals are more sensitive to ISI due to the reduction of the eye openings, and accordingly require precise coefficients of the FFE. However, increasing the resolution of the FFE in a voltage-mode transmitter results in a larger numbers of driver cells and a higher complexity of routing as well, degrading the output bandwidth significantly. The increased routing load for the clock distribution also raises the power dissipation. To improve the resolution of the FFE without degrading the output bandwidth and power efficiency, a two-step FFE consisting of coarse tuning and fine tuning is adopted.

Figure 6.7.1 illustrates the architecture of the proposed 112Gb/s PAM-4 TX. The built-in PRBS generator delivers 128 parallel data through a 128:16 MUX. The data is then split into two 8b branches (MSB and LSB) and fed into FIR generators to create the 4-tap FFE. After serializing by the 8:4 MUXs, coarse tuning for FFE is achieved from the coefficient MUXs. The quarter-rate 4:1 MUXs are directly triggered by the 1-UI pulse generator to improve the output bandwidth. The pre-drivers utilize resistive feedback in the first stage, further reducing the time constant at the MUX output. To obviate the retiming latch, the automatic phase alignment technique is applied on the preceding 8:4 MUXs. The fine-tuning FFE is generated by adjusting the output impedance of each driver cell individually from the reference generators, which are first calibrated with a global resistor calibration loop. An on-chip 14GHz PLL is included to create the differential clocks and then fed into a poly-phase filter (PPF) to generate differential quadrature clocks. Finally, a calibration loop for duty-cycle and phase mismatch between the differential quadrature clocks is adopted to minimize the horizontal eye distortions.

For a voltage-mode driver, the weighting for each input is controlled by the output resistance in each path. As a result, different weightings can be achieved in a single unit cell of the driver by properly adjusting the output resistance. Figure 6.7.2 shows such an arrangement. Here, the driver unit cell incorporates an SST structure with tunable resistors M_N and M_P to maintain output impedance matching and adjust the FFE coefficients simultaneously. A global resistor calibration loop fixes the variable resistor $R_{P,cal}$ and $R_{N,cal}$ with the external precise resistors at power on. The reference generators then adopt four impedance control loops with the replica drivers and variable resistors, creating V_{ref4} for coarse FFE and V_{ref1}~V_{ref3} for fine FFE settings. Each driver cell selects a proper reference voltage by a 4:1 selector, realizing a two-step FFE accordingly. Finally, series inductors L_S further improve the bandwidth in the presence of the driver output capacitance and the pad capacitance.

To relax the timing constraint for the last stage MUX and reduce the power consumption in the clock path, quarter-rate clocking architectures have been adopted [1-4]. However, the retiming latches are required to relax the timing margin for the quarter-rate 4:1 MUX, increasing the power consumption in both the data and clock paths. Although the number of retiming latches can be reduced by using the 8-phase clocks to trigger the preceding 8:4 MUX [6], the required 8-phase frequency divider increases the power consumption to some extent. Additionally, triggered by 8-phase clocks, the preceding retiming latches of the 8:4 MUX make the timing constraint stringent. In this paper, all of the retiming latches for the quarter-rate 4:1 MUX are removed by adopting an automatic phase alignment technique, improving the power efficiency and robustness. Figure 6.7.3(a) illustrates the MUX architecture which consists of an 8:4 MUX triggered by phase interpolators, a phase aligner, a 1-UI pulse generator creating 25% duty cycle clocks CK_1~CK_4, and a 4:1 MUX directly triggered by CK_1~CK_4. To obviate the retiming latch, the preceding 8:4 MUX is triggered by two phase interpolators PI_1 and PI_2, which are adjusted by the phase aligner to ensure good timing margin for the 4:1 MUX. The phase aligner fixes the edge of D_1 with the rising edge of CK_4 by using a DFF-based phase detector (PD) with a finite state machine (FSM). The PI_1 and PI_2 create the orthogonal clocks $CK_{0,7G}$ and $CK_{90,7G}$, respectively, increasing the timing margin from the preceding FIR generator. The related timing diagram is shown in Fig. 6.7.3(b). By aligning the data transition of D_1 with the rising edge of CK_4, each data can be picked up by the corresponding pulses with 1-UI margin. As a result, the 4:1 MUX has enough time for settling to maintain the output eye quality. Figure 6.7.3(c) illustrates the 4:1 MUX, consisting of four tri-state buffers controlled by NAND and NOR gates and sampling data by using 25% duty cycle clocks. This arrangement obviates the use of cascading or stacking devices, improving the output bandwidth significantly.

The TX is fabricated in 40nm CMOS. It consumes 436mW of power at 112Gb/s from 1V and 1.2V supplies. The output channel under the testing setup (wirebond + 20mm PCB trace + end launch connector + dc block + 4'' cable) implies 5.5dB loss at 28GHz. By applying the two-step FFE technique, the output loss is almost compensated with precise FFE settings. Figure 6.7.4(a) and (b) demonstrate the output NRZ waveform with 2^7–1 PRBS at 50Gb/s and 56Gb/s, respectively. The measured eye heights at 50Gb/s and 56Gb/s are 275mV and 265mV, respectively. With the help of duty-cycle and IQ calibration, the horizontal eye distortions are reduced significantly, creating good symmetry for each eye. Figure 6.7.4(c) and (d) illustrate the output PAM-4 waveform at 100Gb/s and 112Gb/s, respectively. By properly adjusting the FFE coefficients, high-quality eye diagrams are achieved. The RLMs at 100Gb/s and 112Gb/s are 97.2% and 97.7%, respectively.

Figure 6.7.5(a) illustrates the locking behavior of the phase aligner, which is measured by connecting the PI codes to a 7b DAC. The calibration is done within 300ns. Figure 6.7.5(b) shows the timing margin under locking by adding an offset code to the PI and measuring the peak-to-peak jitter of the 56Gb/s NRZ output. It reveals more than ±0.5UI margin to maintain output peak-to-peak jitter less than 5ps. The 14GHz PLL phase noise is plotted in Fig. 6.7.5(c), which indicates –109dBc/Hz at 1MHz offset. The integrated rms jitter (from 100Hz to 500MHz) is 208fs. Figure 6.7.5(d) illustrates the normalized amplitude of the post-cursor under different FFE coefficient settings measured from the output pulse response. With the proposed two-step FFE technique, the resolution of the FFE coefficient can be improved without increasing the number of unit cells.

Performance of this work and other state-of-the-art PAM-4 transmitters is summarized in Fig. 6.7.6. With the proposed techniques, this paper achieves a silicon-proven PAM-4 TX delivering >100Gb/s data rate in 40nm CMOS. Figure 6.7.7 depicts the die photograph of the proposed TX, which occupies 0.8×0.7mm^2 active area. The detailed power breakdown is illustrated in Fig. 6.7.7.

Acknowledgements:
The authors would like to acknowledge chip fabrication support provided by National Chip Implementation Center (CIC), Taiwan, R. O. C. This work was supported in part from the Young Scholar Fellowship Program by Ministry of Science and Technology (MOST) in Taiwan, under Grant MOST107-2636-E-155-001.

References:
[1] J. Kim, et al., "A 112Gb/s PAM-4 Transmitter with 3-Tap FFE in 10nm CMOS," *ISSCC*, pp. 102–103, Feb. 2018.
[2] C. Menolfi, et al., "A 112Gb/s 2.6pJ/b 8-Tap FFE PAM-4 SST TX in 14nm CMOS," *ISSCC*, pp. 104–105, Feb. 2018.
[3] E. Depaoli et al., "A 4.9pJ/b 16-to-64Gb/s PAM-4 VSR Transceiver in 28nm FDSOI CMOS," *ISSCC*, pp. 112–113, Feb. 2018.
[4] M. Bassi, et al., "A 45Gb/s PAM-4 Transmitter Delivering 1.3Vppd Output Swing with 1V Supply in 28nm CMOS FDSOI," *ISSCC* pp. 66–67, Feb. 2016.
[5] T. O. Dickson, et al., "A 1.8pJ/b 56Gb/s PAM-4 Transmitter with Fractionally Spaced FFE in 14nm CMOS," *ISSCC*, pp. 118–119, Feb. 2017.
[6] A. A. Hafez, et al., "A 32-to-48Gb/s Serializing Transmitter Using Multiphase Sampling in 65nm CMOS," *ISSCC*, pp. 38–39, Feb. 2013.
[7] G. Steffan, et al., "A 64Gb/s PAM-4 Transmitter with 4-Tap FFE and 2.26pJ/b Energy Efficiency in 28nm CMOS FDSOI," *ISSCC*, pp. 116–117, Feb. 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

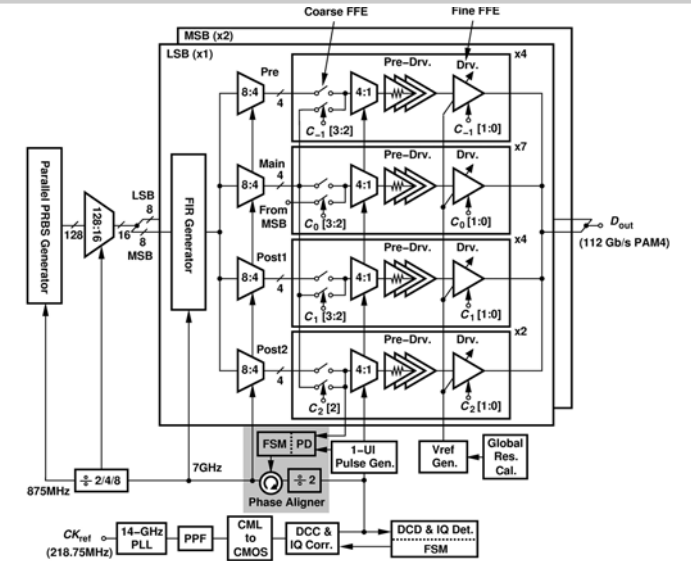


Figure 6.7.1: Block diagram of the TX.

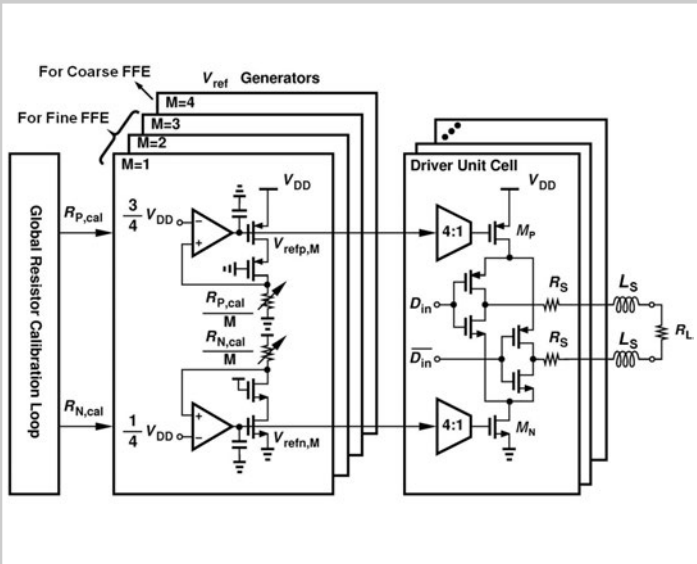


Figure 6.7.2: SST driver with reference generators for coarse and fine FFEs.

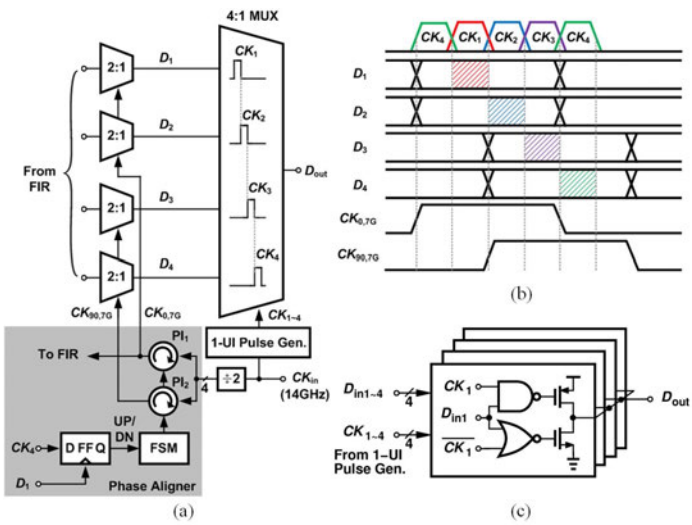


Figure 6.7.3: (a) MUX architecture with automatic phase alignment, (b) timing diagram, (c) schematic of the 4:1 MUX.

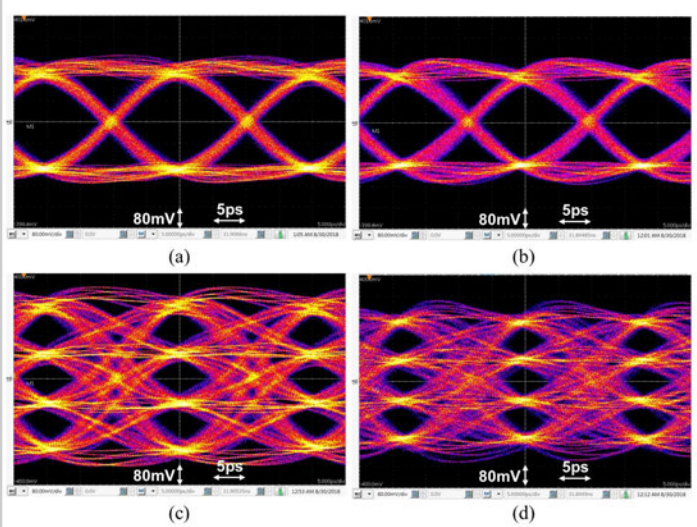


Figure 6.7.4: Measured NRZ output eye diagrams: (a) 50Gb/s, (b) 56Gb/s. Measured PAM-4 output eye diagrams: (c) 100Gb/s, (d) 112 Gb/s.

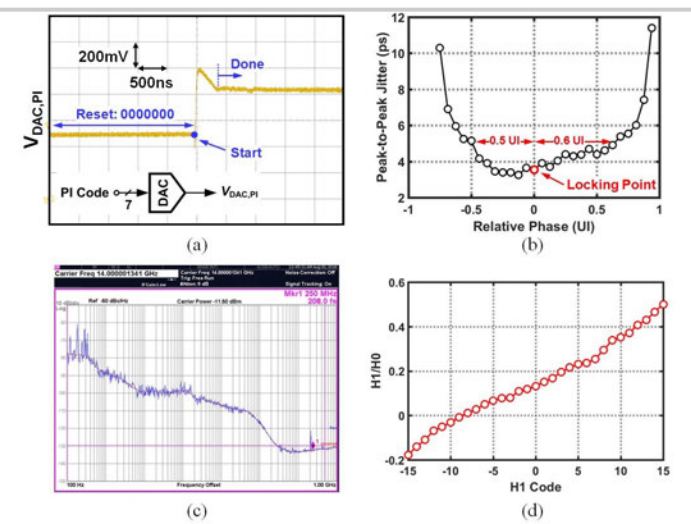


Figure 6.7.5: (a) Locking behavior of phase aligner, (b) output jitter across different phase code, (c) PLL phase noise plot at 14GHz, (d) normalized magnitude of post-cursor.

	[1]	[2]	[5]	[7]	This Work
Architecture	Quarter-rate	Quarter-rate	Half-rate	Quarter-rate	Quarter-rate
Data-rate (Gb/s)	112	112	56	64	112
Clock Source	On-Chip PLL	External	External	On-Chip PLL	On-Chip PLL
Driver Topology	CML	SST	SST	CML	SST
Output Swing w/o FFE	0.75 V_{ppd}	0.92 V_{ppd}	0.9 V_{ppd}	1.2 V_{ppd}	1 V_{ppd}
TX FFE	3-tap	8-tap	3-tap	4-tap	4-tap (Two-Step)
Serialization Ratio	32:1	32:1	16:1	40:1	64:1
RLM (%)	98.5	N/A	N/A	94	97.7
Efficiency (pJ/bit)	2.07	2.6	1.73	2.26	3.89
Area (mm^2)	0.0302	0.095	0.035	N/A	0.56
Technology	10nm FinFET	14nm FinFET	14nm FinFET	28nm FDSOI	40nm CMOS

Figure 6.7.6: Performance summary.

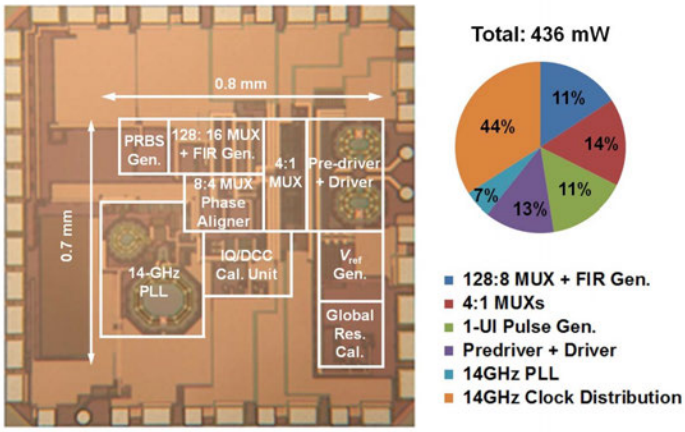


Figure 6.7.7: Die photo and power breakdown.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

6.8 A 36Gb/s Adaptive Baud-Rate CDR with CTLE and 1-Tap DFE in 28nm CMOS

Danny Yoo[1], Mohammad Bagherbeik[1], Wahid Rahman[1], Ali Sheikholeslami[1], Hirotaka Tamura[2], Takayuki Shibasaki[2]

[1]University of Toronto, Toronto, Canada
[2]Fujitsu Laboratories, Kawasaki, Japan

Baud-rate clock-and-data recovery circuits (CDR) are ubiquitous in recent receiver designs as a means of lowering power consumption by sampling the data only once per UI. To further reduce power, prior works in pattern-based baud-rate PD [1] and FD [2] combine clock & data recovery by sharing comparators between the DFE and the PD. However, both CDRs [1, 2] were manually tuned by sweeping the equalization settings of the CTLE and the comparator levels in order to achieve lock. Since the two settings are correlated, it is time-consuming and tedious to sweep the entire solution space. In this paper, we propose an adaptive engine where the CDR system searches and converges to an optimal lock and sampling point. The proposed scheme, implemented in 28nm CMOS and operating at 36Gb/s, relies on the data eye to autonomously adapt the parameters of a CTLE, a 1-tap DFE, and the PD locking point.

Figure 6.8.1 shows a block diagram of a full-rate adaptive receiver. The receiver consists of a CTLE, a 1-tap speculative DFE, which shares its front-end comparators with a baud-rate PD [1], the CDR, and the proposed adaptive engine highlighted at the bottom. The adaptive engine receives the recovered clock (CK_{rec}) from the CDR and rotates it to a new phase (CK_X) in order to adaptively sample the CTLE output at the intersection of the 011 and 110 patterns. The rotation is performed by a phase interpolator (PI) and guided by the PI logic in the quarter-rate system implementation. The heart of the adaptive engine is a *data level loop*, which is a feedback loop similar to the one proposed in [3], that observes the CTLE samples and reconstructs the data level (dLev) with 9b resolution. To do so, the adaptive sampler subtracts the stored dLev from the current CTLE sample, quantizes the difference to 1b, and feeds the resulting bits to a digital filter prior to summing them up in an accumulator. The DAC produces an analog level corresponding to the 9b dLev and feeds it to the sampler as its threshold. The role of the digital filter and the pattern filter is to provide programmability to calculate the average, the maximum, and the minimum of the CTLE output at phase CK_X. A cycle counter (or filter scheduler) schedules 16k clock cycles to execute each of these tasks using the same data level loop. The calculated values of dLev for various filters are then used by the adaptive logic block to guide the CTLE parameter (C) and the 1-tap DFE coefficient (α) that also determines the sampling phase of the PD.

Figure 6.8.2 illustrates the operating principle of the proposed adaptive engine. For a channel with a pulse response of $1 + \alpha D^1 + \alpha_2 D^2 + \alpha_3 D^3 + \cdots$, we co-optimize C and α so as to minimize the residual ISI ($\Sigma_{i=2}^{\infty} |\alpha_i|$) and cancel the first ISI term (α) while maximizing the timing margin at the sampling phase. As shown in Fig. 6.8.2, the residual ISI manifests itself in the crossing thickness (ΔX) of the eye diagram at the intersection of 011 and 110 patterns (CK_X). Therefore, we observe and minimize it over a range of C starting from its highest setting (15) and lowering it towards its lowest setting (0), one step at a time, until we reach ΔX_{min} (when we observe an increase in ΔX, we return to the previous setting which corresponds to ΔX_{min}). To cancel the first ISI term, we observe the average of the CTLE output at CK_X (labeled as X_{avg}) and assign $X_{avg}/2$ as the DFE coefficient α. The DFE coefficient obtained in this way will also determine the sampling phase, which corresponds to the maximum timing margin available for the PD. This is indeed accurate if the CTLE eye has no residual ISI ($\Sigma_{i=2}^{\infty} |\alpha_i|$), and thus all ISIs are equalized for the 0011 repeating pattern. In this case, the rise time of the 011 data sequence equals the fall time of the 110 data sequence, and the phase difference between the zero-crossing of the 0101 pattern and CK_X corresponds to 0.5UI. Highlighted by a thicker black line in the eye diagram of Fig. 6.8.2, the rising data trace between the zero-crossing of the 0101 pattern and CK_X can be approximated as a straight line. Hence, the data point corresponding to $\alpha = X_{avg}/2$ will be at the center of the highlighted horizontal eye opening, with a timing margin of 0.5UI on each side. In other words, CK_{rec} will sample the data at an optimal phase for a 1-tap speculative DFE. To avoid losing lock during the adaptation process, for every value of CTLE parameter C, α is always adjusted to $X_{avg}/2$ such that the PD would lock to the optimal sampling phase throughout the process.

The proposed 36Gb/s adaptive baud-rate CDR is fabricated in 28nm CMOS technology with a 0.9V supply. A detailed block diagram with quarter-rate clocking is shown in Fig. 6.8.3. The analog front-end consists of a CTLE with two identical high-frequency boost stages controlled by the same parameter (C), 8 comparators, which are shared between the DFE and PD implemented in a quarter-rate clocking scheme, and one additional comparator, which is used as a sampler for digital adaptation. The PD operates on $CK_{rec}/2$ while the 1-tap speculative DFE operates on $CK_{rec}/4$, and the digital back-end operates on $CK_{rec}/8$. The CDR loop includes a conventional analog charge pump (CP), loop filter (LF), and a ring VCO. The VCO and the quadrature clock network run on a regulated supply provided by a low-drop out (LDO) regulator.

The testchip is wire-bonded using a QFN package and tested with data transmitted through a Tyco 5" channel. To confirm that the CDR locks with adaptation, we use 35Gb/s PRBS31 and PRBS7 input patterns and observe the spectrums of the recovered clock. Figure 6.8.4 illustrates measured spectrums of the 8.75GHz quarter-rate recovered clocks for both patterns. Also shown in the same figure is the measured jitter tolerance (BER < 10^{-12}) with adaptation turned ON, plotted against IEEE 802.3 masks. The three measured jitter tolerance curves were obtained by injecting sinusoidal jitter to the transmit PRBS pattern generator (SHF 12104A). These curves verify the operation of the receiver under two different channel losses at two different data rates (16.73dB at 35Gb/s and 18.25dB at 36Gb/s, both at their respective Nyquist frequencies) and two different data patterns (PRBS31 and PRBS7).

To confirm that the adaptation engine locks to the optimal settings for the CTLE and the DFE/PD, we use 35Gb/s PRBS31 input data in combination with the Tyco 5" channel. Figure 6.8.5 shows that the adaption engine converges to a CTLE parameter C = 8 and a DFE/PD parameter α = 133mV. A manual sweep of C while keeping α at its adapted value (133mV), and a manual sweep of α while keeping C at its adapted value (8), both degrade the timing margin as we move away from the optimal setting. This degradation is more drastic when we change C because the control logic for C is more coarse in our design (4b resolution compared to α 9b resolution). The timing margin in all cases is defined as the minimum high-frequency jitter tolerance over jitter frequencies in the 10-to-100MHz range.

The total power consumption is measured to be 106.3mW or 3.04pJ/b at 35Gb/s. This total power does not include the digital adaptation, which consumes 6.3mW but is turned off after convergence, and the IO power (IO pad drivers and buffers), which consumes 2.7mW but is used purely for debugging purposes. To ensure that the inductor-less CDR could phase lock with a ring-VCO, the VCO and the clocking network were over-designed in terms of power to lower the phase noise (theoretically, every two-fold increase in VCO power yields 3dB gain in the phase noise). The VCO, LDO and the clock buffers consume 71.2mW of the total 106.3mW, or 67%. Therefore, the reported FOM of 3.04pJ/b could be further improved for this baud-rate scheme. Finally, Fig. 6.8.6 compares this work to recently published prior works for baud-rate CDRs. This work presents a live on-chip adaptation for a baud-rate CDR that combines clock & data recovery by sharing comparators between the DFE and the PD. In addition, this adaptive scheme does not require any specific training sequence.

Figure 6.8.7 shows the die micrograph with the total die area of 1.232mm^2 and its power breakdown per supply domain.

Acknowledgements:
The authors would like to thank CMC Microsystems for providing CAD tools and measurement equipment, MOSIS for technology access, and Hossein Shakiba for careful review and feedback on the manuscript.

References:
[1] T. Shibasaki, et al., "A 56Gb/s NRZ-Electrical 247mW/lane Serial-Link Transceiver in 28nm CMOS," *ISSCC*, pp. 64-65, Feb. 2016.
[2] W. Rahman, et al., "A 22.5-to-32Gb/s 3.2pJ/b referenceless baud-rate digital CDR with DFE and CTLE in 28nm CMOS," *ISSCC*, pp. 120-121, Feb. 2017.
[3] V. Stojanovic, et al., "Adaptive equalization and data recovery in a dual-mode (PAM2/4) serial link transceiver," *IEEE Symp. VLSI Circuits*, pp. 348-351, June 2004.
[4] P. A. Francese, et al., "A 16 Gb/s 3.7 mW/Gb/s 8-Tap DFE Receiver and Baud-Rate CDR With 31 kppm Tracking Bandwidth," *IEEE JSSC*, vol. 49, no. 11, pp. 2490-2502, Nov. 2014.
[5] R. Dokania, et al., "A 5.9pJ/b 10Gb/s serial link with unequalized MM-CDR in 14nm tri-gate CMOS," *ISSCC,* pp. 184-185, Feb. 2015.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

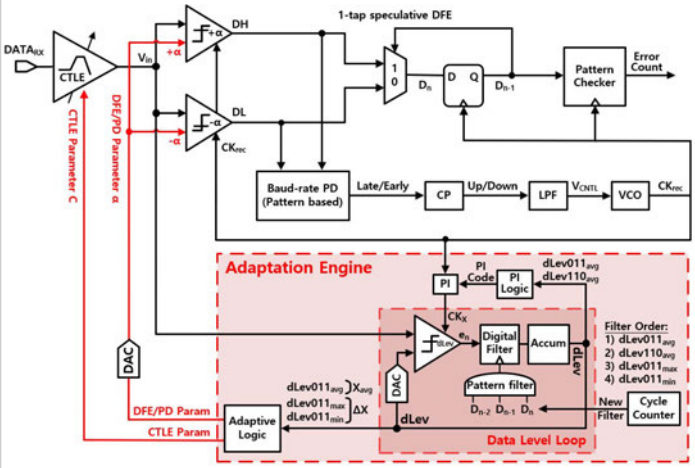


Figure 6.8.1: Proposed architecture for the adaptive baud-rate CDR, shown in a full-rate system.

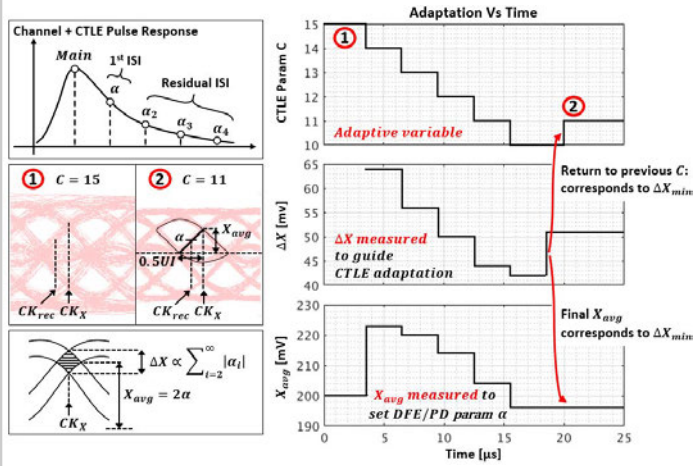


Figure 6.8.2: Adaptation vs. time where CTLE parameter C is adapted to achieve minimum ΔX. For every C, DFE/PD parameter α is determined by observing X_{avg} such that the CDR is adapted in small tick-tock increments to maintain phase lock.

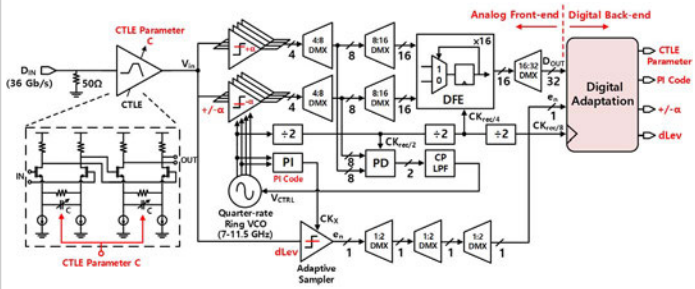


Figure 6.8.3: Schematic of the proposed baud-rate receiver with quarter-rate clocking.

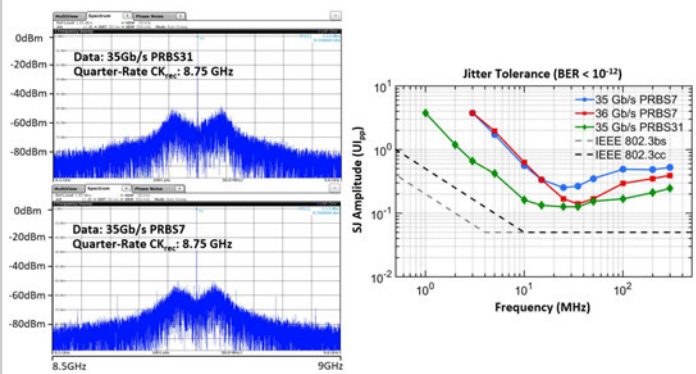


Figure 6.8.4: Measured spectrum of recovered clock and measured jitter tolerance of the CDR locked with adaptation turned ON for PRBS7 (at 35Gb/s and 36Gb/s) and PRBS31 (at 35Gb/s).

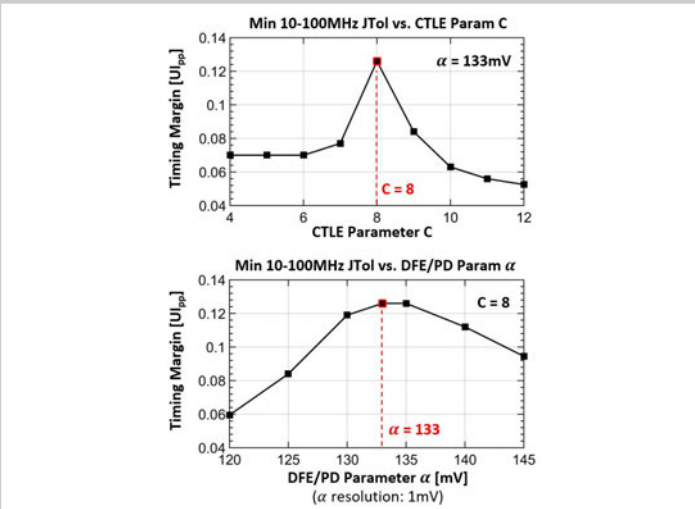
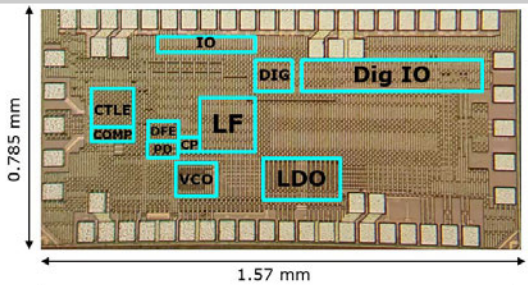


Figure 6.8.5: Measured minimum jitter tolerance in the 10-to-100MHz jitter frequency range for the adapted and adjacent settings of CTLE param (C) and DFE/PD param (α). Adapted settings are highlighted in red.

	(JSSC 14) [4]	(ISSCC 15) [5]	(ISSCC 16) [1]	(ISSCC 17) [2]	This Work
Process	22nm (SOI)	14nm (CMOS)	28nm (CMOS)	28nm (CMOS)	28nm (CMOS)
Supply [V]	0.85, 1.0, 1.15	0.9	0.96	0.9	0.9
Inductor-less	No	Yes	No	Yes	Yes
Data-rate [Gb/s]	16	10	56.2	22.5-32	36
Baud-rate	Yes	Yes	Yes	Yes	Yes
Signalling	NRZ	NRZ	NRZ	NRZ	NRZ
Adaptation	Yes (Offline)	Yes (Live)	Yes (Offline)	No	Yes (Live)
Channel Loss [dB]	34	24	18.4	14.8	18.25
Equalization	CTLE 8-tap DFE	3-tap TX FFE CTLE 4-tap DFE	CTLE 1-tap DFE	CTLE 1-tap DFE	CTLE 1-tap DFE
Total Power [mW]	59.7	59†	142.2	102.0	106.3
FOM [pJ/bit]	3.7	5.9†	2.53	3.19	3.04

† RX + TX power

Figure 6.8.6: Performance comparison table.



Domain	Power[mW]	Blocks
VDDA	18	Comparators, DEMUX, PD, DFE, CP, LF, clock divider, clock buffers
VDD_LDO	71.2	VCO, LDO, clock buffers
VDD_CTLE	14.4	CTLE
VDD_DAC	2.7	Comparator Ref DAC
TOTAL	106.3*	
FOM @35 Gb/s	3.04 pJ/bit	

* Does not include adaptation power (6.3mW) as is turned off after convergence
* Does not include IO power (2.7mW) as is used solely for debugging

Figure 6.8.7: Die micrograph and power breakdown.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 7 Overview: *Machine Learning*

DIGITAL ARCHITECTURES AND SYSTEMS SUBCOMMITTEE

Session Chair: ***Dejan Markovic***
University of California, Los Angeles, Los Angeles, CA

Associate Chair: ***Mahesh Mehendale***
Texas Instruments, Dallas, TX

Interim Subcommittee Chair: Thomas Burd, AMD, Santa Clara, CA

Architectures supporting machine learning for embedded perception and cognition are continuing their rapid evolution, inspired by modern data analytics and enabled by the low energy cost of CMOS processing. This makes it feasible to migrate data analytics toward edge and wearable devices.

This session covers trends in machine learning for modern image and speech recognition and video processing for next-generation mobile/edge devices used in automotive, robotics, and other inference applications. The session features programmable accelerators for Convolutional Neural Networks (CNNs), Deep Reinforcement Learning (DRL), Recurrent Neural Networks (RNNs), and neuromorphic algorithms.

8:30 AM

7.1 An 11.5TOPS/W 1024-MAC Butterfly Structure Dual-Core Sparsity-Aware Neural Processing Unit in 8nm Flagship Mobile SoC

J. Song, Samsung Electronics, Hwaseong, Korea

In Paper 7.1, Samsung introduces an energy-efficient neural processing unit for a mobile product. Their dual-core neural processing unit features 1,024 highly utilized MAC units with in-place feature selection and a butterfly-structure architecture. The processor core is integrated in 5.5mm^2 in 8nm CMOS and achieves 6.94TOPS at 0.8V, and 11.5TOPS/W at 0.5V, on full-layer neural networks.

9:00 AM

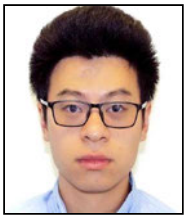
7.2 A 20.5TOPS and 217.3GOPS/mm^2 Multicore SoC with DNN Accelerator and Image Signal Processor Complying with ISO26262 for Automotive Applications

Y. Yamada, Toshiba Electronic Devices & Storage, Kawasaki, Japan

In Paper 7.2, Toshiba presents an image recognition SoC for ADAS applications. The SoC performs ASIL-B-compliant high-performance image recognition and ASIL-D-compliant high-reliability control processes. The 16nm SoC is 94.52 mm^2 and achieves peak performance of 20.5TOPS at 0.8V, and 217.3 GOPS/mm^2 with 10 processors, 4 DSPs, and 8 accelerator variants.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

9:30 AM



7.3 An 879GOPS 243mW 80fps VGA Fully Visual CNN-SLAM Processor for Wide-Range Autonomous Exploration

Z. Li, University of Michigan, Ann Arbor, MI

In Paper 7.3, the University of Michigan describes a low-power real-time CNN-SLAM processor that implements full-visual SLAM on a single chip. SLAM performance is tested on the industrial standard KITTI benchmark that renders a large-scale realistic automobile trajectory over 1km. The processor supports large-scale automotive evaluation with >1000 CNN-keypoints/frame on VGA (640×480) resolution in real-time at 80fps, occupies 10.92mm^2, and consumes 243.6mW from a 0.9V supply in 28nm CMOS.

10:15 AM

7.4 A 2.1TFLOPS/W Mobile Deep RL Accelerator with Transposable PE Array and Experience Compression

C. Kim, KAIST, Daejeon, Korea

In Paper 7.4, KAIST proposes a deep reinforcement learning (DRL) SoC, supporting both CNNs and learning-optimized RNNs. The reusability of weights and inputs allows for reconfiguration of PE mapping with a 49% reduction in memory accesses. Encoding and decoding techniques reduce external data bandwidth by 37% and overall power by 31%. The 65nm 16mm^2 chip achieves a peak 2.16TFLOPS/W at ~0.8V, and 204 GFLOPS at 1.1V with 16b data.

10:45 AM

7.5 A 65nm 0.39-to-140.3TOPS/W 1-to-12b Unified Neural-Network Processor Using Block-Circulant-Enabled Transpose-Domain Acceleration with 8.1× Higher TOPS/mm^2 and 6T HBST-TRAM-Based 2D Data-Reuse Architecture

J. Yue, Tsinghua University, Beijing, China

In Paper 7.5, Tsinghua University unveils a unified block-circulant chip supporting CNN/FC/RNN transpose-domain acceleration. The design features a global-parallel local-serial FFT module and a TRAM-based 2D reuse architecture that shows 16×/12.9× on/off-chip memory access reduction. The 65nm 7.5mm^2 chip achieves a 0.39-to-140.3TOPS/W with 1-to-12b data at 0.54 to 1.15V.

11:15 AM

7.6 A 65nm 236.5nJ/Classification Neuromorphic Processor with 7.5% Energy Overhead On-Chip Learning Using Direct Spike-Only Feedback

J. Park, Seoul National University, Seoul, Korea

In Paper 7.6, Seoul National University describes an on-chip-trainable neuromorphic processor for classification tasks. Training energy of 254.3nJ/image is achieved through a hardware-oriented direct feedback algorithm, out-of-order weight updates, and an update-skipping mechanism. During prediction, the processor achieves 97.83% MNIST recognition accuracy with 236.5 nJ/prediction. The 65nm 10mm^2 processor delivers 3.42TOPS/W at 0.8V.

11:45 AM

7.7 LNPU: A 25.3TFLOPS/W Sparse Deep-Neural-Network Learning Processor with Fine-Grained Mixed Precision of FP8-FP16

J. Lee, KAIST, Daejeon, Korea

In Paper 7.7, KAIST presents a DNN learning processor that incorporates: 1) fine-grained mixed precision (FGMP) data to reduce external memory access and increase throughput while retaining learning accuracy, 2) a sparse DNN learning/inference core for high throughput, 3) an input load balancer (ILB) to improve PE utilization. The fabricated chip occupies 16mm^2 in 65nm CMOS and achieves 25.3TFLOPS/W at 0.78V.

7.1 An 11.5TOPS/W 1024-MAC Butterfly Structure Dual-Core Sparsity-Aware Neural Processing Unit in 8nm Flagship Mobile SoC

Jinook Song[1], Yunkyo Cho[1], Jun-Seok Park[1], Jun-Woo Jang[2], Sehwan Lee[2], Joon-Ho Song[2], Jae-Gon Lee[1], Inyup Kang[1]

[1]Samsung Electronics, Hwaseong, Korea
[2]Samsung Advanced Institute of Technology, Suwon, Korea

Deep learning has been widely applied for image and speech recognition. Response time, connectivity, privacy and security drive applications towards mobile platforms rather than cloud. For mobile systems-on-a-chip (SoCs), energy-efficient neural processing units (NPU) have been studied for performing the convolutional layers (CLs) and fully-connected layers (FCLs) [2-5] in deep neural networks. Moreover, considering that neural networks are getting deeper, the NPU needs to integrate 1K or even more multiply/accumulate (MAC) units. For energy efficiency, compression of neural networks has been studied by pruning neural connections and quantizing weights and features with 8b or even lower fixed-point precision without accuracy loss [1]. A hardware accelerator exploited network sparsity for high utilization of MAC units [3]. However, since it is challenging to predict where pruning is possible, the accelerator needed complex circuitry for selecting an array of features corresponding to an array of non-zero weights. For reducing the power of MAC operations, bit-serial multipliers have been applied [5]. Generally, extremely low- or variable-bit-precision neural networks need to be carefully trained.

Figure 7.1.1 shows the number of operations per element, weight or feature, in Inception-v3. If an element is reused more than another element, it is advantageous for power reasons if the element remains stationary, thereby saving the power for reloading the element. Otherwise, without high reuse, the data should be streaming. As the first layers are mainly CLs, the number of operations per weight is much higher than that per feature. For the last layers, they have small CLs and a FCL, and the number of operations per feature is relatively high. The NPU should alternately operate with stationary weights and streaming features or vice-versa based on the layer type.

In this paper, we present an energy-efficient sparsity-aware neural processing unit with the following features: (1) an energy-efficient butterfly-structure dual-core accelerator having 1,024 MACs, (2) exploitation of three-fold parallelism in computing CLs and FCLs, (3) application of simple feature-map selection and lockstep-free MAC operations based on the sparsity, (4) support for bandwidth-efficient network traversal.

The overall architecture of the NPU in Exynos is described in Fig. 7.1.2. It consists of one NPU controller (NPUC) and two NPU cores. The NPUC has a CPU, a DMA unit, code SRAM and a network controller. The CPU is communicating with the host system of the SoC and defines the network scale for the network controller. The controller automatically configures all modules in the two cores and traverses the network. The DMA unit manages the compressed weights and feature maps in each of the 512KB scratchpads of the cores.

The proposed NPU integrates 1,024 MAC units. For mitigating logic costs and wiring congestion, the butterfly-structure NPU is separated into two cores with different clock signals, because first MAC operations are rapidly computed within a core and last activations are slowly performed with cooperation among two cores. In a core, there are two data-staging units (DSUs) that are buffering features in feature-map holders and storing weights in the memory of dispatchers, as well as decompressing the weight stream. For CLs, a non-zero weight is selected in a dispatcher, and its corresponding 2D array of 16 features is chosen by a feature-map selector. For FCLs, a feature is chosen by the selector, and its corresponding array of 16 weights is selected by the dispatcher. The selected weights and features of two DSUs are dispatched to dual 8b MAC arrays (MAAs). In a dual-MAC having two multipliers and one accumulator, the two weights and two features are multiplied, and their results accumulated. Each of dual MACs has buffers holding 16 partial sums. There are 16 MAAs for a core and 1,024 MACs for two cores. After all MAC operations, MAAs deliver the partial sums to a data-returning unit (DRU), where the partial sums of 8 MAAs are for near DRUs, and the others go through asynchronous bridges for far DRUs. The partial sums of two MAAs are added one-by-one and activated by ReLU or PReLU on the DRU. The DRU writes the array of the output features back to the scratchpad.

Figure 7.1.3 shows how the proposed NPU exploits three-fold parallelism of CLs and FCLs. For CLs, parallelism A comprises dividing the 3D input feature maps into four chunks along channels. Two chunks among them are stored in the 512KB scratchpad. Each of DSUs in a core fetches a chunk in a streaming manner. A weight kernel is also divided along the channel dimension. The weight kernels are stationary until all outputs are calculated. The NPU can exploit 16 weight kernels in parallel on 16 MAAs; this is parallelism B. Moreover, a 4-by-4 patch of output features in a MAA can be computed in parallel through parallelism C. These parallelisms are applied in a similar way for FCLs. An array of input features is divided into four chunks, as in parallelism A, while assigning them to each of the DSUs. The weight matrix is also partitioned into four matrices. The NPU can perform MAC operations with 256 rows of a matrix in one cycle as parallelism B, and 16 output features in an MAA are computed in parallel, as shown in parallelism C. The array of features is stationary until all outputs are calculated, and the weights are streaming, consumed after one MAC operation.

The proposed NPU improves convolution efficiency by skipping weights that are zero, using feature-map selection circuitry shown in Fig. 7.1.4. When the index of a non-zero weight is found in the dispatcher, its corresponding 2D array of features is chosen by in-place selection and circular rotation of column and row directions. The speeds of consuming weights in dispatchers are different for each DSU, because the distribution of non-zero weights is not constant among the chunks (i.e. workload imbalance). The speed difference can slow the convolution because of the required output index synchronization. In order to alleviate this problem, the dual 8b MAC has an auxiliary accumulator. When the two multiplication results are for the same output index, they will be accumulated together. If the indices are different, the multiplication results will be accumulated separately. This removes the lockstep requirement between dispatchers having different distributions of zero-weights.

Given that neural networks are getting deeper, the NPU must carefully use the external memory bandwidth and the scratchpads. A network compiler for the proposed NPU partitions networks into sub-networks and decides the traversal over multiple network layers, as shown in Fig. 7.1.5. For example, feature maps on a sub-network of three layers are determined by the compiler, and the network controller will configure the NPU cores based on the traversal order. The input feature maps of the intermediate layers are forwarded from the output feature maps of the previous layers without producing any additional external memory accesses.

Figure 7.1.6 shows the measurement results for the proposed NPU in Exynos, fabricated in 8nm CMOS technology. The NPU occupies 5.5mm^2 and operates at 0.5-to-0.8V supply voltage, 67-to-933-MHz clock frequency. The tera-operations-per-second (TOPS) of single layers using 5×5, 3×3 and 1×1 kernels are measured with respect to the fraction of zero-weights. The NPU achieves 6.9TOPS and 3.5TOPS with 75% zero-weights for 5×5 and 3×3 kernels, respectively. However, performance for 1×1 kernels is more modest because of the bandwidth limit of streaming. The energy efficiency, as measured for the single layers, is 11.5 and 8.4TOPS/W for 5×5 and 3×3 kernels, respectively. The holistic energy efficiency is measured as 3.4TOPS/W when running Inception-v3 on the NPU. For 8b weights, the NPU achieves 2.06× higher efficiency and 10× higher performance compared to [5]. The chip micrograph of the NPU is shown in Fig. 7.1.7.

Acknowledgements:
We thank Jihyuck Jo, Gil-Yeon Lim, Jintaek Noh and Eungjoo Kim for their invaluable support, and the groups of Nir Zoran and Mickey Karnovsky in Samsung Semiconductor Israel R&D Center for their enthusiastic collaboration.

References:
[1] S. Han, et al., “Deep compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding,” arxiv:1510.00149, 2016, https://arxiv.org/abs/1510.00149.
[2] Y. Chen, et al., "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks,” *ISSCC*, pp. 262-263, 2016.
[3] S. Zhang, et al., "Cambricon-X: An Accelerator for Sparse Neural Networks," *IEEE/ACM Int. Symp. Microarchitecture*, 2016.
[4] G. Desoli, et al., "A 2.9TOPS/W Deep Convolutional Neural Network SoC in FD-SOI 28nm for Intelligent Embedded Systems," *ISSCC*, pp. 238-239, 2017.
[5] J. Lee, et al., " UNPU: A 50.6 TOPS/W Unified Deep Neural Network Accelerator with 1b-to-16b Fully-Variable Weight Bit-Precision," *ISSCC*, pp. 218-219, 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

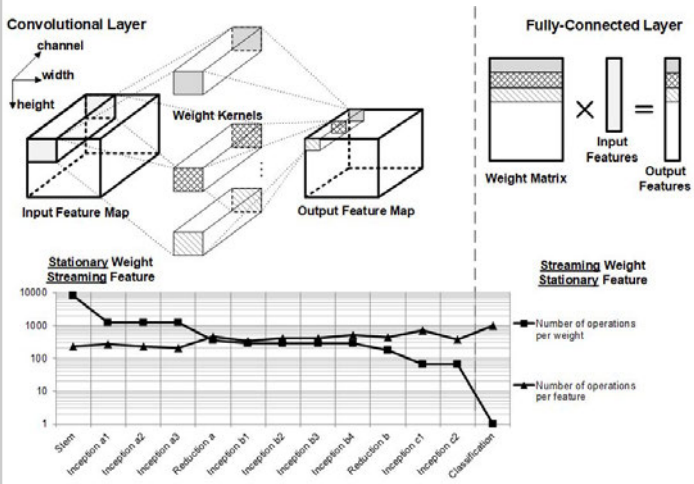


Figure 7.1.1: The number of operations per weight and feature for convolutional and fully-connected layers.

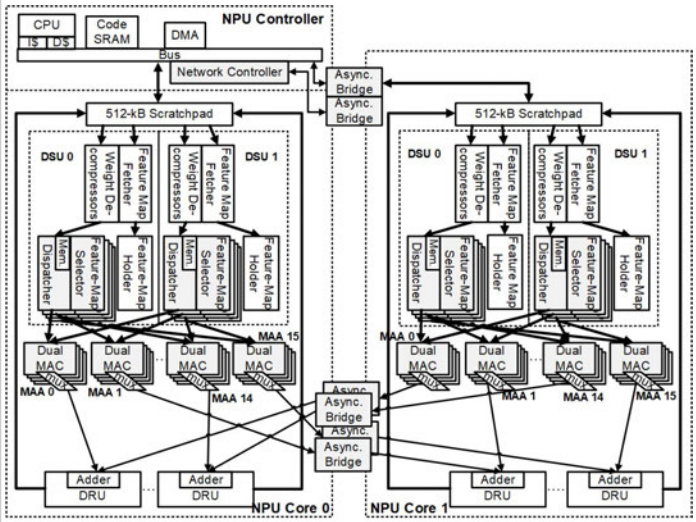


Figure 7.1.2: Overall architecture of the proposed butterfly-structure dual-core neural processing unit.

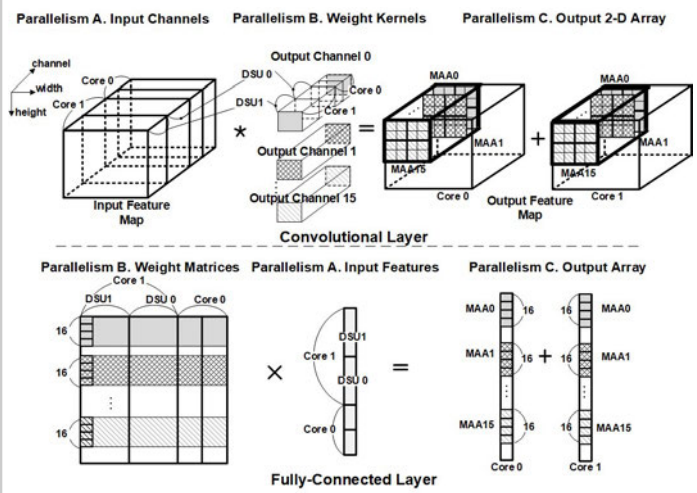


Figure 7.1.3: Threefold parallelism for convolutional and full-connected layers.

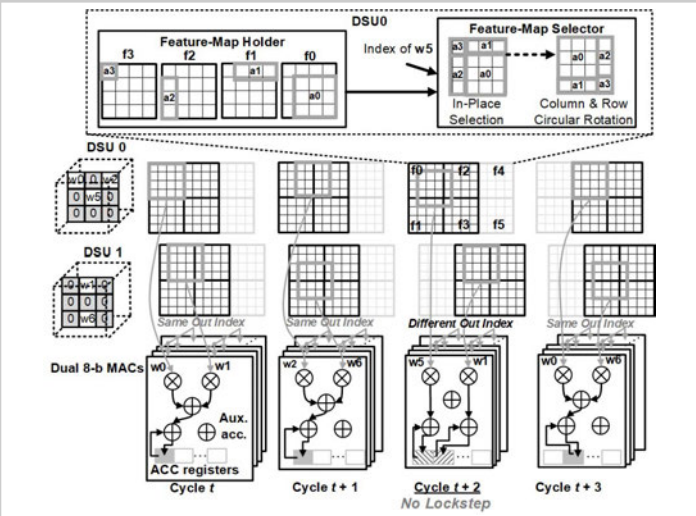


Figure 7.1.4: Feature-map selectors based on in-place selection and circular rotation and dual 8b MACs with lockstep-free operation for skipping weights that are zero.

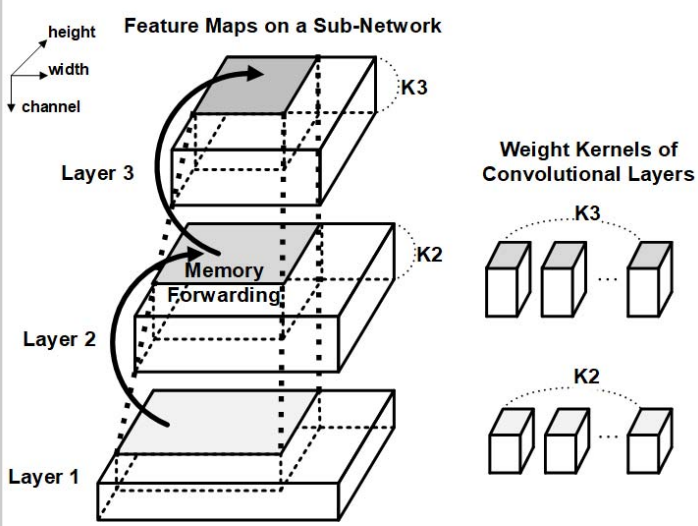
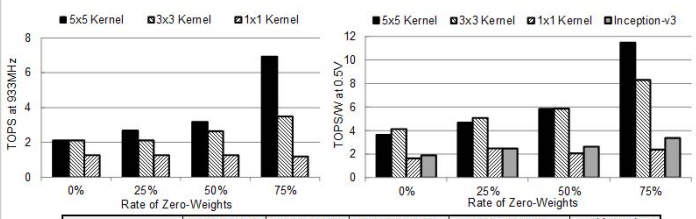


Figure 7.1.5: Network traversal over layers with memory forwarding.



	ISSCC 2016 [2]	MICRO 2016[3]	ISSCC 2017 [4]	ISSCC 2018 [5]	This Work
Purpose	CL	CL, FCL	CL	CL, FCL	CL, FCL
Process (nm)	65	65	28, FDSOI	65	8
Area (mm²)	12.25	6.38	34	16	5.5
Supply Voltage (V)	0.82-1.17	-	0.575-1.1	0.63-1.1	0.5 - 0.8
Working Frequency (MHz)	250	1000	200 - 1175	2 - 200	67 - 933
On-Chip Memory (kB)	181.5	56	5760	256	1568
Bit Precision	16	16	8, 16	1-16	8, 16
Peak Performance (GOPS)	84	544	676 (8b)	7372 (1b), 1382 (4b), 691.2(8b), 345.6 (16b)	1910 (8b) 6937* (8b)
Energy Efficiency (TOPS/W)	0.21	-	2.93 (8b) @ 0.575V	50.8 (1b), 11.6 (4b), 5.57 (8b), 3.08 (16b) @ 0.66V	11.5 (8b) @ 0.5V 4.5 (8b) @ 0.8V
Power (mW)	278 @ 200MHz	954 mW @ 1GHz	39 @ 0.575V	3.2 @ 0.63V 297 @ 1.1V	39 @ 0.5V 1,553 @ 0.8V

* 75% weight zeroes

Figure 7.1.6: Measurement results and performance comparison table.

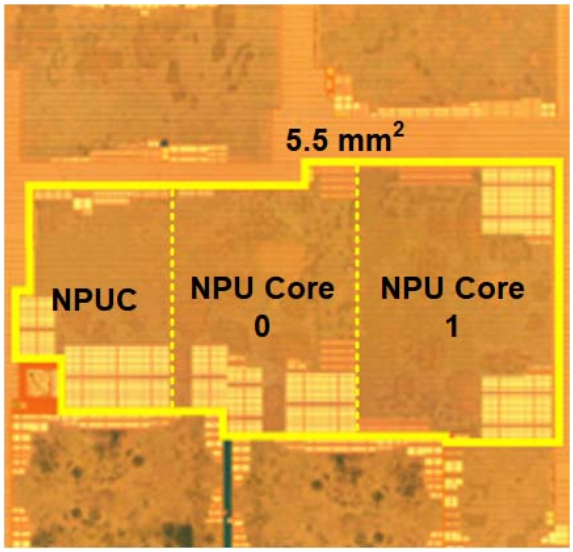


Figure 7.1.7: Chip micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

7.2 A 20.5TOPS and 217.3GOPS/mm² Multicore SoC with DNN Accelerator and Image Signal Processor Complying with ISO26262 for Automotive Applications

Yutaka Yamada, Toru Sano, Yasuki Tanabe, Yutaro Ishigaki, Soichiro Hosoda, Fumihiko Hyuga, Akira Moriya, Ryuji Hada, Atsushi Masuda, Masato Uchiyama, Tomohiro Koizumi, Takanori Tamai, Nobuhiro Sato, Jun Tanabe, Katsuyuki Kimura, Ryusuke Murakami, Takashi Yoshikawa

Toshiba Electronic Devices and Storage, Kawasaki, Japan

Vehicles with features for an advanced driver-assistance system (ADAS) are already available on the market. Technologies for image recognition are essential elements for ADAS applications, and several SoCs performing image recognition for ADAS have been developed [1-3]. As vehicles that support ADAS become more complicated, these technologies need improvement. Recently, deep neural networks (DNNs) have achieved higher recognition accuracy than traditional feature-based matching algorithms, and can be applied to applications such as road detection. Consequently, several image recognition SoCs with DNN processors [4-5] have been developed. Functional safety is important for automotive applications, so several SoCs, such as [2], comply with ISO26262. The requirements for SoCs that support ADAS features are: high performance for detecting various objects, low power use to keep execution stable in the rapidly changing environment of moving vehicles, and high safety to avoid serious accidents.

In this paper, we implement an image-recognition SoC for ADAS applications. Fig. 7.2.1 presents a block diagram of the SoC. As the SoC performs both image recognition processes that need high computing power, and control processes that need high reliability, we partitioned the SoC into two regions. One region is called the "Processing Island" (PI) for image recognition processes, and the other is called the "Safe Island" (SI) for control processes. PI includes two clusters of quad ARM Cortex-A53 processors for application processors, four DSPs for various image processing tasks, eight kinds of hardware-dedicated accelerators, a two-channel 32b LPDDR4 interface, 16Mb on-chip SRAM, and several I/O interfaces. The accelerators are: 1) a DNN accelerator for classification and segmentation, 2) two channel image signal processors (ISPs) for making color images from raw images, 3) AKAZE for feature point tracking, 4) STMAT for matching stereo images, 5) PYRAM for making pyramid images, 6) MATCH for object tracking, 7) AFFINE for making affine transformation images, and 8) HOX for detecting various objects. SI includes two dual-core lock-step (DCLS) ARM Cortex-R4 processors for control processors, a hardware secure module (HSM) for managing security data such as cryptographic keys and performing encryption and decryption, 8Mb on-chip SRAM, and various I/O interfaces. Additionally, there are memory protection units (MPUs) in each accelerator and each memory unit.

Figure 7.2.2 shows an overview of the DNN accelerator. DNN applications require many multiply-accumulate (MAC) computations, and high memory bandwidth for weight and feature-map data. The DNN accelerator has 1024 MAC units in its execution unit, and its performance achieves 1.6TFLOPS. The MAC units support IEEE754 binary-16 format, because automotive applications require more accuracy than consumer applications. The accelerator has two features for reducing bandwidth of external memory: 1) the decompression circuit in the DMA unit can reduce the bandwidth of external memory by loading compressed weight data and decompressing the data; and 2) the local memory to reduce the bandwidth requirements of transferring temporary data. The execution unit accesses only the local memory, and the DMA unit transfers data between the external memory and the local memory.

Figure 7.2.3 shows an overview of the ISP. The ISP supports HDR images because images of automotive applications need high dynamic range to continually detect objects regardless of driving conditions, day or night. The ISP includes not only a demosaic unit, but also several image-quality-correction units, an HDR compression unit, an auto-exposure unit, an auto-white-balance correction unit, and a histogram unit. As all of these functions can be performed without storing temporary data to external memory, the ISP achieves low latency and requires low bandwidth to external memory. The pixel rate of each ISP is 600Mpixels/s, and its performance is over 4.4TOPS.

The goal of functional safety is to reduce the risk of accidents to acceptable levels. The approaches for reducing the risk are to reduce failure rates and increase fault coverage. ISO26262 describes a safety risk level called automotive safety integrity level (ASIL) which is classified from ASIL-A to ASIL-D, based on the risk assessment. To reduce risk, introducing safety mechanisms (SMs) is required. As the processes of SI require more integrity than those of PI, we assume the ASIL of SI (ASIL-D) is higher than that of PI (ASIL-B). Therefore, the SMs of SI are implemented in order to make them more robust than those of PI. The SMs of PI are: 1) parity/ECC circuits with SRAM. Only SRAM in the ISP has ECC circuits, because output data of the ISP is used by many processes; 2) run-time BIST circuits for logic circuits; 3) ECC circuits with the bus for detecting and correcting errors during bus access; and 4) several monitor circuits, such as a clock monitor. The SMs of SI are: 1) ECC circuits with SRAM, 2) duplicated circuits and monitor circuits for detecting faults in the random logic, 3) ECC circuits with the bus, and 4) several duplicated monitor circuits. Furthermore, MPUs of DDR I/F and Work RAM of PI protect the memory region allocated by SI to avoid undesired access from PI. The placement and the connection of the MPUs and the accessibility of each bus master are shown in Fig. 7.2.4.

The run-time BIST of the accelerators can be controlled by the processors. However, it is difficult for the processor to control the run-time BIST of the ISP, because the process of the ISP begins when a new frame is received from an external imager. Therefore, a dedicated controller for run-time BIST of the ISP is introduced. The run-time BIST controller of the ISP is shown in Fig. 7.2.3. As the internal parameters of the ISP are used in several frames, the backup unit backs up and restores the parameters across run-time BIST. The BIST unit executes the run-time BIST process in the V blanking period. Fig. 7.2.5 shows the timing chart of the run-time BIST. The sequence of Case A is the following: to finish processing of a frame, the controller: 1) backs up the parameters, 2) performs run-time BIST, 3) restores the parameters. Then, 4) the ISP waits to execute the main process until receiving the next frame. Tb and Tr, which are the start times of 2) and 3), are preset by the processors. In Case B, when the end time of 1) is later than the start time of 2) due to high bus latency, the controller does not execute run-time BIST and sends an alarm to the control processors.

Figures 7.2.6 and 7.2.7 show the performance and die photograph of the SoC implemented in 16nm. Its size is 94.52mm². Peak performance achieved is 20.5TOPS and total power consumption is 9.78W. Power consumptions in Fig. 7.2.6 are simulations and represent a full utilization scenario. The power consumption of PI measured by the evaluation board is 2.73W, in the case of executing pedestrian and vehicle detection using 2 cores of ARM Cortex-A53, 4 DSPs, and 6 kinds of accelerators, including the DNN and HOX.

References:
[1] J. Tanabe, et al., "A 1.9TOPS and 564GOPS/W Heterogeneous Multicore SoC with Color-based Object Classification Accelerator for Image-Recognition Applications," *ISSCC*, pp. 328-329, 2015.
[2] C. Takahashi, et al., "A 16nm FinFET Heterogeneous Nona-Core SoC Complying with ISO26262 ASIL-B: Achieving 10^{-7} Random Hardware Failures per Hour Reliability, " *ISSCC*, pp. 80-81, 2016.
[3] K. Jason Lee, et al., "A 502-GOPS and 0.984-mW Dual-Mode Intelligent ADAS SoC With Real-Time Semiglobal Matching and Intention Prediction for Smart Automotive Black Box System," *IEEE JSSC*, Vol. 52, No. 1, pp. 139-150, 2017.
[4] J. Lee, et al., "UNPU: A 50.6TOPS/W Unified Deep Neural Network Accelerator with 1b-to-16b Fully-Variable Weight Bit-Precision," *ISSCC*, pp. 218-219, 2018.
[5] G. Desoli, et al., "A 2.9TOPS/W Deep Convolutional Neural Network SoC in FD-SOI 28nm for Intelligent Embedded Systems," *ISSCC*, pp. 238-239, 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

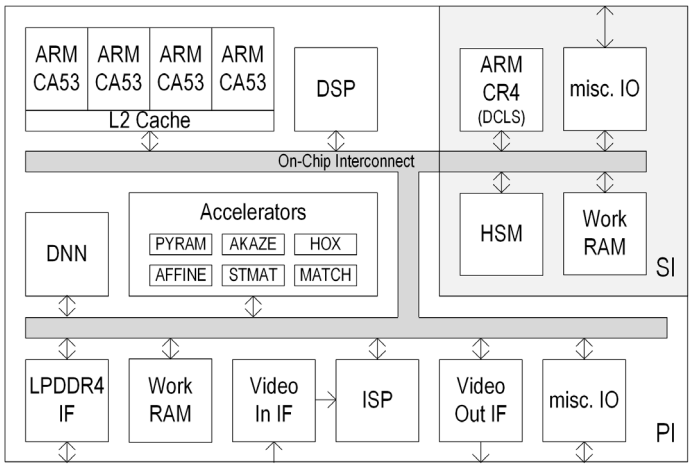


Figure 7.2.1: Block diagram of the SoC.

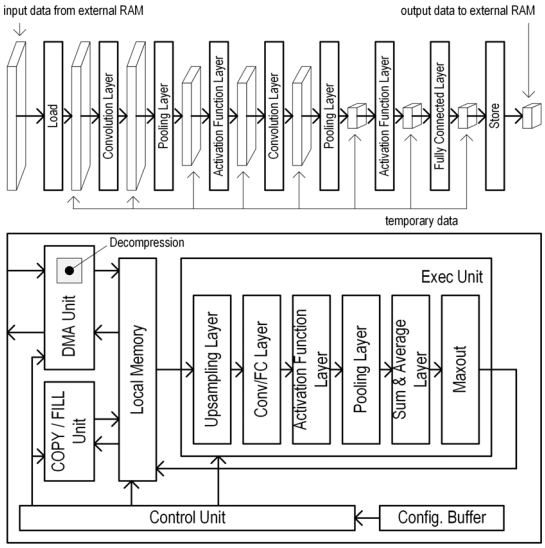


Figure 7.2.2: Example of the DNN process and overview of the DNN accelerator.

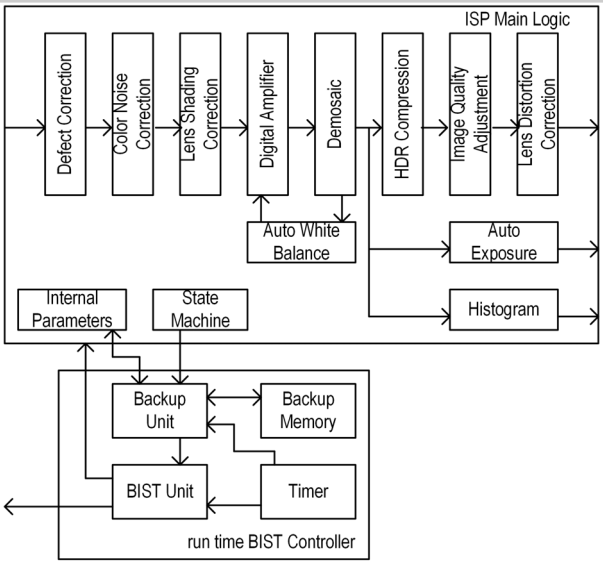
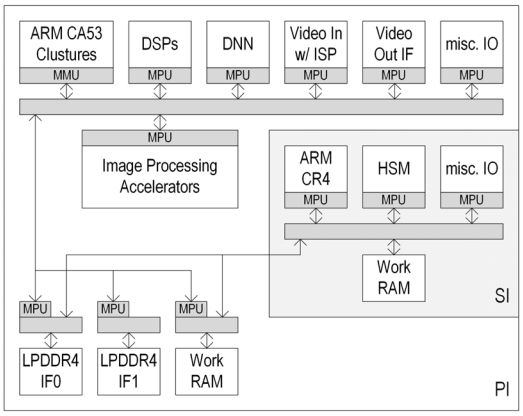


Figure 7.2.3: Overview of the ISP with the run-time BIST controller.



	Work RAM in SI	Work RAM in PI	LPDDR4 IF0	LPDDR4 IF1
Master in SI	Not protected	Not protected	Not protected	Never accessed
Master in PI	Never accessed	Protected	Protected	Protected

Figure 7.2.4: Connection of the MPUs and the accessibility of bus masters in the SoC.

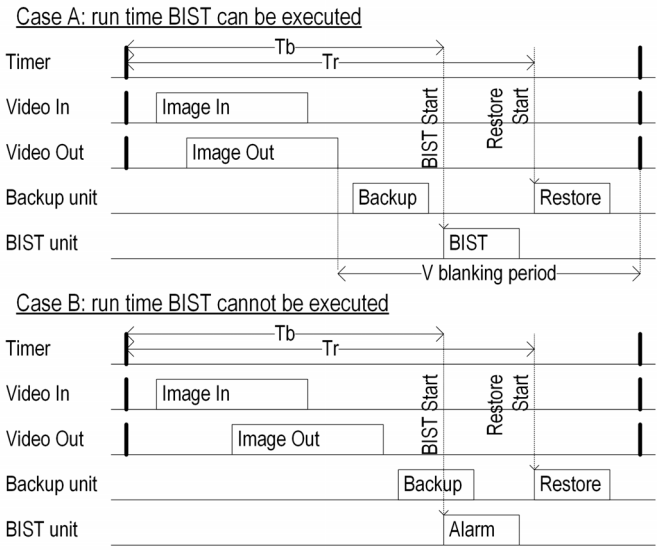


Figure 7.2.5: Timing chart of run-time BIST of the ISP.

Performance Evaluation of the SoC

	Performance [GOPS]	Power [mW]	Efficiency [GOPS/W]
ARM CA53	256	1,286	199.1
DSP	1,024	1,377	743.6
DNN	1,597	1,593	1,002.8
ISP	8,880	1,579	5,623.8
STMAT	3,480	419	8,305.5
AKAZE	2,681	726	3,692.6
MATCH	1,175	254	4,627.6
PYRAM	396	130	3,046.2
AFFINE	204	281	726.0
HOX	842	1,696	496.7
ARM CR4	1	435	2.8
Total	20,537	9,776	2,100.8

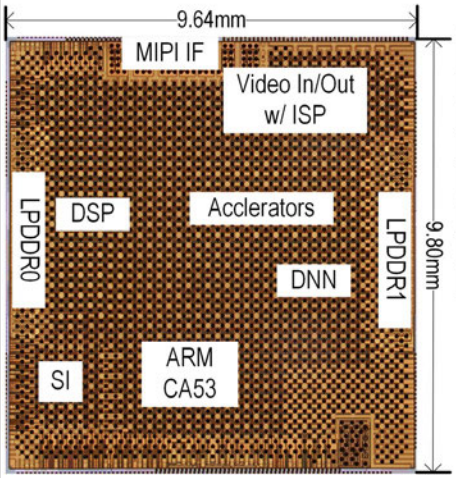
Power consumptions include functional safety logics
*Vdd=0.8V; Process=center; Temp=25°C

Performance Comparison

	This Work	Our prev. work ISSCC 2015 [1]	JSSC 2017 [3]	ISSCC 2017 [4]	ISSCC 2018 [5]
Application	Automotive	Automotive	Automotive	Embedded	IoT
Process	16nm	40nm	65nm	28nm	65nm
Peak Performance [GOPS]	20,537	1,900	502	751	346
Power Efficiency [GOPS/W]	2,101	564	862	2,930	3,080
Area Efficiency [GOPS/mm2]	217.3	18.0	31.4	21.5	21.6
Normalized Area Efficiency*	217.3	65.9	239.3	52.2	347.6

*Normalized Area ratio (16nm : 28nm : 40nm : 65nm) assume to be (1 : 2.43 : 3.66 : 11.08)

Figure 7.2.6: Performance results and comparison.



Process	16nm
Chip Size	94.52mm2
On-Chip SRAM	142.9 M bit
Supply Voltage	
SI Core	0.8V
SI I/O	1.8~3.3V
PI Core	0.8V
PI I/O	1.8~3.3V

Figure 7.2.7: Chip photograph of the SoC.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

7.3 An 879GOPS 243mW 80fps VGA Fully Visual CNN-SLAM Processor for Wide-Range Autonomous Exploration

Ziyun Li, Yu Chen, Luyao Gong, Lu Liu, Dennis Sylvester, David Blaauw, Hun-Seok Kim

University of Michigan, Ann Arbor, MI

Simultaneous localization and mapping (SLAM) estimates an agent's trajectory for all six degrees of freedom (6 DoF) and constructs a 3D map of an unknown surrounding. It is a fundamental kernel that enables head-mounted augmented/virtual reality devices and autonomous navigation of micro aerial vehicles. A noticeable recent trend in visual SLAM is to apply computation- and memory-intensive convolutional neural networks (CNNs) that outperform traditional hand-designed feature-based methods [1]. For each video frame, CNN-extracted features are matched with stored *keypoints* to estimate the agent's 6-DoF pose by solving a *perspective-n-points* (PnP) non-linear optimization problem (Fig. 7.3.1, left). The agent's long-term trajectory over multiple frames is refined by a *bundle adjustment* process (BA, Fig. 7.3.1 right), which involves a large-scale (~120 variables) non-linear optimization. Visual SLAM requires massive computation (>250GOP/s) in the CNN-based feature extraction and matching, as well as data-dependent dynamic memory access and control flow with high-precision operations, creating significant low-power design challenges. Software implementations are impractical, resulting in 0.2s runtime with a ~3GHz CPU+ GPU system with >100MB memory footprint and >100W power consumption. Prior ASICs have implemented either an incomplete SLAM system [2,3] that lacks estimation of ego-motion or employed a simplified (non-CNN) feature extraction and tracking [2,4,5] that limits SLAM quality and range. A recent ASIC [5] augments visual SLAM with an off-chip high-precision inertial measurement unit (IMU), simplifying the computational complexity, but incurring additional power and cost overhead.

This paper presents an accurate, low-power, real-time CNN-SLAM processor that implements full-visual SLAM on a single chip. The proposed major design features are 1) feature extraction with a highly parallelized and programmable CNN engine with 32% better accuracy in feature matching than SIFT; 2) aggressively pruned feature matching with temporal pose prediction and address hashing to eliminate 97% of unnecessary matchings; 3) numerically stable fixed-point implementation with function re-organization and pivoting in the linear solver; and 4) hierarchical memory organization, eliminating external DRAM accesses for BA optimization. We report the ASIC's SLAM performance on the industrial standard KITTI benchmark that renders large-scale automobile trajectory over 1km. The design supports localization and mapping of >1000 keypoints/frame on VGA (640×480) resolution in real-time at 80fps, consuming 243.6mW from a 0.9V supply in 28nm CMOS.

Figure 7.3.1 shows the overall procedure of the SLAM system, consisting of frame-by-frame PnP and multi-frame BA. PnP processing computes the 6-DoF pose (combination of 3D translation $\boldsymbol{T}$ and 3D rotation $\boldsymbol{R}$) for each frame by analyzing the projected coordinate differences of matched keypoints in two subsequent frames. CNN-extracted keypoints are matched with others in the previous frame to establish correspondence of the same points in the environment. The 6-DoF pose is obtained by solving a non-linear optimization to minimize the 2D reprojection error of these matched points. After acquiring the pose of the current frame, all newly added keypoints are projected onto 3D real-world coordinates for PnP processing of the next frame. A frame that contains ≥50% new keypoints is registered as a *keyframe*. To refine the long-term trajectory of the camera, BA is performed over the last 20 keyframes, whenever a new keyframe is registered.

Figure 7.3.2 shows the overall architecture of the processor, including a programmable CNN engine, PnP engine, and BA engine. The CNN engine has a mesh of 512 MAC units in 8 clusters (8×8 MACs each), a 192KB memory for CNN weights/activations, and two interleaved 32KB image buffers for streaming (Fig. 7.3.3, left). The CNN uses 8b weights and does not require any off-chip weight loading. If the current frame is identified as a keyframe, the list of 3D world and 2D projected coordinates of keypoints is sent to the graph memory in the BA engine, where iterative Levenberg–Marquardt (LM) optimization is performed over the last 20 keyframes.

Figure 7.3.3 details the operation of the programmable CNN engine. First, 1D convolutions are performed with shifting input activations (IAs) using a row of 8 MAC units with k (kernel size) cycles. This operation is repeated on a second and third row of IAs to complete the 2D convolution. Pooling and extrema detection are performed in the same fashion as 2D convolution in the MAC array. Partial 2D convolution outputs are accumulated locally in each MAC unit. Using 8 clusters of 8×8 MAC arrays, 8 consecutive input channels are convolved in parallel to accumulate partial output activations with 8 accumulators, and 8 consecutive output channels are processed in parallel to maximize reuse of the same IAs. A 4-layer CNN network used for SLAM evaluation is shown in Fig. 7.3.3 (bottom left). The proposed architecture computes convolution, pooling, and ReLU together rather than treating them as separate layers with buffering of inputs-outputs in between. For keypoint detection, the engine is programmed to compute the difference of Gaussian filters.

As each frame contains ~1000 keypoints, enumerating all possible matchings (~1000×1000) is extremely costly. Addressing this challenge, we propose a *prediction-based* pruned keypoint matching scheme depicted in Fig. 7.3.4. We use a locally linearized camera movement model to predict the new pose of the current frame from the poses of the previous two frames. Based on this pose prediction, the search range for each keypoint is reduced to ~±24 pixels/dimension on the 2D projected image. This eliminates 97% of unnecessary matchings with negligible accuracy degradation (<0.1%). Prediction-based matching efficiency is further improved by employing a hierarchical memory system (492KB) using keypoint position-based hashing to store/load feature descriptors and 3D coordinates. The hash input is the keypoint location, and its output points to the keypoint entry in the memory that stores a list of feature descriptor and 3D coordinates at that location. Each feature has 64 elements, and the feature matching cost is calculated with 8 parallel processing units, taking up to 8 cycles. Early termination reduces this to ~4.8 cycles when feature elements mismatch in early stages (Fig. 7.3.4).

BA processing involves 20 keyframes and 4096 keypoints stored in the hierarchical graph memory (Fig. 7.3.5). Because each keypoint appears on multiple keyframes, each entry is structured to contain a single 3D coordinate and multiple (≤8) 2D projected coordinates associated with different keyframe IDs. A separate FIFO serves to eliminate and insert keypoints into the memory, while matched keypoints are merged into a single entry. In each BA iteration, the keypoints' 3D coordinates are updated according to the frame pose, and very small increments are applied to numerically compute the Jacobian matrix. Computing Jacobians requires an extremely large dynamic range and high precision; thus, prior works [4,5] used double-precision floating-point operations. We reformulate the reprojection so that common offset is subtracted before normalizing the projected 2D points into homogeneous coordinates. This allows a 32b fixed point implementation with ~40% energy reduction, while maintaining numerical stability in computing the Jacobian. After linearization, we construct a sparse Hessian matrix that only has non-zero 6×6 submatrices diagonally that relate to the 6-DoF pose of each keyframe. Thanks to the sparse and deterministic matrix structure, we block partition the Hessian and solve each 6×6 submatrix sequentially instead of solving the full 120×120 linear system. We apply pivoting to Gaussian elimination (GE) to maintain numerical stability in the linear system solver. With pivoting, non-zero elements in other rows are eliminated with the maximum element of the selected row. Row shuffling is performed before back substitution (BS), and GE/BS share 6 parallel computing units.

The SLAM processor is fabricated in 28nm HPC CMOS. Real-time image input and 6-DoF SLAM output are streamed using USB3.0. Fig. 7.3.6 (top left) shows the trajectory produced by the chip for KITTI automobile scenes with >1000 images and >500m range, tracking >1000 keypoints per frame. Fig. 7.3.6 (top right) shows frequency/energy scaling of the chip across voltage. At 0.9V nominal voltage, the VGA processing latency is 12.5ms. It achieves 97.9% accuracy in translation and 99.34% in rotation on KITTI rendering automotive scenes over 1km. The chip consumes 243.6mW to process 80fps VGA at 3.6TOPS/W, marking a 15× improvement in performance and 1.44× in energy efficiency over listed prior works. Prior works use hand-crafted features or off-chip IMU and do not support large-scale KITTI evaluation. Power reduces to 61.8mW for VGA at 30fps @ 0.63V, yielding 48% additional energy efficiency. Fig. 7.3.7 shows the die photo and a performance summary.

Acknowledgements:
We thank TSMC the University Shuttle Program for chip fabrication and Samsung for funding.

References:
[1] K. Tateno, et al., "CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction," *Computer Vision and Pattern Recognition*, vol. 2, 2017.
[2] Z. Li, et al., "A 1920× 1080 30fps 2.3 TOPS/W Stereo-Depth Processor for Robust Autonomous Navigation," *ISSCC*, pp. 62-63, 2017.
[3] J. Yoon, et al., "A Unified Graphics and Vision Processor With a 0.89 μW/fps Pose Estimation Engine for Augmented Reality," *IEEE TVLSI,* vol. 21, no. 2, pp. 206-216, 2013.
[4] I. Hong, et al., "A 27 mW Reconfigurable Marker-Less Logarithmic Camera Pose Estimation Engine for Mobile Augmented Reality Processor," *IEEE JSSC*, vol. 50, no. 11, pp. 2513-2523, 2015.
[5] A. Suleiman, et al., "Navion: A Fully Integrated Energy-Efficient Visual-Inertial Odometry Accelerator for Autonomous Navigation of Nano Drones," *IEEE Symp. VLSI Circuits*, 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

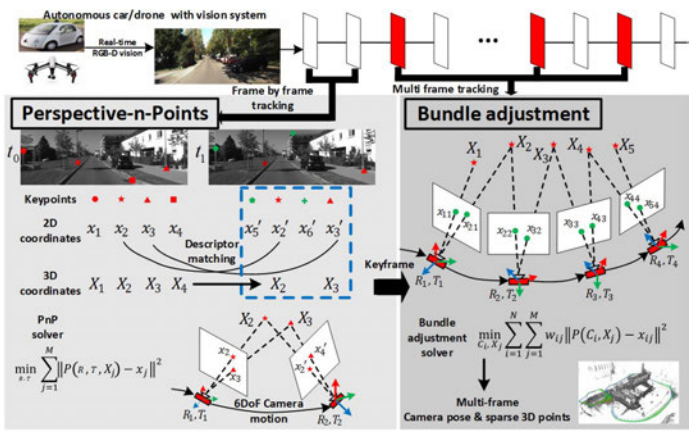


Figure 7.3.1: Proposed visual SLAM system for autonomous vehicles and associated processing procedures.

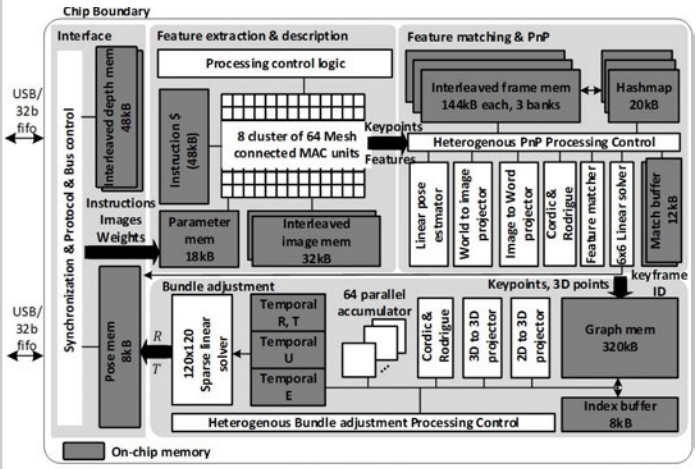


Figure 7.3.2: Overall CNN-SLAM chip architecture.

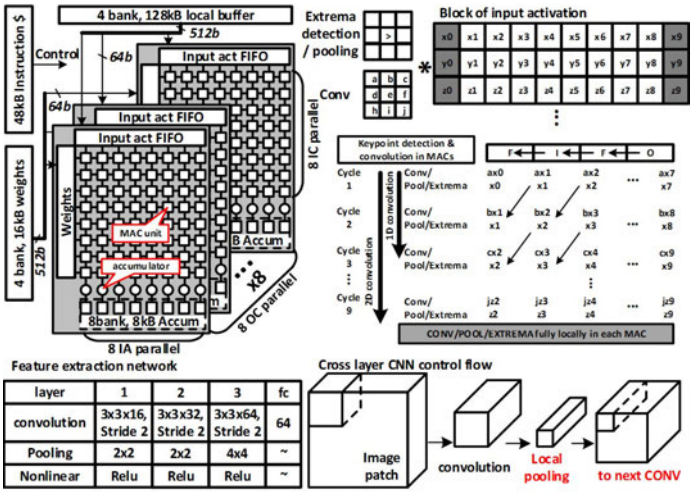


Figure 7.3.3: Proposed CNN processor architecture with reconfigurable keypoint detection and feature description.

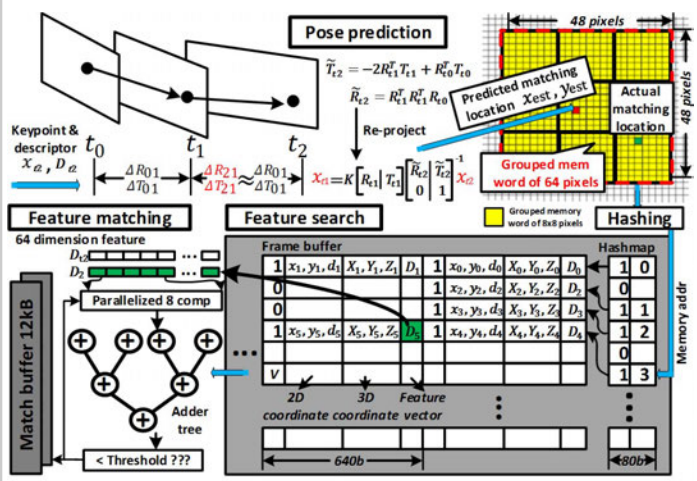


Figure 7.3.4: Proposed frame-based tracking with pruned feature matching with linear pose estimation.

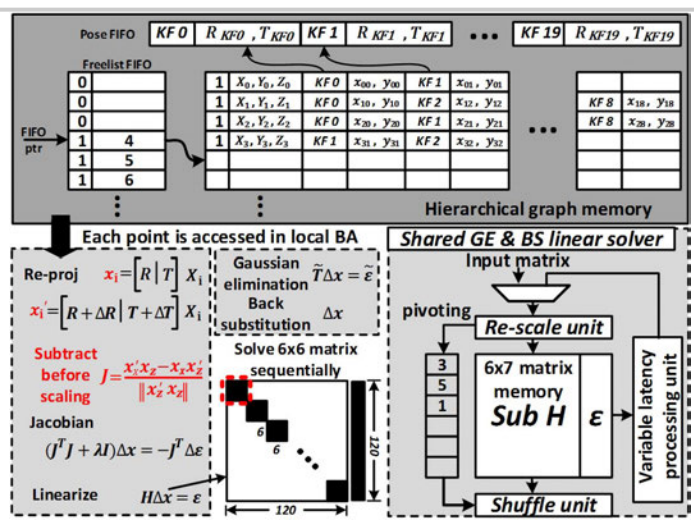
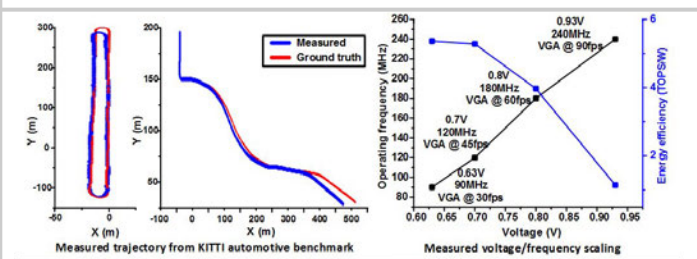
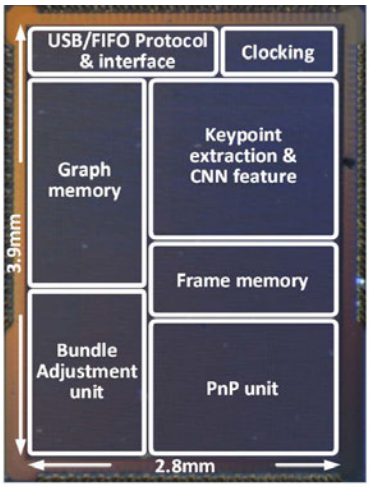


Figure 7.3.5: Hierarchical graph memory organization and proposed numerically stable fixed-point implementation with function re-formulation and pivoting in linear solver.



		This work	TVLSI 2013[3]	JSSC 2015[4]	VLSI 2018[5]
Complete Visual SLAM system	Vision frontend	√ CNN feature	√ Marker recognition	×	√ Harris feature
	Pose tracking	√ Visual	×	√ Visual	× Inertial sensor
	Graph optimization	√ Local BA	×	×	√ Local BA
Technology		28nm	180nm	65nm	65nm
Chip area		10.92mm²	28.75mm²	3.25mm²	20mm²
On-chip memory		1126kB	N.A	96KB	854kB
Frequency		240MHz	200MHz	133MHz	83.3MHz
Throughput & image size		640 X 480, 80fps	640 X 480, 100fps	640 X 480, 95fps	752 X 480, 90fps
Operation range		± 500m	Not reported due to limited range		
Track points / frame		1000	9	33	200
Operating voltage		0.63-0.9V	1.8V	1.2V	1.2V
Performance		879.6 GOPS @ 0.9V, 215MHz 329.8 GOPS @ 0.63V, 90MHz	153.6 GOPS	20.2 GOPS	10.5-59.1 GOPS
Power		243.6 mW @ 0.9V, 215MHz 61.75 mW @ 0.63V, 90MHz	413 mW	27 mW	24 mW* (excluding IMU power)
Energy efficiency		3.6 TOPS/W @ 0.9V, 215MHz 5.34 TOPS/W @ 0.63V, 90MHz	0.37 TOPS/W	0.75 TOPS/W	0.43-2.5 TOPS/W

Figure 7.3.6: Chip measurements and comparison with recent prior works.



	This work
Vision frontend	CNN feature
Pose tracking	PnP
Graph optimization	Local BA
Technology	28nm
Chip area	10.92mm^2
On-chip memory	1126kB
Frequency	240MHz
Throughput & image size	640 X 480, 80fps
Operation range	± 500m
Track points / frame	1000
Operating voltage	0.9V
Performance	879.6 GOPS @ 0.9V, 215MHz 329.8 GOPS @ 0.63V, 90MHz
Power	243.6 mW @ 0.9V 61.75 mW @ 0.63V
Energy efficiency	3.6 TOPS/W @ 0.9V 5.34 TOPS/W @ 0.63V

Figure 7.3.7: Die photo and summary of performance.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

7.4 A 2.1TFLOPS/W Mobile Deep RL Accelerator with Transposable PE Array and Experience Compression

Changhyeon Kim, Sanghoon Kang, Dongjoo Shin, Sungpill Choi, Youngwoo Kim, Hoi-Jun Yoo

KAIST, Daejeon, Korea

Recently, deep neural networks (DNNs) are actively used for object recognition, but also for action control, so that an autonomous system, such as the robot, can perform human-like behaviors and operations. Unlike recognition tasks, real-time operation is important in action control, and it is too slow to use remote learning on a server communicating through a network. New learning techniques, such as reinforcement learning (RL), are needed to determine and select the correct robot behavior locally. Fig. 7.4.1(a) shows an example of a robot agent that uses a pre-trained DNN without RL, and Fig. 7.4.1(b) depicts an autonomous robot agent that learns continuously in the environment using RL. The agent without RL falls down if the land slope changes, but the RL-based agent iteratively collects walking experiences and learns to walk even though the land slope changes.

Deep RL (DRL), a combination of RL with a DNN [1,2], adapts the DNN through two steps: sample collection (SC) and policy update (PU) for continuous control of behavior. First, during the SC step, the agent collects sampled experiences composed of the input state (S_t), the output action (O_t) and the corresponding reward (R_t). Once enough experience samples are collected, the PU step starts to compute the loss (LC), and then, updates the weights (WU) of the DNN. The loss is computed to maximize the reward for the agent's actions, and the weight update is performed with calculated gradients from the loss. As shown in the Fig. 7.4.1(c), the SC step uses a single input feature (IF) batch, necessitating IF reuse for efficiency. On the other hand, the PU step uses multiple IF batches, necessitating W reuse for efficiency. Therefore, reconfigurable HW that can flexibly reuse IF or W is required for DRL. Moreover, the LC and WU steps have the same data reusability, but an additional transpose operation is required for multiplying the transposed W in the WU step. Previous DNN HW [3,4] performed only inference, with specific parallelism assumptions, and a fixed datapath. As such, in prior work, memory accesses and core utilization were not optimized for varied DNN processing cases. Even though [5] is able to tune the DNN, only trivial modifications for fully connected layers (FCLs) are possible with a limited number of special PEs.

We introduce a DRL accelerator optimized for training an RNN and the FCL of a DNN with highly reconfigurable PEs. Since the data-reuse requirements of DNN Ws and IFs for learning is different from that of inferencing, a new reconfigurable PE, Transposable PE, is used to map the Ws and IFs to different memories, BMEM and UMEM, without additional matrix-transpose operations. Reconfiguration of the memory mapping can reduce memory accesses by 49%, during SC and PU processing. In addition, experience samples in the SC step and partial products in the WU step are compressed to shrink storage requirements (139MB/20000 experiences), and consequently, external bandwidth and power are reduced by 37% and 31%, respectively.

Figure 7.4.2 shows the overall architecture of the proposed mobile DRL accelerator. It consists of 4 DRL cores, the top controller with top-shared memory. The top controller and the 4 DRL cores are connected with an on-chip network. Each DRL core has an experience compressor, 16 PE arrays, 16-way activation/1D SIMD units, and a RISC controller. The broadcasting memory (BMEM) and unicasting memory (UMEM) are connected to a row-buffer and column-buffers, respectively, to provide data for the PE array. The output of the PE array is aggregated in the accumulation unit. The 16-way activation unit processes nonlinear functions, and a 16-way 1D SIMD unit adds, multiplies and performs logarithm functions for weight update and loss computation.

Figure 7.4.3 shows the architecture of the Transposable 2D PE array, which performs matrix multiplication. The unicasted B_{11}-B_{41} data at different PEs can be multiplied by A_1 data broadcasted throughout the first row to perform a parallel matrix operation. 16 4×4 transposable PE arrays are integrated and share a row buffer that receives broadcast data from BMEM. Each of the 16 PE arrays is connected to a column buffer that feeds unicast data. The unicasted data is stored in the PE array for reuse, and the broadcasted data is fetched every cycle through the row buffer. The PE performs all necessary operations reusing the unicasted data. And, after all of the unicasted data is reused, new unicasted data is fetched into the PE array through the column buffer. Input features are accessed C_0 (# of output channels) times, and the weights are accessed BA (# of batches for network update) times. In SC, when BA is smaller than C_0, the reuse of IF is more efficient than reuse of W (case (2)). In contrast to SC, in the PU step, the reuse of W is more efficient because of the larger BA (case (3)). In the case of C_0 or BA (case (1)) being smaller than PE array width, IF broadcasting *and* W broadcasting is used to fully utilize the PE array. A floating-point/fixed-point MAC unit is integrated in each PE, which performs multiplication of 16b floats and 16b fixed-point numbers. Four 4b multiplication and two 8b multiplication results are obtained without additional cost, so that numerical weight precision can be varied according to different accuracy/performance requirements.

Figure 7.4.4 shows the memory architecture, the transposed memory controller and the buffer-feeding logic. The 4 cores, each representing different workers in the DRL process, collaborate and compute in parallel with tiled weight matrices and tiled feature matrices. The transposed memory controller distributes the fetched weights into the core's local memory, UMEM and BMEM. Suppose the weights are a sequence of row vectors stored in column order. For the SC and the LC steps, which perform inference, weights should be fed in column-major manner. On the other hand, during the WU stage, transposed weights are required for error propagation. To feed the weights in transposed order, the transposed memory controller feeds the weights in a row-major manner. For the SC stage, the weights are stored in the UMEM and transferred into different columns of the PE array. For the LC and the WU stage, which require batch computations, the weights stored in the BMEM are fed into the PE array in the row direction. With the proposed transposed memory controller and configurable core architecture, the needed computations for DRL are fully supported without the additional burden of weight transposition.

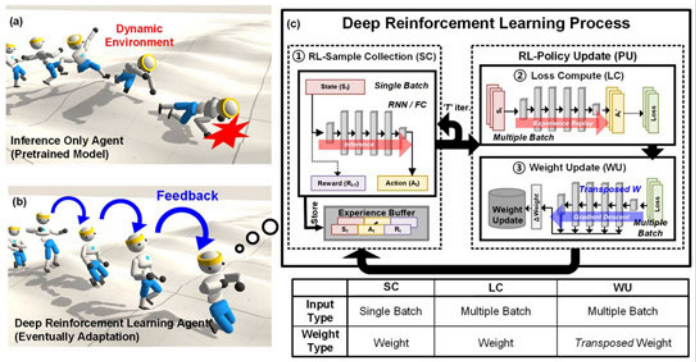
Figure 7.4.5 shows the compression scheme for experiences and partial products (PPs). The experience compressor is composed of an encoder and decoder. Since the range of IF is widely distributed, the IF is represented as 16b float in this work. As shown in the graph, the exponent of an input layer is highly concentrated in few values. The average ratio of 3 most-frequent exponent values (top 1-3 exponents) are 68% and 85% for 20480 experiences and PPs of LC and WU, respectively. A 2b code word indicates whether the current IF node is encoded or not. Codeword 00 indicates that the exponent value cannot be skipped because it does not belong to the top-3 exponents. Codewords 00,01, and 11 indicate that the exponent of the current node belongs to the top-3 exponents, respectively, thereby skipping the exponent value. The experience encoder scans the 1D IF to determine the 3 most-frequent exponent values and then sequentially encodes codewords, fractions, and exponents. The decoder scans code words and re-synthesizes sequential input fractions and exponents to stream-out IF data as a 16b float types. Through this compression approach, the bandwidth for experience and PP is reduced by 37% in the PU step, and power consumption is reduced by 31%.

Figure 7.4.6 shows the measurement results. The top-left figures show the RL results on hill climbing. The red and the black lines indicate the climb distance without DRL and with DRL, respectively. As steps progress, and the angle of the hill increases, the climb distance continuously increases with DRL. The proposed DRL accelerator can operate from 0.67-to-1.1V supply @ 5-200MHz clock frequency. The power consumption at 0.67V and 1.1V are 2.4mW and 196mW, respectively. The table shows a performance comparison with four previous deep-learning SoCs. This work is the only one that supports RL. The DRL accelerator shown in Fig. 7.4.7 is fabricated in 65nm CMOS technology and occupies 16mm^2, achieving 2.16TFLOPS/W energy-efficiency.

Acknowledgements:
This work was supported by the Samsung Research Funding & Incubation Center for Future Technology under Grant SRFC-TB1703-09.

References:
[1] M. Volodymyr, et al., "Human-level control through deep reinforcement learning." Nature 518.7540 (2015): 529.
[2] J. Schulman, et al., "Proximal policy optimization algorithms" arXiv:1709.06560, 2017
[3] J. Lee, et al., "UNPU: A 50.6TOPS/W Unified Deep Neural Network Accelerator with 1b-to-16b Fully-Variable Weight Bit-Precision," *ISSCC*, pp. 218-220, 2018.
[4] K. Ueyoshi et al., "QUEST: A 7.49TOPS Multi-Purpose Log-Quantized DNN Inference Engine Stacked on 96MB 3D SRAM Using Inductive-Coupling Technology in 40nm CMOS," *ISSCC*, pp. 216-218, 2018.
[5] Z. Yuan, et al., "STICKER: A 0.41-62.1 TOPS/W 8bit Neural Network Processor with Multi-Sparsity Compatible Convolution Arrays and Online Tuning Acceleration for Fully Connected Layers", *IEEE Symp. VLSI Circuits*, pp. 33-34, 2018.
[6] B. Fleischer, et al., "A Scalable Multi-TeraOPS Deep Learning Processor Core for AI Training and Inference", *IEEE Symp. VLSI Circuits*, pp. 35-36, 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE



	SC	LC	WU
Input Type	Single Batch	Multiple Batch	Multiple Batch
Weight Type	Weight	Weight	*Transposed* Weight

Figure 7.4.1: Deep reinforcement learning (DRL) process and data type.

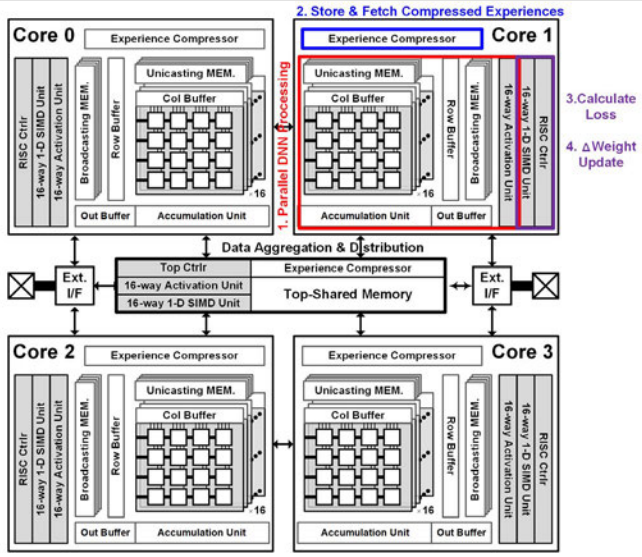


Figure 7.4.2: Overall architecture.

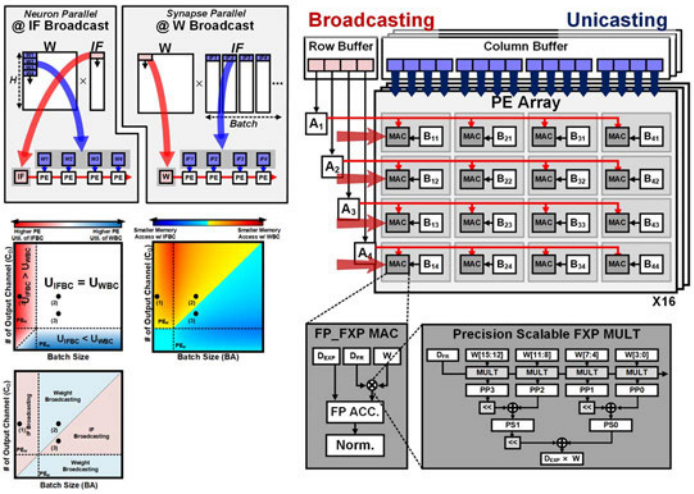


Figure 7.4.3: Transposable PE array architecture.

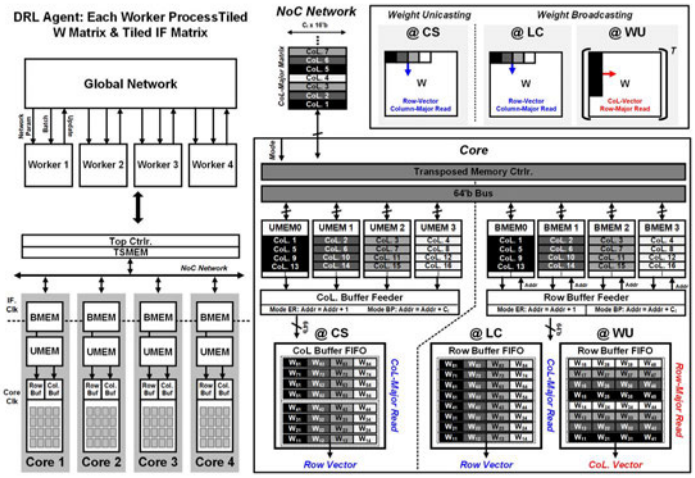
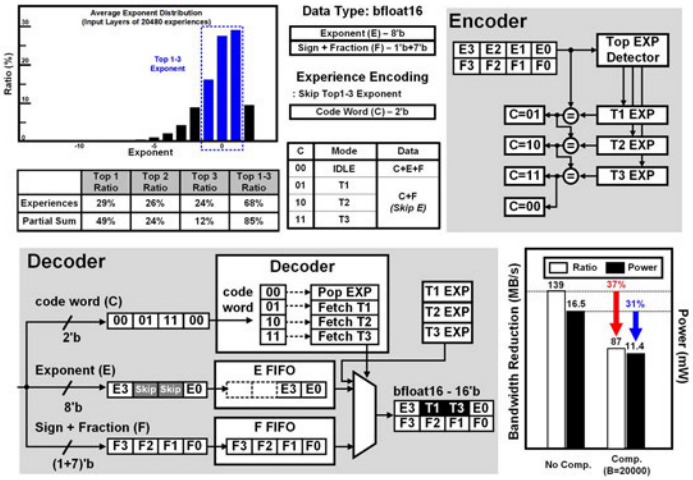


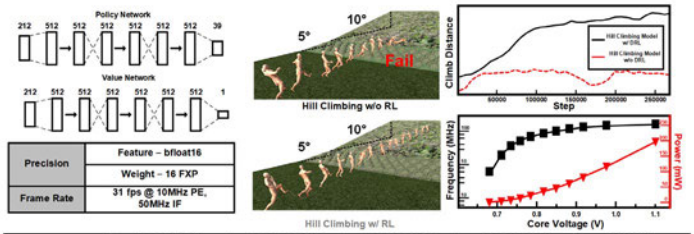
Figure 7.4.4: Transposed memory controller and buffer feeding logic.



	Top 1 Ratio	Top 2 Ratio	Top 3 Ratio	Top 1-3 Ratio
Experiences	29%	26%	24%	68%
Partial Sum	49%	24%	12%	85%

C	Mode	Data
00	IDLE	C+E+F
01	T1	C+F (Skip E)
10	T2	
11	T3	

Figure 7.4.5: Activation compressing using experience encoder and decoder.

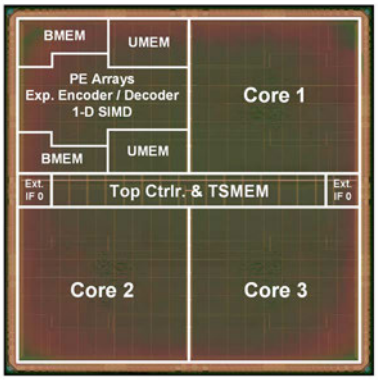


Precision	Feature – bfloat16
	Weight – 16 FXP
Frame Rate	31 fps @ 10MHz PE, 50MHz IF

	ISSCC 2018 [4]	ISSCC 2018 [3]	S. VLSI 2018 [5]	S. VLSI 2018 [6]	This Work
Inference	CNN/RNN/FC	CNN/RNN/FC	CNN/FC	CNN/RNN/FC	CNN/RNN/FC
Learning	-	-	FC Tuning	CNN/RNN/FC	CNN/RNN/FC
Precision	FXP 1-4	Feature: FXP 16 Weight: FXP 1-16	FXP 8	FP16, binary, ternary	Feature: bfloat16 Weight: FXP 4,8,16
Process (nm)	40	65	65	14	65
Area (mm^2)	121.55	16	12	9	16
Frequency (MHz)	75-330	5-200	20-200	750-1500	5-200
Data Compression Scheme	X	X	X	X	O
Energy Efficiency	-	3.06 TOPS/W (16'b) 11.6 TOPS/W (4'b) 50.6 TOPS/W (1'b)	0.411-62.1 TOPS/W	-	2.16 TFLOPS/W (16'b)
Peak Performance	1.96 TOPS (4'b) 7.49 TOPS (2'b)	345.6 GOPS (16'b) 1382 GOPS (4'b) 7372 GOPS (1'b)	**102-5638 GOPS	1500 GFLOPS	204 GFLOPS (16'b)
Normalized Area Efficiency	*6.11 GOPS/mm^2 (4'b FXP)	21.6 GOPS/mm^2 (16'b FXP)	8.5 - 469.83 GOPS/mm^2 (8'b FXP)	*7.73 GFLOPS/mm^2	12.75 GFLOPS/mm^2

* Normalized to 65nm Process
**Maximum Performance with High Sparsity Condition

Figure 7.4.6: Measurement results and comparison table.



Process	65nm 1P8M Logic CMOS
Area	4mm × 4mm
SRAM	448 KB
Supply	0.67V – 1.1V
Power	196 mW @ 200MHz, 1.1V
	2.4 mW @ 10MHz, 0.67V
Precision	Feature – bfloat16
	Weight – 16/8/4'b FXP
Peak Performance	204 GOPS @ 16b Weight

Figure 7.4.7: Chip micrograph and summary.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

7.5 A 65nm 0.39-to-140.3TOPS/W 1-to-12b Unified Neural-Network Processor Using Block-Circulant-Enabled Transpose-Domain Acceleration with 8.1× Higher TOPS/mm² and 6T HBST-TRAM-Based 2D Data-Reuse Architecture

Jinshan Yue[1], Ruoyang Liu[1], Wenyu Sun[1], Zhe Yuan[1], Zhibo Wang[1], Yung-Ning Tu[2], Yi-Ju Chen[2], Ao Ren[3], Yanzhi Wang[3], Meng-Fan Chang[2], Xueqing Li[1], Huazhong Yang[1], Yongpan Liu[1]

[1]Tsinghua University, Beijing, China
[2]National Tsing Hua University, Hsinchu, Taiwan
[3]Northeastern University, Boston, MA

Energy-efficient neural-network (NN) processors have been proposed for battery-powered deep-learning applications, where convolutional (CNN), fully-connected (FC) and recurrent NNs (RNN) are three major workloads. To support all of them, previous solutions [1-3] use either area-inefficient heterogeneous architectures, including CNN and RNN cores, or an energy-inefficient reconfigurable architecture. A block-circulant algorithm [4] can unify CNN/FC/RNN workloads with transpose-domain acceleration, as shown in Fig. 7.5.1. Once NN weights are trained using the block-circulant pattern, all workloads are transformed into consistent matrix-vector multiplications (MVM), which can potentially achieve 8-to-128× storage savings and $O(n^2)$-to-$O(n\log(n))$ computation complexity reduction.

Several challenges exist to achieve such potential improvements. 1) A high-performance and low-power fast Fourier transform (FFT) module supporting variable bit precision is required; 2) a more efficient data-reuse architecture is needed because the original multiplication-and-accumulation (MAC) operations after the FFT induce non-trivial on/off-chip memory access; 3) the data-reuse architecture requires a more power- and area-efficient transpose SRAM (TRAM) compared with previous TRAMs [5-6] with large power/area overhead.

We introduce a 65nm 0.39-to-140.3TOPS/W unified NN processor, STICKER-T, with superior (8.1× higher) TOPS/mm². The major contributions include: 1) A unified block-circulant chip architecture supporting CNN/FC/RNN transpose-domain acceleration; 2) a 1-to-12b 8-to-128-point high-performance low-power global-parallel local-serial FFT module with real activation feature enhancement; 3) a 1-to-12b 2D data-reusable MAC array utilizing a hierarchical-bitline-switching (HBST) 6T TRAM with 2.14× energy and 4.06× array-size savings.

Figure 7.5.2 shows the unified block-circulant-enabled transpose-domain acceleration flow and overall processor architecture. NN weight matrices are pre-trained into block-circulant patterns, where a *NxN* block is represented with only *N* parameters. By adopting an 8-128 block size, the block-circulant pattern provides 8×-128× storage reduction. Once the CNN/FC/RNN are represented with circulant matrix-vector multiplications (MVM), they are equivalent to a time-domain convolution and can be accelerated by a unified processor in the transpose domain. The processor consists a high-performance low-power FFT module, a 2D data-reusable MAC array, and a power/area efficient HBST 6T TRAM. The FFT module is fed by a 16KB activation SRAM and a shared-twiddle-factor (TF) buffer. The TRAM transposes activations from the FFT module and provides them to a 16×16 MAC array with weights from a 64KB SRAM.

Figure 7.5.3 depicts a high-performance low-power global-parallel local-serial FFT module design with real activation feature enhancement. The global-parallel FFT module consists 7 stages of butterfly units (BFU) and twiddle-factor production units (TFP), and each stage consists 32 parallel BFUs and TFPs, which simultaneously support variable-bitwidth FFT from 8×8-point to 1×64-point operations. To avoid a large area overhead, each local TFP is implemented in an area-efficient bit-serial way. The TFP receives 12b data from the upper stage BFU, and executes bit-serial multiplication by shift-accumulation with 1b of TF in each cycle. Results from stages 3-7 are dynamically selected to support 8-to-128-point FFT under different circulant block sizes. The global-parallel local-serial FFT design shows 5.05×/4.36× power/area reduction compared with a parallel FFT design, and 5.3-to-32× performance gains compared with a serial FFT design.

Furthermore, the FFT module adopts three low power techniques: 1) 62 trivial TFs are shared and bit-serially fetched by all TFPs, instead of using one TF register for each TFP; 2) TFPs for non-trivial TFs, e^{j0}, $e^{j\pi/2}$, are removed and implemented as direct-connect or a real/imaginary (R./I.) part swap, which saves 28% of the total TFPs. 3) In real activation feature enhancement, two separate activation vectors are coupled as R./I. parts of a complex vector, and decoupled after FFT operations for 2× performance gain, because activations are real numbers. Considering all these factors, the FFT module further saves power/area by 2.95×/2.93× without performance loss.

Figure 7.5.4 shows the 6T HBST-TRAM-based 2D data-reuse MAC array with bit-serial MAC units. Activations are horizontally written into SRAM after FFT operations, but require vertical reading for better data reuse. STICKER-T adopts a 6T TRAM to enable activation and batch-level weight reuse, by 1 data fetch for 16 multiplications. Compared with a conventional SRAM architecture, our approach requires 16×/12.9× fewer on/off-chip memory accesses, and reduces area by 1.04×. Furthermore, bit-serial MAC units are designed to flexibly support 1-to-12b precision. Activations bit-serially pass a serial-to-parallel (STP) module, and the STP module generates *k*-bit data by *k* cycles of left-shift with sign extension for unused bits. Weights are bit-serially multiplied with 12b activations by shift-accumulation. Meanwhile, the STP prepares next *k* bit activation. MAC results are locally accumulated and stored in 24b registers to reduce SRAM access. In total, the 6T TRAM supported 2D data reuse MAC array with bit-serial MAC units achieves 1.26× power savings (not including the potentially significant reduction in off-chip memory accesses owing to smaller circulant weight storage requirements).

Figure 7.5.5 presents the power/area-efficient 32Kb 6T HBST TRAM, which supports regular mode (RGM) and transpose mode (TPM) read/write operations. Beside typical SRAM sub-circuits and a cell-array with a hierarchical bitline (BL) structure for vertical access in RGM, this macro includes HBST cells, vertical WL drivers (VWLDRV), and a vertical IO (VIO) for horizontal access in TPM. Each local cell array has eight 6T cells with a HBST cell. The HBST cell employs four transistors to connect the local bitline pair (LBL and LBLB) to either horizontal global BL pair (HBL and HBLB) or vertical global BL pair (VBL and VBLB). In RGM, HWL=1, LBL/LBLB is connected to VBL/VBLB and data can be accessed by the HIO. In TPM, the selected VWL is activated to connect the selected LBL/LBLB to HBL/HBLB and the VIO is enabled. Thanks to fully differential sensing and reduced BL loading, this work shows 2.14×/1.29× energy savings, and 1.11×/2.09× delay reduction compared with previous works [5-6]. Apart from the 6T cell design, this work uses a foundry-compact 6T SRAM cell to achieve compact macro area, which leads to 2.05×/4.06× array-size savings compared with 7T [5] and 8T [6] cells using a large-area logic-rule.

STICKER-T is fabricated in a 65nm CMOS technology and measurement results are shown in Fig. 7.5.6. Three pre-trained block-circulant CNN/FC/RNN models are executed by the unified processor, operating at 0.54-to-1.15V with 25-to-200MHz frequency. The average energy efficiency range is 3.76-to-14.4TOPS/W with various block sizes and bit precisions. Energy efficiency is inversely proportional to bit-precision due to the bit-serial design. Larger block size leads to higher energy efficiency due to fewer FFT and MAC operations. The peak energy-efficiency is 140.3TOPS/W under 3b FFT, 1b MAC precision, block-size 128 at (0.54V, 25MHz). Thanks to block-circulant enabled transpose-domain acceleration and efficient circuit implementations, the comparison table shows that STICKER-T achieves the highest (8.1× higher) TOPS/mm² and 4.2× higher energy efficiency at 4b precision compared with a state-of-the-art unified NN processor [4]. Figure 7.5.7 shows the summary table and die photo with 2.5×3.0mm² area.

Acknowledgements:
Supported in part by NSFC under Grant #61674094, #61874066 and #61720106013; and in part by Beijing Innovation Center for Future Chip.

References:
[1] Z. Yuan, et al., "STICKER: A 0.41-62.1 TOPS/W 8bit Neural Network Processor with Multi-Sparsity Compatible Convolution Arrays and Online Tuning Acceleration for Fully Connected Layers," *IEEE Symp. VLSI Circuits*, pp. 33-34, 2018.
[2] D. Shin, et al., "DNPU: An 8.1 TOPS/W Reconfigurable CNN-RNN Processor for General-Purpose Deep Neural Networks," *ISSCC,* pp. 240-241, 2017.
[3] J. Lee, et al., "UNPU: A 50.6 tops/w Unified Deep Neural Network Accelerator with 1b-to-16b Fully-Variable Weight Bit-Precision," *ISSCC,* pp. 218-220, 2018.
[4] C. Ding, et al., "CirCNN: *accelerating and compressing deep neural networks using block-circulant weight matrices*," *IEEE/ACM MICRO*, pp. 395-408, 2017.
[5] K. Bong, et al., "A 0.62 mW Ultra-Low-Power Convolutional-Neural-Network Face-Recognition Processor and a CIS Integrated With Always-On Haar-Like Face Detector," *ISSCC*, pp. 248-249, 2017.
[6] J. Seo, et al., "A 45nm CMOS Neuromorphic Chip With a Scalable Architecture For Learning In Networks Of Spiking Neurons," *CICC*, pp.1-4, 2011.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

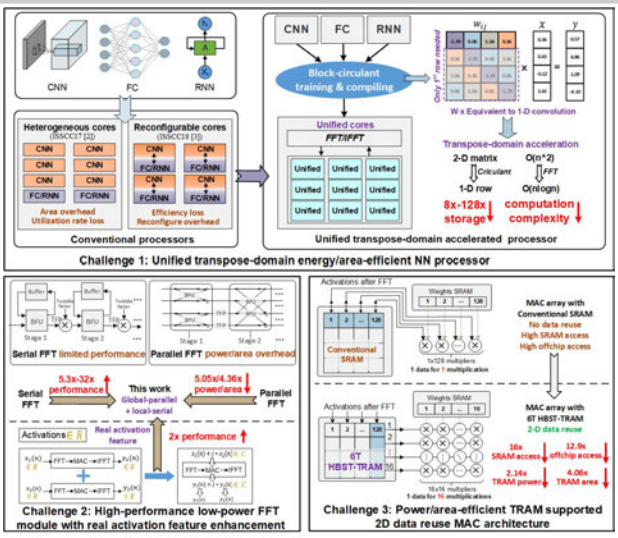


Figure 7.5.1: Challenges for unified transpose-domain block-circulant accelerated NN processor.

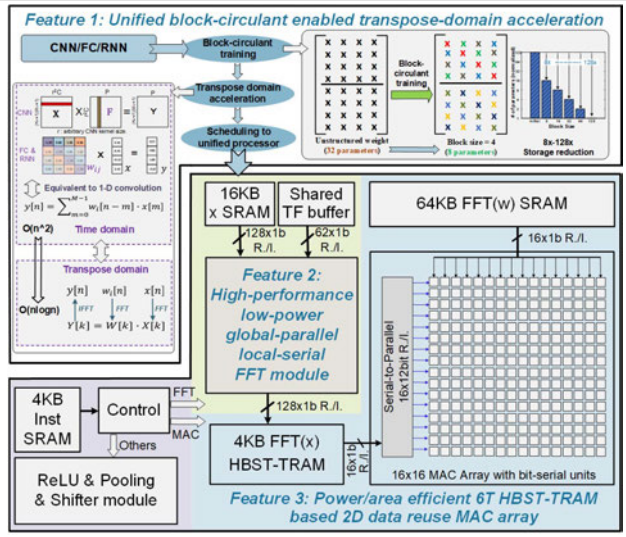


Figure 7.5.2: Unified block-circulant-enabled transpose-domain acceleration flow and overall processor architecture.

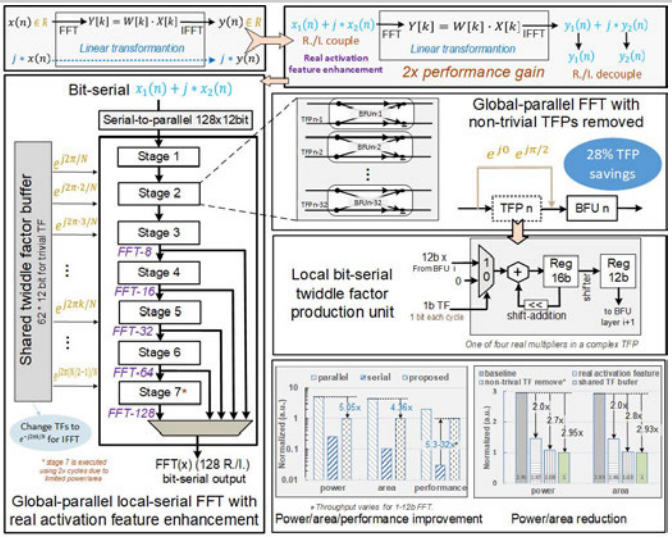


Figure 7.5.3: High-performance low-power global-parallel local-serial FFT module with real activation feature enhancement.

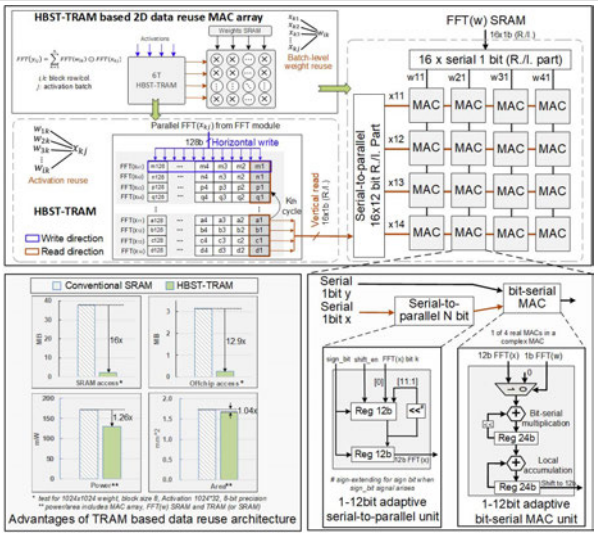


Figure 7.5.4: 6T HBST-TRAM-based 2D data-reuse MAC array with bit-serial MAC units.

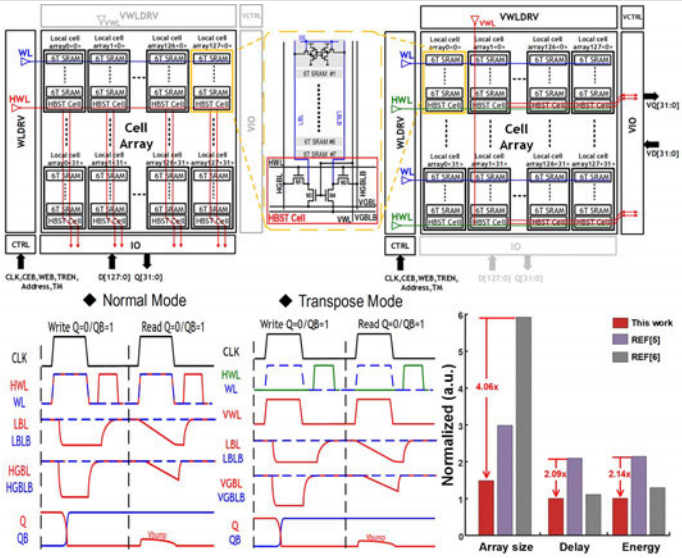
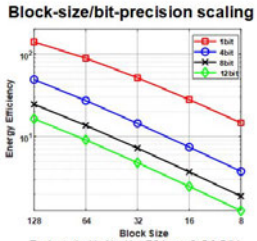


Figure 7.5.5: Power/area efficient 32Kb 6T hierarchical-bitline-switching TRAM.

CNN/FC/RNN models with varied block-size/precision

Network	Tiny YOLO	LSTM	FC
Dataset	DataOJI	TIMIT	MNIST
Layer type	CNN	RNN (LSTM)	FC
Block-size	32	32	128
Accuracy loss	3%	1.93%	1.50%
# of circulant weights	494K	101K	17.4K
Equivalent operations	6.95GOP	6.50MOP	1.11MOP
FFT bit-precision	8bit	8bit	6bit
MAC bit-precision	5bit	4bit	2bit
Execution time (1 in 32 batch)	0.14s	33.8us	6.44us
Energy efficiency	3.76TOPS/W	14.4TOPS/W	12.9TOPS/W



Comparison with state-of-the-art NN processors

	ISSCC17-DNPU[2]	ISSCC18-UNPU[3]	VLSI18-STICKER[1]	This work
Technology	65nm	65nm LP	65nm GP	65nm GP
Area[mm^2]	16	16	12	7.5
Support NN type	CNN+FC+RNN	CNN+FC+RNN	CNN+FC	CNN+FC+RNN
Weight precision	4-7b(Quantized)	1-16b	8b	1-12b
Activation precision	4,8,16b	16b	8b	1-12b
Performance (TOPS)	1.08	7.37	5.6	14.9 1)
Power(mW)	34.6-279	3.2-297	20.5-248.4	13.3-339
Area efficiency (TOPS/mm2)	0.068(4b) **0.79x**	0.086(4b) **1x**	0.47(Conv)	**0.70(4b) 8.1x** 1)
	-	0.46(1b) **1x**	-	**1.99(1b) 4.32x** 1)
Energy efficiency (TOPS/W)	8.1(4b) **0.70x**	11.6(4b) **1x**	62.1 (Conv)	**49.1(4b) 4.2x** 2)
	-	50.6(1b) **1x**	-	**140.3(1b) 2.77x** 2)

One operation (OP) represents one real multiplication or one real addition.
1) @1.15V,200MHz; 2) @0.54V,25MHz. AlexNet FC layer 2, block size 128.

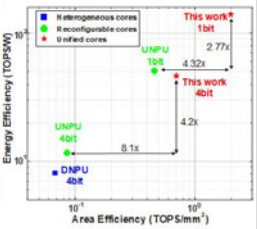
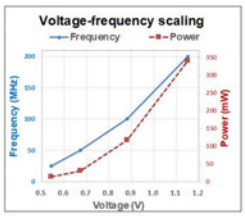


Figure 7.5.6: Measurement results and comparison table.



Technology	TSMC 65nm GP
Chip Area	3.0mmx2.5mm
Core Area	2.46mmx1.96mm
Support NN type	CNN+FC+RNN
Weight Bit-precision	1-12 bit
Activition Bit-precision	1-12 bit
Total SRAM	100kB
HBST-TRAM	4kB
Supply Voltage	0.54-1.15V
Frequency	25-200MHz
Power	13.3-339.2mW
Performance	0.13-14.9TOPS
Energy Efficiency 1)	0.39-140.3TOPS/W

1) minimum @1.15V,200MHz,12bit, block size=8. Maximum @0.54V,25MHz ,3bit FFT, 1bit MAC, block size=128, on AlexNet FC layer 2.

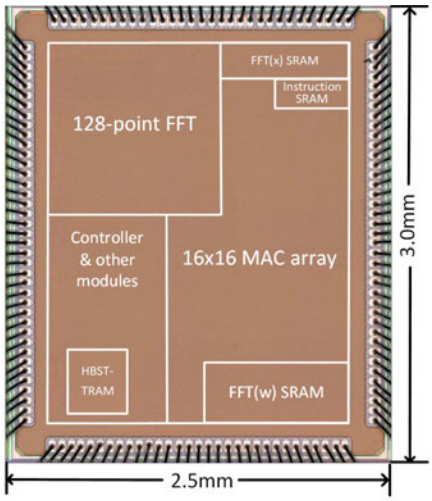


Figure 7.5.7: Die photo and metrics.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

7.6 A 65nm 236.5nJ/Classification Neuromorphic Processor with 7.5% Energy Overhead On-Chip Learning Using Direct Spike-Only Feedback

Jeongwoo Park, Juyun Lee, Dongsuk Jeon

Seoul National University, Seoul, Korea

Advances in neural network and machine learning algorithms have sparked a wide array of research in specialized hardware, ranging from high-performance convolutional neural network (CNN) accelerators to energy-efficient deep-neural network (DNN) edge computing systems [1]. While most studies have focused on designing inference engines, recent works have shown that on-chip training could serve practical purposes such as compensating for process variations of in-memory computing [2] or adapting to changing environments in real time [3]. However, these successes were limited to relatively simple tasks mainly due to the large energy overhead of the training process. These problems arise primarily from the high-precision arithmetic and memory required for error propagation and weight updates, in contrast to error-tolerant inference operation; the capacity requirements of a learning system are significantly higher than those of an inference system [4].

To overcome such challenges, this paper presents an on-chip trainable neuromorphic processor with the advantages of spiking neural networks (SNNs) and DNNs combined (Fig. 7.6.1). SNNs enable highly energy-efficient computing, but suffer from relatively low accuracy of bio-plausible learning rules [5], whereas stochastic gradient descent (SGD) based DNNs provide good accuracy at the expense of high-precision computation during training. To reduce training power drastically, while maintaining high algorithm accuracy, we propose three design approaches: (1) high accuracy spike-based learning suitable for hardware implementation, (2) exploiting algorithm properties to look ahead and skip updates when they are redundant, and (3) out-of-order weight updates for training overhead reduction. Using these techniques, the processor is trained on the handwritten digits (MNIST) dataset and achieves 97.83% classification accuracy, while only consuming 254.3nJ/image during training, demonstrating a mere 7.5% energy increase over inference, which consumes 236.5nJ/image.

The overall architecture of the neuromorphic system (Fig. 7.6.2) consists of a spike converter (as the input layer), two hidden layers, an output layer with 10 target neurons, and control logic. Each hidden layer consists of 200 neurons, two buffer memories totaling 2.2 or 4.8Kb, and an update-value calculation unit with a sigmoid function implemented as LUT. In the hidden neuron cluster, 5 neurons share a 70×200b SRAM for weight storage in order to improve area efficiency. Each hidden neuron accumulates three values: previous layer spikes, output layer spikes, and target spikes. The accumulated value of previous layer spikes is used to determine the firing probability of that neuron, and accumulated values from output layer spikes and target spikes are used to calculate weight update (ΔW) values. After determining the firing probability, the spike converter sends spikes sequentially to the next layer. If the processor is in training, spikes are determined stochastically using pseudo-random threshold values generated by a 10b LFSR. For inference, a fixed threshold of 0.5 is used to generate all spikes.

Inspired by the segregated dendrites (SD) algorithm [6], we propose a hardware-oriented spike-based learning algorithm that simultaneously enables efficient hardware implementation and boosts learning performance. Instead of using the complex neuron architecture in typical SNN algorithms that employ post-synaptic potential values and inter-time frame neuron potential dynamics from differential equations, this neuron architecture uses direct spike values and simple accumulation. Moreover, only a single time frame is applied, which employs a single spike value for calculation rather than averaging over multiple sampled spikes. This neuron architecture improves MNIST recognition accuracy from 96.1 to 97.7%, and the total number of operations per image is reduced by 30%. Applying single-shot training further improves the accuracy to 98.1% and reduces the operations by 96%, as only a single sample is required for all calculations (Fig. 7.6.3). Single-shot modification also helps reduce training overhead by reducing high-precision multiplications and buffer memory size. In back-propagation-based learning, SD or multiple-shot modified SD, previous layer activation values are multiplied by the intermediate ΔW values in each neuron. However, previous layer activation values are single-bit spikes in single-shot modified SD, replacing high-precision multiplications with multiplication by one or zero. The other important property of our modification is that the total number of bits required for feedback is merely 20b: 10b each for output and target neuron spikes, greatly reducing computation and global interconnects. While the single-shot modified SD algorithm was tested on a relatively small-scale problem of classifying handwritten digits, the learning rule can be directly adopted in general fully-connected networks with different network structures and hence can be extended to other applications. For example, the algorithm could serve a key role in fine tuning large networks (typically based on CNN) through transfer learning and training fully-connected layers in the last stages.

As the ΔW values rely on the difference in stochastically sampled output spikes and the teaching signal represented by one-hot-encoded spikes, it is observed that ΔW=0 for all neurons when the output and target spikes exactly match. In order to exploit this observation, we implemented an update-skipping mechanism illustrated in Fig. 7.6.4. A spike pattern analyzer deserializes the output and target spikes, and asserts a skip update signal when they match. This skip update signal blocks inputs to ΔW calculation units and update paths in order to prevent energy dissipation from redundant logic switching, and deasserts the write signal on SRAM and buffer register files. As shown in Fig. 7.6.4, the skip update rate increases from 3.8 to 81.7% in the first epoch as the processor learns to represent images with correct output spikes, and the overhead in training energy decreases from the peak at 25.6% to 8.2% in the last 1,000 images of the first epoch. Averaged over 100 epochs, the skip update rate stands at 94.5% and the training energy overhead is measured as 7.5%.

Figure 7.6.5 describes the out-of-order weight update scheme. When training with in-order updates, the next image cannot be processed until the ΔW from the previous image has been added to the weight matrix. This inevitably leads to idle states in some modules, degrading hardware efficiency. By contrast, out-of-order updates do not wait for the update to take effect and instead directly proceed to process the next image, which allows the processor to remove idle states and reduce training time overhead by 98%. This out-of-order technique increases energy efficiency in a twofold manner: (1) by allowing the processor to reduce training time, the hardware is utilized to the full capacity, and (2) by reusing weight values that are read out from SRAM to process the forward pass, an additional SRAM read operation is not required in the update operation. While these out-of-order updates are beneficial for hardware utilization, simulations show that the statistical performance of classifications do not degrade, due to the inherent robustness of the proposed learning algorithm.

The neuromorphic processor was fabricated in 65nm LP CMOS. It consumes 23.6mW at 0.8V, 20MHz while processing 100K images/s during inference. During training, it consumes 23.1mW at 0.8V, 20MHz with the throughput of 94.3K images/s. Measurements in Fig. 7.6.6 show that the processor consumes only 7.5% more energy for training, providing 7.3× lower training overhead than prior works that offer the on-chip training feature. In addition, the processor achieves 97.83% average classification accuracy (peaking at 98.06%) on the MNIST test dataset with 34% (72% if scaled to technology) lower inference energy than recent inference-only design with similar classification performance. The die photograph and chip details are in Fig. 7.6.7.

Acknowledgements:
This work was supported by the National Research Foundation of Korea (Grant No. NRF-2016R1C1B2016072), the KIST Institutional Program (Program No. 2E27810-18-P034), and the IC Design Education Center (IDEC).

References:
[1] P. Whatmough, et al., "A 28nm SoC with a 1.2GHz 568nJ/prediction Sparse Deep-Neural-Network Engine with >0.1 Timing Error Rate Tolerance for IoT Applications," *ISSCC,* pp. 242-243, 2017.
[2] S. Gonugondla, et al., "A 42pJ/Decision 3.12TOPS/W Roust In-Memory Machine Learning Classifier with On-Chip Training," *ISSCC,* pp. 490-491, 2018.
[3] A. Amravati, et al., "A 55nm Time-Domain Mixed-Signal Neuromorphic Accelerator with Stochastic Synapses and Embedded Reinforcement Learning for Autonomous Micro-Robots," *ISSCC*, pp. 124-125, 2018.
[4] C. Tsai, et al., "A 41.3/26,7pJ per Neuron Weight RBM Processor Supporting On-Chip Learning/Inference for IoT Applications," *IEEE JSSC*, vol. 52, no. 10, pp. 2601-2612, Oct. 2017.
[5] F. Buhler, et al., "A 3.43TOPS/W 48.9pJ/Pixel 50.1nJ/Classification 512 Analog Neuron Sparse Coding Neural Network with On-Chip Learning and Classification in 40nm CMOS," *IEEE Symp. VLSI Circuits*, 2017.
[6] J. Guerguiev, et al., "Towards Deep Learning with Segregated Dendrites," *ELife*, vol. 6, Dec. 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

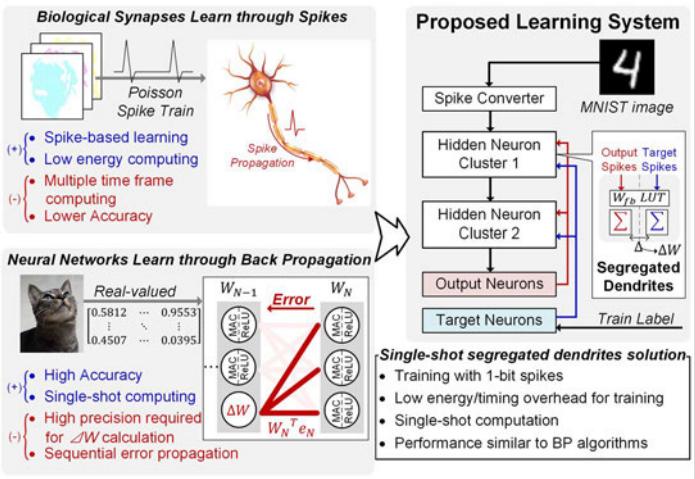


Figure 7.6.1: Overview of learning based on biological models, compared to back propagation. This design integrates the advantages of the two learning methods.

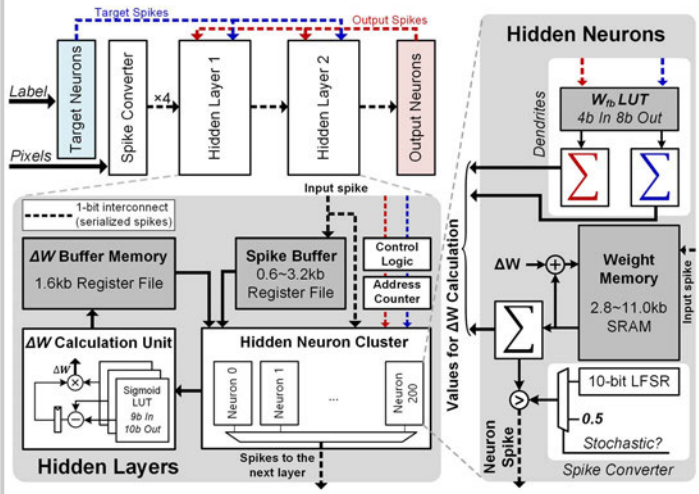
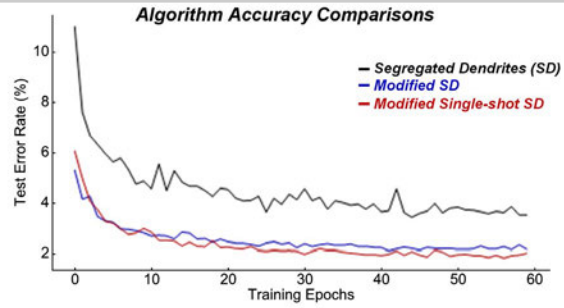


Figure 7.6.2: Architecture of the neuromorphic processor. One-bit spikes are used for learning, lowering computational complexity and reducing data transmission.



	Clock Cycles /Image	MNIST Accuracy	Training Buffer Memory Size	OP/Image
Segregated Dendrites	14,000	96.1%	47.2Kb	29,379KOP
Modified SD	10,010	97.7%	38.4Kb	20,741KOP
Single-shot Modified SD	**212**	**98.1%**	**7.6Kb**	**804KOP**

Figure 7.6.3: Analysis of classification performance and hardware implementation cost across baseline algorithm (SD), modified SD, and single-shot modified SD.

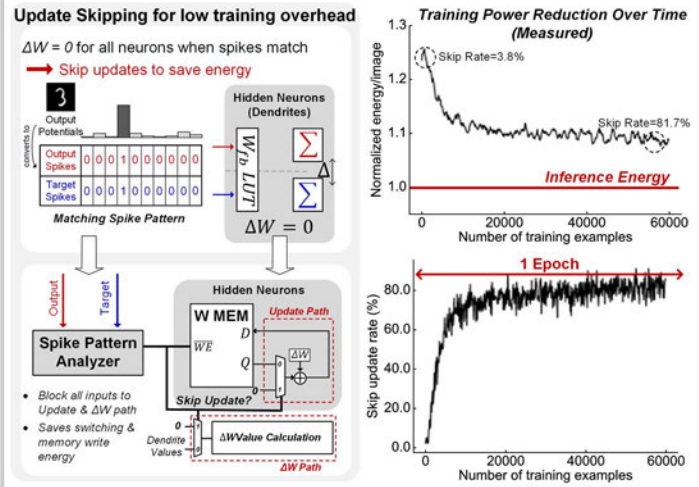


Figure 7.6.4: Update skipping mechanism and hardware implementation along with measured effects.

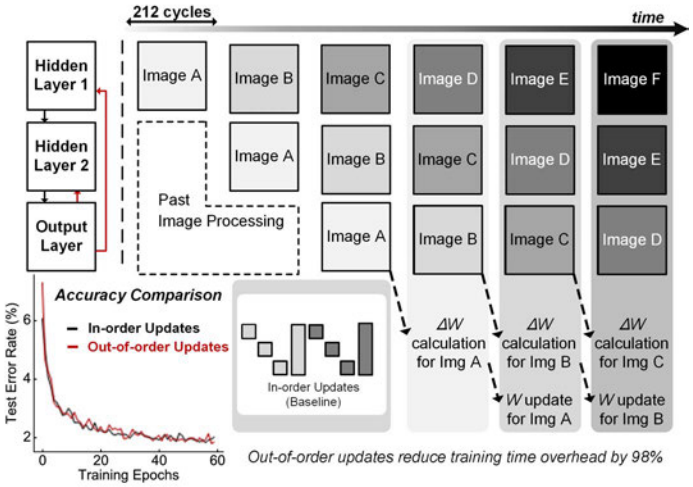
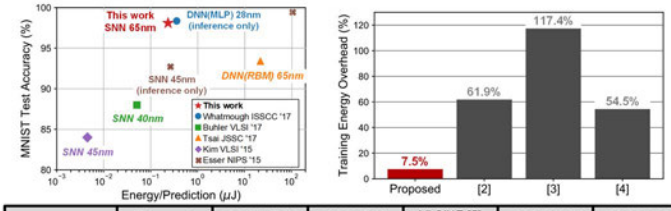


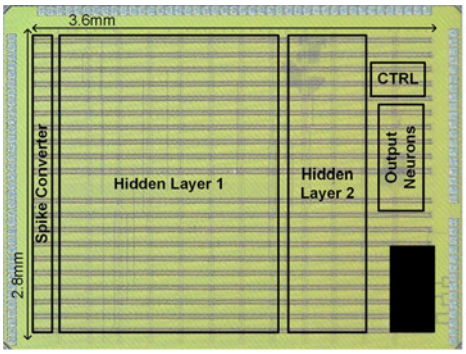
Figure 7.6.5: Out-of-order weight update scheme. While the scheme reduces training time overhead by 98%, performance does not degrade.



	This work	ISSCC'17 [1]	JSSC'17 [4]	VLSI'17 [5]	ISSCC'18 [2]	ISSCC'18 [3]
Technology	**65nm**	28nm	65nm	40nm	65nm	55nm
On-chip Training	**Yes**	No	Yes	Yes	Yes	Yes
MNIST Accuracy	**97.83%**	98.36%	93.4%	88%	N/A	N/A
Algorithm	**SNN**	DNN (MLP)	DNN (RBM)	SNN	SVM	SVM RL
Prediction Energy	**236.5nJ/Image**	360nJ/Image	21.9μJ/Image	50.1nJ/Image	N/A	N/A
Training Overhead	**7.5%***	N/A	54.5%	N/A	61.9%	117.4%
Throughput	**100K img/sec**	15K img/sec	14K img/sec	2.2M img/sec	N/A	N/A
Power	**23.6mW**	20.3mW	310.0mW	87.0mW	6.8mW	0.7mW
Scaled Efficiency**	**7.89TOPS/W**	1.23TOPS/W	3.37TOPS/W	4.90TOPS/W	7.24TOPS/W	6.13TOPS/W

* Average of 100 training epochs
**Scaled Efficiency = Reported Efficiency * (Technology/28nm)

Figure 7.6.6: Performance summary and comparison with prior works.



Technology	65nm LP
V_{dd}	0.8V
Power	23.1mW (Train) 23.6mW (Infer)
Frequency	20MHz
Target Operation	MNIST Recognition
Efficiency	3.42TOPS/W
Core Area	2.8×3.6mm^2
Throughput	100K img/s

Figure 7.6.7: Die photo and summary table.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

7.7 LNPU: A 25.3TFLOPS/W Sparse Deep-Neural-Network Learning Processor with Fine-Grained Mixed Precision of FP8-FP16

Jinsu Lee, Juhyoung Lee, Donghyeon Han, Jinmook Lee, Gwangtae Park, Hoi-Jun Yoo

KAIST, Daejeon, Korea

Recently, deep neural network (DNN) hardware accelerators have been reported for energy-efficient deep learning (DL) acceleration [1-6]. Most prior DNN inference accelerators are trained in the cloud using public datasets; parameters are then downloaded to implement AI [1-5]. However, local DNN learning with domain-specific and private data is required meet various user preferences on edge or mobile devices. Since edge and mobile devices contain only limited computation capability with battery power, an energy-efficient DNN learning processor is necessary. Only [6] supported on-chip DNN learning, but it was not energy-efficient, as it did not utilize sparsity which represents 37%-61% of the inputs for various CNNs, such as VGG16, AlexNet and ResNet-18, as shown in Fig. 7.7.1. Although [3-5] utilized the sparsity, they only considered the inference phase with inter-channel accumulation in Fig. 7.7.1, and did not support intra-channel accumulation for the weight-gradient generation (WG) step of the learning phase. Also, [6] adopted FP16, but it was not energy optimal because FP8 is enough for many input operands with 4× less energy than FP16.

We present an energy-efficient on-chip learning accelerator. Its data precision is optimized, while maintaining training accuracy with fine-grained mixed precision (FGMP) of FP8-FP16. The partial use of narrower bitwidths reduces external memory accesses (EMA) and enhances throughput. In addition, sparsity is exploited with intra-channel accumulation, as well as inter-channel accumulation, to support the 3 DNN learning steps of Fig. 7.7.1 with higher throughput to enhance energy efficiency. Also, the input load balancer (ILB) is integrated to improve PE utilization when faced with workload imbalance caused by irregular sparsity.

Figure 7.7.2 shows the proposed deep-learning neural processing unit (LNPU) composed of 16 sparse DL cores, a central core (CC), a SIMD core, and a top RISC controller. Each sparse DL core has an ILB and 4 PE-lines with 48 PEs that each have an FP8-FP16 configurable MAC and a 4×16b local register file (LRF). The CC aggregates partial-sums (Psum) of each core during feed-forward (FF) and back-propagation (BP) and feeds activation gradients into each core during WG. The CC converts FP8 data into FP16 or vice versa, and performs zero compression and decompression. The SIMD core calculates non-linear functions and batch-normalization.

Figure 7.7.3 illustrates FGMP, selecting the optimal precision as either FP8 or FP16 adaptively for every input operand. The input operands, for example activations or their gradients, are divided into an FP8 group and an FP16 group according to their exponents. Each group is loaded or stored from/to memory and computed with a run-length-compressed format using 2b and 4b skip-indices, respectively. The data near the central region is represented in FP8 to include more data in the FP8 group for lower EMA and higher computation efficiency. The data beyond FP8 is grouped in FP16, and the ratio between FP8 and FP16 is controlled by a threshold, which is adjusted during DNN learning. In addition, FP8 and FP16 groups are adaptively shifted layer by layer and step by step throughout DNN learning to minimize overflow and underflow. On each core, the FP8 group is processed first, and then the FP16 group is processed. MAC units in the sparse DL core can be configured into either 2 of FP8 or an FP16. To minimize area consumption, a multiplier, an aligner and an adder in the MAC are shared by both FP16 and FP8 with only 13% area overhead compared to FP16. Moreover, zeroes arising from FGMP do not impose any latency overhead, due to the sparse DL core. The training of ResNet-18 maintains its accuracy with 90% of computations in the FP8 format, and in this case, EMA is reduced by 38.9%.

Figure 7.7.4 illustrates the datapath for each learning step, and the sparse operation on a PE line. The compressed data on a tile (x,1,ch) is broadcasted to PE lines along with skip indices through the ILB. The other operands are fed into each PE from input registers (I0~I2) through a local router, which enables three neighbor PEs in the same PE line to share operands. The local router is controlled by the skip index to feed proper operands into PEs. In the case of FF and BP, the weight is stored in the column buffer (C-buf) and moves to I0~I2 of each PE line. The PEs perform 1D convolution and the Psums are accumulated in A0~A2 using weights in I0~I2 and activations (FF) or activation gradients (BP) with zero-skipping. Then, the result of Psum in A0~A2 is stored into the LRF (R0~R11) of the corresponding address. The Psum stored in LRF (R0~R11) is spatially aggregated between the other PE lines by the local aggregation unit and the result moves to CC. In the case of WG, the activation gradient on tile (x,1,ch) moves from the CC to the LRF (R0~R11) of the corresponding address. Then, the activation gradients are fed from the LRF to the proper PE through the local router to generate weight gradients. The weight gradients are aggregated temporally on accumulation registers (A0~A2) of each PE, and then, the results are stored to C-buf. With the help of the sparse DL core, ResNet-18 learning throughput is improved by 1.82×.

Figure 7.7.5 shows the ILB. Eight compressed operands are fetched sequentially from DMEM into the ILB according to the address from the address generator units. The skip-index decoder calculates the C-buf address (A_{Cbuf}) to fetch proper weights, and a validation mask to selectively push data (A_{Cbuf}, skip-index, and D_{IN}) into a FIFO. At run-time, the ILB balances input data evenly to each PE accounting for irregular sparsity and various ratios between the FP8 and FP16 group sizes. For example, the PE line2, which does not get any operand from the ILB in the current cycle, takes input data from PE line3 with the help of validation masks. Due to the ILB, PE utilization is increased 2-26% for 10-to-90% input sparsity.

Figure 7.7.6 shows the measurement results. The LNPU can operate at 0.78-to-1.1V supply with a maximum 200MHz clock frequency. The power consumption at 0.78V and 1.1V are 43.1mW and 367mW, respectively. The energy efficiency is 3.48TFLOPS/W (FP8) for 0.0% sparsity and 25.3TFLOPS/W (FP8) for 90% sparsity. The accelerator achieves 5.84TFLOPS/W with 8.2fps and 2.74TFLOPS/W with 32.8fps on the VGG-16 conv-layer inference benchmark at FP8 precision. It supports not only DNN inference, but also learning for various DNN structures. Owing to the FGMP and sparse DL core, energy efficiency is improved 2.08× compared to dense FP16 operation without any degradation in learning accuracy for ResNet-18. The energy efficiency is 4.4× higher than the NVIDIA V100 GPU and its normalized peak performance is 2.4× higher than a previous DNN learning processor [6].

The LNPU is fabricated using 65nm CMOS technology and occupies 16mm^2 die area, as shown in Fig. 7.7.7. The DNN processor handles sparsity and offers fine-grained mixed precision for learning and shows high normalized peak performance compared to the state-of-the-art DNN learning processors.

References:
[1] K. Ueyoshi, et al., "QUEST: A 7.49 TOPS Multi-Purpose Log-Quantized DNN Inference Engine Stacked on 96MB 3D SRAM Using Inductive-Coupling Technology in 40nm CMOS," *ISSCC*, pp. 216–218, 2018.
[2] J. Lee, et al., "UNPU: A 50.6TOPS/W Unified Deep Neural Network Accelerator with 1b-to-16b Fully-Variable Weight Bit-Precision," *ISSCC*, pp. 218-219, 2018.
[3] Z. Yuan, et al., "STICKER: A 0.41-62.1 TOPSW 8bit Neural Network Processor with Multi-Sparsity Compatible Convolution Arrays and Online Tuning Acceleration for Fully Connected Layers," *IEEE Symp. VLSI Circuits*, pp. 33-34, 2018.
[4] E, Park, et al., "Energy-efficient Neural Network Accelerator Based on Outlier-aware Low-precision Computation," *ACM ISCA*, pp. 688-698, 2018.
[5] J. Albericio, et al., "Cnvlutin: Ineffectual-Neuron-Free Deep Neural Network Computing," *ACM ISCA*, 2016.
[6] B. Fleischer, et al., "A Scalable Multi-TeraOPS Deep Learning Processor Core for AI Training and Inference," *IEEE Symp. VLSI Circuits*, pp. 35-36, 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

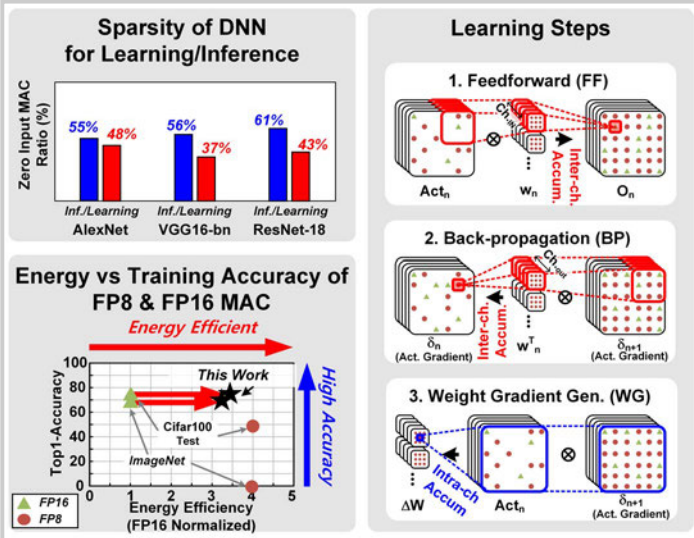


Figure 7.7.1: Data sparsity during DNN learning and accuracy-efficiency trade-off for data-type of PEs.

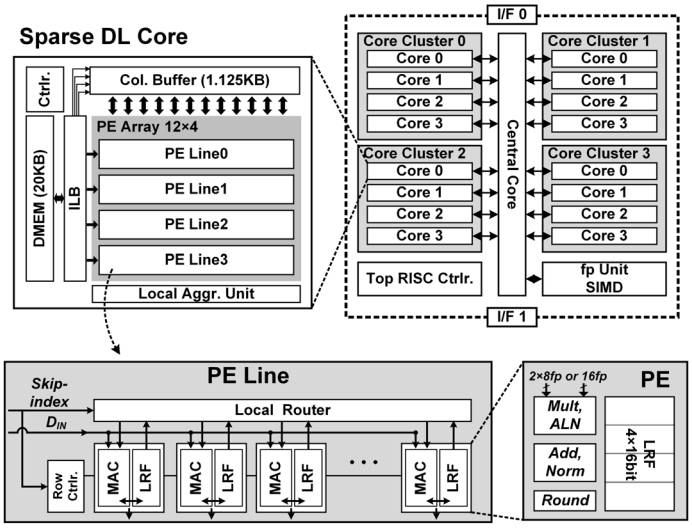


Figure 7.7.2: Overall architecture.

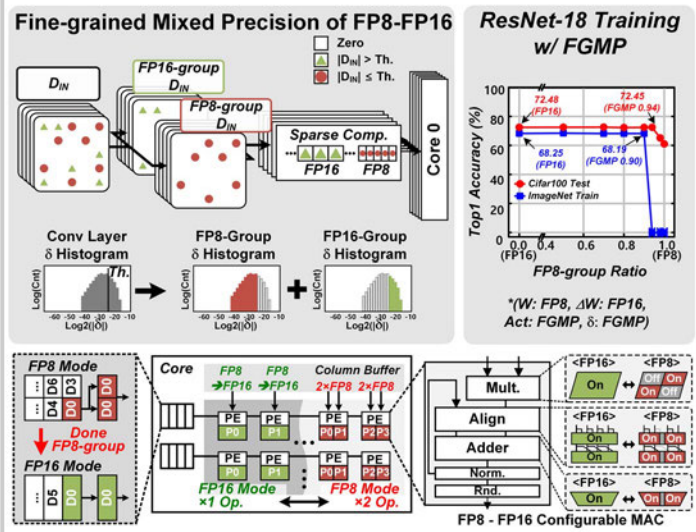


Figure 7.7.3: Fine-grained mixed precision (FGMP) between FP8 and FP16.

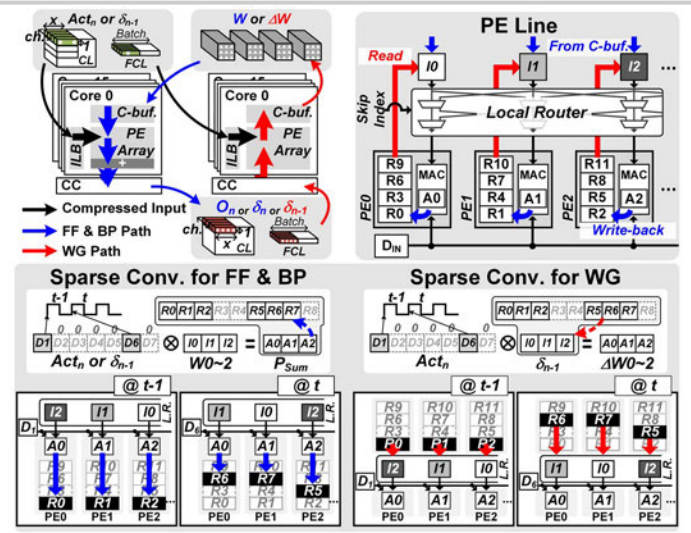


Figure 7.7.4: Data-path for three learning steps and operation of sparse DL core on PE line.

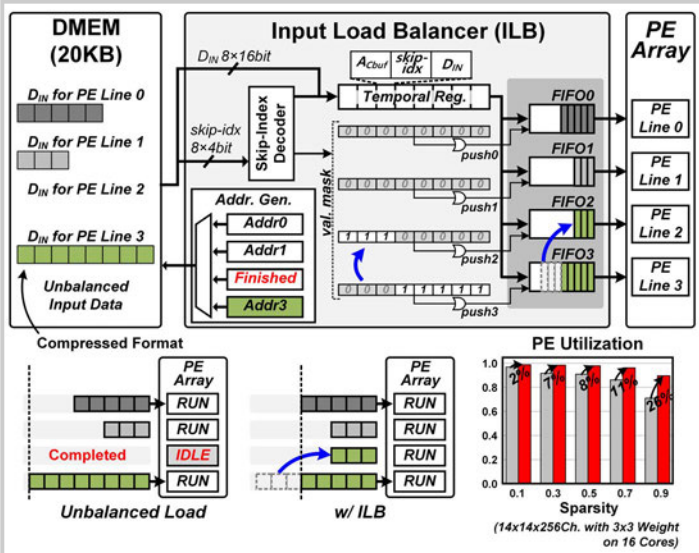
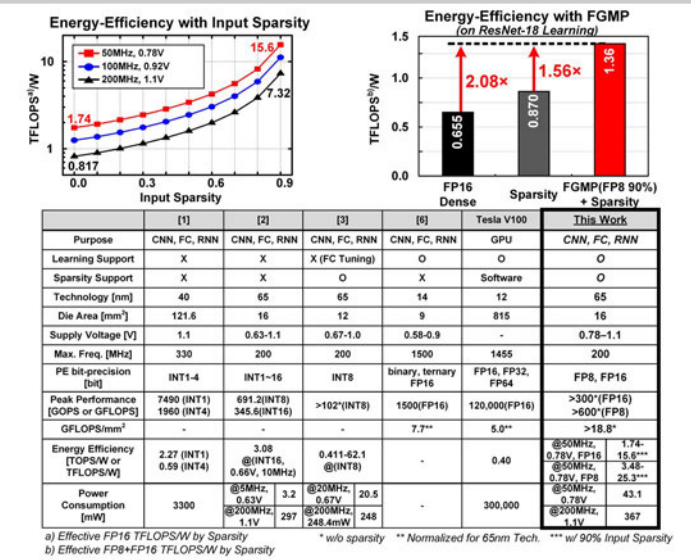


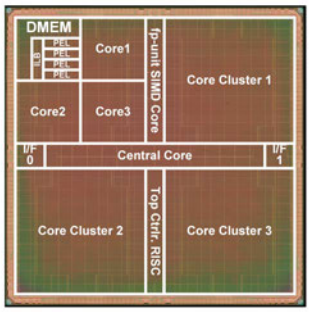
Figure 7.7.5: Input load balancer for PE utilization.



	[1]	[2]	[3]	[6]	Tesla V100	This Work
Purpose	CNN, FC, RNN	CNN, FC, RNN	CNN, FC, RNN	CNN, FC, RNN	GPU	CNN, FC, RNN
Learning Support	X	X	X (FC Tuning)	O	O	O
Sparsity Support	X	X	O	X	Software	O
Technology [nm]	40	65	65	14	12	65
Die Area [mm²]	121.6	16	12	9	815	16
Supply Voltage [V]	1.1	0.63-1.1	0.67-1.0	0.58-0.9	-	0.78–1.1
Max. Freq. [MHz]	330	200	200	1500	1455	200
PE bit-precision [bit]	INT1-4	INT1~16	INT8	binary, ternary FP16	FP16, FP32, FP64	FP8, FP16
Peak Performance [GOPS or GFLOPS]	7490 (INT1) 1960 (INT4)	691.2(INT8) 345.6(INT16)	>102*(INT8)	1500(FP16)	120,000(FP16)	>300*(FP16) >600*(FP8)
GFLOPS/mm²	-	-	-	7.7**	5.0**	>18.8*
Energy Efficiency [TOPS/W or TFLOPS/W]	2.27 (INT1) 0.59 (INT4)	3.08 @(INT16, 0.66V, 10MHz)	0.411-62.1 @(INT8)	-	0.40	@50MHz, 0.78V, FP16: 1.74-15.6*** @50MHz, 0.78V, FP8: 3.48-25.3***
Power Consumption [mW]	3300	@5MHz, 0.63V: 3.2; @200MHz, 1.1V: 297	@20MHz, 0.67V: 20.5; @200MHz, 248.4mW: 248	-	300,000	@50MHz, 0.78V: 43.1; @200MHz, 1.1V: 367

a) Effective FP16 TFLOPS/W by Sparsity
b) Effective FP8+FP16 TFLOPS/W by Sparsity
* w/o sparsity ** Normalized for 65nm Tech. *** w/ 90% Input Sparsity

Figure 7.7.6: Measurement results and performance comparison table.



	Specifications		
Technology	65nm 1P8M CMOS		
Die Area	4mm × 4mm (16mm^2)		
SRAM	372 KB		
Supply Voltage	0.78V ~ 1.1V		
Frequency	~ 200MHz		
Data Type	FP8, FP16		
Power Consumption (mW)	43.1mW @ 0.78V, 50MHz		
	367mW @ 1.1V, 200MHz		
Power Efficiency [TFLOPS/W]		FP16	FP8
	50MHz, @0.78V	1.74 -15.6* TFLOPS/W	3.48-25.3* TFLOPS/W
	200MHz, @1.1V	0.817-7.32* TFLOPS/W	1.63-11.9* TFLOPS/W

*Effective TFLOPS/W with 90% Input Sparsity

Figure 7.7.7: Chip micrograph and performance summary.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 8 Overview: *DC-DC Converters*

POWER MANAGEMENT SUBCOMMITTEE

Session Chair: ***Gerard Villar Pique***
NXP Semiconductor, Eindhoven, The Netherlands

Associate Chair: ***Jason Stauth***
Dartmouth Colege, Hanover, NH

Subcommittee Chair: Yogesh Ramadass, Texas Instruments, Santa Clara, CA

DC-DC converters are rapidly evolving towards more efficient and compact solutions. In this sense, hybrid converters bring interesting benefits by the smart combination of inductors and capacitors to efficiently transfer energy from input to output. At the same time, full integration of DC-DC converters is enabled by small technology nodes down to 14nm. On the other hand, this session will also reflect the increasing interest towards automotive applications and their demands in terms of wide input voltage range and EMI.

8:30 AM

8.1 A 93.8% Peak Efficiency, 5V-Input, 10A Max I_{LOAD} Flying Capacitor Multilevel Converter in 22nm CMOS Featuring Wide Output Voltage Range and Flying Capacitor Precharging

C. Schaef, Intel, Hillsboro, OR

In Paper 8.1, Intel presents a 4-level flying capacitor multilevel converter that operates from a 5V supply and delivers up to 10A load current with a wide output voltage range. Using a 22nm process, a 10nH inductor, and 5MHz switching frequency, the design is able to achieve up to 93.8% efficiency at a 1.8V output voltage.

9:00 AM

8.2 A Continuous-Input-Current Passive-Stacked Third-Order Buck Converter Achieving 0.7W/mm^2 Power Density and 94% Peak Efficiency

A. Abdulslam, University of California, San Diego, La Jolla, CA

In Paper 8.2, the University of California, San Diego presents an input-passive-stacked buck DC-DC converter that processes power via a stack of two inductors and one capacitor connected to the input. Using small off-chip passive components the design achieves 86.6% conversion efficiency at 0.7W/mm^2 power density.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

9:30 AM

8.3 A 10.9W 93.4%-Efficient (27W 97%-Efficient) Flying-Inductor Hybrid DC-DC Converter Suitable for 1-Cell (2-Cell) Battery Charging Applications

C. Hardy, University of Colorado, Boulder, CO

In Paper 8.3, the University of Colorado presents a step-down hybrid converter that uses an input flying inductor and a switched-capacitor network, suitable for charging one- or two-cell batteries from a 9V input. Parasitic inductance of a USB cable is utilized in place of a discrete inductor, reducing on-board power dissipation to 630mW while delivering 7.2W.

10:15 AM

8.4 Fully Integrated Buck Converter with 78% Efficiency at 365mW Output Power Enabled by Switched-Inductor-Capacitor Topology and Inductor Current Reduction Technique

N. Tang, Washington State University, Pullman, WA

In Paper 8.4, Washington State University describes a fully integrated DC-DC hybrid converter in standard 65nm CMOS technology. The implemented topology reduces the amount of current flowing through the inductor, making it more suitable to lower-quality fully integrated inductors. The converter is able to provide 730mW/mm^2 while running at 450MHz.

10:45 AM

8.5 A Fully Integrated Voltage Regulator in 14nm CMOS with Package-Embedded Air-Core Inductor Featuring Self-Trimmed, Digitally Controlled Variable On-Time Discontinuous Conduction Mode Operation

C. Schaef, Intel, Hillsboro, OR

In Paper 8.5 Intel presents a fully integrated voltage regulator in 14nm CMOS with a 2.5nH package-embedded air-core inductor and using variable ON-time discontinuous conduction mode (DCM) operation. Operating at 70MHz, the design achieves conversion efficiencies between 75% and 88% with 500mA load current at a range of output voltages, and improved light-load efficiency.

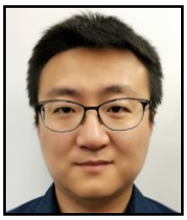
11:15 AM

8.6 A Fully Integrated 85%-Peak-Efficiency Hybrid Multi-Ratio Resonant DC-DC Converter with 3.0-to-4.5V Input and 500µA-to-120mA Load Range

P. Renz, Leibniz University Hannover, Hannover, Germany

In Paper 8.6, Leibniz University Hannover introduces a fully integrated hybrid multi-ratio resonant DC-DC converter with on-chip inductance and capacitors, suitable for wearable applications. It supports 3.0V-to-4.5V input, 500µA-to-120mA output current and achieves a peak efficiency of 85%. The dynamic load range is 19× wider and the efficiency enhancement factor is 12.8% better than prior art.

11:45 AM



8.7 A 2MHz 4-to-60V_{IN} Buck-Boost Converter for Automotive Use Achieving 95% Efficiency and CISPR 25 Class 5 Standard

J. Xue, Analog Devices, Milpitas, CA

In Paper 8.7, Analog Devices presents a 2MHz 4-to-60V_{IN} automotive-use silent-switcher buck-boost converter. The converter achieves CISPR 25 Class-5 limits. Balanced peak-current mode control minimizes both switching and conduction loss, resulting in 95% power efficiency and >2.5× power density with 10× switching frequency.

8.1 A 93.8% Peak Efficiency, 5V-Input, 10A Max I_{LOAD} Flying Capacitor Multilevel Converter in 22nm CMOS Featuring Wide Output Voltage Range and Flying Capacitor Precharging

Christopher Schaef[1], Sheldon Weng[1], Beomseok Choi[2], William Lambert[2], Kaladhar Radhakrishnan[2], Krishnan Ravichandran[1], James Tschanz[1], Vivek De[1]

[1]Intel, Hillsboro, OR; [2]Intel, Chandler, AZ

Power delivery for compute SoCs has become a major challenge due to increasing current demands, shrinking form factors, as well as highly dynamic load patterns. Traditional inductive topologies like buck converters still require large inductances since switching frequencies cannot be increased without degrading efficiency. As a result, converter sizes are difficult to scale and transient responses are slow. On the other hand, switched capacitor converters (SCCs) are promising for small form factors and offer fast transient response since they do not require large inductors. However, SCCs can achieve high efficiency only at fixed conversion ratios, thus limiting their usage since a wide continuous range of input and output voltages are needed in SoC platforms. Hybrid converters offer the combined benefits of conventional buck and SC topologies [1-3]. They can provide superior power density and efficiency by reducing both inductance and capacitance requirements significantly. Since they use stacked low voltage transistors like SCCs, higher switching frequencies can be achieved without degrading efficiency. In addition, soft-charging of the flying capacitors through the inductor eliminates charge-sharing losses present in pure SCCs.

Several hybrid converter topologies have been previously demonstrated [1,2]. In this paper, we present a 4-level FCML topology implemented in a 22nm CMOS process with three cascoded complementary switch pairs (See Fig. 8.1.7). The powertrain, bootstrapped gate-drivers, level shifters and high-resolution 3-phase PWM are integrated on chip, while the digital controller is implemented on an FPGA. The converter operates at switching frequencies of 5 to 10MHz and demonstrates (i) startup precharging of the flying capacitors, (ii) closed-loop control with flying capacitor balancing, and (iii) operation across two nominal conversion ratios.

The powertrain schematic with bootstrapped gate drivers is shown in Fig. 8.1.1. It consists of three high-side PMOS and three low-side NMOS switches. Each switch is implemented with a cascode of two thick-gate transistors ($V_{gs,max}$ =1.8V, $V_{ds,max}$ = 1.2V). The input capacitors, C_{in}, and flying capacitors C_{x1} and C_{x2} are implemented with 6x0402 ceramic capacitors on the package, and hold $1/3V_{in}$ and $2/3V_{in}$ respectively. Gate-drive supplies for S2/S5 are generated with on-chip bootstrapping switches that charge the capacitors C_{g2}/C_{g5} when the power switches S1/S5 are turned on. Moreover, S3/S4 require nested bootstrapping, which uses C_{g2}/C_{g5} as the source when S2/S5 are turned on to charge C_{g3}/C_{g4}. Capacitors C_{g1-6} are implemented with a combination of on-chip MIM capacitors and a 01005 ceramic capacitor on package.

The FCML converter operates by configuring the flying capacitors in different states to produces different voltage levels at the switching node V_x, as illustrated in Fig. 8.1.2. A 4-level FCML topology can produce voltages of 0, $1/3V_{in}$, $2/3V_{in}$ and V_{in} at V_x without changing the flying capacitor voltages. This allows the FCML topology to produce a continuous wide output voltage range that is not constrained by the nominal conversion ratio as in other SC topologies. Switches S1-S6 operate in complimentary pairs (S1/S6, S2/S5, S3/,S4) where each pair is controlled with a PWM signal that is phase-shifted by 120° relative to the neighboring switch pair. Each PWM transition of one switch pair creates a voltage transition on V_x, and produces an effective switching frequency that is 3× higher than the PWM frequency. This results in a significant switching loss reduction compared to other topologies since each device is only turned on and off every third switching period. Measured waveforms in Fig. 8.1.2 illustrate operation in two different duty-cycle ranges where V_x either transitions between 0 and $1/3V_{in}$ (D<1/3) or $1/3V_{in}$ and $2/3V_{in}$ (1/2<D<2/3).

In order to generate the three phase-shifted PWM signals with sufficient resolution, a hybrid PWM is implemented on chip (Fig. 8.1.3). It uses a counter and delay line to create the PWM signals. The 8b delay line is powered by an on-chip LDO controlled by a 10b DAC that can change the supply voltage, and thus modulate the delay. A digital DLL controls the LDO to match the delay line to the reference clock period, which controls the PWM frequency. In this implementation, each PWM cycle is divided into 6 clock periods, i.e. the three PWMs are phase shifted by two clock periods. Since the delay line is shared between the three PWMs, careful timing of the delay-line mux is required as duty cycles of the three PWMs can vary even in steady state for flying capacitor balancing. The PWM is tested by digitally sweeping the duty-cycle command and measuring the output voltage. Measurements show that good linearity is achieved even at the crossover points when the converter transitions from the 1/3 to the 2/3 operating regime.

One of the main challenges of SC and hybrid converters has been startup from 0 to nominal input voltage, since all flying capacitors have to be kept at a specific voltage to prevent over-voltage on the power devices. Figure 8.1.4 shows startup waveforms for a 0-to-5V V_{in} transition. The power stage is kept in specific states to charge the flying capacitors as the input voltage ramps up. While the input voltage is below 1/3 of the nominal value, switches S1, S2, S5 and S6 are turned on, thus connecting both flying capacitors directly to V_{in}. After V_{in} rises above 1/3 of its nominal value, S2 and S5 are turned off which stops charging of C_{x1} but continues charging of C_{x2} until V_{in} reaches 2/3 of its nominal value. Then all switches are turned off. After V_{in} has reached its nominal value, the converter enters active operation. One advantage of this method is that it does not need an additional precharge circuit which can be area-consuming depending on the required ramp rate. In this implementation, delay and accuracy of the off-chip instrumentation limited ramp-rate and settling errors whereas a fully integrated implementation could achieve even higher speed and accuracy with this technique. However, this method does require some knowledge of the nominal input voltage prior to startup.

Another key challenge for hybrid converters is the charge balance on the flying capacitors. Since the flying capacitors are charged and discharged through the inductor, their voltages can drift from their nominal values if charge and discharge cycles are not matched. While most hybrid converters, including FCML, have natural balancing mechanisms, imbalance can still occur as illustrated in Fig. 8.1.4. Hence a balancing control loop is required to always ensure balanced operation. Here, a control loop is implemented by sensing V_{in} as well as both flying capacitor voltages V_{cx1} and V_{cx2}. By adjusting the duty cycles of pwm1 and pwm3 relative to the nominal duty cycle, as defined by the main output-voltage control loop, the charge and/or discharge of the flying capacitor can be controlled, thereby returning their voltages to a balanced state. This balancing control loop operates at much slower speed compared to the output voltage controller, so as not to interfere with output voltage regulation. The main output voltage control loop is identical to a traditional buck converter. The compensator is implemented digitally and defines the nominal duty cycle for all three PWMs. Due to the small inductance and high switching frequency, the converter achieves fast transient response as illustrated in Fig. 8.1.4 for a 0-to-6A output current step. Measurements of V_{in}, V_{cx1}, V_{cx2} and V_{out} were performed with 4 channels of off-chip difference amplifiers and ADCs which feed in to the digital controller as illustrated in Fig. 8.1.1. All measurements were taken with 4×4.7μF of output decoupling that were placed together with the load (not included on the same package).

Efficiency is measured across a range of load currents for a V_{in} of 5V, V_{out} of 1.2 and 1.8V, and switching frequencies of 5 and 10MHz as shown in Fig. 8.1.5. A peak efficiency of 93.8% is achieved at 3A for an output voltage of 1.8V at 5MHz and above 89% efficiency is maintained up to the maximum load current of 10A. At 1.2V output a peak efficiency of 90.5% is achieved at 4A load current.

The hybrid converter demonstrated in this work uses a significantly smaller inductance of only 10nH compared to previously reported designs and operates at a much higher switching frequency. In addition, the power density and the load current of 10A are the highest among previously reported hybrid converters as illustrated in Fig. 8.1.6.

Acknowledgements
The authors would like to thank Ye Seul Nam, Jimin Yao and Shawna Liff for assembly, Dan Lake, Robert Flory and Sheldon Sun for test platform design, Tri Huynh and Linda Low for mask design, Jin Feng and Huzaifa Rampurawala for digital design, Sirisha Kale for integration support and Adriana Sanabria Borbon for her internship contributions.

References:
[1] W. Liu et al., "10.3 A 94.2%-Peak-Efficiency 1.53A Direct-Battery-Hook-Up Hybrid Dickson Switched-Capacitor DC-DC Converter with Wide Continuous Conversion Ratio in 65nm CMOS," *ISSCC,* pp. 182-183, Feb. 2017.
[2] C. Schaef et al., "10.2 A Digitally Controlled 94.8%-Peak-Efficiency Hybrid Switched-Capacitor Converter for Bidirectional Balancing and Impedance-Based Diagnostics of Lithium-Ion Battery Arrays," *ISSCC,* pp. 180-181, Feb. 2017.
[3] Y. Karasawa et al., "A 92.8% Efficiency Adaptive-On/Off-Time Control 3-Level Buck Converter for Wide Conversion Ratio with Shared Charge Pump Intermediate Voltage Regulator", *IEEE Symp. VLSI Circuits,* 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

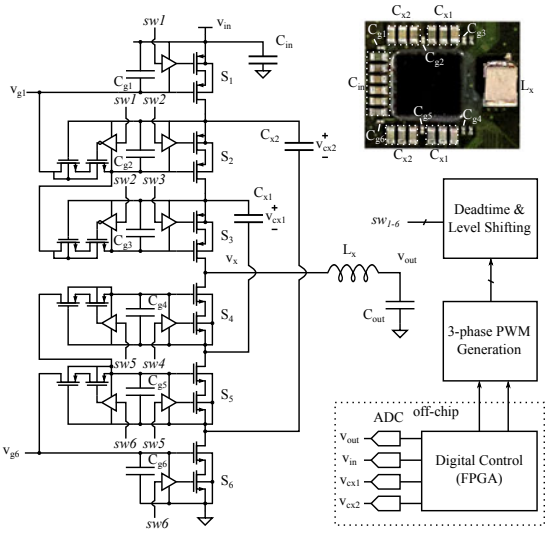


Figure 8.1.1: Circuit schematic of 4-level FCML powertrain with block diagram of control loop and photograph of co-packaged components.

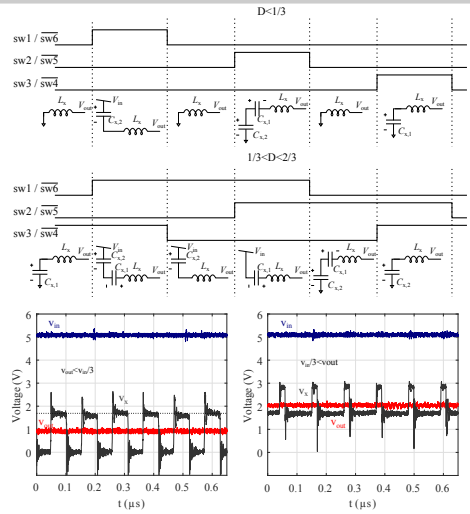


Figure 8.1.2: PWM switch signals for different operating duty cycles (top) along measured switch node and input/output voltage waveforms showing converter operation in two nominal conversion ratio regions.

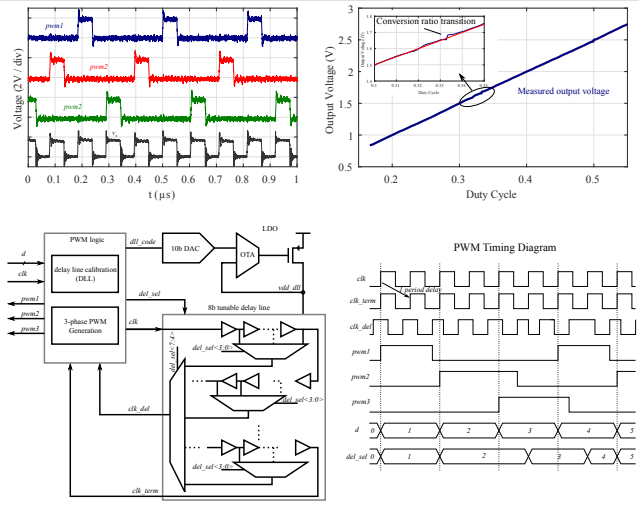


Figure 8.1.3: Overview of on-chip 3-phase PWM generation with measured waveforms of PWMs and switch node voltage. Output voltage vs duty-cycle measurements show good linearity across a wide output voltage range.

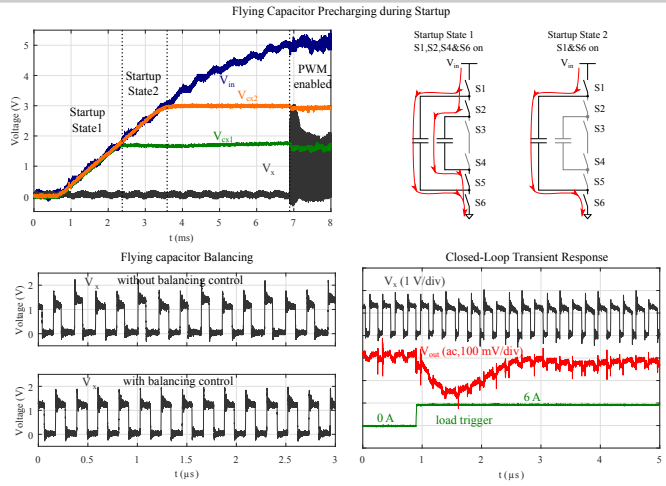


Figure 8.1.4: Measured waveforms during startup show precharging of flying capacitors through different power-stage switching states. Switch node and input/output voltage waveforms demonstrate balancing control as well as closed-loop response to a 6A load transient.

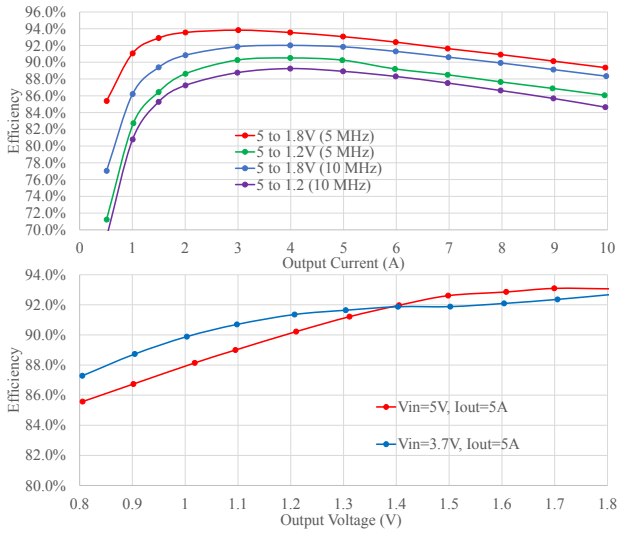


Figure 8.1.5: Measured Efficiency across output current and input/output voltage.

	This Work	[1]	[2]	[3]
Topology	4L FCML	4:1 Dickson	2:1 ReSC	3L FCML
Input Voltage	5 V	4.2 V	7.2 V	5 V
Output Voltage	1.8 V	1 V	3.6 V	1.5 V
Startup Precharge	Yes	No	No	No
Balancing	Yes	Yes	No	Yes
Multiple Ratios	Yes	No	No	Yes
Peak Output Current	10 A	1.05 A	1 A	0.8 A
Peak Efficiency @ I_{out}	93.8% @ 3 A	94.2% @ 0.18 A	94.8 % @ 0.3A	80%
Efficiency @ I_{max}	89.4 %	88.3 %	90.5%	78%
Inductance	10 nH	470 nH	100 nH	220 nH
Flying capacitance	2 x 13.2 µF	3 x 22 µF	2x 0.47µF	1µF
Output Decoupling	18 µF	22 µF	2 x 1 µF	2.4 µF
Sw. Frequency	5 MHz	500 kHz	1 MHz	4.4-11.8 MHz
Die Area current density	1.45 A/mm2	0.26 A/mm2	0.17 A/mm2	0.4 A/mm2
Vol. power density*	198 W/cm3	85 W/cm3	19 W/cm3	-
Total z-height	1.2 mm	3 mm**	2 mm	-

* considers volume of bounding box including all passives
** estimated from photograph

Figure 8.1.6: Comparison Table with recently published hybrid dc-dc converters.

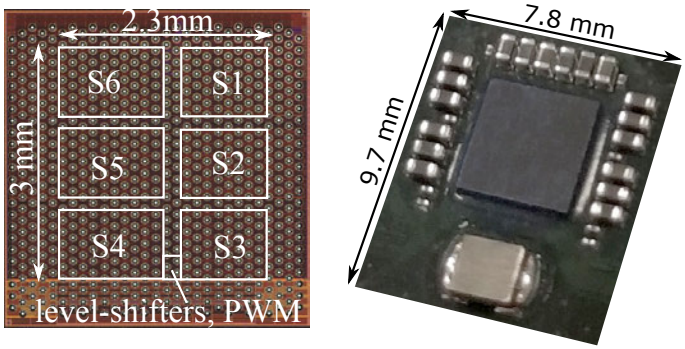


Figure 8.1.7: Annotated die micrograph and package photograph with dimensions. Active die area for powerstage, gate drivers, level-shifters and PWM is 6.9mm².

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

8.2 A Continuous-Input-Current Passive-Stacked Third-Order Buck Converter Achieving 0.7W/mm² Power Density and 94% Peak Efficiency

Abdullah Abdulslam, Patrick P. Mercier

University of California, San Diego, La Jolla, CA

The power density and efficiency of power-management integrated circuits (PMICs) is playing an increasingly important role in the miniaturization of modern computing platforms. Small inductors can be used to help miniaturize buck DC-DC converters, however as noted in Fig. 8.2.1 (lower left), small inductors tend to have high DC resistance (DCR) – greater than a comparably-sized CMOS switch – such that they ultimately limit the achievable power density of miniaturized buck converters to <0.4W/mm² when including the area of both the passives and the PMIC [1,2]. While recent work in switched-capacitor (SC) converters has shown that high power density is achievable, it is only possible when employing exotic ultra-high-density capacitors and when operating over a small number of conversion ratios; increasing the number of ratios to support the needs of dynamic voltage scaling loads (e.g., 0.4 to 1.2V) alongside use of conventional capacitor technologies degrades power density to <<0.1W/mm² [3]. Hybrid converters, which process power with both capacitors and inductors, offer an attractive means to potentially increase power density and/or efficiency. Resonating an SC circuit with an inductor can, for example, dramatically increase power density to 0.9W/mm² in [4], however with limited, though improved in [4], ability to regulate beyond the nominal SC ratio at high efficiency. Hybrid multilevel converters can regulate to arbitrary output voltages, though still achieve power density <0.3W/mm² [5], in part due to the increased number of passives and associated higher routing complexity, and in part due to the large conversion ratio in the design in [5].

To further improve power density while supporting regulation of arbitrary output voltages, this paper presents a hybrid DC-DC converter topology, shown in Fig. 8.2.1, that splits the inductor from a conventional buck converter into two half-sized inductors that are stacked on top of and below the input coupling capacitor, which is now flying, to form a passive-stacked 3rd-order buck (PS3B) converter. Unlike the conventional buck converter where the small inductor, which has a larger DCR than the switches in a well-designed comparably-sized PMIC, processes power on the high-current/low-voltage side of the converter, the stacked inductors in the PS3B process power on the low-current/high-voltage side of the converter. Even when accounting for the larger DCR imposed by the half-sized inductors, the overall conduction loss of the PS3B converter is up to 1.4× lower than an equivalent-area conventional buck converter. The input current in the PS3B is also continuous, in stark contrast to the fully pulsated input current of a conventional buck converter, which helps reduce the area of the passives required for filtering. Compared to a 2-phase buck converter, the PS3B has the same theoretical conduction losses at D=0.5, though, importantly, with fewer power transistors (2 vs. 4) and fewer power pins (3 vs. 4), which facilitates less routing complexity that, when combined with the lack of area and/or energy-consuming inter-phase alignment circuits (e.g., series current sensing resistors or equivalent), enables an overall increase in power density compared to prior-art when including the area of the passives [6].

Figure 8.2.2 shows a detailed circuit diagram of the converter along with its operating principle. In phase $Ø_1$, V_{OUT} is connected to V_{X1} via M_1, and thus L_1 experiences a positive voltage of (V_{IN}-V_{OUT}) and is energized, L_2 experiences a negative voltage (V_{OUT}-V_{IN}) and is deenergized, while C_F discharges to the output via I_2. In phase $Ø_2$, V_{OUT} is connected to V_{X2} via M_2, and thus L_2 experience a positive voltage (V_{OUT}) and is energized, L_1 experience a negative voltage (-V_{OUT}) and is deenergized, and C_F is charged via I_1. Thus, V_{X1} switches between V_{OUT} and (V_{OUT}+V_{IN}), while V_{X2} switches between (V_{OUT}-V_{IN}) and V_{OUT}. Accordingly, based on inductor volt-second balance, the voltage on C_F is balanced at V_{IN} while V_{OUT}=DV_{IN}, which is exactly the same relationship as a conventional buck converter, though in this case via three reactive elements – hence it is a 3rd-order buck.

Since V_{X1} and V_{X2} switch between non-standard voltages, M_1 and M_2 are driven with level-shifted signals generated via on-chip MOS-based bootstrapping capacitors, C_{boot1} and C_{boot2}. Here, the top plate of C_{boot1} is nominally charged to (V_{OUT}+V_{IN}) during phase $Ø_2$ via M_{13}, while the bottom plate of C_{boot2} is nominally charged to (V_{OUT}-V_{IN}) during phase $Ø_1$ via M_{23}, with the other terminals of both capacitors connected to V_{OUT} at all times. A deadtime circuit is used to avoid short-circuit current by transitioning from $Ø_1$ to $Ø_2$ according to the following sequence: 1) C_{boot1} is disconnected from M_1 as M_{12} turns off, and the gate capacitance of M_1 is discharged to V_{OUT} since M_{11} turns on; 2) C_{boot2} is disconnected from V_{X2} as M_{23} turns off; 3) C_{boot2} discharges the gate capacitance of M_2 to (V_{OUT}-V_{IN}) as M_{22} turns on; and 4) C_{boot1} is connected to V_{X1} for its top plate to be charged to (V_{OUT}+V_{IN}). Transitioning from $Ø_2$ to $Ø_1$ is done through the reverse manner. Although the voltages at V_{X1} and V_{X2} operate over V_{IN} or below ground, all of the MOSFETs, the bootstrapping capacitors, and the flying capacitor only ever experience a blocking voltage of V_{IN}, thereby ensuring reliable operation. Importantly, the parasitics of the power MOSFETs experience the same ΔV as a conventional buck, and thus do not suffer from additional switching losses. To mitigate potential ringing on the MOSFET driver lines from bondwire inductance, a small on-chip MIM capacitor, $C_{F,on\text{-}chip}$, is placed in parallel to the off-chip C_F.

Figure 8.2.3 shows the controller used to generate the four control signals of the converter. Two-stage level-shifters shift the control signals from {0,V_{IN}} to the appropriate levels for the drivers. All of the drivers and the level shifters are powered by either V_{OUT} or by one of the two bootstrapping capacitors, and thus no external voltage sources are needed. The timing between the different control signals is adjusted carefully through a deadtime circuit so that the dis/charging of MOSFETs' gate capacitance and the bootstrapping capacitors is done in the proper sequence. As shown in Fig. 8.2.3, C_{boot2} is implemented in deep *n*-well since its bottom-plate experiences a negative voltage, while C_{boot1} is implemented as a usual MOS capacitor. Interestingly, the bulk of the chip is biased at V_{OUT} instead of ground for convenience, since the power MOSFETs and most of the driver switches are referenced to V_{OUT} (though through use of deep *n*-wells this is not strictly necessary).

The PS3B converter is implemented in a 1.85mm² 0.18µm CMOS chip, which is wirebonded to a 5mm² interposer comprising two 1.3mm² 240nH inductors (L_1 and L_2), two 0.5mm² 470nF capacitors (C_F and C_{OUT}), and all necessary routing and wirebonds as shown in Fig. 8.2.3. The converter operates with V_{IN}=1.8V, which is a rail commonly found in mobile, desktop, and server-grade computing platforms, and generates SoC-compatible voltages between 0.4 and 1.5V at currents up to 2.5A. Since the input current is continuous, no additional input capacitors are needed, saving additional area compared to conventional implementations. Measurement results in Fig. 8.2.4 show that V_{X1} and V_{X2} switch as expected at up to 6.5MHz to generate V_{OUT}, and under PWM control the converter has <50mA input current ripple, <20mV output voltage ripple, and can respond to a 0-to-1A current step with <170mV droop in <20µs. The converter achieves a peak power density of 0.7W/mm² at an efficiency of 86.6%, and a peak efficiency of 94% at a power density of 0.18W/mm², with both figures including the full area of the converter, passives, and wirebonding/routing. If considering only the area of the passives, assuming advanced packaging is available to reduce interconnect area, then the converter achieves a power density of 1W/mm². Compared to prior art summarized in Figs. 8.2.5 and 8.2.6, the converter achieves >2× higher power density or 16% higher efficiency amongst designs that efficiently operate across a wide, SoC-compatible voltage range, and that report the size of the passives. A photograph of the assembled PMIC is shown in Fig. 8.2.7.

Acknowledgements:
This work was supported by pSemi Corporation.

References:
[1] S. J. Kim et al., "High Frequency Buck Converter Design Using Time-Based Control Techniques," *IEEE JSSC,* vol. 50, no. 4, pp. 990-1001, 2015.
[2] Texas Instruments, TPS8268150 Data sheet Rev C., "TPS8268x 1600-mA High-Efficiency MicroSiP™ Step-Down Converter Module," Oct. 2014, Revised June 2015, Accessed Oct. 2018, <www.ti.com/lit/gpn/tps8268150>.
[3] Y. Jiang et al., "A 0.22-to-2.4V-Input Fine-Grained Fully Integrated Rational Buck-Boost SC DC-DC Converter Using Algorithmic Voltage-Feed-In (AVFI) Topology Achieving 84.1% Peak Efficiency at 13.2mW/mm²," *ISSCC,* pp. 422-424, Feb. 2018.
[4] C. Schaef et al., "A Variable-Conversion-Ratio 3-Phase Resonant Switched Capacitor Converter with 85% Efficiency at 0.91W/mm² Using 1.1nH PCB-Trace Inductors," *ISSCC,* pp. 1-3, Feb. 2015.
[5] W. Liu et al., "A 94.2%-Peak-Efficiency 1.53A Direct-Battery-Hook-Up Hybrid Dickson Switched-Capacitor DC-DC Converter with Wide Continuous Conversion Ratio in 65nm CMOS," *ISSCC,* pp. 182-183, Feb. 2017.
[6] B. Lee et al., "A 25MHz 4-Phase SAW Hysteretic DC-DC Converter with 1-Cycle APC Achieving 190ns T_{settle} to 4A Load Transient and Above 80% Efficiency in 96.7% of the Power Range," *ISSCC*, pp. 190-191, Feb. 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

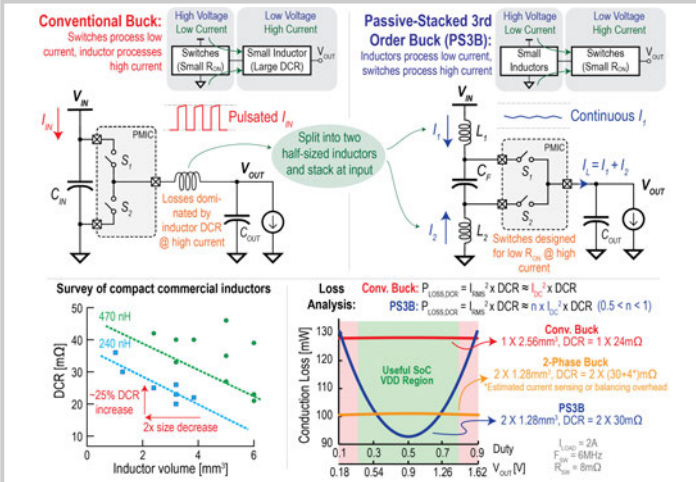


Figure 8.2.1: A conventional buck converter (top left) vs. the PS3B converter (top right) designed to operate small inductors, which have high DCR (bottom left), at the high-voltage low-current side of the converter to minimize conduction losses (bottom right).

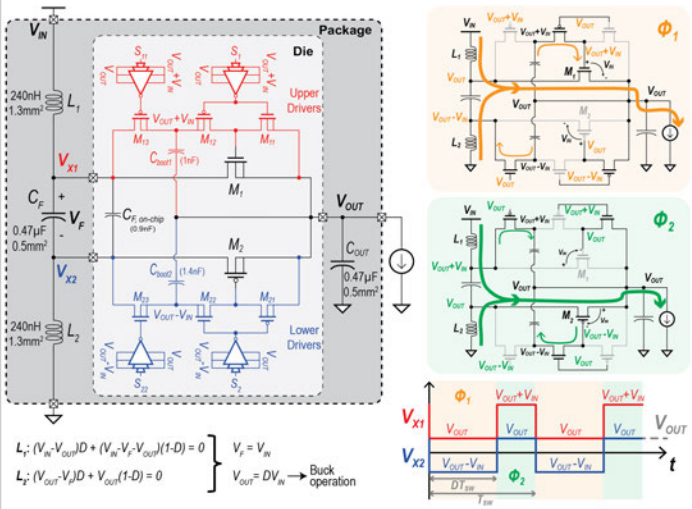


Figure 8.2.2: Schematic of the PS3B converter, along with associated timing diagrams.

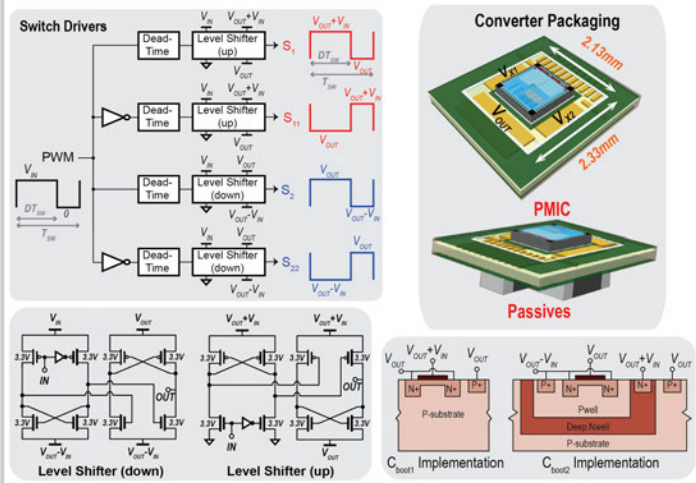


Figure 8.2.3: Generation of level-shifted signals from an input PWM signal (top left); level-shifter schematics (bottom left); 3D render of converter packaging (top right); bootstrap capacitor cross-section (bottom right).

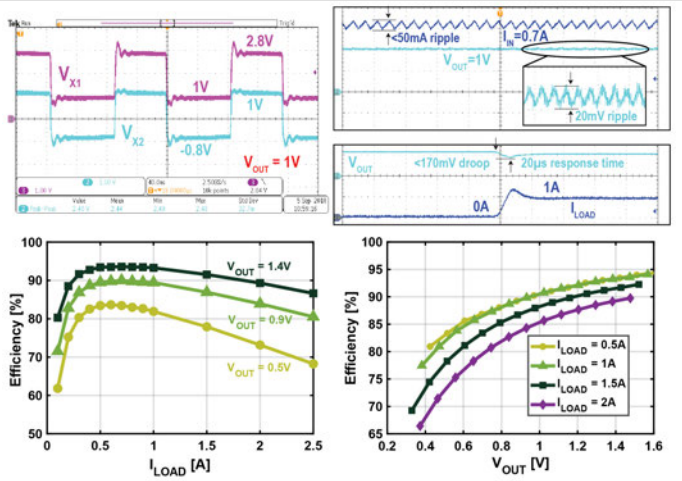


Figure 8.2.4: Measured internal waveforms of the converter (top left); measured input current ripple, output voltage ripple, and response time during a 0-to-1A load step (top right); efficiency vs. load current (bottom left) and output voltage (bottom right).

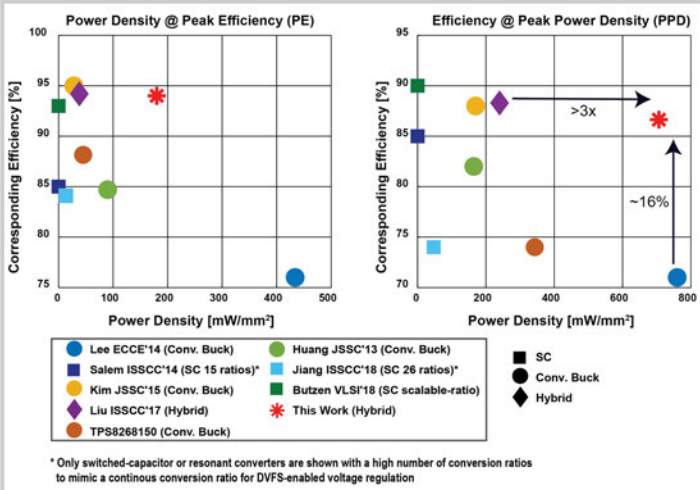


Figure 8.2.5: Measured efficiency and power density compared to prior-art converters that can efficiently regulate across SoC-compatible voltages and that report the total converter area (including the area of the passives and the interconnection between them).

	Kim JSSC'15	Lee ISSCC'17	Liu ISSCC'17	Jia JSSC'18	Jiang ISSCC'18	TI TPS8268150	This Work
Topology	Buck	Buck (4-phase)	Hybrid (Multilevel)	Buck	SC (26 ratios)	Buck	**Hybrid (PS3B)**
Technology	65nm	350nm	65nm	65nm	65nm	NR	**180nm**
Input Voltage [V]	1.8	3.3	3.0 – 4.5	1.1	0.22 – 2.4	2.3 - 5.5	**1.8**
Output Voltage [V]	0.6 – 1.5	0.3 – 2.5	0.3 – 1.0	0.3 – 0.86	0.85 – 1.2	1.5	**0.5 – 1.5**
I_{LOAD} (MAX) [A]	0.6	6	1.53	0.04	0.08	1.6	**2.5**
Input Current	Pulsated	Pulsated	Pulsated	Pulsated	Pulsated	Pulsated	**Continuous**
PMIC Area [mm²]	5	1.88	4.05	0.13	2.42	6.67	**1.85**
Total Footprint [mm²]	5	NR	6**	0.13	2.42	6.67	**5**
Peak Efficiency (PE)	95.5%	88.1%	94.2%	73%	84.1%	88%***	**94%**
Power Density @ PE * [W/mm²]	0.03***	NR****	0.04**	NR	0.013	0.045***	**0.18 (0.25**)**
Peak Power Density (PPD) * [W/mm²]	0.17	NR****	0.24	0.27	0.034***	0.36***	**0.7 (1.0**)**
Efficiency @ PPD	88%	NR****	88.3%	NR	74%***	74%***	**86.6%**

NR = Not Reported
* Unless otherwise specified, power densities are computed with respect to the total converter footprint including the passives and routing
** Computed by summing the area of the passive component footprints
*** Estimated from measurement results
**** The area of the passives are not reported

Figure 8.2.6: Comparison to prior-art converters that operate across SoC-compatible voltage ranges.

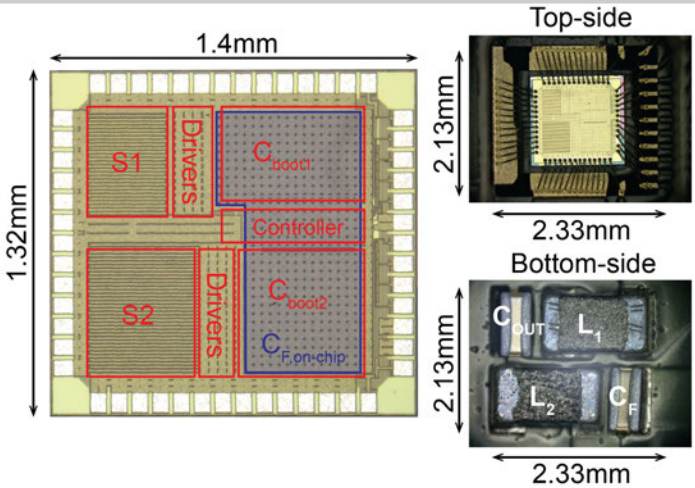


Figure 8.2.7: Micrograph of the fabricated stacked-passives 3rd-order buck converter (left); top side of the wirebonded prototype (top right); bottom side of the prototype (bottom right). The converter occupies 5mm² including the PMIC, passives, and the package area including the wirebonds and pads.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

8.3 A 10.9W 93.4%-Efficient (27W 97%-Efficient) Flying-Inductor Hybrid DC-DC Converter Suitable for 1-Cell (2-Cell) Battery Charging Applications

Casey Hardy, Hanh-Phuc Le

University of Colorado, Boulder, CO

The annual increase in performance and feature density of mobile products has driven the need for batteries with higher capacity and thus higher power-delivery solutions to maintain sensible charging times. The USB-C power-delivery specification was developed to meet these higher demands by enabling the input supply V_{BUS} voltage at a higher setting, e.g. 9V, to allow higher power delivery, up to 100W, while minimizing current levels and associated IR drops across the USB-C cable. To bridge the increased gap from this input voltage to charge one- and two-cell battery products, a power- and space-efficient DC-DC converter is desirable.

The key challenge in realizing such converters is achieving the needed output power while meeting product form-factor and on-board thermal constraints. While the traditional Buck converter has been widely used [1], hybrid converter topologies reported in [2,3] are potential candidates for this application. However, these topologies have the inductor located at the high-current output path, leading to a disadvantageous trade-off between inductor size and its DCR losses and saturation current, which gets worse with increasing demands for high charging current and compact space for USB-C chargers. Therefore, a topology that relocates the inductor to the lower-current input side while still providing the required step-down conversion is preferable. Moving the inductor to the input also introduces the possibility of utilizing the abundant parasitic inductance of a USB cable, often 100s of nH, to replace the role of the inductor. This USB cable charger solution can completely remove a bulky magnetic component and associated heat loss from on-board and spread it across the length of the cable, a relatively inexpensive part of the system. The hybrid topology presented in [4,5] illustrates the benefits of moving the inductor to the input, but its conversion capability is limited to the traditional 5V USB applications. In this paper, we demonstrate an extension of this topology for larger conversion ratios to efficiently support one- or two-cell batteries from a higher 9V USB-C V_{BUS} setting.

Figure 8.3.1 shows the block diagram and power-stage topology of the proposed converter. It consists of an input flying inductor L that feeds current into a network of two synchronous switched capacitors $C_{F1,F2}$ and six switches SW1-SW6. NMOS switches, SW1 through SW5, form the main power stage of the converter while a small PMOS switch, SW6, is used to source charge to capacitor C_{O3} to power internal circuitry.

The converter is operated in two duty-cycle-controlled phases, ϕ_1 and ϕ_2, as illustrated in Fig. 8.3.2 assuming V_{O1} output is configured as the main power delivery path, i.e. for one-cell battery applications. The timing diagram in Fig. 8.3.2 uses measured waveforms to illustrate and validate the converter operation. In ϕ_1, SW2, SW4 and SW6 are OFF while SW1, SW3, and SW5 are ON, connecting the flying inductor between V_{IN} and V_{O2} to charge its current at the rate of (V_{IN}-V_{O2})/L. C_{O2} receives charge from the inductor current and C_{F1} while C_{F2} discharges into V_{O1}. In ϕ_2, SW1, SW3, and SW5 are turned OFF while SW2 and SW4 are ON connecting switching node V_{XN}, i.e. the bottom plates of C_{F1} and C_{F2}, to V_{O1} and V_{X1P} to V_{O3}. V_{O3} is higher than V_{IN}, allowing the inductor current discharge to C_{F1} and V_{O1} at the rate of (V_{O3}-V_{IN})/L. During this phase, while C_{F1} is soft-charged by the inductor current, the charge drawn from V_{O2} though C_{F2} charges this capacitor up and flows to the load. Small SW6 is also on in ϕ_2 to draw a small charge to V_{O3} to power internal level shifters and drivers. A programmable deadtime is added at the transition from ϕ_1 to ϕ_2 to soft-charge parasitic capacitors C_{P1}, C_{P2}, and C_{P3}. Performing inductor volt-second and capacitor charge-balance analyses, the conversion ratio for V_{O1} can be calculated as V_{O1}/V_{IN} = 1/(3-D) where D is the duty cycle of SW1. V_{O2} and V_{O3} are two and three times the unloaded value of V_{O1}, respectively. The operation and analysis of the converter with V_{O2} as the loaded output can be described similarly. A regulation feedback loop can be wrapped around either the V_{O1} or V_{O2} to provide regulated output voltage ranges suitable for charging either a single- or two-cell battery pack.

In this hybrid topology, the back-end switched-capacitor (SC) network limits the blocking voltage of power switches to only V_{O1}, a fraction of V_{IN}, enabling the use of low-voltage 5V devices for lower $R_{DS,ON}$ and thus higher efficiency. To leverage this topological benefit, careful design considerations are given to gate drivers, level-shifters, and signal routing that leverage available voltage rails, V_{O1}, V_{O2}, and V_{O3} to ensure power-switch operation within their gate voltage ratings. As shown in Fig. 8.3.3, the gate signals G1, G3, G5, and G6 for SW1, SW3, SW5, and SW6, respectively, can be a version of ϕ_1 or ϕ_2 in V_{O1}-GND, V_{O2}-V_{O1}, or V_{O3}-V_{O2} voltage domains using a single level shifter or a dual level shifter. To reduce the latency of these level shifters while keeping their area small, MIM capacitors are used to couple signals between different voltage domains. Weak NMOS transistors are added to the cross-coupled inverters to ensure their rail-to-rail operation. Different from other switches, SW2 and SW4 are unique in that they require their gate signals, G2 and G4, to cross two voltage domains, i.e. 0V to V_{O2} or V_{O1} to V_{O3}. In order to meet these signal ranges while not exceeding the device gate oxide voltage rating, independently controlled push-pull gate driver stages are utilized with a dedicated non-overlap signal generator to minimize shoot-through currents.

A specific design methodology and loss modeling are developed for the proposed converter to optimize the switching frequency, the total size, and partitioning of the power switches for optimal efficiency. Using this model, the converter's loss breakdown when switched at 2.3 MHz for a 3.6V/3A load is calculated as shown in Fig. 8.3.3.

A hybrid converter prototype was implemented in 7.37mm^2 of a 1P6M 0.13µm BCD process. Flip-chip bumping with a careful pinout structure (Fig. 8.3.4 left) was used to minimize package and PCB parasitics and related losses. The converter was first tested using a 1uH 5.5mΩ-DCR discrete inductor. Figure 8.3.4 (right) shows that it achieves peak efficiency of 93.4% at 3.6V/5.8W and 97% at 7.6V/16.7W, and delivers peak power of 10.9W/91.4% and 27.4W/96.4% when regulated at V_{O1} and V_{O2}, respectively.

In this demonstration, a simple voltage-mode type-1 compensator was implemented for closed-loop regulation. Measured transient responses at 1A and 0.5A load steps when the converter is regulated at V_{O1} and V_{O2} are shown in Fig. 8.3.5 (top). Output droops and overshoots are maintained below 120mV, or 3.3%, for V_{O1} at 3.6V and 120mV, or 1.7%, for V_{O2} at 7.2V.

The converter is also demonstrated with the on-board inductor replaced by the parasitic inductance of a 1m USB cable that is measured at 522nH inductance and 272mΩ resistance as shown in Fig. 8.3.5 (bottom). Accounting for on-board power losses only, since the inductor-related losses are distributed across the cable, the converter achieves peak on-board efficiency of 92.4% (97.6%) and maximum power of 7.2W (14.4W) when loaded at V_{O1} (V_{O2}). A shorter cable with lower parasitic inductance could be utilized but at the cost of degraded efficiency due to the higher ripple current. A preliminary experiment on interference to USB communication proves a reliable data connection and file transfer via the USB cable while the proposed hybrid converter is utilizing the cable parasitic inductance and delivering 3.6V/6.5W.

The proposed hybrid converter is compared with state-of-the-art converters in Fig. 8.3.6. An annotated die micrograph is shown in Fig. 8.3.7.

Acknowledgements:
The authors would like to thank Kilby Labs of Texas Instruments for sponsoring this project and in particular Yogesh Ramadass and Kevin Scoones for their valuable comments and suggestions during the development of this work.

References:
[1] Texas Instruments, "1S 5A Fast Charger MaxCharge™ Technology for High Input Voltage and Adjustable USB OTG Boost," *BQ25890 Datasheet*, 2018.
[2] W. Liu et al., "A 94.2%-Peak-Efficiency 1.53A Direct-Battery-Hook-Up Hybrid Dickson Switched-Capacitor DC-DC Converter with Wide Continuous Conversion Ratio in 65nm CMOS," *ISSCC*, pp. 182-183, Feb. 2017.
[3] C. Schaef and J. T. Stauth, "A Highly Integrated Series–Parallel Switched-Capacitor Converter With 12 V Input and Quasi-Resonant Voltage-Mode Regulation," *IEEE JESTPE*, vol. 6, no. 2, pp. 456-464, 2018.
[4] G. Seo and H. Le, "An Inductor-less Hybrid Step-Down DC-DC Converter Architecture for Future Smart Power Cable," *IEEE APEC*, pp. 247-253, 2017.
[5] G. Gabian et al., "Hybrid Buck Converter Optimization and Comparison for Smart Phone Integrated Battery Chargers," *IEEE APEC*, pp. 2148-2154, 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

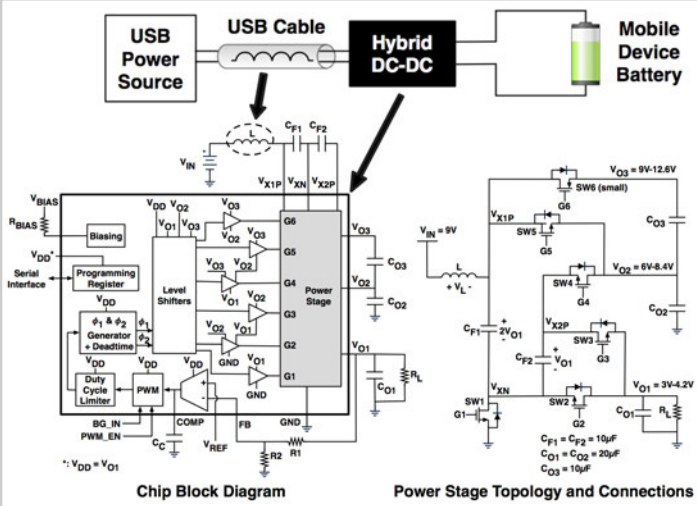


Figure 8.3.1: Block diagram and topology of the proposed hybrid DC-DC converter suitable for USB battery charger applications. The load and regulation are connected to V_{01} in this figure.

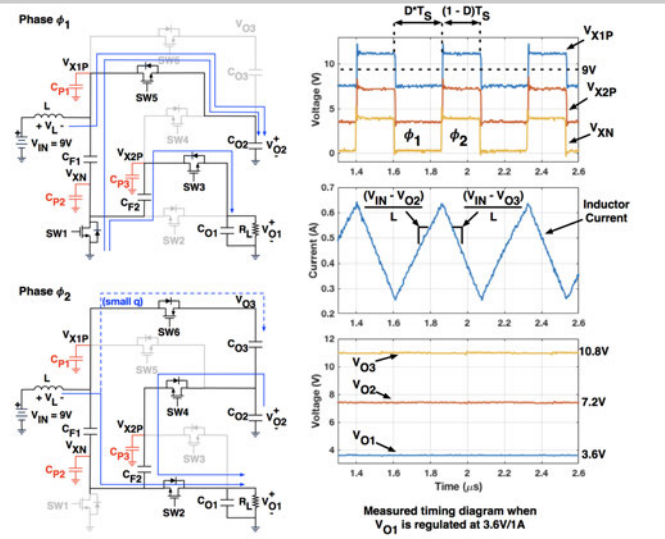


Figure 8.3.2: Power-stage operation and measured steady-state waveforms using a 1μH discrete inductor.

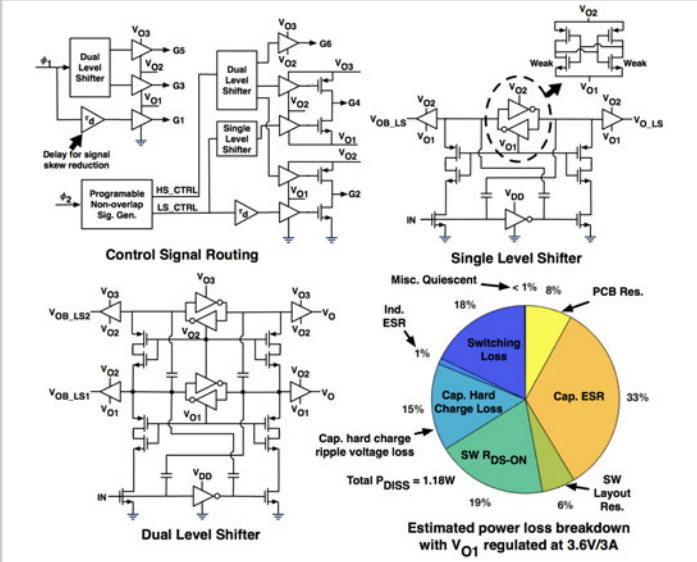


Figure 8.3.3: Circuit sub-blocks and power-loss breakdown.

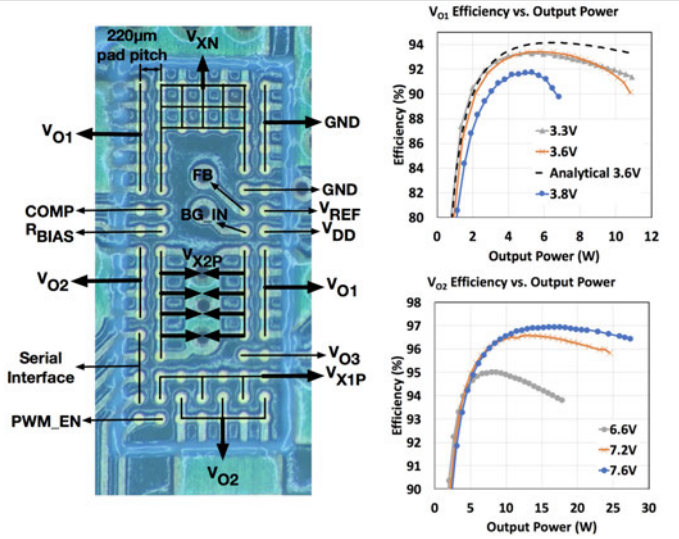


Figure 8.3.4: Flip-chip PCB footprint with annotated signal routing and measured efficiency for V_{01} and V_{02} using a 1μH 5.5mΩ-DCR discrete inductor.

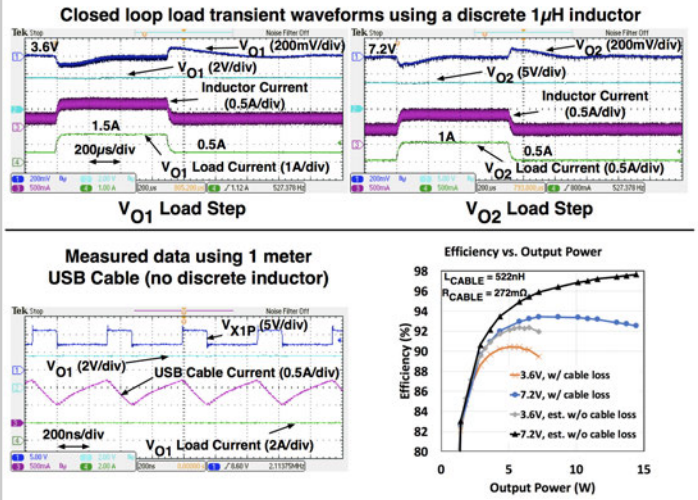


Figure 8.3.5: Load transients for V_{01} and V_{02} using a 1μH discrete inductor and measured performance using a 1-meter USB cable.

	This Work	[1]	[2]	[3]	[4]	[5]
Topology	**Hybrid**	Buck	Hybrid	ReSC	Hybrid	Hybrid
Inductor location	**Input**	Output	Output	Output	Input	Input
Input voltage [V]	**9**	3.9 - 14	3 - 4.5	12	5	5 - 7
Output voltage [V]	**3 - 4.2 (6 - 8.4)**	2.8 - 4.6	0.3 - 1	3.5 - 3.8	3.3 - 3.8	4
Integrated regulation	**Yes**	Yes	Yes	Yes	No	No
Peak output power [W]	**10.9 (27.4)**	19*	1.24*	4.62	15**	20*
Overall peak efficiency w/ discrete inductor [%]	**93.4 (97)^**	92*	94.2	87	—	97*
Peak on-board efficiency w/ USB cable [%]	**92.4 (97.6)^^**	—	—	—	91.5*	—
Peak power density [W/mm²]	**1.48 (3.72)**	—	0.31*	0.47	—	6.1***
Support two-cell charging	**Yes**	No	No	No	No	No
Switching frequency [MHz]	**2.3**	1.32 - 1.68	0.2 - 5	1.7	2	1
Process	**130nm BCD**	—	65nm CMOS	180nm CMOS	Discrete	LDMOS

* Estimated from reported measurement results, ** With USB cable, *** Includes power switch area only
^ 1μH discrete inductor, ^^ 522nH cable inductance

Figure 8.3.6: Comparison with prior art.

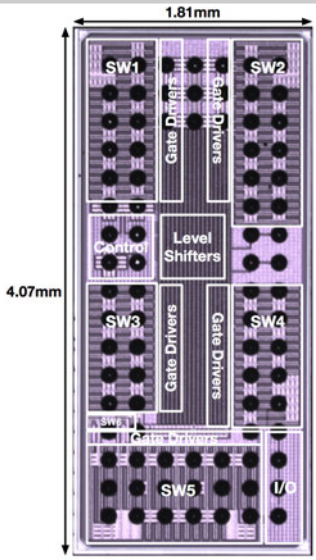


Figure 8.3.7: Annotated die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

8.4 Fully Integrated Buck Converter with 78% Efficiency at 365mW Output Power Enabled by Switched-Inductor-Capacitor Topology and Inductor Current Reduction Technique

Nghia Tang[1], Bai Nguyen[1], Yangyang Tang[2], Wookpyo Hong[1], Zhiyuan Zhou[1], Deukhyoun Heo[1]

[1]Washington State University, Pullman, WA
[2]HiSilicon, Shenzhen, China

Fully integrated buck converters can improve performance and reduce the power consumption of system-on-chip by providing point-of-load regulation with dynamic voltage scaling [1]. As a core component of a buck converter, an inductor with large inductance and small resistance is desirable for high power efficiency, but on-chip inductor integration is challenging due to silicon-area constraints and parasitic effects. Common techniques to integrate on-chip inductors include using on-die spiral inductors [1-3], package bond wires [4], and magnetic cores [5,6], resulting in inductors on the order of nH with resistances of several hundred mΩ. Switching frequencies approaching 100MHz and beyond are often used to reduce the current ripple of such small inductors [1-6] at the cost of switching loss. A special hybrid topology, known as a 3-level converter, has been shown to reduce inductor current ripple and the associated power loss by doubling the effective switching frequency [1]. Compared to the conventional buck converter, the 3-level topology uses two more power switches that contribute toward switching loss and conduction loss. Some other hybrid topologies have demonstrated unique characteristics and benefits [7], but none have been proposed to improve the performance of fully integrated buck converters.

This paper presents a switched-inductor-capacitor (SIC) step-down converter that can improve power efficiency by reducing the average current of on-chip inductors relative to the load current. Figure 8.4.1 illustrates a diagram of the converter consisting of a power stage, switch drivers, and a feedback controller. The core devices of the power stage include inductor L, flying capacitor C_{FLY}, output capacitor C_{OUT}, and three power switches that are controlled in two phases by a PWM signal with frequency f_{SW} and duty-cycle D. In contrast to the conventional buck converter, the SIC topology places the inductor at the input instead of the output, effectively reducing the inductor average current relative to the load current. Thus, for a specific load, the SIC converter endures lower power loss from its inductor. The working principle of the SIC converter in steady-state is shown in Fig. 8.4.2. In phase 1, switches ϕ_{1HI} and ϕ_{1LO} are turned on for a duration of D/f_{SW}, and current I_L through inductor L increases by ΔI_{L1} with slope S_{L1}. In phase 2, switch ϕ_2 is turned on for a duration of $(1-D)/f_{SW}$, and I_L decreases by ΔI_{L2} with slope S_{L2}. The relationship between V_{IN} and V_{OUT} can be derived as $V_{OUT}=V_{IN}/(2-D)$ by applying inductor volt-second balance. Therefore, the voltage conversion ratio, $VCR=V_{OUT}/V_{IN}$, can be adjusted from 1/2 to 1 by controlling the duty-cycle D from 0 to 1. The error amplifier produces an error signal from the difference between V_{OUT} and the reference voltage, V_{REF}. From the error signal and a 450MHz ramp signal, the comparator outputs a PWM signal to drive the power switches such that V_{OUT} is regulated to V_{REF}.

The average inductor current of the SIC converter can theoretically be expressed as $I_{L,DC}=VCR\times I_{OUT}$ because $V_{IN}\times I_{IN}=V_{OUT}\times I_{OUT}$ and $I_{IN}=I_{L,DC}$. In contrast, the average inductor current of the conventional buck converter and the 3-level converter is equal to I_{OUT}. Since VCR is smaller than 1 for a step-down conversion, the SIC converter has lower average inductor current and lower conduction loss than the other two converters. Figure 8.4.3 illustrates the current flows in each of the three converters. All current paths in the 3-level converter include three resistive elements in series, i.e., two power switches and one inductor. In contrast, any current path in the conventional buck converter and the SIC converter only includes two resistive elements in series, i.e., two power switches in series or one power switch and one inductor in series. Thus, the power switches in the conventional buck converter and the SIC converter can be sized smaller for lower conduction loss and lower switching loss. With proper switch sizes and switching frequency, the SIC converter can outperform the conventional buck converter and the 3-level converter in terms of power efficiency.

To verify the concept, all three types of converters using identical inductors were simulated under the same supply voltage and load conditions, and the performance metrics were compared by observing the conduction loss of the inductors and the power switches. The inductors have conduction loss arising from the average current ($I_{L,DC}$) and current ripple (ΔI_L), and the power switches have both conduction loss and switching loss. The buck converter, the 3-level converter, and the SIC converter were designed to operate at 654MHz, 178MHz, and 450MHz, respectively, to have similar inductor current ripple and similar power losses from the switches. Figure 8.4.4 shows the simulated inductor currents, power losses, and power efficiencies of the three converters. The SIC converter outperforms the other two at any VCR between 1/2 and 1 due to the lower average current through the inductor and fewer resistive components in the current paths.

The SIC converter was fabricated in a 65nm CMOS process with an area of 1.3×0.5mm^2, excluding test pads. By using a top metal layer, the inductor was designed to have an inductance of 0.85nH and a resistance of 340mΩ at 450MHz. The flying capacitor C_{FLY} and the output capacitor C_{OUT} were 1.72nF and 3.1nF, respectively, implemented by metal-oxide-semiconductor capacitors and metal-oxide-metal capacitors, with a capacitance density of 10nF/mm^2. The substrate-coupling capacitance of the bottom plate of C_{FLY} is 30pF. The power loss from switching this parasitic capacitance at 450MHz is 2.4mW, 3.3mW, 4.3mW, and 5.5mW at V_{OUT} = 600mV, 700mV, 800mV and 900mV, respectively, which contributes to less than 10% of the total power loss. Charge-sharing loss between the C_{FLY} and C_{OUT} is found to be insignificant, on the order of microwatts. Due to conduction loss of the inductor and power switches, V_{OUT}=600mV can be produced with a duty cycle of 12%. Figure 8.4.5 shows the measurement results, including power efficiency, load transient response, reference tracking, and output voltage ripple at V_{IN}=1.2V. A peak efficiency of 78% is measured at an output voltage of 900mV and a load current of 406mA. The transient response overshoot is between 16.1% for a 3.6× load step at 900mV and 20% for a 6.1× load step at 600mV. The reference tracking response shows a rising time of 40ns and a settling time of 300ns. In steady-state, the output voltage ripple is below 56mV for the entire range of output voltages and load currents. Figure 8.4.6 summarizes the performance of the SIC converter against existing fully integrated buck converters. The SIC converter delivers the highest power and current density while maintaining comparable power efficiency to the best of the class.

Acknowledgements:
This work was supported in part by the U.S. NSF under Grants CNS-1705026, the gift funding from Intel Corporation, and the NSF Center for Design of Analog–Digital Integrated Circuits (CDADIC).

References:
[1] W. Kim, D. Brooks, G.-Y. Wei, "A Fully-Integrated 3-Level DC-DC Converter for Nanosecond-Scale DVFS", *IEEE JSSC,* vol. 47, no. 1, pp. 206-219, Jan. 2012.
[2] H. K. Krishnamurthy et al., "A 500 MHz, 68% Efficient, Fully On-Die Digitally Controlled Buck Voltage Regulator on 22nm Tri-Gate CMOS," *IEEE VLSI Circuits Symp.,* pp. 210-211, 2014.
[3] J. Wibben and R. Harjani, "A High-Efficiency DC-DC Converter Using 2 nH Integrated Inductors," *IEEE JSSC,* vol. 43, pp. 844-854, Apr. 2008.
[4] M. Kar et al., "An All-Digital Fully Integrated Inductive Buck Regulator With A 250-MHz Multi-Sampled Compensator and a Lightweight Auto-Tuner in 130-nm CMOS," *IEEE JSSC,* vol. 52, no. 7, pp. 1825-1835, July 2017.
[5] L. G. Salem and P. P. Mercier, "A Single-Inductor 7+7 Ratio Reconfigurable Resonant Switched-Capacitor DC-DC Converter with 0.1-to-1.5V Output Voltage Range," *IEEE CICC*, pp. 1-4, 2015.
[6] H. K. Krishnamurthy et al., "A Digitally Controlled Fully Integrated Voltage Regulator with On-Die Solenoid Inductor with Planar Magnetic Core in 14nm Tri-Gate CMOS," *ISSCC,* pp. 336-337, 2017.
[7] G. Seo and H. Le, "An Inductor-less Hybrid Step-Down DC-DC Converter Architecture for Future Smart Power Cable," *IEEE APEC,* pp. 247-253, 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

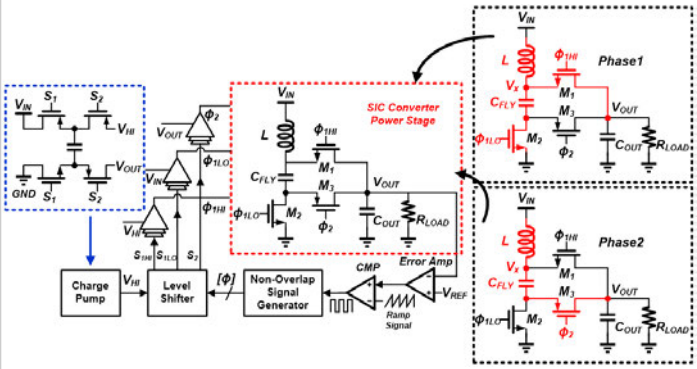


Figure 8.4.1: Block diagram and basic operation of the SIC converter.

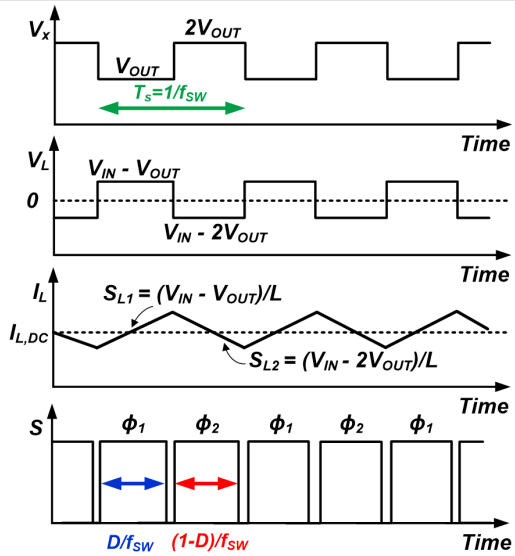


Figure 8.4.2: Steady-state operation of the SIC converter.

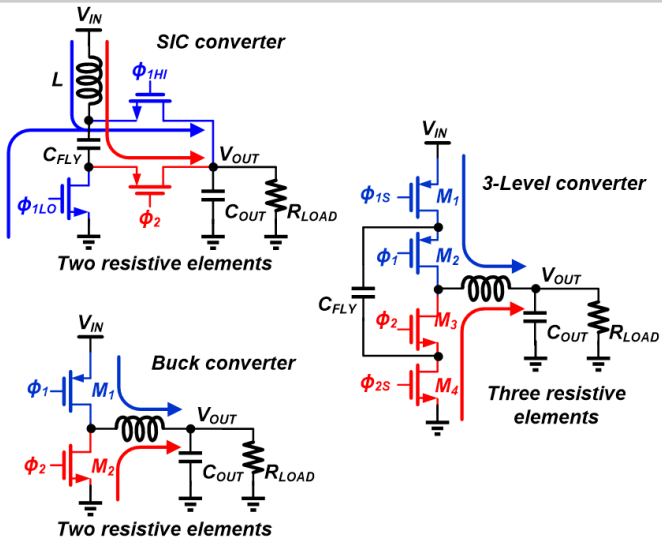


Figure 8.4.3: Current paths in the SIC converter, the 3-level converter, and the buck converter.

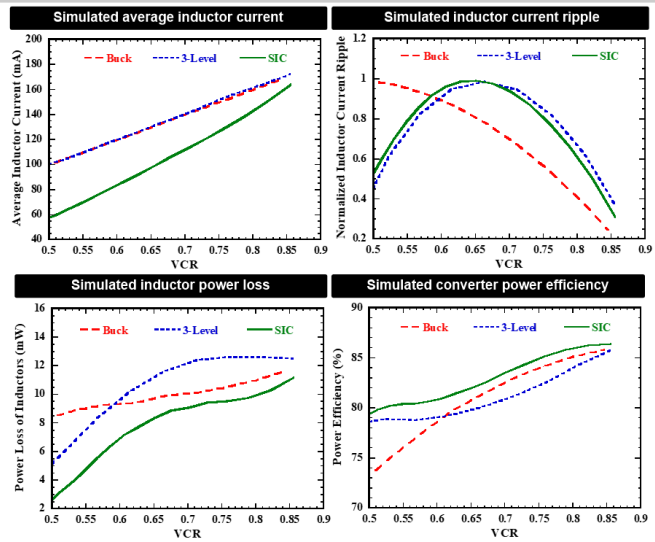


Figure 8.4.4: Comparison of the inductor current, current ripple, and power loss, and the power efficiency of the three converters in simulation.

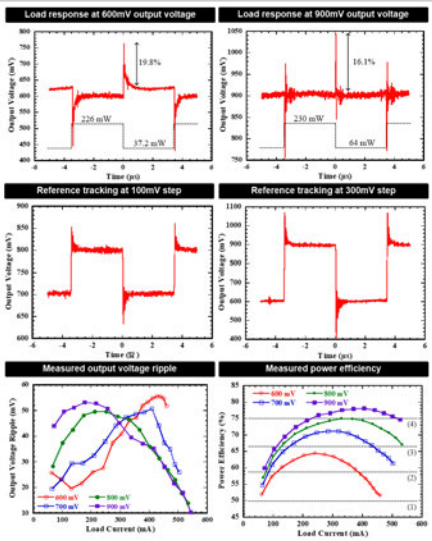


Figure 8.4.5: Measured load response, reference tracking, output voltage ripple, and power efficiency at V_{IN}=1.2V.

	[1]	[2]	[3]	[5]	This Work
Process (nm)	130	22	130	180	**65**
Topology	3-Level	Conventional Buck	Conventional Buck	$R^2SC^\star$	**SIC**
Supply Voltage (V)	2.4	1.5	1.2	1.8	**1.2**
Output Voltage (V)	0.4 – 1.4	1	0.3 – 0.9	0.1 – 1.5	**0.6 – 0.9**
Frequency (MHz)	50 – 240	500	170	1 - 100	**450**
Number of Phases	4	1	1	1	**1**
Total Ind. (nH)	4	1.5	4	18	**0.85**
Total Cap. (nF)	28	10	5.2	3	**4.82**
Max Load Current (mA)	800	250	350	120	**533**
Max Output Power (mW)	1000	250	315	210	**475**
Output Ripple (mV)	< 100	~ 10	~ 40	NA	**< 56**
Area (mm^2)	5	1.5	1.5	NA	**0.65**
Power Density (mW/mm^2)	200	167	210	210	**730**
Current Density (mA/mm^2)	160	167	233	140	**820**
Eff. @ Max Power (%)	63	N/A	76	NA	**74.6**
Peak Eff. (%)	77	68	77.9	73.3	**78**
EEF**	0.15 – 0.5	0.02	0 – 0.56	-0.06 – 0.36	**0.04 – 0.22**

* Recursive resonant switched capacitor converter (R^2SC) with an on-package inductor

** Efficiency enhancement factor: $EEF = 1 - Efficiency_{LDO}/Efficiency_{Converter}$

Figure 8.4.6: Performance summary of the SIC converter and comparison with prior art.

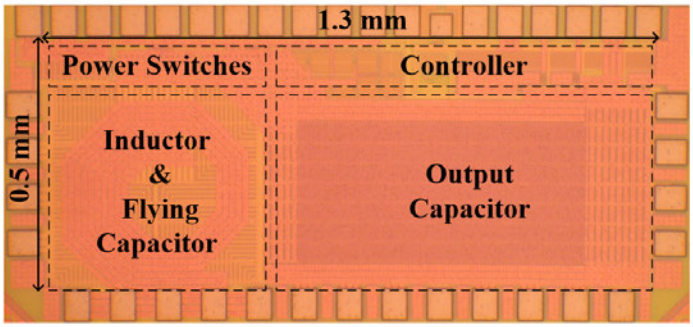


Figure 8.4.7: Die micrograph of the SIC converter fabricated in a 65nm CMOS process.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

8.5 A Fully Integrated Voltage Regulator in 14nm CMOS with Package-Embedded Air-Core Inductor Featuring Self-Trimmed, Digitally Controlled Variable On-Time Discontinuous Conduction Mode Operation

Christopher Schaef[1], Nachiket Desai[1], Harish Krishnamurthy[1], Sheldon Weng[1], Huong Do[2], William Lambert[2], Kaladhar Radhakrishnan[2], Krishnan Ravichandran[1], James Tschanz[1], Vivek De[1]

[1]Intel, Hillsboro, OR
[2]Intel, Chandler, AZ

Fully Integrated Voltage Regulators (FIVR) with package-embedded air-core inductors [1] or on-die solenoid inductors with planar magnetic core [2] promise efficient power delivery and fine-grain wide-range DVFS in complex SoCs while providing fast transient response. The FIVR must provide high conversion efficiency across a wide operating range of output voltages and load currents, including light to medium loads, to maximize the overall energy efficiency of the SoC across different power states. Phase shedding and switch scaling have been used for high-frequency FIVR designs with pulse-width modulation (PWM) control in continuous conduction mode (CCM) to maintain high efficiency for large load currents [1-5], and pulse-frequency modulation (PFM) and hysteretic control have been used to achieve high efficiency across light to medium loads [3-5]. In this paper, we present an FIVR in 14nm CMOS with a 2.5nH air-core inductor embedded in an ultrathin coreless package (200μm thick) (Fig. 8.5.7), featuring self-trimmed, soft-switched and digitally controlled variable ON-time DCM operation up to 70MHz to achieve high conversion efficiencies across light to medium load currents ranging from 5mA to 500mA and wide 0.7-1.2V output voltage range. The FIVR uses a cascoded thin-gate powertrain (Fig. 8.5.1) to support input voltages up to 2Vmax with the cascode bias rail set at $V_{in}/2$ which consumes <1uA at light load. A small thick-gate device is connected across the inductor to dampen oscillations when the power stage is in a high-impedance state. The output voltage is monitored by a comparator with sub-ns response time which triggers an inductor current pulse when the output drops below the reference voltage. A resistor divider with a feedforward capacitor is used to achieve fast response time.

One of the key challenges for efficient DCM operation at high switching frequencies is fast and accurate detection of the zero-crossing of the inductor current. A comparator across the low-side NMOS switch is commonly used for this purpose. However, low-power comparators, especially in scaled process nodes, exhibit large random offsets compared to the tens of mV voltage drop across the switch. This can degrade efficiency and create EMI/RFI issues as well. In addition, compensating for the comparator and gate driver delays pose major challenges for accurate and efficient zero current detection (ZCD). Since the current in a 2.5nH inductor can undergo large negative excursions in a very short period of time, the conversion efficiency is governed strongly by the accuracy and speed of the ZCD. While the ZCD comparator offset and delay need to be small, its overall power consumption needs to be minimized as well in order to achieve high conversion efficiency at light loads (<10mA).

The ZCD comparator (Fig. 8.5.2) implemented here uses power gating, auto-zeroing and digital self-trimming to reduce power consumption while maintaining high accuracy and speed. The comparator is power gated between DCM pulses to save bias currents. During this idle period, only the input bias branch is kept ON to enable fast transition to active state. Once an inductor current pulse is initiated, the comparator enters an auto-zero mode while the high-side PMOS power switch is turned on. In this mode, an internal compensation network for the comparator is activated to allow stable operation in feedback and sample the offset voltage onto the offset-storage capacitor at the input. After the high-side PMOS switch turns off, the compensation network and feedback connection are disabled, and the comparator enters the "compare" mode.

Furthermore, an additional capacitor, C_{trim}, is pre-charged through a 5b capacitive DAC to a digitally controlled voltage. During the high-to-low side transition, C_{trim} is connected to the offset-storage capacitor to introduce a small controllable offset. This offset is controlled through an additional loop which uses the residual current detector (Fig. 8.5.2) to adjust the comparator trip-point, thereby correcting for any remaining offset and circuit delays. Positive or negative residual current can be detected by sampling the voltage at V_x after the low-side switch has turned off; a negative/positive residual current will lead to a positive/negative voltage at V_x. Measurements (Fig. 8.5.2) demonstrate convergence of the loop to a very low residual current after the offset register is reset.

An additional challenge for high frequency DCM operation stems from the resonance effects of the distributed output power delivery network and on-package ceramic capacitors (Fig. 8.5.3). This can cause undesirable retriggering of PFM pulses, as shown from measurements in Fig. 8.5.3, resulting in significantly higher output voltage ripple under light load conditions. In order to suppress retriggering, we have implemented a programmable forced off-time that prevents a new pulse from being triggered for a certain time period after the previous pulse has ended. Measurements show that ~1ns off-time is sufficient to prevent double pulses (Fig. 8.5.3).

Constant ON-time DCM operation causes large variations in inductor peak current across a wide output voltage range of 0.7 to 1.2V, thus degrading conversion efficiency at low output voltages when the inductor peak current far exceeds the nominal target value. For example, an ON-time of 7.5ns limits the inductor peak current to 1.2A at the maximum 500mA output load at 1.2V. At 0.7V output voltage, the same ON-time causes the inductor peak current to increase to 2.7A, thus impacting efficiency. We have implemented a digitally controlled ON-time generator that utilizes the digital input/output voltage commands available from the SoC power management unit to compute the correct ON-time for the specific operating load range and inductor peak current target. Measurements demonstrate that the inductor peak current is held approximately constant after an output voltage change (Fig. 8.5.4), as evidenced by a constant pulse frequency before and after the voltage transition, as opposed to a constant ON-time scheme. In addition, the conversion efficiency measurements (Fig. 8.5.5) across a wide range of output voltages demonstrate the effectiveness of the proposed digitally controlled variable ON-time scheme, compared to a constant-ON time implementation.

Transient performance of the FIVR control loop for both reference voltage steps and load current transients are measured (Fig. 8.5.4). Fast FIVR response, enabled by the high-speed comparator, to a 200mA loading and unloading transient is demonstrated using an on-chip load with sub-ns turn-on time.

Figure 8.5.5 shows the FIVR conversion-efficiency measurements across a wide range of output voltages and load currents, as well as the major loss components. At 500mA load current, efficiencies of 88%, 82% and 75% are achieved for output voltages of 1.2V, 1V and 0.8V respectively. The efficiency is fairly constant across 5mA-to-500mA loads since the PFM controller power overhead of 33uA is low, but it drops off significantly below 5mA. As illustrated by the measured spectral components of the inductor current and the inductor ac resistance characteristics (Fig. 8.5.5), this efficiency degradation is largely caused by the ac-resistance loss in the low quality-factor air-core inductor embedded in the coreless ultrathin package (see Fig. 8.5.7) becoming the dominant loss component under light load conditions. For higher loads, a larger portion of the inductor current spectrum is at dc, thus causing inductor losses to be lower. At light load the dc component becomes small and the ac-resistance losses are dominant. Comparisons with previously reported FIVRs are shown in Fig. 8.5.6.

Acknowledgements:
The authors would like to thank Ye Seul Nam, Jimin Yao and Shawna Liff for assembly, Eduardo Quijano Centeno, Dan Lake and Sheldon Sun for test platform design, Tim Nolen and Linda Low for mask design, Satish Yada, Changgeng Xi for integration and Sally Safwat Amin for her internship contributions.

References:
[1] E. A. Burton et al., "FIVR — Fully Integrated Voltage Regulators on 4th Generation Intel® Core™ SoCs," *APEC,* pp. 432-439, 2014.
[2] H. K. Krishnamurthy et al., "A Digitally Controlled Fully Integrated Voltage Regulator with On-Die Solenoid Inductor with Planar Magnetic Core in 14nm Tri-Gate CMOS," *ISSCC,* pp. 336-337, Feb. 2017.
[3] H. K. Krishnamurthy et al., "A 500 MHz, 68% Efficient, Fully On-Die Digitally Controlled Buck Voltage Regulator on 22nm Tri-Gate CMOS," *IEEE Symp. VLSI Circuits,* pp. 1-2, 2014.
[4] H. K. Krishnamurthy et al., "A Digitally Controlled Fully Integrated Voltage Regulator with 3D-TSV Based On-Die Solenoid Inductor with Backside Planar Magnetic Core in 14nm Tri-Gate CMOS," *IEEE Symp. VLSI Circuits,* pp. C148-C149, 2017.
[5] S. S. Kudva and R. Harjani, "Fully Integrated On-Chip DC-DC Converter with a 450× Output Range," *IEEE JSSC,* pp. 1940-1951, 2011.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

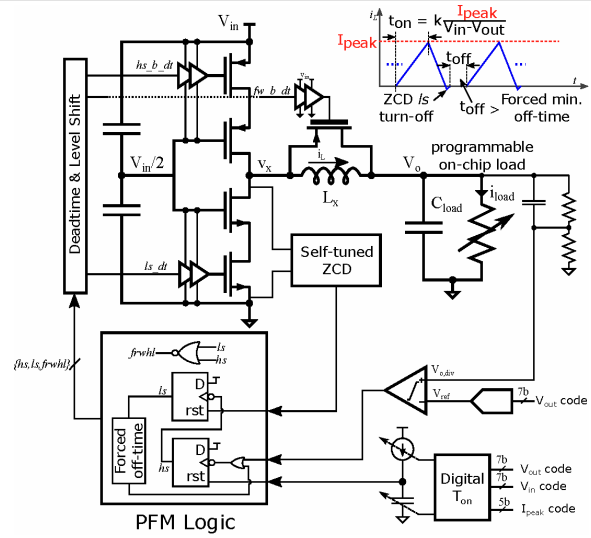


Figure 8.5.1: System Diagram showing cascoded powertrain and main controller blocks.

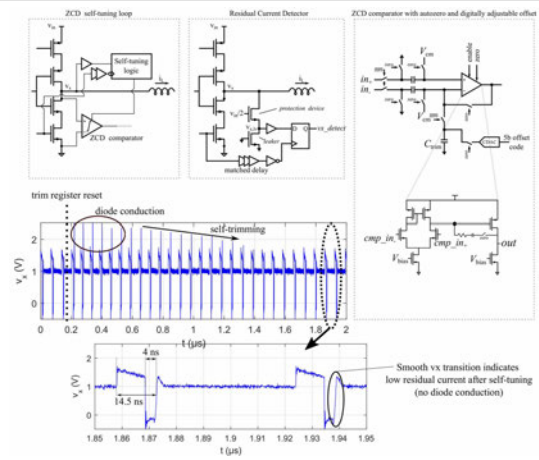


Figure 8.5.2: Circuit diagrams of the Zero-Current Detection block with main comparator (right), residual current detector (center) and digital offset-tuning loop (left). The bottom graph shows measured waveforms of the switching node voltage V_x (probed on-package) during convergence of the self-tuning loop after resetting the offset register.

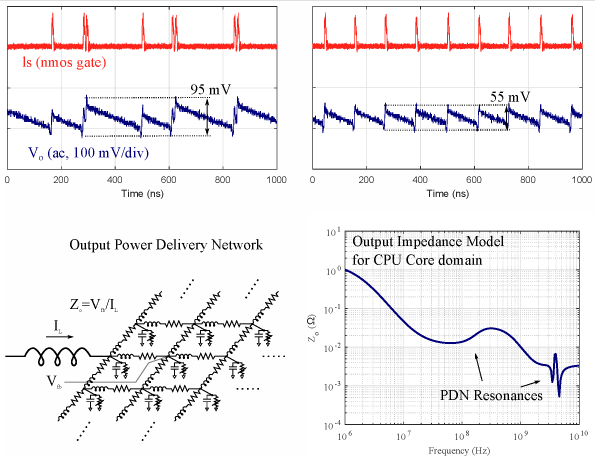


Figure 8.5.3: Measurement data of output voltage ripple in light-load PFM operation. Double-triggering can be observed (top left) due to PDN-resonance effects. An impedance plot of an actual CPU core PDN model shows source of resonant effects. Programmable off-time is effective in preventing retriggering (top right).

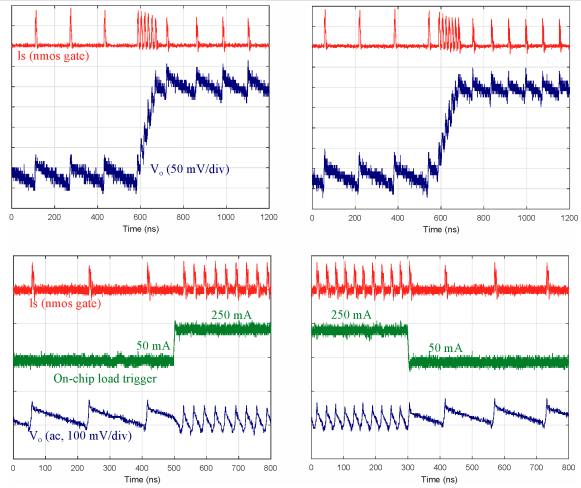


Figure 8.5.4: Transient measurement waveforms for reference steps with (top left) and without (top right) the automatic on-time adjustment. Output current loading (bottom right) and unloading (bottom left) transient waveforms created with high speed on-chip load (50ps rise time).

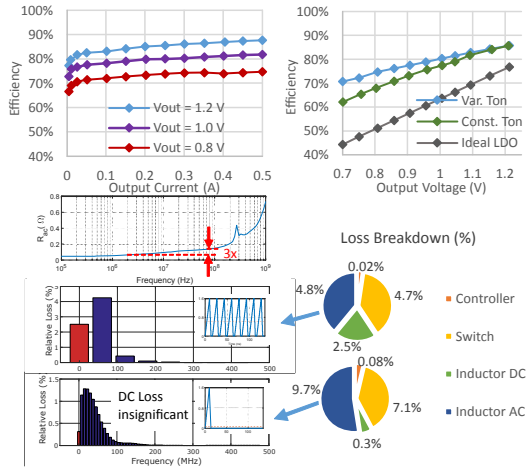


Figure 8.5.5: Measured efficiency data vs load current for different output voltage (top left). Efficiency vs output voltage for constant and variable (auto-adjusted) on-time. Inductor power-loss spectrum shows higher ac-loss components for lighter load conditions.

	[4]	[5]	[1]	[2]	This Work
Process	14 nm	130 nm	22 nm	14 nm	14 nm
Inductor Tech.	Magnetic (3D-TSV)	Air-core (on-chip)	Air-core (in-package)	Magnetic (on-chip)	Air-core (in-package)
V_i / V_o	1.2 / 0.9 V	1.2 / 0.88 V	1.7 / 1.05 V	1.5/1 V	1.6 / 1.2 V
I_o	25 mA	30 mA	16 A	360 mA	500 mA
Inductance	5 nH	2 nH	- nA	1.5 nH	2.5 nH
On-die C_o	6 nF	5 nF	112 nF	5 nF	10 nF*
Eff@Imax	80 %	56 %	90 %	80 %	88 %
Eff@1% Imax	- (nA)	- (nA)	<40%	<36%	72 %
Control	PWM/Hys/PFM	PFM	CCM	CCM	PFM (DCM)
Frequency	90 MHz	200 MHz	140 MHz	100 MHz	70 MHz

*additional on-package decoupling on load domain: 150nF

Figure 8.5.6: Comparison table with previously published integrated voltage regulators.

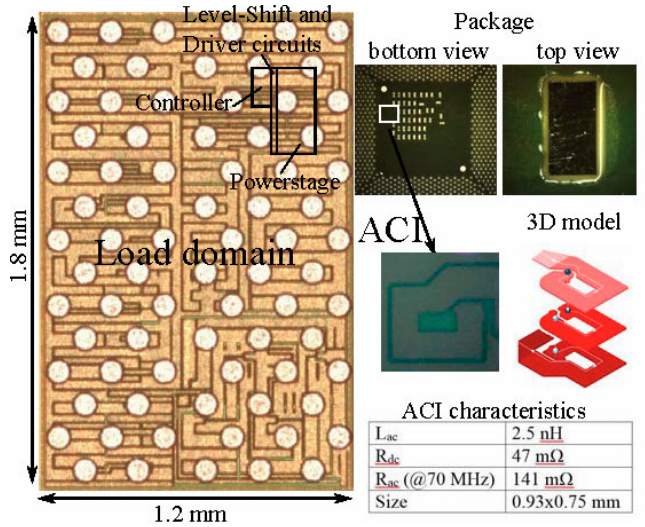


Figure 8.5.7: Die micrograph with main FIVR blocks highlighted along with illustration of the in-package 3-turn air-core inductor.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

8.6 A Fully Integrated 85%-Peak-Efficiency Hybrid Multi-Ratio Resonant DC-DC Converter with 3.0-to-4.5V Input and 500μA-to-120mA Load Range

Peter Renz[1], Maik Kaufmann[1], Michael Lueders[2], Bernhard Wicht[1]

[1]Leibniz University Hannover, Hannover, Germany
[2]Texas Instruments, Freising, Germany

DC-DC converters for applications like wearables require an ultra-compact and flat module size. Hybrid converters are a promising converter class that supports integration of inductive and capacitive components, while minimizing losses and improving power density. Since many of these applications are often in sleep mode, high efficiency has to be achieved from high to low output power. Fully integrated 3-level buck converters [1,2] do not maintain good efficiency over the full load range, since they operate at high switching frequencies, required for inductive PWM operation with small inductors. The converter in [3] achieves better efficiencies, but has a small conversion ratio. Resonant SC converters [4,5] reach high efficiencies due to lower switching frequencies and better passive component utilization. However, low-power operation is not supported [1,3-5] or it suffers from low efficiencies [2]. Moreover, only [4] supports a wide input voltage range based on mixed inductive and resonant operation, but still has an efficiency drop over varying input voltages. Hybrid converters, combining an SC cell with an external LC output filter, operated with inductive PWM control [6], achieve high efficiencies over wide input voltage, but they do not support low-power operation.

This paper presents a fully integrated hybrid DCDC converter, shown in Fig. 8.6.1. A resonant multi-ratio SC power stage with three different conversion ratios (2/3, 1/2, 1/3) supports an input voltage range of 3.0 to 4.5V, covering the full Li-Ion range, and a wide load current range of 500μA to 120mA. The target output is set between 1.5 and 1.8V and can be scaled to 1.3V and lower. All passive components are fully integrated on-chip, including the buffer capacitors C_{in}, C_{out} and the power inductor L. In order to maintain both high efficiency and low output voltage ripple over a wide input and load range, this work utilizes switch-conductance regulation (SwCR). Therefore, each of the power switches consists of a segmented 8-bit binary weighted 2×1.5V transistor stack, level shifters, logic and a charge pump (CP) for gate-overdrive generation. A relaxation oscillator generates the resonance switching frequency $f_{sw,res}$ depending on the conversion ratio. In order to scale the frequency-depended charge-pump losses, the CP control block varies the CP switching frequency according to the Sw<0:7> value of the segmented switches, which is the load of the CP.

The wide input voltage range in combination with wide output power is achieved by a resonant multi-ratio control, Fig. 8.6.2. For the 1/2 SC ratio, the two flying capacitors are always in parallel. For the 1/3 and 2/3 ratio the capacitors are connected in series during one of the phases, resulting in a factor 4/3 higher resonance frequency and different duty cycle. This is automatically adjusted by a switchable current source in the relaxation oscillator. The fine regulation controls the equivalent output resistance R_{out} by varying the switch resistance R_{sw} and the switching frequency f_{sw} as shown in Fig. 8.6.2 (right). Measured data points confirm the expected resonant operation with the on-chip spiral inductor. As a main benefit, the converter offers zero current switching and efficient resonant charging, operating at resonance frequency $f_{sw,res}$. This also results in a small on-chip C_{in}. Moreover the converter maintains a continuous current flow, which results in low output voltage ripple. Switching losses are minimized by varying the width of the power switches depending on the load (SwCR). 8-bit resolution is needed to achieve a V_{out} control deviation of less than 10mV. At low output currents (higher R_{sw}), the converter regulates the switching frequency f_{sw} instead SwCR, because the R_{out} curve, Fig. 8.6.2 (right), approaches the characteristics of a conventional SC converter. This reduces bottom-plate losses and thereby maintains high efficiency at low power. The transition to SC mode happens when all bits Sw<0:7> reach zero. In SC mode, all switch segments are off, except for one remaining offset segment.

Figure 8.6.3 shows the implementation of the segmented power switches and driver circuits. The low-side switches can be directly controlled with the logic signals. A charge-pump CP generates the gate overdrive voltage for each high-side switch. Most signals within the CP are self-timed, only one GND-referred clock signal φ_{CP} is needed. The clock signal is transmitted to the high-side domain by a fast capacitive level shifter, which results in very energy efficient high-frequency signal shifting. In contrast, pulsed cascode level shifters are used for SwCR, since the controller sets the switch conductance information Sw<0:7> at a rate, which is far below f_{sw}. These level shifters are optimized to have ~8× lower static losses compared to the fast dynamic level shifters.

Stacked 2×1.5V low voltages transistors ($V_{DS,max}$=4.4V) result in lower losses and smaller layout area than 5V NMOS/PMOS switches. Even though the stacked 2×1.5V power switch solution requires a dedicated CP for gate-overdrive generation with additional losses, it has factor of 2 and 3.6 lower total energy loss E_{loss} compared to a 5V NMOS/PMOS switch, respectively, shown in Fig. 8.6.4 (top). Losses at the stacked 2×1.5V switches do not scale down linearly with the number of activated switches, because of the voltage swing at the parasitic capacitances of the deactivated segments, depicted in Fig. 8.6.4 (bottom). This work proposes two additional switches S_G and S_{GS}, which decouple the gate of the upper transistor and short-circuit the capacitance C_{GS2}. When a switch segment is inactive, S_G is off and S_{GS} is on. This leads to ~2× lower effective capacitance and significantly lower charge redistribution losses. At low output current, this results in significant efficiency improvement, e.g. 4% at 30mA.

The converter was fabricated in a 0.13μm BCD process. Two different options for the passives have been implemented, a fully integrated option with an on-chip spiral inductor of 9nH and a highly integrated option with a 10nH SMD inductor placed on top of the die, as shown in Fig. 8.6.7. The on-chip inductor has a DC resistance of R_{DC}=280mΩ and a quality factor of Q=6 at 35.5MHz. 1nF MIM capacitance is used for each C_{fly1} and C_{fly2}. An underlying series of a high-resistively biased p-well and n-well minimizes the influence of the parasitic bottom-plate capacitance. A 180pF MIM capacitance is used for C_{in} and a 10nF MOS capacitor for C_{out}.

The measured efficiency in Fig. 8.6.5 shows a flat run over the output current, which is due to effective scaling of the switching losses by the SwCR. High efficiency at very low power is maintained by dual-mode operation, which enters SC mode below 20-to-30mA load. The design achieves a peak efficiency of 85% with the integrated 9nH planar inductor and 88.5% with a 10nH in-package inductor at power densities of 0.033W/mm² and 0.054W/mm². The converter operates at resonance switching frequencies of up to 35.5MHz (1/2 ratio) and 47.5MHz (x/3 ratio). Waveforms of V_{sw}, φ1 and φ2 (bottom center) confirm the resonant operation, in line with the curves in Fig. 8.6.2. As expected, due to SwCR a small output voltage ripple <35mV is achieved (50mV in low-power), shown in Fig. 8.6.5 (bottom left). The design was successfully tested down to V_{out}=1.2V (1/3 ratio). The comparison with prior art in Figs. 8.6.5 (bottom right) and 8.6.6, show that this fully integrated converter achieves a superior peak efficiency, a 19× wider dynamic load-current range, and a 12.8% better efficiency enhancement factor EEF. The wide input voltage range is fully suitable for Li-Ion operation.

References:
[1] W. Kim et al., "A Fully-Integrated 3-Level DC-DC Converter for Nanosecond-Scale DVFS," *IEEE JSSC*, vol. 47, no. 1, pp. 206-219, Jan. 2012.
[2] W. Godycki et al., "Part-Time Resonant Switching for Light Load Efficiency Improvement of a 3-Level Fully Integrated Buck Converter," *ESSCIRC,* pp. 163-166, Sept. 2014.
[3] H. K. Krishnamurthy et al., "A Digitally Controlled Fully Integrated Voltage Regulator with On-Die Solenoid Inductor with Planar Magnetic Core in 14nm Tri-Gate CMOS," *ISSCC*, pp. 336-337, Feb. 2017.
[4] C. Schaef et al., "A Variable-Conversion-Ratio 3-Phase Resonant Switched Capacitor Converter with 85% Efficiency at 0.91W/mm² Using 1.1nH PCB-Trace Inductors," *ISSCC*, pp. 1-3, Feb. 2015.
[5] K. Kesarwani et al., "A 2-Phase Resonant Switched-Capacitor Converter Delivering 4.3W at 0.6W/mm² with 85% Efficiency," *ISSCC*, pp. 86-87, Feb. 2014.
[6] W. C. Liu et al., "A 94.2%-Peak-Efficiency 1.53A Direct-Battery-Hook-Up Hybrid Dickson Switched-Capacitor DC-DC Converter with Wide Continuous Conversion Ratio in 65nm CMOS," *ISSCC*, pp. 182-183, Feb. 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

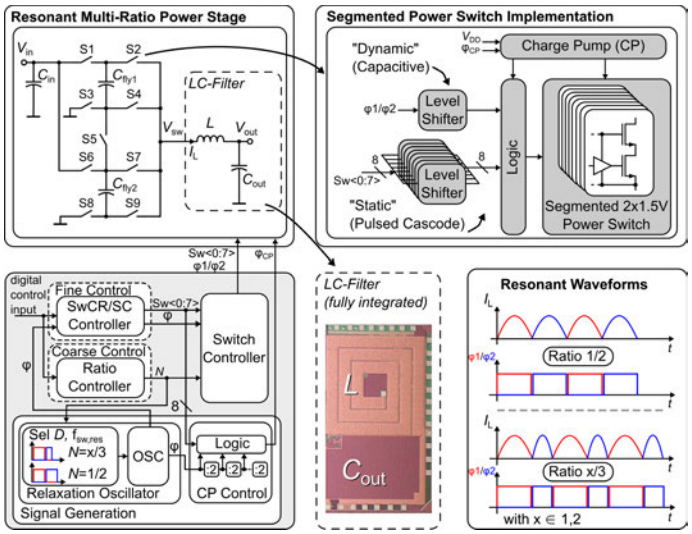


Figure 8.6.1: Block diagram of the resonant multi-ratio SC converter.

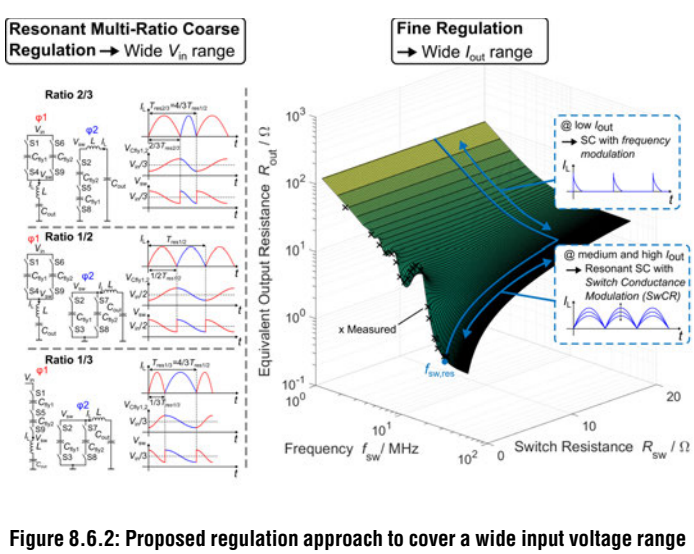


Figure 8.6.2: Proposed regulation approach to cover a wide input voltage range and a wide load range.

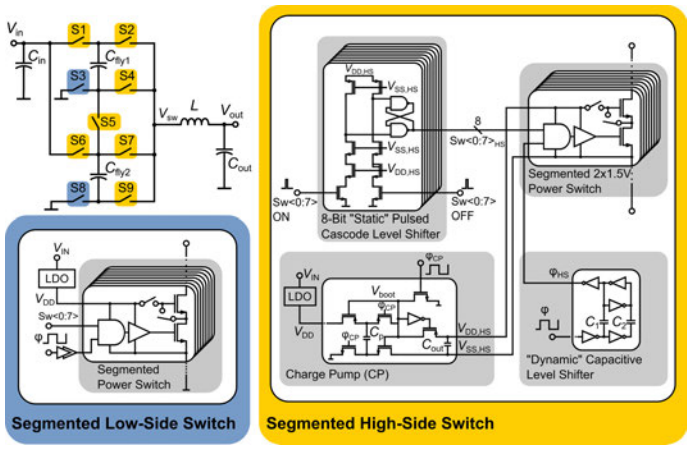


Figure 8.6.3: Implementation of segmented power switches and necessary driver circuits.

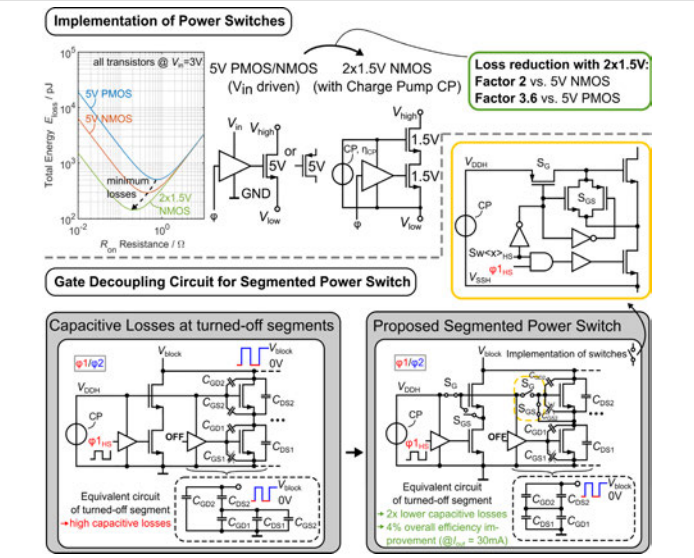


Figure 8.6.4: Implementation of power switches for high-efficiency operation.

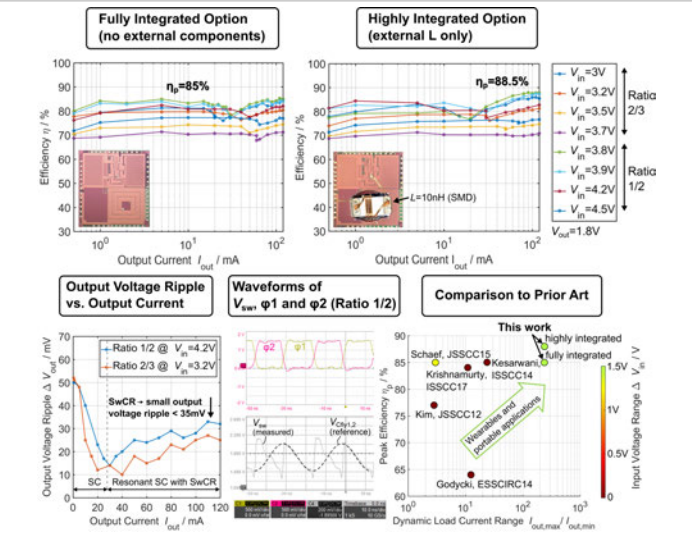


Figure 8.6.5: Efficiency measurements, output voltage ripple vs. output current, Waveforms of V_{sw}, ϕ1, ϕ2 and comparison to prior art.

	Fully Integrated Converters				Highly Integrated		
	Kim, JSSC12	Godycki, ESSCIRC14	Krishnamurthy, ISSCC17	**This work (no ext. comp.)**	Schaef, ISSCC15	Kesarwani, ISSCC14	**This work (ext. L)**
V_{IN} / V	2.4	1.8	1.5	**3.0-4.5**	3.7-5	6/3.6	**3.0-4.5**
V_{OUT} / V	0.4-1.4	0.7	0.9-1.15	**1.5-1.8**	1.2-2.5	3/1.8	**1.5-1.8**
I_{out} / mA	300-850	16-200	30-330	**0.5-120**	500-1.5e3	50-1.2e3	**0.5-120**
f_{sw} / MHz	50-200	N/A	100	35-50	20-50	20-40	35-50
C_{fly}	18nF	1.5nF	n.a.	2nF	24nF	18nF	2nF
C_{out}	10nF	1.9nF	5nF	10nF	6nF	11nF	10nF
C_{in}	1nF	not reported	1nF	**180pF**	6nF	11nF	**180pF**
Resonant operation	no	no	no	**yes**	yes	yes	**yes**
Technology	130nm	65nm	14nm	130nm	180nm	180nm	130nm
Ratio	n.a	1	1	**3**	1	1	**3**
L_{total}	4nH	3nH	1.5nH	**9nH**	3x5.5nH (external)	11nH (external)	10nH (external)
η_p / % @ Conversion Ratio @ V_{OUT}	77 @2.2 @1.1V	64 @2.6 @0.7V	84 @1.3 1.15V	**85 @2.1 @1.8V**	85 @2 @2.5V	85 @2 @3V	**88.5 @2.1 @1.8V**
EEF @ η_p / %**	40.5	39.2	8.7	**45.7**	42.7	41.2	**47.9**
Fully-integrated	yes	yes	yes	**yes**	no	no	**no**
Load current dynamic range*	2.83	12.5	11	**240**	3	24	**240**

* $I_{out,max}/I_{out,min}$ (with η > 60% or $I_{out,min}$ specified) ** $1-(\eta_{LDO}/\eta_{conv})$

Figure 8.6.6: Comparison to state-of-the-art publications.

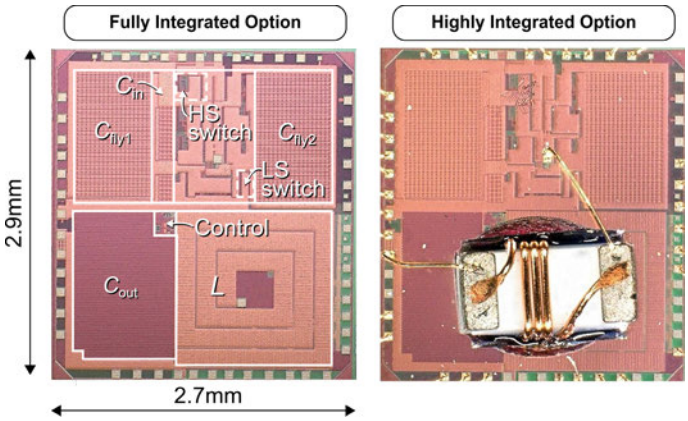


Figure 8.6.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

8.7 A 2MHz 4-to-60V_{IN} Buck-Boost Converter for Automotive Use Achieving 95% Efficiency and CISPR 25 Class 5 Standard

Jing Xue, Min Kyu Song, Xugang Ke, Min Chen, Leonard Shtargot

Analog Devices, Milpitas, CA

High-voltage highly efficient switching power converters have gained high popularity in automotive applications. Powered by the car battery (V_{IN}), the converters need to experience a wide V_{IN} range from 4V to 60V under harsh cold-crank and load dump environments. This requires the converter to accomplish a fine output voltage (V_{OUT}) regulation, seamless mode transition for V_{IN} above, below or equal to V_{OUT}. Conventional buck or boost converters can only achieve monotonic step-down or step-up conversion. Other solutions such as standalone boost and buck, or SEPIC converters [1,2] either require multiple cascaded power stages or suffer from lower system efficiency, larger footprint and higher bill-of-material cost. The buck-boost converter thusly becomes a promising topology by its rather simple power-stage structure and wide conversion ratio [3-5]. However, the electromagnetic interference (EMI), power loss, and complicated control scheme are challenging with its inherent four-switch operation.

As the switching frequency increases, the high di/dt and dv/dt switching transitions incur severe EMI noise, which remains as a major roadblock to meet the strict EMI standard in automotive applications [6]. The switching ON and OFF actions of power converters are critical to the EMI. To mitigate this issue, a silent-switcher-enabled switching buck-boost controller is presented, as shown in Fig. 8.7.1. When switchers are alternating ON and OFF, current flows in a loop that consists of switches (S_A, S_B and S_C, S_D) and the respective decoupling capacitors. The di/dt of this loop is limited only by parasitic inductances, thus much greater than the di/dt of the inductor current (I_L) ripple, resulting in a significant impact on the EMI performance of the complete circuit. One way to reduce the EMI radiation is to minimize this hot loop, however, PCB areas, package size, and thermal design requirements causes certain limitations to the minimum hot-loop size. Other solutions include use of a metal shield or slowing down the switching strength, but both cause undesirable trade-offs such as size, efficiency, and range of the conversion ratio.

The proposed design exhibits the effect of the metal shield but without the actual shield, overcoming the above challenges. The key idea is as follows: by positioning two symmetrical decoupling capacitors closely to both V_{IN} and V_{OUT} sides, the dual hot-loops are formed, as shown in Fig. 8.7.1. Consequently, the large di/dt-induced current flowing though the hot loop can be equally divided and the divided currents will flow in opposite directions to each other, thus enabling magnetic radiation cancellation among themselves.

To realize the above architecture, the proposed controller implements a novel direct I_L sensing scheme shown in Fig. 8.7.1. By removing a sense resistor typically placed between the bottom switches (S_B, S_C) to ground [5], the dual hot-loops can be formed in reverse directions that enables the cancellation effect mentioned above. Additionally, the hot loop itself can be further minimized, resulting in improved overall EMI performance. On the other hand, with the minimized hot loops and ground impedance, along with the significantly reduced parasitic inductance across the power switches/capacitors, the ring at switching nodes (V_{SW1}, V_{SW2}) can be effectively suppressed to realize high reliability and low EMI for automotive use. Furthermore, a balanced peak-current mode buck-boost controller is implemented in the proposed design. This scheme allows the seamless transitions to a pure two-switching operation, which effectively improves both the EMI and the efficiency.

Figure 8.7.2 illustrates the operating principle of the balanced peak-current mode buck-boost controller. When $V_{IN} << V_{OUT}$, the converter operates in peak-boost current mode in the boost region. Only two switches, S_C and S_D are used by the peak-boost current control to regulate V_{OUT}, while S_A is ON and S_B is OFF to minimize the switching loss from S_A and S_B. As V_{IN} increases, the energizing phase (AC) keeps decreasing while the de-energizing phase (AD) keeps increasing. When $V_{IN} <= V_{OUT}$, the conventional boost converter cannot regulate V_{OUT} due to practical issues. Therefore, a fixed BD phase is inserted in to balance the AC and AD phase, and the converter operates in peak-boost current mode in the buck-boost region. When $V_{IN} = V_{OUT}$, the duration of the AC and BD phase are the same, and the proposed converter is switched from peak-boost to peak-buck current mode. In this mode, S_C and S_D are switched under the fixed duty ratio while S_A and S_B are switched by the peak-buck current control. As V_{IN} increases, the converter operates in peak-buck current mode in the buck region. S_C is OFF and S_D is ON all the way and the switching loss from S_C and S_D are minimized. The hysteresis window is designed to enhance the converter's robustness during mode transition. Figure 8.7.2 also emphasizes the energy transferred to the output. In contrast to conventional AC-BD operation [3], the proposed converter ensures the maximum energy transferred to the output by including the AD phase. The larger the AD phase, the more energy can be transferred from V_{IN} to V_{OUT} directly. Therefore, the amplitude of I_L and corresponding conduction loss can be minimized in the buck-boost region. Thereby, with proposed peak-buck peak-boost current control, the converter can achieve high power efficiency over a wide V_{IN} range by reducing both switching loss and conduction loss.

To sense the I_L precisely for a cycle-to-cycle peak current control with a high common-mode voltage up to 60V switching at 2MHz, a direct inductor current sensor with high CMRR is implemented as shown in Fig. 8.7.3. A low temperature-coefficient resistor (R_{sns}), is placed in series with the inductor (L) to sense the I_L. In peak-buck mode (V_{mod}="1"), S_A and S_B switch alternatively. When S_A is turned ON, the high voltage drain-extended MOS transistors M_1 and M_2 are triggered ON after a blanking time delay and further pass the voltage level across R_{sns} to the current sensor. The peak buck voltage-to-current (V2I) converter generates two identical sinking currents $I_{pbuk}=k(V_{COMP}-V_{slp})$. The I_{pbuk} then creates a voltage drop of $V_{R1}=kR_1\times(V_{COMP}-V_{slp})$ across R_1. The comparator stays low until V_{SNS} is higher than V_{R1}. Then S_A is OFF and S_B starts turning ON. Notice that another mirrored sinking current I_{pbuk} is also generated and fed into the current sensor. Therefore, the currents into the differential node are balanced and offset current due to the parasitic resistors can be cancelled to improve the current-sensing accuracy. On the other hand, in peak-boost mode (V_{mod}="0"), the peak boost V2I converter generates two identical sinking currents I_{pbst}, creating a voltage drop of $V_{R1}=k(V_{OS}+V_{COMP}-V_{slp})$. To achieve a smooth inductor current profile during mode transition, an offset voltage is implemented in peak boost mode to compensate the V_{slp} slope change. Once V_{SNS} is larger than V_{R1}, S_C is OFF and S_D is ON. In this way, the peak I_L is detected and further compared to the voltage of V_{R1} that is determined by V_{COMP} and V_{slp}, realizing an effective peak-buck peak-boost current mode control.

The proposed controller was fabricated in a 0.35µm BCD process. The converter supports a wide V_{IN} from 4V to 60V and regulates V_{OUT} at 12V with maximum output power of 48W. In Fig. 8.7.4, the measured waveforms show transitions between different operation regions. It is demonstrated that the converter achieves fine V_{OUT} regulation as well as seamless transitions between different operation regions. Figure 8.7.5 shows measured peak and average radiated EMI performance with automotive CISPR 25 Class-5 limits. In Fig. 8.7.6, a peak power efficiency of 95% can be achieved at 2MHz f_{SW}. Figure 8.7.6 also provides the comparisons with the state-of-the-art wide-V_{IN} power converters. This work achieves >6.8× and >5.9× reductions in the inductance and the capacitance and improves the power density by >2.5×. The converter accomplishes a similar power efficiency at >10× higher f_{SW} compared to prior art. It is the only 2MHz converter to achieve CISPR 25 Class-5 limits for automotive applications in the comparison table. Figure 8.7.7 shows the die micrograph.

Acknowledgements:
The authors would like to thank Keith Szolusha and his group for valuable discussion and EMI testing support.

Reference:
[1] Analog Devices, LTC7813, "Low I_Q, 60V Synchronous Boost+Buck Controller," Datasheet, 2016.
[2] Analog Devices, LT8711, "Micropower Synchronous Multitopology Controller with 42V Input Capability," Datasheet, 2017.
[3] C. Zheng et al., "A 10MHz 92.1%-Efficiency Green-Mode Automatic Reconfigurable Switching Converter with Adaptively Compensated Single-Bound Hysteresis Control," *ISSCC,* pp. 204-205, Feb. 2010.
[4] X. Hong et al., "98.1%-Efficiency Hysteretic-Current-Mode Noninverting Buck-Boost DC-DC Converter with Smooth Mode Transition," *IEEE Trans. Power Electron.*, vol. 32, no. 3, pp. 2008-2017, Mar. 2017.
[5] Texas Instruments, LM5175-Q1, "42V Wide Vin Synchronous 4-Switch Buck-Boost Controller," Datasheet, 2016.
[6] X. Ke et al., "A 10MHz 3-to-40V VIN Tri-Slope Gate Driving GaN DC-DC Converter with 40.5 dBµV Spurious Noise Compression and 79.3% Ringing Suppression for Automotive Applications," *ISSCC,* pp. 430-431, Feb. 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

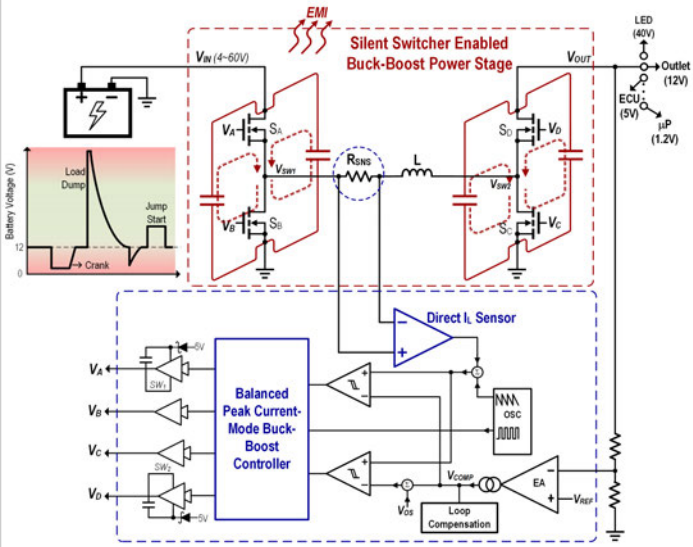


Figure 8.7.1: System architecture of the silent-switcher buck-boost converter.

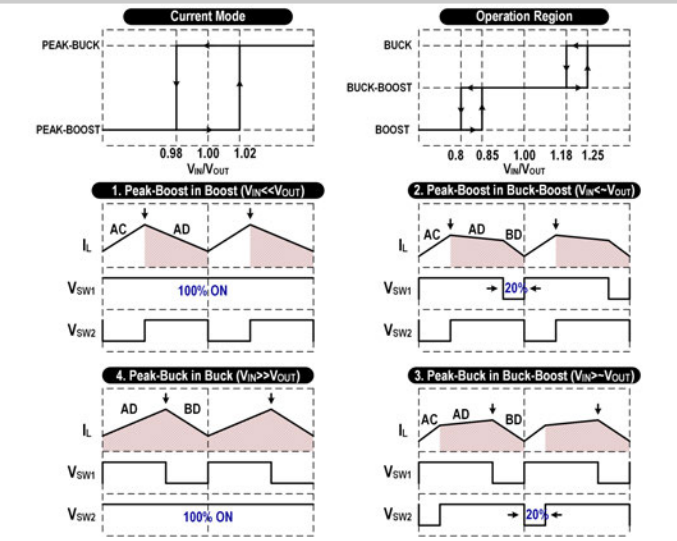


Figure 8.7.2: Operation waveforms of silent-switcher peak-buck peak-boost current mode control.

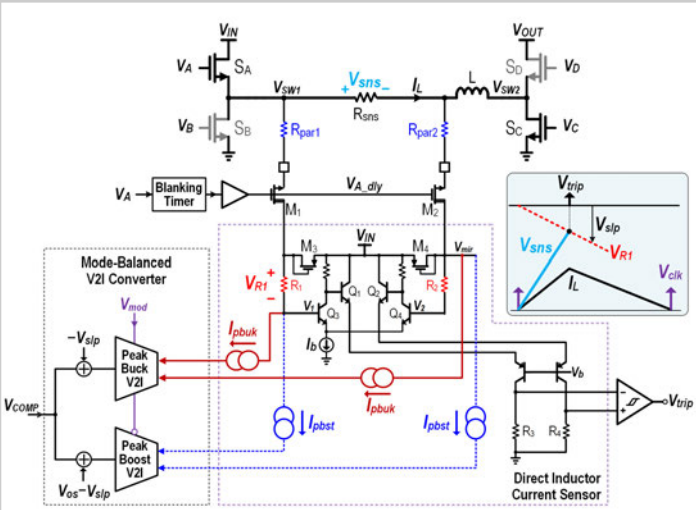


Figure 8.7.3: Circuit diagram of direct inductor current sensor with mode-balanced V2I converter.

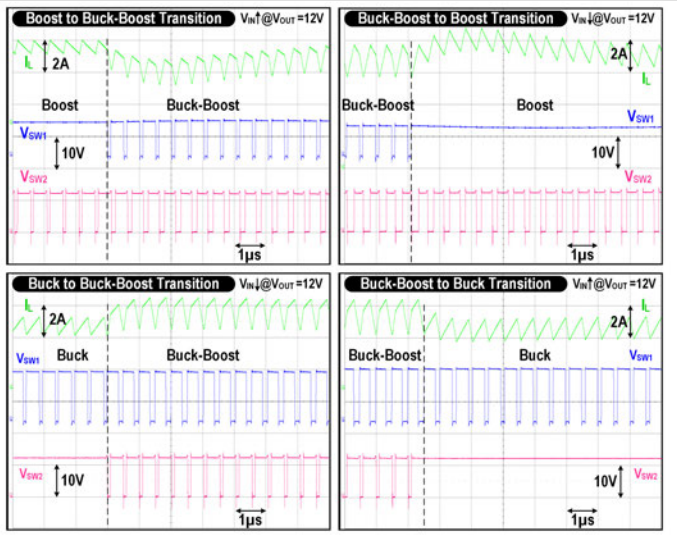


Figure 8.7.4: Measured transition waveforms between different operation regions.

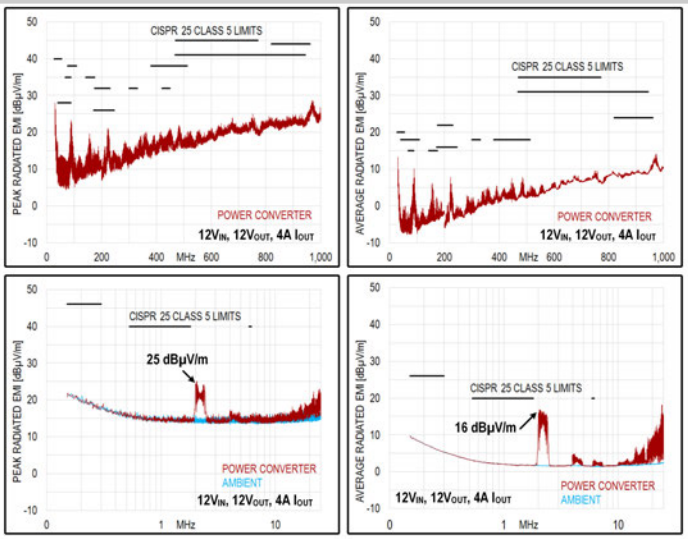
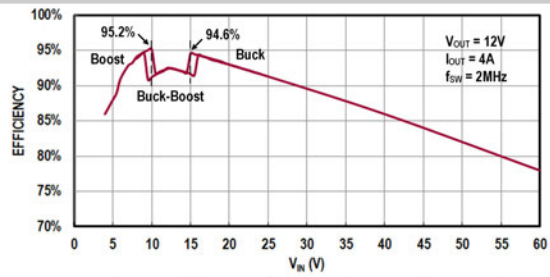


Figure 8.7.5: Measured peak and average radiated EMI performance from 25MHz to 1000MHz and 0.1MHz to 25MHz, respectively.



Design		[1]	[2]	[3]	[4]	[5]	This Work
Process		0.60µm BCD	0.35µm BCD	0.13µm CMOS	0.35µm CMOS	N.A.	0.35µm BCD
Topology		Boost+Buck	SEPIC	Buck-Boost	Buck-Boost	Buck-Boost	Buck-Boost
V_{IN} (V)		8-60	4.5-40	0.9-1.8	2.5-5	3.5-42	4-60
V_{OUT} (V)		0.5-2.5	2	0.9-2.2	3.3	0.8-55	0.8-60
f_{SW} (kHz)		200	200	10000	1660	200	2000
# of Passive Components		2L 2C	2L 1C	1L 1C	1L 1C	1L 1C	1L 1C
L (µH)		2.2, 10	8.2, 15	1	1	6.8	1
C (µF)		74.2, 113	66	1	10	130	22
Power Density (W/mm²)	P_{Max} (W)	120	48	N.A.	N.A.	60	48
	PCB Area (mm²)	2670	1650	N.A.	N.A.	1790	590
Peak Efficiency (%)		97%	90%	92%	98%	98%	95%
EMI Performance		Moderate	Moderate	Poor	Poor	Marginal	Best

Figure 8.7.6: Measured power efficiency and comparison table.

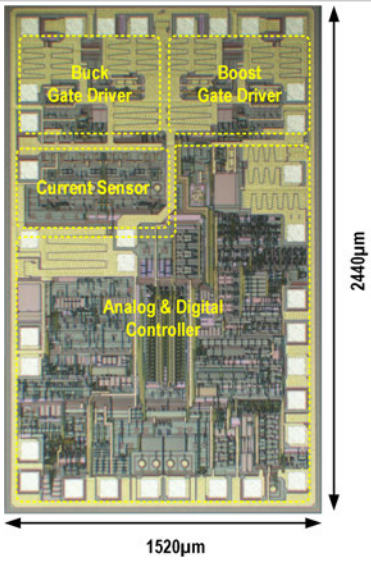


Figure 8.7.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 9 Overview: *High-Frequency Transceivers for Radar and Communications*

WIRELESS SUBCOMMITTEE

Session Chair: ***Nagendra Krishnapura***
Indian Institute of Technology Madras, Chennai, India

Associate Chair: ***Yuu Watanabe***
Waseda University, Atsugi, Japan

Subcommittee Chair: Stefano Pellerano, Intel, Hillsboro, OR

Radars, beamforming systems, and mm-wave radios are at the cutting edge of present-day integrated circuit design. This session opens with an invited paper that outlines technological challenges in developing automotive surround-view radar and describes the implementation of a 79GHz FMCW radar system. This is followed by papers describing state-of-the-art transceivers for automotive FMCW radars, UWB radars for vital sign detection, mm-wave beamforming systems, and an ultra-high-data-rate 300GHz transceiver.

INVITED PAPER **8:30 AM**

9.1 Toward Automotive Surround-View Radars

C-M. Hung, MediaTek, Hsinchu, Taiwan

In Paper 9.1, MediaTek describes future trends and technological challenges in developing automotive surround-view radar systems. A 79GHz FMCW radar transceiver integrated with DSP enables a full radar system in a 16mm×25mm package.

9:00 AM

9.2 A 192-Virtual-Receiver 77/79GHz GMSK Code-Domain MIMO Radar System-on-Chip

V. Giannini, Uhnder, Austin, TX

In Paper 9.2, Uhnder describes a 77/79GHz MIMO radar SoC with 12 transmitters and 8×2 receivers resulting in 192 virtual receivers. A radar system using the 28nm CMOS IC demonstrates 7.5cm range resolution, 1° angular resolution, and 0.099km/h Doppler resolution.

9:30 AM

9.3 A 680μW Burst-Chirp UWB Radar Transceiver for Vital Signs and Occupancy Sensing up to 15m Distance

Y-H. Liu, imec - Netherlands, Eindhoven, The Netherlands

In Paper 9.3, imec, Delft University of Technology, and Eindhoven University of Technology describe a UWB radar for occupancy and vital sign detection. The proposed burst-chip operation and time-domain pre-distortion enable 100× power reduction and a fast chirp slope of 0.7GHz/40μs with a 0.5MHz rms error. The radar detects heart rate at 5 meters and respiration at 15 meters while consuming only 680μW.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

10:15 AM

9.4 A 145GHz FMCW-Radar Transceiver in 28nm CMOS

A. Visweswaran, imec, Leuven, Belgium

In Paper 9.4, imec describes a 145GHz FMCW-Radar transceiver with integrated antennas in 28nm CMOS. The transmitter has an EIRP of 11dBm with a 3dB bandwidth of 127 to 154GHz. The receiver has an effective isotropic noise figure of 6dB and a 3dB bandwidth of 138 to 151GHz. TX leakage is suppressed via delay control on chip.

10:30 AM

9.5 An 80Gb/s 300GHz-Band Single-Chip CMOS Transceiver

S. Lee, Hiroshima University, Higashihiroshima, Japan

In Paper 9.5, Hiroshima University, the National Institute of Information and Communication Technology, and Panasonic describe a 300GHz transceiver in 40nm CMOS. A data rate of 80Gb/s over a distance of 3cm is demonstrated using a pair of transceiver ICs with 24dBi horn antennas at each end.

10:45 AM

9.6 A 42.2Gb/s 4.3pJ/b 60GHz Digital Transmitter with 12b/Symbol Polarization MIMO

C. Thakkar, Intel, Hillsboro, OR

In Paper 9.6, Intel describes a 60GHz digital polar transmitter with on-chip pulse shaping. The 28nm CMOS IC with on-PCB dual polarization antenna transmits spectrally-shaped 28.2Gb/s, 8b/symbol data with 5.6dBm average output power and -21.3dB EVM while consuming 272mW, as well as Nyquist-rate 42.2Gb/s, 12b/symbol data with 6dBm average output power and -24.2dB EVM while consuming 182mW.

11:15 AM

9.7 A Scalable 71-to-76GHz 64-Element Phased-Array Transceiver Module with 2×2 Direct-Conversion IC in 22nm FinFET CMOS Technology

S. Pellerano, Intel, Hillsboro, OR

In Paper 9.7, Intel describes a 71-to-76GHz 64-element phased array with a direct-conversion 2×2 transceiver in 22nm FinFET CMOS technology. In this work, which is the first demonstration of a phased-array in FinFET technology, 4 ICs are co-packaged in a single module that includes 16 patch antennas, and 4 such modules are tiled on a PCB to form the 64-element array. The array has a measured EVM of -20dB for 7.2Gb/s 16QAM at 35dBm effective isotropic radiated power.

11:45 AM

9.8 A 28GHz 20.3%-Transmitter-Efficiency 1.5°-Phase-Error Beamforming Front-End IC with Embedded Switches and Dual-Vector Variable-Gain Phase Shifters

J. Park, KAIST, Daejeon, Korea

In Paper 9.8, KAIST and Hanbat National University describe a 28GHz beamforming front-end IC in 65nm CMOS. It has 6b phase control resolution with 1.5° rms phase error, a 16dB gain control range with 1.4° peak phase variation. The transmitter has a P_{1dB} of 13.9dBm and a peak efficiency of 20.3%, and the receiver has a 4.58dB noise figure.

9.1 Toward Automotive Surround-View Radars

Chih-Ming Hung[1], Alex TC Lin[1], BC Peng[1], Hua Wang[2], Jui-Lin Hsu[1], Yen-Ju Lu[1], Weishow Hsu[1], Jing-Hong Conan Zhan[1], Brian Juan[1], Chi-Hang Lok[1], Sam Lee[1], PC Hsiao[1], Qiang Zhou[2], Mark Wei[1], Hsiang-Yun Chu[1], Yu-Lun Chen[1], Chao-Ching Hung[1], Kevin Fong[1], Po-Chun Huang[1], Pierce Chen[1], Sheng-Yuan Su[1], Yan-Jiun Chen[1], Kehou Chen[1], Chun-Chao Tung[1], Yi-Jhan Hsieh[1], Tzung-Chuen Tsai[1], Yi-Fu Chen[1], Wei-Kuo Hsin[1], Liang Guo[3], Hanfei Liu[3], Dapeng Jin[3]

[1]MediaTek, Hsinchu, Taiwan
[2]MediaTek, San Jose, CA
[3]MediaTek, Hefei, China

The future of driving extends from physical mobility and enjoyment today to having more services enabled by wireless connectivity, clean and green technologies, secure transactions, and so forth. Whether the ownership is based on physical vehicles or services, or whether the drivers are human or robots, one demand that will never change is to have better safety at affordable cost. Among available sensor technologies for the advanced-driver-assistance-system (ADAS), radar is indispensable due to its unique capability in robustness against environmental impacts, long-range detection, sufficient range resolution, and simultaneous multi-depth detection. Those are very crucial since camera, Lidar and ultrasonic sensors perform poorly under severe weather conditions, and an autonomous vehicle would become partially blinded without radars. There are several automotive radar applications such as front radars responsible for autonomous cruise control and automatic emergency braking, as well as corner radars responsible for blind-spot detection (BSD), cross-traffic alert (CTA), and the like. A new class of applications comprehending ultra-short range sensing and 360° surround view for parking assistance, door-opening alert, etc. is emerging. In this paper, requirements of the new applications will be examined, which will be further broken into system and circuit specification. A new system including application-driven algorithm, hardware and software designs will be presented to fulfill the new demands.

The aforementioned front and corner radars are often activated only when vehicle speed is above 30km/h. As a result, the minimum range detection is often greater than 25 to 50cm. The field of view (FoV) for elevation is typically very narrow to avoid clutters and false alarms such as from ground and bridges. In contrast, for surround-view radars illustrated in Fig. 9.1.1, which can be activated in all conditions including slowly maneuvering vehicles in tight space and being able to detect toddlers below a bumper, the minimum range detection needs to be <10cm with <5cm range resolution, and both azimuth and elevation FoV need to be very wide. A comparison for elevation FoV among various types and designs of commercially available sensors is shown in Fig. 9.1.1. A new and cost-effective sensor is required to fulfill the requirements. For maximum range detection, in the case of BSD, CTA, parking assistance (PA), etc., a few traffic lanes of 15-to-20m radius is sufficient.

Designing mm-wave circuits for automotive radars operating between 76 and 81 GHz is challenging [1-3]. However, that only comprises around 25% of the overall radar sensor development. The rest is on software and algorithms in order to comprehend the ever-increasing complexity of driving scenarios. New demands on imaging radar capability further increase the digital circuits and software content. Shown in Fig. 9.1.2 is an overview of frequency-modulated continuous-wave (FMCW) radar operations. A frequency sweep (chirping signal) is sent out by a radar transmitter. The transmitted signal is also used to downconvert the received echo. Spectral analysis on the intermediate frequency (IF), or beat frequency f_b, reveals radial range (distance) information of the target from the sensor. In Fig. 9.1.2, there are four peaks in the RX IF spectrum corresponding to four distinct targets at difference distance. There are multiple methods to estimate target velocity involving different FMCW waveforms. One method utilizing multiple chirps (a frame) is illustrated in Fig. 9.1.2. Each range FFT bin corresponding to a specific target is collected across all chirps in a frame. Applying FFT on the accumulated complex range data would characterize the movement signature of the target over time, i.e. velocity, and is often cited as the Doppler FFT. The range data from multiple chirps establishes a plane in the radar data structure. When there are multiple receivers to form a phased array, a data cube can be constructed. Angle of arrival can be derived from further signal processing of the data cube across multiple receivers. The interference from other FMCW radars can affect the desired detection in various ways as shown in Fig. 9.1.2. In the radar receiver chain, a bandpass filter is commonly included. Interference appears at very close proximity and out-of-band frequencies can be rejected. However, those appearing in band would become ghost targets or increase the noise floor/skirt degrading receiver sensitivity. Interference mitigation remains as an active research topic.

The surround-view radar requirements, fundamental operations and cost effectiveness lead to a vehicle-level solution using multiple radar sensors each with one TX and one RX. Each sensor reports radial range and velocity, as well as signal strength and confidence level of the detection. Low-level radar signal processing is performed within each sensor while application-level signal processing is done in an external microcontroller unit, which also interacts with vehicle operations such as through a controller area network. A block diagram of the individual sensor including high-level specification of intermediate stages is shown in Fig. 9.1.3. Antennas are integrated within the chip leaving no conductive mm-wave signal beyond the IC package. The fully integrated CMOS SoC contains mm-wave transceivers, frequency and waveform generators, self-diagnosis and calibration functions, and digital baseband (DBB) including hardware accelerators and radar signal processing unit. Compared to typical mm-wave wireless transceivers, unique requirements for automotive radars include full-duplex operations within the same frequency band, fast and linear frequency sweep (chirping) over wide bandwidth, high junction temperature up to 150°C, high reliability standard, and functional safety. Fast chirping can facilitate detection under high velocity as well as a variety of applications such as micro Doppler signature extraction. The receiver adopts a mixer-first design considering the reflected wave from a vehicle's own bumper is the largest constant interferer.

The antennas are formed with arrays of folded dipole elements on fan-out wafer-level package (FOWLP) that has been widely used for automotive radars (Fig. 9.1.4). Loss tangent from FOWLP is not as low as that of Rogers material. However, for ultra-short-range radars (USRRs), in-package antennas are ideal because their sizes are small. Low loss for the short interconnect within a package is more beneficial than that for the antenna elements. The chip is directly mounted on a standard low-cost PCB. The size of the fully assembled sensor is 22×16mm^2 which is the smallest complete automotive radar sensor reported to date. The size is significantly dominated by the mechanical constraints of the housing and connector. Therefore, the whole sensor structure has to be designed and simulated together to guarantee antenna performance. The final radiation patterns are also shown in Fig. 9.1.4. Both horizontal and vertical FoV are wide as desired. Effects from retainers and installation onto a bumper vary with car models, and thus are not included in this illustration.

Chirp linearity defined as maximum frequency error over the total bandwidth of a frequency sweep is a key parameter for FMCW operations (Fig. 9.1.5). The chirp nonlinearity can be decomposed into multiple frequency components. The effects could be modeled as FM modulation over the received IF tones producing smeared beat signals or FM tones around the beat frequency. As a result, inaccurate detection, ghost targets, and sensitivity degradation could occur, degrading the overall radar performance. Due to a finer range resolution, USRR requires better than +/-0.05% chirp linearity, which is tougher than that for long range detection [1]. Segmented varactors are used in the VCO design. At 100MHz/μs fast sweep over 4GHz mm-wave bandwidth, chirp linearity for USRR can be satisfied. Limited by digital logic, this design can achieve 340MHz/μs over 4GHz bandwidth. The corresponding peak chirp linearity is +/-0.2%.

The chip was implemented in a standard 65nm CMOS process. A die photo is shown in Fig. 9.1.7 along with photos for the complete PCB and sensor module. The silicon is 9mm^2 and the package is 6.5×6mm^2. The chip has passed automotive qualification and over 60 thousand devices have been characterized. Under 100% duty cycle, the RF/analog and DBB consume 700 and 150mW, respectively. In a typical radar operation, the average power consumption is between 25 and 40mW, respectively. The overall performance is summarized in Fig. 9.1.6 together with a comparison to prior publications. The system demonstration proves the feasibility of proximity sensing using one-TX one-RX radar sensors for automotive surround-view applications.

Acknowledgements:
The authors would like to thank all the members contributing to this work including Chia-Yu Lin, Jack Lu, Scott Kuan, Peter Wu, Ray Lin, Sy-Been Wang, Ben C.-Y. Wu, J.-L. Chen, Yi-Jen Huang, Brad Liu, N.-W. Liu, Min Yang, Vincent Chiang, Lion Huang, W.-Y. Ma, Jay Lin, Dan Liu, Paul Ko, K.-T. Chan, Samny Huang, P.-C. Ho, Y.-J. Wang, Chung-Fa Lee, W.-Z. Wu, Peter Shih, Benson Lin, Yu-Li Hsueh, J.-R. Chen, Chi-Hsueh Wang, Jeff Suen, Zheng Zeng, Charles Chen, Simon Teng, Tao Cheng, and Peter Huang.

References:
[1] B. P. Ginsburg et al., "A Multimode 76-to-81GHz Automotive Radar Transceiver with Autonomous Monitoring", *ISSCC*, pp. 158-159, Feb. 2018.
[2] F. Belfiore et al., "A 76-to-81GHz Packaged Transceiver for Automotive Radar with FMCW Modulator and ADC", *European Radar Conference*, pp. 143-146, Oct. 2017.
[3] T. Fujibayashi et al. "A 76- to 81-GHz Multi-Channel Radar Transceiver", *IEEE JSSC*, vol. 52, no. 9, pp. 2226-2241, 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

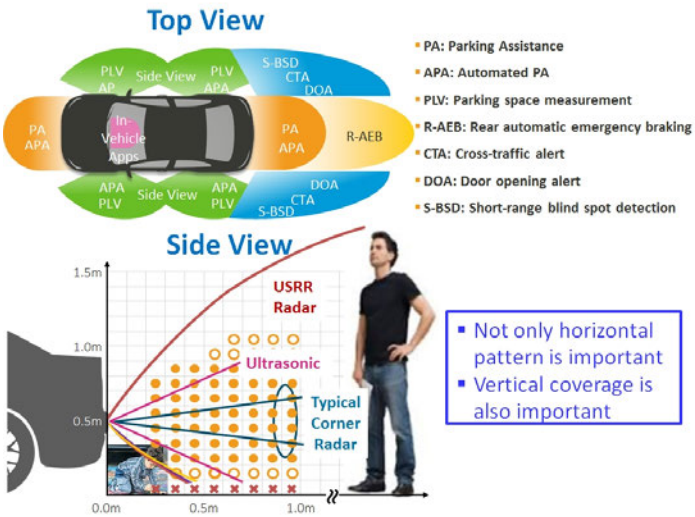


Figure 9.1.1: Use cases for surround-view radars. Detection in elevation is critical but generally lacking from available solutions today.

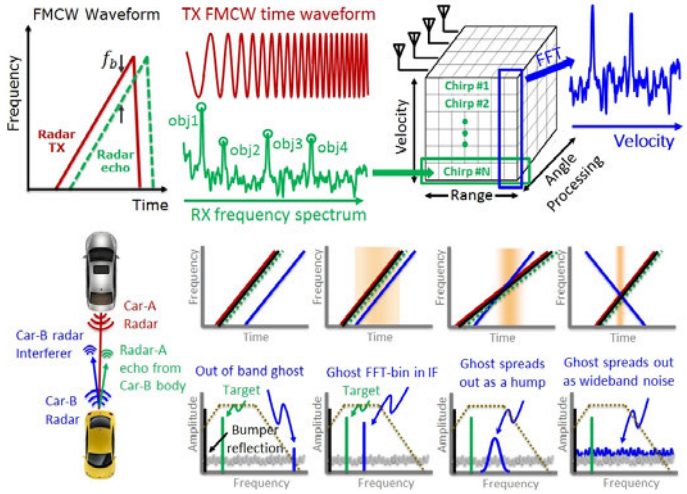


Figure 9.1.2: Overview of radar operations from transmitting a chirp signal to radar signal processing in the digital baseband. Also shown are effects of interference scenarios from other radars.

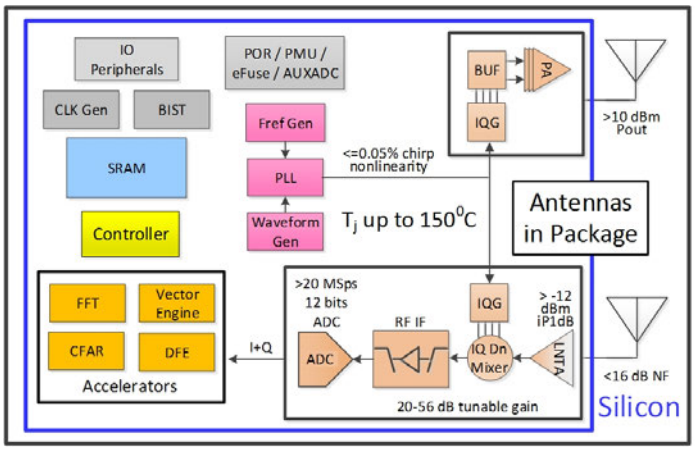


Figure 9.1.3: A block diagram of the mm-wave SoC with high-level specification. The chip supports up to 150°C junction temperature.

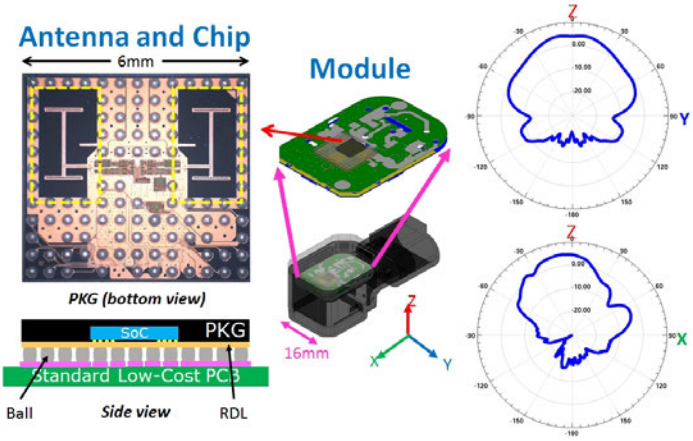


Figure 9.1.4: In-package antennas designed with PCB and module housing. Also shown are azimuth and elevation radiation patterns of the fully assembled sensor.

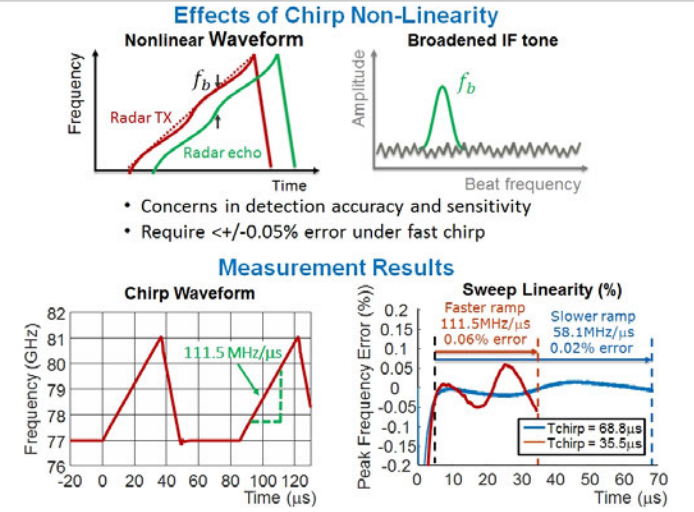


Figure 9.1.5: Impact of chirp non-linearity and measurement results of this work. The nonlinearity is better than the prior arts capable of >4GHz one-shot sweep.

All under worst case temperature and frequency unless specified as "(typ)"

	This Work	'18 ISSCC[1]	'17 JSSC[3]	'17 EuRAD[2]	'14 EuMA [4]	'12 TMTT[5]
Process Technology	65nm CMOS	45nm CMOS	130nm SiGe BiCMOS	130nm SiGe BiCMOS	65nm CMOS	180nm SiGe BiCMOS
Number of TX/RX	1/1	3/4	1/6	3/4	2/4	1/4
Integrated Antenna	Yes	No	No	No	No	No
PA Output Power (dBm, w/o antenna)	13	10.8	13	13 (typ)	11 (typ)	11
RX NF (dB)	13	18	12	9 (typ)	12 (typ)	15
RX iP1dB (dBm)	-7.5	-7	+3	-5 (typ)	-	-5
Chirp Linearity (peak % @ MHz/µs)	0.02 @ 58.1 0.06 @ 111.5	0.06 @ 100	-	0.2%	-	-
Phase Noise @1MHz offset (dBc/Hz)	-91	-91	-99	-95 (typ)	-84	-97
Max. Junction Temperature (°C)	150	125	125	-	90	125
Self Diagnosis	Yes	Yes	Partial	Yes	No	Partial
RF Area (mm²)	2.3	22	12.7	-	33.6	18.5
Peak RF Power Dissipation (W)	0.7	3.5	1.8	2.7	1.1	2.54

[1] B. P. Ginsburg et al., "A Multimode 76-81GHz Automotive Radar Transceiver with Autonomous Monitoring", *2018 ISSCC*, San Francisco, CA, 2018.
[2] F. Belfiore et al., "A 76 to 81GHz packaged transceiver for automotive radar with FMCW modulator and ADC", *2017 EURAD*, 2017.
[3] T. Fujibayashi, et al. "A 76- to 81-GHz Multi-Channel radar Transceiver", *JSSC*, v52, n9, pp. 2226-2241, 2017.
[4] T. Shimura, et al., "Multi-Channel Low-Noise Receiver and Transmitter for 76-81 GHz Automotive Radar Systems in 65 nm CMOS," *EuMC.*, pp. 596-599, 2014.
[5] S. Tretta et al., "An RCP Packaged Transceiver Chipset for Automotive LRR and SRR Systems in SiGe BiCMOS Technology", *IEEE TMTT*, 2012.

Figure 9.1.6: Summary and comparison table. All reported performance parameters for this work are under FMCW system operations and from the worst-case junction temperature.

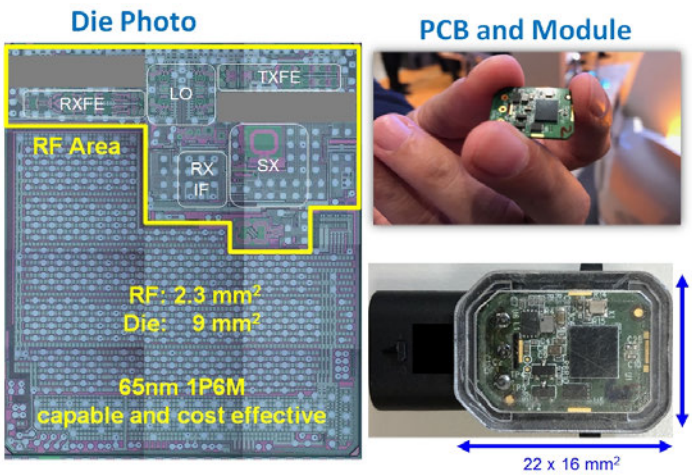


Figure 9.1.7: Die micrograph, PCB and sensor-module photos.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

9.2 A 192-Virtual-Receiver 77/79GHz GMSK Code-Domain MIMO Radar System-on-Chip

Vito Giannini, Marius Goldenberg, Aria Eshraghi, James Maligeorgos, Lysander Lim, Ryan Lobo, Dave Welland, Chung-Kai Chow, Andrew Dornbusch, Tim Dupuis, Struan Vaz, Fred Rush, Paul Bassett, Hong Kim, Monier Maher, Otto Schmid, Curtis Davis, Manju Hegde

Uhnder, Austin, TX

MIMO radars transmit selected waveforms on N TX physical antennas: when they receive them back on M RX physical antennas, they are able to reconstruct an array with (N×M) virtual receive elements, thereby obtaining a finer spatial/angular resolution. Fully integrated mm-wave radar transceivers such as [1-3] implement up to 12 Virtual Receivers (VRX) and can be expanded to larger time-domain MIMO arrays only by relying on costly PCB-based implementations.

We propose a code-domain MIMO radar capable of processing up to 192 VRX: we put special emphasis on architecting a platform that is sufficiently versatile to cover diverse applications, while still minimizing costs and meeting the most challenging link-budget requirements. Figure 9.2.1 shows the block diagram of the proposed radar: the SoC features a 77/79GHz-band transceiver with 12 TX and 8 RX channels that can be time-multiplexed to 2 sets of antennas to get coverage of both azimuth and elevation profiles. In this prototype, the 15.2-to-16.2GHz LO is generated by an external PLL: the differential LO input is converted to quadrature by using a polyphase filter (PPF); the filter output is buffered and distributed to all TX/RX channels with a differential 100Ω-terminated transmission line. Finally, all channels use limiting amplifiers and Sub-Harmonic Quadrature-Injection-Locked Oscillators (SH-QILO) to generate the mm-wave carrier at 5× the injected LO frequency. Digital signal processing and other radar signal computational power is supported on the same chip by two CPUs with floating point support and two DSPs. The overall software-defined hardware pipeline is capable of up to 20 Tera-Ops baseband processing.

Figure 9.2.2 shows the architecture of the proposed TX: a GMSK digital modulator feeds a zero-IF quadrature mm-wave upconverter. When compared to BPSK radar implementations [1], this TX is more spectrally efficient due to the absence of the large adjacent frequency sidelobes. Compared to FMCW radars [2], we shift the modulation complexity/precision to the high-speed data-converters and the DSP, where performance scales well with the CMOS technology node. A quadrature GMSK up-converter generates a constant-envelope phase-modulated signal and uses a fully saturated power amplifier. Code-domain MIMO allows to achieve more power on target by simultaneously using multiple TX channels with different codes. Hence, for a given link budget, designing smaller PAs becomes a viable option, which allows to more easily fulfill the most stringent lifetime requirements. Digital beamforming is an available mode of operation that allows to trade field of view for higher EIRP. At baseband, a programmable Pseudo-Random-Noise (PRN) code is generated, digitally GMSK modulated and fed into current-steering 2/4/8 Gs/s DACs whose current outputs are amplified and filtered by reconstruction filters. The inset in Fig. 9.2.2 shows the schematic of this current-mode biquadratic filter: transistors M1p/m, M2p/m and M3p/m form a negative feedback that translates into a low input impedance at nodes iip and iim. Transistors M4p/m provide the scaled output current to the mixer. This architecture implements a current-mode gm-C biquad with programmable cut-off frequency (0.25 to 2GHz) and current gain (0 to 14dB). The current-mode double-balanced mixer performs the frequency upconversion to the 77/79 GHz bands.

The RX uses a zero-IF architecture followed by 2Gs/s SAR ADCs, a digital front-end and a data control unit, which handles the radar data cube computations including range correlation, doppler processing and beamforming. As shown in Fig. 9.2.3, the RX front-end is built around the 3-winding planar transformer T0. The first LNA stage is duplicated into LNA11 and LNA12: at their inputs, these amplifiers are connected to two different antenna ports while their outputs are power-combined into T0. This solution allows to time-multiplex the RX to azimuth or elevation antennas, avoiding the use of lossy mm-wave SPDT switches. The RX front-end needs to tackle a large modulated self-interference coming from the radar's own TXs, which may both saturate the RX and increase the correlation noise floor. This challenge is even more stringent in MIMO radars where multiple TX active channels may constructively sum up at specific RX inputs. We propose a cancellation scheme comprising a Pattern Generator Unit (PGU), Analog and Digital Interference Cancellation Units (A/D-ICU). In the A-ICU, a background channel estimation runs continuously: the resulting signal path delays and impulse response are then applied to the upconverted mm-wave feedback whose output is power-combined into the transformer T0. The D-ICU estimates the complex sum of the wideband self-interferer presented at each receiver input and further removes the residual interferer before the correlators. The Radar Data Control Unit can process up to 1024 range bins for a single radar scan. The Winogard FFT engine can manage up to 800 million points per second Doppler throughput. The beam-forming engine throughput is 1.6 billion-beams/s and implements a covariance-matrix SVD decomposition engine that can support Direction-of-Arrival algorithms such as Multiple Signal Classification (MUSIC), Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) and Capon's algorithm engine to further improve the angular resolution.

The radar SoC prototype was implemented in a CMOS 28nm HPC technology and packaged in a Fan-Out Wafer-Level Package. The die micrograph is shown in Fig. 9.2.7: the SoC occupies 71mm^2 within which the mm-wave transceiver itself occupies 14mm^2, resulting in one of the smallest-area per VRx to date. The overall power consumption varies with the use case and it is typically 15W from 0.9/1.3/1.8V supplies. The mm-wave RF characterization was performed on a test PCB with WR12 waveguide interfaces for each TX and RX channel. The radar is tested with an external PLL, and the 77.5GHz TX phase-noise at 1MHz offset is -110dBc/Hz.

Figure 9.2.4 shows measurement results for the RX self-interference cancellation unit. These system measurements were performed on an antenna board that uses Substrate-Integrated-Waveguide (SIW) antennas. During this test, 8 TXs were operating at the same time in MIMO mode. The resulting RX output is shown in the FFT, IQ and correlation domains, before and after cancellation. We report more than 30dB cancellation which allows to concurrently lower the correlation noise floor by about 20dB compared to when the self-interference is present.

Figure 9.2.5 shows 4 different static radar SoC measurements. Figure 9.2.5(a) shows a MIMO measurement for a car moving at 25 mph boresight on an air field: the radar graph shows 22dB SNR for an instantaneous detection at 290m. In Fig. 9.2.5(b) the radar is operating in phased-array mode and shows detections for a person moving boresight up to 280m. Figure 9.2.5(c) shows results of beam-forming (dashed green line) and MUSIC (blue line) in an anechoic chamber: the radar resolves the two -10dBsm corner reflectors within about ±1-degree angle. Figure 9.2.5(d) shows how the digital beamforming sidelobe performance (red) can be improved after windowing (blue) and using a hardware-based Spatially-Variant Apodization (SVA) (green). The key mm-wave performance is provided in Fig. 9.2.6 together with a comparison with the recent state-of-the-art for PMCW/FMCW radars. At the package output reference plane, the TX combined output power is about +19.6dBm (125C/80GHz), when 12 TX are active at the same time. The RX NF measured at the ADC output is about 16dB (125C/80GHz).

A MIMO radar SoC is presented that uses a novel digital-modulation architecture and achieves state-of-the-art mm-wave performance. The SoC achieves the best angular resolution, while being the most integrated CMOS radar sensing platform compared to the state of the art shown in Fig. 9.2.6.

Acknowledgements:
The authors would like to thank for their contributions A. Sengottuvel, Y. Ramasamy, M. Ali, A.E. Ertan, A. Sathyanarayana, R. Rao, J.P. Bordes, J. Preussner, and J. Schroeder.

References:
[1] D. Guermandi et al., "A 79GHz 2×2 MIMO PMCW Radar SoC in 28nm CMOS," *IEEE ASSCC*, pp. 105-108, 2016.
[2] B. P. Ginsburg et al., "A Multimode 76-to-81GHz Automotive Radar Transceiver with Autonomous Monitoring," *ISSCC*, pp. 158-160, 2018.
[3] T. Fujibayashi et al., "A 76-to-81-GHz Multi-Channel Radar Transceiver," *IEEE JSSC*, vol. 52, no. 9. pp. 2226-2241, Sept. 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

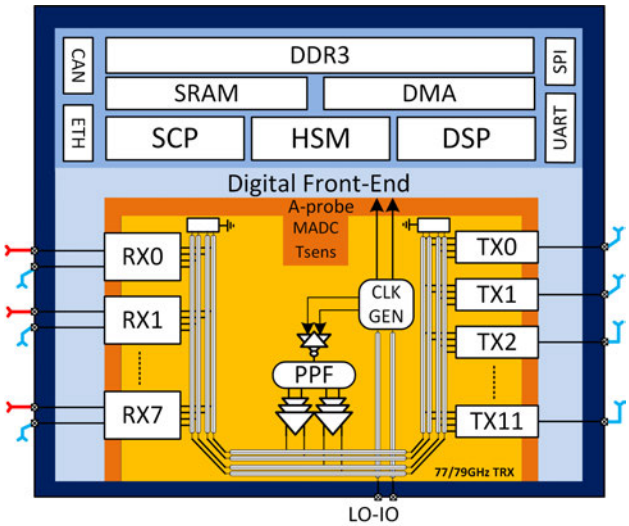


Figure 9.2.1: Radar SoC block diagram, including the 12TX / 8×2 RX 77/79 GHz MIMO transceiver, the digital signal processing, memory and interfaces.

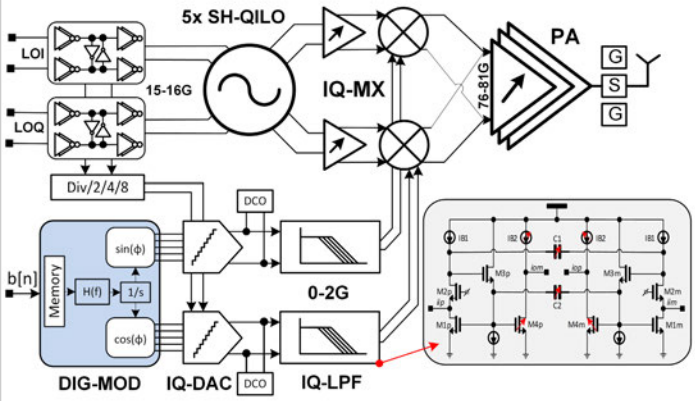


Figure 9.2.2: TX Block Diagram and schematic of the wideband current-mode biquadratic cell.

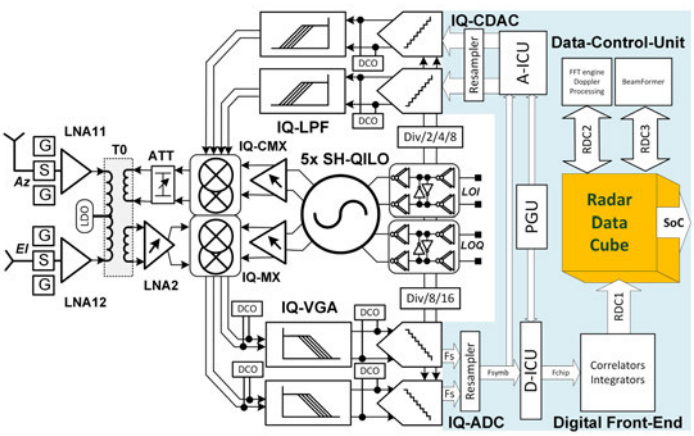


Figure 9.2.3: RX architecture, including details of the double-antenna-input LNA, the cancellation feedback, as well as the digital front-end.

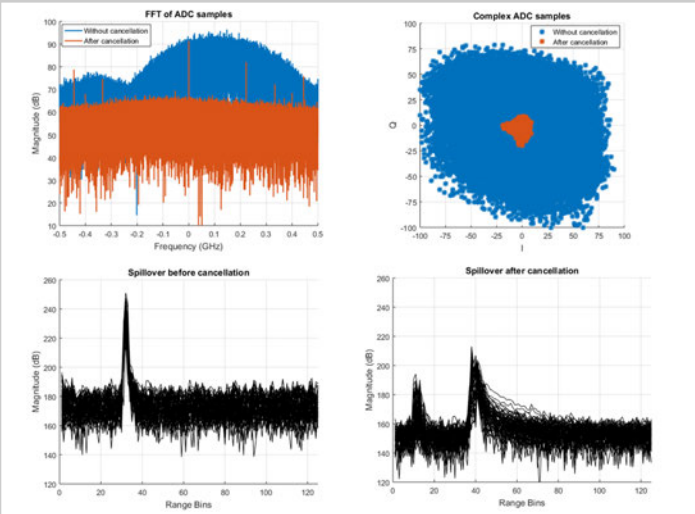


Figure 9.2.4: GMSK Spillover cancellation for 8 TX MIMO channels operating at the same time. The RX outputs are shown in the FFT, IQ and correlation domains before and after cancellation.

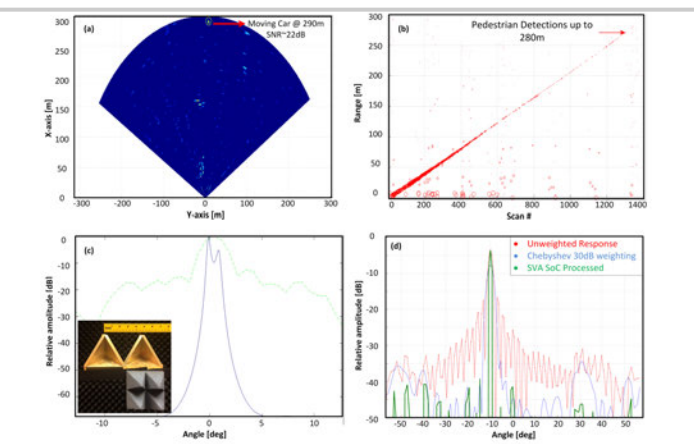


Figure 9.2.5: Radar-on-Chip measurements: (a) 25mph moving car MIMO detection at 290m with 22dB SNR; (b) Beamforming pedestrian detections up to 280m; (c) -10dBsm corner reflectors in anechoic chamber after beamforming and Music; (d) angle sidelobes performance after digital beamforming (red), using a Chebyshev window (blue) or applying Spatially-Variant Apodization (SVA) in hardware.

		This work	[1]	[2]	[3]
Technology		CMOS 28nm	CMOS 28nm	CMOS 45nm	SiGe 130nm
Modulation		GMSK-PMCW	BPSK-PMCW	FSK-FMCW	-
mm-wave bands		77/79 GHz	77/79 GHz	77/79 GHz	77/79 GHz
Physical mm-wave antenna IOs TX / RX / max VRx	#	12 / 8(x2) / 192	2 / 2 / 4	3 / 4 / 12	1 / 6 / 6
TX Max Combined Output Power	dBm	19.6	11.5	15.8	13
RX Noise Figure	dB	16	12	17.8	13.5
RX IP1dB	dBm	-6.7	-	-5	+3
RX IF BW	GHz	0.25 to 2	0.25 to 1	15	-
RF TRX summary					
TRX Silicon Area breakdown	mm^2	14	7.9	22	12
TRX Area / (# VRx)	mm2/VRx	0.14 (96)	1.97 (4)	1.83 (12)	1 (12)
TRX Power consumption (3TX x 4RX, 12 VRx, 100%)	W	3.5	-	3.5	1.8
TRX Power consumption (12TX x 8RX, 96 VRx, 100%)	W	9.5	-	-	-
Power consumption / (# VRx)	mW/VRx	100 (96)	250 (4)	292 (12)	300 (6)
SoC system summary					
Range Resolution	cm	7.5	7.5	-	-
Angular Resolution	deg	1	5	-	-
Doppler Resolution	Km/h	0.099	-	-	-
Baseband processing	Tera-op/s	20	-	-	-
Total Silicon Area	mm^2	71		-	-
Typical Power Consumption	W	15		-	-

Figure 9.2.6: Summary and comparison for 125C/80GHz performance at the package output.

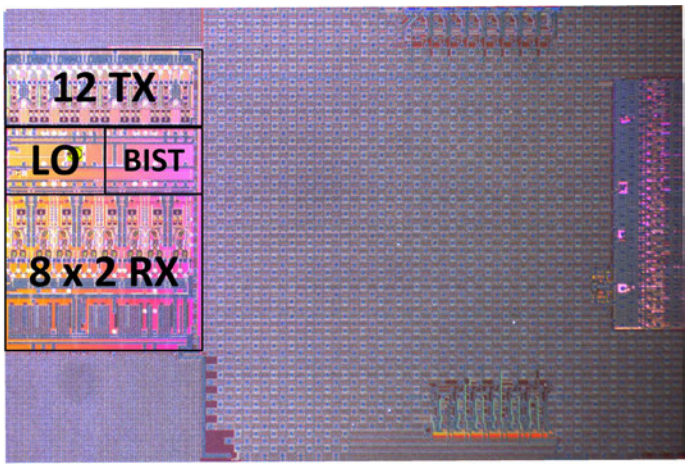


Figure 9.2.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

9.3 A 680μW Burst-Chirp UWB Radar Transceiver for Vital Signs and Occupancy Sensing up to 15m Distance

Yao-Hong Liu[1], Sunil Sheelavant[1,2], Marco Mercuri[1], Paul Mateman[1], Johan Dijkhuis[1], Wilfried Zomagboguelou[1,3], Arjan Breeschoten[1], Stefano Traferro[1], Yan Zhan[1], Tom Torf[4], Christian Bachmann[1], Pieter Harpe[3], Masoud Babaie[2]

[1]imec - Netherlands, Eindhoven, The Netherlands
[2]Delft University of Technology, Delft, The Netherlands
[3]TU Eindhoven, Eindhoven, The Netherlands
[4]imec, Leuven, Belgium

For remote vital signs and occupancy detection in many smart home/building applications, radar sensors are a preferred option over cameras, due to privacy preservation and robustness to ambient light conditions. These radars not only need to provide precise range and vital signs information over meters distance, but also preferably can operate on a battery up to a few months or even years, for cost and practical reasons (like smoke detectors). State-of-the-art remote vital-sign sensors typically use an impulse-radio UWB (IR-UWB) radar [1,2] because it provides a range resolution <20cm. However, their power consumption is typically in the order of 100's of mW, preventing long-term maintenance-free battery-powered operations. Although mains power can be used to supply such radars, this is not always available, depending on the location and the building type, and the installation cost (e.g., power routing) is significantly higher than for battery-powered ones. In this work, a burst-chirp radar with an energy-efficient chirp generation is proposed, leading to a record-low power consumption of 680μW.

The radar frequency is chosen to be 7GHz UWB for the following reasons: availability of sufficient BW to reach a fine range resolution, possibility of low-power circuit implementation, and lower path loss compared to mm-wave bands. However, UWB has very tough peak (0dBm/50MHz), average (-41dBm/MHz), and sideband (<-70dBm/MHz) power spectral density (PSD) regulations, and the BW is much narrower in certain regions (e.g., 6 to 8.5GHz in Europe). Hence, it is challenging for the IR-UWB radars to fulfill the sidelobe PSD without compromising detection distance [1,2]. Moreover, the power consumption of IR-UWB transceivers does not reduce with the duty cycling. E.g., the IR-UWB RX needs to be continuously active for >66ns to receive a 2ns impulse reflected from a 10m distance (33× longer than the signal duty-cycle). Frequency-modulated continuous-wave (FMCW) radars [3][4] have low sideband emission (<-30dB), but the continuous operation of FMCW radars can easily violate the UWB average PSD requirement compared to IR-UWB radars with a similar detection distance. In this work, an FMCW radar is heavily duty-cycled to a short burst to meet UWB regulations. However, such burst operation also poses several design challenges in FMCW radars: fast chirp generation, frequency pulling, and settling overhead mitigation.

The maximum unambiguous velocity of the targets determines the required pulse repetition interval (PRI) in FMCW radars. A PRI of 1.3ms is chosen here to capture any fast movement of an indoor subject (e.g., sudden falling) with a Doppler velocity up to 7m/s. With 0dBm PA output power, a duty cycling factor should be less than 3% to meet the UWB average PSD restriction, thus resulting in a 40μs chirp time. All the radar circuits are disabled outside the burst-chirp period (Fig. 9.3.1), which significantly reduces the average power consumption by ~33×, whereas the chirp speed should also be increased by the same ratio. Although the 2-point modulation PLL presented in [5] can provide a chirp slope up to 940MHz/29μs, long settling time and high power consumption of the fractional-N PLL significantly increase the energy overhead in a heavy duty-cycling system. A fast chirp generation based on an open-loop DCO is employed in this work, since a precise center frequency is not required and PLLs typically do not rule the fast chirp behavior in a short burst. As shown in Fig. 9.3.1, the DCO is first locked to an initial frequency f0 within 5μs by an FLL consuming 500μW. Then the loop is opened, and the DCO frequency starts to chirp to the targeted final frequency f1. The digital PA (DPA) is only enabled during DCO chirping (Fig. 9.3.1). If the DPA is enabled while the DCO is not chirping, even a 100ns duration would introduce a spur in the spectrum and violate the regulations. On the other hand, if the DPA is activated/deactivated while the DCO is chirping, the frequency pulling would introduce a large chirp frequency error. To alleviate the frequency pulling while avoiding spurs, the DPA is digitally ramped up/down at the start/end of each chirp, and the timing of this procedure is precisely controlled by a finite-state-machine (FSM) with a 20ns time resolution.

Figure 9.3.2 shows the proposed LC-DCO with an embedded Domino chirp generation sampled at 50MHz. The DCO consists of a 9b binary FLL bank and a 9b unary chirp bank. The chirp bank has a digitally-controlled capacitor array with 32 columns and 16 rows, and each unit has a 0.5fF capacitance. If each unit is controlled separately, there will be up to 512 digital controls routing inside the DCO, which would significantly increase parasitics. This will diminish not only the chirp speed, but also the DCO start-up margin. In this work, the dynamic-latch-based shift register is tightly integrated with a unit capacitor, so the entire DCO becomes very compact. To minimize clock activity and also supply droop, only one row at a time has an active clock while the other rows remain static. The chirp slope can reach maximum 700MHz/10μs. During the burst, the FLL, DCO and chirp generation consume only 3mW DC power.

To meet the targeted displacement accuracy for heartbeat detection at several meters, an RMS frequency error must be <2MHz. However, the chirp non-linearity due to the nature of $\sqrt{LC}$ (Fig. 9.3.3) can introduce a frequency error of more than 10MHz. Such non-linearity is predictable, since the FLL precisely defines the initial C0, and the inductance L is fairly constant over PVT. A simple digital predistortion (DPD) technique could have been applied, but the capacitor units in Domino chirp generation are not individually accessible. As conceptually illustrated in Fig. 9.3.3 (with exaggeration), a time-domain DPD (T-DPD) is proposed to correct the curvature of the chirp by modulating the number of the clock periods T_{CLK} in each frequency step *k*. This number is calculated by a 3rd-order polynomial, which is implemented in the digital domain with simple hardware to approximate the inverse of the square-root relation. The polynomial's coefficients are only adapted for each wafer (i.e., one-time calibration) to overcome the process variation of the unit capacitors (ΔC). The linearized results are insensitive to supply and temperature variations, thanks to the metal-finger capacitors.

An HPF in the RX is typically required in FMCW radars to suppress the TX-RX leakage, but its slow settling significantly increases energy consumption overhead in burst operation. Thanks to the fast chirp slope of 700MHz/40μs, an HPF corner as high as 200kHz can be selected, resulting in a quick settling within 5μs. Unlike the ADCs in many direct-sampling IR-UWB RXs that typically sample at 10s of GHz, and consume 10s of mW [1], the presented radar RX uses a 9b SAR-ADC sampled at 12.5MHz consuming only 30μW.

The chip was fabricated in 40nm CMOS, and the die micrograph is shown in Fig. 9.3.7. Thanks to a digital-intensive implementation, the transceiver occupies a core area of only 0.48mm^2. The TX output spectrum, the time-domain waveform of the TX burst, and its corresponding frequency demodulation are shown in Fig. 9.3.4. When PRI is selected to be 1.3ms, the 0dBm TX has average and sideband PSDs of -45dBm/MHz and -85dBc/MHz, respectively, which are well below the UWB indoor regulations internationally. As shown in Fig. 9.3.5, the chirp non-linearity is dramatically reduced when T-DPD is applied, and the results are fairly consistent against supply and temperature variations. The measured RMS error among 5 different dies in the same wafer is 0.3 to 0.5MHz. The measured RX P_{1dB} and noise figure are -35dBm and 12.5dB, respectively. Figure 9.3.5 also shows the RX output de-chirping FFT with rectangular windowing. A 3m cable delay and a 50dB attenuator are connected in between TX output and RX input. The measured sidelobe is lower than -12.8dB after applying T-DPD, where the ideal level is -13.5dB [4]. The presented radar together with off-the-shelf antennas (each 16cm^2 size, 3-to-6dB gain) can detect human respiration and heartbeat up to 15meter and 5meter distances, respectively. The peak power consumption is broken down in Fig. 9.3.5. Even lower average power consumption can be achieved with longer PRI for slow-moving scenarios. Figure 9.3.6 summarizes the performance and compares with state-of-the-art IR-UWB and FMCW radars. The presented radar achieves the longest vital-sign detection distance among state of the art, while consuming at least 100× less power, enabling many battery-powered and handheld radar sensing applications.

References:
[1] N. Andersen et al., "A 118mW Pulse-Based Radar SoC in 55nm CMOS for Non-Contact Human Vital Signs Detection," *IEEE JSSC,* pp. 3421-3433, Dec. 2017.
[2] T.-S. Chu et al., "A Short-Range UWB Impulse-Radio CMOS Sensor for Human Feature Detection," *ISSCC,* pp. 294-296, 2011.
[3] L. Lou et al., "A 253mW/Channel 4TX/4RX Pulsed Chirping Phased-Array Radar TRX in 65nm CMOS for X-Band Synthetic Aperture Radar Imaging," *ISSCC,* pp. 160-162, 2018.
[4] Y. Wong et al., "A 260mW Ku-Band FMCW Transceiver for Synthetic Aperture Radar Sensor with 1.48GHz Bandwidth in 65nm CMOS Technology," *IEEE T-MTT,* pp. 4385-4399, Nov. 2017.
[5] H. Yeo et al., "A 940MHz-Bandwidth 28.8ms-Period 8.9GHz Chirp Frequency Synthesizer PLL in 65nm CMOS for X-band FMCW Radar Applications," *ISSCC,* pp. 238-240, 2016.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

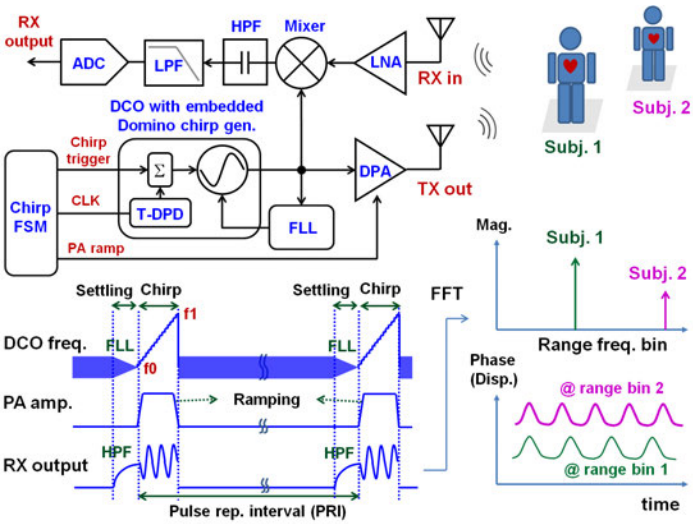


Figure 9.3.1: Simplified block diagram and concept illustration of the proposed burst-chirp radar.

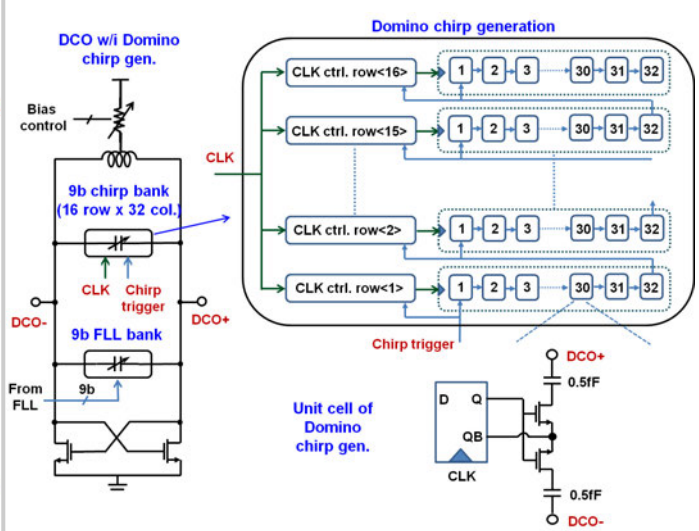


Figure 9.3.2: Simplified schematic of the DCO with Domino chirp generation.

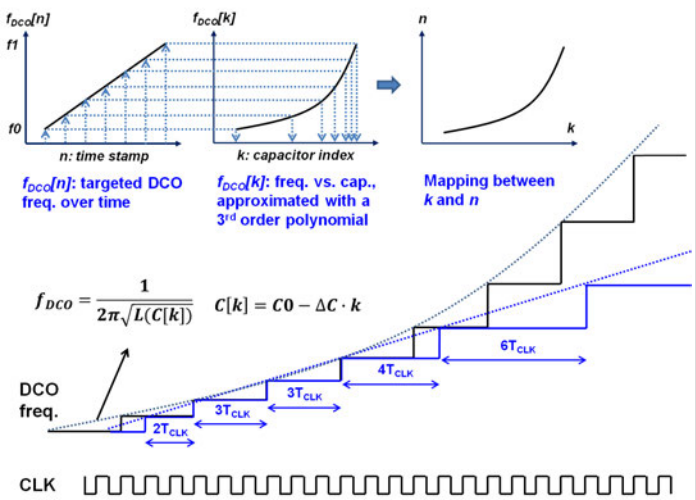


Figure 9.3.3: Concept illustration of the proposed time-domain chirp pre-distortion (with exaggeration).

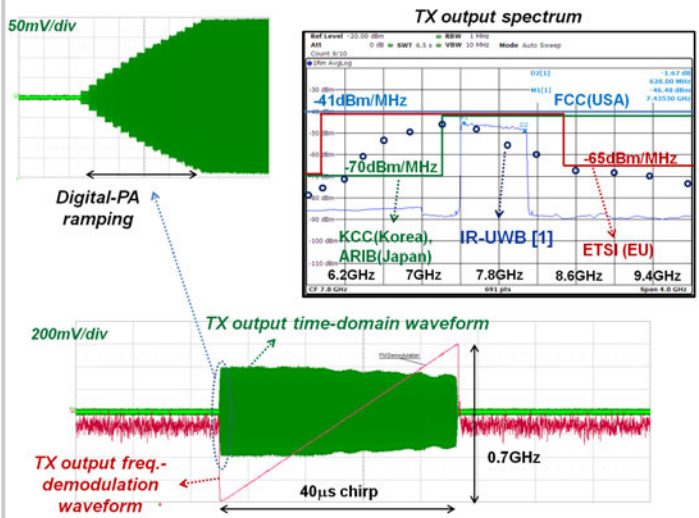


Figure 9.3.4: Measured time-domain waveforms of TX burst and frequency demodulation, and the TX output spectrum.

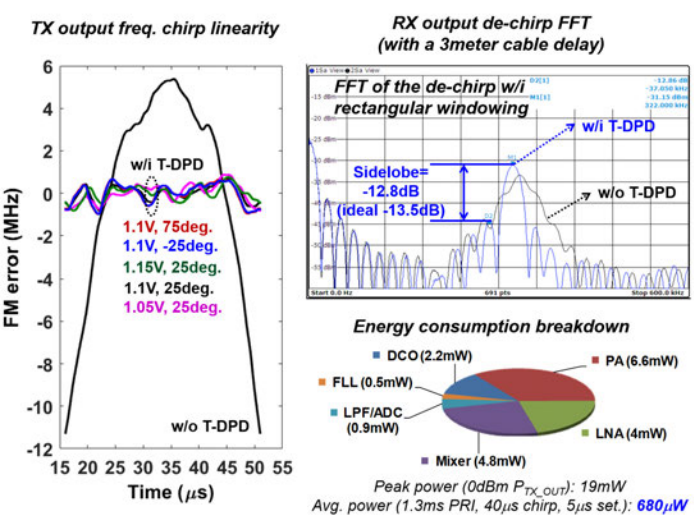


Figure 9.3.5: Measured chirp frequency non-linearity, RX output de-chirp FFT, and the peak power consumption breakdown.

	This work	[1] Andersen, JSSC'17	[2] Chu, ISSCC'11	[3] Lou, ISSCC'18	[4] Wang, JSSC'17	[5] Yeo, ISSCC'16
Tech.	40nm	55nm	130nm	65nm	65nm	65nm
Freq. (GHz)	6.8-8.2GHz	7.2-8.5GHz	2-5GHz	10GHz	15GHz	8.4-9.4GHz
UWB indoor regulation Compliancy	FCC/ETSI/ KCC/ARIB	FCC/ETSI/ KCC/ARIB	N.A.	Outdoor	Outdoor	-
UWB sideband PSD	**-85dBm/MHz**	-70dBm/MHz	-	-	-	-
Modulation	Burst chirp	Impulse	Impulse	FMCW saw-tooth*	FMCW saw-tooth	FMCW triangular
VDD	1/1.1V	1.8V	1.9V	1.2V	1.2V	1.1-1.3V
Mod. BW	0.7-0.8GHz	1.4GHz	0.9GHz	1GHz	1.48GHz	0.95GHz
Pulse/chirp period	10-40µs	1.8ns	3ns	0.2-1ms	1-10ms	28µs
Pulse rep. interval	1.3ms	N.A.	100ns	1ms	1-10ms	28µs
TX P_{OUT}	0dBm	0.7dBm	N.A.	10dBm	13dBm	-
TX RMS freq. error	0.3-0.5MHz	-	-	0.2-1.1MHz	0.16MHz	1.7MHz
RX FFT sidelobe	-12.8dB	-	-	-12.9dB	-12.7dB	-
Heartrate det. distance	5m	5m	-	-	-	-
Breath det. distance	**15m**	9m	0.75m	-	-	-
Die area	**1.8mm²**	8.6mm²	11.9mm²	7.8mm²	4mm²	0.18mm²***
Peak DC power: Chirp gen.	**2.7mW**	11.5mW	N.A.	N.A.	41mW	14.8mW
Peak DC power: Total	**19mW**	N.A.	N.A.	141mW/ch**	148mW**	-
DC power active det.	**<0.68mW**	118mW	695mW	141mW/ch**	148mW**	14.8mW***

* Non-continuous saw-tooth chirp for synthetic aperture radar, but the radar is continuous active.
** PA power consumption is excluded for fair comparison. *** PLL only.

Figure 9.3.6: Summary and comparison table of state-of-the-art IR-UWB and X-band FMCW radars.

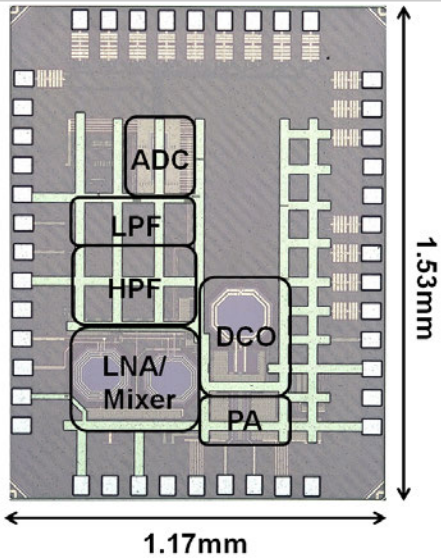


Figure 9.3.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

9.4 A 145GHz FMCW-Radar Transceiver in 28nm CMOS

Akshay Visweswaran, Kristof Vaesen, Siddhartha Sinha, Ilja Ocket, Miguel Glassee, Claude Desset, Andre Bourdoux, Piet Wambacq

imec, Leuven, Belgium

Indoor radar applications detecting people, vital signs and minute gestures require a high range resolution. This paper presents a 145GHz FMCW radar transceiver with on-chip antennas in 28nm bulk CMOS. An RF bandwidth of 13GHz yields an 11mm range resolution, and the high RF carrier permits greater velocity and MIMO-angular resolution. An external chirp signal at 16.1GHz is upconverted to 145GHz via a cascade of two frequency triplers in the RX and TX. For MIMO operation, a central chirp signal from a sub-sampling PLL integrated in 28nm CMOS [1] is distributed to multiple TRX chips on a PCB, eliminating mm-wave signal routing. The radar operates over a 0.15-to-10m range using fast sawtooth chirps of 5-to-50µs duration. The beat frequency at IF fits between 400kHz and 15MHz. The on-chip TX leakage produces a strong beat near DC, suppressed via active highpass filtering. However, group delays of the PA and LNA can shift the beat into the RX passband. This is overcome by delaying the 16.1GHz RX chirp on chip. The chip dissipates 500mW in continuous mode, with the option of a low-power, duty-cycled FM transmission mode (for close ranges) enabled via fast powering built into the transceiver.

Despite operation at approximately $f_T/2$, thin metallization and substrate losses affecting passives, and stringent density rules, the TX EIRP is 11dBm. The RX achieves a noise figure and maximum conversion gain of 8dB and 84dB, respectively. Beyond 100GHz, interfacing off-chip antennas on high-frequency laminates via flip-chip or wire-bonding gets prohibitive due to insertion loss, limited bandwidth, and lack of reproducibility. To overcome these, advanced packaging technologies (e.g., eWLB, FO-WLP, InFO) and on-chip antennas are viable solutions. This work explores sub-arrayed dipoles on chip that mitigate deleterious substrate modes when driven in-phase, yielding an efficiency of 62% compared to 52% for a single dipole.

The TX (see Fig. 9.4.1(a)) contains a PA driven by a cascade of transformer-coupled gain stages and two triplers. The 16GHz pre-drivers are current-biased differential amplifiers. Phase inversion for outer-code sequencing in the MIMO configuration is implemented in stage A_0 (Fig. 9.4.2(a)) driving the first tripler; and is enabled by the selection of diff-pair sets $M_{1,2}$ or $M_{3,4}$ via switches M_{S1} and M_{S2}, respectively. The 48GHz amplifier (A_1) is a pseudo-differential stage operating in Class-AB. The two triplers (see Fig. 9.4.2(b)) are based on harmonic self-mixing and circumvent the selectivity of injection-locked LC multipliers. The Class-AB operation of $M_{1,2}$ generates a second-harmonic ($2f_0$) voltage on node 'x', maximized via inductance L_{ss} resonant at $2f_0$. The differential f_0 drive at the gates of $M_{1,2}$, and the $2f_0$ swing at node 'x' generate antiphased 3rd-harmonic signals in each device via their nonlinear transfer. The fractional bandwidth for both triplers is 15%. A common-mode trap at the transformer load (see Fig. 9.4.2(b)) eliminates drain-modulation of transistor currents at $2f_0$ and ensures a clean supply distribution.

The 145GHz PA and LNA comprise four transformer-coupled, neutralized, pseudo-differential stages (see Fig. 9.4.2(c)). MOM capacitors, C_n, are chosen for unconditional stability over process corners at the expense of G_{MAG}. The PA, characterized standalone, has a P_{sat} of 7dBm. The LNA, optimized for noise (antenna customized to its Z_{opt} of 42Ω) and 1dB compression, is stagger-tuned for a flat frequency response. The PA and LNA draw 130mA and 30mA, respectively.

The sub-arrayed dipoles are drawn in the Al-passivation layer (see Fig. 9.4.2(d)). An artificial magnetic conductor is formed by the pattern metal-filled oxide, 100um thick substrate, and PCB reflector. It is optimized for phase coherence of direct and reflected paths. Partial cancellation of the dominant substrate modes is achieved when the dipoles, separated by 450µm, are driven in phase. TX and RX antennas are kept 3.7mm apart to minimize crosstalk (see Fig. 9.4.2(e)). Minimized bondwires and supply-decoupling networks on the back of the PCB reduce interference. The chip is surrounded by a 250µm thick RF absorber (Emerson and Cuming™ BSR-1, 25).

The RX (see Fig. 9.4.1(a)) comprises the 0.9V RF front-end, analog processing at IF and a replica of the TX buffering and upconversion. The LNA is interfaced to the RX mixer via transformer T_1 (see Fig. 9.4.3(a)). The switching quad (M_{3-6}) is loaded by a voltage-regulation loop that high-pass filters the TX-leakage beat and suppresses DC offsets due to limited mixer port isolation at 145GHz. The impedance of $M_{7,8}$ is lowered by the loop gain (illustrated in Fig. 9.4.3(c)), forming a first-order highpass filter. The dominant pole in this loop is at the output of folded-cascode amplifiers $A_{1,2}$, using Miller capacitance C_c. The RX passband corner frequency is 400kHz. Correlated noise from V_{CM} is subtracted at the differential output. The mixer drives a G_m stage (see Fig. 9.4.1(b)), which is AC-coupled to shunt-feedback amplifier A_{F1}. This interface prevents NF degradation, increases filtering, and transitions seamlessly from a 0.9V to a 1.8V supply. The G_m stage comprises an NMOS differential pair with PMOS loads. The AC-coupling is also designed for 400kHz, yielding an overall 40dB/dec filter roll-off (see Fig. 9.4.3(b)). A_{F1} is a two-stage opamp (see Fig. 9.4.3(c)). The degenerated CMFB amplifier introduces a zero ($R_{CM}C_{CM}$) to stabilize the loop. Because of AC-coupling, when V_{CM} is zero at startup, M_1 & M_2 are cut off. Ramping V_{CM} does not turn on A_{F1}. Current bleeders (M_3 and M_4), placed across the input pair trigger the CMFB loop with an increase of V_{CM}, and subsequently bias $M_{1,2}$ via the resistive feedback. The DAC frequency of the digital PLL (chirp generator) and off-chip ADC/DSP is 80MHz. The aliases and noise above 40MHz are suppressed by a fully differential, 6th-order Butterworth filter. The implementation and simulated performance of the filter are shown in Fig. 9.4.3(d), which uses the amplifier of Fig. 9.4.3(c) (sans current bleeders) for its three stages. Feedback stages A_{F1} and A_{F2} cumulatively provide an 8-bit gain programmability of 60dB. The programmable delay, τ, in Fig. 9.4.1(b) (0 to 80ps, 13ps steps) is designed at 16GHz using multiplexed inverters (see Fig. 9.4.4(a)) to compensate the LNA and PA delays. Delaying the reference chirp prior to mixing eliminates the TX-leakage tone by shifting it to a lower frequency with sufficient suppression.

The radar transceiver (see Fig. 9.4.7) is characterized using the Keysight™ PNAx and Vdi™ D-band extension module. Far-field measurements at 145GHz of power versus distance, r, are shown in Fig. 9.4.4(b), and agree with the expected $1/r^2$ gradient. The average EIRP is 11dBm with a ±0.5dB spread. The TX operates with a saturated PA and Fig. 9.4.4(b) shows a 3dB bandwidth of 127 to 154GHz.The RX response of Fig. 9.4.5(a), measured at 10MHz IF, shows a 3dB bandwidth of 138 to 151GHz. Different gain settings show a programmability of 57dB (±1dB). The measured TX-RX isolation is shown in Fig. 9.4.4(d). The RX noise is expressed as the effective isotropic noise figure (EINF) [2] since measuring the antenna standalone at 145GHz is not a trivial undertaking. The EINF is ~6dB at 10MHz IF. Adding the simulated antenna gain of 2.5dBi shows good agreement with the simulated RX NF of 8dB. The NF at 400kHz is 5dB higher due to flicker noise. The IF response measured with a 145GHz carrier is shown in Fig. 9.4.5(b). MIMO captures (2×2 assembly, Fig. 9.4.1(c)) of a heartbeat emulator at 2m using a 10GHz chirp signal are shown in Fig. 9.4.5(c). The speaker diaphragm driven by a Raspberry-Pi™ board has a 1mm displacement and is excited at 1Hz and 5Hz as seen from the ADC-Zedboard™-Matlab™ capture. The phase-angle variation is limited by fixed objects in the same range bin, as explained in the inset. MIMO experimentation along the azimuth is shown in Fig. 9.4.5(d). The angle of incidence of three closely spaced objects is evaluated using the MUSIC classification algorithm.

Compared to state-of-the-art integrated radars (see Fig. 9.4.6), this work shows a superior EIRP and NF with a low power consumption.

Acknowledgement:
The work is supported by Panasonic and the EC H2020 grant 783119.

References:
[1] Q. Shi et al., "A Self-Calibrated 16GHz Subsampling-PLL-Based 30µs Fast Chirp FMCW Modulator with 1.5GHz Bandwidth and 100kHz rms Error," *ISSCC*, Paper 26.1, Feb. 2019.
[2] K. Khalaf et al., "A 60-GHz 8-Way Phased-Array Front-End With T/R Switching and Calibration-Free Beamsteering in 28-nm CMOS", *IEEE JSSC*, vol. 53, no. 7, pp. 2001-2011, July 2018.
[3] J. Grzyb et al., "A 210–270-GHz circularly polarized FMCW radar with a single-lens-coupled SiGe HBT chip," *IEEE Trans. THz Sci. Technol.*, vol. 6, pp. 771-783, 2016.
[4] B. Ginsburg et al., "A 160GHz Pulsed Radar in 65nm CMOS," *IEEE JSSC.*, vol. 49, no. 4, pp. 984-995, April 2014.
[5] B. Ginsburg et al., "A Multimode 76-to-81GHz Automotive Radar Transceiver with Autonomous Monitoring," *ISSCC,* pp. 158-159, 2018.
[6] H.J. Ng et al., "Miniaturized 122 GHz System-on-Chip Radar Sensor with On-Chip Antennas Utilizing a Novel Antenna Design Approach," *IEEE MTT-S Int. Microwave Symp.,* 2016.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

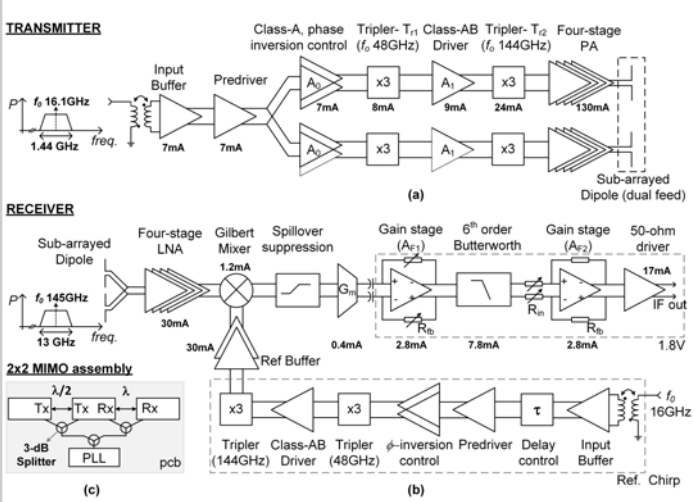


Figure 9.4.1: Block diagram of the 145GHz FMCW transceiver, showing (a) the transmitter, (b) the receiver, and (c) the 2×2 MIMO assembly on a PCB.

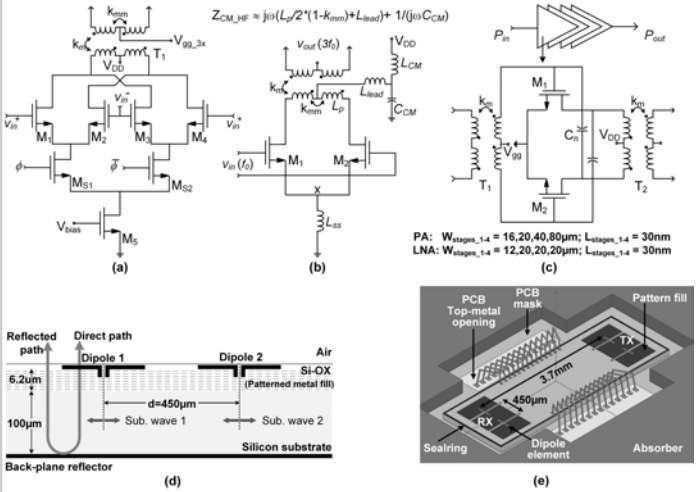


Figure 9.4.2: Circuit diagram of (a) phase inversion stage, A_0, of Fig. 9.4.1(a), (b) frequency tripler (c) PA and LNA. (d) Cross section of the sub-arrayed dipole, (e) 3D view of the chip-mount and packaging.

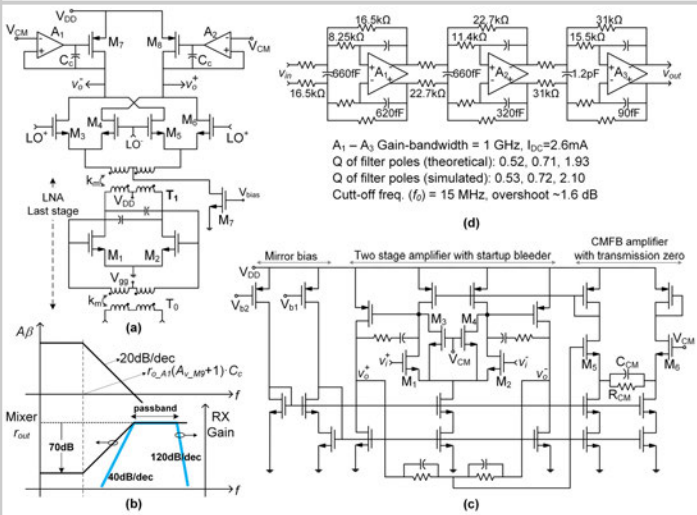


Figure 9.4.3: Circuit diagram of (a) RX mixer and TX-leakage suppression filter, (b) Illustration of the RX frequency response, (c) two-stage amplifier, A_{F1}, of Fig. 9.4.1(b), (d) 6th-order Butterworth filter and simulation results.

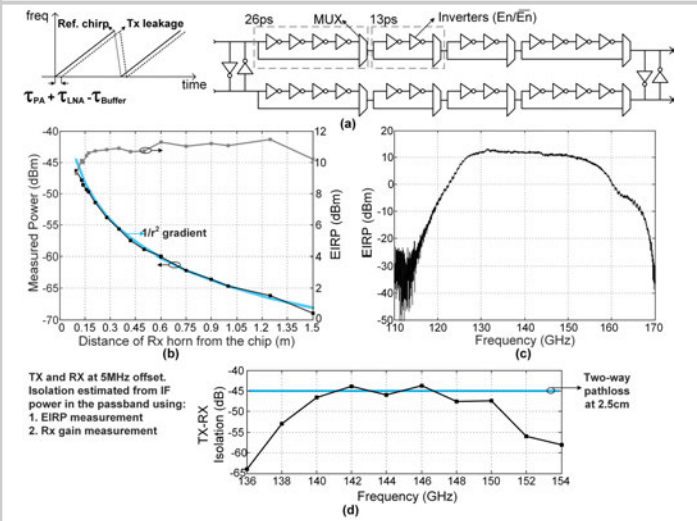


Figure 9.4.4: (a) Programmable on-chip delay for leakage suppression, (b) far-field measures of EIRP for varying distance (c) transfer characteristic of the transmitter, (d) TX-RX leakage measurement.

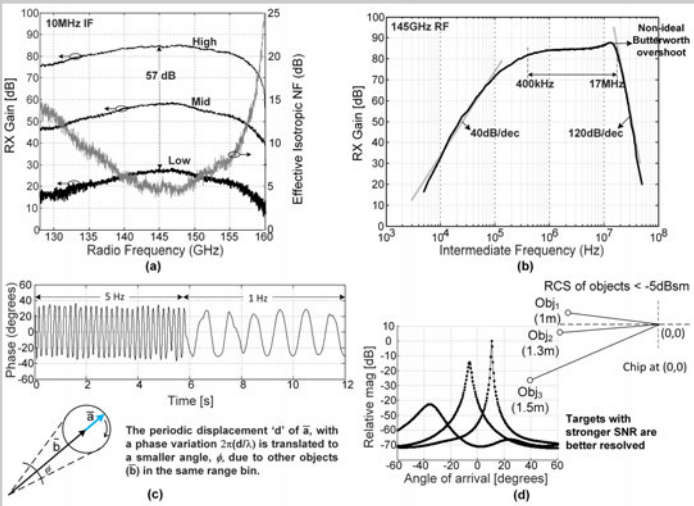
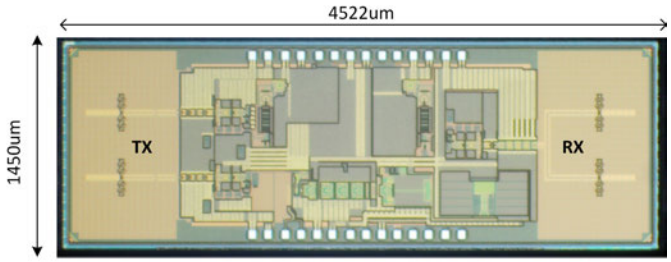


Figure 9.4.5: (a) RX gain and EINF, (b) IF response of the receiver, (c) MIMO capture of heartbeat emulator at 2m, with 1mm displacement, (d) MIMO experiment showing angular resolution of targets using the MUSIC algorithm.

Reference	This Work	Wuppertal, Trans. THz 2016 [3]	TI, JSSC 2014 [4]	TI, ISSCC 2018 [5]	IHP, IMS 2016 [6]	I. Nasr, Google, IFX, JSSC 2016
Technology	28nm CMOS	0.13μm SiGe	65nm CMOS	45nm CMOS	0.13μm SiGe	0.35μm SiGe
Radar type	FMCW	FMCW	Pulsed	FMCW	FMCW	FMCW
Antenna	On Chip	On Chip	In Package	In Package	On Chip	In Package
RF frequency [GHz]	145	240	160	79	122	60
RF Bandwidth [GHz]	13	60	7	4	5.8	7
Range resolution [cm]	1.1	0.3	2.1	3.8	2.6	2.1
Channels	1TX - 1RX	1TX - 1RX	4TX - 4RX	3TX - 4RX	1TX - 1RX	2TX - 4RX
Single-path TX Power/EIRP [dBm]	11.5 (EIRP)	5	4	10.8	5	4
RX Gain [dB]	80	10	42.5	NA	NA	19
Noise Figure [dB]	8	21	22.5	18	NA	9.5
IF Bandwidth [MHz]	17	18	100	15	NA	1
Chip size [mm²]	6.5	3.2	20	22	10.35	20.2
Power diss. reported for all channels [mW]	500	1800	2200+º	3500++	NA	990*

+ Includes on-chip PLL and ADC
++ Includes on-chip PLL, ADC and DSP
º Power dissipation in pulsed operation
* RF front-end only with on-chip VCO

Figure 9.4.6: Table comparing this work with state-of-the-art integrated radars.



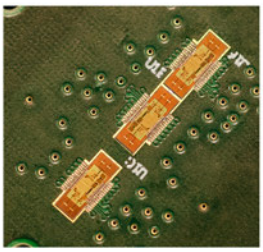


Figure 9.4.7: Die micrograph and zoomed-in view of the MIMO assembly board showing the three TRX chips.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

9.5 An 80Gb/s 300GHz-Band Single-Chip CMOS Transceiver

Sangyeop Lee[1], Ruibing Dong[1], Takeshi Yoshida[1], Shuhei Amakawa[1], Shinsuke Hara[2], Akifumi Kasamatsu[2], Junji Sato[3], Minoru Fujishima[1]

[1]Hiroshima University, Higashihiroshima, Japan
[2]National Institute of Information and Communications Technology, Koganei, Japan
[3]Panasonic, Yokohama, Japan

IEEE Standard 802.15.3d, published in October 2017, defines a high-data-rate wireless physical layer that enables up to 100Gb/s using the lower THz frequency range between 252 and 325GHz (hereafter referred to as the "300GHz band"). It stipulates that the 300GHz band be channelized into thirty-two 2.16GHz-wide channels (Fig. 9.5.1) or a smaller number of wider channels whose bandwidths are all integer multiples of 2.16GHz. This paper presents a CMOS transceiver (TRX) chip targeted at channels 49 through 51 and 66 of 802.15.3d (Fig. 9.5.1). The TRX was fabricated using a 40nm CMOS process. There have been reports on solid-state transceivers (TRXs) operating in or near the 300GHz band [1-6]. Some of these [1-3] were TX/RX or block-level chipsets, which can enjoy more flexibility in design and independent optimization of TX and RX. They successfully achieved ≥64Gb/s. On the other hand, single-chip TRXs [4-6] did not always reveal achievable data-rates nor were capable of supporting quadrature amplitude modulation (QAM). Nevertheless, eventual development of full-featured single-chip TRXs is desirable especially for applications requiring deployment of many TRXs, as is envisioned implicitly by 802.15.3d. The single-chip QAM-capable CMOS TRX presented herein is an outcome of efforts in that direction.

Since the 40nm CMOS process that we used has a relatively low f_{max}, we adopt PA- and LNA-less architecture. A simplified schematic of the TRX is shown in Fig. 9.5.1. It operates either in TX or RX mode. These modes share the TX part, which serves as an LO multiplier chain in RX mode. The TX mode is architecturally similar to the mixer-last TX in [7], but the input is baseband (BB) as opposed to IF. The input signal is upconverted by a quadrature modulator and the IF is superposed on LO (~133GHz) by the modulator itself. The resulting (LO±IF) are squared by frequency doublers (or square mixers [7]), producing $(LO\pm IF)^2$. $(LO+IF)^2$ and $(LO-IF)^2$ are then fed, respectively, to the positive (+) and negative (–) ports of a rat-race balun variant (double-rat-race [7]), and the difference (Δ) port outputs the desired RF_{TX} signal, 4LO·IF (~266GHz). In RX mode, the BB input ports are terminated appropriately and the modulator works only as an LO buffer. The doublers generate LO^2 and the sum (Σ) port of the double-rat-race feeds the RX part with $2LO^2$ (~266GHz). The mixer-first direct-conversion RX part consists of a low-noise mixer (LNM) and BB amplifiers.

The quadrature modulator in the shared TX part is required to leak a certain amount of LO to the output, and therefore well-balanced double-balanced mixers cannot be used. There exists an optimum leaked LO power that gives the highest RF power in TX mode (Fig. 9.5.2). To make the amount of LO leakage controllable, we use what could be called a "semidoubly-balanced quadrature mixer," shown in Fig. 9.5.2. It is schematically similar to a double-balanced mixer but BB inputs are not balanced. What would have been the "BB inverse ports" are used to tune the coefficient α in the output, (αLO+IF). Different values of α are chosen in TX and RX modes. The last stage of the IF amplifier is a stacked doubler driver having large output voltage amplitude (Fig. 9.5.2). This configuration contributes significantly to high RF (LO) power in TX (RX) mode. The stacked driver and the doubler were optimized together as a single block. This split of the IF amplifier into the doubler driver and the rest facilitated wideband design.

Noise performance is of utmost importance for the dedicated RX part. However, conversion gain of the downconversion mixer cannot be greater than unity (0dB) because of low f_{max}. Clearly, the noise factor and conversion loss of the mixer must both be as low as possible, but performance metrics that require bona fide gain (>1), such as the noise measure, do not help quantify and compare performance of various mixer configurations. We arrived at the somewhat ordinary low-noise mixer configuration, shown in Fig. 9.5.3, by looking at the Engberg-Gawler factor given by $EG=(1-1/G_1)[1-M_1/(F_2-1)]$ [8], where G_1 and M_1 are the available (conversion) gain and the (SSB) noise measure of the 1st stage of two cascaded stages, respectively, and F_2 is the noise factor of the 2nd stage. In our case, the "1st stage" includes the mixer and a buffer amplifier (Fig. 9.5.3), and the "2nd stage" is the BB amplifier. The latter was optimized independently of the "1st stage" and therefore F_2 has a fixed value. EG serves as a useful performance metric even when $G_1<1$ and $M_1<0$: the larger (more to the right on a real number line) the value of EG, the better. The LO signal, LO^2, for downconversion coming from the TX part and RF_{RX} from the receiving antenna port are combined by a rat-race and applied to the MOSFET gates of the "1st stage." Simulated noise figure (NF) and conversion gain of the "1st stage" are fairly flat over a 20GHz BB bandwidth (Fig. 9.5.3). This was accomplished by high-frequency gain peaking of the inter-stage matching network (Fig. 9.5.3).

Figure 9.5.4 shows the schematic of the entire TRX. LO_{IN} is supplied from an external signal generator. After being split into in-phase and anti-phase branches, it goes through frequency triplers before reaching the quadrature modulators (SDBQMs). Branch-line quadrature hybrids are used to produce quadrature phase in TX and RX parts. The four square mixers perform final upconversion. The double-rat-race power-combines RF and cancels out LO^2 in TX mode, whereas it power-combines LO^2 in RX mode. Transmitting and receiving antennas are to be connected to RF_{TX} and RF_{RX} ports, respectively. The RX part has I and Q branches.

Figure 9.5.5 shows the TRX characterization setup and measurement results. The TRX chip was mounted on a printed circuit board (PCB). LO_{IN} and dc power were supplied to the chip through bond wires. BB input signals were generated by an arbitrary waveform generator and were fed through a multi-tip probe. TX output signal RF_{TX} was led to a WR3.4-band block downconverter via a waveguide probe, and a spectrum analyzer measured the power. The frequency dependence of TX output power was measured by sweeping the frequency of the sinusoidal BB input at a fixed LO^2 frequency, f_{LO2}. The input-power dependence of the TX output power was measured at a fixed BB frequency of 5GHz. The observed LO leakage reduction can be understood analogously to "blocking" in a receiver, in which a powerful interferer (IF in Fig. 9.5.1 in this case) accompanying a weaker desired signal (LO in Fig. 9.5.1) causes gain compression. RX conversion gain and NF were measured using a multifunction vector network analyzer and a frequency extender through observing one of the four BB outputs. TX mode and RX mode consumed dc power of 890mW and 897mW, respectively.

Figure 9.5.6 shows the wireless link measurement setup. A pair of 24dBi horn antennas were used. Output BB signals from the RX were analyzed using a real-time oscilloscope with vector signal demodulation and channel equalization capabilities. Signal constellations and power spectra are shown for frequencies and bandwidths corresponding to channels 49, 50, and 66 of 802.15.3d (Fig. 9.5.1). A data-rate of 80Gb/s over a distance of 3cm was achieved with 16QAM. Recently reported THz TRXs are compared in Fig. 9.5.6. A chip micrograph of our TRX is shown in Fig. 9.5.7. The die size is 4.92×2.25mm^2.

Acknowledgment:
This work was financially supported by the Ministry of Internal Affairs and Communications of Japan.

References:
[1] I. Kallfass et al., "Towards MMIC-Based 300 GHz Indoor Wireless Communication Systems," *IEICE Trans. Electron.*, vol. E98-C, no. 12, pp. 1081-1090, Dec. 2015.
[2] P. Rodríguez-Vázquez et al., "A 65 Gbps QPSK One Meter Wireless Link Operating at a 225–255 GHz Tunable Carrier in a SiGe HBT Technology," *IEEE Radio Wireless Symp.*, pp. 146-149, Jan. 2018.
[3] H. Hamada et al, "300-GHz 100-Gb/s InP-HEMT Wireless Transceiver Using a 300-GHz Fundamental Mixer," *IEEE IMS*, pp. 1480-1483, June 2018.
[4] J. Al-Eryani et al., "Fully Integrated Single-Chip 305–375-GHz Transceiver with On-Chip Antennas in SiGe BiCMOS," *IEEE Trans. Terahertz Sci. Technol.*, vol. 8, no. 3, pp. 329-339, May 2018.
[5] S. Kim et al., "300 GHz Integrated Heterodyne Receiver and Transmitter with On-Chip Fundamental Local Oscillator and Mixers," *IEEE Trans. Terahertz Sci. Technol.*, vol. 5, no. 1, pp. 92-101, Jan. 2015.
[6] J.-D. Park et al., "A 260 GHz Fully Integrated CMOS Transceiver for Wireless Chip-to-Chip Communication," *IEEE Symp. VLSI Circuits*, pp. 48-49, June 2012.
[7] K. Takano et al., "A 105Gb/s 300GHz CMOS Transmitter," *ISSCC*, pp. 308-309, Feb. 2017.
[8] J. S. Engberg and G. B. Gawler, "Significance of the Noise Measure for Cascaded Stages," *IEEE Trans. Circuit Theory*, vol. 16, no. 2, pp. 259-260, May 1969.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

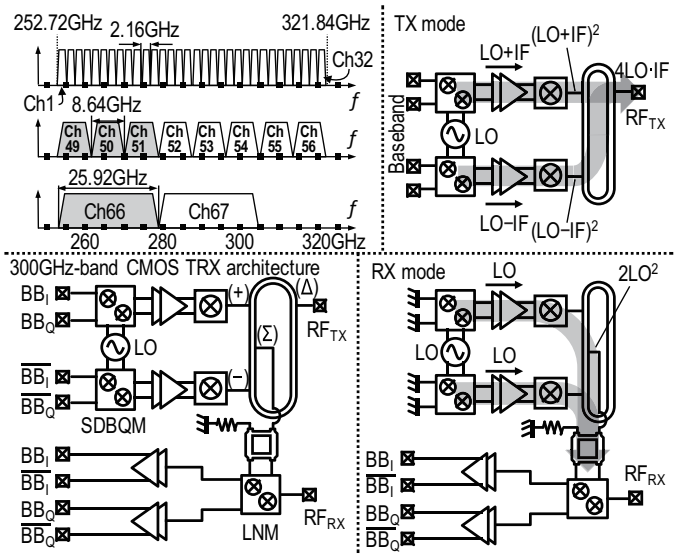


Figure 9.5.1: 802.15.3d channelization, TRX architecture, and TX & RX modes.

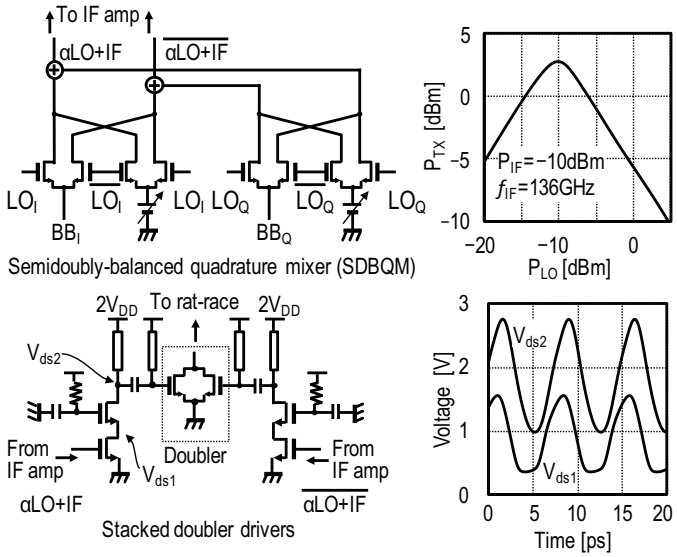


Figure 9.5.2: Quadrature modulator (SDBQM); stacked doubler driver at IF.

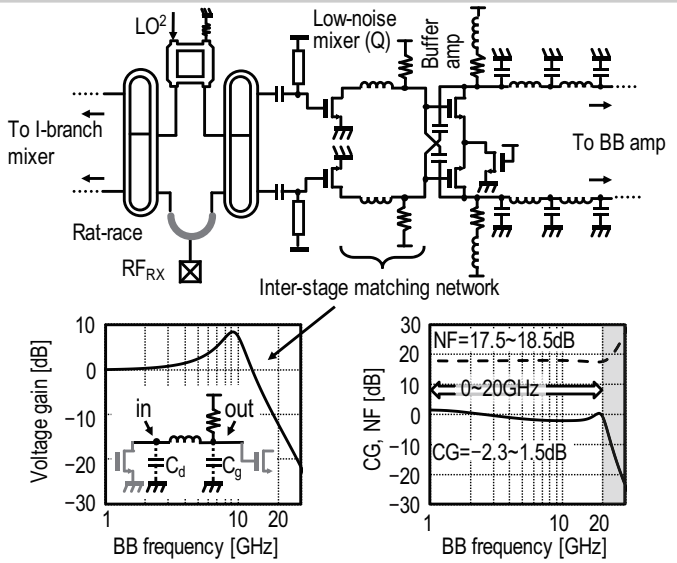


Figure 9.5.3: RX low-noise mixer and buffer amplifier.

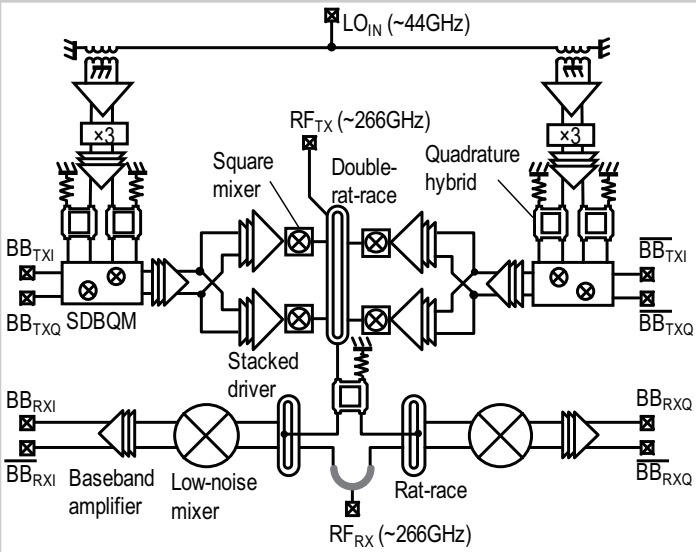


Figure 9.5.4: Schematic of 300GHz-band CMOS transceiver.

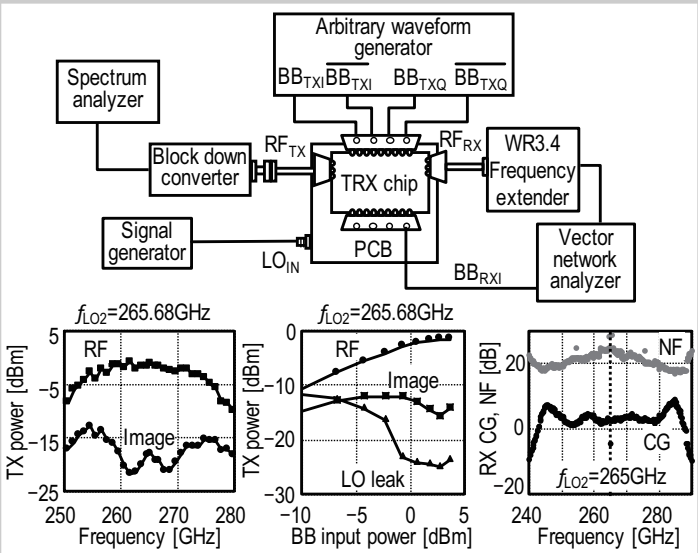
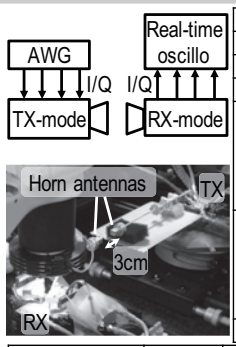


Figure 9.5.5: TRX characterization setup and measurement results.



Channel	Ch. 49	Ch. 50	Ch. 66
Center freq.	257.04GHz	265.68GHz	265.68GHz
Modulation	16QAM	16QAM	16QAM
Data rate	28.16Gb/s	28.16Gb/s	80Gb/s
Constellation (Equalized)			
Spectrum			
EVM	10.9%rms	11.3%rms	12.0%rms

	[1]	[2]	[3]	[4]	[5]	This work
Technology	35nm GaAs	0.13μm SiGe HBT	80nm InP	0.13μm SiGe BiCMOS	250nm InP	**40nm CMOS**
Freq. [GHz]	300	240	287	340	298	**265.68**
Modulation	QPSK	QPSK	16QAM	FM/AM	-	**16QAM**
Single chip	N	N	N	Y	Y	**Y**
Pout [dBm]	−4	8.5	9.5	0.1	−2.3	**−1.6**
Pdc [W]	-	1.96	-	1.7	0.45	**1.79**
Data rate [Gb/s]	64	65	100	-	-	**80**

Figure 9.5.6: Wireless measurement setup, constellations, and comparison.

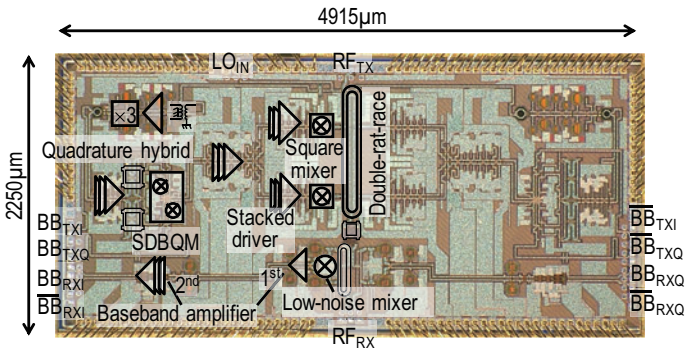


Figure 9.5.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

9.6 A 42.2Gb/s 4.3pJ/b 60GHz Digital Transmitter with 12b/Symbol Polarization MIMO

Chintan Thakkar, Stefan Shopov, Anandaroop Chakrabarti, Shuhei Yamada, Debabani Choudhury, James Jaussi, Bryan Casper

Intel, Hillsboro, OR

With an intent to galvanize high-speed wireless PAN connectivity in modern-day handheld computing devices around the unlicensed 57-to-71GHz band, the IEEE 802.11ay standard [1] supports 2-to-4× bonding of 2.16GHz channels (i.e. 4.32 to 8.64GHz), constellations up to 64QAM, and spatial MIMO. To support such wideband TX, prior linear 60GHz solutions [2] employ baseband modulation followed by a linear upconversion and PA with saturated power of ~10dBm. For OFDM and high-order single-carrier (SC) modulations, the PA operates in significant power back-off, leading to poor average output-to-DC (Pout-to-Pdc) power efficiency (<1.5%). Additionally, for spectral compliance, these solutions use off-chip pulse-shaping with 8-to-12b, 25-to-65GS/s arbitrary waveform generators (AWG). Implementing such high-performance DACs on-chip would incur significant power overhead in handheld devices.

To improve TX power efficiency, digital I/Q [3] or polar modulated [4,5] solutions using a 60GHz switching PA have been shown to lower power consumption by turning off PA segments at power back-off. The TX in [3,4] operate at ≤1.76Gsym/s with a 16-QAM data-rate of 6 to 7Gb/s, which is far below the 802.11ay mandated >20Gb/s throughput. These designs also have limited amplitude modulation (AM) depth (full-scale to LSB ratio) of <14dB, which is insufficient to implement SC-64QAM and/or OFDM. While the digital TX in [5] demonstrates 27.8Gb/s using 3.46Gsym/s 16QAM and dual-polarization (DP) MIMO, it lacks pulse shaping for spectral compliance. This work (Fig. 9.6.1) demonstrates an energy-efficient digitally modulated polar TX in 0.9V 28nm CMOS supporting both SC and OFDM signaling. With DP-MIMO, the design applies on-chip configurable pulse shaping for spectrally compliant, low EVM throughput up to 28.8Gb/s with 3.52Gsym/s 16QAM, and achieves baud-rate signaling with low EVM up to 42.2Gb/s with 5.28Gsym/s 16QAM and 3.52Gsym/s 64QAM.

While a polar TX is power efficient, supporting multi-channel bonding with even modest oversampling ratios for pulse shaping necessitates direct digital modulation of mm-wave amplitude/phase modulation (AM/PM) blocks at switching rates >10GHz. The PM in [5] is implemented using an LC injection-locked oscillator with >3dB of PM-AM distortion and a capacitor DAC that trades off phase-range with switching speed, limiting operation to <3.5GHz. To avoid such limitations, high-speed PM in this work is achieved by blending quadrature LO phases that are amplitude weighted using current-steering 5b+5b I+Q modulated Gilbert-quads (Fig. 9.6.2). At the PM input, LO frequency configurability is achieved with a broadband 2/3rd LO frequency input CML-based quadrature divider, followed by Class-C biased frequency/phase tripling and amplification. At the PM output, PM-AM distortion is minimized by limiting amplitude variation to within ±0.3dB (measured in foreground at the PA output) from the subset of all possible I+Q combinations, while simultaneously achieving a measured RMS / worst-case step of 1.46° / 5.23° at 63GHz. At this frequency, the LO chain and PM consume 42.5mW.

The PM signal is fed to a 2-stage driver amplifier to maximize input swing into the differential digital PA (DPA) for AM. For higher drain efficiency than cascode switching [5], digital AM is implemented using tail switching with minimal voltage headroom (Fig. 9.6.2, similar to [3,4]). When transitioning to the on and off states, switching speed is set by the tail switch resistance and large-signal 1/Gm of the LO pair respectively. While this topology is scalable to high-speed modulation, it typically suffers from LO feedthrough at low AM settings due to Cgd coupling, which limits the AM depth (13.5dB in [3]). In order to achieve a higher extinction ratio to support up to OFDM/SC 64QAM, this work uses compact, per-segment Cgd neutralization to improve AM depth to 35.3dB (measured) with 6b binary segmentation. At 63GHz, the DPA is measured using a probed die with ±7.7° of AM-PM distortion (foreground corrected in the PM), 11.5dBm maximum output power, and state-of-art 40%/24% peak/-6dB drain efficiency. Including the driver chain and DPA, the AM path consumes 51mW of peak power.

Within both AM/PM paths, spectral shaping in-band is implemented on-chip with a 4-tap FIR by dividing each mm-wave switching segment into 4 equal weight legs (similar to [6]) (Fig. 9.6.3) for the equivalent of 4-fold linear interpolation and zero-order hold. Spectral shaping can be turned on/off using selection logic to switch between filtering or baud-rate signaling modes. To push digital spectral aliases out in frequency for sufficient out-of-band rejection, prior digital TX solutions implement the equivalent of 6-to-8X samples per UI [3,4] at 1.5 to 1.76Gsym/s. However, with higher symbol rates associated with channel bonding, the digital switching and routing power of a high oversampling ratio (OSR) would be significant. To reduce this overhead, the OSR is limited to 4×. Spectral aliases, that are now closer in frequency, are attenuated by employing current-starved inverter-based programmable slew-rate control on the digital AM/PM drivers. This overlays a lowpass filter centered around the LO frequency that, for marginal in-band attenuation, provides substantial out-of-band rejection.

To communicate with 2× higher spectral efficiency, this TX supports spatial MIMO in a single antenna footprint [5] with a dual TX chain. The DP antenna (Fig. 9.6.7) is designed on a 6-layer, 22-mil PCB with RO4350B/RO4450F core/prepreg. A proximity coupled excitation with stacked patch structure is used to achieve the desired bandwidth (57 to 71GHz) on both V and H-pols. A differential excitation topology enhances polarization isolation to >35dB based on simulated radiation patterns (Fig. 9.6.4). To maintain high signaling bandwidth across the TX circuitry and antenna, the 28nm CMOS TX prototype is copper-pillar bumped and flip-chip attached to the PCB antenna. Digital polar modulation is implemented on-chip using per-pol 10b+6b PM+AM SRAM, 32:1 serialization per-bit and programmable PM-AM digital modulation delay control.

The DP TX-antenna PCB assembly is measured over-the-air (OTA) by an RX DP circular horn antenna and V/H LNA+mixer chains to downconvert to an IF between 10 and 13GHz (selected based off symbol-rate). With both TX pols on and programmed with uncorrelated random symbol sequences, V/H IF measurements are simultaneously performed with an 80GS/s real-time oscilloscope, followed by linear bandwidth equalization of the downconversion path. With a TX LO of 63GHz, baud-rate SC signaling, and 2-to-3 tap TX pre-distortion, simultaneous V/H DP-MIMO constellations of 5.28Gsym/s (3× channel bonding) 16QAM (8b/symbol, 42.2Gb/s aggregate), 3.52Gsym/s (2× channel bonding) 32QAM (10b/symbol, 35.2Gb/s aggregate) and 3.52Gsym/s (2× channel bonding) 64QAM (12b/symbol, 42.2Gb/s) are measured with an average EVM over 4096 points per-pol and with respect to RMS amplitude of -20.1dB, -24.3dB and -24.7dB respectively (Fig. 9.6.4). With a 4× OSR, and on-chip FIR and slew-rate digital driver control enabled, simultaneous V/H DP-MIMO constellations of 3.52Gsym/s (2-channel bonding) SC-16QAM (28.2Gb/s) are measured with an average EVM of -21.3dB. Figure 9.6.5 shows the corresponding spectral compliant frequency response, validating on-chip pulse-shaping capability. Similarly, 1.76Gsym/s (1-channel) DFT-s-OFDM 64QAM is measured with -25.5dB EVM and a spectrally compliant response, validating the design of high AM/PM depth and resolution. The TX also achieves spectrally-shaped 7.04Gb/s SC-16QAM per-pol across three 802.11ay frequency channels centered at 60.48, 62.64, and 64.80GHz, and EVM variation <1.5dB, thus demonstrating LO configurability.

Compared to the state-of-art linear 60GHz TX with off-chip pulse shaping [2] (Fig. 9.6.6), this polar TX while transmitting spectrally shaped 28.2Gb/s DP 16QAM achieves 1.8×/3.1× better Pout-to-Pdc efficiency including/excluding pulse shaping digital power from clock/data distribution and AM/PM drivers. Compared to the prior SC-16QAM digital TX with on-chip pulse shaping [3], this work includes OFDM signaling, achieves 21.8dB higher AM depth, 4.7× higher data-rate, 5.1dB better TX EVM and 1.5× better Pout-to-Pdc efficiency. Compared to the prior baud-rate signaling 8b/symbol DP-MIMO TX [5], this work achieves 1.5× higher throughput, 1.8× better power efficiency, and 12b/symbol DP-MIMO. Such a power-efficient TX for spectrally-dense signaling could be compelling for high-performance PAN connectivity.

Acknowledgements:
The authors thank S. Daneshgar and K. Dasgupta for technical discussions, B. Horine and T. Nguyen for assistance with prototyping and testing, and M. Haycock for management support.

References:
[1] IEEE 802.11ay [Online]: http://www.ieee802.org/11/Reports/tgay_update.htm
[2] J. Pang et al., "A 128-QAM 60GHz CMOS Transceiver for IEEE 802.11ay with Calibration of LO Feedthrough and I/Q Imbalance," *ISSCC*, pp. 424-425, Feb. 2017.
[3] J. Chen et al., "A Digitally Modulated mm-Wave Cartesian Beamforming Transmitter with Quadrature Spatial Combining," *ISSCC*, pp. 232-233, Feb. 2013.
[4] K. Khalaf et al., "Digitally Modulated CMOS Polar Transmitters for Highly-Efficient mm-Wave Wireless Communication," *IEEE JSSC*, vol. 51, no. 7, pp. 1579-1592, July 2016.
[5] S. Daneshgar et al., "A 27.8Gb/s 11.5pJ/b 60GHz Transceiver in 28nm CMOS with Polarization MIMO," *ISSCC*, pp. 166-167, Feb. 2018.
[6] R. Bhat, J. Zhou, and H. Krishnaswamy, "A >1W 2.2GHz Switched-Capacitor Digital Power Amplifier with Wideband Mixed-Domain Multi-Tap FIR Filtering of OOB Noise Floor," *ISSCC*, pp. 234-235, Feb. 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

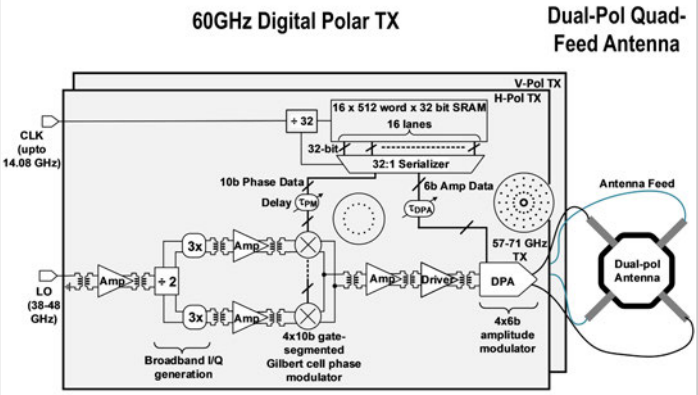


Figure 9.6.1: Block diagram of 60GHz dual-polarization (DP) MIMO digital polar TX circuitry, and differential quad-feed antenna for high cross-polarization isolation.

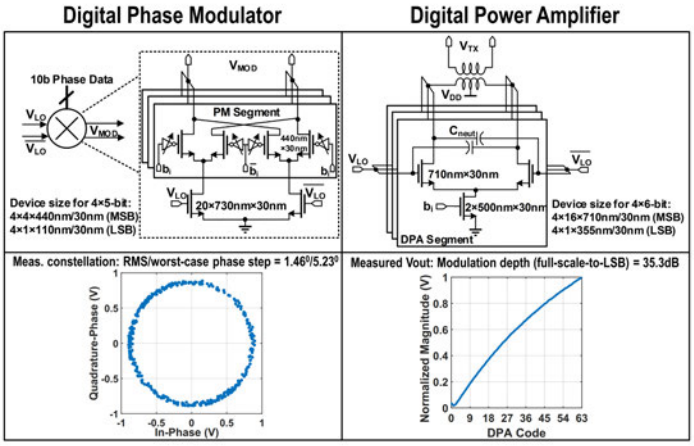


Figure 9.6.2: Schematics of digital current-switched phase modulator and digital tail-switched power amplifier with built-in 4-tap FIR segmentation, and measured amplitude/phase modulation (AM/PM) characteristics.

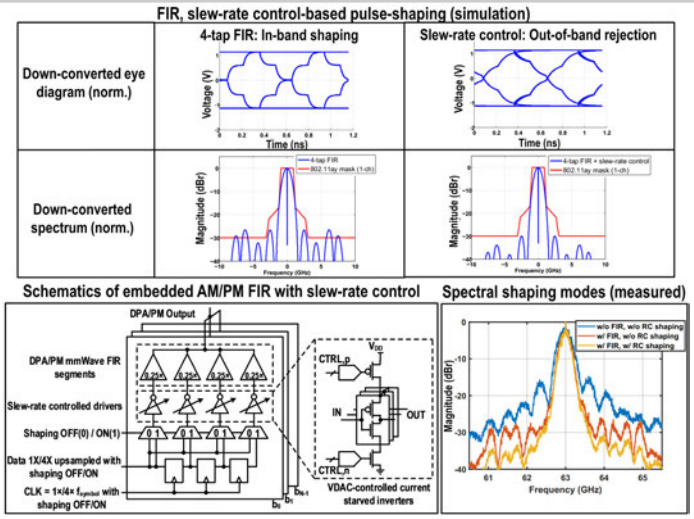


Figure 9.6.3: Pulse shaping using AM/PM-embedded mm-wave FIR for in-band shaping and slew-rate control for alias attenuation: simulation (top), schematics (bottom, left) and measurements (bottom, right).

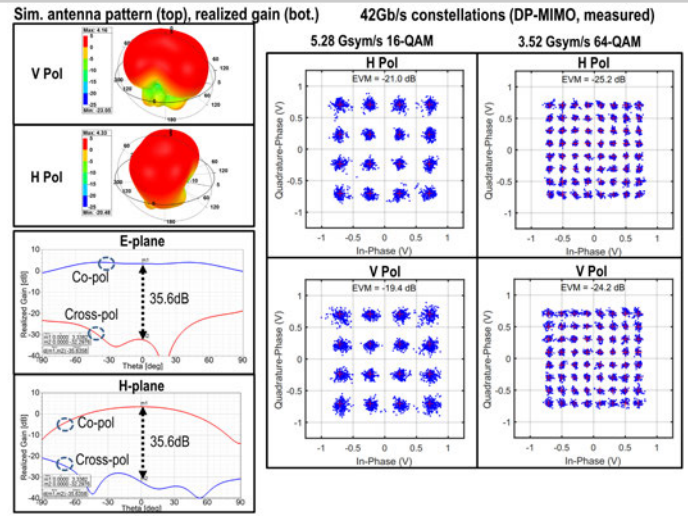


Figure 9.6.4: Simulated V/H antenna radiation patterns and realized gains (left), and measured 42Gb/s TX constellations of baud-rate simultaneous V/H DP-MIMO with 5.28Gsym/s 16QAM and 3.52Gsym/s 64QAM (right).

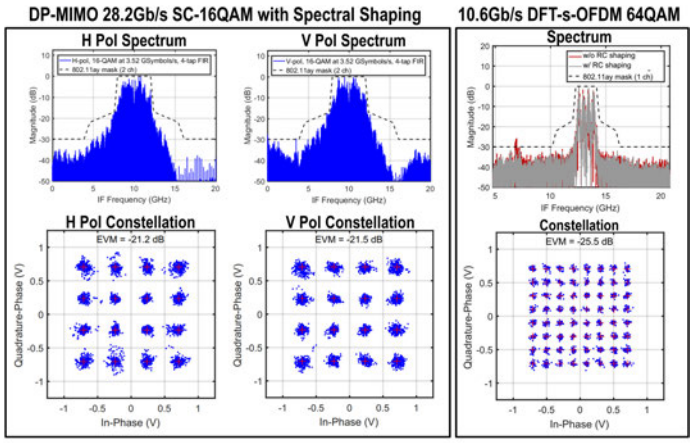


Figure 9.6.5: Measured TX spectrum (top), constellations (bottom) of pulse-shaped 3.52Gsym/s simultaneous V/H DP-MIMO 16QAM (28.2Gb/s, left), and 1.76Gsym/s 64-QAM OFDM (10.6Gb/s, right) with/without slew-rate control of AM/PM digital drivers.

Metric	[2] ISSCC 2017	[3] ISSCC 2013	[5] ISSCC 2018	This Work	
CMOS Technology	65nm	65nm	28nm	28nm	
TX Architecture	Linear	Digital I/Q	Digital Polar	Digital Polar	
Antenna	Single-pol off-PCB horn	Single-pol 2X PCB patches	Dual-pol PCB patch	Dual-pol differential PCB patch	
Pout,max	10.5dBm	9.6dBm*	4.0dBm	11.5dBm	
PA DE (max. / -6dB back-off)	Unavailable	28.5 / 15%	Unavailable	40 / 24%	
Modulation Depth	Not applicable	13.5dB	18dB	35.3dB	
Spectral Shaping	Off-chip Keysight AWG	On-chip	None	On-chip	None
Samples per UI	>7.1×	8×	1×	4×	1×
Modulation	SC-64QAM	SC-16QAM	SC-16QAM	SC-16QAM	SC-64QAM
MIMO Streams	1	1	2	2	2
Spectral Efficiency	6 b/sym	4 b/sym	8 b/sym	8 b/sym	12 b/sym
Symbol Rate	7.04 Gsym/s	1.50 Gsym/s	3.48 Gsym/s	3.52 Gsym/s	3.52 Gsym/s
Data-Rate	42.2 Gb/s	6.0 Gb/s	27.8 Gb/s	28.2 Gb/s	42.2 Gb/s
Avg. TX EVM	-25.5 dB	-16.2 dB	Unavailable	-21.3 dB	-24.7 dB
Pout,avg./pol.	4.0 dBm	0.7 dBm*	Unavailable	5.6 dBm	6.0 dBm
Pdc (total)	169 mW*	65 mW*	210 mW	272 mW	182 mW
Pout,avg/Pdc/pol.	1.5%*	1.8%*	–	2.7%	4.4%
Power Efficiency	4.0 pJ/b*	10.8 pJ/b*	7.6 pJ/b	9.7 pJ/b	4.3 pJ/b

*Excludes pulse shaping power, *Normalized to single-element

Figure 9.6.6: Performance summary and comparison with state-of-the art 60GHz TXs.

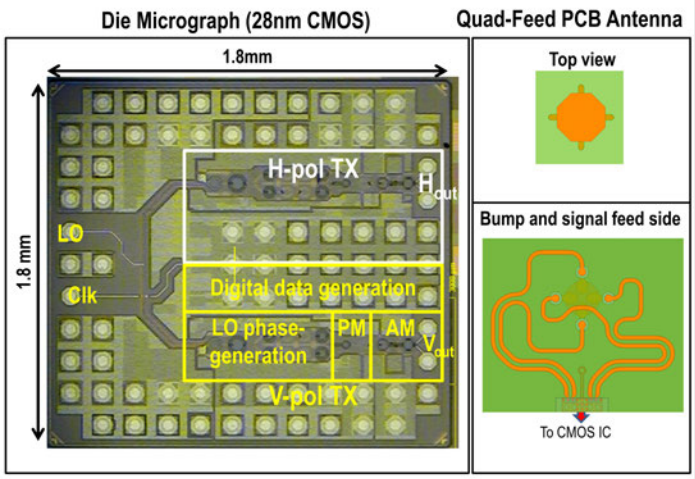


Figure 9.6.7: 28nm CMOS die micrograph (left), and quad-feed PCB antenna design (right).

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

9.7 A Scalable 71-to-76GHz 64-Element Phased-Array Transceiver Module with 2×2 Direct-Conversion IC in 22nm FinFET CMOS Technology

Stefano Pellerano[1], Steven Callender[1], Woorim Shin[1], Yanjie Wang[2], Somnath Kundu[1], Abhishek Agrawal[1], Peter Sagazio[1], Brent Carlton[1], Farhana Sheikh[1], Arnaud Amadjikpe[1], William Lambert[3], Divya Shree Vemparala[1], Mark Chakravorti[1], Satoshi Suzuki[1], Robert Flory[1], Chris Hull[1]

[1]Intel, Hillsboro, OR; [2]Hillsboro, OR; [3]Intel, Chandler, AZ

Fifth-generation cellular communication standards (5G) target Gb/s data-rates, pushing the industry beyond the sub-6GHz bands. Tens of GHz of spectrum are available in the frequency bands from 30 to 300GHz. To maintain acceptable link budgets with sufficient antenna apertures, arrays are typically required at these frequencies and electrical beam steering is needed to retain spatial coverage. For such complex systems, highly-integrated, low-cost and energy-efficient SoCs are desirable to enable volume deployment. FinFET technologies are an ideal candidate to tackle this challenging integration, given the excellent balance between density and RF/mm-wave performance that has been recently demonstrated [1].

The block diagram of the 2×2 transceiver IC is shown in Fig. 9.7.1. Integrating 4 elements minimizes mm-wave routing losses on the package, since IC size is comparable with required antenna pitch at E-band. Stacked vias are used to connect each of the four mm-wave die bumps to the respective antenna feed, avoiding any mm-wave routing on package and achieving a simulated loss of <0.5dB. A 2×2 unit array also allows for a more cost-effective known-good-die solution for larger arrays. The adopted direct-conversion architecture reduces power consumption and area as IF mixers/LOs and resonators are not required [2]. Transformer-based matching networks are adopted for wide bandwidth, bias implementation and more practical EM-modeling compared to LC networks. The majority of the analog baseband is designed in current-mode to achieve the targeted 2GHz RF bandwidth with sufficient linearity. The integrated T/R switch uses two $\lambda/4$ transmission lines, a dedicated shunt switch at the LNA input for TX-mode and the common-source transistors of the PA final stage rearranged as a shunt switch in RX-mode. Both switches are on the primary side of the transformer-output matching networks so that the residual leakage inductance seen at the antenna side effectively shortens the transmission line lengths to $<\lambda/4$, reducing front-end losses. Simulated T/R switch loss is 1.8/2dB for TX/RX.

Input/output transformer matching networks of the differential LNA also act as baluns (Fig. 9.7.2). The LNA is based on a stacked two-stage neutralized common-source amplifier [2], taking advantage of excellent G_{max}/NF_{min} performance of the FinFET process at low-V_{ds} to reduce power consumption. Simulated gain, NF and IIP3 at 73GHz are 14.9dB, 5dB and -17.6dBm respectively for 11mW. The single-ended LNA output drives a single-balanced passive mixer, followed by a coarse VGA (cVGA) with 18dB gain control in 6dB steps and a transconductance amplifier with 6.7% gain step. A current-mode, 3rd-order Chebychev-II filter suppresses adjacent/alternate channels by >10/30dB. An active IQ vector modulator performs baseband phase-shifting and calibrates IQ phase/amplitude imbalance, achieving better than 1.5 degree resolution with <7mW power in <0.04mm² area. The phase-shifter is followed by a fine VGA (fVGA) with 4% gain step that can be used for array tapering. The signal currents from the 4 RXs are first combined and then brought off-chip by a voltage driver with 24dB gain control in 6dB steps. Two automated DC offset calibration loops are implemented (Fig. 9.7.1). Measured and simulated gain and noise figure (integrated over 300kHz to 1.1GHz) for a single receiver (including T/R switch loss) are shown in Fig. 9.7.2, together with the measured EVM vs RX input power for a 4Gb/s 16-QAM single-carrier modulation. Measured RX IIP3 is -18dBm.

The input IQ TX signal is first converted into current by a transconductance amplifier and then filtered to suppress DAC images (Fig. 9.7.1). The signal is then split by 4 to each transmitter, where it is phase-shifted by the IQ vector modulator and amplified by two VGAs with same range and resolution as in the RX. Two automated loops remove DC offset and suppress carrier leakage at the TX output. An active IQ double-balanced mixer upconverts the signal and drives the mm-wave PA. The 3-stage PA is implemented with neutralized differential pair common-source Class-AB amplifiers (Fig. 9.7.3), with a simulated P_{sat} and peak PAE of +10.9dBm and 16% respectively. Measured and simulated TX P_{sat} (including T/R switch loss) across E-band is shown in Fig. 9.7.3.

Analog baseband beamforming allows for low loss, low-power and low-area phase-shifters and together with direct-conversion architecture is the first step towards a fully integrated digital beamforming solution. To address mm-wave LO distribution challenges that come with this architecture, 4 PLLs have been integrated, one local to each transceiver (Fig. 9.7.1). This approach greatly improves design scalability, since each transceiver can be "stepped" 4 times to implement the array. The 2.4GHz reference signal is distributed to the 4 PLLs using CMOS buffers. Any residual phase/delay mismatch in the reference distribution (or between the PLLs) is compensated by the overall array phase calibration required for beamforming. A subsampling PLL with 24GHz quadrature complementary VCO (QVCO) is adopted for low power (Fig. 9.7.4). A QVCO is more efficient than a VCO with a polyphase filter for quadrature generation, especially if an in-phase coupling technique [3] is used to minimize FoM degradation typically introduced by quadrature coupling (<1.5dB in this design). A ping-pong architecture in the sampler maintains constant loading on the QVCO and minimizes reference spurs. Instead of using a power-hungry 24GHz divider for the frequency locking loop (FLL) required for the subsampling PLL to lock reliably, the aliased frequency error generated by a replica sampler driven by the other quadrature phase of the QVCO is used [$CK_{alias}(Q)$ in Fig. 9.7.4]. Current pulses are generated at each rising edge of $CK_{alias}(Q)$ and injected into the same loop filter controlled by the phase-locking loop. As the frequency error is reduced to zero, the loop gain of the frequency loop is also suppressed, handing-off the VCO control to the phase-locking loop. Aliased frequency error can be $\pm f_{ref}/2$ at most, so the frequency detector can be low power by operating at <1.3GHz. The QVCO must be initialized within $f_{ref}/2$ from the target frequency to avoid locking to the wrong integer ratio. This is addressed by selecting the correct coarse tuning band at power-on after a one-time calibration. An automatic band selection loop monitors QVCO control voltage, updating the coarse tuning if it drifts outside an acceptable range. An 8b $\Delta\Sigma$ DAC performs a fine sub-degree IQ phase calibration by offsetting the two tank resonances in the QVCO [4]. Each QVCO drives two IQ frequency triplers based on a doubler+mixer architecture [5]. Driving the tripler with IQ phases enhances conversion efficiency: the $3f_{LO}$ components generated by two separate mechanisms – the $2f_{LO}$ of the doubler mixing with the f_{LO} of the mixer and the DC component of the doubler mixing with the $3f_{LO}$ of the mixer – add in phase. The measured PLL locks in <6µs from cold start over 63 to 77GHz with a phase noise of -105/-109dBc/Hz @ 1/10MHz (73GHz) for 30mW power (including frequency triplers).

The 2×2 transceiver IC is implemented in 22nm FinFET technology and measures 2.1×2.4mm² (Fig. 9.7.7). Four ICs are co-packaged in a 12-layer 8.1×7.5mm² module that includes 16 patch antennas; 4 packages are tiled on a PCB to form the 64-element array (Fig. 9.7.7). A package variant with ground-signal-ground probe pads replacing the patch antennas is used for conductive measurements to characterize individual transceivers. One-to-four/four-to-one resistive splitters/combiners are used on the PCB and on each package to distribute the baseband analog IQ signals and the 2.4GHz reference to each IC. After LO leakage, amplitude and phase calibrations, TX beam patterns are measured in an anechoic chamber for the 64-element array (Fig. 9.7.5). Measured array gain is 35.4dB, 0.7dB from the theoretical value (Fig. 9.7.3), demonstrating proper array calibration, sufficient isolation between elements and thermal stability. Beamforming performance is stable over time, with 64 PLLs locked coherently. Measured P_{eirp} at P_{sat} is +44.4dBm at 73GHz with 9.47W power consumption. Fig. 9.7.5 shows the measured TX EVM vs P_{eirp} over symbol rates (16QAM) and steering angles. TX constellation (-24dB EVM) and output spectra are also shown in Fig. 9.7.5 for a 6Gb/s 16-QAM modulation and broadside beam with P_{eirp}=+33.6dBm. Fig. 9.7.6 summarizes performance and compares it with state-of-the-art.

Acknowledgments:
The Authors would like to thank Jimin Yao, Marwan Abu-mahaimed, SriChaitra Chavali, Huong Do, Tian Qiu, Ye Seul Nam, Vipul Mehta, Bob Sankman, Hector Carlo, Denise Souza and Brad Jackson for packaging and assembly support, Ada Secasiu, Hyung-jin Lee and Umesh Jalan for die fabrication and preparation, Cuong Le, Cristan Paulino and Tom Palomino for layout support, Phong Bach and Jin Feng for physical digital design, Justin Mccallister and Shuhei Yamada for testing support.

References:
[1] W. Shin et al., "A Compact 75 GHz LNA with 20 dB Gain and 4 dB Noise Figure in 22nm FinFET CMOS Technology," *IEEE RFIC*, pp. 284-287, 2018.
[2] G. Mangraviti et al., "A 4-Antenna-Path Beamforming Transceiver for 60GHz Multi-Gb/s Communication in 28nm CMOS," *ISSCC*, pp. 246-247, Feb. 2016.
[3] X. Yi et al., "A 57.9-to-68.3 GHz 24.6 mW Frequency Synthesizer with In-Phase Injection-Coupled QVCO in 65nm CMOS Technology," *IEEE JSSC*, vol. 49, no. 2, pp. 347-359, Feb. 2014.
[4] T. Siriburanon et al., "A Low-Power Low-Noise mm-Wave Subsampling PLL Using Dual-Step-Mixing ILFD and Tail-Coupling Quadrature Injection-Locked Oscillator for IEEE802.11ad", *IEEE JSSC*, vol. 51, no. 5, pp. 1246-1260, May 2016.
[5] C.-N. Kuo et al., "A K-Band CMOS Quadrature Frequency Tripler Using Sub-Harmonic Mixer," *IEEE MWCL*, vol. 19, no. 12, pp. 822-824, Dec. 2009.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

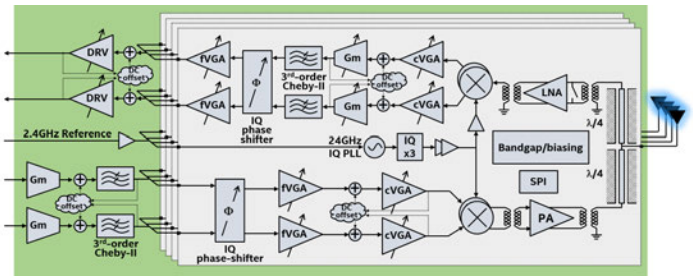


Figure 9.7.1: Integrated 2×2 transceiver IC simplified block diagram.

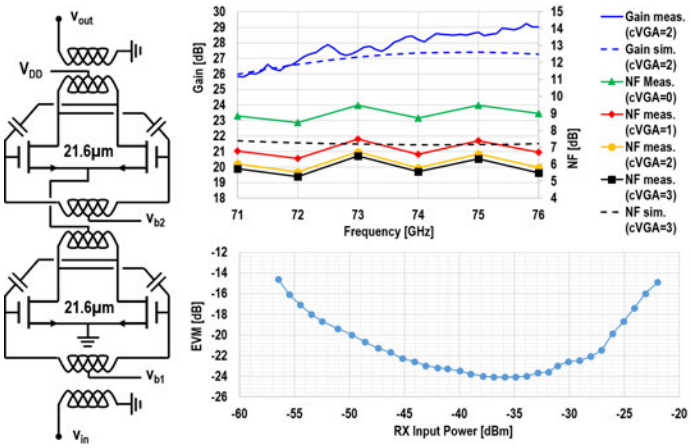


Figure 9.7.2: LNA schematic, simulated and measured receiver gain and NF (integrated over 300kHz to 1.1GHz) and measured receiver EVM vs input power for a 4Gb/s, 16-QAM single-carrier modulated signal at 73GHz (single receiver, conductive).

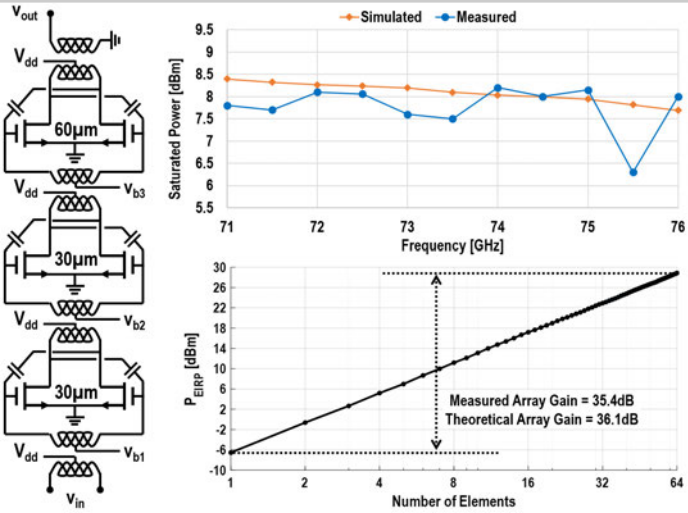


Figure 9.7.3: Simplified PA schematic (T/R switch circuitry not shown) with simulated and measured (conductive) single transmitter P_{sat} across E-band (top plot). Measured TX array gain for the 64-element array at back-off (bottom plot).

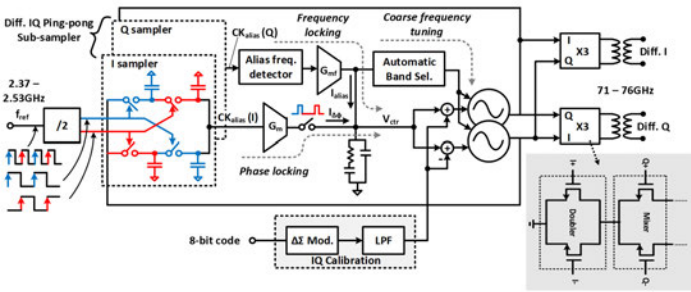


Figure 9.7.4: Simplified block diagram of the 24GHz subsampling PLL with quadrature VCO and IQ frequency triplers.

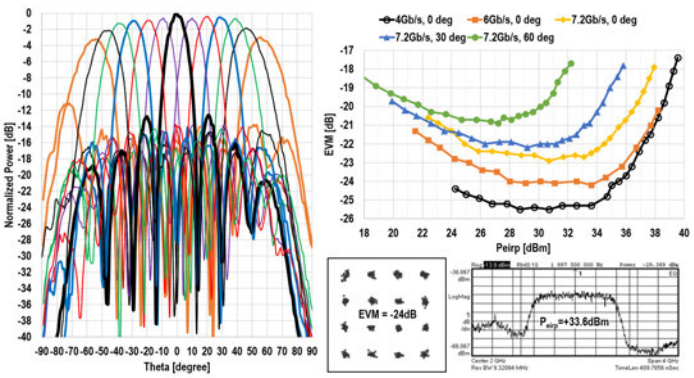


Figure 9.7.5: Measured transmitter beampatterns and EVM vs P_{eirp} for the 64-element array for different symbol rates (16QAM) and steering angles (azimuth, H-plane); measured transmitter 16-QAM constellation (-24dB EVM) and spectrum (at the reference receiver) for 6Gb/s and broadside beam with P_{eirp}=+33.6dBm. Spectral replica near DC is due to image tones on the LO path of the mm-wave downconverter used for the reference receiver.

	This Work	[2]	[Shahramian ISSCC'18]	[Sowlati ISSCC'18]	[Garcia RFIC'14]	[Dunworth ISSCC'18][a]	[Kim RFIC'17]
Tech. [nm]	22nm FinFET	28nm CMOS	180nm SiGe	40nm CMOS	130nm SiGe	28nm LP-RF CMOS	28nm LP CMOS
Center Freq. [GHz]	71-76	57-66	80-100	57-66	86-96	28	28
Elements per IC	4	4	16TX, 8RX	12	16TX,32RX	24	8
Measured Array Size	64 (single pol.)	4 (single pol.)	256 (single pol.)[b]	144 (single pol.)	-	4 (dual pol.)	8 (dual pol)
Architecture	Baseband Beamforming Direct-Conversion	Baseband Beamforming Direct-conversion	RF Beamforming Direct-Conversion	RF Beamforming 10.65GHz IF	RF Beamforming 10GHz IF & Direct-conversion	RF Beamforming 6.5GHz IF	RF Beamforming Direct-Conversion
LO Input/Ref. [GHz]	2.37-2.53	0.027	1.7	0.22	1-4	N/A (uses XO)	3.1
Phase Shifter Res. [bit]	8	6	9	6	5	3	3
RF BW [GHz]	10 TX[c], 5 RX	8	-	9	6 TX[c], 17 RX	7.5 TX, 5.5 RX	2.2[e]
Baseband IQ BW [GHz]	2	1.76	-	-	-	-	0.5
Single TX Gain [dB]	26	30	-	45	23 (from IF)	34-44	48 (2 patches)
Single RX Gain [dB]	36.7 (max)	62	80	23	30 (RF front-end)	32-34 (4 channels)	44 (2 patches)
Single TX P_{sat} [dBm]	8 (w/ TR sw.)	9	8	5 (w/ TR sw.)	2 (w/ TR sw.)	14 (w/ TR sw.)	10.5 (w/ TR sw. and pack.)
Single RX NF [dB]	6 (w/ TR sw. and pack.)	4.8-6.2 (w/ pack.)	8	7 (w/ TR sw.)	8.2 (w/ TR sw.)	4.4-4.7 (w/ TR sw.)	6.7 (w/ TR sw. and pack.)
Array TX P_{eirp} [dBm]	+44.4 (P_{sat})	+22.5 (cl.-AB)[f] +24 (cl.-A)[f]	+59 (P_{sat})[f]	+51 (max)	-	+35 (peak)	+31.5 (P_{sat})
Array EVM [dB]	TX -25.5 @ P_{eirp}=+34dBm (4Gb/s, 16QAM) -20 @ P_{eirp}=+36.5dBm (7.2Gb/s, 16QAM)	TX -21 @ P_{eirp}=+24dBm (7Gb/s, 16QAM) TX-to-RX -20 @ 1m (10.5Gb/s, 64QAM)	TX -26 @ P_{eirp}=+52dBm (0.5Gb/s, 32QAM)[f] TX-to-RX -21.9 @ P_{eirp}=+52dBm (10Gb/s, 16QAM)[f]	-	-	-	RX -33.3 (floor) (10MHz LTE, 64QAM) TX -33.1 @ P_{eirp}=0dBm (10MHz LTE, 64QAM)
TX Power/el [mW]	148 (@ P_{sat})	167 (cl.-A), 106 (cl.-AB)	275	58	169/181 (IF/BB output)	90 (@ P_{sat}=6dBm)	85 (@ P_{sat}=3dBm)
RX Power/el. [mW]	168[d]	104	225	46	106	42	50
Die Size [mm²]	2.1 x 2.4	2.63 x 3	8 x 4.4	4.68 x 4.68	6.6 x 6.7	4.65 x 5.97	2.8 x 2.6
Die Size/el. [mm²]	1.26	1.97	1.47	1.83	0.92	1.16	0.91

[a] Measurement limited by PLL locking range [b] From visuals [c] Estimated from figure [d] All measurements use external 20-23GHz LO [e] Including 50ohm output driver (1 per IC)

Figure 9.7.6: Performance summary and comparison with prior art for phased-array solutions above 60GHz or with similar architecture in CMOS.

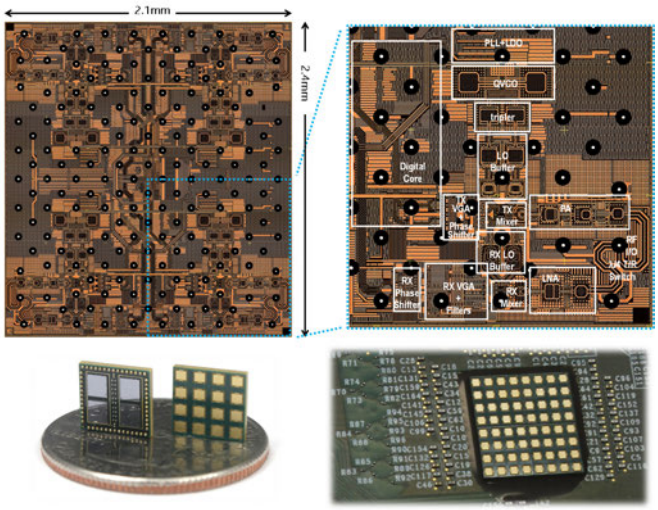


Figure 9.7.7: 2×2 transceiver die micrograph, with details of the single transceiver (top). 16-element module with 4 ICs (front and back) and 64-element system with 4 modules mounted on PCB (bottom).

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

9.8 A 28GHz 20.3%-Transmitter-Efficiency 1.5°-Phase-Error Beamforming Front-End IC with Embedded Switches and Dual-Vector Variable-Gain Phase Shifters

Jinseok Park*[1], Seungchan Lee*[1], Dongho Lee[2], Songcheol Hong[1]

[1]KAIST, Daejeon, Korea
[2]Hanbat National University, Daejeon, Korea
*Equally-Credited Authors (ECAs)

Millimeter-wave beamforming front-end ICs have been studied intensively as the service of 5G wireless communication is scheduled to begin in the near future [1-4]. The ICs include circuit elements such as PAs, LNAs, phase shifters, variable gain blocks, and switches to support antenna arrays for RF/hybrid beamforming. Due to the large number of antennas required for beamforming, the beamforming IC should include as many circuit elements as possible in a chip. The IC also needs high phase- and gain-control resolutions not only for controlling the beams precisely but also for error corrections and calibrations [1]. However, higher-bit controls of the phase and gain as well as high transmitting power increase the chip size in conventional structures, posing a trade-off between them. The front-end IC structure proposed here increases both transmitting power and gain/phase resolutions without increasing either the chip size or the power consumption.

Figure 9.8.1 shows the 28GHz beamforming front-end IC structure with embedded switches, a dual-vector variable-gain phase shifter (VGPS), and distributed gain control functions. The antenna switch is embedded in the PA output matching circuit and the LNA input matching circuit, reducing the size and the insertion loss of the switch compared to a separately designed switch. The VGPS is implemented using the proposed dual-vector phase shifter to enable control of the phase and gain within a single block, allowing the IC to be small and to consume little power regardless of the resolution of gain and phase. The PA has triple power modes to be used for gain control and to further reduce the amount of power consumption when the gain is reduced. In the mid-power mode (MPM), half of the drive amplifier cells are turned off. In the low-power mode (LPM), the supply voltage of the power stage changes from 2.2V to 1.2V additionally. To improve the maximum linear power, an integrated linearizer is used. The linearizer is connected between common-gate (CG) transistors' sources as in Fig. 9.8.1. This acts as a variable resistor according to the input power. Since resistance of the linearizer increases with the input power, the linearizer has gain-expansion and phase-lag characteristics, which compensate the PA with gain-compression and phase-lead characteristics. As a result, the AM-AM and AM-PM characteristics are improved. This principle is also applicable for all PA modes to maintain high linear power. The LNA also has a gain control function with low phase variation. In the second stage of the LNA, a parallel PMOS is connected to the differential output nodes as in Fig. 9.8.1, and gain control voltage is applied to the gates of both the CG NMOS and the PMOS. The phase of the LNA is automatically compensated during the gain control process via the complementary impedance changes of the NMOS and PMOS. With those blocks, the gain control functions are distributed into VGPS and multimode PA for TX and VGPS and variable gain LNA (VGLNA) for RX, reducing the chip size and power consumption.

Figure 9.8.2 shows the proposed dual-vector VGPS, which consists of a quadrature allpass filter type IQ generator based on [5], main and auxiliary vector generators, and a transmission-line-transformer-based summing circuit at the output. The gain and phase are controlled by summing the two vectors as determined by the four tail currents flowing along the current mirrors of I_{main}, Q_{main} ,I_{aux}, and Q_{aux} paths. The total current must remain constant to ensure matching between the summing circuit and the two vector generators and to have a constant output impedance regardless of the gain and phase states, which is necessary for linear vector summations. The current-mapping table in Fig. 9.8.2 shows how to set the four tail currents to make the total current constant for all phase and gain steps. The magnitudes of I_M and θ respectively determine the gain and phase states, where $-1 \le \theta \le 1$ and $0.5 \times I_{total} \le I_M \le I_{total}$. The DACs for phase and gain control are designed to generate currents according to the mapping table. Here, eight and seven bits are assigned for phase and gain control, respectively. The figure shows that the proposed VGPS can generate constellations in circles of variable radii.

Figure 9.8.3 shows TX and RX continuous-wave (CW) performances, where both input and output embedded switches are included in the measurements. TX and RX gain are 18.6dB and 14.8dB with RF 3dB BW of 5.2GHz and 3.8GHz, respectively. The NF of RX at 28GHz is 4.58dB, which is increased only by 1dB compared to simulated NF of LNA without switch and PA. NF variations are 4.58 to 5.8dB with 14dB gain control. The OP_{1dB} of TX is 13.95 to 14.6dBm from 26 to 30GHz, which is reduced only by approximately 0.42 to 1.15dB compared to measured OP_{1dB} of the PA without switch and LNA. TX and PA efficiency at saturated output power are 20.3% and 31.6%, respectively. The embedded antenna switch has TX-to-antenna isolation of 25dB and antenna-to-RX isolation of 32dB at 28GHz in the RX and TX modes, respectively.

Figure 9.8.4 shows measured phase- and gain-control performances. The dual-vector VGPS has 5.625° phase resolution and 0.5dB gain-control step. The static constellation diagram measured at 28GHz demonstrates that the PA power mode can be used as a gain-control scheme. The gain control by MSBs is implemented with the PA power mode, and that by LSBs is done with the dual-vector VGPS. This improves gain-control dynamic range, and reduces power consumption by 43% from 78mW to 43mW when 8dB gain is reduced. The power consumption here includes that of all blocks including bias circuits. The PA power-mode change causes additional phase errors, which are compensated by the high-resolution dual-vector VGPS. The measured rms gain and phase errors at the maximum gain are less than 0.25dB and 1.5° across 24 to 30GHz, respectively. The total rms phase error from simultaneous gain and phase changes is less than 2.8° across 26 to 30GHz. The VGLNA is used for extra gain control of the RX, and peak phase variation of the RX is 1.4° for 16dB gain control. IMD3s of the TX with three power modes are measured with an 80MHz tone spacing.

Figure 9.8.5 shows EVMs and ACLRs measured with 80MHz 64-QAM and 256-QAM (9.45 and 9.7 dB PAPR) WLAN 802.11.ac signals. The front-end IC achieves an average output power of 8.45/7.1dBm, a drain efficiency of 17.7/14.5%, and TX total power consumption of 82.8mW (77mW) with an EVM of -25dB (-30dB) at 28.5 GHz for 64-QAM (256-QAM) signals. In the low-power mode, the IC has an average power of 2.8dBm (0.6dBm) and drain efficiency of 11.2% (7.3%). Pout/PA efficiency remains at more than 7.8dBm/15.8% (6.3dBm/12.6%) across 26 to 30GHz for 64-QAM (256-QAM) OFDM signals.

Figure 9.8.6 compares this work with previously reported state-of-the-art 28GHz front-end ICs. The power consumption and efficiency are normalized to those of a single channel, and the chip size is compared while excluding the pads. This front-end IC has the highest OP_{1dB}, TX efficiency, and linear output power under a 64-QAM signal. It also has higher linear output power satisfying -30dB EVM with a 256-QAM signal. With the dual-vector VGPS and distributed gain-control functions, this work has achieved higher phase control resolution with the lowest rms phase and gain error, as well as a wide gain-control range with the smallest phase error. By using the PA mode as a gain-control scheme, the chip's power consumption is further reduced by 19% and 43% when the gain is reduced by 4dB and 8dB, respectively. Considering the high output power and phase resolution, it has lower power consumption and a smaller chip size. A die micrograph is shown in Figure 9.8.7, and the core chip occupies an area of $0.67mm^2$.

Acknowledgements:
The authors would like to thank Seunghoon Kang, Euntaek Sung, and Seunghun Wang for their help. This work was supported by IITP (2017-0-00409, Development on mm-Wave Beamforming IC for 5G Communication) and an NRF(2017R1A2A1A05001361) grant funded by the Korea government (MSIT).

References:
[1] B. Sadhu et al., "A 28-GHz 32-Element TRX Phased-Array IC with Concurrent Dual-Polarized Operation and Orthogonal Phase and Gain Control for 5G Communications," *IEEE JSSC*, vol. 52, no. 12, pp. 3373-3391, Dec. 2017.
[2] H. Kim et al., "A 28-GHz CMOS Direct Conversion Transceiver with Packaged 2 x 4 Antenna Array for 5G Cellular System," *IEEE JSSC*, vol. 53, no. 5, pp. 1245-1259, May 2018.
[3] J. D. Dunworth et al., "A 28GHz Bulk-CMOS Dual-Polarization Phased-Array Transceiver with 24 Channels for 5G User and Basestation Equipment," *ISSCC*, pp. 70-72, Feb. 2018.
[4] K. Kibaroglu et al., "A Low-Cost Scalable 32-Element 28-GHz Phased Array Transceiver for 5G Communication Links Based on a 2 × 2 Beamformer Flip-Chip Unit Cell," *IEEE JSSC*, vol. 53, no. 5, pp. 1260-1274, May 2018.
[5] K-J. Koh and G.M. Rebeiz, "0.13-μm CMOS Phase Shifters for X-, Ku-, and K-band Phased Arrays", *IEEE JSSC*, vol. 42, no. 11, pp. 2535-2546, Nov. 2007.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

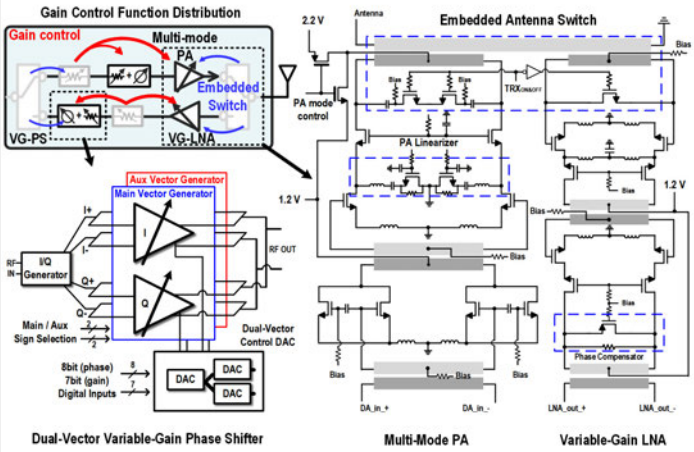


Figure 9.8.1: IC architecture with a block diagram of dual-vector VGPS and schematics of embedded antenna switch, multimode linear PA, and VGLNA.

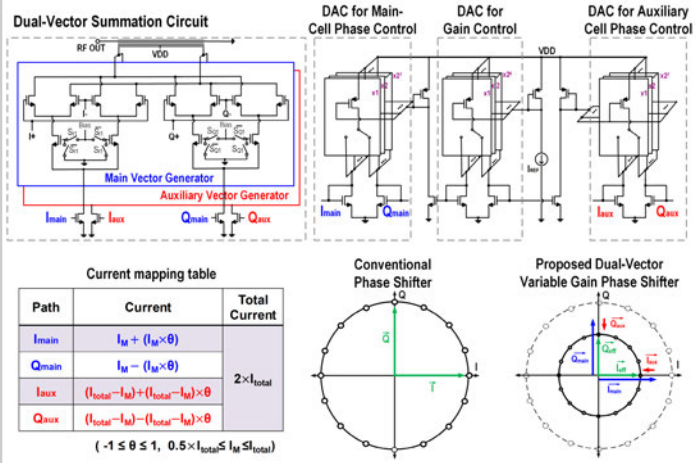


Figure 9.8.2: The proposed dual-vector VGPS; schematics of dual-vector VGPS circuit and its control DACs, Current mapping table, and Vector constellation.

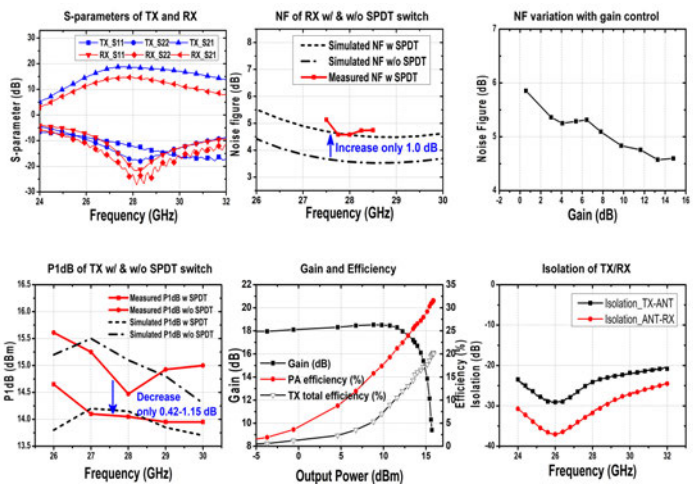


Figure 9.8.3: Measured RX and TX CW performance; S parameters of TX/RX, NF of RX, NF variation with gain control, P_{1dB} , TX efficiency, and Isolations of TX/RX.

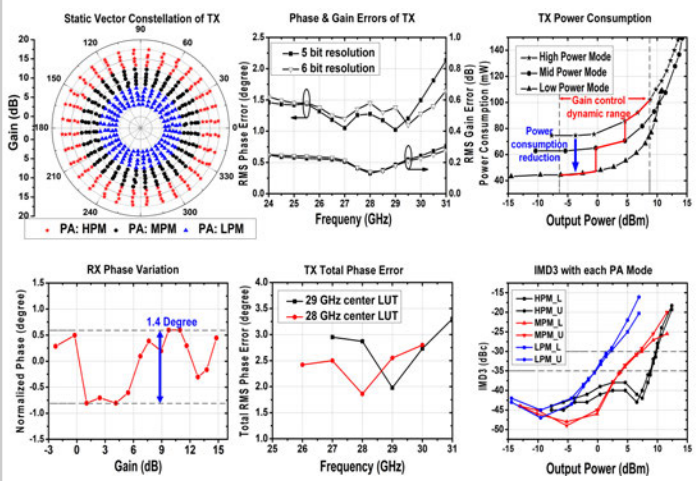


Figure 9.8.4: Measured phase/gain control performance; TX vector constellation, phase/gain errors, power consumption, RX phase variation, TX phase error, and IMD3.

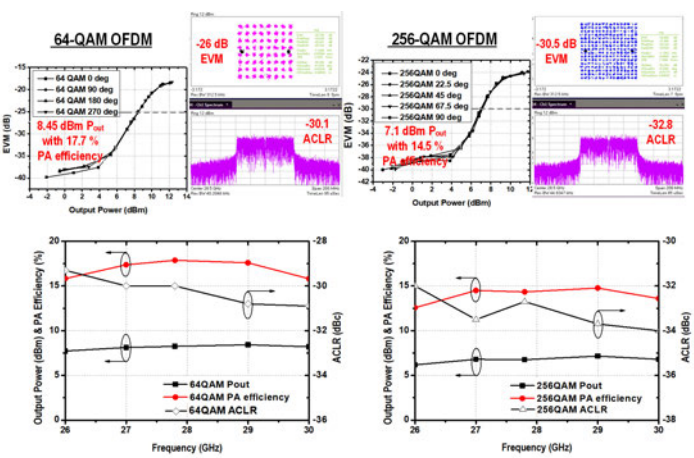


Figure 9.8.5: Measured EVMs and ACLRs with 64-QAM and 256-QAM OFDM signals; Linear power, PA efficiency, and ACLR of TX (64QAM and 256QAM).

Parameter	This Work		[1] JSSC17	[2] JSSC18	[3] ISSCC18	[4] JSSC18
Process	65 nm Bulk CMOS		130 nm SiGe BiCMOS	28 nm LP CMOS	28 nm LP-RF CMOS	180 nm SiGe BiCMOS
Tx OP1dB (dBm)	13.9		13.5	9.5	12	10.5
RF 3 dB Bandwidth (GHz)	5.2 (TX) / 3.8 (RX)		2 (TX) / 1.5 (RX)	2.2	7.5 (TX) / 5.5 (RX)	7 (TX) / 6 (RX)
Phase / Gain Control Resolution (bit)	8 (2bit for cal.) / 7		5 / 5	3 / -	3 / 3	6 / 4
Peak Phase Variation by Gain Control / Gain Control Range	**1.4° / 16 dB**		8 ° / 8 dB	- / 12 dB	- / 9 dB	6 ° / 14 dB
RMS Phase Error due to Phase Control (°)	**<1.5**		1.5	7	-	5
RMS Gain Error (dB)	**<0.25**		1.4 **	-	-	0.9
Peak Tx Efficiency (%)	**20.3**		13.8	9.5	20	5.6 †
Tx 64-QAM Pout (dBm)	**8.45**	**2.8**	-	2	6	2.5
64-QAM Power Consumption (mW)	**82.8**	**53**	-	66	90	-
256-QAM Pout (dBm)	**7.1**	**0.6**	-	-	-	-
256-QAM Power Consumption (mW)	**77**	**49**	-	-	-	-
Noise Figure (dB)	4.58		6	6.7	3.8 *	4.6
Rx Power per Channel (mW)	44		103.1	40	42	130
1 Channel Core size (mm²)	0.67		3.13 *	0.52 *	0.6 *	2.09 *

* Graphically estimated, ** Peak gain error, † Efficiency at P1dB.

Figure 9.8.6: Comparison table with published state-of-the-art 28GHz front-end ICs.

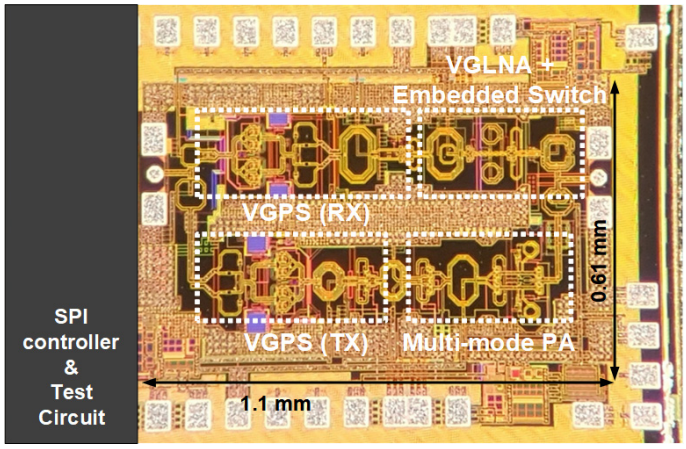


Figure 9.8.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 10 Overview: *Sensor Interfaces*

ANALOG SUBCOMMITTEE

Session Chair: ***Tim Piessens***
ICsense, Leuven, Belgium

Associate Chair: ***Wen-Chieh Wang***
MediaTek, Hsinchu City, Taiwan

Subcommittee Chair: Kofi Makinwa, Delft University of Technology, Delft, The Netherlands

This session highlights advances in state-of-the-art sensor interfaces. In the first paper, a highly-integrated energy measurement system is shown that uses self-calibration to achieve high robustness. A second paper describes a high-resolution accelerometer that achieves a noise floor of 22ng/√Hz. The system consists of a MEMS accelerometer, an HV driver ASIC and a low-voltage signal processing ASIC. Next, two integrated resistor-based temperature sensors are presented, one that operates over the largest reported temperature range (-40ºC to 180ºC) and one that achieves a state-of-the-art energy efficiency of $20fJ/K^2$.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

8:30 AM

10.1 An Energy Measurement Front-End with Integrated In-Situ Background Full System Accuracy Monitoring Including the Current and Voltage Sensors

S. Danesh, Analog Devices, Edinburgh, United Kingdom

In Paper 10.1, Analog Devices presents an energy measurement system that measures both line voltages and currents. The system continually monitors its accuracy by injecting reference stimuli. It overcomes the dynamic range challenge associated with detecting these stimuli in the presence of unknown loads, achieving better than 1000ppm inaccuracy for both voltage and current measurements.

9:00 AM

10.2 A 22ng/√Hz 17mW MEMS Accelerometer with Digital Noise-Reduction Techniques

Y. Furubayashi, Hitachi, Tokyo, Japan

In Paper 10.2, Hitachi presents a 17mW MEMS accelerometer. By using a servo-signal-leakage cancellation scheme, and a high-linearity 1b bandpass DSM DAC, it achieves a very low noise floor of 22ng/√Hz.

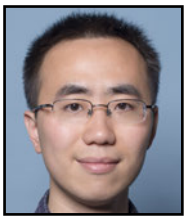
9:30 AM

10.3 A 0.12mm² Wien-Bridge Temperature Sensor with 0.1°C (3σ) Inaccuracy from -40°C to 180°C

Ç. Gürleyük, Delft University of Technology, Delft, The Netherlands

In Paper 10.3, Delft University of Technology describes a Wien bridge temperature sensor that is used to explore the temperature and stress dependence of various on-chip resistors. It achieves an inaccuracy of 0.1°C from -40°C to 180°C.

9:45 AM



10.4 A Wheatstone-Bridge Temperature Sensor with a Resolution FoM of 20fJ·K²

S. Pan, Delft University of Technology, Delft, The Netherlands

In Paper 10.4, Delft University of Technology presents an FIR-DAC resistor-based temperature sensor. It achieves a resolution of 160μK in a 10ms conversion time, and a state-of-the-art energy efficiency of 20fJ·K².

10.1 An Energy Measurement Front-End with Integrated In-Situ Background Full System Accuracy Monitoring Including the Current and Voltage Sensors

Seyed Danesh[1], William Holland[1], Joe Spalding[1], Michael Guidry[2], J.E.D. Hurwitz[1]

[1]Analog Devices, Edinburgh, United Kingdom; [2]Analog Devices, Wilmington, MA

Millions of utility electricity meters are deployed globally to determine billable energy consumption. These devices are factory calibrated, and then perform measurement without interruption over their lifetime. Once in the field their accuracy is unknown due to component aging and high voltage/current events[1]. Meters are periodically replaced as a preventive measure, leading to the unnecessary replacement of perfectly good meters while some inaccurate meters may remain in field. This paper presents an energy-measurement front-end that enables continuous background Condition Base Monitoring (CbM) over the lifetime of a meter. Crucially it monitors the performance of its voltage and current sensors, which are typically the largest sources of inaccuracy and drift. During normal operation, small test signals ("stimuli"), with adaptable but well-defined characteristics are injected into the sensors. Their amplitudes at the front-end's output are then extracted to determine the gain accuracy of the full signal chain. In order to do this, the system must meet 3 main challenges. First, it must be able to generate and inject highly stable and accurate stimuli signals into the sensors. Second, it must have a wide dynamic range in order to accurately extract the stimuli in the presence of unknown and significantly larger load signals. Third, it must incorporate all the digital signal processing involved with stimuli signal selection, extraction and removal. This paper describes the circuit techniques used to address the first two challenges.

Figure 10.1.1 shows a system and circuit view of the implemented energy measurement front-end. As is typical in single-phase meters, line voltage is measured via a potential divider, phase current via a shunt resistor, and neutral current via a current transformer (CT). The phase current and voltage are used to determine energy usage, while the neutral current is used to detect faults or potential electricity theft. For all 3 sensors, the test stimuli, a current (I_s) or a voltage (V_s), follow the same signal path as the corresponding load signals, I_{load} or V_{load}, being measured. The signal measured by the front-end is then the sum of the load signal and the reference stimuli. The front-end's digital backend detects and then removes the stimuli prior to the determination of load voltage and line current which are then used in the metrology functions. The digital backend also defines the stimuli, which are two-level signals with well-controlled rise and fall times, whose frequency and phase are dynamically adjusted on the basis of the type of load signals present. However, most of their frequency content is concentrated around 50/60Hz, i.e. close to the load signals' dominant frequency. Compared to typical load currents (100A RMS) and voltages (230V RMS), the corresponding stimuli (16mA peak and 150mV peak) are at the system noise floor. These must then be measured to ~0.1% accuracy, setting the required dynamic range to 138dB (I) and 127dB (V).

The signal chain for the current stimuli (Fig. 10.1.2) starts with a stable reference voltage V_{ref} derived from a chopped (500kHz) substrate PNP bandgap reference with curvature correction and filtering [2], with a secondary output V_{ref-n}. This is converted to a stable 15uA reference current by a switched-capacitor V-to-I circuit whose output is then modulated and amplified to ~30mA by a current driver. V_{ref-n} has a small negative temperature coefficient (TC) to cancel the positive TC of the switched capacitors. Controlled by a 2MHz clock (derived from an external crystal oscillator), two 7pF metal-metal capacitors, C_1 and C_2, are alternately charged to V_{ref-n} by the currents I_1 and I_2. At the end of each of their respective charging phases, each capacitor is connected to an integrator, which accumulates the difference between V_{ref-n} and the final capacitor voltage before being reset prior to the start of the next charging cycle. Via a degenerated NMOS, the integrator generates a current that tops up (by ~20%) the current of a resistor-defined current DAC to create the currents needed in the loop (I_1 and I_2). The same currents when not charging the capacitors are switched to the output I_{ref} so in contrast to [3], this scheme avoids any output current-mirror errors, as both the output (I_{ref}) and the charging currents (I_1, I_2) are obtained from the same current sources. Due to mismatch in C_1 and C_2, and differences in I_1 and I_2 a small ripple (~0.1%) in I_{ref} can be observed (at 2MHz). However, it is outside the system's band of interest, and the integrator ensures that the output current (the average of I_1 and I_2) is well-defined.

Figure 10.1.3 shows the driver that generates the current stimuli. Its output compliance is set by the signal swing on the sensors, which can be either positive or negative w.r.t. I_{load}. A split-supply scheme is used, in which the reference and part of the driver are on a 1.8V supply, while the output stage is supplied by a divide by 2 or 3 charge-pump to efficiently deliver the large currents required. The driver is a current multiplier based on a chopped opamp and two matched resistors. To realize the 2000:1 ratio necessary to generate 30mA from a 15uA input, the resistors are realized with a large programmable array of 3kΩ unit elements, which are carefully laid out to match routing, parasitics and aging characteristics. The programmability allows for trade-off of magnitude of stimuli current and speed of operation. The current multiplier's input is a square-wave modulated version of I_{ref} to which a pedestal current I_{ped} (~I_{ref}/16) is added to ensure that the driver is always in its linear operating region. A linear-slew-rate control circuit sets the rise and fall times of the resulting current stimuli.

For phase current, the small shunt size (typically 140μΩ) presents a noise floor challenge, on top of the dynamic-range challenge. A low-noise, high-input-impedance, chopped (96kHz), current-feedback instrumentation amplifier (Fig. 10.1.4) was used to boost the shunt voltage and drive a delta-sigma ADC. Its input stage consists of two amplifiers (A1 and A2), which use a single PMOS to combine their input (gate) and feedback (source) signals, allowing them to handle inputs ~150mV below GND from a single supply. Compared to a switched-capacitor approach, e.g. [4], the LNA confers better immunity to RF interference. The input devices $MP_{1/2}$ are in their own well, biased by $MP_{3/4}$ at a V_{gs} above their source voltages to increase their V_T. This allows both $MP_{1/2}$ and cascodes $MP_{5/6}$ to be driven by the same gate voltages, while still allowing their drain current to be folded to a push-pull output stage (B). The bodies of MP_{1-6} track the input, and the feedback network maintains constant current through the input stage, helping THD and situating the amplifier output at one V_t above the input without a CMFB circuit.

To monitor the accuracy of the voltage channel, a small voltage stimulus is applied at the bottom of a high-impedance voltage divider (1000:1) via a chopped slew-rate-controlled Class-A amplifier (Fig. 10.1.5). This ensures that the stimulus sees the same transfer function as the mains voltage. To define the rising edge of the stimuli, a pull-up current (I_P) is used to charge a capacitor via S_{RP} until comparator C_P fires and a stable reference voltage V_{refp} is switched in through S_{HP}. Similarly, the slew-rate of the falling edge is defined by the current (I_n) and comparator C_N until V_{refn} is switched in. The amplitude of the stimuli is thus directly linked to the stable output of the bandgap reference.

The complete design including the measurement front-end, stimuli drivers, and signal processing was fabricated in 0.18μm CMOS, with a die area of 11.48mm^2. The IC is capable of measuring Active and Reactive Energy over a 3000:1 dynamic range to 0.1% accuracy, and over a 10,000:1 dynamic range to 0.25% accuracy, while measuring RMS current over a 1000:1 dynamic range to 0.1% accuracy (tested to 100A). All metering performance was achieved while the accuracy of the current and voltage sensors was continuously monitored in the background. To characterise its performance, the front-end was built into 90 electricity meters. Its monitoring accuracy was then tested while the meters were subject to a range of typical application load current and voltage levels. As shown in Fig. 10.1.6, the front-end can measure the accuracy of each sensing channel to better than 0.1% even in the presence of load signals. The testing was done at room temperature, with different load signal levels. However, the reference stimuli's TC will limit the system's tracking accuracy. Figure 10.1.6 shows the temp-co of the current stimuli paths (Bandgap + V-to-I + stimuli driver), the V-to-I contribution being <10ppm/°C. The TC of the complete stimuli path is dominated by the TC of the single-temperature-trimmed bandgap reference, and the TC of the voltage channel stimuli was found to match that of the current channel closely. The LNA's input referred noise was measured as 7.8nV/√Hz at 50Hz and TC was 4.4ppm/°C mean, +/-5.4ppm/°C sigma from measurement of 380 multi-lot parts while consuming 1.2mA from a 2.5V supply. This compares favourably to the performance of state-of-the-art instrumentation amplifiers [4,5]. Typical electricity meters are Class 1, requiring energy measurement accuracy greater than 1%. Over temperature, the proposed energy measurement front-end allows energy measurement accuracy to be monitored with better than 0.5% accuracy.

References:

[1] S. Zielger et al., "Current Sensing Techniques: A Review", *IEEE Sensors Journal*, vol. 9, no. 4, pp. 354-376, May 2009.

[2] G. Ge et al., "A Single-Trim CMOS Bandgap Reference With a 3σ Inaccuracy of ±0.15% From –40°C to 125°C", *IEEE JSSC*, vol. 46, no 11, pp. 2693-2701, 2010.

[3] G. Torelli et al., "Tracking Switched-Capacitor CMOS Current Reference", *IEE Proceedings – Circuits, Devices and Systems*, vol. 145, no 1, pp. 44-47, 1998.

[4] H. Wang et al., "5.7 A 19nV/√Hz-Noise 2μV-Offset 75μA Low-Drift Capacitive-Gain Amplifier with Switched-Capacitor ADC Driving Capability", *ISSCC,* pp. 98-99, Feb. 2017.

[5] F. Sebastiano et al., "A 0.07mm^2 2-Channel Instrumentation Amplifier with 0.1% Gain Matching in 0.16μm CMOS", *ISSCC*, pp. 294-295, Feb. 2014.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

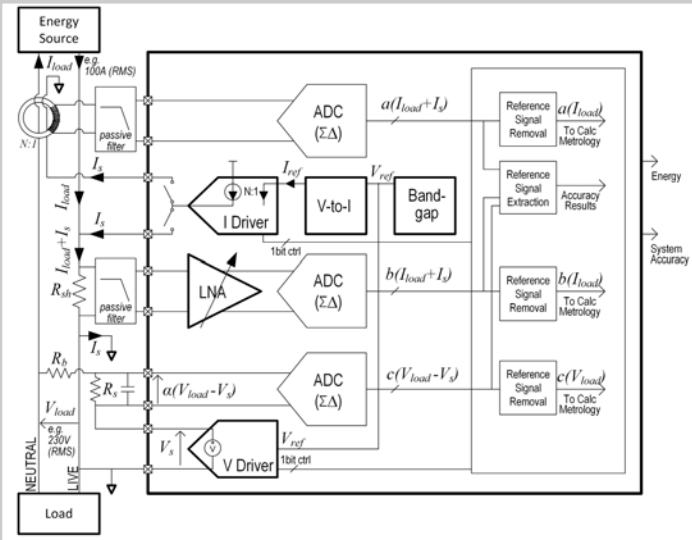


Figure 10.1.1: Energy measurement front-end with integrated accuracy monitoring of sensors and system.

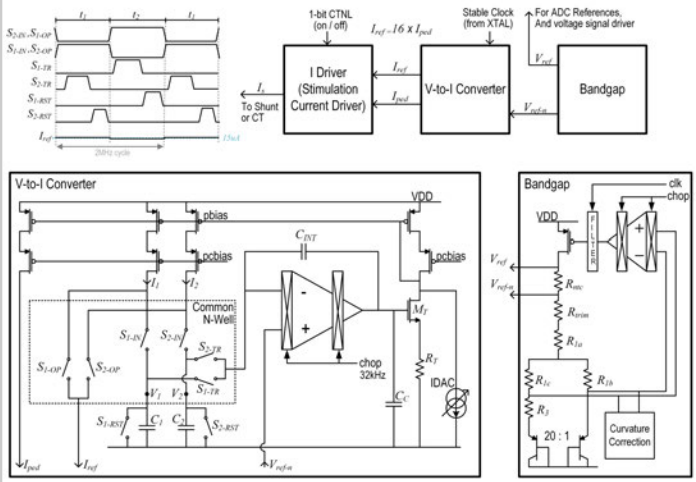


Figure 10.1.2: Current channel stimuli reference signal generator.

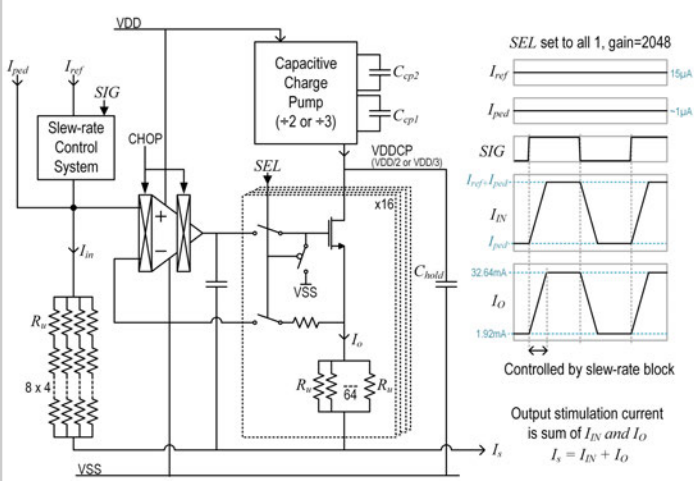


Figure 10.1.3: Current channel stimuli signal driver circuit.

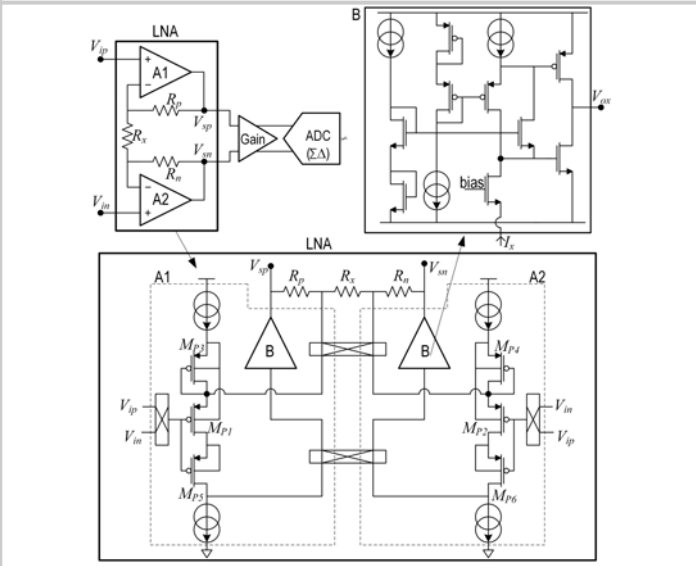


Figure 10.1.4: Low-Noise Amplifier for current measurement path.

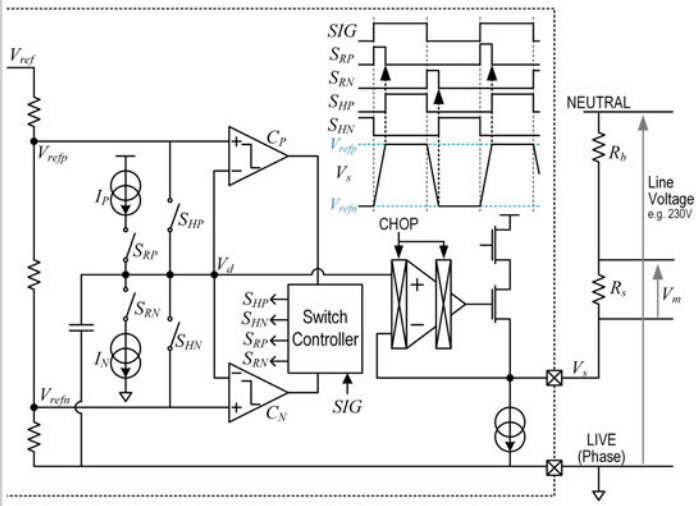


Figure 10.1.5: Voltage channel stimuli reference signal driver into potential divider.

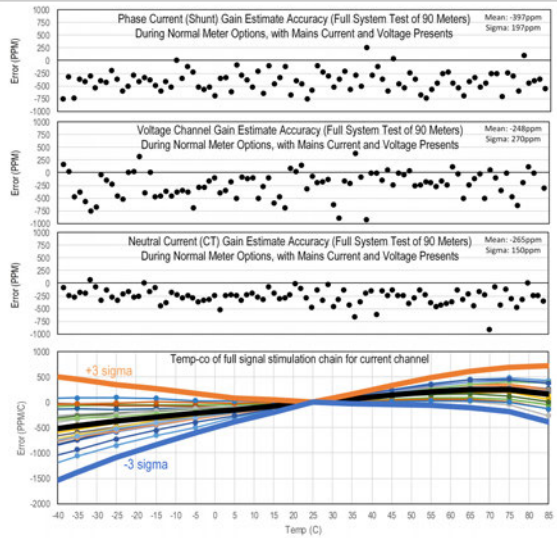


Figure 10.1.6: Characterisation results of silicon temp-co and full energy accuracy measurement system.

10

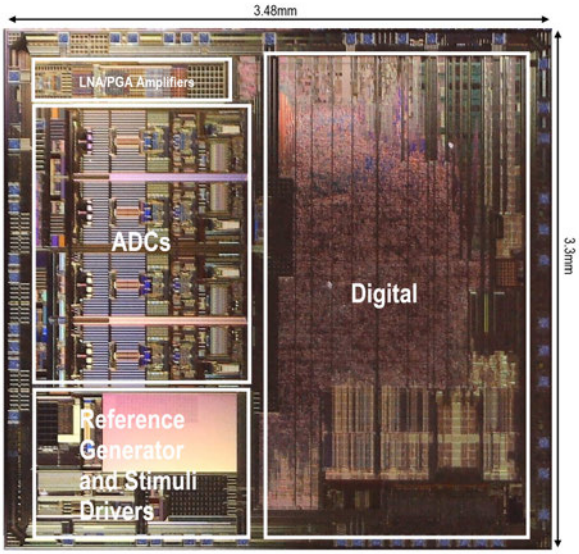


Figure 10.1.7: Chip micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

10.2 A 22ng/√Hz 17mW MEMS Accelerometer with Digital Noise-Reduction Techniques

Yuki Furubayashi, Takashi Oshima, Taizo Yamawaki, Keiki Watanabe, Keijiro Mori, Naoki Mori, Akira Matsumoto, Hideto Kazama, Yudai Kamada, Atsushi Isobe, Tomonori Sekiguchi

Hitachi, Tokyo, Japan

In emerging applications such as next-generation oil and gas exploration and infrastructure monitoring, a large number of highly sensitive accelerometers need to be deployed. The fully integrated MEMS accelerometer in this work achieves both the extremely low noise level (22ng/√Hz) required by these applications and sufficiently low power (17mW) enabling battery-based operations while meeting the required input range and bandwidth. Low noise is conventionally achieved either by the use of geophones, which have limited input bandwidth, or by the use of MEMS accelerometers, which dissipate much more power [1]. The highlights of this MEMS accelerometer are its MEMS structure with isolated detection and servo capacitors, noise reduction in the digital domain based on awareness of unfamiliar noise mechanisms and its low-noise high-voltage (HV) DACs.

As shown in Fig. 10.2.1, the accelerometer consists of a MEMS chip and a detection IC as well as an HV IC for driving a heavy MEMS device. The MEMS chip includes two pairs of capacitors for differential detection, a pair of capacitors for servo control, and two pairs of capacitors for compensation of the gravitational acceleration [2]. One electrode of each capacitor (bonded with the insulation layers) moves together in accordance with the input acceleration, thereby forming a mechanical mass. A change in capacitance (ΔC) of the detection capacitors occurs due to the imbalance between input acceleration applied to the mass and electrostatic force generated on the mass by the servo capacitors. The detection IC detects ΔC and generates a servo signal quantized to 1b to facilitate the design of a following HV servo DAC. The driver IC includes the 1b HV servo DAC, which converts the 1b servo signal to differential binary HV signals for the servo capacitors. With this feedback loop, the 1b servo signal tracks the input acceleration and thus serves as the output signal. To reduce the power, the detection IC operates at 1.2V generated by an LDO from a 1.4V supply. The gravitational acceleration in an arbitrary direction was initially compensated with the electrostatic force generated by applying the output voltage of the 10V 9b DAC to the gravity-compensation capacitors, which enabled a small transient response of the feedback loop and thus the low-voltage operation.

Every kind of noise had to be reduced to far below conventional levels. The Brownian noise generated by the collision of air molecules with the mass was reduced by small perforations formed over the entire mass. The 1/f noise of the analog chain in the detection IC was shifted out of band by modulating ΔC with a 500kHz pulse V_{DT}, generated and applied to the detection capacitors by the driver IC, and by demodulating it before the PID controller. The thermal noise of the front C/V amplifier was suppressed by not reducing the op-amp feedback factor (by using the insulation layers to isolate the servo and gravity-compensation capacitors from the op-amp input). The in-band 1b quantization noise was suppressed by three integrations of the PID controller with four zeros for stability. The voltage follower in the driver IC outputs a 7V reference voltage with a 12V supply, and either this reference or ground is connected to the pair of servo capacitors complimentarily in accordance with the 1b servo signal to implement the 1b HV servo DAC. HV MOS transistors with L of 8μm and W of several tens of millimeters were adopted for the input differential pair of the voltage-follower op-amp to reduce the reference 1/f noise.

In the quest for extremely low noise, two other kinds of noise were found. One is servo-signal leakage, which is inherently blocked by the above-described MEMS structure but is still present due to mismatch of the two parasitic capacitances (C_{PP} and C_{PN}) located between the input node of the servo signal and the input nodes of the op-amps. Even ΔC_P (=C_{PP}-C_{PN}) of 0.1fF affects the noise level below 100ng/√Hz. As shown in Fig. 10.2.2, the developed accelerometer cancels the servo-signal leakage in the digital domain after a 1MS/s 13b SAR ADC. Since the servo-signal leakage passes through the analog BPF in the detection IC along with the modulated ΔC signal, its waveform is changed at the ADC output. To mimic the waveform, the 1b servo signal, which is a source of the servo-signal leakage, is shaped by a 16-tap FIR filter (which emulates the analog BPF) before it is subtracted from the ADC output for cancellation. The tap coefficients of the FIR filter (a_k) extracted by the LMS algorithm are proportional to ΔC_P while replicating the frequency response of the analog BPF under PVT variation. The up-sampled operation of the FIR filter to approximate a continuous-time filter was found to be unnecessary, since the extracted coefficients reflect the discrete-time to continuous-time transformation. The other kind of noise is generated by interaction between the high-frequency components of the 1b quantization noise, S(f) and the higher-order mechanical resonances of the mass, H(f). As shown in Fig. 10.2.3, the high-frequency mass vibration proportional to H(f)S(f) is further coupled with S(f) (while the electrostatic force is generated) and added to the in-band noise. To reduce this noise, a digital bandpass delta-sigma modulator (DBPDSM) is used to generate the 1b servo signal instead of the 1b quantizer alone, since the former can create a high-frequency notch in S(f) while preserving the low-frequency signal components. To minimize the overlap between H(f) and S(f), and hence the above noise, the position of the notch is placed in the region where the higher-order resonances are densely located by using the appropriate parameter G of the DBPDSM.

The MEMS chip was fabricated on 6-inch Si/SOI/Si bonding wafers. The mechanical mass weighs 18 milligrams and its fundamental mechanical resonance is located at 595Hz with Q value of 1500. The detection and driver ICs were fabricated with the 0.13μm CMOS and 0.35μm HV CMOS processes, respectively. The three chips were mounted on the board to form the accelerometer. The measured spectra of the 1b servo signal under DC input acceleration in the cases when either a DBPDSM or a 1b quantizer is used are shown in the upper left of Fig. 10.2.4, verifying the presence of the expected notch for the former case. The measured output spectra for the two cases are shown in the upper right of Fig. 10.2.4, where the low noise level was obtained only with DBPDSM, verifying the technique shown in Fig. 10.2.3. The measured output spectra with and without the servo-signal-leakage cancellation are shown in the lower left of Fig. 10.2.4, where the noise was reduced from 170 to 30ng/√Hz by the cancellation, verifying the technique shown in Fig. 10.2.2. The output spectrum measured at a dedicated site with small ambient vibrations is shown in the lower right of Fig. 10.2.4, verifying only 22ng/√Hz noise level on average from 10 to 400Hz and 400Hz input bandwidth along with the measured flat gain shown in the upper left of Fig. 10.2.5. The measured output vs. DC input acceleration shown in the upper right of Fig. 10.2.5 verifies the input range of 0.55g. The response to arbitrary input acceleration in the time domain is compared with that of a dedicated apparatus placed together in the bottom graph of Fig. 10.2.5, verifying good agreement. The measured total power is 17mW, of which the detection IC consumes 11mW from 1.4V/1.8V supplies and the driver IC consumes 6mW from 12V/1.8V supplies. As shown in Fig. 10.2.6, the developed accelerometer achieved outstanding FoM and 9× lower noise level compared to state-of-the-art low-noise low-power MEMS accelerometers. Die micrographs are shown in Fig. 10.2.7.

Acknowledgement:
The authors would like to thank Dr. Tatsuyuki Saito, Mr. Kazumasa Ando, Mr. Yasushi Tainaka, Dr. Tatemi Ido, Mr. Sohei Ohmori, Ms. Chisaki Takubo, Mr. Noriyuki Sakuma and Dr. Masaru Kokubo of Hitachi Ltd. for their invaluable contributions to this project. The authors would also like to thank Prof. Hitoshi Mikada of Kyoto University for his exemplary guidance and Prof. Akito Araya of University of Tokyo for his technical assistance during the measurement.

References:
[1] J. Laine et al., "A High-Sensitivity MEMS-Based Accelerometer," *The Leading Edge*, 33(11), 1234-1242, 2014.
[2] Y. Kamada et al., "A Low-Cost, 30 ng/√Hz Noise Floor MEMS Accelerometer for Large Sensor Network," *IEEE MEMS*, pp. 936-939, 2018.
[3] M. Pastre et al., "A 300Hz 19b DR Capacitive Accelerometer Based on a Versatile Front End in a 5th-Order ΔΣ Loop," *ESSCIRC*, pp. 288-291, Sept. 2009.
[4] X. Wang et al., "A 1.2μg/√Hz-Resolution 0.4μg-Bias-Instability MEMS Silicon Oscillating Accelerometer with CMOS Readout Circuit," *ISSCC*, pp. 476-477, Feb. 2015.
[5] H. Xu et al., "A Closed-Loop ΣΔ Interface for a High-Q Micromechanical Capacitive Accelerometer with 200 ng/√Hz Input Noise Density," *IEEE JSSC*, vol. 50, no. 9, pp. 2101-2112, Sept. 2015.
[6] H. Xu et al., "A 10 mW, 0.4 μg/√Hz, 700 Hz ΣΔ High-Order Electromechanical Modulator for a High-Q Micromechanical Capacitive Accelerometer," *IEEE Sensors Journal*, vol. 18, no. 3, pp. 1187-1194, Feb. 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

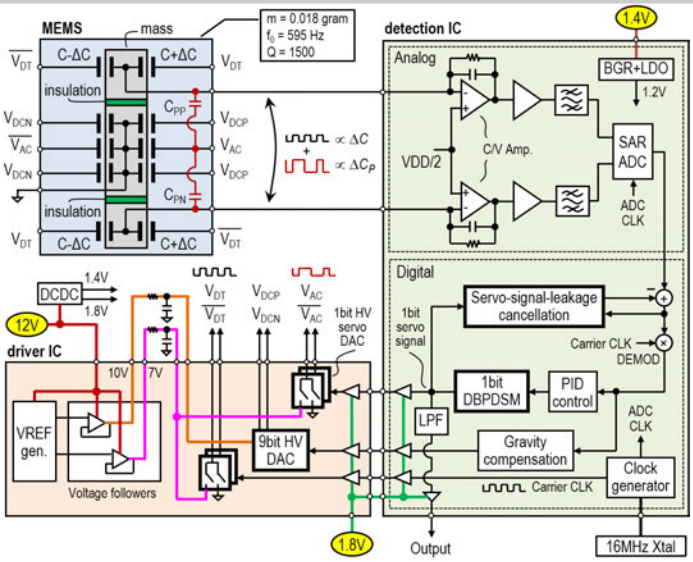


Figure 10.2.1: Overall block diagram of MEMS accelerometer.

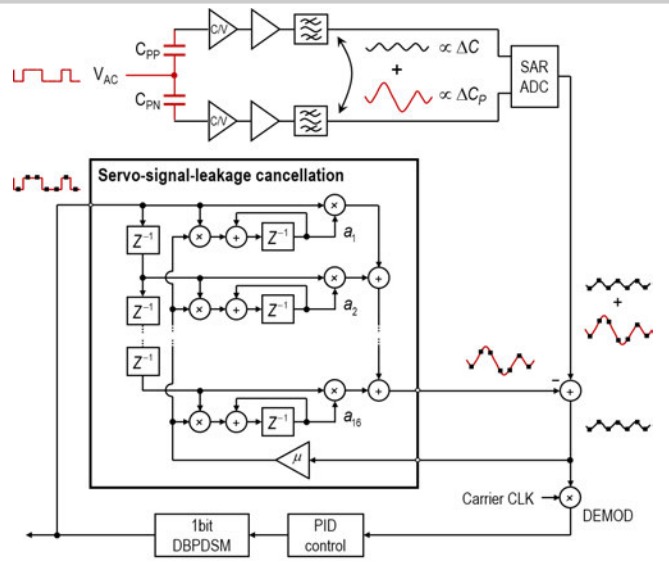


Figure 10.2.2: Digital cancellation of servo-signal leakage.

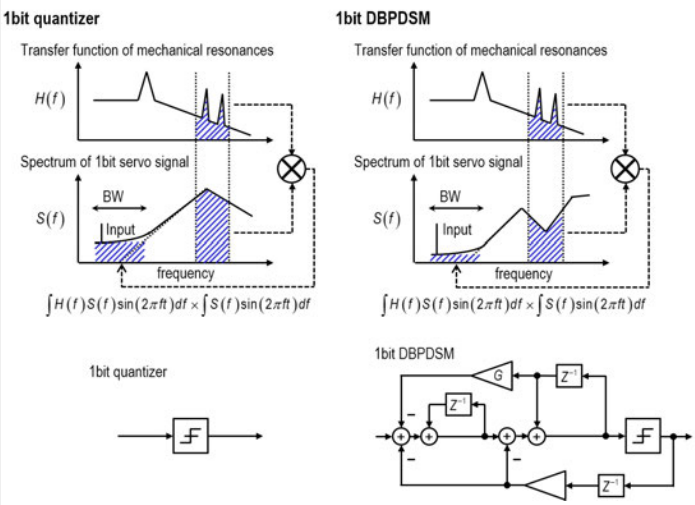


Figure 10.2.3: Effect of digital bandpass delta-sigma modulator.

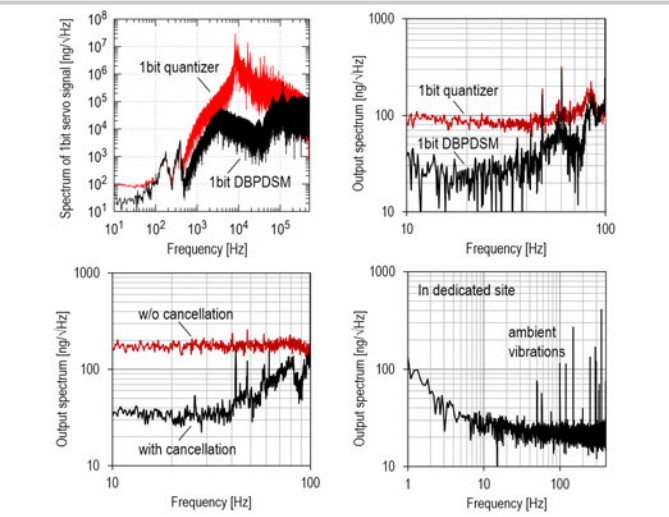


Figure 10.2.4: Measured spectra to verify the effect of DBPDSM (upper left and right), to verify the servo-signal-leakage cancellation (lower left), and to verify the noise level (lower right).

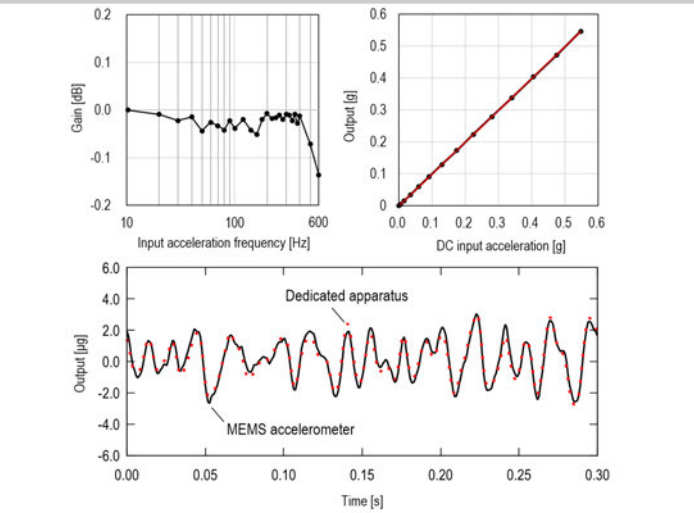


Figure 10.2.5: Measured gain dependency on input acceleration frequency (upper left), measured output vs. DC input acceleration (upper right), and measured time-domain response (bottom).

Parameter	This work	ESSCIRC2009 [3]	ISSCC2015 [4]	JSSC2015 [5]	Sensors J.2018 [6]
Process	0.13μm CMOS 0.35μm HV CMOS	0.6μm CMOS	0.35μm CMOS	0.5μm CMOS	0.5μm CMOS
Supply (V)	1.4 / 1.8 / 12	3.3 / 9	1.5	7	5.0
Power (mW)	17	12	4.4	23	10
Input range (g)	±0.55	±11	±20	±1.2	±0.7
Input bandwidth (Hz)	400	300	10	300	700
Noise (ng/√Hz)	22	1150	1200	200	400
Area (mm^2)	17.6 26.0	9.7	6.0	7.8	9.0
FoM1 (mW*ng/√Hz)	374	13800	5280	4600	4000
FoM2 (mW*ng/Hz)	19	797	1670	266	151

・FoM1 is defined by Noise (ng/√Hz)×Power (mW).

・FoM2 is defined by Noise (ng/√Hz)×Power (mW) / √(Bandwidth (Hz)) as seen in [5] and [6].

Figure 10.2.6: Performance summary and comparison with prior art.

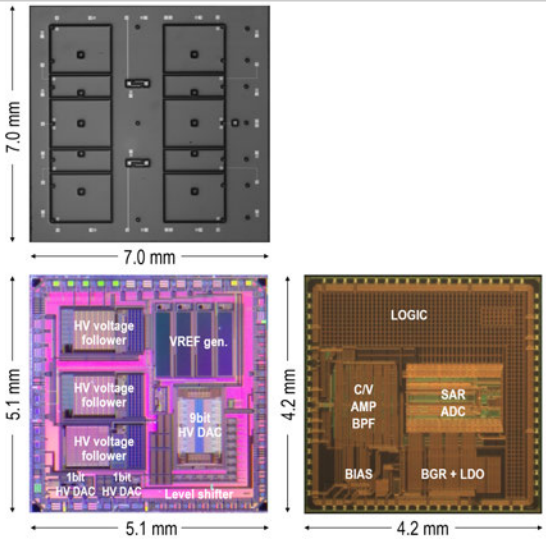


Figure 10.2.7: Die micrographs of MEMS (upper), driver IC (lower left), and detection IC (lower right).

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

10.3 A 0.12mm² Wien-Bridge Temperature Sensor with 0.1°C (3σ) Inaccuracy from -40°C to 180°C

Sining Pan, Çağrı Gürleyük, Matheus F. Pimenta, Kofi A. A. Makinwa

Delft University of Technology, Delft, The Netherlands

Resistor-based temperature sensors can achieve much higher resolution and energy efficiency than conventional BJT-based sensors [1], but they typically occupy more area (>0.25mm²) and have lower operating temperatures (≤125°C) [2-4]. This work describes a 0.12mm² resistor-based sensor that uses a Wien-bridge (WB) filter to achieve 0.1°C (3σ) inaccuracy from -40°C to 180°C. Compared to a state-of-the-art WB sensor [4], it occupies 6× less area and achieves comparable relative accuracy over a 76% wider operating range.

As shown in Fig. 10.3.1, the heart of the sensor is a Wien Bridge (WB) bandpass filter made from polysilicon resistors (R_{WB} = 64kΩ) and MIM capacitors (C_{WB} = 5pF). Unlike Wheatstone-bridge sensors [2,3], WB sensors only require one type of resistor, resulting in higher accuracy, and facilitating the experimental characterization of undocumented resistor properties such as 1/*f* noise and stress sensitivity. The filter is driven by a fixed-frequency square wave (f_{drive} = 500kHz) so that its temperature-dependent phase-shift ϕ_{WB} can be digitized by a phase-domain delta-sigma modulator (PDΔΣM) [3]. In the PDΔΣM, phase detection is performed by a chopper in the feedback loop of the 1st integrator (Fig. 10.3.2). Depending on the state of the bitstream BS, the chopper is driven by one of two reference phases $\phi_{0,1}$ (90°±30° w.r.t. to the phase of f_{drive}) such that the DC component of the integrated current is either positive or negative. The integration capacitor C_{int} must then be dimensioned to filter out the resulting ripple and ensure that the output of the 1st integrator does not clip. In [4], this required a 180pF capacitor that occupied more than half of the sensor's area.

To reduce the size of C_{int}, R_{WB} can be increased, but this will be at the expense of worse resolution. A better approach is to increase the output swing of the 1st integrator. In this work, the 1st integrator is built around a two-stage Miller-compensated opamp based on current-reuse amplifiers (Fig. 10.3.2). The 1st stage provides good energy efficiency, while the 2nd uses high-V_T devices to efficiently provide a near rail-to-rail output swing. Compared to the conventional choice of two common-source stages, the second stage provides twice the output current for the same bias current. Together with the doubling of R_{WB}, this allows the value of C_{int} to be reduced from the 180pF used in [4], to 23pF. At room temperature (RT), the amplifier draws 14μA and has a gain bandwidth product of 17MHz. To further reduce area, the 2nd integrator and the feed-forward coefficient are realized in a switched-capacitor manner, thus avoiding the large resistors used in [4]. The associated folded-cascode amplifier draws only 2.5μA at RT. On-chip logic generates the drive signal f_{ref} and the phase references $\phi_{0,1}$ from an external 6MHz frequency reference.

However, reducing C_{int} will increase the opamp's closed-loop input impedance Z_{in} ($\propto 1/(C_{int}\cdot GBW)$). This is in series with the WB and is thus a source of spread and 1/*f* noise. To minimize spread, a constant-Gm biasing circuit based on a resistor of the same type as R_{WB} ensures that Z_{in} tracks R_{WB} over a wide temperature range. Although the opamp is effectively chopped, the 1/*f* noise present in its bias current will modulate Z_{in}, and thus R_{WB}, causing residual 1/*f* noise. To minimize this, the core of the biasing circuit was realized with large PMOS devices (W/L = 40μm/5.5μm), and critical current mirrors were realized with the standard NPN transistors available in the chosen process (Fig. 10.3.2). Simulations show that the 1/*f* corner of the sensor is then about 1Hz and that R_{WB} is less than 1% of Z_{in} over corners.

Three pairs of identical sensors based on silicided p-poly (SP), unsilicided n-poly (NP) and high-resistive poly (HRP) resistors were fabricated on the same die in a 0.18μm CMOS process (Fig. 10.3.7). This facilitates the use of differential measurements to distinguish sensor drift from the inevitable ambient temperature drift. Each sensor consumes 29μA from a single 1.8V supply and occupies 0.12mm², of which the WB occupies 25%. PSDs of the bitstream outputs of the three sensors are shown in Fig. 10.3.3 (top) based on differential data captured during a 100s measurement interval. Figure 10.3.3 (bottom) shows the calculated resolution of each sensor versus conversion time. Due to its greater temperature coefficient (TC), the SP sensor exhibits the best resolution: 460μK in a 10ms conversion time, corresponding to a 110fJ·K² resolution FoM. For longer conversion times, sensor resolution is limited by 1/*f* noise. The corner frequencies are ~4Hz (NP and HRP), and ~1Hz (SP). However, the SP corner frequency is limited by the readout electronics.

10 chips (60 sensors) from a single batch were packaged in ceramic DIL and characterized from -40°C to 180°C. To correct for the inherent cosine nonlinearity of the PDΔΣM and the non-linear relationship between ϕ_{WB} and R_{WB} (Fig. 10.3.1, bottom), a 7th-order polynomial is used to translate the decimated output of each sensor into an equivalent sensor resistance R_{WB}. Since the temperature dependency of polysilicon resistors is comparatively linear, this approach minimizes the residual error after a 1st-order fit [4]. Figure 10.3.4 (top) shows the resulting temperature dependence of each resistor type. The following RT TCs were extracted: 0.31%/°C (SP), -0.15%/°C (NP) and -0.10%/°C (HRP), which agree with the process documentation. After a 1st-order fit to compensate for process spread, followed by the use of a fixed 6th-order polynomial to correct for the systematic non-linearity of the sensors, their residual spread is shown in Fig. 10.3.4 (bottom). The sensors achieve 3σ inaccuracies of 0.1°C (SP), 0.4°C (NP) and 0.9°C (HRP) from -40°C to 180°C. After a correlated 1-point trim [3], the SP sensor achieves an inaccuracy of 0.4°C (3σ), which is comparable with that of BJT-based sensors [5].

To observe the effects of mechanical stress, 10 chips from the same batch were characterized in injection-molded plastic QFN packages. Figure 10.3.5 (top) shows the average dependency of R_{WB} for both the ceramic and plastic packaged chips. After using the same non-linearity correction polynomials determined for the ceramic packaged chips, the inaccuracy after a 1st-order fit increases by only 0.2°C for the SP sensors, but to 1.4°C for the NP sensors, and even 2.5°C for the HRP sensors (Fig. 10.3.5 (bottom)). The sharp inflexion in all the inaccuracy plots around 100°C is probably due to the effects of moisture on the plastic packages [7]. Of the three resistor types, the SP resistor is clearly the least stress sensitive, exhibiting a packaging shift similar to that of BJT-based sensors [6].

In Fig. 10.3.6 the performance of the SP sensor is summarized and compared to the state-of-the-art. Compared to a state-of-the-art WB sensor [4], this design has a 76% larger operating range and occupies 6× less area. The latter is achieved at the expense of somewhat less relative inaccuracy [1]. Compared to a state-of-the-art WhB sensor [3], it achieves better accuracy and occupies 2× less area, but is less energy efficient. Measurements in plastic packages show that the SP sensors are quite insensitive to packaging stress (< 0.2°C packaging shift) and obtain good accuracy after a 1-point trim. They can thus replace BJT-based sensors in applications where both a wide temperature range and a high resolution are required.

Acknowledgments:
The authors would like to thank Zu-yao Chang for his help with the measurements, and the Advanced Packaging Center for expediting the plastic packaging.

References:
[1] K. A. A. Makinwa, "Smart Temperature Sensor Survey", [Online]. Available: http://ei.ewi.tudelft.nl/docs/TSensor_survey.xls
[2] C. H. Weng et al., "A CMOS Thermistor-Embedded Continuous-Time Delta-Sigma Temperature Sensor With a Resolution FoM of 0.65 pJ°C²," *IEEE JSSC*, vol. 50, no. 11, pp. 2491-2500, Nov. 2015.
[3] S. Pan and K. A. A. Makinwa, "A 0.25 mm²-Resistor-Based Temperature Sensor With an Inaccuracy of 0.12 °C (3σ) From -55 °C to 125 °C," *IEEE JSSC*, Oct. 2018.
[4] S. Pan et al., "A Resistor-Based Temperature Sensor with a 0.13pJ·K² Resolution FOM," *IEEE JSSC*, vol. 53, no. 1, pp. 163-174, Jan. 2018.
[5] K. Souri et al., "A 40μW CMOS Temperature Sensor with an Inaccuracy of ±0.4° C (3σ) from– 55° C to 200°C," *ESSCIRC*, pp. 221-224, Sept. 2013.
[6] B. Yousefzadeh et al., "A BJT-Based Temperature Sensor with a Packaging-Robust Inaccuracy of ±0.3°C (3σ) from -55°C to +125°C After Heater-Assisted Voltage Calibration," *ISSCC*, pp. 162-163, Feb. 2017.
[7] U. Ausserlechner et al., "Drift of Magnetic sensitivity of smart Hall sensors due to moisture absorbed by the IC-package," *IEEE Sensors*, pp. 455-458, Oct. 2004.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

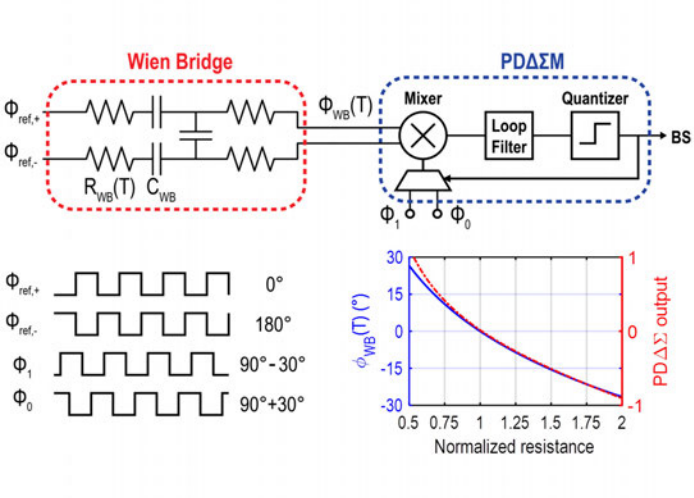


Figure 10.3.1: CTΔΣ readout of a Wien-bridge temperature sensor (top), waveforms, phase response of the Wien bridge and BS average of the PDΔΣM (bottom).

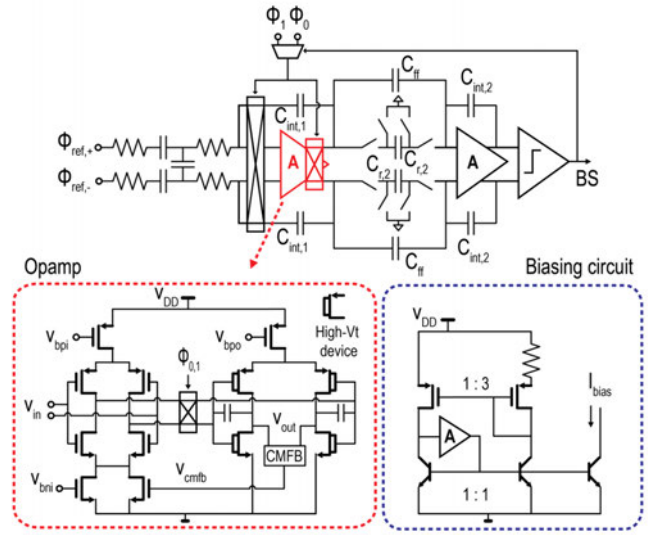


Figure 10.3.2: Block diagram of the PDΔΣM. Simplified diagram of the 1st-stage amplifier and the biasing circuit.

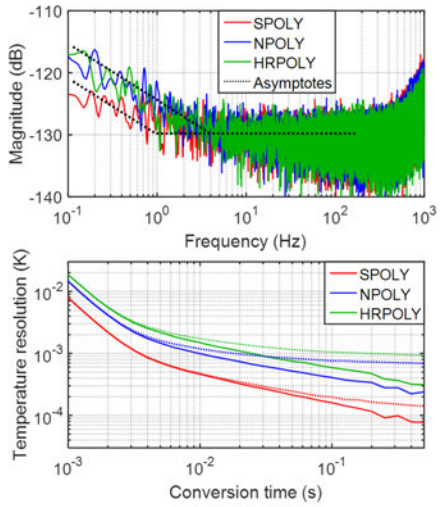


Figure 10.3.3: Bitstream spectra (100s interval, Hanning window) (top); Resolution vs. conversion time for 1s (solid lines) and 100s (dashed lines) intervals (bottom).

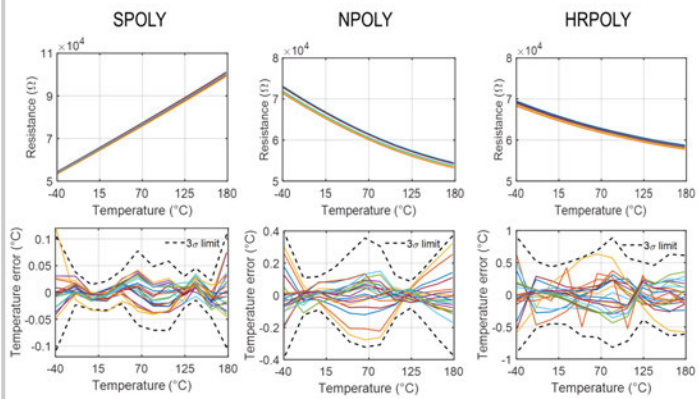


Figure 10.3.4: Extracted sensor resistance R_{WB} with ceramic packaging (top); Inaccuracy after a 1st-order fit and systematic non-linearity correction (bottom).

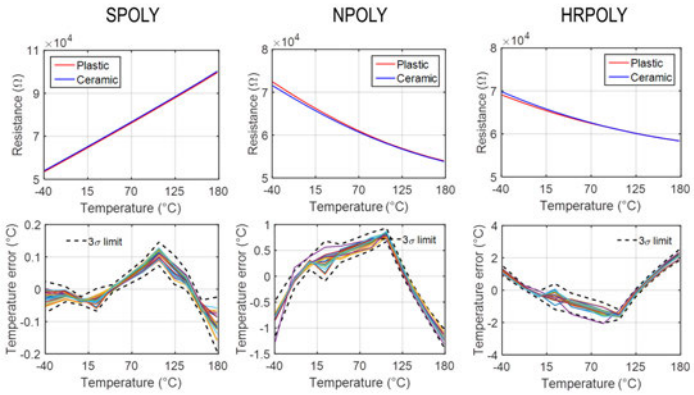


Figure 10.3.5: Average resistance vs. temperature of three types of resistors in different packages (top); Inaccuracy of plastic packaged sensors after a 1st-order fit and system non-linearity correction from ceramic packaged sensors (bottom).

	[2]	[3]	[4]	[5]	This Work	
Sensor type	Resistor WhB	Resistor WhB	Resistor WB	BJT	**Resistor WB**	
Technology	0.18µm	0.18µm	0.18µm	0.16µm	**0.18µm**	
Area [mm²]	0.43	0.25	0.72	0.1	**0.12**	
Temperature range	-40°C to 125°C	-55°C to 125°C	-40°C to 85°C	-55°C to 200°C	**-40°C to 180°C**	
Trimming points	2 [a]	2 [b]	2 [b]	1	**1**	**2 [b]**
3σ inaccuracy [°C]	0.4	0.12	0.03	0.4	**0.4**	**0.11**
Relative inaccuracy	0.48%	0.13%	0.05%	0.33%	**0.36%**	**0.10%**
Supply voltage [V]	1.5	1.8	1.8	1.6	**1.8**	
Power [µW]	65	94	160	35	**52**	
Conversion time [ms]	0.1	5	5	4.2	**10**	
Resolution [mK]	10	0.29	0.41	20	**0.46**	
Resolution FoM [fJ·K²] [c]	650	40	130	59000	**110**	

[a] 1-point trim and 1st-order fit. [b] 1st-order fit. [c] Energy / Conversion x (Resolution)².

Figure 10.3.6: Summary of measurement results and comparison with previous work.

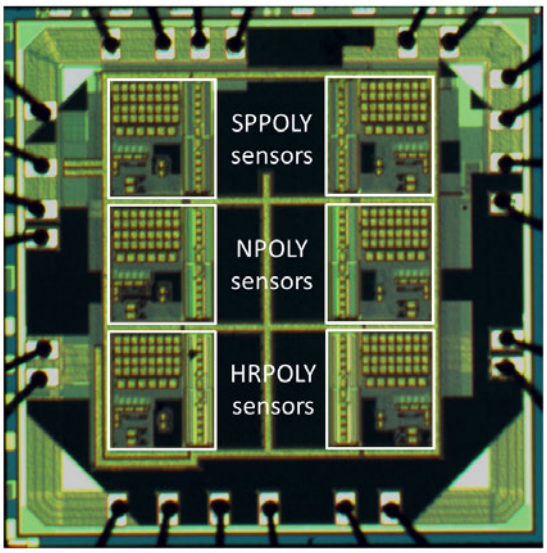


Figure 10.3.7: Micrograph of six temperature sensors on one die.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

10.4 A Wheatstone Bridge Temperature Sensor with a Resolution FoM of 20fJ·K²

Sining Pan, Kofi A. A. Makinwa

Delft University of Technology, Delft, The Netherlands

Temperature sensors intended for embedded applications should be both energy- and area-efficient. The combination of Wheatstone-bridge (WhB) sensors and continuous-time ADCs has proven to be highly energy efficient [1,2]. However, their area (> 0.25mm²) is still larger than that of comparable BJT-based sensors [3,4]. This paper presents a temperature sensor that uses an FIR-DAC ADC to reduce its area and increase energy-efficiency, both by 2× compared to the state-of-the-art [5].

A simplified model of the sensor is shown in Fig. 10.4.1 (top). To maximize its sensitivity, the WhB consists of resistors R_p (105kΩ, silicided p-poly) and R_n (100kΩ, non-silicided n-poly) with positive and negative temperature coefficients (TCs), respectively [1,2]. Its temperature-dependent output current I_{sig} is then digitized by a continuous-time delta-sigma modulator (CTΔΣM), which balances I_{sig} with the current I_{DAC} generated by a 2b resistive DAC R_{DAC} (4×720kΩ, non-silicided n-poly).

The use of a multi-bit DAC reduces the swing of the loop filter input current I_{err}, which decreases the power dissipation of the 1st integrator and the area of C_{int}. In [2], a zoom ADC is used, and the DAC state is determined by combining the result of an initial coarse SAR conversion with the output of a 1b CTΔΣM. Extra logic is then required to implement the SAR conversion, as well as the data-weighted-averaging (DWA) and segment-averaging schemes used to mitigate DAC mismatch and amplifier non-linearity, respectively.

In this design, a 1b CTΔΣM drives a FIR-DAC, thus ensuring 1b linearity without the need for extra logic [6]. For the same DAC resolution, the resulting I_{err} swing will then be about 2× less than that in the zoom ADC of [2], since the FIR-DAC does not require over-ranging (Fig. 10.4.1, bottom). As a result, both the size and area of C_{int} can be reduced, as well as the power dissipation of the 1st integrator .

Due to the uncorrelated spread of R_p and R_n, the nominal range of the WhB output current I_{sig} will spread significantly from batch to batch. To compensate for this, and so make optimal use of the input dynamic range of the modulator, a 4b batch trim is applied to R_p (~5.7kΩ/step). As shown in Fig. 10.4.2, the trimming scheme ensures that only one switch is in series with the selected segment of R_p. Compared to the trimming scheme in [1], this minimizes temperature-sensing errors due to the switches' finite on-resistance.

To achieve sub-mK resolution in a short conversion time (t_{conv} = 10ms), the modulator employs a 2nd-order feedforward architecture (Fig. 10.4.2). This also reduces the swing at the output of the 1st integrator, and thus reduces the area of C_{int1}. The 2nd stage consists of a switched-capacitor integrator (C_{S2} and C_{int2}) and a feedforward path (C_{FF2}). An extra FIR filter ($C_{FIR,C}$), is used to compensate for the delay introduced by the FIR-DAC [6].

As shown in Fig. 10.4.3, the opamp used in the 1st integrator consists of two current-reuse stages. This maximizes the noise efficiency of the input stage, and, compared to the use of two common-source amplifiers, halves the output stage bias current for a given maximum output current. The output stage uses high threshold devices to achieve a large output swing (~1.5V at room temperature (RT)). This, in turn, allows the area of C_{int1} to be further reduced. Compared to the OTA used in [2], which has a closed-loop input impedance of 1/gm, the closed-loop input impedance of the opamp is much lower, reducing its input swing and thus improving its linearity. The input stage is chopped to suppress its offset and 1/*f* noise, and so the chopping frequency f_{chop} must be chosen such that quantization noise is not downconverted to DC. In this design, f_{chop} can be set to either f_s = 500kHz [2] or, by exploiting the spectral notches of the FIR-DAC, to $f_s/8$ [6]. From simulations, the chopped opamp has a residual 1/*f* corner frequency of 2Hz, and a DC gain of 80dB, while its input/output stages consume 15μW/11μW, respectively. The 2nd stage is built around a cascoded telescopic OTA, which also has 80dB gain, but consumes only 3.5μW.

Two identical sensors were fabricated on the same die in a 0.18μm CMOS process (Fig. 10.4.7). This allows their resolution to be accurately estimated via differential measurements, which effectively reject ambient temperature drift. Each sensor consumes about 44μA (41μA analog and 3μA digital) from a 1.8V supply, and occupies 0.12mm², 60% of which is occupied by the WhB and the DAC. The sensors share the same clock generation circuit (0.003mm²). To further conserve area, C_{int1} (27pF, MIM) is located directly above the WhB. For flexibility, the sinc² decimation filters are implemented off-chip.

20 samples from one wafer (40 sensors) were mounted in ceramic DIL packages and characterized in a temperature-controlled oven. The packages were mounted in good thermal contact with a large aluminum block. After a batch trim, the residual spread from sample to sample is less than ± 3% full scale at RT. To mimic the effect of batch-to-batch spread, the sensor's output was characterized over temperature for two different trim code settings (Fig. 10.4.4 (top left)). After a 1st-order fit to compensate for process spread [2], the resulting systematic non-linearity differs by less than 3mK for the two trim-code settings, it also agrees well with simulations (less than 0.1°C difference) and so can be robustly corrected by a fixed 5th-order polynomial. The sensor then achieves an inaccuracy of 0.14°C (3σ) from –55°C to 125°C for both trim code settings (Fig. 10.4.4 (bottom)). Its supply sensitivity is quite low: ~0.03°C/V from 1.6 to 2V at RT. This remains constant up to about 500Hz, above which it exhibits a first-order roll-off due to the finite bandwidth of the 1st integrator's opamp.

FFTs of the sensor's bitstream output are shown in Fig. 10.4.5 (top). As designed, the sensor's noise is dominated by the WhB, and so changing f_{chop} from f_s to $f_s/8$ has no significant effect on its resolution. The observed 1/*f* noise (~20Hz corner frequency) is mainly due to the non-silicided poly resistors of the WhB [2]. Computing the standard deviation from the output of a single sensor over a long (20s) interval results in an estimated resolution that mainly reflects oven drift (Fig. 10.4.5 (bottom)). The effect of oven drift can be suppressed by computing the standard deviation from the difference in the output of the two sensors on each die. Alternatively, a shorter measurement interval can be used. As shown in Fig. 10.4.5 (bottom), over a 1s interval, the results of the single-ended and differential approaches agree well, resulting in a resolution of 160μK (rms) for t_{conv} = 10ms. With the differential approach, the effect of the sensor's own 1/*f* noise can be clearly seen: the resolution gets worse for longer (20s) measurement intervals.

In Fig. 10.4.6, the performance of this FIR-DAC WhB sensor is summarized and compared with state-of-the-art. It achieves the best FoM, improving on the state-of-the-art [2] by 2×. It is also the most compact high-resolution (sub-mK) temperature sensor, and is smaller than most precision BJT sensors [3,4]. These features make the sensor well suited for embedded applications in which both high-resolution and accuracy are required.

Acknowledgment:
The authors would like to thank Hui Jiang for his helpful comments and his contributions to the sensor layout.

References:

[1] C-H. Weng et al., "A CMOS Thermistor-Embedded Continuous-Time Delta-Sigma Temperature Sensor With a Resolution FoM of 0.65 pJ°C²," *IEEE JSSC*, vol. 50, no. 11, pp. 2491-2500, Nov. 2015.
[2] S. Pan and K. A. A. Makinwa, "A 0.25 mm²-Resistor-Based Temperature Sensor With an Inaccuracy of 0.12 °C (3σ) From -55°C to 125°C," *IEEE JSSC*, Oct. 2018.
[3] B. Yousefzadeh et al., "A BJT-Based Temperature-to-Digital Converter With ±60 mK (3σ) Inaccuracy From -55°C to +125°C in 0.16μm CMOS," *IEEE JSSC*, vol. 52, no. 4, pp. 1044-1052, April. 2017.
[4] M. K. Law et al., "A 1.1μW CMOS Smart Temperature Sensor With an Inaccuracy of ±0.2°C (3σ) for Clinical Temperature Monitoring," *IEEE Sensors Journal.*, vol. 16, no. 8, pp. 2272-2281, April, 2016.
[5] K.A.A. Makinwa, "Smart Temperature Sensor Survey", [Online]. Available: http://ei.ewi.tudelft.nl/docs/TSensor_survey.xls
[6] S. Billa et al., "Analysis and Design of Continuous-Time Delta–Sigma Converters Incorporating Chopping," *IEEE JSSC*, vol. 52, no. 9, pp. 2350-2361, Sept. 2017.
[7] M. H. Roshan et al., "A MEMS-Assisted Temperature Sensor With 20-μK Resolution, Conversion Rate of 200 S/s, and FOM of 0.04 pJK²," *IEEE JSSC*, vol. 52, no. 1, pp. 185-197, Jan. 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

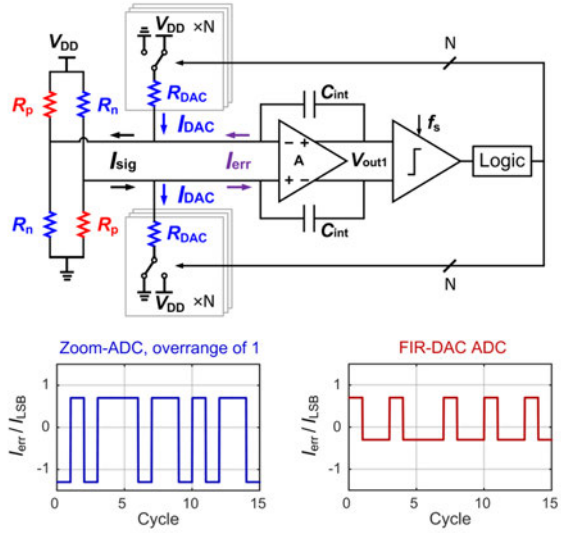


Figure 10.4.1: CTΔΣ readout of a differential Wheatstone bridge temperature sensor (top), error current I_{err} over time (bottom).

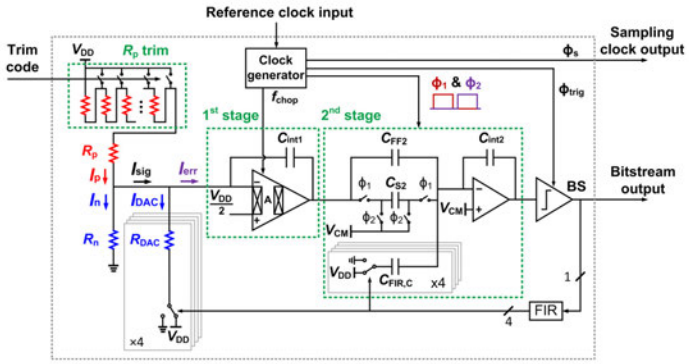


Figure 10.4.2: Simplified single-ended system block diagram.

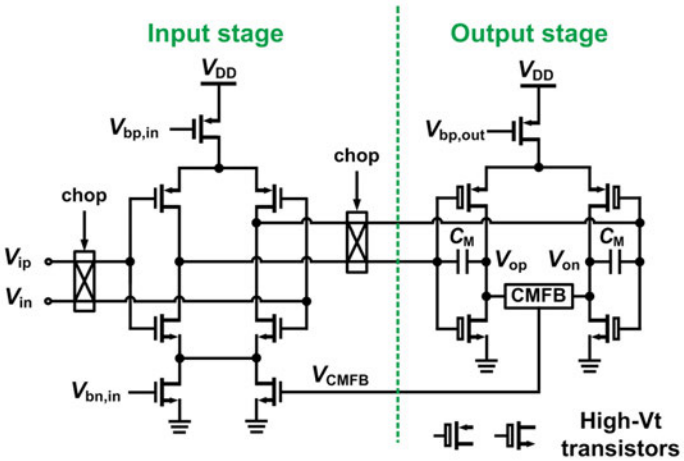


Figure 10.4.3: Schematic diagram of the 1st-stage opamp.

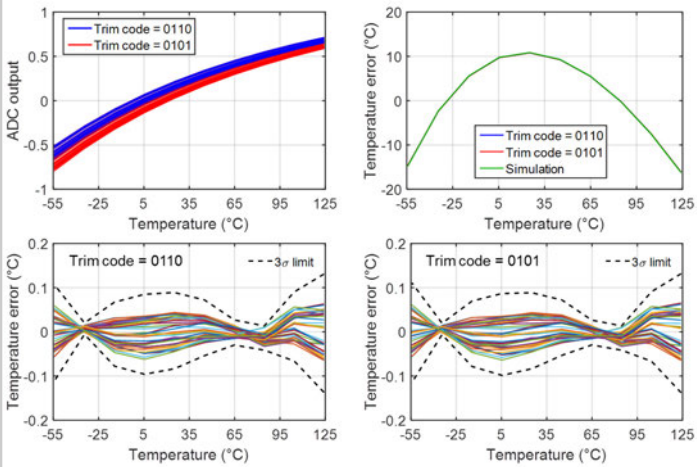


Figure 10.4.4: The sensor's characteristic (top left), systematic nonlinearity after a 1st-order fit (top right) and temperature error after individual 1st-order fit and fixed nonlinearity removal (bottom).

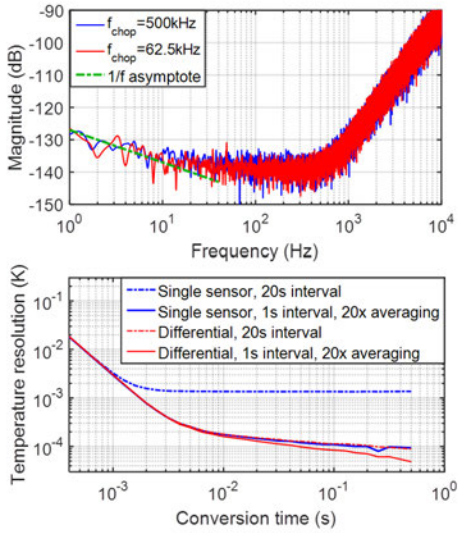


Figure 10.4.5: PSD of the sensor's bitstream (top), resolution vs. time plot (bottom).

	[7]	[3]	[4]	[1]	[2]	This work
Sensor type	Dual-MEMS Resonator	BJT	BJT	Resistor WhB	Resistor WhB	**Resistor WhB**
CMOS Technology	0.18µm	0.16µm	0.18µm	0.18µm	0.18µm	**0.18µm**
Area [mm²]	0.54	0.16	0.198	0.43	0.25	**0.12**
Temperature range	-40°C to 85°C	-70°C to 125°C	25°C to 45°C	-45°C to 125°C	-55°C to 125°C	**-55°C to 125°C**
3σ inaccuracy [°C] (Trimming points)	--	0.06 (1)	0.2 (1)	0.4 (2 [a])	0.12 (2 [b])	**~0.14 (2 [b])**
Supply voltage [V]	1.6	1.6	1 (analog) 1.8 (digital)	1.5	1.8	**1.8**
Power consumption [µW]	13000	7	1.1	65	94	**79**
Conversion time [ms]	5	5	500	0.1	5	**10**
Resolution [mK]	0.02	15	10	10	0.29	**0.16**
Resolution FoM [fJ·K²] [c]	40	7300	55000	650	40	**20**

[a] 1 point trim + 1st order fit. [b] 1st order fit. [c] FoM = Energy / Conversion x (Resolution)².

Figure 10.4.6: Performance summary and comparison with previous work.

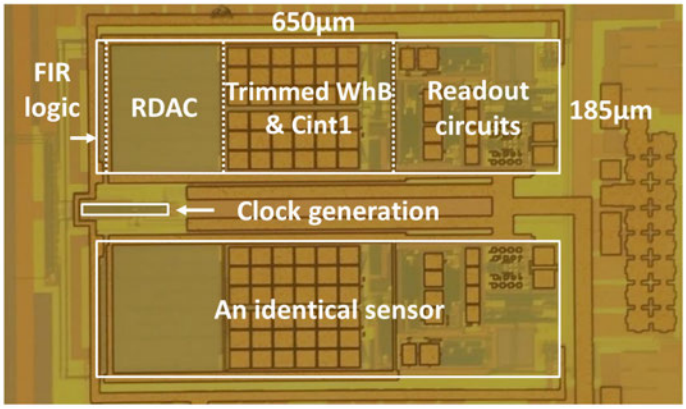


Figure 10.4.7: Die micrograph of the fabricated temperature sensor.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 11 Overview: *Diagnostics*

IMMD SUBCOMMITTEE

Session Chair: ***Joonsung Bae***
Kangwon National University, Chuncheon, Korea

Associate Chair: ***Michael Kraft***
KULeuven, Leuven, Belgium

Subcommittee Chair: Chris Van Hoof, IMEC, Leuven, Belgium

This session comprises five papers about medical diagnostic systems-on-chip that advance the state-of-the-art for biosensing, neural and ultrasound imaging. The first paper describes an ultrasound ASIC for an echocardiography system with low power consumption and extremely high integration density. The second paper presents a biosensor array based on a 3-electrode system used for voltammetry measurements for DNA analysis. The third paper demonstrates a biosensor for cancer diagnostics based on bio-functionalized microneedles using highly sensitive capacitive detection techniques. The fourth paper describes a frontend circuit for a magnetoresistive biosensor that can tolerate high mismatch between sensors. The fifth paper presents a minimally invasive CMOS fluorescence neural imaging probe based on 512 single-photon-avalanche diodes.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

10:15 AM

11.1 A 5.37mW/Channel Pitch-Matched Ultrasound ASIC with Dynamic-Bit-Shared SAR ADC and 13.2V Charge-Recycling TX in Standard CMOS for Intracardiac Echocardiography

J. Lee, KAIST, Daejeon, Korea

In Paper 11.1, KAIST, University of California at Berkeley, National University of Singapore and Singapore Institute of Neurotechnology present a pitch-matched ultrasound ASIC for an intracardiac echocardiography (ICE) system. The presented work achieves high integration density with a 250×250μm^2 unit cell and the low power consumption of 5.37mW with 5Vpp TX.

10:45 AM

11.2 A CMOS Biosensor Array with 1024 3-Electrode Voltammetry Pixels and 93dB Dynamic Range

A. Manickam, InSilixa, Sunnyvale, CA

In Paper 11.2, InSilixa Inc. describes a CMOS 32×32 biosensor array with integrated 3-electrode voltammetry 100×100μm^2 pixels, including a CMOS-compatible amorphous carbon electrode and temperature control. It achieves 280fA$_{rms}$ input-referred noise in a 0.1-to-20Hz bandwidth and 93dB dynamic range.

11:15 AM

11.3 A Capacitive Biosensor for Cancer Diagnosis Using a Functionalized Microneedle and a 13.7b-Resolution Capacitance-to-Digital Converter from 1 to 100nF

S. Song, Yonsei University, Seoul, Korea

In Paper 11.3, Yonsei University presents a capacitive 0.87mm^2 biosensor that detects a cancer biomarker using an electrochemical reaction on functionalized microneedles. Employing a CDC featuring 0.1nJ/step, it measures the capacitance between the microneedles with 13.7b resolution from 1 to 100nF and can detect the biomarker with 15fM (rms) resolution from 0.1pM to 1nM.

11:45 AM

11.4 A Fast-Readout Mismatch-Insensitive Magnetoresistive Biosensor Front-End Achieving Sub-ppm Sensitivity

X. Zhou, University of California, San Diego, La Jolla, CA

In Paper 11.4, the University of California San Diego demonstrates a fast-readout 180nm 0.249mm^2 CMOS front-end ASIC for a magnetoresistive biosensor with high mismatch insensitivity. Consuming 1.39mW/ch, it achieves 22.7× faster readout time, better than 7.8× lower baseline, and 2.3× lower power compared to the state of the art, while being able to tolerate up to 10% mismatch between sensors.

12:00 PM

11.5 A 512-Pixel 3kHz-Frame-Rate Dual-Shank Lensless Filterless Single-Photon-Avalanche-Diode CMOS Neural Imaging Probe

C. Lee, Columbia University, New York, NY; KIST, Seoul, Korea

In Paper 11.5, the Korea Institute of Science and Technology, Columbia University and the California Institute of Technology present a minimally invasive fluororence neural imager displacing only 0.012mm^2 tissue. It is based on two 70μm thin and 3.2mm long shanks with 256 single-photon-avalanche diodes each with 25μm pitch achieving spatial reconstruction of light sources as small as 60μm apart using angle sensitive gratings.

11.1 A 5.37mW/Channel Pitch-Matched Ultrasound ASIC with Dynamic-Bit-Shared SAR ADC and 13.2V Charge-Recycling TX in Standard CMOS for Intracardiac Echocardiography

Jihee Lee[1], Kyoung-Rog Lee[1], Benjamin E. Eovino[2], Jeong Hoan Park[3], Liwei Lin[2], Hoi-Jun Yoo[1], Jerald Yoo[3,4]

[1]KAIST, Daejeon, Korea
[2]University of California, Berkeley, CA
[3]National University of Singapore, Singapore, Singapore
[4]Singapore Institute for Neurotechnology, Singapore, Singapore

Intracardiac echocardiography (ICE) is an ultrasound sonogram that visualizes the anatomical structure of the heart in real time, with a mm-scale catheter inserted through the intracardiac vessels, and guides surgical intervention for atrial septal defect (ASD) closure. To achieve high-quality medical imaging, an ICE system must meet stringent power consumption requirements with low-noise operation. Since an ASIC and ultrasound transducers are tightly bonded through flip-chip or direct integration, an ultrasound unit TRX channel must be pitch-matched to each transducer channel [1,2]. A piezoelectric Micromachined Ultrasound Transducer (pMUT) is a suitable ultrasound transducer for implantable sensor applications, since it does not need high (~200V) bias that is a must in capacitive MUTs (cMUT) [3]. However, pMUT devices suffer from process variation, which leads to low image quality, and to date, no work addresses this issue for both TX/RX in real time. To meet all these requirements at once, we present a 6×6 TRX pitch-matched pMUT ASIC with a standard CMOS-compatible 13.2V HV pulser, on-chip per-pixel calibration scheme, and a Dynamic Bit-Shared (DBS) ADC for portable ICE applications.

Figure 11.1.1 shows the proposed ICE system. The catheter is 1.5m in length and 3mm in diameter; it integrates a 2D 6×6 5MHz 250μm-pitch pMUT array [3] and a pitch-matched 6×6 TRX ASIC in the tip. To ensure signal integrity even with a long wire, an on-chip DBS ADC digitizes the results on each channel prior to being transmitted. To prevent any tissue damage from overheating, low power consumption is a must, which also opens the door to further implantable ultrasound imaging systems.

Figure 11.1.2 shows the ASIC architecture. It consists of a 36-channel ultrasound TRX, a digital controller, a calibration block, a PMU, and a PLL. The unit ultrasound TRX channel is composed of a Charge-Recycling HV (CRHV) pulser, a TX/RX switch, an LNA, a time-gain compensation (TGC) amplifier, an analog delay cell, and a 10b SAR ADC. In order to reduce power consumption and compensate for pMUT process variation (aberration), 3 key features are proposed: (1) standard CMOS-compatible CRHV pulser, (2) a 4-channel 10b DBS SAR ADC, and (3) a large-dynamic-range TGC amplifier with automatic *per-pixel* calibration. An on-chip aberration compensation block is integrated into the calibration block to compensate for phase inequalities due to unsymmetrical geometry and velocity distribution of the pMUT array [4]. Moreover, the Received Signal Strength Indicator (RSSI) dynamically adapts the TGC gain depending on its input signal magnitude.

Figure 11.1.3 shows a schematic diagram of the CRHV pulser, along with its cross-sectional view. Previous works adopt HV process technology to facilitate HV operation, which increases the process cost, occupies a bulkier physical area with additional parasitic components, and most importantly, is not compatible with the most advanced process nodes [5]. In contrast, the CRHV pulser is fully compatible with standard CMOS, where the stacked I/O transistors always operate within its normal operation voltage range. By stacking M_1-M_6 transistors with staggered well potentials, stress to each transistor is dispersed, thereby compatible with standard CMOS process. The CRHV pulser operates in 2 phases: In the charging phase, the pMUT's outer and inner electrodes are connected to V_{HIGH} and ground, respectively. In the redistribution phase, the outer and the inner electrodes are shorted, re-using theoretically half the charge from the previous phase; this reduces the TX power, which dominates the unit TRX channel power consumption, by 32.8%. pMUTs inherently suffer from mismatch, and their acoustic pressure varies even with an identical excitation voltage, resulting in aberration [4]. To equalize the TX acoustic pressures, the gate voltage of M_1 and M_6 are controlled by the DAC with a 3b digital code, resulting in the real-time *per-pixel* calibration of TX channels by up to $3.3V_{pp}$. As a result, the standard variation of pulse-to-echo (P2E) maximum amplitude is reduced from 0.0304 to 0.0037, corresponding to an 87.8% reduction in aberration (Fig. 11.1.3).

Noting that the majority (~90%) of power dissipation in the RX path comes from the ADC, we propose a DBS SAR ADC (Fig. 11.1.4) to save power. As adjacent ultrasound pixels receive similar echo results with phase delay, they share a coarse ADC after subarray delaying. The 10b DBS SAR ADC shares 0/2/4/6 MSB with adjacent 2×2 channels; the coarse ADC pre-defines the switch polarity of the fine ADC, therefore there is a greater reduction in power consumption as more bits are shared. The analog delay cells perform subarray beamforming with 5ns resolution, and the remaining delay difference <5ns between channels changes the number of sharable bits of the coarse ADC by inverse proportion. The SNDR is measured to be 56.85dB with 9.15b ENOB. The DBS-ADC consumes 0.85mW with 6 shared coarse bits (vs. 1.23mW without bit-sharing), where up to 30.9% power reduction can be achieved.

Figure 11.1.5 shows the circuit schematic for closed-loop RX path with attenuation-adaptive gain/noise control. The RSSI is composed of 10 amplifier-rectifier (AR) cells and 9 bias filters. It measures the magnitude of input signals in a log scale to control the gain and noise [6]. The required dynamic range of the RX is 90dB to penetrate 4cm in human body The signal attenuation is roughly 40dB as the acoustic signal penetrates the body with an attenuation rate of 1dB/cm/MHz and the minimum display resolution is 50dB. As 90dB dynamic range is power consuming for the RX path, the TGC adaptively changes its gain and noise based on the signal magnitude measured by the RSSI. The RSSI can measure an input signal from -105dBm to -35dBm logarithmically and consumes 0.66mW; the RSSI is shared throughout 36 channels.

Figure 11.1.6 shows the B-mode image result of a phantom and the comparison table with other recent ultrasound ASIC works. An imaging phantom is constructed with polydimethylsiloxane (PDMS), where 3 wires are inserted inside. The phantom is placed on the top of the transducer array, and three wires are located 25mm, 28mm, and 35mm above from the bottom. The 36-channel pMUTs insonify 5MHz ultrasound pulse with $5V_{pp}$ to the phantom, and pulse-to-echo signals are recorded from -30° to +30° beam steering angle with 5° spacing. The recorded digital RX data shows clear echo signals when the beam is steered towards the wire. A B-mode imaging in the lateral direction with 60° field of view shows the presence and the position of the 3 wires in the phantom.

Figure 11.1.7 shows the chip micrograph and the performance summary. The ASIC in standard 180nm 1P6M CMOS occupies $11.75mm^2$ area, and TX consumes 3.31mW and 9.26mV for $5V_{pp}$ and $13.2V_{pp}$ mode, respectively. With the DBS ADC, ADC and RX power consumption are reduced by 30.9%, 28.6%, respectively, which is only 0.95mW per channel. The automatic calibrated TGC with RSSI adapts its gain and noise corresponding to its input signal magnitude. Compared to the previous ultrasound systems shown in the comparison table (Fig. 11.1.6), the proposed ultrasound ASIC achieves the highest integration scale which integrates *all* TX, RX, and ADC into $250\times250\mu m^2$ unit channel, and the lowest power consumption (5.37mW with $5V_{pp}$ TX), which effectively prevents the tissue overheating.

Acknowledgement:
The authors acknowledge Dr. Judyta B. Tillak and Ms. Luna Yue Liang for their technical discussions. The work was funded by NUS Hybrid-Integrated Flexible Electronic Systems Program.

References:
[1] M. Chen, et al., "A Pixel-Pitch-Matched Ultrasound Receiver for 3D Photoacoustic Imaging with Integrated Delta-Sigma Beamformer in 28nm UTBB FDSOI," *ISSCC Dig. Tech. Papers*, pp. 456-457, Feb. 2017.
[2] C. Chen, et al., "A 0.91mW/element Pitch-Matched Front-End ASIC with Integrated Subarray Beamforming ADC for Miniature 3D Ultrasound Probes," *ISSCC Dig. Tech. Papers*, pp. 186-187, Feb. 2017.
[3] B.E. Eovino, et al., "Broadband Ring-Shaped pMUTS Based on an Acoustically Induced Resonance," *IEEE MEMS*, pp. 1184-1187, Jan. 2017.
[4] J. Tillak, et al., "A 2.34μJ/scan Acoustic Power Scalable Charge-Redistribution pMUT Interface System with On-Chip Aberration Compensation for Portable Ultrasonic Applications," *IEEE ASSCC*, pp. 189-192, Nov. 2016.
[5] G. Jung, et al., "Single-Chip Reduced-Wire Active Catheter System with Programmable Transmit Beamforming and Receive Time-Division Multiplexing for Intracardiac Echocardiography," *ISSCC Dig. Tech. Papers*, pp. 188-189, Feb. 2018
[6] J. Jang, et al., "A 540μW Duty Controlled RSSI with Current Reusing Technique for Human Body Communication," *IEEE TBioCAS*, vol. 10, no. 4, pp. 893-901, Aug. 2016.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

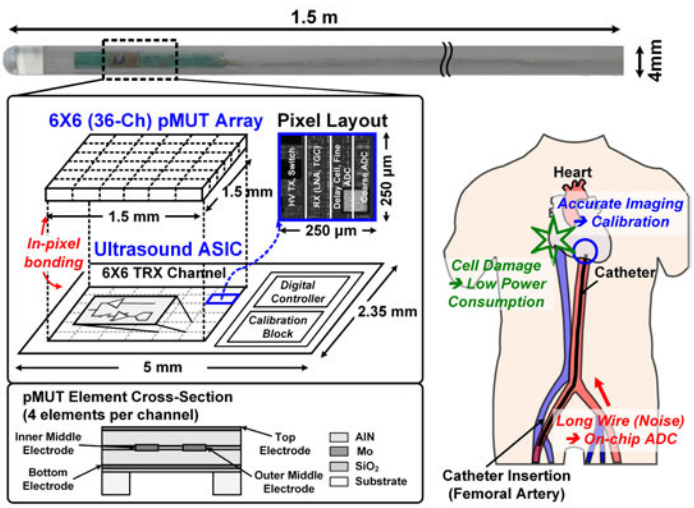


Figure 11.1.1: Proposed intracardiac echocardiography (ICE) system.

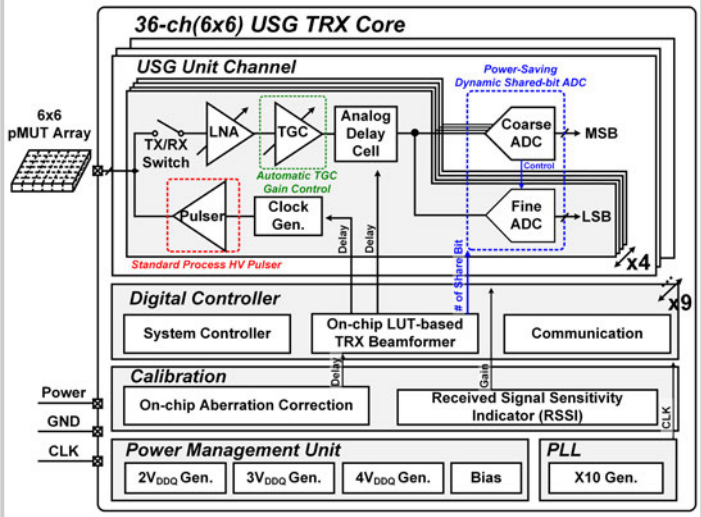


Figure 11.1.2: Overall architecture of ICE ASIC.

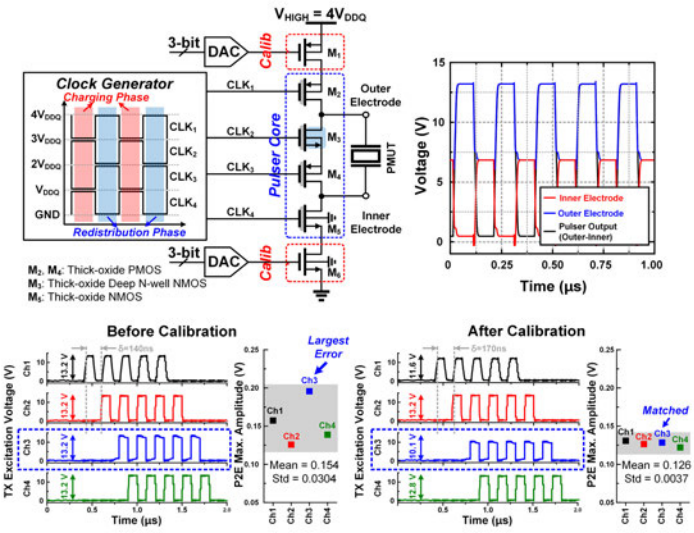


Figure 11.1.3: $4V_{DDQ}$ Charge Recycling High Voltage (CRHV) pulser with standard CMOS process.

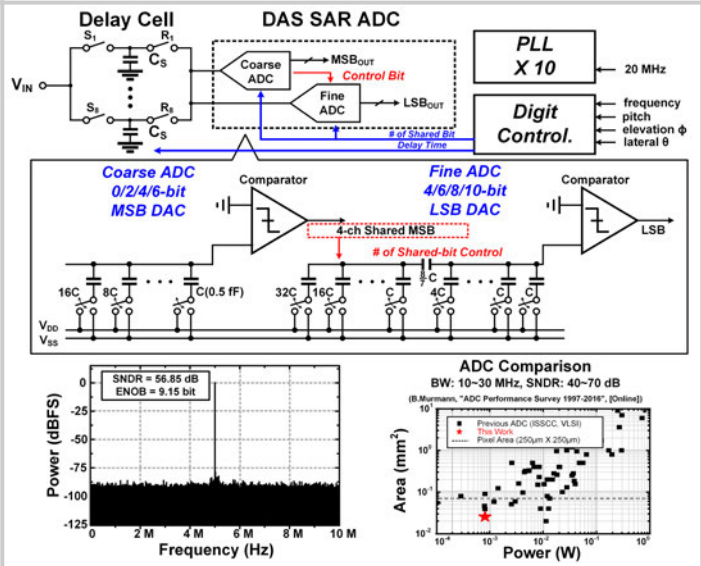


Figure 11.1.4: Power saving 4-channel Dynamic Bit-Shared (DBS) SAR ADC.

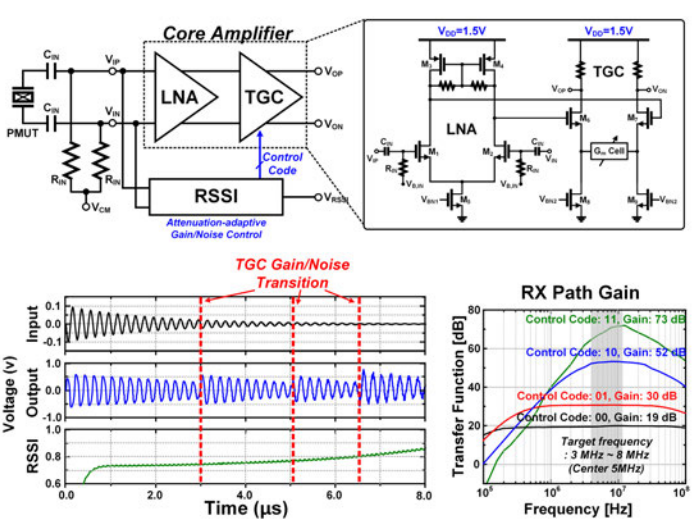
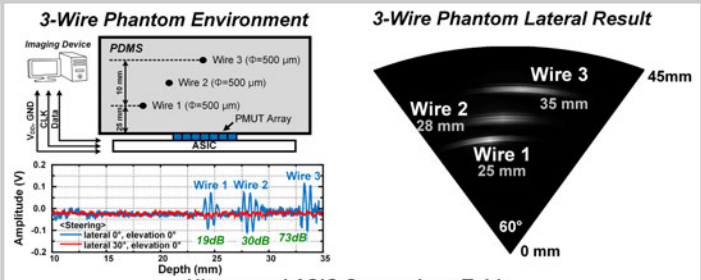


Figure 11.1.5: 90-dB dynamic range RX with attenuation-adaptive gain/noise calibration.

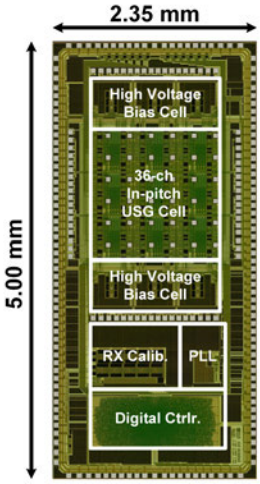


Ultrasound ASIC Comparison Table

	J.Y. Um ISSCC'14	J.Tillak ASSCC'15	M.C.Chen JSSC'17	G.Jung ISSCC'18	C.Chen ISSCC'18	This Work
Process	0.13 μm	65 nm	28 nm	0.18 μm HV	0.18 μm	**0.18 μm**
Transducer	2D CMUT	PMUT	2D CMUT	PZT / CMUT	2D PZT	**2D PMUT**
Pitch-Matched	No	No	Yes (250 μm)	No	Yes (150 μm)	**Yes (250 μm) ***
TX/RX per ch.	Only RX	TX / RX	Only RX	TX / RX	RX	**TX / RX**
ADC per ch.	X	X	O	X	X (1ADC/4ch)	**O**
Area/ch.	0.303 mm²	0.13 mm²	0.0625 mm²	0.45 mm²	0.026 mm²	**0.0625 mm²**
Max. TX Voltage	-	6 V	-	60 V	-	**13.2 V**
TX Power/ch.	-	N/A	-	N/A	-	**3.31 mW ($5V_{pp}$)** **9.26 mW ($13.2V_{pp}$)**
RX Power/ch.	17.81 mW	N/A	17.5 mW	N/A	0.91 mW	**0.95mW ****
Total Power/ch.	17.81 mW	N/A	17.5 mW	6.26 mW	0.91 mW	**5.37 mW (TX $5V_{pp}$)** **10.33 mW (TX $13.2V_{pp}$)**
TRX BF	RX	TX	RX	TX	RX	**TX / RX**
TRX On-chip Cal.	No	Yes	No	No	No	**Yes**

*This work is the *only* ASIC which integrated both TX & RX in a pitch-matched channel **6 shared coarse bit

Figure 11.1.6: B-mode imaging result of 3-wire phantom and ultrasound system comparison table.



Process			180 nm 1P6M Standard CMOS
Die Size			2.35 x 5.0 mm^2
Supply Voltage			1.5 V
Transducer			2D PMUT
TX		Frequency	1/2/5/10 MHz
		Power	3.31 mW ($5V_{pp}$)
			9.26 mW ($13.2V_{pp}$)
		Amplitude	Max. 13.2 V_{pp}
RX	LNA+ TGC	Gain	29 / 30 / 42 / 53 dB
		BW	3 ~ 13 MHz
		Power	82.2 μW
	Delay Cell	Delay min.	5 ns
		Delay max.	1.28 μs
	ADC	Sampling Rate	20 MHz
		SNDR	56.85 dB
		Power	0.85 mW (6-bit shared)
			1.23 mW (0-bit shared)
	RSSI	Dynamic Range	-105 ~ -35 dBm
		Power	0.66 mW
Digital Controller		Frequency	20 MHz / 200 MHz
		Power	4.38 mW

Figure 11.1.7: Chip micrograph and ASIC performance summary.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

11.2 A CMOS Biosensor Array with 1024 3-Electrode Voltammetry Pixels and 93dB Dynamic Range

Arun Manickam[1], Kae-Dyi You[2], Nicholas Wood[1], Lei Pei[1], Yang Liu[1], Rituraj Singh[1], Nader Gamini[1], Davood Shahrjerdi[2], Robert G. Kuimelis[1], Arjang Hassibi[1]

[1]InSilixa, Sunnyvale, CA
[2]New York University, New York, NY

Electro-analytical (E-chem) biosensors offer unique advantages over widely used optical biosensors and can be considered ideal for low-cost, mass-deployable point-of-care (PoC) diagnostic devices [1]. They possess fully electronic and real-time transduction methods and require little or no external instrumentation. Despite these advantages, two fundamental challenges have hampered broad adoption of E-chem biosensors: IC design and transducer (surface) material suitability. E-chem sensors require both low-noise and high detection dynamic range (DDR) front-end circuits and must also accommodate electrode-electrolyte interfaces with significant PVT variations and temporal drifts [2-5]. The absence of CMOS-compatible bio-electronic interfaces with adequate chemical and thermal stability has been another impediment [6].

We present a fully integrated CMOS E-chem biosensor array (Fig. 11.2.1) that addresses these core challenges. The disposable biosensor module comprises a 7×9mm^2 CMOS E-chem IC, mounted on a PCB substrate and encapsulated by a fluidic cap (~50µl reaction volume). The IC includes a 32×32 array of 3-electrode voltammetry E-chem sensor pixels, an on-chip temperature control system, and peripheral circuitry for array readout and testing. Surface functionalization is accomplished by covalent immobilization of DNA to amorphous carbon (a-C) electrodes.

In 3-electrode voltammetry, the current passing through working electrode (WE), I_w, is measured when varying V_w across its electrode-electrolyte interface by connecting the counter electrode (CE) and the reference electrode (RE) in a potentiostat configuration. In biosensing applications, I_w consists of two components, I_X the small faradaic reduction-oxidation (redox) current generated by electro-active labels, and $I_b = C_w\, dV_w/dt$, the background polarization current originating from C_w, the WE electrode-electrolyte capacitance.

As depicted in Fig. 11.2.1, each pixel includes a dedicated WE, and a shared CE and RE. The electrodes are created by covering the exposed top metal of the CMOS stack with ~100nm of a-C on top of a 10-to-15nm titanium (Ti) adhesion layer. Sputtered a-C surfaces, unlike widely used gold (Au) surfaces [6], are CMOS-compatible and enable DNA probe immobilization via stable covalent bonds. No external electrode is required, as the generated I_w is independent of a-C surface variations. This is accomplished by making the areas of shared CE and RE much larger than pixel WE (ratio>430) to ensure that, in all test conditions, the impedances of CE and RE interfaces remain negligible compared to WE.

A high DR ΔΣ current detection sensor is integrated in every 100×100µm^2 pixel to measure I_w and produce a noise shaped digital output with an oversampling frequency of $f_{\Sigma\Delta}$ (Fig. 11.2.2). The CTIA integrates I_w while continually setting V_w. A clocked comparator acts as 1b ADC and quantizes the CTIA output (V_{out}) and, based on its value, a DAC adds or subtracts a quantized charge (current) at the CTIA input. During array readout, the row and column decoders read the single-bit digital output of all pixels sequentially and stream them off-chip using a LVDS circuit for decimation. To implement the thermal control system, a uniformly distributed resistive heater is integrated in the top metal layer that can deliver up to 10W of power from a 5V supply. To estimate the local temperature, 13 temperature sensor pixels are situated within the array to monitor the reverse-biased current of a CMOS N_{well}/P_{sub} diode.

In Fig. 11.2.3, we show the circuit schematic and the timing diagram of the ΔΣ current sensing pixel. The CTIA uses a high bandwidth (~30MHz) gain-boosted folded-cascode OTA optimized for low-noise performance. This topology was selected to mitigate the adverse effects of a large C_w (varying between 40pF to 1nF) that not only increases the noise contribution of the OTA, but also simultaneously slows down the DAC speed by increasing the discharge time of C_d. This OTA has a large-area input differential pair to reduce its input-referred voltage offset, which in turn reduces the offset current $I_{os} \approx C_d V_{os} f_{\Sigma\Delta}$. The output of the CTIA, V_{out}, is connected to a Yukawa comparator and D_{out}, its latched output, drives the shared row bus through a tri-state buffer when selected by the ROW signal. D_{out} is also used as the input to the in-pixel DAC to select the feedback current of $C_d\,(V_{d1} - V_w) f_{\Sigma\Delta}$ when $D_{out} = 1$ or $C_d\,(V_{d2} - V_w) f_{\Sigma\Delta}$ when $D_{out} = 0$.

Several configuration and testing capabilities are implemented. The pixels can select V_{w1} or V_{w2}, two externally generated waveforms, using the PROG signal as the source for V_w, to allow testing of specific pixels. In addition, the conversion gain can be set "high" or "low" by changing C_d between 100 and 250fF via the global GAIN signal, which results in LSBs of 1.25 and 3.125fA, respectively. The feedback capacitance of the CTIA, C_f, is changed in accordance with C_d to keep ΔV_{out} constant. Finally, for test purposes, an asynchronous analog sample and hold circuitry is available at every pixel, to read V_{out} independently using dedicated column decoders and analog column amplifiers.

In Fig. 11.2.4a-c, we report the measured electrical performance of this chip including the pixel input-referred noise power spectral density (PSD), pixel I_{os}, I_b and C_w distributions, and temperature sensing diode current. The integrated noise within the 0.1-to-20Hz bandwidth is 280fA rms in 100mM KCl). This noise floor translates to ~93dB DDR in low gain mode, considering the upper limit, set by the DAC, of ±12.5nA. In this configuration, the usable detection range for I_x becomes >98% of the full scale as the worst-case background current (max $\{I_b + I_{os}\}$) is <250pA. Figure 11.2.4d shows the average current measured by the temperature sensor pixels, I_t, which is dominated by linear I_{os} and exponential I_d at lower and higher temperatures, respectively.

The measured results for 3-electrode voltammetry experiments with varying concentrations of methylene blue (MB) redox molecule in the solution are shown in Fig. 11.2.5a. V_w is cycled between ±0.75V with a scan rate of 0.5V/s. The slope in I_x peak using 6 different MB concentrations is interpolated to be ~4pA/µM, which, considering the pixel noise, translates to a lower detection limit 70nM for MB.

DNA "melt" experiments between 35 and 80°C were performed to both demonstrate array-based DNA detection and to prove the stability of immobilized probes on the WE. As shown in Fig. 11.2.5b, a stem-loop DNA hairpin probe [6] with MB label is spotted on multiple WEs and the chip increases the temperature gradually and continuously records I_w. The MB label is close to the WE surface at lower temperature and moves away from the surface at higher temperatures, when the DNA hairpin structure opens, causing a reduction in I_x.

In Fig. 11.2.5c, kinetics of DNA hybridization was monitored in real time using different DNA sequences, one acting as the capture probe and one acting as the negative control, based on the protocol in [6]. If a complementary sequence is added to the reaction, the DNA hairpin probe captures it, which in turn opens the hairpin, and again pushes MB away from the surface to reduce I_x. Our results demonstrated this phenomenon with τ~410s while the negative control signal was unchanged.

A performance summary and comparison to previous work is reported in Fig. 11.2.6. As evident, this system satisfies all the stringent electronic and thermal requirements of biosensing. In addition, it offers a bio- and CMOS-compatible transducer and surface material suitable for high volume manufacturing.

References:

[1] J. Wang, "Electrochemical Biosensors: Towards Point-of-Care Cancer Diagnostics," *Biosensors and Bioelectronics,* vol. 21, no. 10, pp. 1887-1892, 2006.
[2] M. Schienle, et al., "A Fully Electronic DNA Sensor with 128 Positions and In-Pixel A/D Conversion," *IEEE JSSC,* vol. 39, no. 12, pp. 2348-2445, 2004.
[3] P. M. Levine, et al., "Active CMOS Sensor Array for Electrochemical Biomolecular Detection," *IEEE JSSC,* vol. 43, no. 8, pp. 1859-1871, 2008.
[4] C. Yang, et. al., "Amperometric Electrochemical Microsystem for a Miniaturized Protein Biosensor Array," *IEEE TBioCAS, vol.* 3, no. 3, pp. 160-168, 2009.
[5] H. M. Jafari, et al., "Nanostructured CMOS Wireless Ultra-Wideband Label-Free PCR-Free DNA Analysis SoC," *IEEE JSSC* vol. 49, no. 5, pp. 1223-1241, 2014.
[6] C. Fan, et al., "Electrochemical Interrogation of Conformational Changes as a Reagentless Method for the Sequence-Specific Detection of DNA," *PNAS, vol.* 100, no. 16, pp. 9134-9137, 2003.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

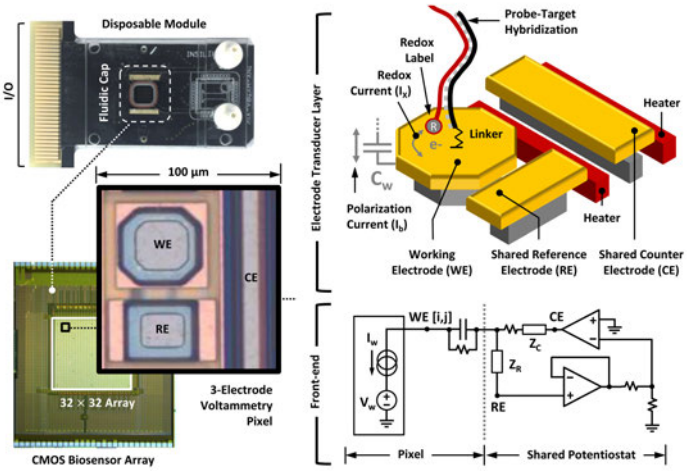


Figure 11.2.1: Disposable module, biosensor array and 3-electrode voltammetry pixels.

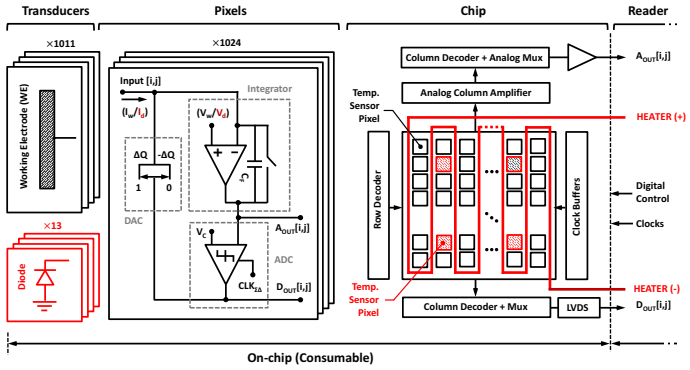


Figure 11.2.2: Array and pixel-level architecture.

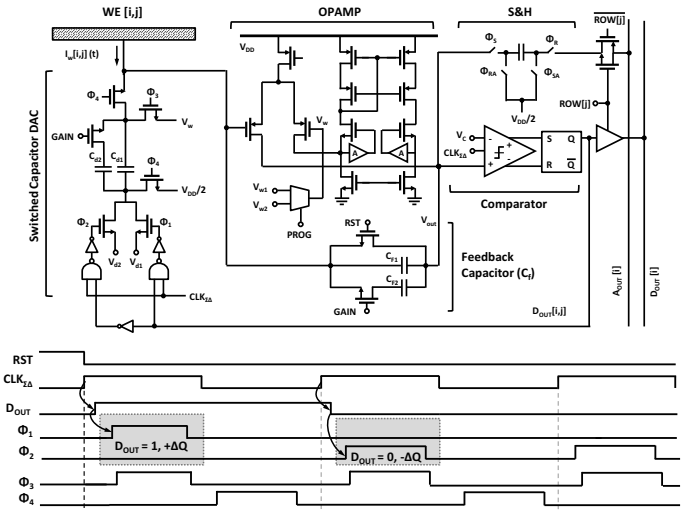


Figure 11.2.3: Circuit schematic and timing diagram of the ΔΣ current sensing pixel.

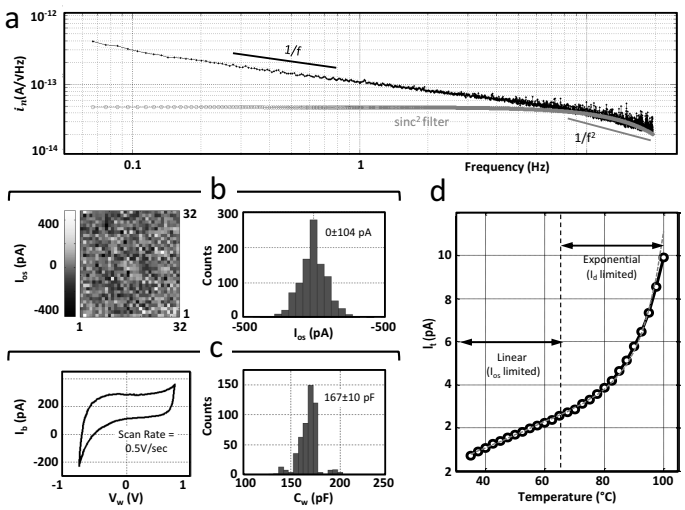


Figure 11.2.4: Measured electrical performance.

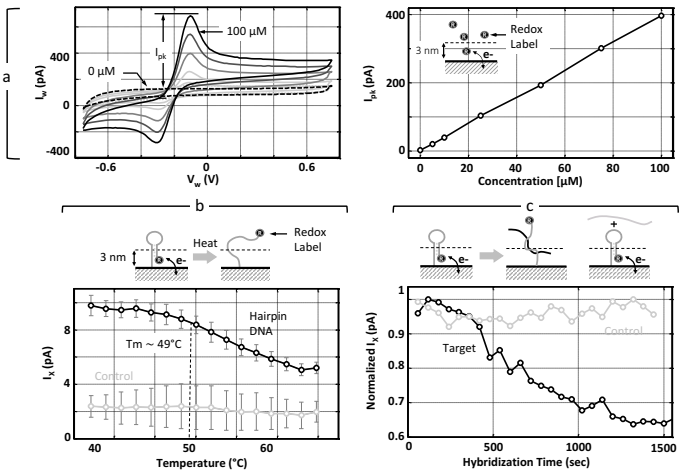


Figure 11.2.5: Detecting redox labels in (a) solution, (b) DNA heat switch structure, and (c) DNA label-free hybridization assay.

Specification	JSSC 2004 [2]	JSSC 2008 [3]	TBIOCAS 2009 [4]	JSSC 2014 [5]	This work
Sensor Signal Chain	Current-to-Frequency Converter	CTIA + Dual Slope ADC	CTIA + Analog Out	Current Conveyer + Dual Slope ADC	ΣΔ Current Detector
Pixel Area (µm²)	100,000†	250,000†	-	60,000	10,000
Electrode Material	Ti/Pt/Au	Ti/Au	Ti/Au	Ni/Pd/Au	Al/Ti/a-C
Number of Pixels	128	16	-	54	1024
I_{max} (nA)	100	250	1 10,000	2.2, 59, 350	12.5
Sensitivity (pA)	10	240	6 60,000	8.6, 132, 1050	0.28
Dynamic Range (dB)	80	60	44	48, 53, 51†	93#
Readout Speed (Hz)	7	2500	100	3400	50
CMOS Fabrication Process	0.5 µm	0.25 µm	0.6 µm	0.13 µm	0.25 µm
Power (mW)	-	10†	3	0.042 / pixel	0.25 / pixel

† approximated value # 0.1 – 20 Hz Bandwidth

Figure 11.2.6: Performance comparison table.

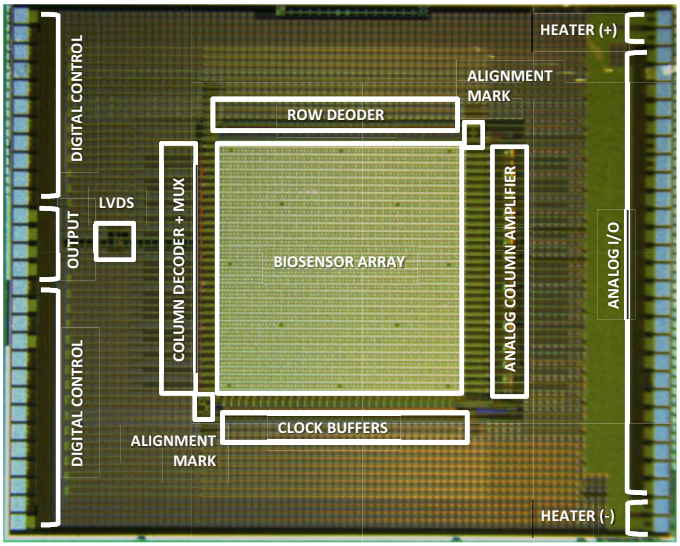


Figure 11.2.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

11.3 A Capacitive Biosensor for Cancer Diagnosis Using a Functionalized Microneedle and a 13.7b-Resolution Capacitance-to-Digital Converter from 1 to 100nF

Seungwoo Song, Jukwan Na, Moonhyung Jang, Hyeyeon Lee, Hyesoo Lee, Yongbeom Lim, Heonjin Choi, Youngcheol Chae

Yonsei University, Seoul, Korea

A malignant tumor consists of rapidly growing cancer cells, and requires a dedicated blood supply to provide oxygen and nutrients. Therefore, vascular endothelial growth factor (VEGF), a signal protein produced by cells stimulating angiogenesis, is considered as a key biomarker in clinical diagnosis of cancers [1,2]. There are already existing methods for the VEGF detection requiring advanced instruments and complex protocols [2]. Recently, significant progress has been achieved in biosensors for the detection and quantification of VEGF using synthetic receptors [2]. In particular, a capacitive biosensor detects the change of dielectric properties when the receptor binds to VEGF, and capacitance change can be used to quantify the reactions. However, the sensitivity of the capacitive biosensors still needs to be improved for use in cancer diagnosis.

This paper presents a capacitive biosensor for cancer diagnosis in human blood, which detects VEGF with a 15fM rms resolution over a wide detection range from 0.1pM to 1nM. The sensor is implemented with functionalized microneedles with peptide receptors, and the electrochemical reaction with VEGF changes the capacitance between two microneedles. A large capacitance variation of the microneedle is handled by an adaptive 5b current stimulator, and a small capacitance change due to the reactions can then be readout by a high-resolution capacitance-to-digital converter (CDC) based on a continuous-time delta-sigma modulator (CTΔΣM), achieving 13.7b resolution in the capacitance range from 1 to 100nF.

Synthetic peptides allow site-specific labeling with protein-peptide interactions. We developed a high-affinity peptide with VEGF by acetylation to the peptide V_{114} sequence as specified in [1]. It is directly assessable from the blood and constant agents remain when coupled to VEGF. Microneedles are fabricated with a Si substrate using a combination of bottom-up and top-down approaches to provide a vertical height of >600μm that can reach below dermis (Fig. 11.3.1). Au-coated microneedles are functionalized with the peptide receptor. After the VEGF-peptide reaction, the dielectric constant between two microneedles changes, and thus the capacitance change ΔC_{VEGF} represents the VEGF concentration. The sensor circuit can be approximated by a series combination of the sensor capacitor (C_{VEGF}) and the fluid impedance (Z_f). Since the sensor's 1/*f* noise corner is about 100Hz, the driving frequency is selected with a capacitively-operated and low noise frequency from 0.1 to 10kHz.

Compared to relatively small change in ΔC_{VEGF} (0.4% per 1pM), the sensor exhibits a large baseline capacitor C_b (tens of nF), the value of which also varies. To address this issue, the CDC digitizes C_{VEGF} in two steps (Fig. 11.3.2). First, a coarse slope ADC measures the output signal V_{out}, and increases the number of input currents I_0 in a 5b current stimulator. When V_{out} reaches the reference voltage V_{ref}, the counter output D_C of the slope ADC is stored and the current value nI_0 is set. V_{out} is then precisely measured by a fine readout, which is implemented with a low-pass filter (LPF) and a CTΔΣM. The current stimulator is implemented with a DC current source and chopper switches, and the driving frequency is determined by the chopping frequency f_c. The modulated current nI_{0pulse} is applied to the sensor and a triangular signal appears at the sensor output. After demodulation chopper, an output signal V_{out} with an amplitude of $n \times I_0/C_{VEGF} \times f_c$ is obtained and the DC component of the V_{out} represents the sensor's capacitance. Each current source (I_0 =1μA) is implemented using a 3-stack regulated cascode structure with a 2.5V supply to achieve high PSRR (>75dB) and high output impedance (200MΩ).

The V_{out} contains a large amount of $2f_c$ harmonics, thus requiring an adequate filtering to avoid aliasing. In contrast to conventional approaches that require high-order narrow-BW LPFs, the 1st-order LPF can be used due to the intrinsic anti-aliasing of the following CTΔΣM. The sensor output is directly connected to an input resistor R_I (72kΩ) of the LPF, as shown in Fig. 11.3.3(top). Although the constant current from R_I reduces V_{out} amplitude, this can be compensated with the help of adaptive stimulator. To deal with the finite gain error and noise of the opamp, the LPF employs a negative-R (NR) at virtual ground [3]. After applying NR, the current loss due to the finite response of the opamp is compensated with the NR current. The opamp noise, both thermal and 1/*f*, is also attenuated by the NR. This is because the NR shapes the opamp's input-referred noise to have a high-pass filter response and attenuates the noise at baseband. Since the 1/*f* noise of the opamp are mitigated by the NR, the NR becomes the main source of 1/*f* noise. This issue is solved by shifting the chopper demodulator from the sensor output to the LPF's virtual ground, because the 1/*f* noise of the NR is then modulated to f_c. Figure 11.3.3(bottom) shows a simplified schematic of the CTΔΣM, implemented using a 1b 3rd-order CIFF structure with resistive DACs (RDACs). The CTΔΣM has a sampling frequency f_s of 5.12MHz and a direct path from V_{LPF} to the quantizer is omitted to improve anti-aliasing property. The loop filter consists of three active-RC integrators with NR assistants, which relax the requirements of opamps, such as DC-gain, GBW and slew-rate. A resistive adder is realized at the quantizer input and an excess-loop-delay is also compensated by a direct feedback path to the quantizer input.

The CDC is realized in 65nm CMOS process (Fig. 11.3.7), occupying 0.87mm². The supply voltage is 2.5V for the current stimulator and 1.2V for the rest. It consumes 270.1μW, 159.7μW of which is dissipated in the stimulator. Figure 11.3.4(top) shows the output PSD of the CDC for a sensor capacitor of 8.2nF with f_c = 250Hz. The Q-noise is shaped properly regardless of f_c harmonics. Since the f_c harmonics are well placed in the out-of-band, and they can be removed in a digital backend. It achieves a capacitance resolution (C_{res}) of 0.49pF rms in a ΔΣ conversion time of 120ms (13.7b resolution). Figure 11.3.4(bottom) shows the measured C_{res} over a wide capacitance range from 1 to 100nF. Due to the proposed techniques, the DR is enhanced by 32× and C_{res} is maintained below 0.02%.

Figure 11.3.5 shows *in vitro* measurements for the complete sensor system (receptor + microneedle + IC) with human blood serum or phosphate-buffered saline (PBS). A capacitance variation is measured by increasing VEGF concentration in the range of 0.1pM to 1nM. Two sensors (with or without receptors) are measured to confirm the affinity of the peptides. Without receptors, the sensor achieves a small change in capacitance (4.2% with 1nM VEGF). This improves to 91.4% with the application of receptors. The sensor's resolution is measured to be <15fM rms over all detection range. As shown in Fig. 11.3.5(right), the VEGF selectivity is also confirmed by comparison with other proteins (IgG, Con A and CTx) at the same injection level.

Due to unique features of this work, it is hard to make fair comparison with other biosensors for VEGF detection [2]. Most of them focus only on electrochemical sensors, i.e., without readout ICs. Since the system relies on EIS principle for biomarkers, it is compared to current state-of-the-art works on bio-impedance or physical sensor interfaces (Fig. 11.3.6). Compared with bio-impedance interfaces [4,5], this work achieves higher resolution and DR with a competitive energy-efficiency FoM_{DR} (0.1nJ/step). This work achieves even higher DR than that of a recent work on physical sensor interface [6], but with significantly better energy-efficiency and 17× less area.

Acknowledgements:
This work was supported by a Nano.Material Technology Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2014M3A7B4051594) and the IDEC, Korea.

References:
[1] W. J. Fairbrother, et al., "Novel Peptides Selected to Bind Vascular Endothelial Growth Factor Target the Receptor-Binding Site," *Biochemistry* 37, pp.17754–17764, Dec. 1998.
[2] S. Dehghani, et al., "Aptamer-Based Biosensors and Nanosensors for the Detection of Vascular Endothelial Growth Factor (VEGF): A Review," *Biosensors and Bioelectronics*, vol. 110, pp. 23-37, July 2018.
[3] M. Jang, et al., "A 55μW 93.1dB DR 20kHz/BW Single-Bit CT ΔΣ Modulator with Negative R Assisted Integrator Achieving 178.8dB FoM in 65nm CMOS," *IEEE Symp. VLSI Circuits*, pp. C40–C41, June 2017.
[4] N. Helleputte, et al., "A 345μW Multi-Sensor Biomedical SoC with Bio-Impedance, 3-Channel ECG, Motion Artifact Reduction, and Integrated DSP," *IEEE JSSC*, vol. 50, no. 1, pp. 230-244, Jan. 2015.
[5] H. Ha, et al., "A Bio-Impedance Readout IC with Frequency Sweeping from 1k-to-1MHz for Electrical Impedance Tomography," *IEEE Symp. VLSI Circuits*, pp. C174–C175, June 2017.
[6] G. Qu, et al., "A 0.28mΩ-Sensitivity 105dB-Dynamic-Range Electrochemical Impedance Spectroscopy SoC for Electrochemical Gas Detection", *ISSCC Dig. Tech. Papers*, pp. 286-287, Feb. 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

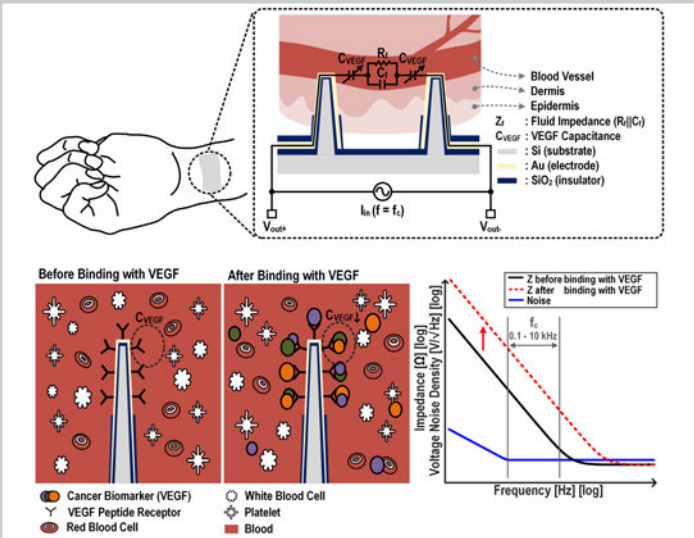


Figure 11.3.1: Sensing the cancer biomarker (VEGF) using the peptide-functionalized microneedle.

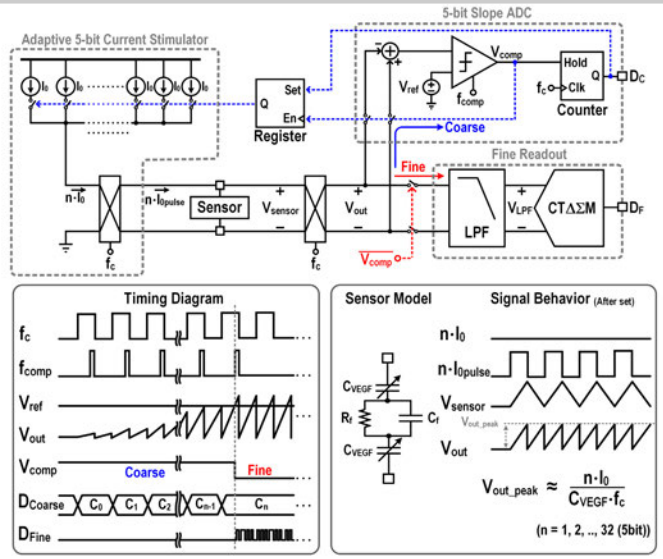


Figure 11.3.2: Block diagram and operation of the sensor readout.

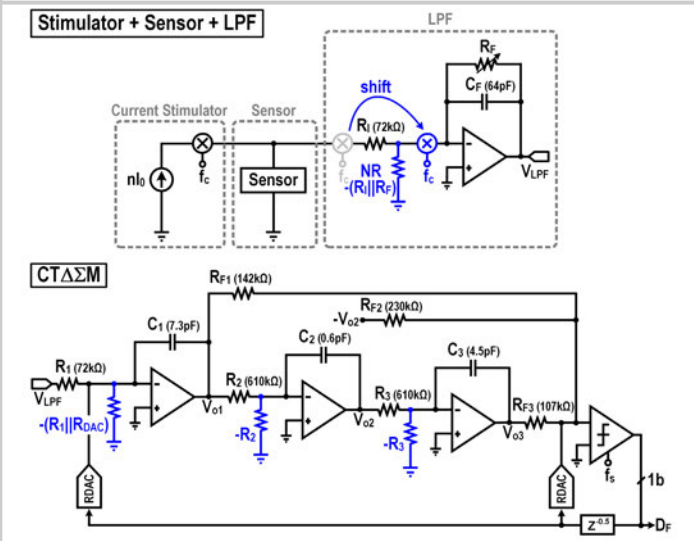


Figure 11.3.3: Simplified (single-ended) circuit diagram of the LPF (top) and CTDSM (bottom).

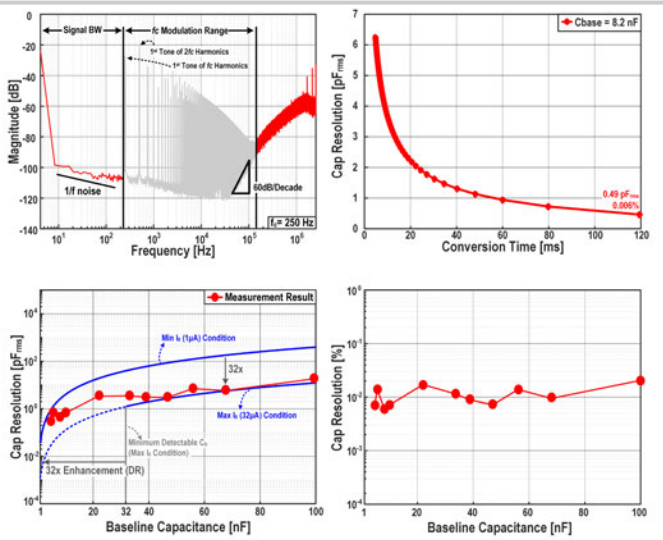


Figure 11.3.4: Measured results: PSD, resolution vs. conversion time, and capacitance resolution under various baseline capacitance.

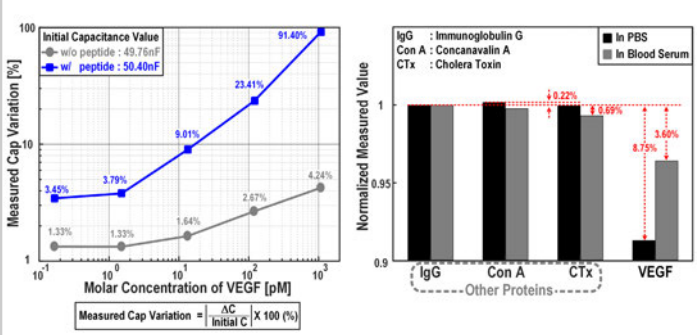


Figure 11.3.5: *In vitro* measurement results: (left) Measured capacitance variation (%) with respect to VEGF concentration and (right) comparison results with other proteins (IgG, Con A, and CTx) at the same level of injection (0.17nM).

Features		This work	JSSC15 Helleputte [4]	VLSI17 Ha [5]	ISSCC18 Qu [6]
Application		**Cancer Diagnosis**	Bio Impedance	Bio Impedance	Gas Detection
Sensing Target		**C**	Z	Z	R
Architecture	Stimulator	**Adaptive IDAC (Pulse)**	IDAC (Pulse)	IDAC (Pulse)	VDAC (Sine)
	Readout	**LPF + CTΔΣM**	IA + PGA + LPF + DTΔΣM	IA + LPF + SAR	PGA + LPF + SAR
Technology		**65nm**	180nm	180nm	180nm
Area [mm²]		**0.87**	-	2.52	15.12
[1]DR [bit]		**18.3**	12.5	11.0	17.1
[2]Resolution [bit]		**13.7**	12.5	11.0	17.1
Conversion time [ms]		**120**	6.7	9	100
Power [μW]	Stimulator	**159.7**	32.4	63.0	20600
	Readout	**110.4**	25.6	18.7	
Supply Voltage [V]		**1.2 / 2.5**	1.2	1.2 / 1.8	3
[3]FoM_{DR} [nJ/Step]		**0.10**	'0.05	'0.22	14.66
[4]$FoM_{Resolution}$ [nJ/Step]		**2.44**	'0.05	'0.22	14.66

$^{1}\text{Resolution} = \log_2\left(\frac{\text{Baseline impedance}}{\text{Minimum detectable impedance}}\right)$

$^{2}\text{DR} = \log_2\left(\frac{\text{Maximum detectable baseline impedance}}{\text{Minimum detectable impedance}}\right)$

$^{3}FoM_{Resolution} = \frac{\text{Power x Conv.t}}{2^{\text{resolution}}}$

$^{4}FoM_{DR} = \frac{\text{Power x Conv.t}}{2^{DR}}$

(Conv.t = Conversion Time)

* Calculated value: In case of impedance measurements, the power consumption of the single (for real) readout is included.

Figure 11.3.6: Performance summary and comparison with other state-of-the-art works.

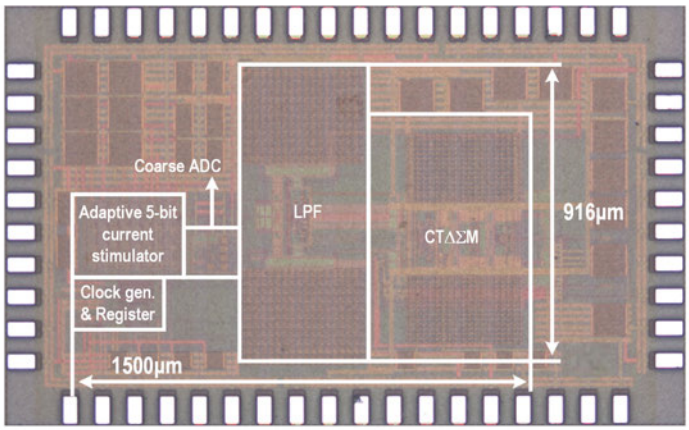


Figure 11.3.7: Chip micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

11.4 A Fast-Readout Mismatch-Insensitive Magnetoresistive Biosensor Front-End Achieving Sub-ppm Sensitivity

Xiahan Zhou, Michael Sveiven, Drew A. Hall

University of California, San Diego, La Jolla, CA

Magnetic biosensors have drawn significant research interest as a means to replace large optical instrumentation commonly found in centralized diagnostic laboratories. Magnetic biosensing is attractive since biological samples lack a magnetic background thus enabling high sensitivity measurements. Figure 11.4.1 shows a magnetic immunoassay where: 1) the surface of a magnetic sensor is coated with capture molecules (e.g., antibodies) specific to the biomarker (e.g., antigen) of interest, 2) the sample, detection molecules, and magnetic nanoparticles (MNPs) are added and self-assemble thus tethering the MNPs to the sensor, and 3) the sensor is read out by applying a magnetic field, H_A, that magnetizes the superparamagnetic MNPs producing a stray field that perturbs the local magnetic field. The underlying magnetoresistive (MR) sensor detects this change in field as a miniscule ppm change in resistance, R_{sig}, that is directly proportional to the number of MNPs. Unfortunately, R_{sig} is superimposed on the baseline resistance, R_0, resulting in a large baseline-to-signal ratio (R_0/R_{sig}>10,000) that requires a very high dynamic range (DR>120dB) front-end for quantitative detection [1,2].

Traditional techniques such as a Wheatstone bridge are limited by the sensor-to-sensor matching, which is often no better than 5 to 10%, even on the same die [3]. Other techniques such as magnetorelaxometry solve the baseline issue by self-referencing using a pulsed magnetic field but require high-speed digitization to capture the fast (<1μs), transient relaxation signal and need averaging to deal with the noise [2]. Magnetic double-modulation is another technique that separates the signals spectrally and reduces the 1/*f* noise from the sensors and AFE; but most implementations perform the down modulation and lock-in digitally thus still requiring high DR ADCs and only improve the baseline-to-signal ratio by an order of magnitude [3-5]. This paper presents a CMOS front-end for MR biosensors that can tolerate up to 10% sensor mismatch and achieves sub-ppm sensitivity with an input-referred noise PSD of 46.4nT/√Hz, input-referred baseline of <0.235mT, power consumption of 1.39mW, and readout time of 11ms owing to: 1) a double modulation bias scheme with reference sensors that reduces baseline/signal to <20 and a built-in down-modulator that relaxes the ADC bandwidth, 2) fast settling duty-cycled resistors (DCRs) that reduce the settling time by 40× compared to traditional DCRs, and 3) a high-frequency interference-rejection (HFIR) sampling technique that reduces out-of-band interference and thus the required ADC DR.

The chip architecture is shown in Fig. 11.4.2 where a sinusoidal bias current at frequency f_C is injected into a giant magnetoresistive (GMR) sensor array placed inside of an external Helmholtz coil that generates a sinusoidal magnetic field at frequency f_H. Output spectra shows that the signal of interest is modulated to $f_C \pm f_H$ while the R_0 baseline remains at f_C. A mux selects an active and a reference sensor that are fed to the programmable gain amplifier (PGA) which rejects the common-mode signal thus eliminating most of the baseline (>90%). The PGA also down-modulates the signal to DC relaxing the ADC bandwidth. However, the sensor mismatch (in both MR and R_0) causes baselines at DC and f_H that are still up to 20× and 800× larger than the signal of interest, respectively. Left unaddressed, these would still require a high DR ADC; instead, we propose a sampling scheme that reduces the interference before it is seen by the ADC.

The PGA is implemented by a capacitive feedback amplifier, where the gain C_1/C_2 is set by a 2b programmable capacitor (Fig. 11.4.3). Inputs V_S and V_{ref} contain large common-mode swing and the outputs also have large swing mainly due to the interference at f_H. The DC operating points were chosen to maximize both swings. Since the output swing is highly dependent on the sensor mismatch, the PGA can be programmed for high gain in low mismatch scenarios to improve the SNR. The OTA uses switched-capacitor common-mode feedback (SC-CMFB) to maximize the output swing and integrates choppers at the cascode node for down-modulation. This amplifier is designed for fast settling needed when multiplexing sensors. The sudden closing of the row select switches (Sel: 0→1) causes a large transient at V_{in}, which takes a long time to settle when using a traditional DCR or a pseudo-resistor for biasing since the bias resistors must be large for low noise. To address this, the switches controlled by φ_{dcr} remain closed during and after sensor selection for 0.1ms, thus providing a low impedance path from V_{in} to V_{cmi} enabling fast settling. The fast settling DCRs reduce the settling time by 40×.

The ADC is implemented by a 1st-order incremental ΔΣ modulator (Fig. 11.4.4) with an oversampling ratio of 10,000. As shown in Fig. 11.4.2, the interference at the ADC input can be as large as 200mV, while the MR baseline and signal at DC are <10mV. While out-of-band, this interference eats into the DR and cannot be easily low-pass filtered without enduring a prohibitively long settling time. Instead, we propose a feed-forward high-pass filtering technique in the sampling network. A high-pass filtered version of the signal is first sampled onto the bottom plate of C_4. Subsequently, the signal is sampled onto the top plate of C_4, thus cancelling the high frequency interference. The HFIR is implemented by a pseudo-differential switched-capacitor high-pass filter followed by a chopped buffer to reduce offset and 1/*f* noise. A modified correlated double sampling (CDS) technique is applied to remove the OTA offset and 1/*f* noise. Specifically, switches $S_{1,2}$ were added to isolate C_6 from V_{ac} so that C_6 only samples the offset from the OTA during the sampling phase. A 16b counter and a shift register decimate the bitstream and serialize the output, respectively.

The chip in 180nm CMOS consumes 1.39mW excluding the sensor bias (ADC: 69%, PGA: 19%, digital blocks and bias: 12%). Figure 11.4.5 shows electrical measurement results where the PGA settling time is 50μs with the proposed DCR compared to a traditional DCR and pseudo-resistor which take 2ms and >14ms, respectively. The ADC DR for various sensor mismatches is shown where this design achieves significant improvements (>74dB for >2% mismatch) due to the HFIR technique that otherwise saturates the ADC. The ADC has an ENOB of 12.1b in the 100Hz bandwidth and provides 16dB interference suppression at f_H=10kHz. The fast readout time can be traded off for sensitivity, where this design achieves 0.98ppm sensitivity (147μΩ), which corresponds to ~2200 MNPs (d=50nm) with a SNR of 6dB.

The chip was connected to GMR sensor arrays to demonstrate the magnetic immunoassay. One sensor array was functionalized with biotinylated-bovine serum albumin (Biotin-BSA) and the other human Interleukin 6 (IL-6), a cancer biomarker. Both chips had BSA negative controls to monitor nonspecific binding and the reference sensor was covered by epoxy. Figure 11.4.6 shows measured real-time binding curves clearly differentiating the active and control sensors. The chip performance is summarized and compared to state-of-the-art solutions. Compared with other magnetometry-based designs, this work achieves 22.7× faster readout time, >7.8× lower baseline, and 2.3× lower power (see Fig. 11.4.6). Although relaxometry-based designs inherently eliminate the baseline, their lower signal amplitude and higher input-referred noise preclude high-sensitivity measurements [2]. This chip can tolerate up to 10% sensor mismatch, making it compatible with commercial tolerances for MR sensor fabrication. The sub-ppm sensitivity, low power consumption, and fast readout make this work ideal for multichannel, high sensitivity, point-of-care diagnostics. Figure 11.4.7 shows a die photo.

Acknowledgements:
This work was supported in part by Qualcomm, Inc. and the National Science Foundation (grant ECCS-1454608). The authors would like to thank MagArray, Inc. for providing the GMR sensors.

References:
[1] H. Wang, et al., "A Frequency-Shift CMOS Magnetic Biosensor Array with Single-Bead Sensitivity and no External Magnet," *ISSCC Dig. Tech. Papers,* pp. 438-439, Feb. 2009.
[2] S. Gambini, et al., "A 10 kPixel CMOS Hall Sensor Array with Baseline Suppression and Parallel Readout for Immunoassays," *IEEE JSSC,* vol. 48, no. 1, pp. 302-317, 2013.
[3] T. Costa, et al., "A CMOS Front-End with Integrated Magnetoresistive Sensors for Biomolecular Recognition Detection Applications," *IEEE TBioCAS,* vol. 11, no. 5, pp. 988-1000, 2017.
[4] S.J. Han, et al., "A High-Density Magnetoresistive Biosensor Array with Drift-Compensation Mechanism," *ISSCC Dig. Tech. Papers,* pp. 168-169,594, Feb. 2007.
[5] D.A. Hall, et al., "A 256 Pixel Magnetoresistive Biosensor Microarray in 0.18 um CMOS," *IEEE JSSC,* vol. 48, no. 5, pp. 1290-1301, 2013.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

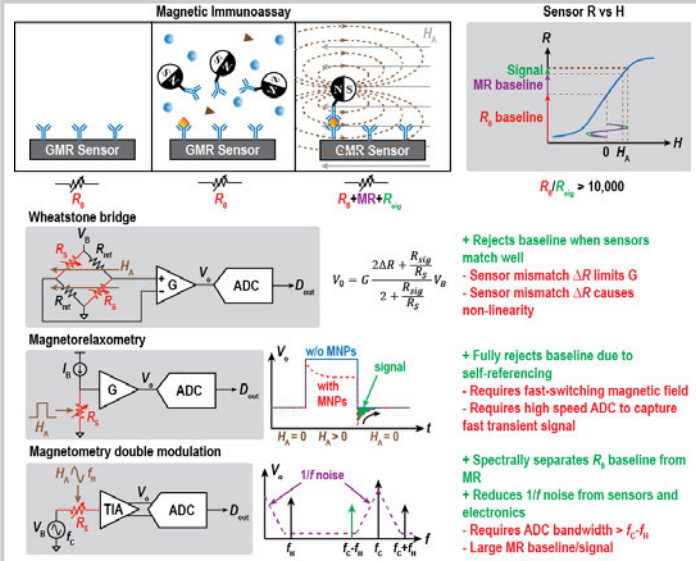


Figure 11.4.1: Magnetic immunoassay and prior readout architectures.

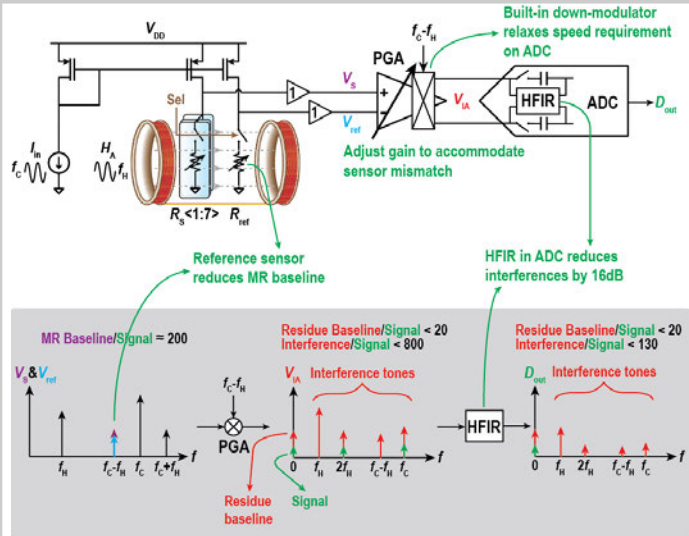


Figure 11.4.2: Architecture of the analog front-end and baseline/interference rejection diagram.

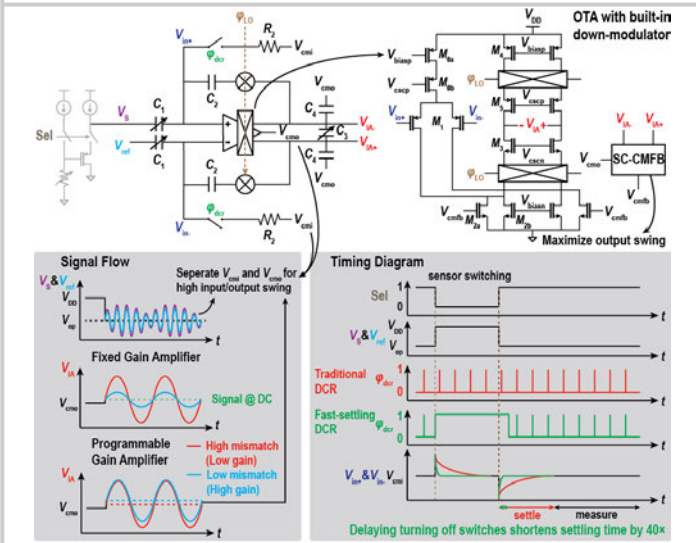


Figure 11.4.3: Schematic of the programmable gain amplifier and timing diagram.

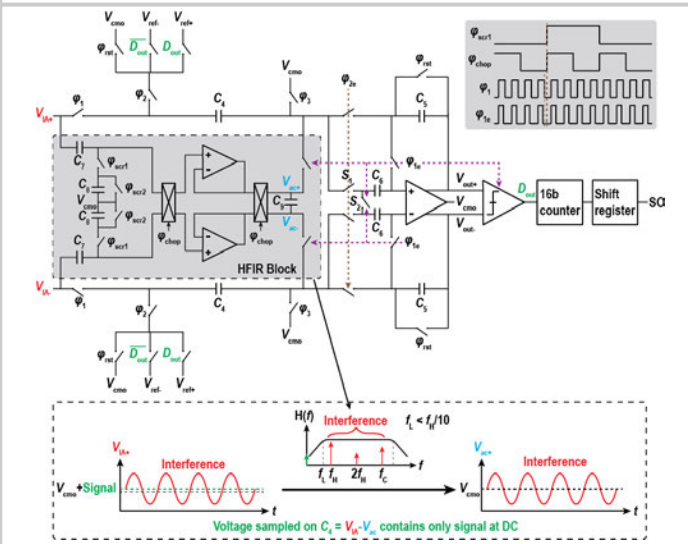


Figure 11.4.4: Schematic of the incremental ΔΣ ADC with high-frequency interference-rejection sampling.

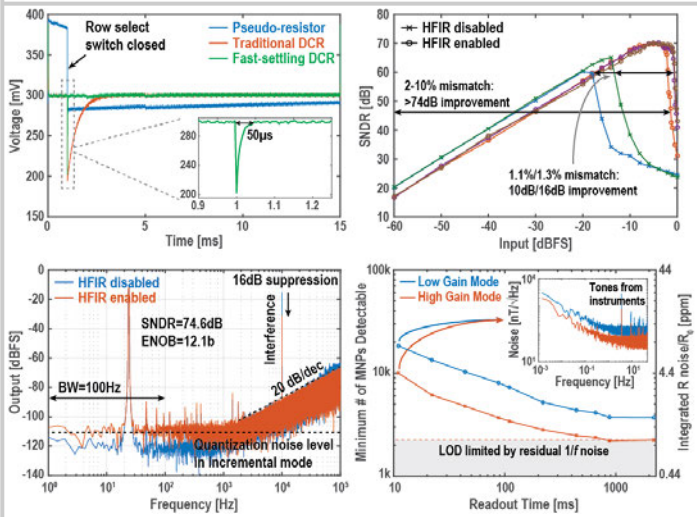
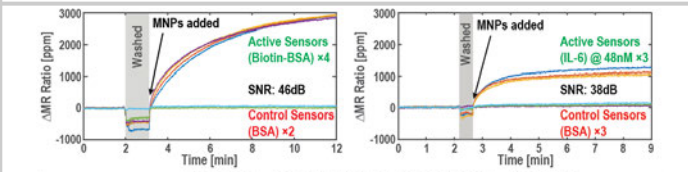


Figure 11.4.5: Measured PGA settling transient (top left), ADC dynamic range (top right), ADC output spectrum (bottom left), and system sensitivity (bottom right).



	ISSCC 2009 [1]	JSSC 2013 [2]	TBioCAS 2017 [3]	ISSCC 2007 [4]	JSSC 2013 [5]	This Work
Sensor Type	LC	Hall	GMR	GMR	GMR	GMR
Sensor Resistance (kΩ)	N/A	N/A	0.85	N/A	1.92	0.15
MR Ratio (%)	N/A	N/A	5.37	N/A	9.2	9.04
MNP Size (nm)	1,000	1,000	250	50	50	50
Technology (µm)	0.13	0.18	0.35	0.25	0.18	0.18
AFE Architecture	LC oscillator	V/I converter	Amplifier	Mixer + PGA	TIA	PGA with Mixer
ADC Architecture	VCO-based	Inc. ΔΣ	No ADC	No ADC	ΔΣ	Inc. ΔΣ + HFIR
Input-referred Noise (nT/√Hz)	N/A	270	11.5ψ	N/A	49	107.1 (low gain) 46.4 (high gain)
Sensor Mismatch Tolerance	N/A	N/A	N/A	N/A	N/A	10% (low gain) 2.5% (high gain)
Power/Ch (mW)*	N/A	0.825	4.9ψ	N/A	3.15	1.39
Area/Ch (mm²)	N/A	0.012	3.17	N/A	0.219	0.249
Input-referred Baseline (mT)	N/A	0.007	1.84	N/A	7.09	< 0.235
Readout Time/Ch (ms)	400‡	50	1,000	250	250	11

* Does not include sensor bias and magnetic field generator. ψ Does not include power and noise from ADC.
‡ Read from figures.

Figure 11.4.6: Magnetic assay real-time measurements (top) and comparison table (bottom).

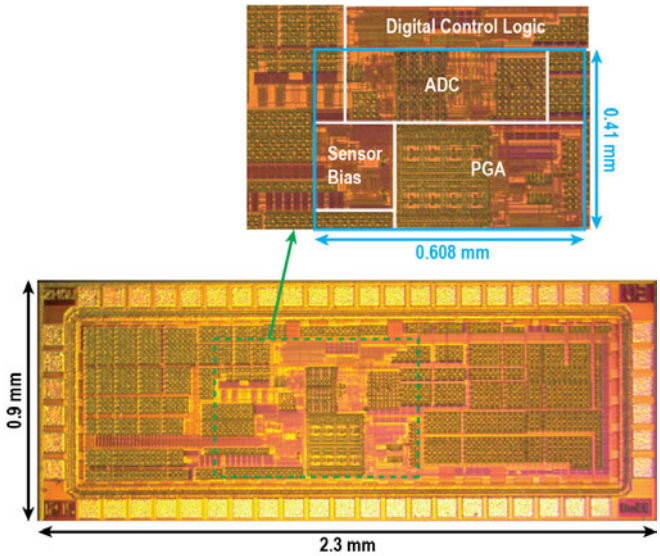


Figure 11.4.7: Die photo.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

11.5 A 512-Pixel 3kHz-Frame-Rate Dual-Shank Lensless Filterless Single-Photon-Avalanche-Diode CMOS Neural Imaging Probe

Changhyuk Lee[1,2], Adriaan J. Taal[1], Jaebin Choi[1], Kukjoo Kim[1], Kevin Tien[1], Laurent Moreaux[3], Michael L. Roukes[3], Kenneth L. Shepard[1]

[1]Columbia University, New York, NY
[2]KIST, Seoul, Korea
[3]California Institute of Technology, Pasadena

Optical functional neural imaging has revolutionized neuroscience with optical reporters that enable single-cell-resolved monitoring of neuronal activity *in vivo*. State-of-the-art microscopy methods, however, are fundamentally limited in imaging depth by absorption and scattering in tissue even with the use of the most sophisticated two-photon microscopy techniques [1]. To overcome this imaging depth problem, we develop a lens-less, optical-filter-less, shank-based image sensor array that can be inserted into the brain, allowing cellular-resolution recording at arbitrary depths with excitation provided by an external laser light source (Fig. 11.5.1). Lens-less imaging is achieved generally by giving each pixel a spatial sensitivity function, which can be introduced by near-field or far-field, phase or amplitude masking. Since probe thickness must be less than 70μm to limit tissue damage and far-field masks are characterized by distances on the order of 200μm between the mask and the detector [2], we employ a near-field amplitude mask formed by Talbot gratings in the back-end metal of the CMOS process, which gives each pixel a diffraction-grating-induced angle-sensitivity [3]. Filter-less fluorescence imaging is achieved with time-gated operation in which the excitation light source is pulsed and pixel-level time-gated circuitry collects photons only after the excitation source has been removed.

The implantable neural imager takes the form of two shanks, each 3.2mm long and containing 256 single-photon-avalanche-diode (SPAD) pixels in two rows at 25μm pitch and 6.3% fill factor (Fig. 11.5.1). The shanks are thinned to 70μm in a substrate back-etch process. Each shank defines an imaging field of view of approximately 3.2mm×200μm×200μm, limited by scattering in tissue. An off-chip FPGA controls the time-gating and data acquisition.

An externally supplied SPAD bias high voltage (V_{CA}) and an active-reset bias voltage (V_{RST}) set the SPADs in Geiger mode with an excess bias voltage determined by $V_{OD}=V_{CA}-V_{BD}$, where V_{BD} is the breakdown voltage (Fig. 11.5.2). A variable SPAD dead time, determined by a digital delay (3ns to 1μs) controlled by bias current $I_{hold\text{-}on}$, is used to limit afterpulsing probability, which is reduced to 1.5% at 10ns. A detection circuit, that ignores rising edges of the diode anode voltage associated with diode gated turn-off, is used to trigger a pixel-level, 5b photon-event counter. The quenching and active-reset circuits are capable of supplying excess bias voltages (V_{OD}) between 0 and 5 V, enabling the dynamic adjustment of photon detection probability (PDP) and dark-count rate (DCR). Functioning SPAD yield is consistently better than 95% across all tested shanks. A global time-gating shutter scheme ensures simultaneous recording from all SPADs, ensuring integrity of time-correlated photon detection. The ON signal is level shifted to V_{RST} in order to toggle the SPAD between V_{CA} and $V_{CA}-V_{RST}$, allowing the SPAD to be turned on and off synchronously to an external fast-pulsed excitation laser source.

We employ a single supply connection along the entire shank and mitigate potential supply noise issues with 250fF of local decoupling capacitance per pixel. The imaging probe, dominated by the charging of SPAD diode capacitance at quenching events, consumes a total of 97μW under dark conditions and 68mW in full output saturation lighting. Digital circuitry and I/O require approximately 4mW during readout. In addition, each individual pixel can be programmed on or off, lowering operating power while recording over subsets of pixels. The SPAD array displays excellent output linearity over almost three orders of magnitude of input light intensity (Fig. 11.5.3b). A V_{OD} of 1V, which provides the best trade-off between DCR and PDP, is employed consistently in our measurements.

To verify time-gated fluorescence imaging, we image fluorophores (Alexa Fluor 488, 500-to-540nm emission, 4ns lifetime) on beads using a mode-locked Ti-Sapphire pulsed laser, frequency doubled to 480nm as an excitation source. Conventional two-photon microscopy provides ground truth data on bead locations. By synchronizing an 8b digital-to-time converter (DTC) to the 80MHz laser pulse, time-gating in the SPADs achieves a temporal resolution of 350ps (Fig. 11.5.3c). A variable duty cycle of 10 to 50% allows the SPAD to be moved out of Geiger mode after sufficient decay of the fluorescent signal, avoiding dark count and unnecessary power dissipation. Emitted fluorescent signal intensities from a 15μm microsphere (Thermo Fisher F8844), corresponding to soma dimensions, are measured to be on the order of 1nW. For an isotropic radiating source located 200μm away from the 7.7μm-diameter SPAD, an SNR of 40dB dictates a maximum frame rate of approximately 3kfps. Higher frame-rates, ultimately bounded by the laser repetition rate, are possible under high signal-level conditions. At circuit-level, framerate is bounded by readout overhead requiring two FPGA 100MHz clock cycles per pixel.

While much of the novelty of this work lies in the form factor and application of this imager, this design compares favorably with other recent SPAD imagers with conventional planar array formats [3-5] (Fig. 11.5.4) which displaces a volume of only 0.012mm^3 on the implanted shanks.

Thirty-two SPAD ensembles of 16 pixels (Fig. 11.5.5a), each consisting of orthogonal combinations of angular modulation frequencies (β=12, 20), direction (x, y), and quadrature phase (=0°, 90°, 180°, 270°), are used to maximize spatial diversity for computational image reconstruction [3]. The angular modulative linear mapping is measured in two-dimensional angular (θ,ϕ) coordinates and subsequently parametrically fit for use in image reconstruction (Fig. 11.5.5b). The resulting angular amplitude response function is affected by three factors. In addition to the angular modulation due to the near-field diffraction effect of the gratings, windowing effect rejects photons for incident angles larger than 60 degrees from normal. Finally, the PDP at a given wavelength is lower for obliquely incident photons due to a deeper effective multiplication layer. Overall transmission efficiency for the SPADs with amplitude gratings is 16% over all incident angles (Fig. 11.5.5c).

Spatial resolution assessment was performed by positioning two micropipettes over the shank, simulating neural point sources with their extreme tips filled with Alexa Fluor 488 (Fig. 11.5.6). A third pipette is completely filled and overlaid perpendicular to the shank close to the base, simulating a neural structure. Inverse imaging is performed using a Tikhonov regularization, enforcing non-negativity and smoothness with total-variance denoising. The image is further sharpened by L1-minimization through soft-thresholding.

Resolution in this experiment ($\Delta x \approx 45$μm, $\Delta y \approx \Delta z \approx 60$μm) is extracted by the 34% (or 1) radius of estimation confidence ellipsoid. The ellipsoid radius in each direction is estimated by computing the spatial covariance from a larger volume cutout. The resolution is limited by integrated dark count rate, integrated photon shot noise at high excitation intensities, spectral spread of the fluorophore which challenges the monochromatic assumption in the pixel angular sensitivity, and pixel-to-pixel deviation of angular response from the median fitted response. The denser longitudinal (x-axis) SPAD population yields a better resolution along that axis.

Future functional imaging will enable blind-source separation techniques to be used to directly infer neural sources taking advantage of their temporal properties without the need to perform this inverse imaging, better exploiting the high temporal resolution of the probe imagers over traditional multiphoton laser-scanning microscopy.

Acknowledgments:
This work was supported by the National Institutes of Health under Grant U01NS090596, by the Defense Advanced Research Projects Agency (DARPA) under Contract N66001-17-C-4012, and by the U. S. Army Research Laboratory and the U. S. Army Research Office under Contract W911NF-12-1-0594. We would also like to thank TSMC Foundry for their full support in testchip fabrication.

References:
[1] H. Wang, et al., "LOVTRAP, An Optogenetic System for Photo-induced Protein Dissociation", *Nature Methods*, vol. 13, no. 9, p. 755–758, Sept. 2016.
[2] J.K. Adams, et al., "Single-Frame 3D Fluorescence Microscopy with Ultraminiature Lensless FlatScope", *Science Advances*, vol. 3, no. 12, Dec. 2017
[3] C. Lee, et al., "A 72 × 60 Angle-Sensitive SPAD Imaging Array for Lens-less FLIM," *Sensors*, vol. 16, no. 9, p. 1422, Feb. 2016.
[4] M. Perenzoni, et al., "A 160 × 120 Pixel Analog-Counting Single-Photon Imager With Time-Gating and Self-Referenced Column-Parallel A/D Conversion for Fluorescence Lifetime Imaging" *IEEE JSSC*, vol. 51, no. 1, p.155-167, Jan. 2016.
[5] R.M. Field, et al., "A 100 fps, Time-Correlated Single-Photon-Counting-Based Fluorescence-Lifetime Imager in 130 nm CMOS," *IEEE JSSC*, vol. 49, no. 4, p. 867-880, Apr. 2014.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

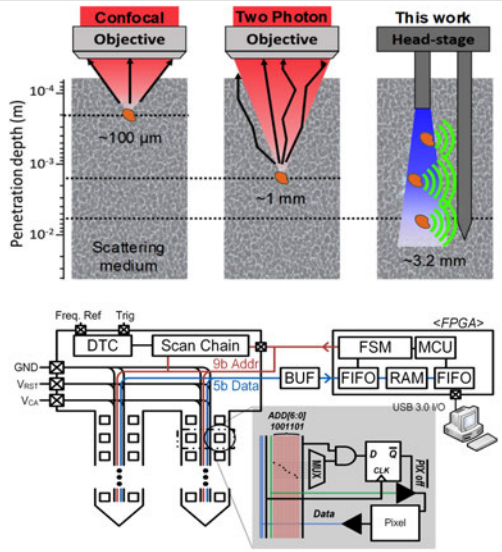


Figure 11.5.1: Concept diagram and system architecture.

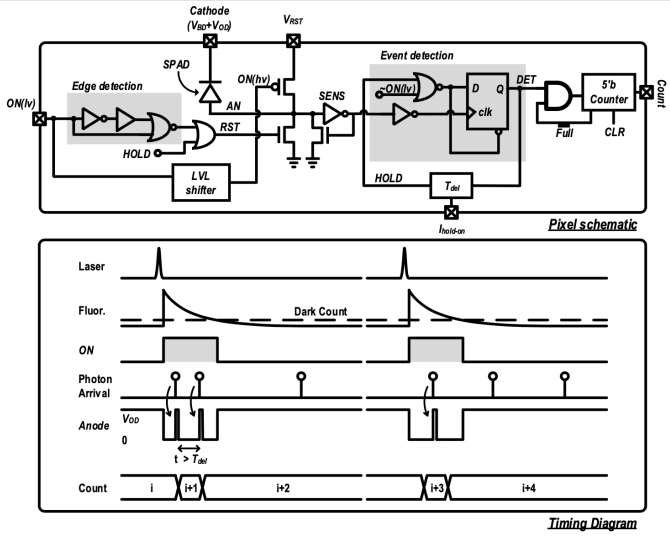


Figure 11.5.2: Schematic of the in-pixel active quenching circuit and timing diagram in the case of a detection event.

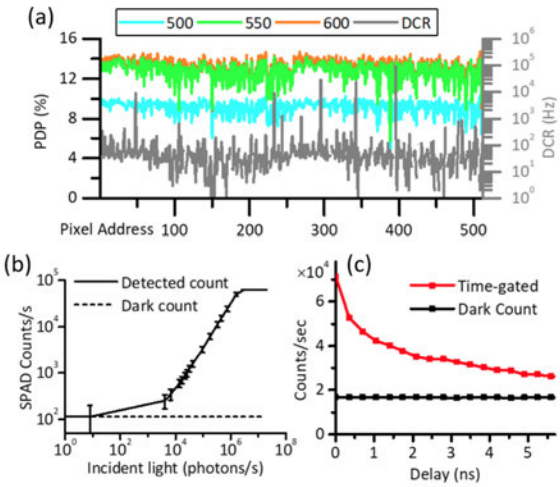


Figure 11.5.3: SPAD performance: (a) array-wide PDP and DCR variability at 1V V_{OD} over fluorescence emission wavelengths (b) SPAD linearity over almost three order of magnitude in incident light intensity, and (c) time-gated measurement of 4ns fluorescent lifetime.

	[3]	[4]	[5]	THIS WORK
PIXEL PITCH	35um	50um	48um	25.3um
PIXEL COUNT	72x60	160 x 120	64x64	4x128
POWER CONS.	83.8mW	157mW	26W	72mW
TEMPORAL RESOLUTION	71ps	194ps	62.5ps	350ps
DCR MEDIAN	400Hz @1.2V	580Hz @3V	544Hz@1.5V	40Hz @1V
PDP	2.7% @1.2V	-	30% @1.5V	10% @1V
FILL FACTOR	9.6%	21%	0.77%	6.3%
FRAMERATE	-	486fps	100fps	3Kfps

Figure 11.5.4: Comparison with recent work. PDP and DCR are reported at excess bias voltage.

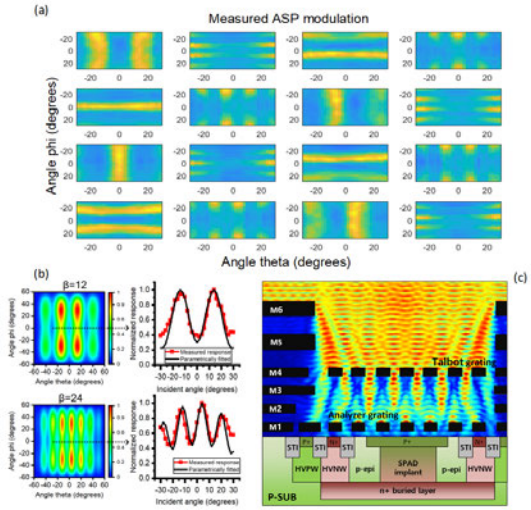


Figure 11.5.5: (a) Measured angle-sensitivity of A-SPAD in angular coordinates of two selected pixels. (b) fitted sensitivity and comparison with measurement (c) illustration of angle sensitive SPAD.

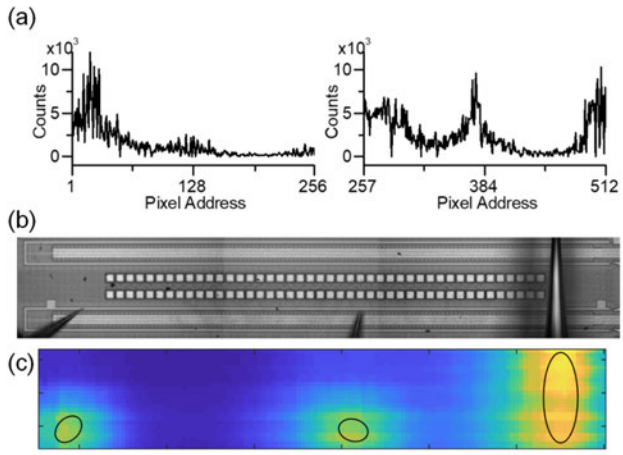


Figure 11.5.6: (a) Recorded data from micropipette experiment (b) Stitched photomicrograph of die with three positioned pipettes. The base of the shank is on the right, the center pads are use of real estate unrelated to probe. Higher addresses correspond to bottom shank in photo. (c) volumetric reconstruction above chip surface at z=200µm with their respective error covariance ellipses.

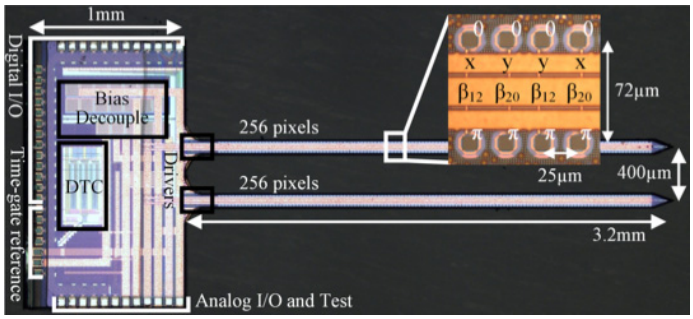


Figure 11.5.7: Photomicrograph after etching and thinning of die.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 12 Overview: *Emerging Technologies*

TECHNOLOGY DIRECTIONS SUBCOMMITTEE

Session Chair: ***Wim Dehaene***
Kuleuven-MICAS, Leuven, Belgium

Associate Chair: ***Sriram Vangal***
Intel, Hillsboro, OR

Subcommittee Chair: Makoto Nagata, Kobe University, Kobe, Japan

This session highlights advances in non-volatile memory integration, battery protection circuits, and flexible thin-film transistor-based memories. Paper 12.1 demonstrates the first-known integrated non-volatile FPGA, CPU and STT-MRAM with the lowest average power consumption. A novel internal short-circuit detector presented in Paper 12.2 that enables Li-ion battery protection and safety using c-axis aligned crystalline IGZO FETs. Paper 12.3 advances the state-of-the-art of TFT-IGZO memory design by increasing SRAM speed by 5.5× and its storage capacity by 4×.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

8:30 AM

12.1 An FPGA-Accelerated Fully Nonvolatile Microcontroller Unit for Sensor-Node Applications in 40nm CMOS/MTJ-Hybrid Technology Achieving 47.14μW Operation at 200MHz

M. Natsui, Tohoku University, Miyagi, Japan

In Paper 12.1, Tokohu University describes a fully nonvolatile FPGA-accelerated microcontroller. In addition to the embedded FPGA, it integrates a nonvolatile 32b ARM-Cortex-M0 CPU and a 64kB STT-MRAM fabricated using a 40nm CMOS/MTJ-hybrid process resulting in 47.14μW power consumption at 200MHz operating frequency.

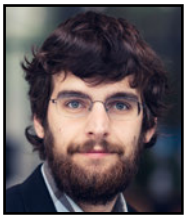
9:00 AM

12.2 Micro Short-Circuit Detector Including S/H Circuit for 1hr Retention and 52dB Comparator Composed of C-Axis Aligned Crystalline IGZO FETs for Li-Ion Battery Protection IC

T. Ikeda, Semiconductor Energy Laboratory, Kanagawa, Japan

In Paper 12.2, Semiconductor Energy Laboratory presents a micro short-circuit detector for battery protection using n-channel c-axis aligned crystalline IGZO FETs. It has a S/H circuit capable of voltage retention for more than 1hr and an analog current load comparator with 52dB gain.

9:30 AM



12.3 Memory Solutions for Flexible Thin-Film Logic: up to 8kb, >105.9kb/s LPROM and SRAM with Integrated Timing Generation Meeting the ISO NFC Standard

F. De Roose, KU Leuven, Heverlee, Belgium; imec, Heverlee, Belgium

In Paper 12.3, imec presents a flexible IGZO SRAM and LPROM with integrated decoder and timing circuitry. The circuits presented are: a 512b SRAM with 242kb/s throughput, a 1kb LPROM operated at 140kb/s, compliant with the ISO NFC standard, and an 8kb self-timed LPROM accessed at 44kb/s.

12.1 An FPGA-Accelerated Fully Nonvolatile Microcontroller Unit for Sensor-Node Applications in 40nm CMOS/MTJ-Hybrid Technology Achieving 47.14μW Operation at 200MHz

Masanori Natsui[1], Daisuke Suzuki[1], Akira Tamakoshi[1], Toshinari Watanabe[1,2], Hiroaki Honjo[1,2], Hiroki Koike[1,2], Takashi Nasuno[1,2], Yitao Ma[1], Takaho Tanigawa[1,2], Yasuo Noguchi[1,2], Mitsuo Yasuhira[1,2], Hideo Sato[1], Shoji Ikeda[1,2], Hideo Ohno[1], Tetsuo Endoh[1,2], Takahiro Hanyu[1]

[1]Tohoku University, Miyagi, Japan
[2]JST/ACCEL, Tokyo, Japan

Recently, the demand for low-power, high-performance microcontroller units (MCUs) for power-supply-critical sensor node applications has been increasing. In response to this demand, the use of nonvolatile memory elements for realizing MCUs for sensor node applications has been actively researched and developed. The latest nonvolatile MCUs (NV-MCUs) demonstrated 32b operation at 30MHz [1] and 8b operation at 100MHz [2]. However, this performance level is not suitable for sensor node applications that process large numbers of received signals and extract valuable information from them immediately to reduce the amount of transfer data to a data center. The use of various nonvolatile devices has also been proposed. However, these devices exhibit critical drawbacks when applied to sensor node applications, including limited endurance and low compatibility with standard CMOS. A spintronics-based nonvolatile device with unlimited endurance, a short switching time, and CMOS compatibility is a promising candidate for designing a low-power, high-performance NV-MCU.

In this work, we demonstrate a fully nonvolatile MCU using perpendicular magnetic tunnel junction (MTJ) device technology. The main contributions of this work are summarized as follows. (1) A 32b NV-MCU integrating a programmable accelerator based on a nonvolatile FPGA (NV-FPGA), a nonvolatile CPU (NV-CPU), and an STT-MRAM is fabricated using a hybrid process of 40nm standard CMOS and 39nm perpendicular MTJ devices on a 300mm wafer fabrication line. (2) The power dissipated during the standby state is reduced to 0.7μW with the use of fine-grained power gating based on complete nonvolatility. (3) The use of an NV-FPGA-based programmable accelerator achieves the best energy efficiency of the intermittent operations among of reported works. (4) Average power consumption of 47.14μW at an operating frequency of 200MHz is confirmed from the measurement results of the fabricated chip, which provides both the highest processing performance and the highest energy efficiency for highly functional sensor nodes powered by harvested energy.

The architecture of this NV-MCU is shown in Fig. 12.1.1. The NV-MCU consists of a 32b NV-CPU, a 64kB STT-MRAM, an NV-FPGA, and peripheral modules including a power management unit (PMU), an ADC, timers, and I/O interfaces connected by an advanced high-performance bus (AHB). A perpendicular MTJ device is adopted with a CoFeB-MgO-CoFeB structure to realize the nonvolatile memory function whose resistive state is changed via spin-transfer torque switching. The NV-CPU, whose instruction set is fully compatible with an ARM Cortex-M0 microprocessor, has 476 MTJ-based nonvolatile flip-flops (MFFs) to maintain the intermediate data when there is no power supply, which eliminates backup/restore overhead caused by data transfer between the memory and the CPU. A memory access controller (MEM4X) accelerates the data transfer between the STT-MRAM and other modules by up to four times by incorporating both memory-access multiplexing and code-length-aware prefetching techniques, which enables to relax the performance requirement for the STT-MRAM.

Figure 12.1.2 shows the structures of both the embedded STT-MRAM and the NV-FPGA. The embedded STT-MRAM consists of 2-transistor-2-MTJ memory cells, which achieve both low power and a compact footprint (16 F^2) by applying a shared-write-bit-line (WBT) structure. This STT-MRAM supports the read/write mode of variable data length, and ECC function can be easily applied by utilizing its simple array configuration. Measurement results indicate that the STT-MRAM performs 50MHz read/write operations for consecutive memory addresses at room temperature. The embedded NV-FPGA has seven DSP blocks assuming application as a programmable signal processing accelerator as well as 1,176 six-input look-up tables (LUTs) following modern commercial FPGA devices. Because all the tiles constituting the NV-FPGA have nonvolatile functions and can each be power gated, the unnecessary power consumption of unused tiles is eliminated. In addition, the NV-FPGA can be reconfigured based on the processing task required and can be utilized as an accelerator to realize high-speed and energy-efficient processing.

Figure 12.1.3 shows both a Shmoo plot and the measured waveform of an FPGA-accelerated operation executed in the NV-CPU. The Shmoo plot shows that the NV-MCU achieves 100MHz and 200MHz operations with internal supply voltages of 1.1V and 1.3V, respectively. Data exchange between the NV-CPU and the NV-FPGA is performed via the reserved addresses on the STT-MRAM. After raising an enable signal, which represents the beginning of the data transfer from the NV-CPU to the NV-FPGA, input data is stored to the STT-MRAM via the AHB. Each time a certain amount of processing using this input data is completed by the NV-FPGA, another enable signal is sent, and the output from the NV-FPGA is written to the memory via the AHB in a similar manner.

Figure 12.1.4 shows the impact of nonvolatile power gating and FPGA-based acceleration on the relationship between the power dissipation and the intermittent operation interval. Because the NV-MCU is entirely nonvolatile, we can apply power gating immediately after the operation has completed, which reduces the static power dissipation in the standby state to 0.7μW. An average power consumption of 100μW can be achieved at an intermittent operation interval of 50.91ms, which meets the requirements for a general sensor node application powered by harvested energy. Furthermore, by replacing a subset of the sequential processing performed by the NV-CPU with parallel processing performed by the NV-FPGA, the energy efficiency can be improved even further. As a result, average power consumption of 100μW can be achieved at an intermittent operation interval of 23.38ms, which is 54% shorter than the interval demonstrated when the NV-FPGA is not used.

Figure 12.1.5 compares the energy dissipation in typical operations executed in the sensor node both with and without NV-FPGA-based acceleration. While the power consumption of the NV-FPGA increases, the power consumed by both memory read/writes and CPU operations is reduced because of the improvement in the computational efficiency. In addition, because completing the operations and shifting to a standby state takes less time, the total energy efficiency is improved further (e.g., 11.5× for an FIR filter). As a result, the energy efficiency is improved by the NV-FPGA in either case.

Figure 12.1.6 shows a comparison of our NV-MCU with other MCUs with embedded nonvolatile memories in recent studies [1-5]. Both a test chip micrograph of this NV-MCU and its specifications are shown in Fig. 12.1.7. The NV-MCU demonstrated both the highest operating frequency and the smallest active power (in μW/MHz) among those in the literature. Moreover, the NV-MCU achieved a sufficient small leakage power as the other MCUs even if it is manufactured with advanced process and incorporates the FPGA-based accelerator, because of its fine-grained nonvolatile power gating. As a result, the NV-MCU has the smallest average power consumption for applications involving the intermittent operation of sensor nodes.

Acknowledgements:
The authors thank Y. Takako of Focal Agency for excellent technical assistance. The authors also thank Prof. K.C. Smith for valuable comments to improve the quality of the paper. Part of this research was supported by ImPACT Program of CSTI, STT-MRAM R&D program under Industry-Academic collaboration of CIES consortium, and JSPS KAKENHI Grant Number 16KT0187 and 17KK0001.

References:
[1] T. Onuki, et al., "Embedded Memory and ARM Cortex-M0 Core Using 60-nm C-Axis Aligned Crystalline Indium-Gallium-Zinc Oxide FET Integrated With 65-nm Si CMOS," *IEEE JSSC*, vol. 52, no..4, pp. 925-932, 2017.
[2] Y. Liu, et al., "A 65nm ReRAM-Enabled Nonvolatile Processor with 6× Reduction in Restore Time and 4× Higher Clock Frequency Using Adaptive Data Retention and Self-Write-Termination Nonvolatile Logic," *ISSCC Dig. Tech. Papers*, pp. 84-86, Feb. 2016.
[3] Y. Tsuji, et al., "Sub-μW Standby Power, <18 μW/DMIPS@25MHz MCU with Embedded Atom-switch Programmable Logic and ROM," *Symp. VLSI Tech.*, pp. T86-T87, 2015.
[4] V. Singhal, et al., "A 10.5μA/MHz at 16MHz Single-Cycle Non-Volatile Memory Access Microcontroller with Full State Retention at 108nA in a 90nm Process," *ISSCC Dig. Tech. Papers*, pp. 148-150, Feb. 2015.
[5] N. Sakimura, et al., "A 90nm 20MHz Fully Nonvolatile Microcontroller for Standby-Power-Critical Applications," *ISSCC Dig. Tech. Papers*, pp.184-186, Feb. 2014.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

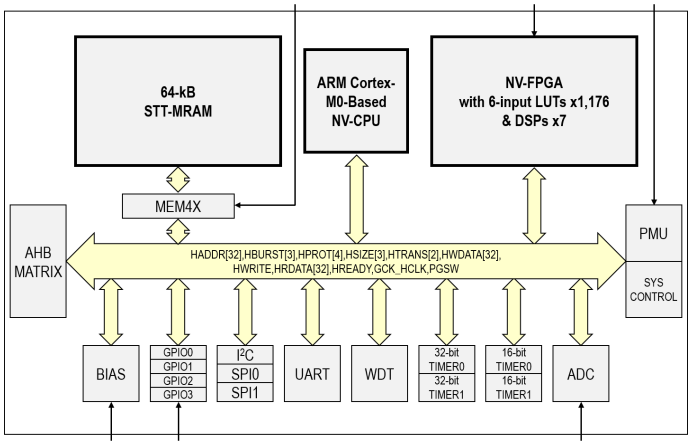


Figure 12.1.1: NV-MCU architecture.

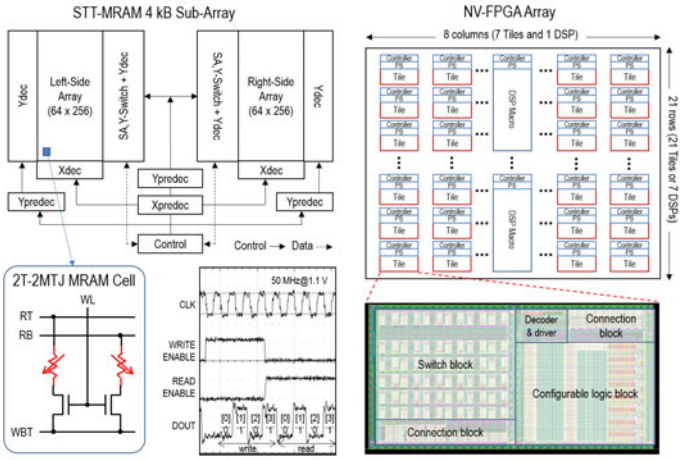


Figure 12.1.2: Embedded STT-MRAM and NV-FPGA.

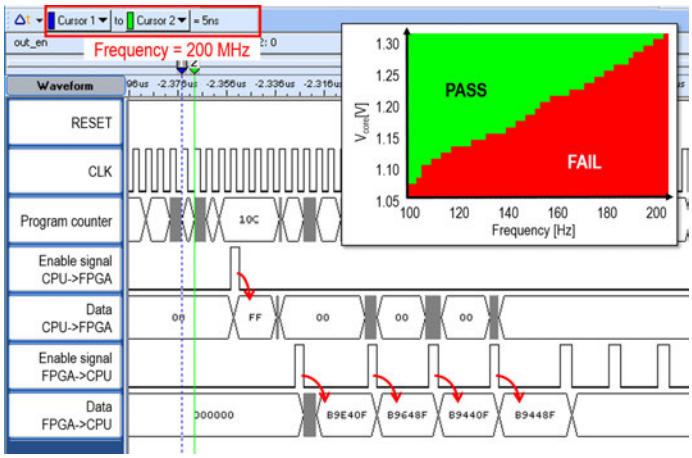


Figure 12.1.3: Measured waveform of FPGA-accelerated operation execution in ARM-Cortex-M0-based 32-bit nonvolatile CPU.

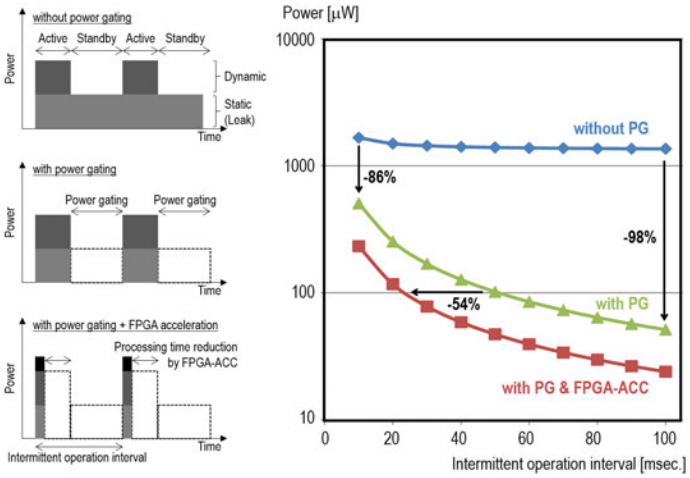


Figure 12.1.4: Impact of nonvolatile power gating and FPGA-based acceleration on the relationship between power dissipation and intermittent operation interval.

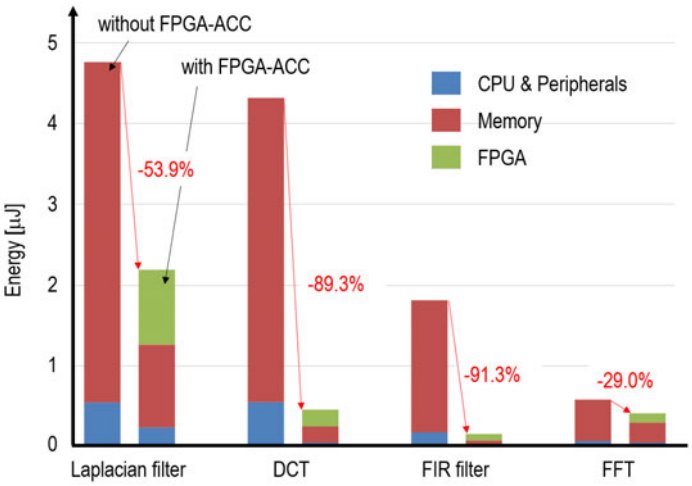
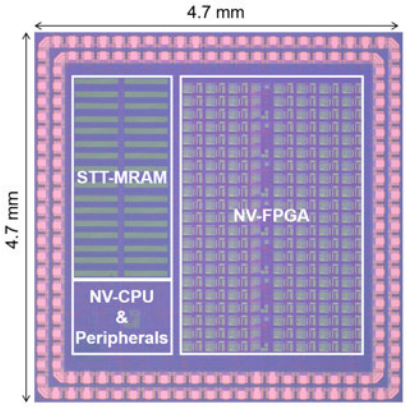


Figure 12.1.5: Breakdown of energy dissipation for typical operations.

Performance metrics	This work	JSSC2017 [1]	ISSCC2016 [2]	VLSI2015 [3]	ISSCC2015 [4]	ISSCC2014 [5]
Architecture	**Cortex-M0 32 bit**	Cortex-M0 32 bit	8051 8 bit	MSP430 16 bit	MSP430 16 bit	MSP430 16 bit
MOS technology	**40-nm LVT,SVT,HVT**	65-nm Low Leakage	65-nm SVT	65-nm SOTB	90-nm SVT,HVT	90-nm SVT
Memory technology	**2T-STT-MRAM 64 kB**	CAAC-IGZO 8 kB	ReRAM 8 kB nvSRAM 4 kB	Atom-switch 16 kB	N/A	3T-SpinRAM 64 kB
Flip-flops in CPU	**Nonvolatile**	Nonvolatile	Nonvolatile	Volatile	Volatile	Nonvolatile
Embedded programmable logic	**Yes (6-input LUTs x 1,176 & DSPs x 7)**	No	No	Yes (4-input LUTs x 2,000)	No	No
Supply voltage [V]	**1.8/3.3 (Peri.) 1.1-1.3 (Core)**	3.3(CAAC-IGZO) 1.1(CMOS)	3 (HV) 0.8 (Core)	0.41	0.72	1.8-3.3 (Dvcc) 1.0 (Core)
Frequency [MHz]	**200**	30	100	25	16	20
Active power [μW/MHz]	**26.7**	28.2	33	33	28.3	145
Standby power [μW]	**0.7**	0.009	N/A	0.7	0.32	1.2
Average power @intermittent operation [μW]*	**47.14**	143.3	>2900	583.2	499.8	2560

* Intermittent operation interval of 50 ms is assumed.

Figure 12.1.6: Comparison of the performance of the proposed MCU with that of other MCUs with embedded nonvolatile memories.



MOS Tech. Node	40-nm LVT, SVT, HVT
MTJ Tech. Node	39-nm perpendicular (Electrically determined size)
Supply Voltage	1.1-1.3 V (Core) 1.8/3.3 V (Peripherals)
MRAM Capacity	64 kB (4 kB sub-array x 16)
MRAM Sub-Array Structure	2T-2MTJ cell 256 cols. x 64 rows x 2
FPGA Capacity	1,176 six-input LUTs 7 DSPs
Transistor Count	4.8M
MTJ Count	1.5M

Figure 12.1.7: Prototype die micrograph and specifications.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

12.2 Micro Short-Circuit Detector Including S/H Circuit for 1hr Retention and 52dB Comparator Composed of C-Axis Aligned Crystalline IGZO FETs for Li-Ion Battery Protection IC

Hiroki Inoue[1], Takeshi Aoki[1], Fumika Akasawa[1], Toshiki Hamada[1], Toshihiko Takeuchi[1], Kousei Nei[1], Takako Seki[1], Yuto Yakubo[1], Kei Takahashi[1], Shuji Fukai[1], Takahiko Ishizu[1], Munehiro Kozuma[1], Ryota Tajima[1], Takanori Matsuzaki[1], Takayuki Ikeda[1], Makoto Ikeda[2], Shunpei Yamazaki[1]

[1]Semiconductor Energy Laboratory, Kanagawa, Japan
[2]University of Tokyo, Tokyo, Japan

Li-ion batteries are primarily used as power sources in electronic devices and electric vehicles and offer substantial conveniences to consumers. However, fires have broken out likely due to micro short-circuit (also called internal or soft short-circuit) in Li-ion batteries [1]. The micro short-circuit is a failure mode where Li metal first precipitates on a negative electrode and then reaches a positive electrode; eventually a short-circuit occurs between the negative and positive electrodes and the battery voltage slightly decreases. Repetitive occurrences of micro short-circuit will generate heat and lead to serious accidents with a fire or an explosion in the worst-case scenario. Micro short-circuit detection methods are still in the research phase, and applications of battery voltage monitoring systems with AI are expected [1,2]. This situation demands urgent research and development of micro short-circuit detectors for battery protection ICs.

In this study, a micro short-circuit detector composed only of n-channel c-axis aligned crystalline indium–gallium–zinc oxide FETs (CAAC-IGZO FETs) are formed on a battery protection IC. The micro short-circuit detector has a S/H circuit capable of long-term voltage retention and a current source load comparator that are made of analog circuits. The CAAC-IGZO FET is fabricated using a thin-film process and is a next-generation device that exhibits an extremely low off-state current [3]. A memory module that takes advantage of the extremely low off-state current of the CAAC-IGZO FET has been reported [4]. The memory cell of this memory module comprises CAAC-IGZO FETs and a capacitor. The capacitor can retain accumulated charge for a long time. Thus, a S/H circuit capable of long-term voltage retention can be achieved. Since the CAAC-IGZO FET is fabricated using the thin-film process, it can be readily formed on different types of supporting substrates, such as Si wafers and flexible substrates [3,5]. The CAAC-IGZO thin-film process is a promising fan out wafer level package (FOWLP) technology where interconnect wires and functional circuits including memory can be formed on a Si wafer. Therefore, a CAAC-IGZO FET micro short-circuit detector can be formed at a desired position on a Si CMOS battery protection IC.

Figure 12.2.1 shows the technology of a 0.36μm top-gate self-aligned CAAC-IGZO FET stacked on a Si supporting substrate. This structure eliminates overlap between the top gate, source, and drain electrodes, and reduces parasitic capacitance caused by the overlap. This smaller parasitic capacitance suppresses charge injection and feedthrough and increases the sampling accuracy of the S/H circuit. Device types used in this study are dual-gate and back-gate FETs. The dual-gate FET has topology where the top gate is connected to the back gate and has better gate controllability, i.e., higher on-state current and lower off-state current than a FET only with a top gate. Conversely, the back-gate FET has topology where the top gate and back gate voltages can be independently controlled. In the back-gate FET, V_t can be shifted positively via the application of a negative voltage to the back gate. The back-gate FETs exhibit extremely low off-state currents. Dual-gate FETs are used in circuits other than the S/H circuit to achieve high gain due to their high on-state current. The back-gate FET is used in the S/H circuit to retain voltage for a long time.

As shown in Fig. 12.2.2, the CAAC-IGZO FET micro short-circuit detector can be formed on the Si CMOS battery protection IC because it is composed only of CAAC-IGZO FETs. The structure enables the layout area reduction of the detector. The Si CMOS battery protection IC first detects over-charge/discharge and discharging/charging over-currents, then judges whether the battery abnormality is true or false via a delay circuit, and finally outputs a Dout/Cout signal to control external MOSFETs to protect the battery.

As shown in Fig. 12.2.3, the CAAC-IGZO FET micro short-circuit detector has the S/H circuit and the comparator. V_{batin} and V_{sh} are generated via V_{bat} divided using resistors to satisfy $\Delta V_{os}=V_{batin}-V_{sh}$. The S/H circuit composed of CAAC-IGZO FETs can retain V_{sh} for a long time. During this retention, the comparator compares V_{batin} with V_{sh} in a continuous manner. When there is no micro short-circuit, the comparator outputs a high voltage as V_{out} because V_{batin} is higher than V_{sh} and thus detects no abnormal operation. Conversely, when micro short-circuit occurs and V_{batin} temporarily becomes lower than V_{sh}, the comparator outputs a low voltage as V_{out} and thus detects the micro short-circuit, which indicates abnormal operation.

The comparator is dynamic and serves as a current source load amplifier with high gain, offset-cancellation, and bias current retention while combined with the S/H circuit. The comparator has two-stage current source load amplifiers and a latch circuit LATCH, as shown in Fig. 12.2.4. To detect micro short-circuit, the comparator needs to detect a potential difference of approximately 10mV. Since only n-channel CAAC-IGZO FETs are used, a diode-connected load differential amplifier or a resistance load differential amplifier is typically used. However, both differential amplifiers have low gain and therefore cannot detect a potential difference of approximately 10mV. In this study, we developed a differential amplifier with high gain by retaining gate–source voltage V_{gs} in the S/H circuit to achieve a current source load. Similarly, offset-cancellation and bias current retention are achieved via the retention of offset voltage V_{offset} and bias voltage V_{bias} in the S/H circuit. These voltages can be retained for a long time owing to the extremely low off-state current of the CAAC-IGZO FET. The operation timing of the comparator is as follows. During a S/H period, V_{gs} of M1–M4, V_{offset}, and V_{bias} are sampled and retained in the S/H circuit. During a comparison period, when LAT is set high, the comparator compares V_{in+} to V_{in-}. When EN is set high, the comparator outputs V_{out}.

Figure 12.2.5 shows the measured accuracy of the comparator in the micro short-circuit detector. The V_{out} toggle is observed after V_{in-} is set at 1.5V and a triangle wave of ±10mV is input to V_{in+}. Even though the detection accuracy of the comparator depends on V_{bat}, the comparator can detect 10mV even when V_{bat} ranges from 2.5 to 4.0V. The gain is 52dB at V_{bat}=4V; thus, a voltage drop of several tens of millivolts caused by micro short-circuit can be detected.

Figure 12.2.6 shows the retention performance of the micro short-circuit detector. First, V_{sh} of 1.5V is sampled and retained in the S/H circuit. After the retention, a triangle wave is input to V_{batin} in 10mV steps and a comparison is made in each step. Retention of V_{sh}, V_{gs}, V_{offset}, and V_{bias} that depends on the retention time is indirectly monitored via continuous monitoring of detected voltage at which the comparator toggles V_{out}. The measurement results suggest retention for more than 1h at V_{bg} of –2V and the adequate sampling period for micro short-circuit. Figure 12.2.7 shows a test chip micrograph from this study and a comparison table showing the performance of battery protection ICs.

References:
[1] M. Seo, et al., "Detection Method for Soft Internal Short Circuit in Lithium-Ion Battery Pack by Extracting Open Circuit Voltage of Faulted Cell," *Energies*, p. 1669, June 2018.
[2] M. Ismail, et al., "Battery State of Charge Estimation Using an Artificial Neural Network," *IEEE Transp. Electrific. Conf. and Expo*, pp. 342–349, June 2017.
[3] T. Ishizu, et al., "A 140MHz 1Mbit 2T1C Gain-Cell Memory with 60nm Indium–Gallium–Zinc Oxide Transistor Embedded Into 65nm CMOS Logic Process Technology," *Dig. Symp. VLSI Circuits*, pp. 162–163, June 2017.
[4] S. Maeda, et al., "A 20ns-Write 45ns-Read and 10^{14}-Cycle Endurance Memory Module Composed of 60nm Crystalline Oxide Semiconductor Transistors," *ISSCC Dig. Tech. Papers*, pp. 484–485, Feb. 2018.
[5] K. Myny, et al., "A Flexible ISO14443-A Compliant 7.5mW 128b Metal-Oxide NFC Barcode Tag with Direct Clock Division Circuit from 13.56MHz Carriers," *ISSCC Dig. Tech. Papers*, pp. 258–259, Feb. 2017.
[6] I. Lee, et al., "A 179-lux Energy-Autonomous Fully-Encapsulated 17-mm^3 Sensor Node with Initial Charge Delay Circuit for Battery Protection," *Dig. Symp. VLSI Circuits*, pp. 251–252, June 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

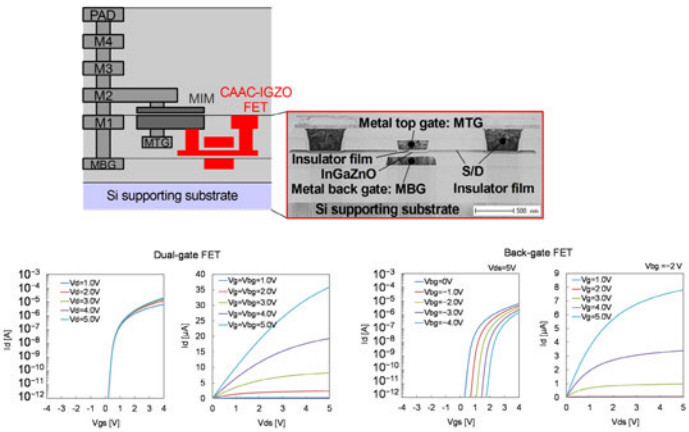


Figure 12.2.1: The technology of a 0.36μm top-gate self-aligned CAAC-IGZO FET stacked on a Si supporting substrate. Cross-section of the CAAC-IGZO FET and its electrical characteristics (W/L=0.36μm/0.36μm).

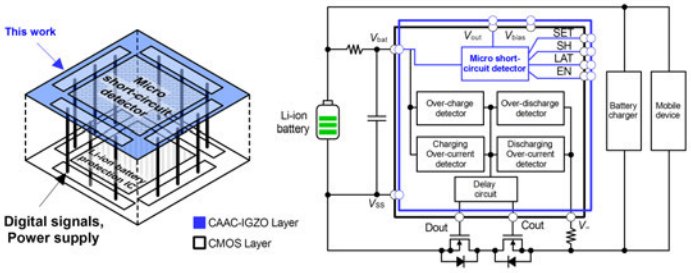


Figure 12.2.2: Chip configuration of the CAAC-IGZO FET micro short-circuit detector formed on the Si CMOS battery protection IC.

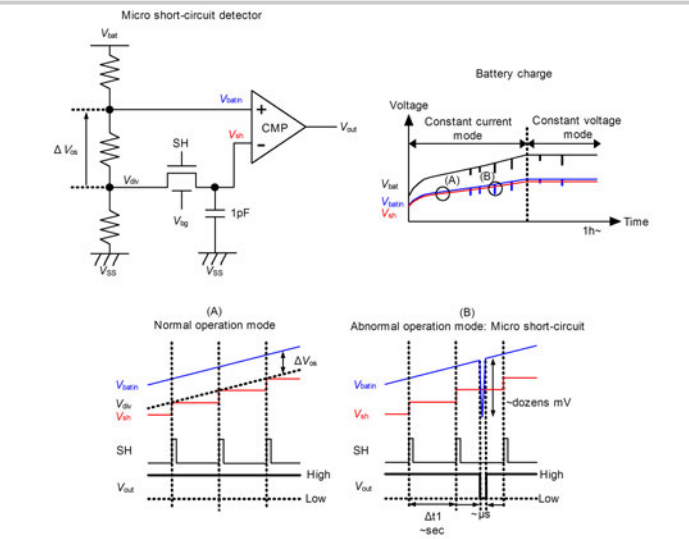


Figure 12.2.3: Micro short-circuit detector configuration and the operation timing diagram.

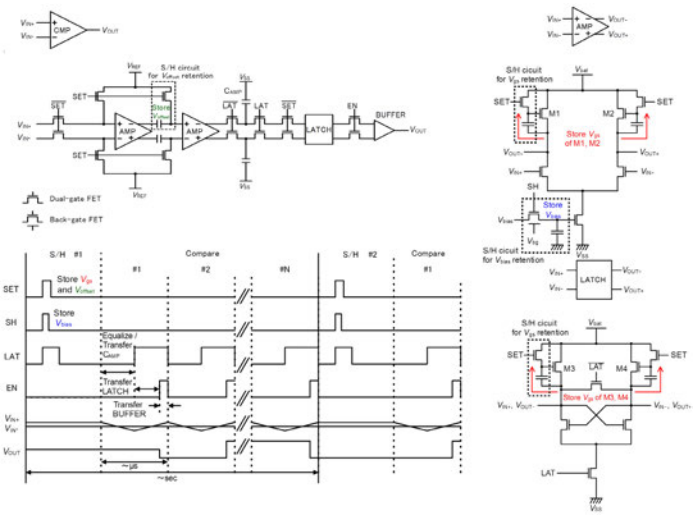
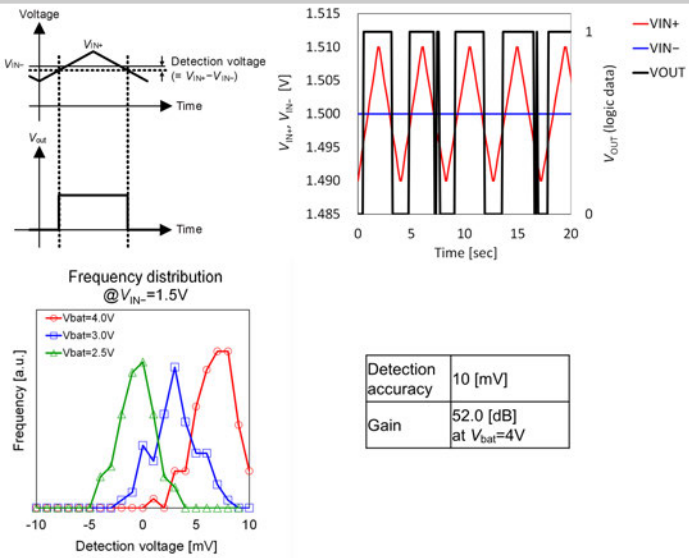
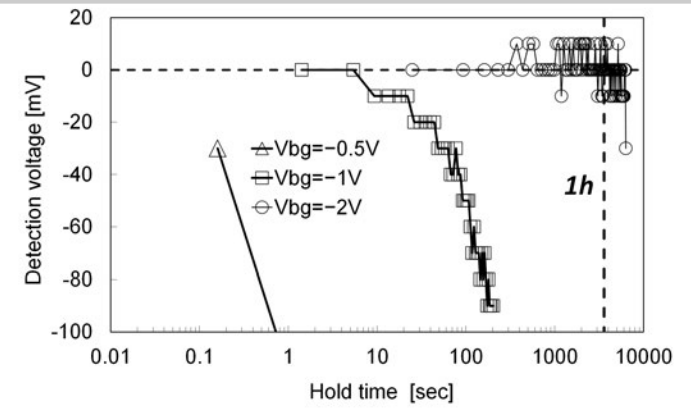


Figure 12.2.4: Current source load CAAC-IGZO FET comparator configuration and the operation timing diagram.



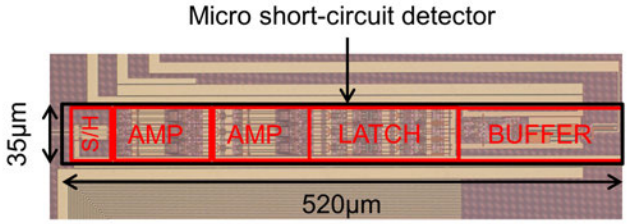
Detection accuracy	10 [mV]
Gain	52.0 [dB] at V_{bat}=4V

Figure 12.2.5: Measurement results for the detection accuracy.



		C [F]	ΔV [V]	Δt [sec]	I_{cut} [A]
V_{bg}	−0.5V	1p	100m	< 1	> 10^{-13}
	−1.0V			< 300	10^{-16}
	−2.0V			> 10000	< 10^{-17}

Figure 12.2.6: Measurement results for the hold time.



	[5]	[6]	This work
Device	Oxide FET	CMOS	CAAC-IGZO FET
Technology [nm]	1500	180	360
Power supply or V_{bat} [V]	3/6	3.6–4.1	2.5–4.0
#Tr.s	1.7k	N/A	1k
Frequency [Hz]	13.56M	–	4k
Power [μW]	7500	0.3	10
Micro short-circuit func.	–	no	yes

Figure 12.2.7: Micro short-circuit detector chip micrograph and a comparison of battery protection ICs.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

12.3 Memory Solutions for Flexible Thin-Film Logic: up to 8kb, >105.9kb/s LPROM and SRAM with Integrated Timing Generation Meeting the ISO NFC Standard

Florian De Roose[1,2], Jan Genoe[1,2], Auke J. Kronemeijer[3], Kris Myny[2], Wim Dehaene[1,2]

[1]KU Leuven, Heverlee, Belgium
[2]imec, Heverlee, Belgium
[3]Holst Centre/TNO, Eindhoven, The Netherlands

Thin-film transistor (TFT) technologies have long been used predominantly for display fabrication and are attractive for large area, low cost and flexible circuit applications. Thanks to the improving performance of thin-film metal-oxide NFC tags and data processing chips on foil [1,2], fabs are considering the large-scale production of flexible logic circuits. However, these systems require a memory, but metal-oxide technology lacks reliable, large memory arrays. Today, data storage is limited to ROMs, flipflops and SRAMs. No memory array has been demonstrated with sufficient storage capacity and speed within the typical power and area budget. This paper demonstrates the first large, fast and low-power memory array in flexible metal-oxide technology, comparable to the Si Intel 4000 series in the seventies [3].

One of the most demanding TFT applications today is an NFC tag according to the ISO14443 standard. Sending NDEF messages requires at least 1kb of memory at a rate of 105.9kb/s. However, no metal-oxide-based memory of this size has been shown before. Another hot topic for thin-film circuits is computation, e.g. implementing encryption. Even the most lightweight encryption algorithms require a rewritable memory bank and a large one-time-programmable memory for storing the key. The largest rewritable memory banks shown are a 4x8b register bank [2] and a 128b SRAM [4]. The largest ROM memory is 144 bits [2]. To enable encryption on foil, larger and faster memories are needed.

The circuits in this paper use an a-IGZO ESL technology [4], a proven and reliable but slow technology. The channel length L_{min} is 15µm (reference) or 9µm (scaled). Figure 12.3.1 shows the performance of the diode load inverter depending on the supply. Diode load logic is fast and robust, at the cost of high power consumption [1]. To improve robustness, the backgate (BG) is used to increase the VT of the drive transistor (inset Fig. 12.3.1). The average inverter has a delay of 2.75µs, while consuming 24.1µW. In samples, speed correlates very closely with power, since the pull-up current determines both.

This work focuses on two types of memory: Laser Programmable ROM (LPROM) and SRAM. Figure 12.3.2 shows the architecture of all matrices. Three LPROM versions are demonstrated. The reference design uses a two-phase readout with precharge to increase speed and reduce cell size [4]. A single multi-input NOR gate is used in the decoder (DEC), increasing speed compared to [4]. To improve the readout speed, a version is designed with a sense amplifier (SA) on the bitline. A pass gate multiplexer (PGM) serves as column decoder to reduce the number of SAs and reduce total power consumption. Third, a timing generation circuit was added to the reference design to show the feasibility and performance of a completely integrated system. The SRAM arrays have 512 bits, in 8 bitlines of 64 cells. The matrix dimensions are chosen asymmetrically to match the bitline behaviour to a larger system, i.e. 64×64.

Figure 12.3.3 shows the schematics of the new blocks. The pass gate MUX consumes less power compared to a diode load implementation since no static leakage path exists. Furthermore, the delay in the MUX is reduced, because the discharge delay of the bitline is dominant with respect to the RC delay of the MUX, whereas the diode load MUX adds two gate delays. Secondly, the sense amplifier (SA) is introduced in the LPROM matrices to decrease readout time and energy. Contrary to SRAM, there are no differential signals available per cell as input of the SA. Therefore, a dummy line is added. The dummy line is like other bitlines, but the pull-down transistor is only half the size. A dummy half-sized transistor is added to match the capacitance. The bitline discharges at half the speed of a regular line, creating a matched reference for the SA. The output of the dummy line also passes through the same PGM to closely match delays.

Finally, a timing generation circuit was designed to generate the precharge (Pre) and wordline enable (WLEN) signal from a clock. Figure 12.3.3 shows the schematic of this circuit, while Fig. 12.3.5 shows the timing diagram. A dummy line is added to the matrix. The dummy line is identical to any bitline with additional load to ensure it is always slower than the other bitlines. This line operates as a reference to monitor the speed of the operation. Since it operates under the same conditions, the timing of this line matches the real bitline behaviour. First, the incoming clock signal is converted in a short Start Pulse. This signal is used to start the precharge phase by setting the precharge flipflop. A sense inverter with a high trip point, set by the BG, is added on the dummy line. When the dummy line reaches this high trip point, the precharge flipflop is reset, and the wordline flipflop is set. Therefore, the WLEN signal is activated and the dummy line discharges. When the dummy line is sufficiently discharged, a sense inverter with low trip point resets the WLEN, ending the readout cycle.

In Fig. 12.3.4, the effect of the different changes on the LPROM circuits are shown. All data is measured at 10V supply, and one die with 24 different arrays is shown in Fig. 12.3.7. Starting from the literature reference, first the size is increased to 1024 bits, which increases the delay. Additionally, an improved decoder reduces delay and power. Next, the size is again increased 8-fold, with a delay increase of only a factor 2, and a power increase of less than 4, demonstrating the effectiveness of the employed techniques. In the next step, the L_{min} is reduced from 15µm to 9µm, including similar scaling of wiring and parasitics. This increases readout speed by a factor 2.35, while the power increases slightly, by 7.8% to 6.1mW, due to a slight decrease of VT for shorter channel lengths. Two variants are made based on this design. Introducing the PGM and SA, speed increases by 30% and power reduces by 36%. When the integrated timing circuit is used, the delay increases by 50% compared to the ideal timing applied in the lab due to the required design margins. The energy consumption decreases by 18% because the output drivers are engaged for a shorter time.

Figure 12.3.5 shows the resulting timing for the 8kb LPROM with SA and the SRAM memory at 10V supply. The Shmoo plots show the improvement in speed when using higher supply voltages. The 8kb LPROM with SA operates at 87kb/s. The performance of the 1kb LPROM memory meets the NFC spec using NDEF messages, where the required data rate is 105.9kb/s. This memory operates at 7V at 140kb/s, with only 0.55mW power consumption. However, if 15V supply is used, the speed increases to 0.5Mb/s at 8.95mW power consumption.

Figure 12.3.6 shows a comparison with the state of the art and the Intel Si 4000 chipset of the early '70s. The SRAM memory is 4× larger than previously published metal-oxide memories, operates at 10V supply, and is >5.5× faster. The performance of LPROM also significantly improves: even though internal timing is used, the size increased from 128b to 8kb, and the supply is only 10V, the speed is increased to 44kb/s at a power of only 1.67µW/b, a 21-fold reduction. If a sense amplifier is combined with a pass gate mux, the speed can be increased to 87kb/s. Finally, we have shown an NFC-compatible 1kb memory which operates from 7V supply at 140kb/s, using only 0.540µW/b. Despite the large difference in mobility between pMOS (~450cm²/Vs) and metal-oxides (~10cm²/Vs), the performance of both SRAM and LPROM with similar L_{min} is very similar.

Acknowledgements:
The authors are very grateful for the received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 644331 (PING project) and No 716426 (FLICs project). F.D.R thanks the Agency for Innovation by Science and Technology in Flanders (IWT-Vlaanderen) for financial support.

References:
[1] K. Myny, et al., "The development of flexible integrated circuits based on thin-film transistors," *Nature Electronics*, vol. 1, pp. 30-39, 2018.
[2] K. Myny, et al., "A Thin-film microprocessor with inkjet print-programmable memory," *Scientific Reports*, vol. 4:7398, pp. 1-6, 2014.
[3] A. Armstrong, et al., "Intel 4004 – 45th Anniversary Project", 2016. [Online]. Available: www.4004.com. [Accessed: 09-Sept.-18]
[4] F. De Roose, et al., "A Thin-Film, a-IGZO, 128b SRAM and LPROM Matrix With Integrated Periphery on Flexible Foil," *IEEE J. of Solid-State Circuits*, vol. 52, no. 11, pp. 3095-3103, 2017.
[5] J. Avila-Niño, et al., "Stable organic static random access memory from a roll-to-roll compatible vacuum evaporation process," *Organic Electronics*, vol. 31, pp. 77-81, 2016.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

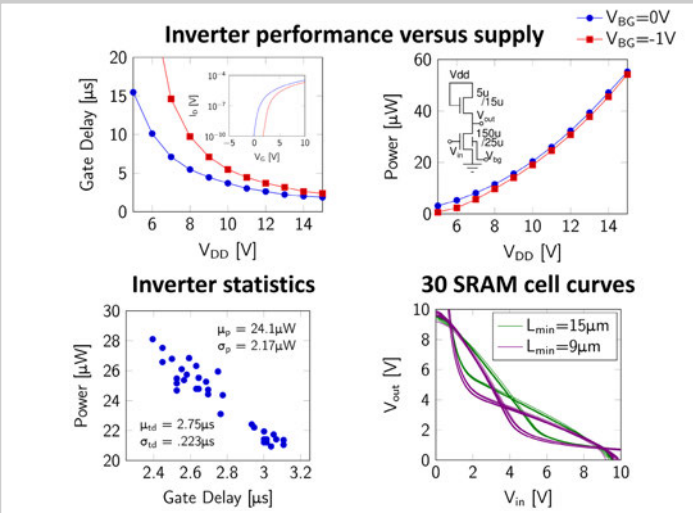


Figure 12.3.1: Performance of the 15µm inverter in the a-IGZO ESL technology: speed, power and variability at 10V V_{DD}. Inset: transfer curve of a TFT with VBG=0V, -1V. Butterfly diagram of 30 SRAM cells with varying scaling.

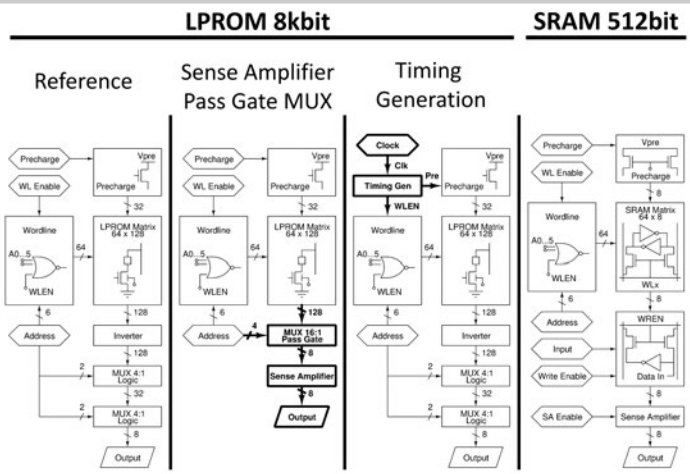


Figure 12.3.2: The architecture of the different fabricated memories, demonstrated for the largest array size.

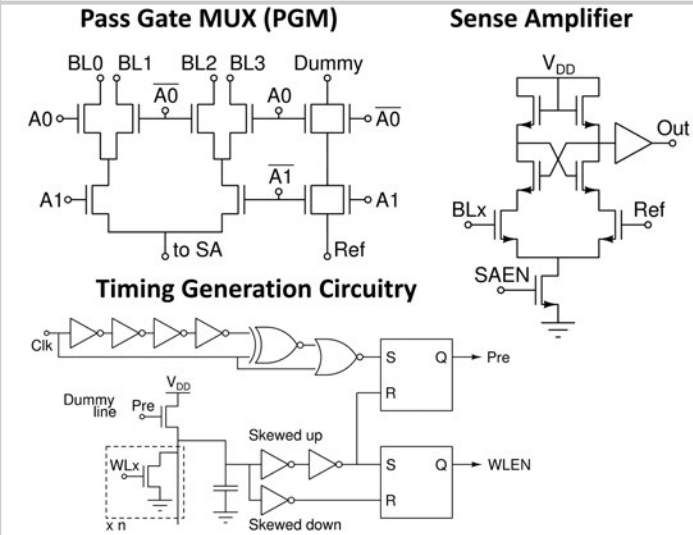


Figure 12.3.3: Schematics of the introduced components. The PGM is shown for a 4:1 configuration but is extended to a 1:16 configuration in the 8kb matrix.

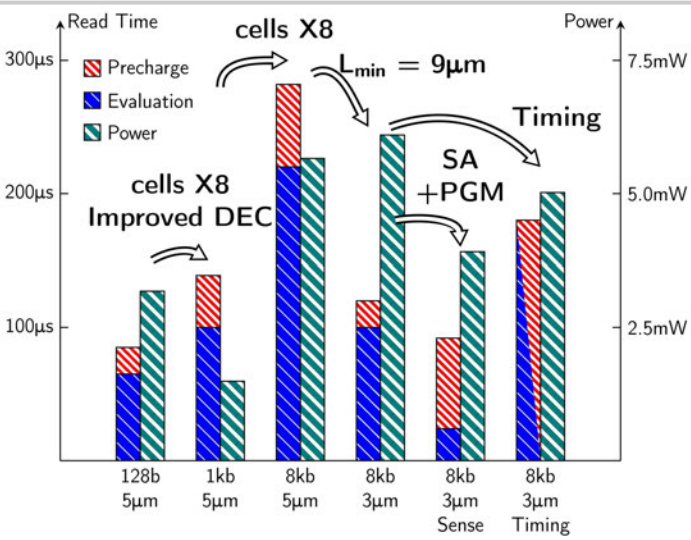


Figure 12.3.4: The measured effect of different LPROM circuit changes on measured speed and power consumption at 10V supply.

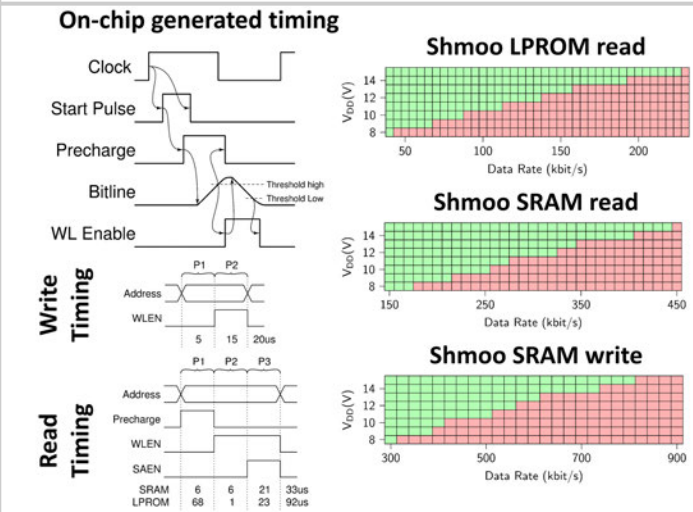


Figure 12.3.5: Operating method of the timing generation block. Optimal measured timing for writing and reading the SRAM (0.5kb) and LPROM with SA (8kb) at VDD=10V. Measured Shmoo plots for said circuits.

LPROM

	This work, NFC memory	This work, with SA	This work, with timing	De Roose et al., 2017 [4]	Myny et al., 2014 [2]	Intel 4001, 1971 [3]
Technology L_{min}	a-IGZO 9µm	a-IGZO 9µm	a-IGZO 9µm	a-IGZO 9µm	Hybrid compl. 5µm	PMOS 10µm
Matrix size	1kbit	8kbit	8kbit	128bit	144bit	4kbit
Supply voltage	7V	10V	10V	15V	10V	15V
Data rate	140kbit/s	87kbit/s	44kbit/s	25kbit/s	5.2kbit/s	750kbit/s
Power per bit	0.540µW	1.46µW	1.67µW	35.1µW	N.A.	54.9µW
Area efficiency*	29.0%	55.4%	49.6%	4.9%	21.1%	13.7%
Chip size	2.28x2.23mm²	5.43x3.91mm²	5.25x4.52mm²	1.5x2.5mm²	8.97x6.93mm²	4.0x2.6mm²

SRAM

	This work	De Roose et al., 2017 [4]	Avila-Nino et al., 2016 [5]	Intel 4002, 1971 [3]
Technology L_{min}	a-IGZO 9µm	a-IGZO 15µm	Organic 50µm	PMOS 10µm
Matrix size	512bit	128bit	16bit	320bit
Supply voltage	10V	15V	40V	15V
Data rate write	400kbit/s	72.7kbit/s	11.1kbit/s	375kbit/s
Data rate read	242kbit/s	30.8kbit/s	N.A.	375kbit/s
Power per bit	125.9µW	98.4µW	N.A.	797µW
Area efficiency*	23.8%	28.5%	No periphery	19.3%
Chip size	6.71x3.18mm²	3.94x3.13mm²	> 4x4mm²	3.3x3.8mm²

*Area efficiency is defined as matrix area / total area

Figure 12.3.6: Comparison Table. The circuits shown outperform thin-film state of the art in almost all specifications, and is quite comparable to the intel MCS4 platform.

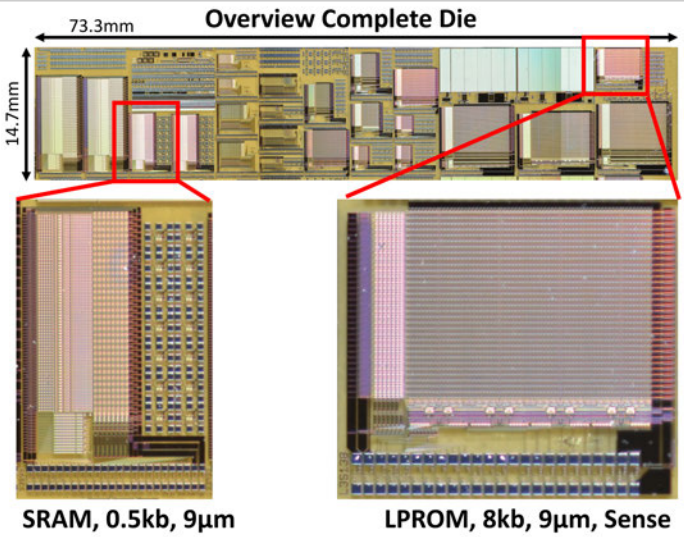


Figure 12.3.7: Photograph of the die area with all memory matrices. Closeup of the matrices in the comparison table.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 13 Overview: *Non-Volatile Memories*

MEMORY SUBCOMMITTEE

Session Chair: ***Yan Li***
Western Digital, Milpitas, CA

Associate Chair: ***Jongmin Park***
SK Hynix, Icheon-si, Gyeonggi-do, Korea

Subcommittee Chair: Jonathan Chang, TSMC, Hsinchu, Taiwan

The five papers in this session include the most advanced non-volatile memory technologies. With large 3D NAND cells, the 4bits/cell flash is going mainstream in many applications. In the 3D NAND memory, the memory stack grows up to 128 layers to achieve higher density. The new 4-plane architecture of putting circuits under the array to gain the highest write throughput of 132MB/s is presented. 1.2Gb/s IO speed is achieved in the 6th Generation VNAND. In the emerging memory, an embedded non-volatile ReRAM macro is done using 22nm FinFET technology. A 7Mb embedded STT-MRAM macro is designed in 22FFL FinFET technology.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

10:15 AM

13.1 A 1.33Tb 4-bit/Cell 3D-Flash Memory on a 96-Word-Line-Layer Technology

N. Shibata, Toshiba Memory, Yokohama, Japan

In Paper 13.1, Toshiba Memory Corp delivers the highest capacity flash memory chip. A 1.33Tb 4-bit/cell (QLC) 3D-Flash memory in a 96-Word-Line-Layer technology that achieves 8.5Gb/mm^2 is developed.

10:45 AM

13.2 A 3.6Mb 10.1Mb/mm^2 Embedded Non-Volatile ReRAM Macro in 22nm FinFET Technology with Adaptive Forming/Set/Reset Schemes Yielding Down to 0.5V with Sensing Time of 5ns at 0.7V

P. Jain, Intel, Hillsboro, OR

In Paper 13.2, Intel Corp reveals array design in 22nm FinFET process with an array density of 10.1Mb/mm^2, leveraging a low voltage ReRAM technology. ReRAM eNVM technology enables superior read performance and byte writeabiilty feature.

11:15 AM

13.3 A 7Mb STT-MRAM in 22FFL FinFET Technology with 4ns Read Sensing Time at 0.9V Using Write-Verify-Write Scheme and Offset-Cancellation Sensing Technique

L. Wei, Intel, Hillsboro, OR

In Paper 13.3, Intel Corp demonstrates Write-Verify-Write (WvW) scheme and offset cancellation sensing technique to achieve 7Mb STT-MRAM arrays in 22FFL FinFET Technology. Die by die tuning of a thin film precision resistor used as reference provides large sensing margin.

11:30 AM

13.4 A 512Gb 3-bit/Cell 3D 6th-Generation V-NAND Flash Memory with 82MB/s Write Throughput and 1.2Gb/s Interface

D. Kang, Samsung Electronics, Hwasung, Korea

In Paper 13.4, Samsung Electronics presents a 512Gb 3b/cell 3D V-NAND flash memory featuring 6th generation V-NAND technology. To improve read and write performance, wordline and bitline setup time reduction techniques are developed.

11:45 AM

13.5 A 512Gb 3-bit/Cell 3D Flash Memory on 128-Wordline-Layer with 132MB/s Write Performance Featuring Circuit-Under-Array Technology

C. Siau, Western Digital, Milpitas, CA

In Paper 13.5, Western Digital presents the A 512Gb 3b/cell 3D-flash memory on 128-word-line-layer technology. 4 planes architecture had been proposed to attain performance of 132MB/s with CMOS under array architecture.

13.1 A 1.33Tb 4-bit/Cell 3D-Flash Memory on a 96-Word-Line-Layer Technology

N. Shibata[1], K. Kanda[1], T. Shimizu[1], J. Nakai[1], O. Nagao[1], N. Kobayashi[1], M. Miakashi[1], Y. Nagadomi[1], T. Nakano[1], T. Kawabe[1], T. Shibuya[1], M. Sako[1], K. Yanagidaira[1], T. Hashimoto[1], H. Date[1], M. Sato[1], T. Nakagawa[1], H. Takamoto[1], J. Musha[1], T. Minamoto[1], M. Uda[1], D. Nakamura[1], K. Sakurai[1], T. Yamashita[1], J. Zhou[1], R. Tachibana[1], T. Takagiwa[1], T. Sugimoto[1], M. Ogawa[1], Y. Ochi[1], K. Kawaguchi[1], M. Kojima[1], T. Ogawa[1], T. Hashiguchi[1], R. Fukuda[1], M. Masuda[1], K. Kawakami[1], T. Someya[1], Y. Kajitani[1], Y. Matsumoto[1], N. Morozumi[1], J. Sato[1], N. Raghunathan[2], Y. L. Koh[2], S. Chen[2], J. Lee[2], H. Nasu[2], H. Sugawara[2], K. Hosono[1], T. Hisada[1], T. Kaneko[1], H. Nakamura[1]

[1]Toshiba Memory, Yokohama, Japan
[2]Western Digital, Milpitas, CA

Since 3D-Flash memory took over for 2D-Flash memory, chip capacity has continuously improved [1-3]. In the 2D-Flash era, 2b/cell (MLC) offered higher performance and reliability, while a 3b/cell (TLC) offered the lowest cost. Thanks to a larger feature size, 3D Flash cell reliability is much better than that of 2D. As a result, TLC 3D-Flash became the mainstream non-volatile memory, since it satisfies most market requirements in both performance and reliability. To meet the continuously growing market demand for higher capacity and lower cost, a 4b/cell (QLC) 3D-Flash memory in a 96-WL-layer technology is presented. It achieves an 8.5Gb/mm^2 area capacity, which is 42~50% greater than the 3D-Flash memory reported in [2,3]. A chip micrograph and a table summarizing key features is shown in Fig. 13.1.7. The total 1.33Tb capacity is the highest single Flash memory chip capacity reported thus far.

To realize QLC a wide V_{th} window is required to store 16 distinct cell levels. One option is to expand the range of V_{th} into the negative region, but negative WL voltages require a triple-well process, which is costly; hence, the source-bias negative sense (SBNS) scheme was proposed [4], in which the cell source (SRC) is biased to a higher potential than the WL for reading of negative V_{th}. However, the amount of negative V_{th} is limited by the sense amplifier's supply voltage, since the signal voltage in the amplifier becomes higher due to the boosted SRC. In this work a new SBNS with clock control (C-SBNS) is implemented to further expand the negative V_{th} region, while maintaining a low supply voltage. Figure 13.1.1 illustrates the clock control concept. Sensing starts with the PRECHARGE phase where BL and SEN nodes are precharged to SRC + ΔV_{BL} and V_{DD}. The BOOST phase begins with CLK1 and CLK2 high, boosting of SEN and OP-amp threshold. The SENSE phase follows, with XXL high; by enabling the XXL switch the SEN level develops depending on the cell data. SEN is not discharged for an off-cell (except due to the cell's subthreshold leakage) and discharged to SRC + ΔV_{BL} + ΔBLC + ΔXXL for an on-cell. At the end of the BOOST phase, SEN in Fig. 13.1.1 is kicked down by CLK1 being low. Note that the amplifier's threshold is shifted up, to a level that is higher than the SEN node for an on-cell. As a result, a higher SRC voltage can be implemented. With this C-SBNS technique, a wider range of negative V_{th} for QLC can be sensed even at a low amplifier supply voltage.

For 3D-Flash using a MONOS cell and a larger memory cell size, the cell-to-cell coupling effect [5] is less than for 2D-Flash. Using a full-sequence (FS) programming method delivers higher programming performance and is commonly used for both MLC and TLC operations because of the reduced V_{th}-distribution degradation after the programming of neighboring cells. Using QLC, for 3D-Flash, presents other challenges to realizing a narrower V_{th} distribution: program disturb and initial charge loss. A two-step programming (16 levels programmed in 1st-step, and 16 in the 2nd) method, which is used in QLC 2D-Flash [5], can suppress these V_{th} degradation factors [3]. However, programming throughput is degraded because of the two 16-level programming steps. To overcome this challenge, a new two-step (8-16) programming method is developed in this work. Figure 13.1.2 shows the programming order and V_{th} transitions of this two-step (8-16) method. (1) cells on the WLn of string0 are programmed to 8 coarse levels. (2) cells in the WLn of string0 are programmed to the final 16 levels after the programming of neighboring cells on WL_{n+1} and the other strings' cells (string1-3) are programmed from initial level to coarse levels and the neighboring cells on WL_{n-1} and the other strings' cells (string1-3) are programmed from coarse levels to final levels. In comparison to a conventional 16-16 two-step program approach, the verify time is improved since only 8 levels are verified in step (1). Figure 13.1.3 shows the measured BER and programming time (t_{PROG}) of these programming methods. The BER dramatically improves for both two-step programming methods, 8-16 and 16-16, as compared to FS. Note that the BER difference between these two-step programming methods is small while t_{PROG} of the 8-16 two-step programming method is 18% better than the conventional 16-16 two-step method. The measured 16-level V_{th} distribution using the 8-16 programming method is also shown in Fig. 13.1.3.

Since random-access read time is critical in achieving higher system level performance, we employed a page/state dependent WL overdrive concept. The WL overdrive technique [1] enables the reduction of the larger 3D WL RC delays. The benefit of this concept is further magnified in QLC due to the many read levels, compared to TLC/MLC. For a QLC page-by-page read 1b out of 4b in a cell is read in one read sequence, and to distinguish the 16 QLC levels it is necessary to change 15 WL levels; therefore, the average WL level transition is 15 WL levels per 4 pages, which is 3.75. In contrast, for a TLC page-by-page read 1b out of 3b in a cell is read in one read sequence, and to distinguish the 8 TLC levels it is necessary to change 7 WL levels; therefore, the average WL level transitions is 7 WL levels per 3 pages, that is 2.33. Finally, for a MLC page-by-page read 1b out of 2b in a cell is read in one read sequence, and to distinguish the 4 MLC levels it is necessary to change 3 WL levels; therefore, the average WL level transitions is 3 WL levels per 2 pages, that is 1.5.The amount of WL voltage change in each WL transition is different as shown in Fig. 13.1.4. For example, the voltage change at the WL transition from "MP2" to "MP3" state is larger than that of "MP1" to "MP2". In the proposed scheme, in "MP3" state read the WL overdriving voltage and duration time is set to a larger value than that of a "MP2" state read. Using optimized WL overdrive settings for page and state the average WL transient time is successfully reduced by 8%.

To further improve total system performance, an independent plane read (IPR) scheme is proposed. For a conventional scheme, the WL address for all selected planes (for multi-plane operations) is required to be the same. During a program operation, the program voltage (V_{PGM}) is applied to the target WL while the neighboring 23 WLs are charged to intermediate voltages to reduce the high-electric-field stress. The remaining 72 (96-24) WLs are biased to a fixed voltage V_{PASS} from a high-voltage switch (HVSW) [6]. To bias the various voltages for 24 WLs, a total of 24 control gate (CG) driver circuits are needed. Therefore, to minimize the area impact of the CG driver the WL address selection between planes is required to be same in a conventional scheme. IPR removes this requirement, which results in improved system flexibility and performance. Note that during a read, lower WL voltages, compared to programming, are needed. Therefore, only 12 WLs are biased to various voltages while the remaining 84 (96-12) WLs are biased to a fixed voltage V_{READ} from the HVSW. IPR is implemented without any additional CG driver circuits, by using a practical arrangement of existing CG driver circuits. As shown in Fig. 13.1.5(a&b), 24 CG-driver circuits are used to drive a common set of 24 WLs between planes during programming. During read, these same 24 CG driver circuits are partitioned into 2 sets of 12, independently driving 12 WLs of each plane.

In addition to SLC mode, mixed use of QLC, TLC and SLC is also supported (Fig. 13.1.6a-c). For system implementation convenience, a flexible protocol of SLC and TLC modes are introduced by adding simple pre-set commands. These modes can be applied dynamically for each block as shown in Fig. 13.1.6(d). For example, SLC mode can be used for storing data when high-speed operation is required. Subsequently, these data can be moved to QLC blocks as needed.

To further provide flexibility for system implementation, during multi-plane read, this chip is designed to allow for mixing selection of QLC pages (lower/middle/upper/top) between planes. It is also possible to select one QLC page in one plane and on TLC/SLC page in the other as shown in Fig. 13.1.6(e). For example, a QLC lower page (user data) read in Plane0 can be done concurrently with an SLC (system table) read in plane1. This is realized, as an extension of the IPR implementation, without adding independent control circuits for each plane, since the sub-clock timing for various reads is virtually common. By leveraging this added flexibility, the system performance is further improved.

Acknowledgements:
The authors would like to thank H. Mukai, H. Saito, T. Shigeoka, I. Kawabata, Y. Takeuchi, K. Shimizu, K. Ino, and S. Yoshikawa, as well as the Design, Layout, CAD, Device, Evaluation, Test, and Process teams.

References:
[1] J. W. Im, et al., "A 128Gb 3b/cell V-NAND flash memory with 1Gb/s I/O rate," *ISSCC*, pp. 130-131, 2015.
[2] H. Maejima, et al., "A 512Gb 3b/Cell 3D Flash Memory on a 96-Word-Line-Layer Technology," *ISSCC*, pp. 336-337, 2018.
[3] S. Lee, et al., "A 1Tb 4b/cell 64-Stacked-WL 3D NAND Flash Memory with 12MB/s Program Throughput," *ISSCC*, pp. 340-341, 2018.
[4] G.G. Marotta, et al., "A 3bit/Cell 32Gb NAND Flash Memory at 34nm with 6MB/s Program Throughput and with Dynamic 2b/Cell Blocks Configuration Mode for a Program Throughput Increase up to 13MB/s," *ISSCC*, pp. 444-445, 2010.
[5] N. Shibata, et al., "A 70nm 16Gb 16-level-cell NAND flash memory," *JSSC*, vol. 43, no. 4, pp. 929-937, Mar. 2008.
[6] K. Kanda, et al., "A 120mm2 16Gb 4-MLC NAND Flash Memory with 43nm CMOS Technology," *ISSCC*, pp. 430-431, 2008.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

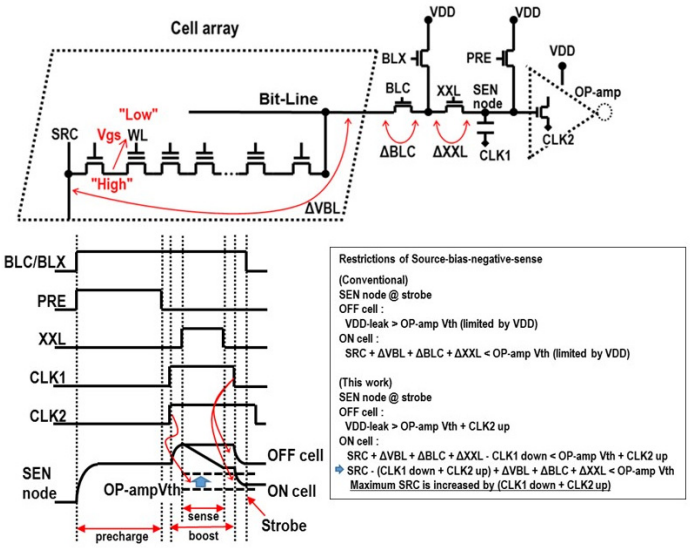


Figure 13.1.1: Source-bias negative sense with clock-control (C-SBNS).

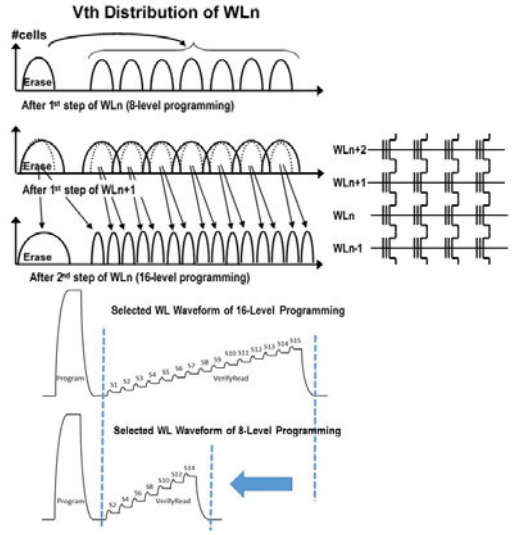


Figure 13.1.2: Proposed two-step (8-16) programing method: programming order & V_{th} transition.

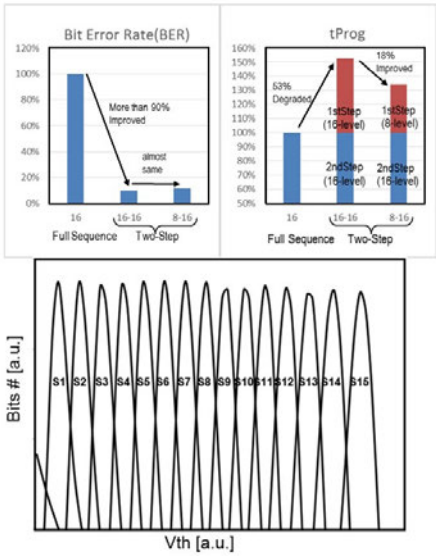


Figure 13.1.3: Measured BER, performance and 16-level V_{th} distribution.

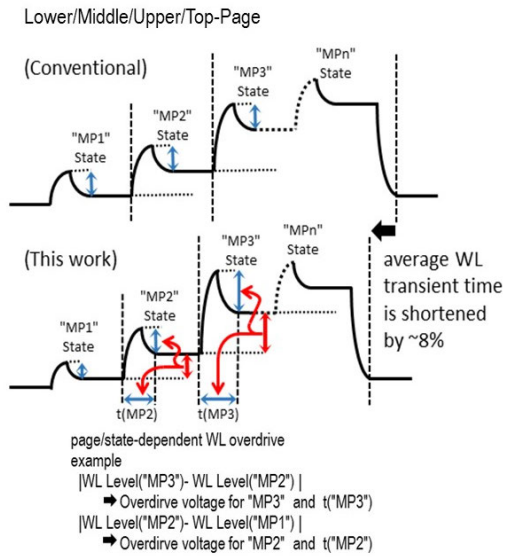


Figure 13.1.4: Page and state dependent WL overdrive.

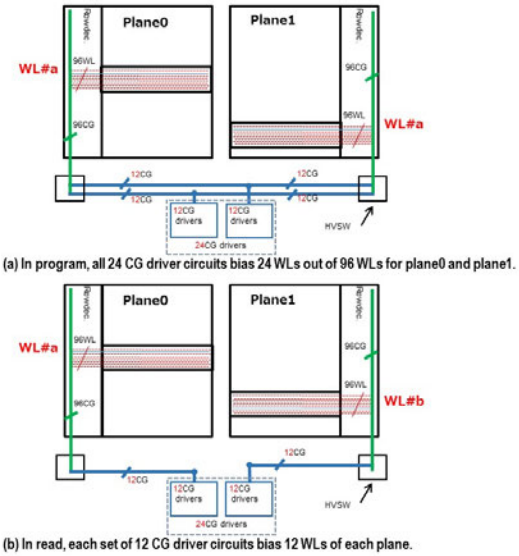


Figure 13.1.5: Independent plane read (IPR).

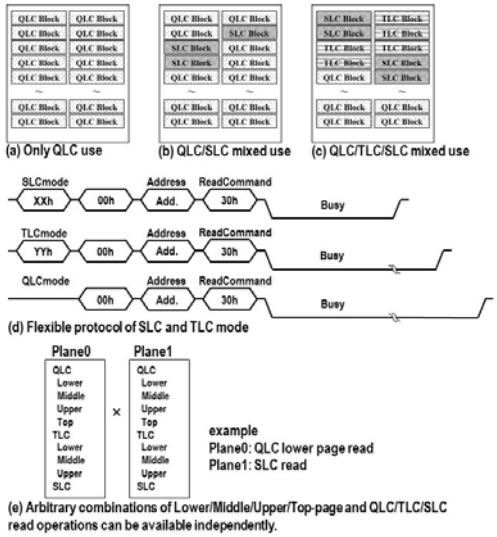
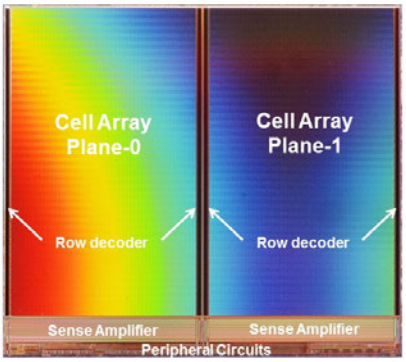


Figure 13.1.6: Arbitrary combinations of lower/middle/upper/top-page and QLC/TLC/SLC read operations.



Bit per cell	4
Density	1.33Tb
Technology	96-WL-Layer 3D Flash Memory
Organization	(16kB + ECC) /Page 1536(=96x4x4) Pages/Block (3642 + EXT) Blocks/Plane 2 Planes
Program Throughput	9.3MB/s
tR	160us (Ave.)
Burst Cycle Time	800Mb/s/pin
Power Supply	Vcc: 2.30V to 3.6V Vccq:1.2V

Figure 13.1.7: Micrograph and key features of 1.33Tb 4-Bit/Cell (QLC) 3D-Flash memory.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

13.2 A 3.6Mb 10.1Mb/mm² Embedded Non-Volatile ReRAM Macro in 22nm FinFET Technology with Adaptive Forming/Set/Reset Schemes Yielding Down to 0.5V with Sensing Time of 5ns at 0.7V

Pulkit Jain, Umut Arslan, Meenakshi Sekhar, Blake C. Lin, Liqiong Wei, Tanaya Sahu, Juan Alzate-vinasco, Ajay Vangapaty, Mesut Meterelliyoz, Nathan Strutt, Albert B. Chen, Patrick Hentges, Pedro A. Quintero, Chris Connor, Oleg Golonzka, Kevin Fischer, Fatih Hamzaoglu

Intel, Hillsboro, OR

A resistive RAM (ReRAM) macro is developed as a low-cost, magnetic-disturb-immune option for embedded, non-volatile memory for SoCs used in IoT and automotive applications. We demonstrate the smallest ReRAM subarray density of 10.1Mb/mm² in a 22nm low-power process. The subarray uses nominal-gate FINFET logic devices, with material innovations to allow low-voltage switching without impacting transistor reliability. Prior art features larger bit cell size or array density, and uses 28 or 40nm technology nodes [1-4]. The smallest read-sense time (t_{SENSE} = 5ns@0.7V) is demonstrated, compared to previous works [2]. An optimized pulse-width (PW) voltage-current write-verify-write (PVC-WVW) sequence helps in mitigating endurance and variability. A flexible and low-area TFR (thin-film resistor) based reference scheme enables optimization of forming, write yield, retention and endurance tradeoffs by skewing different verify and read resistances. A temperature-constant current source and a reference resistance help in the precise control of the forming/set current and the verify/read operations. Compared to area-inefficient bandgap circuits and temperature sensors, the in-situ TFR was used due to its low area, flexibility and seamless integration into the SoC. The memory bank uses a single supply coming from an in-situ charge pump (CP) that is shared across the macro.

Figure 13.2.1 shows the polarity of forming, set and reset operations, and the measured read resistance vs. write voltage (R-V) characteristic for a typical device. Forming and set operations use current control through the analog driver for precise ReRAM filament control. Reset is done via a digital driver. A source follower is kept in the reset direction, on purpose, to provide maximum drive during forming without hurting the periphery reliability. The R-V trace illustrates the low voltage needed to switch a bit in either the set or reset direction.

Figure 13.2.2 shows the architecture of the 3.6Mb functional unit. It consists of 6 blocks, and each block consists of a mid-logic, CP, level-shifter interface and two subarrays. The mid-logic shifter includes scan chains, and logic for address decoding and to generate write-verify sequences. The CP provides a regulated and adjustable subarray voltage, which is shared across the two subarrays. The subarray gets V_{NOM}/V_{CCHI} during write and forming operations, and V_{NOM} during a read. A subarray consists of 16+2-ECC global columns (GC), which are stacked vertically and two sectors horizontally, S0 and S1 . The GC is floor-planned to minimize the mismatch between S0 and S1 during read and set operations. Each GC consists of 32-local columns × 256 rows on each side of the offset cancelling current sense amplifier (OC-CSA) and the central write driver (WD) in an open-BL architecture. WD is also placed on the far side to maximize current drive. RD_{EN} and WR_{EN} multiplexers provide the ability to control read and write operations. The reference resistance is placed on the far right-side, and serves to provide reference for both sectors, since one sector is activated at a time. The BL and SL widths are kept different, to minimize any WL-dependent read current as the reference does not track this. The multiplexer, CTRL, PCH and WL controls in the two sectors are carefully separated/synchronized to ensure there is median leakage cancellation from the unselected sector during a read operation. Repeaters and metal-to-over-poly straps are use to achieve a fast WL slew rate. The write driver is gated by the XOR of SA_{OUT} and D_{IN} ($D_{OUT} \oplus D_{IN}$) to enable a verify before every write operation. This scheme ensures a precise application of PW, voltage and current setting for every bit, across every operational cycle and data polarity, optimizing the endurance of that bit

Figure 13.2.3 shows key circuit innovation details. The programmable WD consists of: (1) a binary-weighted cascode-current based set WD (SWD), which is biased using a shared set/form reference current generator $ISET_{REF}$; (2) a stacked reset WD (RWD); (3) additional gating to allow for the write driver to be disabled when $D_{OUT} \oplus D_{IN}$ is high for WVW. The simulation for $ISET_{REF}$ is shown in the center of Fig. 13.2.3, which shows less than 0.5% temperature and a 15% process sensitivity. The $ISET_{REF}$ generator amplifier is write-gated and biased at 150μA to minimize mismatch; this is less than the overall write power budget. The verify circuit consists of the sense amplifier (SA), an SA latch, an XOR gate and a TFR reference (R_{REF}). The latter is a long resistor ladder, with programmable control of the ground shunt to achieve the desired read and verify levels.

The start and end of set and reset operations are sensitive to PW, V_{CCHI} and current. Particularly, the end of set is more sensitive to PW and $ISET_{REF}$; while, the end of reset is more sensitive to PW and V_{CCHI}. The write-time is specified to be 10μs across process and the operating temperature range, a very flexible sequence of 12 steps (3 PW, 2 currents, and 2 voltages) was applied to leverage the best yield-endurance-retention without requiring die-to-die trimming. This scheme ensures that for every data polarity, an optimal PW, voltage and current setting is used. The voltage was kept at the outermost loop of the WVW steps, to minimize periphery stress. PW was kept at the innermost loop to avoid overset and overreset of the ReRAM filament during programming. Figure 13.2.4 shows an example set/forming simulation using the PVC-WVW scheme. It zooms into the combined pulse-width (PW) with I_{REF} algorithm steps for clarity; Two different polarity write examples are shown to illustrate the WVW sequence. A bit on IO<0> switches in the first write step. D_{OUT} on the subsequent read is compared against D_{IN}, and since it is a match, the subsequent write is prevented. A harder to write bit on IO<1>, on the other hand switches in the third write step in the illustration. We also illustrate how the subarray's V_{NOM} toggles to V_{CCHI} during a write operation.

Figure 13.2.5 shows the proposed OC-CSA scheme. The key features, relative to prior-art [2,3] are (1) a TFR based R_{REF} (instead of a reference voltage/current) that tracks the ReRAM resistance and optimal read-level target, independent of temperature; (2) a high-gain differential-cascode stage; and (2) common-mode feedback to increase the operating range of R_{REF}. The topology consists of a cascoded load first stage, with a cross-coupled latch as the second stage. A clamp transistor ensures low-voltage sensing against read disturb; it clamps the voltage on nodes *inD* and *inR* that are connected to the bitcell. In phase 1, the sampling phase, the SA offset and bias is captured in the capacitors connected to gates of M_{CS} and M_{CASC}. In phase 2, the amplification phase, the capacitors self-bias M_{CS} and M_{CASC} and generate 2x (I_{DATA}–I_{REF}). Finally, in phase 3, the second-stage latch resolves the signals to full rail. Waveforms for some corner cases for a high-resistance state (HRS) verify, a low-resistance state (LRS) verify, and a final read are shown in Fig. 13.2.5. A common-mode feedback transistor (CMFB) ensures that the first-stage amplifier remains in saturation around the common-mode level, especially during the HRS verify case of 100kΩ that suffers from a low-current margin.

Figure 13.2.6 shows the bit-error rate (BER) from the application of the 12-step WVW algorithm. The BER is an aggregate of eight set/reset switching events. A programmable R_{REF} allows for flexibility in choosing the verify- and read-level after manufacturing, which makes it easy to obtain the best BER, while preserving endurance and retention. The voltage-t_{SENSE} Shmoo for a 3.6Mb-ReRAM array tested over a wide range of temperature, –40 to +125°C, is shown in Fig. 13.2.6. The test chip works down to 0.8V, with the smallest (vs. [2]) t_{SENSE} = 6.7ns across temperature. The array also supports a wide supply range, down to 0.5V, with t_{SENSE} = 5ns at 0.7V.

Figure 13.2.7 shows the die-microphotograph of the 3.6Mb-ReRAM array and a feature summary table; the 0.3Mb subarray layout is also shown. This work shows a feasible ReRAM, with the smallest area, with the highest array density (10.1Mb/mm²) that is achieved in a 22nm FINFET technology. This work also demonstrates, that with the aid of the proposed PVC-WVW scheme and OC-CSA, a 10^{-5} BER across –40 - 125°C; the 5ns t_{SENSE} @0.7V is also better than prior-art.

References:

[1] Y. Hayakawa, et al., "Highly reliable TaOx ReRAM with centralized filament for 28-nm embedded application," *VLSI*, pp. 14-15, 2015.

[2] C-C. Chou, et al., "An N40 256K×44 embedded RRAM macro with SL-precharge SA and low-voltage current limiter to improve read and write performance," *ISSCC*, pp. 478-480, 2018.

[3] M-F. Chang, et al., "Low VDDmin Swing-Sample-and-Couple Sense Amplifier and Energy-Efficient Self-Boost-Write-Termination Scheme for Embedded ReRAM Macros Against Resistance and Switch-Time Variations," *JSSC*, vol. 50, no. 11, pp. 2786 – 2795, Nov. 2015.

[4] M. Ueki, et al., "Low-Power Embedded ReRAM Technology for IoT Applications," *VLSI*, pp. 108-109, 2015.

[5] M. Meterelliyoz, et al., "2nd generation embedded DRAM with 4X lower self refresh power in 22nm Tri-Gate CMOS technology," *VLSI*, 2014.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

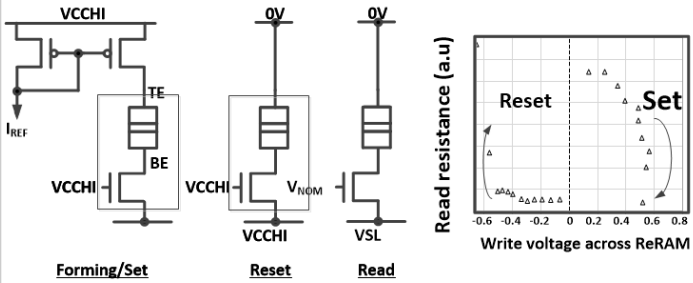


Figure 13.2.1: ReRAM operation polarity and single device measured R-V characteristic.

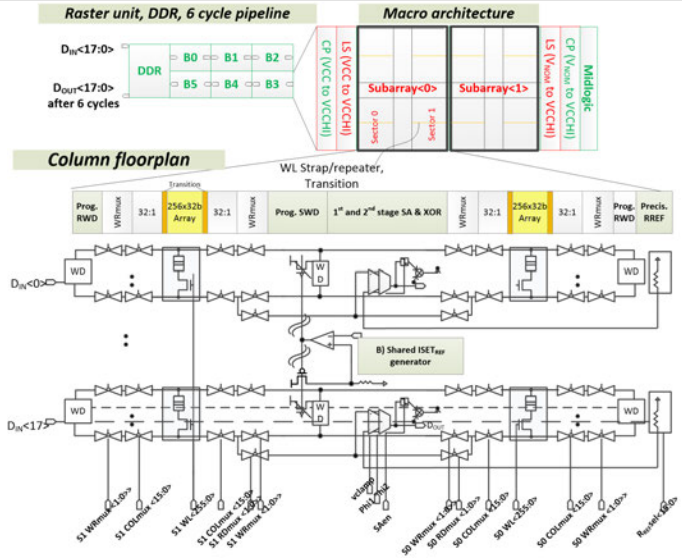


Figure 13.2.2: 3.6Mb chip architecture, focusing on the 0.3Mb subarray.

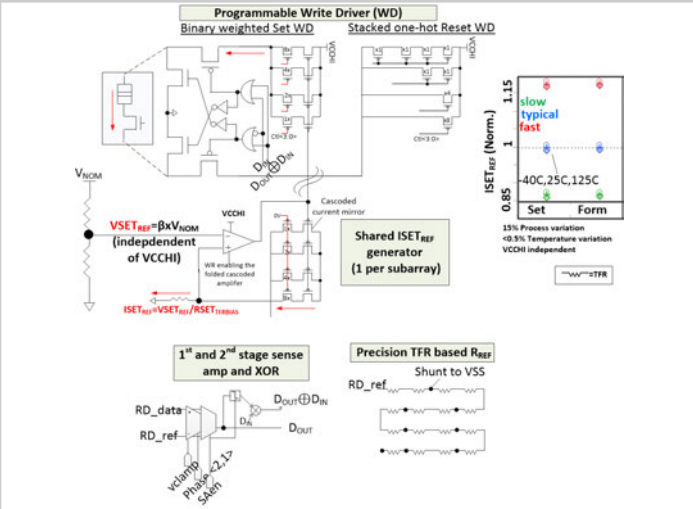


Figure 13.2.3: Key circuits including the Programmable Set and Reset Write driver, $ISET_{REF}$, XOR and precision TFR based R_{REF}. $ISET_{REF}$ variability for Forming and Set is shown on the top right.

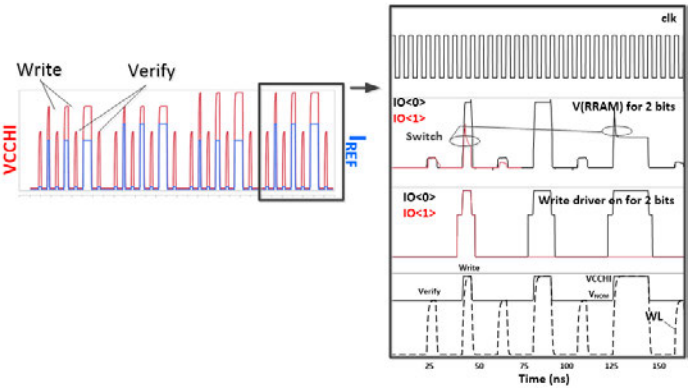


Figure 13.2.4: WVW scheme details and simulation of two example bits.

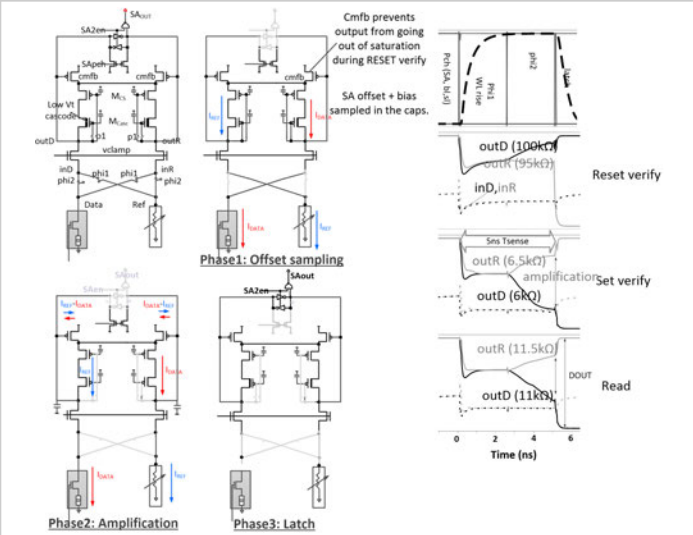


Figure 13.2.5: Two stage coupled SA scheme and example Read, Set and Reset-Verify simulations.

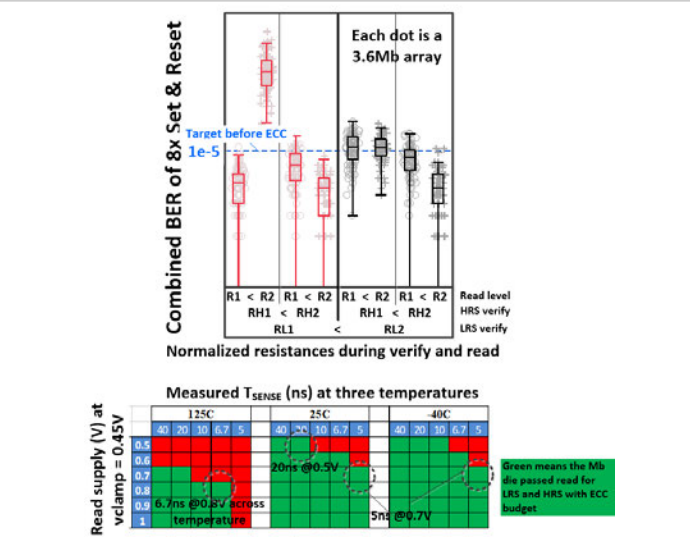


Figure 13.2.6: Mixed WVW sequence applied to get BER at different Verify and Read levels. Read Shmoo at three temperatures.

Process node	22nm low power FINFET
Density	10.1Mb/mm^2
Write time	<10µs
Read time @ 25C	<5ns @ 0.7V, <10ns @ 0.5V
Read voltage	<0.5V
Read energy/bit	<1pJ
Measured BER	<1e-5

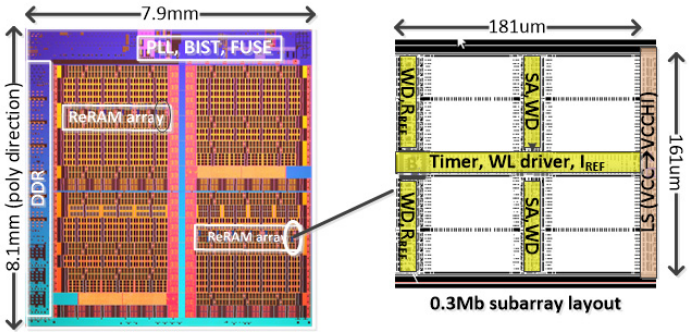


Figure 13.2.7: Design summary and die microphotograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

13.3 A 7Mb STT-MRAM in 22FFL FinFET Technology with 4ns Read Sensing Time at 0.9V Using Write-Verify-Write Scheme and Offset-Cancellation Sensing Technique

Liqiong Wei, Juan G. Alzate, Umut Arslan, Justin Brockman, Nilanjan Das, Kevin Fischer, Tahir Ghani, Oleg Golonzka, Patrick Hentges, Rawshan Jahan, Pulkit Jain, Blake Lin, Mesut Meterelliyoz, Jim O'Donnell, Conor Puls, Pedro Quintero, Tanaya Sahu, Meenakshi Sekhar, Ajay Vangapaty, Chris Wiegand, Fatih Hamzaoglu

Intel, Hillsboro, OR

STT-MRAM has been emerging as a very-promising high-density embedded non-volatile memory (eNVM) [1, 2]. Embedded Flash memory has been the leading eNVM technology, but STT-MRAM has been developed as a better solution for continuing scaling, speed and cost. This paper presents a write-verify-write (WvW) scheme and a programmable offset cancellation sensing technique that achieves a high-yield, high-performance and high-endurance 7Mb STT-MRAM arrays in a 22FFL FinFET technology [3]. The developed technology supports a wide range of operating temperatures between -40 - 105°C. Compared to prior-art [4,5], the two-stage current-sensing technique with a die-by-die tuning of a thin-film precision resistor that is used as a reference can significantly improve the sensing margin during verify and read operations. Read disturb for reference cells is eliminated as there is no MTJ in the reference path.

Figure 13.3.1 illustrates the organization of the 144KB STT-MRAM block. Each block has 4 subarrays sharing one charge pump circuit and control logic. The charge pump circuit can support a WL over-drive voltage during the write operation and an under-drive voltage at temperatures larger than 85°C. The 36KB subarray features 258b/BL and 288b/WL in a butterfly-array configuration. Subarray density, including the ECC bits, is 10.6Mb/mm^2. Each logical IO uses 32:1 column interleaving, with write drivers located on both sides of the 256-row column. The offset-cancellation sense amplifier with a programmable reference resistor (a thin-film precision resistor) is in the center of the column IO to balance the sensing margin of the two array sectors. Each 0.0486um^2 1T1MTJ bitcell is 216×225nm^2, with 2-polysilicon WLs, one Metal 4 BL, and one Metal 1 SL. To improve the WL slew rate, multiple straps are inserted to connect the two polysilicon wordlines to Metal 5. The cross-sectional TEM of the MTJ array embedded between Metal 2 and Metal 4 of a 9-Metal CMOS process is shown in Fig. 13.3.1. The tunnel-magneto-resistance ratio (TMR) of the MTJs is 180% at 25°C, with a target device critical dimension (CD) between 60-80nm [6].

Figure 13.3.2 shows the details of the write circuit using a programmable WvW scheme. WvW control signals are generated in mid-logic, where the write-to-verify time, the number of write pulses, the write pulse width, the write settings, and total write time can be programmed. The write current is determined by the write settings. The center write driver is located in the column IO, which has both strong- and weak-write settings, while the edge write driver only has strong settings to save area.

Figure 13.3.3(a) shows the relative write current versus the available 16 settings, and the maximum current at different WL locations. Writing from both sides of the array can significantly reduce the voltage drop due to interconnect resistance; therefore, the write current is observed to be independent of the row address. Figure 13.3.3(b) illustrates the WvW timing diagram: during verify, the read D_{OUT} is compared with the D_{IN} to be written, upon match the verify passes and the corresponding write driver is disabled; otherwise, it is enabled along with a boosted WL. In the example presented in Fig. 13.3.3(b), the first verify has a mismatch result; hence, the first write is attempted with a lower write current, but in the example, the cell is not switched. As a consequence, the 2nd verify fails and the circuit attempts a 2nd and stronger write pulse. Due to the stronger write current, the cell in this example is switched from a high-resistance state (AP-state) to a low-resistance state (P-state); this results in passing the 3rd verify and the write driver being disabled for the 3rd write.

Figure 13.3.4(a) shows the read-sensing circuit, with the offset-cancellation technique and its read-timing diagram (see Figure 13.3.4(b)). First, BL and SL are equalized and pre-discharged to 0V, and the sense amplifier output is charged to V_{CC}. Then, the WL is driven high for one cycle. When the temperature is higher than 85°C, WL underdrive is enabled to minimize the impact of unselected bitcell leakage to the current sensing margin. The first phase is evaluation, where PH1 is ON and PH2 is OFF. SL remains at 0V and BL is clamped low enough to minimize read disturb. C_{G0} and C_{G1} are gate capacitors. Voltage for C_{G1} is determined by the bitcell current and voltage for C_{G0} is given by the current of the reference resistor (R_{REF}). If the bitcell is in a high-resistance state (AP-state) and the cell current is smaller than reference resistor, $V_{CG1} > V_{CG0}$. The second phase is for amplification, where PH2 is ON and PH1 is OFF. Since the AP-state bitcell's current is weaker and the PMOS load is stronger due to a lower V_{CG0}, V_{OUT} goes high. Given that the same pair of PMOS load and NMOS clamp transistors are used during the two phases, the mismatch between data and reference branches, due to process variation, is cancelled and the sensing margin is significantly improved. Finally, SAE turns ON and the outputs of the final sense amplifier are rectified to supply levels. The reference resistor is made using a thin-film precision resistor, which is less sensitive to process variation. Die-by-die programming is utilized to optimize the reference resistance for maximum yield. A smaller address space is searched, with coarse and fine tunings at hot and cold temperatures, to find the optimal reference resistance values, within a limited test time. The bit error rate (BER) versus the deviation from the picked optimum R_{REF} setting for a given die is shown in Figure 13.3.4(c); showing that the selected R_{REF} setting is optimal and there is enough sense margin for both hot and cold temperatures.

Combining the read and write design techniques described above, the yield for MB-size arrays is significantly enhanced. Figure 13.3.5(a) shows that the resulting yield rate can be >99.997%, across temperature, at a supply voltage of 1.1V; this meets the ECC design target BER. The voltage-read sensing time (t_{SEN}) shmoo for 7Mb STT-MRAM arrays tested at various temperatures is shown in Figure 13.3.5(b). The design achieves a 4ns read sensing time at 0.9V. The array is also tolerant of a wide supply power range, down to 0.6V with an 8ns read sensing time.

Figure 13.3.6(a) provides the BER after 10^6 cycles of write with and without the WvW design technique, as well as the read disturb rate at various clamp voltages (V_{CLAMP}) (see Fig. 13.3.6(b)). No endurance driven errors are detected using the WvW design technique and no read disturb is observed after 10^{12}-read cycles at the optimal clamping voltages.

Figure 13.3.7 shows the die photograph of the test chip with eight 7Mb perpendicular-STT-MRAM arrays (56Mb/die total). The chip is fabricated in a 22FFL FinFET technology. The chip also implements an SRAM array, a PLL, an eFuse, a BIST and DDR IOs, showing that the integration of STT-MRAM has no impact on the functionality or yield of these blocks.

References:

[1] Y. Lee, et al., "Embedded STT-MRAM in 28-nm FDSOI Logic Process for Industrial MCU/IoT Application," *VLSI*, pp. 181-182, 2018.
[2] Y. J. Song, et al., "Highly Functional and Reliable 8Mb STT-MRAM Embedded in 28nm Logic," *IEDM Tech. Dig.*, pp. 663-666, 2016.
[3] B. Sell, et al., "22FFL: A High Performance and Ultra Low Power FinFET Technology for Mobile and RF Applications," *IEDM Tech. Dig.*, pp. 686-688, 2017.
[4] Y. Shih, et al., "Logic Process Compatible 40nm 16Mb, Embedded Perpendicular MRAM with Hybrid-Resistance Reference, sub-uA Sensing Resolution, and 17.5ns Read Access Time," *VLSI*, pp. 79-80, 2018.
[5] Q. Dong, et al., "A 1Mb 28nm STT-MRAM with 2.8ns Read Access Time at 1.2V VDD Using Single Cap Offset Cancelled Sense Amplifier and In-situ Self-Write-Termination," *ISSCC*, pp. 480-481, 2018.
[6] O. Golonzka, et al., "MRAM as Embedded Non-Volatile Memory Solution for 22FFL FinFET Technology," *IEDM Tech. Dig.*, 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

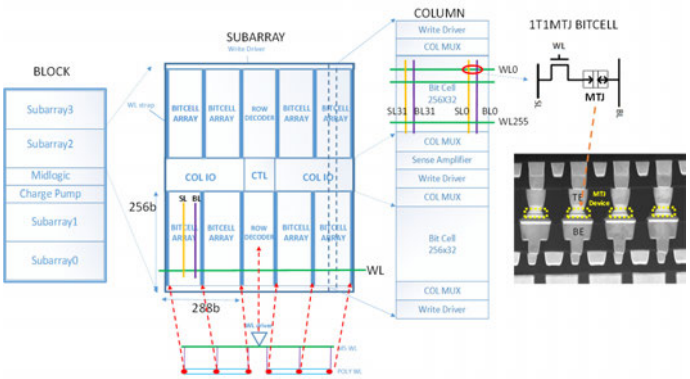


Figure 13.3.1: STT-MRAM block, subarray, column and bitcell diagrams with TEM of MTJ array.

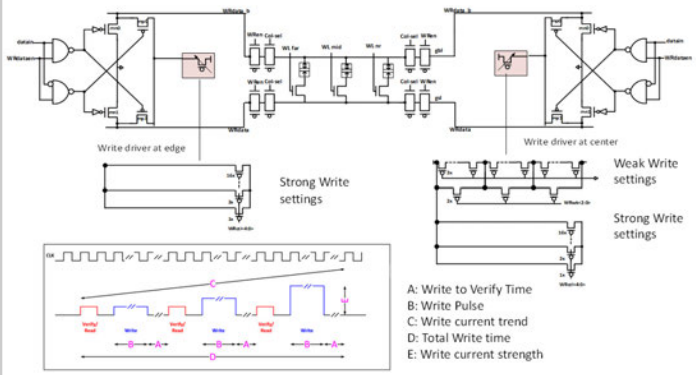


Figure 13.3.2: Write circuit with programmable write verify write scheme.

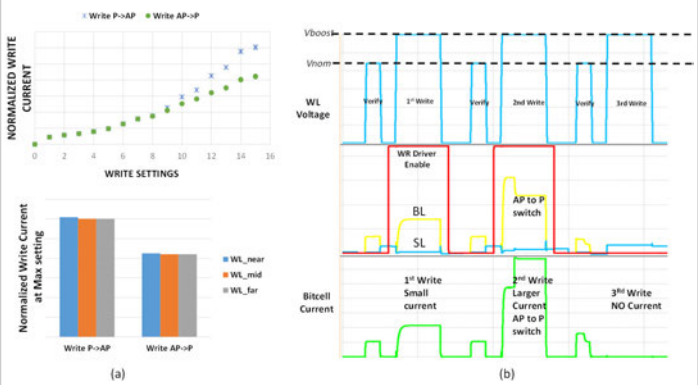


Figure 13.3.3: Write current at different settings and WLs, and WvW timing diagram.

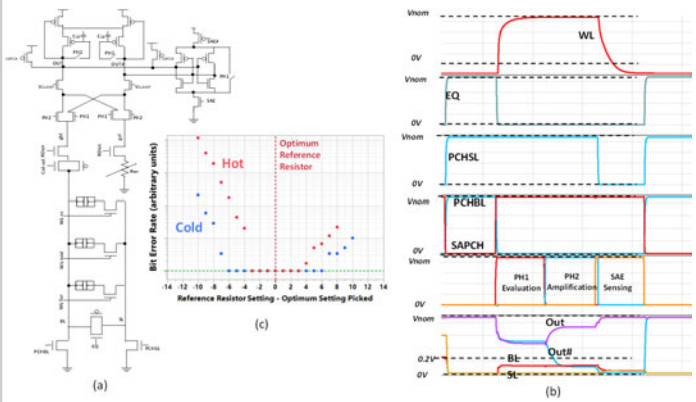


Figure 13.3.4: Read Sensing Circuit and Read Timing Diagram with BER at the deviation of optimal R_{REF} setting picked for a given Die.

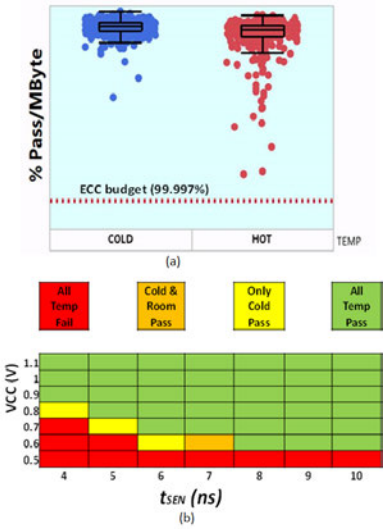


Figure 13.3.5: Bit Yield at 1.1V for Hot and Cold and Read Shmoo plot for 7Mb Array with ECC enabled.

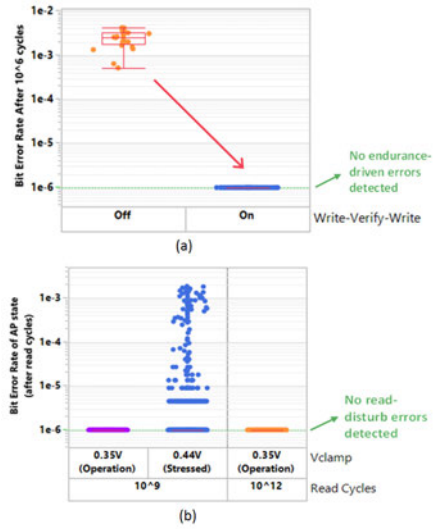


Figure 13.3.6: Write Endurance with and without WvW, and Read disturb at different clamp voltages.

Technology	22FFL FinFET Technology
Memory	Perpendicular STT-MRAM
TMR	>180%@ 25C
Cell type	1T1MTJ
Cell size	$0.0486\mu m^2$
Capacity	7Mb
Subarray Density (Incl. ECC bits)	10.6 Mbits/mm^2
Read Sense Time	4ns@ 0.9V, 8ns@ 0.6V
Bit Yield	> 99.997%
Retention	200C 10 years
Write Endurance	>1E06
READ Disturb	>1E12
Temp Range	-40C to 105C

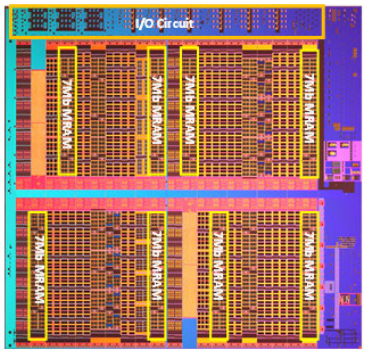


Figure 13.3.7: Die Photo and Summary Table.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

13.4 A 512Gb 3-bit/Cell 3D 6th-Generation V-NAND Flash Memory with 82MB/s Write Throughput and 1.2Gb/s Interface

Dongku Kang, Minsu Kim, Su Chang Jeon, Wontaeck Jung, Jooyong Park, Gyosoo Choo, Dong-kyo Shim, Anil Kavala, Seung-Bum Kim, Kyung-Min Kang, Jiyoung Lee, Kuihan Ko, Hyun-Wook Park, Byung-Jun Min, Changyeon Yu, Sewon Yun, Nahyun Kim, Yeonwook Jung, Sungwhan Seo, Sunghoon Kim, Moo Kyung Lee, Joo-Yong Park, James C. Kim, Young San Cha, Kwangwon Kim, Youngmin Jo, Hyunjin Kim, Youngdon Choi, Jindo Byun, Ji-hyun Park, Kiwon Kim, Tae-Hong Kwon, Youngsun Min, Chiweon Yoon, Youngcho Kim, Dong-Hun Kwak, Eungsuk Lee, Wook-ghee Hahn, Ki-sung Kim, Kyungmin Kim, Euisang Yoon, Won-Tae Kim, Inryoul Lee, Seung hyun Moon, Jeongdon Ihm, Dae Seok Byeon, Ki-Whan Song, Sangjoon Hwang, Kye Hyun Kyung

Samsung Electronics, Hwasung, Korea

Data storage is one of the hottest discussion topics in today's connected world. The amount of data growth is expected to be exponential, while budget and space remain constricted. Since the transformation of storage device from planar NAND to 3D V-NAND [1], the areal density of semiconductor storage devices has continuously evolved and has surpassed the density of magnetic hard drives. By providing the largest storage capacity in the smallest footprint, 3D V-NAND has been leading the data center revolution in recent years. However, 3D-technology scaling faces several technical challenges [2]. (1) As the number of WL stacks increases the channel-hole etch process becomes a limit, since the total WL-mold height increases. (2) Interference between cells increases since the distance between WLs becomes smaller. (3) Faster data transfer speeds are required to support higher IO bandwidth.

This paper presents a 512Gb, 3b/cell, flash memory featuring the 6th generation V-NAND technology, which integrates more than 120-WL layers. An 16kB two-plane cell array with 8kB sub-plane architecture is implemented. Read and write performance is improved to 820 and 82MB/s. To achieve this it was critical to minimize the WL and BL setup time for read and verify (during programming) operations. Also, due to the increased number of WL layers, variation in program and erase characteristics, as well as power dissipation became more serious issues. Therefore, a customized V_{th} window management scheme and a low-power sensing scheme are implemented. Finally, to meet the 1.2Gb/s IO performance target, a channel training scheme and an enhanced on-die termination schemes were developed. Figure 13.4.1 summarizes the key features of this work.

To improve read and write throughput the BL precharge time should be minimized. As shown in Fig. 13.4.2(a) (1) V_1 is applied to BLSHF to force V_{BL1} to BL as a first step; (2) V_2 is applied to BLSHF to force V_{BL2} to BL as shown in Fig 13.4.2(b). To accelerate BL precharge, we need to ensure that $V_1 > V_2$ to keep $V_{BL1} > V_{BL2}$. However, when a V_1 to V_2 transition occurs, the sudden disconnect from the page buffer to BL across BLSHF can be observed when the BL voltage near the page buffer is higher than V_2-V_t(BLSHF). This degrades the overall precharge performance because the BL is not precharged during that period. To solve this problem, enhanced BL precharge scheme is proposed as shown in Fig. 13.4.2(c). In this scheme, rather than applying V_2 as a step function from V_1, V_2 is applied gradually to promote a continuous current flow in any situation across BLSHF, thereby minimizing the BL precharge time.

For a 3b/cell V-NAND, a program pulse is followed by seven times of verify operations to form the seven distinctive V_{th} states. The verify operations are a large overhead for a program operation, since the program pulse time is typically far less than the verify time for a programming cycle. As shown in Fig. 13.4.3(a), when a verify (or target) voltage is applied to WL_N, WL_{N+1} and WL_{N-1} are subjected to the V_{READ} voltage just like the other unselected WLs. Therefore, from WL_N's perspective, a couple of capacitances exist looking towards the upper and lower WL directions. In general, WL_{N+1} always maintains an erase state when WL_N is programmed and verified. By utilizing this principle, we can apply a lower voltage than V_{READ} to WL_{N+1} without affecting the sensing current for the verify operation.

In this work, to minimize the capacitive coupling between WL_N and WL_{N+1}, we apply a copy of the selected WL voltage to WL_{N+1}, as shown in Fig. 13.4.3(b). This technique improves the programming throughput by reducing the WL setup time for the verify operations.

An increase in the number of WL stacked layers results in degraded channel-hole variation [3], as shown in Fig. 13.4.4(a & b). Hence, to minimize this effect, an adaptive erase scheme is used to minimize erase V_{th} variation, as shown in Fig. 13.3.4(c). Generally, a cell's reliability to repetitive erase-and-program operations is inversely proportional to a cell's V_{th} window. Hence, the V_{th} distance between an erase state and the highest programed state is required to be constant, since all WLs should be able to handle the same amount of erase-and-program cycles. Hence, to achieve balanced reliability characteristic for all WLs, we apply a progressive V_{th} window scheme as shown in Fig. 13.4.4(d).

Power consumption is another imminent issue for increasing WL stacks. As shown in Fig. 13.4.5(a), pre-pulse operation is required for every sensing operation to initialize a channel-hole potential. Otherwise, the channel voltage might fluctuate, depending on the situation, due to the accumulated voltage when consecutive sensing operations are performed. When the accumulated channel voltage exceeds a certain level, it might also introduce V_{th} disturbance for cells. However, during a pre-pulse operation, the WL to channel-hole capacitance is instantaneously inflated multiple times, since the WL-channel capacitances for all SSLs (shown in Fig. 13.4.5(b)) become effective. This increases power consumption. (Fig. 5(b)) To overcome this excessive power consumption, we developed a random pre-pulse sensing scheme, which is shown in Fig. 13.4.5(c). This scheme reduces the occurrence of pre-pulse operations to lower the average current consumption. Also, sporadic execution of pre-pulse operations prevents channel voltage accumulation.

To support a 1.2Gb/s toggle 4.0 double-data rate, it was necessary to overcome pin-to-pin load mismatch and read-clock duty-cycle error. Hence, the input and output interface is implemented with read and write DQ training and duty-cycle correction (DCC) training circuits. For instance, write DQ training improves t_{DSH} (data setup and hold time margin) by delaying each pin at the host controller according to its load mismatch, as shown in Fig. 13.4.6(a). Similarly, read training compensates t_{DVW} (data valid window) by adjusting the input strobe for each pin in the controller, as shown in Fig. 13.4.6(b). When training the controller uses a designated data pattern for V-NAND and reads this pattern from the V-NAND to perform validation. In addition to the read and write training, an explicit and implicit DCC improves t_{DVW} by correcting the read clock's duty cycle error, which is caused by channel losses, as shown in Fig. 13.4.6(c). Explicit DCC training is performed with a specific command sequence, whereas implicit DCC training is performed during normal read operations. Finally, to enhance signal integrity (SI) for a multi-die-stack package, two types of on-die-termination (ODT) schemes are supported, as shown in Fig. 13.4.6(d). In the first (designated) ODT scheme, designated chips (CHIP2 and CHIP4) of different packages are used for termination when the nODTx signal is enabled. However, in the second (hybrid-CE) scheme, the termination is enabled with the chip enable signal (nCEx). These termination techniques provide additional options to achieve better SI with lower termination power. The termination modes (read or write) are detected by reading the read clock (nREx) with nODTx signal.

References:
[1] J. Jang, et al., "Vertical Cell Array using TCAT (Terabit Cell Array Transistor) Technology for Ultra High Density NAND Flash Memory," *IEEE VLSI*, pp. 192-193, 2009.
[2] D. Kang, et al., "256 Gb 3 b/Cell V-nand Flash Memory With 48 Stacked WL Layers," *JSSC*, vol. 52, no. 1, pp. 210–217, Jan. 2017.
[3] H. Kim, et al., "Evolution of NAND Flash Memory: From 2D to 3D as a Storage Market Leader," *IEEE Intl. Memory Workshop*, 2017.
[4] C. Kim, et al., "A 512-Gb 3-b/Cell 64-Stacked WL 3-D-NAND Flash Memory," *JSSC*, vol. 53, no. 1, pp. 124-133, Jan. 2018.
[5] H. Maejima, et al., "A 512Gb 3b/Cell 3D Flash Memory on a 96-Word-Line-Layer Technology," *ISSCC*, pp. 336-337, 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

	ISSCC 2018 [5]	ISSCC 2019
Bits per cell	3	3
Density	512Gb	512Gb
Pagesize	16 KB/Page	16 KB/Page
I/O Bandwidth	Max. 1Gb/s	Max. 1.2Gb/s
tBERS	3.5ms (Typ.)	3.5ms (Typ.)
tR	58us	45us
Program Throughput	57MB/s	82MB/s
Vcc	2.30V to 3.6V	2.35V to 3.6V
Vccq	1.8V	1.2V

Figure 13.4.1: Feature Summary.

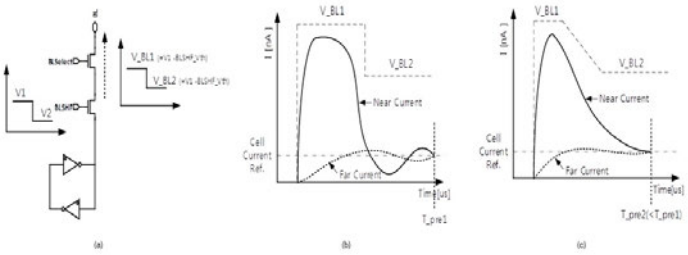


Figure 13.4.2: Enhanced BL precharge scheme.

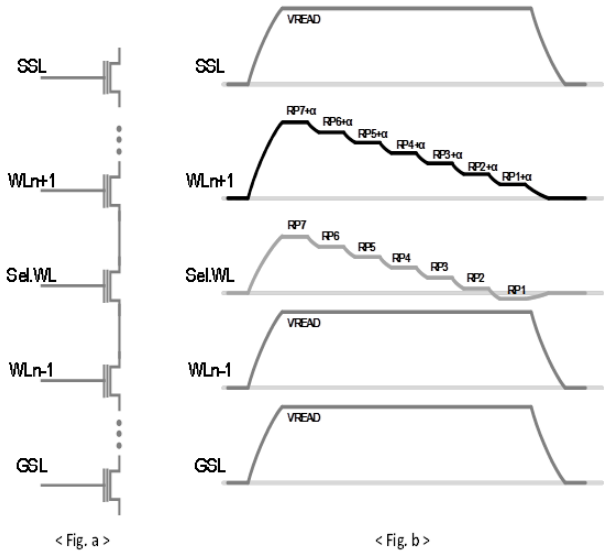


Figure 13.4.3: Couple-capacitance-minimizing technique.

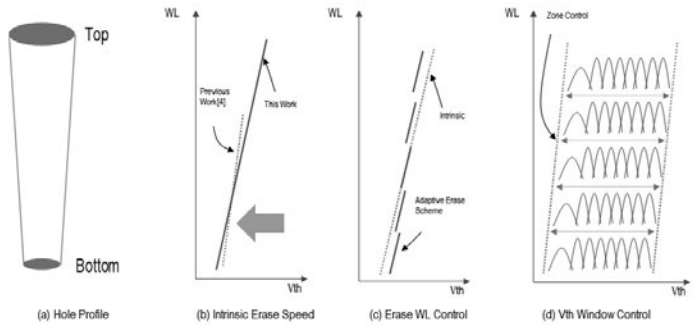


Figure 13.4.4: Progressive V_{th} window scheme.

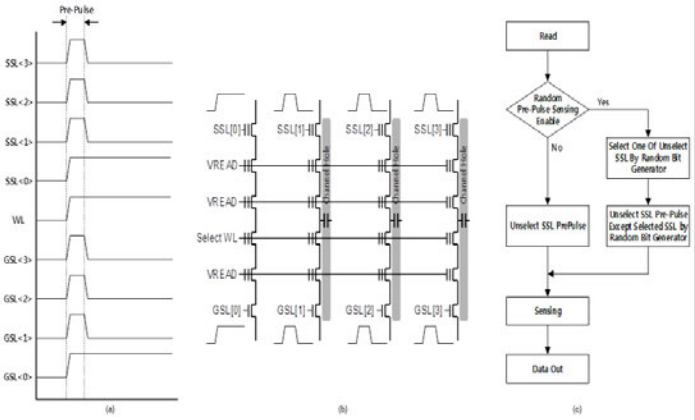


Figure 13.4.5: Random pre-pulse sensing scheme.

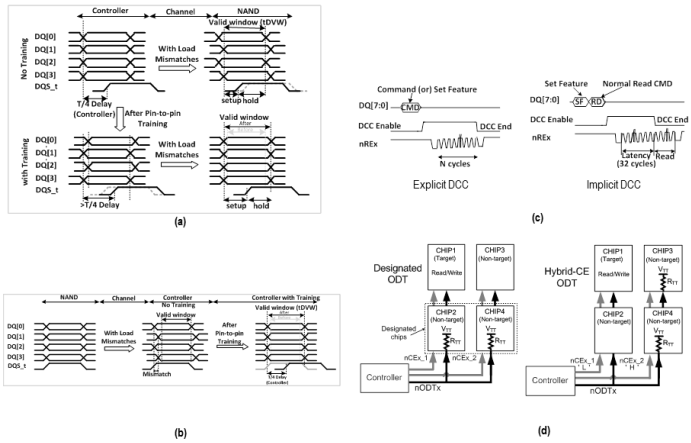


Figure 13.4.6: Toggle 4.0 (a) write training (b) write training (c) Explicit & Implicit DCC Training (d) designated and hybrid-CE ODT.

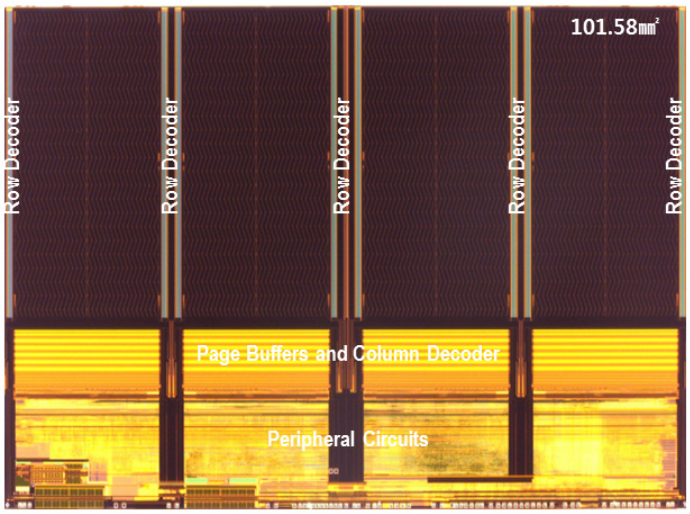


Figure 13.4.7: Die photograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

13.5 A 512Gb 3-bit/Cell 3D Flash Memory on 128-Wordline-Layer with 132MB/s Write Performance Featuring Circuit-Under-Array Technology

Chang Siau[1], Kwang-Ho Kim[1], Seungpil Lee[1], Katsuaki Isobe[2], Noboru Shibata[2], Kapil Verma[1], Takuya Ariki[1], Jason Li[1], Jong Yuh[1], Anirudh Amarnath[1], Qui Nguyen[1], Ohwon Kwon[1], Stanley Jeong[1], Heguang Li[1], Hua-Ling Hsu[1], Tai-yuan Tseng[1], Steve Choi[1], Siddhesh Darne[1], Pradeep Anantula[1], Alex Yap[1], Hardwell Chibvongodze[1], Hitoshi Miwa[1], Minoru Yamashita[1], Mitsuyuki Watanabe[1], Koichiro Hayashi[1], Yosuke Kato[1], Toru Miwa[1], Jang Yong Kang[1], Masatoshi Okumura[1], Naoki Ookuma[1], Muralikrishna Balaga[1], Venky Ramachandra[1], Aki Matsuda[1], Swaroop Kulkani[1], Raghavendra Rachineni[1], Pai K. Manjunath[1], Masahito Takehara[1], Anil Pai[1], Srinivas Rajendra[1], Toshiki Hisada[2], Ryo Fukuda[2], Naoya Tokiwa[2], Kazuaki Kawaguchi[2], Masashi Yamaoka[2], Hiromitsu Komai[2], Takatoshi Minamoto[2], Masaki Unno[2], Susumu Ozawa[2], Hiroshi Nakamura[2], Tomoo Hishida[2], Yasuyuki Kajitani[2], Lei Lin[1]

[1]Western Digital, Milpitas, CA
[2]Toshiba Memory, Yokohama, Japan

Advancements in 3D-Flash memory-layer-stacking technology has enabled density scaling that circumvents the lithography limitations which have prevented 2D-NAND Flash memory from scaling [1]. Bit densities as high as 5.95Gb/mm² on a single die were recently reported [2], where a 512Gb NAND Flash was built on 96 layers of memory. As memory density increases, with memory layers increasing from 96 layers to 128 layers, higher bit density can be achieved by adopting larger capacity die; however, NAND performance per bit density is reducing with the 2-plane architecture. In this work, we propose a 512Gb 3b/cell 128WL-layer NAND Flash, with a bit density of 7.80Gb/mm²: a 31% improvement over the previously reported. Three key performance improving technologies have been implemented. (1) A 4-plane architecture with circuit under array (CUA) technology to improve performance per bit density. (2) A multi-die peak-power management (PPM) system to manage peak-power consumption in the system, via the ZQ pin. (3) A 4KB-page-read mode to reduce power consumption. Figure 13.5.1(a) summarized the key features and Fig. 13.5.1(b) shows the die photograph and the floorplan for this work. Figure 13.5.2 shows a table comparing this work to previous work.

Previous work [3] has reported a CUA architecture, which divides the array into 32 tiles and requires higher array interconnect overhead. In this work, the array is arranged into 4 planes, thereby minimizing array interconnect. Each plane consists of 683 blocks plus spare blocks, each 24MB in size. Each block has 1536 pages, each 16KB in size. Planes are divided in the BL direction to reduce BL length. The WL length is kept at 16KB length using the even/odd combined decoding structure [4]. Block selection signals are routed on top of array to WL-transfer gates, which reside under WL hookup Area. Connections are made by contacts outside of the array, resulting in no die area impact. In a traditional implementation, four planes of NAND memory will incur an area penalty of 15% due to duplication of the sense amplifier and data latches (SADL). In a CUA implementation, SADL will be hidden under the array, which reduces the area impact to below 1%. With the BL length cut in half, the BL *RC*-component is reduced by 4×, reducing the BL settling time by 31%. In a 3b/cell technology, program-verify operations occupy 60% of the total programming time (t_{PROG}). Figure 13.5.3(a) shows that with a reduced BL settling time, we estimate t_{PROG} to be improved by 16%. At the same time, 4-plane parallel operation will boost performance by another 100%, compared to a 2-plane device. We measured programming performance at 132MB/s with four planes operating in parallel. Figure 13.5.3(b) shows the normalized fail bit count (FBC) vs the average read latency (t_R). This work achieves a 56μs t_R.

The 4-plane architecture increases performance, but it also has a draw back with respect to peak power consumption. The peak current increases by 100% due to the increased WL and sense amplifier loading. For a system using multiple NAND die, the simultaneous peak power consumption by the multiple die can cause system power supply to droop and cause functional failure. A traditional system-level approach is to assume a peak-current consumption for the duration of t_{PROG}, and that a limited number of NAND die are activated for parallel operations. Typically, the peak power consumption is 3.5× higher than the average power consumption, and is only active for less than 10% of the total operation period. Command-based peak-power management was proposed [5] to resolve this issue, but this approach increases system management overhead by requiring the system to monitor the status and to issue stop/resume commands. In this work, we propose a multi-die peak-power management (PPM) scheme utilizing a shared ZQ-pad connection. This allows the NAND die in a system to schedule peak-power operations without involving the system management, as shown in Fig. 13.5.4. This proposal skews the sub-operation of NAND die such that peak-power consumption does not occur at the same time. Once peak-power consumption between different die is skewed, the system will be able to fully utilize the available power without considering the possibility of peak-power collisions among multiple die in the system, which would result in power droops. Figure 13.5.5 shows an example of a 3-die system with skewed peak-power, which allows for higher average power consumption while lowering the peak-power consumption. Based on the operation current consumption profile, we define an index of current consumption for each sub-operation of a program, read, or erase operation. This index is then translated by the state machine into a ZQ pin pull-up strength. Before each sub-operation begins on a NAND die, the state machine will issue a current consumption code to the ZQ pin and increase the ZQ pin pull-up strength. At the same time, a polling operation commences to sense the voltage on the ZQ pin. If the voltage on the ZQ pin, which accumulates the pull-up current from all of the NAND die in the system, is above a pre-set level the sub-operation will be postponed. If the voltage on the ZQ pin is lower, then the total peak current consumed in the system is within limits and the sub-operation of the polling NAND die will commence. In the event that there are multiple die polling at the same time, a randomized delay is used to reschedule polling, as shown in Fig. 13.5.6; indefinite sub-operation postponing can be avoided with this scheme. PPM allows the system to enable all NAND die to operate at the same time without introducing power droops due to peak power consumption from multiple die in the system.

To further reduce power consumption, we introduce 4KB-page read operation. As opposed to [3], where a 4KB read is achieved by enabling 2 tiles in tile array; in this work the array is arranged into planes and additional transistors are added in the sense amplifier to facilitate consecutive 4KB selection, as shown in Fig. 13.5.7. Each 4KB-page will be pre-charged and sensed at the same time using the all-BL (ABL) sensing method. Two transistors have been added to facilitate this operation: the BLC4K transistor for BL selection and the NLO4K transistor for BL discharging. For a selected 4KB BL, BLC4K will be turned on, thus biasing BL to the BL bias voltage. The unselected 12KB BLs, the BLC4K transistor will be turned off while the NLO4K transistor will be turned on to bias the unselected BLs to the cell source level. Average current consumption is reduced by 40% compared to a 16KB read. To prevent a BL on the 4KB boundary from becoming the timing bottleneck in a 4KB read, the control of BLC4K and NLO4K are designed such that there is always 16-BL past the 4KB boundary that are biased to the BL bias voltage, thus ensuring that all 4KB BLs see the same pre-charge conditions.

Acknowledgements:
The authors would like to thank the design, layout, verification CAD, device, test, and process teams, as well as Kazuaki Isobe and Farookh Moogat for their support.

References:
[1] H. Tanaka, et al., "Bit Cost Scalable Technology with Punch and Plug Process for Ultra High Density Flash Memory," *VLSI*, pp.14-15, 2007.
[2] H. Maejima, et al., "A 512Gb 3b/Cell Flash Memory on a 96-Word-Line-Layer Technology," *ISSCC*, pp. 336-337, 2018.
[3] T. Tanaka, et al., "A 768Gb 3b/cell 3D-Floating-Gate NAND Flash Memory," *ISSCC*, pp. 142-144, 2016.
[4] R. Yamashita, et al., "A 512Gb 3b/cell flash memory on 64-word-line-layer BiCS technology," *ISSCC*, pp. 196-197, 2017.
[5] M. Sako, et al., "A Low-Power 64Gb MLC NAND-Flash Memory in 15nm CMOS Technology," *ISSCC*, pp. 128-129, 2015.
[6] S. Lee, et al., "A 1Tb 4b/Cell 64-Stacked-WL 3D NAND Flash Memory with 12MB/s program throughput," *ISSCC*, pp. 340-341, 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Features	
Capacity and die size	512Gb, 66mm2
Bit density	7.80Gb/mm2
Organization	3b/Cell (16KB + ECC) / Page, 1536 Pages / Block, (683 + EXT) Blocks / Plane, 4 Planes
tR Read	56us
Program Performance	132MB/s
I/O	1.066Gbps DDR, X8
Power Supply	VCC: 2.3V to 3.6V VCCQ: 1.2V, 1.8V

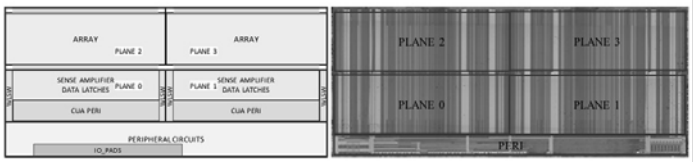


Figure 13.5.1: a) Key features (top). b) Chip Floorplan and Die photo (bottom).

ISSCC	This work	[2]	[6]	[4]
Technology	128 WL Layers	96 WL Layers	64 WL Layers	64 WL Layers
Capacity	512Gb	512Gb	1Tb	512Gb
# of bit/cell	3	3	4	3
# of planes	4	2	2	2
Program Performance	132MB/s	57MB/s	12MB/s	55MB/s
Die Size mm²	66	86	182	132
Bit density Gb/mm²	7.80	5.95	5.63	3.88

Figure 13.5.2: Performance and bit density comparison with previous works.

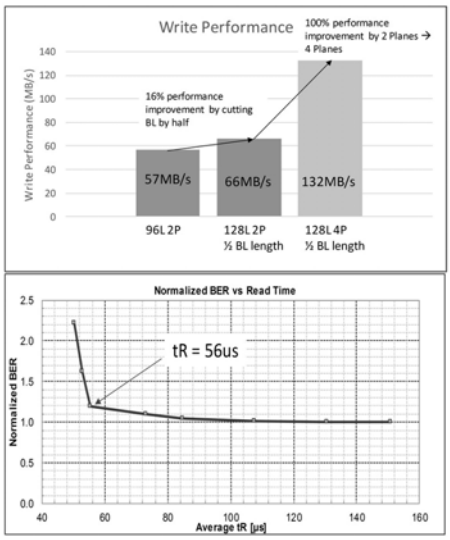


Figure 13.5.3: a) Write performance (top). b) Normalized Fail Bit count vs. t_R (bottom).

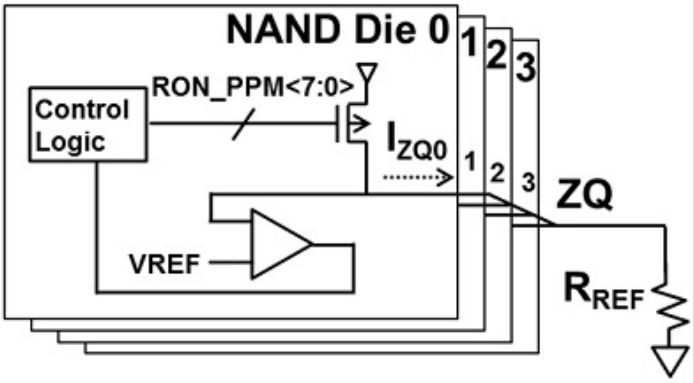


Figure 13.5.4: Multi-Die Peak Power Management scheme utilizing comparator used for ZQ calibration.

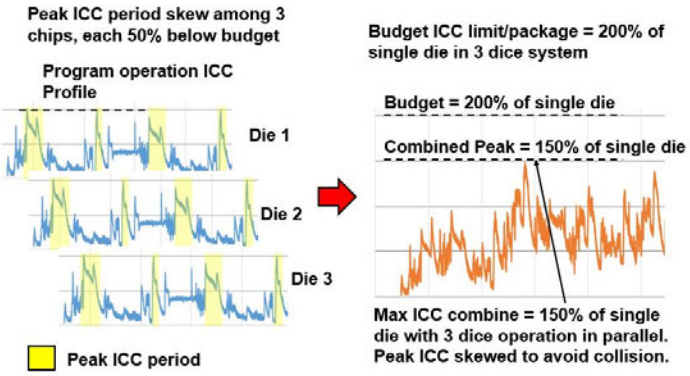


Figure 13.5.5: Peak Power Management scheme skew peak current consumption period to avoid peak current collision. Peak power budget can be fully utilized.

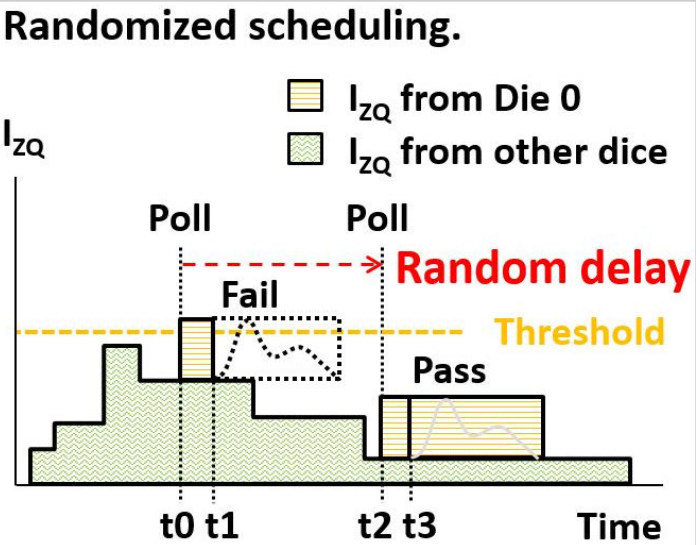


Figure 13.5.6: ICC polling with random delay.

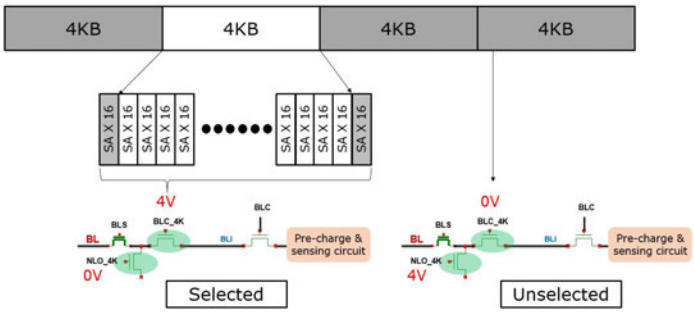


Figure 13.5.7: BL biasing condition for BLC4K and NLO4K for selected and unselected case. Edge words are biased at the same condition as selected word.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 14 Overview:
Machine Learning and Digital LDO Circuits
DIGITAL CIRCUITS SUBCOMMITTEE

Session Chair: ***Vivek De***
Intel, Beaverton, OR

Associate Chair: ***Ping-Ying Wang***
CMOS-Crystal Technology, Zhubei City, Hsinchu County, Taiwan

Subcommittee Chair: Edith Beigné, Facebook, Menlo Park, CA

In this session, seven papers highlight developments in machine learning and digital low-dropout (LDO) linear regulators. The papers demonstrate a hybrid digital and mixed-signal computing platform for swarm robotics, bi-directional memory delay lines to perform time-domain MAC operations, hybrid in-/near-memory compute SRAM and resistive RAM for/with resilience techniques. The digital LDO papers present a computational regulation scheme, a sub-nA wide-dynamic-range implementation and a universal modular hybrid LDO in 14nm CMOS.

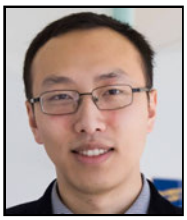
1:30 PM

14.1 A 65nm 1.1-to-9.1TOPS/W Hybrid-Digital-Mixed-Signal Computing Platform for Accelerating Model-Based and Model-Free Swarm Robotics

N. Cao, Georgia Institute of Technology, Atlanta, GA

In Paper 14.1, Georgia Institute of Technology presents a 1.1-to-9.1TOPS/W unified computation platform fabricated in 65nm CMOS for swarm robotics. It demonstrates an energy-efficient hybrid-digital-mixed-signal (HDMS) circuit architecture, to enable energy-resolution scalability across a variety of swarm algorithms and swarm sizes. The chip consumes 0.3 to 3.4μW and occupies 2mm^2 area.

2:00 PM



14.2 A Compute SRAM with Bit-Serial Integer/Floating-Point Operations for Programmable In-Memory Vector Acceleration

J. Wang, University of Michigan, Ann Arbor, MI

In Paper 14.2, the University of Michigan introduces a general-purpose hybrid in-/near-memory compute SRAM (CRAM) that combines an 8T transposable bit cell with vector-based bit-serial in-memory arithmetic to accommodate a wide range of bit-widths, from single to 32 or 64b, and operation types, including integer and floating point addition, multiplication and division, providing the flexibility and programmability necessary for evolving software algorithms, ranging from neural networks to graph and signal processing. The energy efficiency is 0.55TOPS/W for multiply and 5.27TOPS/W for addition.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

2:30 PM

14.3 A 43pJ/Cycle Non-Volatile Microcontroller with 4.7μs Shutdown/Wake-up Integrating 2.3-bit/Cell Resistive RAM and Resilience Techniques

T. F. Wu, Stanford University, Stanford, CA

In Paper 14.3, Stanford University demonstrates a non-volatile microcontroller through on-chip monolithic integration of Resistive RAM and silicon CMOS. It achieves fine-grained temporal power gating (average transition of 4.7μs active to shutdown, 200ns back); average active energy of 43pJ/cycle; multiple bits-per-cell RRAM in a complete chip. New resilience techniques achieve a 10-year lifetime of continuous inference. The chip area is 2.5×4.5mm^2.

3:15 PM

14.4 All-Digital Time-Domain CNN Engine Using Bidirectional Memory Delay Lines for Energy-Efficient Edge Computing

A. Sayal, University of Texas, Austin, TX

In Paper 14.4, the University of Texas, Austin describes a 12.08TOPS/W energy-efficient CNN engine implemented in a 40nm CMOS featuring bi-directional memory delay lines to perform time-domain MAC operations with multi-precision kernel weight and pooling support. The design is all-digital without requiring any capacitors, A/D converters, and supports near-threshold voltage operation and 16× performance boost with 4 input encoding modes. The chip consumes 31μW and occupies 0.124mm^2 area.

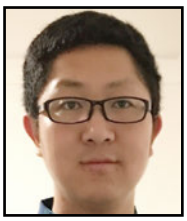
3:45 PM

14.5 A 0.6-to-1.1V Computationally Regulated Digital LDO with 2.79-Cycle Mean Settling Time and Autonomous Runtime Gain Tracking in 65nm CMOS

X. Sun, University of Washington, Seattle, WA

In Paper 14.5, the University of Washington describes computational power regulation, exploiting computing for high-speed regulation. Measurements indicate a 2.79-cycle mean settling time under nominal conditions.

4:15 PM



14.6 A 745pA Hybrid Asynchronous Binary-Searching and Synchronous Linear-Searching Digital LDO with 3.8×10^5 Dynamic Load Range, 99.99% Current Efficiency, and 2mV Output Voltage Ripple

S. Li, University of Virginia, Charlottesville, VA

In Paper 14.6, the University of Virginia demonstrates a 65nm sub-nA, wide-dynamic-load-range digital LDO with a hybrid asynchronous binary-searching and synchronous linear-searching control scheme. The hybrid DLDO achieves 3.8×10^5 dynamic load range, 99.99% current efficiency, and 2mV output voltage ripple, consuming 745pA quiescent current.

4:45 PM

14.7 A Modular Hybrid LDO with Fast Load-Transient Response and Programmable PSRR in 14nm CMOS Featuring Dynamic Clamp Tuning and Time-Constant Compensation

X. Liu, Intel, Hillsboro, OR

In Paper 14.7, Intel presents a universal modular hybrid LDO in 14nm CMOS that can be efficiently configured to provide any desired combination of PSRR and load transient response. Measurement results demonstrate up to 42dB of PSR, drawing 27μA with zero ripple.

14.1 A 65nm 1.1-to-9.1TOPS/W Hybrid-Digital-Mixed-Signal Computing Platform for Accelerating Model-Based and Model-Free Swarm Robotics

Ningyuan Cao, Muya Chang, Arijit Raychowdhury

Georgia Institute of Technology, Atlanta, GA

Artificial swarm intelligence, inspired by biological studies of insects, ants and other organisms, present an emerging computing paradigm, where seemingly simple elements interact with each other to collectively solve challenging problems. In particular, swarm robotics, where multiple robots co-ordinate in real-time to solve diverse problems such as pattern-formation, cooperative reinforcement learning (RL), path-planning etc. [1], find extensive uses in exploration, reconnaissance and disaster relief. This is partly motivated by the robustness of swarm dynamics to failures and malfunctions of individual robots. Successful hardware demonstrations of neuro-inspired algorithms on edge-devices [2-6] is now leading to the emergence of intelligence and control in swarms as the next frontier. Although certain swarm algorithms rely on real-time learning (e.g., cooperative RL) representing a *model-free* approach, many powerful algorithms that have been developed over the past two decades (e.g., pattern formation) rely on a mathematical structure and represent a more traditional *model-based* approach. The next generation of swarm hardware needs to support both of these approaches. In this paper, we identify the commonalities and shared compute primitives across a variety of *model-based* and *model-free* swarm algorithms and present a unified, fully-programmable, energy-efficient and scalable platform capable of real-time swarm intelligence.

While *model-based* applications (pattern formation and coordinated obstacle avoidance) require calculation of vector fields, *model-free* applications (multi-robot predator-prey, patrolling and exploration) rely on RL-based training of neural-network (NN) models (Fig. 14.1.1). The fact that both of these approaches require linear algebraic kernels, as well as non-linear processing units (e.g., trigonometric units in vector fields and non-linear activation in NNs), motivates us to develop a common computing platform to support both. The principal algorithmic approaches and computing primitives are summarized in Fig. 14.1.1.

Figure 14.1.2 illustrates the architecture of the test-chip fabricated in 65nm CMOS. It features a 16KB data-cache, 27B instruction cache, a processing unit and peripheral circuits for control, data-load and write-back. The test-chip interfaces with a raspberry-Pi platform consisting of integrated sensors (inertial sensors and ultrasonic distance sensors) and LoRa (Long Range) radios for decentralized, peer-to-peer communication among mobile robotic vehicles in a swarm. The processing unit consists of two principal modules (1) a non-linear function evaluator (NFE) and (2) a linear processing unit (LPU). We realize a variety of relevant non-linear functions of input (x) using a piecewise linear approximation, where each piece is characterized by a slope (g_{ref}) and an offset (y_{ref}). The number of linear segments of each function depends on the number of inflection points and the (g_{ref}, y_{ref}) for each segment is stored in the data cache as look-up table (LUT). Data-volume is reduced by aliasing periodic functions to their fundamental ranges and comparing x with the period boundaries, as illustrated in Fig. 14.1.2. The test-chip provides support for 7 non-linear functions determined by the control signal, PU[2:0]. Architecturally, the chip supports high-level functions, which are scanned in as 8b instructions to the instruction cache. Each instruction is decoded and the relevant control signals and data read from the data cache are transmitted to the processing unit. Once an output is generated, it is written back to the data cache, enabling a sequential execution of tasks. The large data cache allows *model-free* algorithms and a 3-layered neural network for RL for up to 20 agents.

Swarm algorithms are characterized by dynamically varying environments and need to support various swarm sizes. A detailed study of *model-based* and *model-free* algorithms reveals that as the swarm sizes increase, vector dimensions, as well as the range of operand values required for correct execution also increase. This is shown in Fig. 14.1.3 for path-planning and predator-prey problems with 2 and 20 agents. The required bit-precision as a function of swarm size and algorithm is summarized. This motivates the design of an energy-scalable LPU capable of easy reconfiguration and high energy efficiency across the range. Time-domain mixed-signal (TD-MS) MACs [2] have been shown to offer voltage-scalability and energy-benefits compared to digital counterparts for lower data-width. However, with increasing operand size, the energy/MAC increases non-linearly and surpasses digital logic, as shown in the energy-map and corresponding table. To address this issue, we propose a hybrid-digital-mixed-signal (HDMS) circuit, where a 5b TD-MS multiplier (4 data bits and 1 sign bit represented in the signed magnitude format [2]) is nested within a digital shift-add loop, such that execution is purely TD-MS for bit-width<=5b and hybrid of TD-MS and digital (5b TD-MS followed by a sequence of digital shift and adds) for 6 ≤ bit-width ≤ 8. This creates an energy map with high efficiency across the range and results in 81% (for 3b operations) to 31% (for 8b operations) average energy/MAC reduction compared to digital implementations. The LPU consists of a 3-stage pipeline and supports 6 linear operations of one or multiple operands with varying latencies.

The TD-MS 5b multiplier kernel consists of a digital-to-pulse converter (DPC) which converts operand (A) to a delay ($A.T_0$) via two matched digital-to-time-converters (DTC) followed by an XOR, as shown in Fig. 14.1.4. The DPC output gates a digitally-controller-oscillator (DCO) with a control word (B), which produces a frequency ($B.F_0$). The DCO clocks an 8b counter to produce ($A.B.T_0.F_0$). We match F_0 and T_0 by using the same logic gates and the output captures $A.B$. Even if there is a mismatch between F_0 and T_0, the output gets scaled and the algorithms are robust against such scaling. The DCO inverter can be further calibrated via analog control signals (DCO_BL and DCO_BH). The TD-MS kernels allow 3-to-5b reconfiguration without any overhead and easy interface to digital memory. Further, the digital shift-and-add circuits allow seamless transition from 6-to-8b. Measured responses of the DCO and DPC show acceptable linearity of less than 1.2 lsb (worst-case) at V_{CC}=1.0V and V_{CC}=0.6V.

Figure 14.1.5 illustrates the measured F_{MAX} and logic-power dissipation showing functionality down to 0.36V and a peak logic power of 3.2μW (1.9μW) for 8b (5b) operation. Measured energy/op shows excellent energy-resolution scalability reaching a peak of 0.22pJ/MAC (at 3b) and 1.76pJ/MAC (at 8b). We measure the average arithmetic energy efficiency as a function of V_{CC} and record 9.1TOPS/W (3b) to 1.1TOPS/W (8b). We plot the energy break-down of the computation unit where the LPU (NFE) consumes 88% (12%) of the logic power. The power distribution across the various blocks of the LPU are shown.

We benchmark the test-chip and note competitive figures-of-merit for an emerging application. The test-chip is mounted on a robotic car (Fig. 14.1.6) and interfaces with a Raspberry-Pi, motor-controllers, sensors and LoRA radios. We implement 4 template swarm algorithms (*model-based*: path planning and pattern formation, and *model-free*: predator prey and joint exploration). In Fig. 14.1.6 we demonstrate (1) two robots (R1 and R2) capture two targets (T1 and T2) collaboratively, (2) three robots starting from three random locations converge to a triangular pattern, and (3) cooperative RL where the number of iterations required to perform obstacle avoidance reduces with the number of training examples. We further benchmark the energy/task and the number of actions taken per second for the template problems, illustrating a wide diversity and complexity of applications. The die-shot and the chip-characteristics are shown in Fig. 14.1.7.

Acknowledgements:
This project was supported by the Semiconductor Research Corporation under grant JUMP CBRIC task ID 2777.006.

References:
[1] M. La et. al., "Multirobot Cooperative Learning for Predator Avoidance," *IEEE Trans. on Control Systems Technology*, vol. 23, no. 1, pp. 52-63, 2015.
[2] A. Amravati et. al., "A 55nm Time-Domain Mixed-Signal Neuromorphic Accelerator with Stochastic Synapses and Embedded Reinforcement Learning for Autonomous Micro-Robots," *ISSCC*, pp. 124-125, 2018.
[3] B. Moons et. al., "Envision: A 0.26-to-10TOPS/W subword-parallel dynamic-voltage-accuracy-frequency-scalable Convolutional Neural Network processor in 28nm FDSOI," *ISSCC*, pp. 246-247, 2017.
[4] J. Sim et. al., "A 1.42TOPS/W Deep Convolutional Neural Network Recognition Processor for Intelligent IoE Systems," *ISSCC*, pp. 264-265, 2016.
[5] S. Choi et. al., "A 9.02mW CNN-Stereo-Based Real-Time 3D Hand-Gesture Recognition Processor for Smart Mobile Devices," *ISSCC*, pp. 220-221, 2018.
[6] Y. Chen et. al., "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," *IEEE JSSC*, vol. 52, no. 1, pp. 127-138, 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

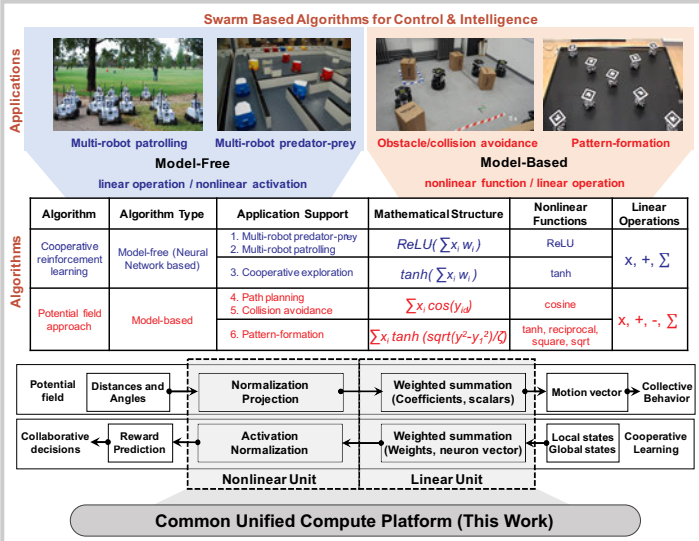


Figure 14.1.1: A common unified compute platform supporting model-based and model-free algorithms in swarm robotics.

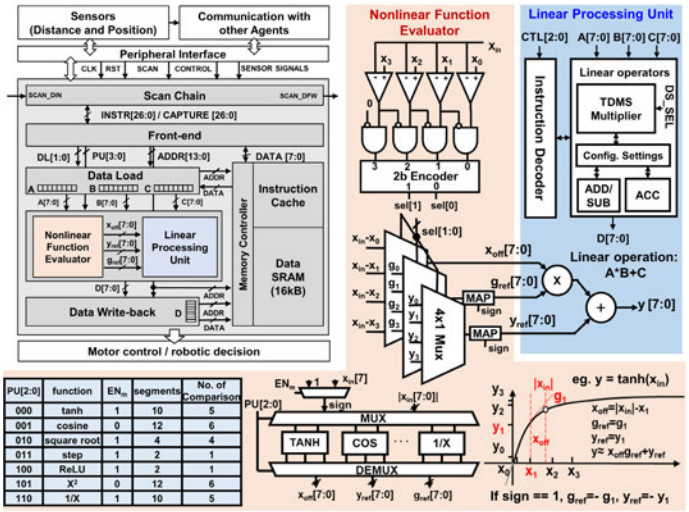


Figure 14.1.2: System architecture illustrating on-chip memory, compute unit and peripheral circuits.

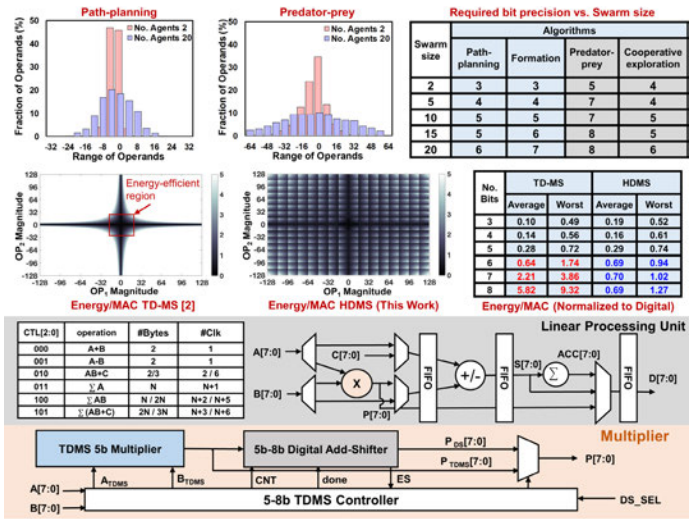


Figure 14.1.3: Hybrid-digital-mixed-signal LPU provides energy-resolution scalability across a variety of algorithms and swarm sizes.

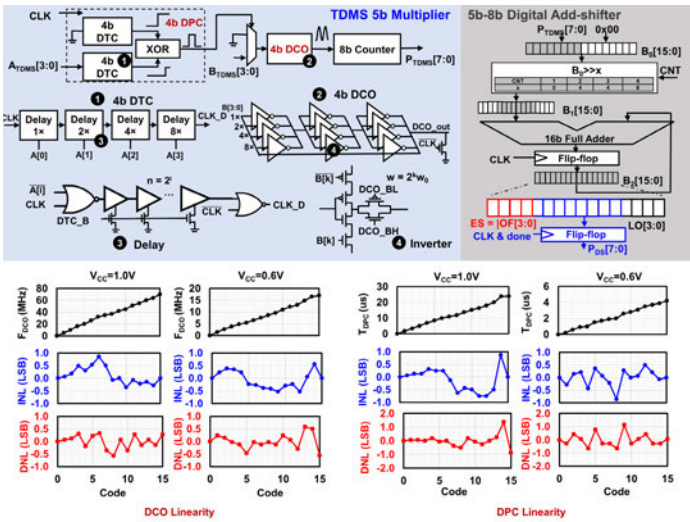


Figure 14.1.4: HDMS-based LPU illustrating various design components and measured linearity of DCO and DPC.

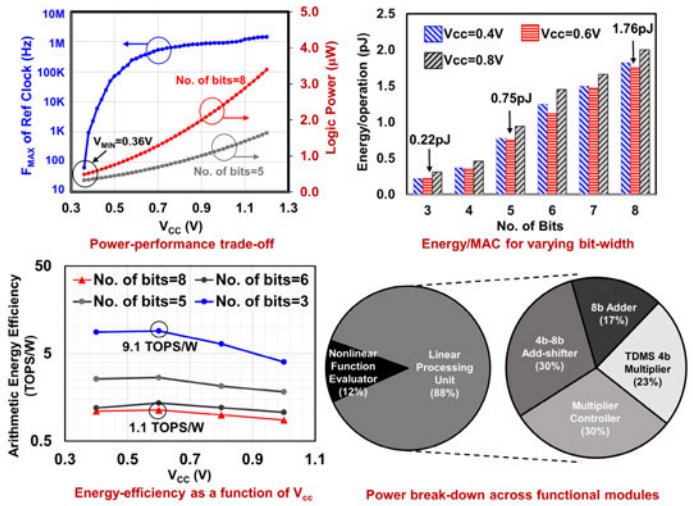


Figure 14.1.5: Measured power, performance and energy-efficiency.

	This work	[2]	[3]	[4]	[5]	[6]
Application	Swarm intelligence	Autonomous micro-robotics	CNN Inference	DNN Inference	CNN Inference	CNN Inference
Optimization algorithm	Cooperative RL/potential field	Reinforcement Learning	none	none	none	none
Learning/Training	Online in real time	Online in real time	offline	offline	none	offline
Technology	65nm	55nm	180nm	65nm	65nm	65nm
Area	2mm²	3.4mm²	3.3mm²	16mm²	16mm²	16mm²
On-die SRAM	16 KB	200 B	144 KB	36 KB	490.5 KB	181.5 KB
Resolution	3-8b	6b	4b-16b	16b	16b	16b
Power	0.3-3.4 uW	650 uW	7.5-300 mW	45 mW	6.57 mW	278 mW
Frequency	1KHz-1.5MHz	67.5 MHz	200 MHz	125 MHz	10 - 100 MHz	200 MHz
Supply voltage	0.4-1V	0.4-1V	1V	1.2V	0.7-1.2V	0.82-1.17V
Performance/Watt	1.1-9.1 TOPS/W	3.12 TOPS/W	0.26-10TOPS/W	1.42TOPS/W	11.8 - 19.7 GOPS	0.21TOPS/W

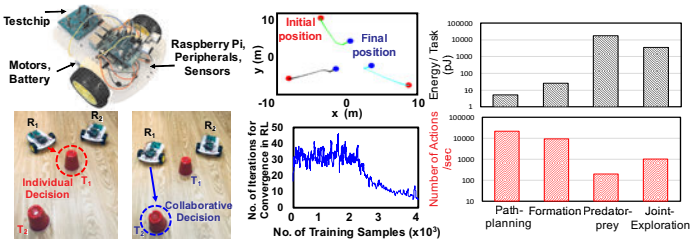
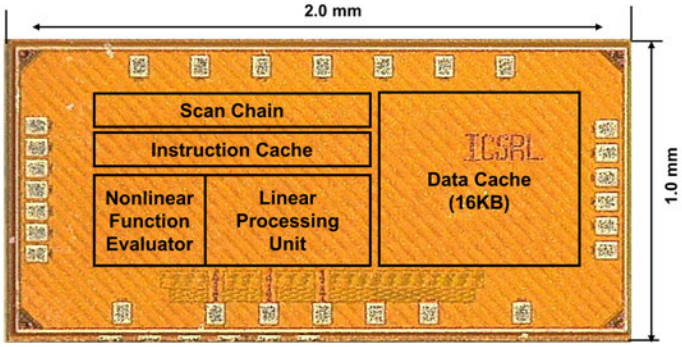


Figure 14.1.6: Benchmarking and application to mobile swarm robotics.



Chip Characteristics	
Technology	65nm 1P9M CMOS
Die area	1mm*2mm
Testing interface	QFN package
Pin Count	28

Figure 14.1.7: Die shot and chip characteristics.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

14.2 A Compute SRAM with Bit-Serial Integer/Floating-Point Operations for Programmable In-Memory Vector Acceleration

Jingcheng Wang, Xiaowei Wang, Charles Eckert, Arun Subramaniyan, Reetuparna Das, David Blaauw, Dennis Sylvester

University of Michigan, Ann Arbor, MI

Data movement and memory bandwidth are dominant factors in the energy and performance of both general purpose CPUs and GPUs. This has led to extensive research focused on in-memory computing, which moves computation to where the data is located. With this approach, computation is often performed on the memory bit-lines in the analog domain using current summing [1-3], which requires expensive analog-to-digital and digital-to-analog conversions at the array boundary. In addition, such analog computation is very sensitive to PVT variations, limiting precision. More recently, full-rail (digital) binary in-memory computing was proposed to avoid this conversion overhead and improve robustness [4, 5]. However, both prior in-memory approaches suffer from the same major limitations: they accelerate only one type of algorithm and are inherently restricted to a very specific application domain due to their limited and fixed bit-width precision and non-programmable architecture. Software algorithms, on the other hand, continue to evolve rapidly, especially in novel application domains, such as neural networks, vision and graph processing, making rigid accelerators of limited use. Furthermore, most available SRAM in today's chips is located in the caches of CPUs or GPUs. These large CPU and GPU SRAM stores present an opportunity for extensive in-memory computing and have, to date, remained largely untapped.

In this paper, we present a general purpose hybrid in-/near-memory compute SRAM (CRAM) that combines the efficiency of in-memory computation with the flexibility and programmability necessary for evolving software algorithms. CRAM augments conventional SRAM in a CPU with vector-based, bit-serial [6, 7] in-memory arithmetic. It can accommodate a wide range of bit-widths, from single to 32b or 64b, and operation types, including integer and floating point addition, multiplication and division. To maintain compatibility with CPU/GPU operation, CRAM writes/reads operands conventionally with horizontal word-lines and vertical bit-lines. Then, using a transposable bitcell [8], CRAM operates directly on the stored operands in memory with additional horizontal compute bit-lines. This enables the same bit position from two vectors elements to be simultaneously accessed on a single bit-line. Logic operations are performed on the bit-line (in-memory), while small additional in-column logic (near-memory with 4.5% SRAM bank area overhead) enables carry-propagation between successive bit-serial calculations, enabling multi-bit arithmetic operations in SIMD fashion across all vectors of elements. To maintain versatility, the memories can function either as traditional or compute memories. The approach was implemented in a small IoT processor in 28nm CMOS, consisting of a Cortex-M0 CPU and 8 CRAM banks of 16KB each (128KB total). The system achieves 475MHz operation and, with all CRAMs active, produces 30GOPS or 1.4GFLOPS on 32b operands for graph, neural, and DSP applications.

Figure 14.2.1 shows the overall organization of the IoT processor. The ARM core can access all 8 memory banks and load/store data using the horizontal word-lines and vertical bit-lines. Then, in-memory instructions can be streamed from one bank to one or more compute-configured banks, while the M0 simultaneously performs other processing with the remaining memory banks. Banks performing in-memory computing use the horizontal compute bit-lines (CBLs) and vertical compute word-lines (CWLs).

Figure 14.2.2 shows the architecture of the 128×256 CRAM sub-array, which is one quarter of a 16KB CRAM macro. An 8T transposable bitcell is used to provide bidirectional access. Fig. 14.2.2 shows an example operation of the data flow for a 1b addition performed in 1 cycle of the bit-serial computation. Here, we add the second bit positions of vector A (A_1=0) and vector B (B_1=1) with carry-in C (=1) from the previous cycle, and store the result back to vector D. First, the CRAM instruction decoder receives the ADD instruction and the 3 column addresses of bits A_1, B_1 and D_1. It activates the CWLs of A_1 and B_1 simultaneously to compute 'A AND B' on CBL and '$\overline{A}$ AND $\overline{B}$' on CBLB. Since A=0 and $\overline{B}$=1, both CBL and CBLB discharge. Then, after the dual sense amps, the results propagate to the near-memory logic located at the end of each CBL. The NOR gate generates 'A XOR B', which combined with C_{in} from the carry latch produces Sum=0 and C_{out} =1. Sum is then written back to D, and C_{out} is stored in the carry latch, which provides C_{in} for the next cycle, thus completing one full bit-serial addition in one clock cycle.

Figure 14.2.3, left, shows how two vectors of 2b numbers (A and B) are added bit-by-bit starting from the least significant bit (LSB). Note that while only one bit of a multi-bit operand is processed in each cycle, all compute bit-lines operate simultaneously, resulting in massive parallelism (2048 CBLs in our design). Subtraction is performed by first inverting B and then adding to A with C_{in} pre-set to 1. As shown in Fig. 14.2.3, multiplication is more complicated as it requires predication. For this, the tag latch (Fig. 14.2.2) is used to enable the write-back driver, resulting in a conditional copy/addition. First, 4 empty columns in the array are reserved for the product and initialized to zero. In the first cycle, the LSB of the multiplier is loaded to the tag latch. In cycles 2 and 3, the multiplicands are copied to product columns only if their tag is 1. In cycle 4, the second bit of the multiplier is loaded to the tag latch. In the next 2 cycles, for rows with tag = 1, the multiplicands are added to the *second* and *third* bits of the product, shifting the multiplicands by 1 to account for the multiplier bit position. Finally, we store C_{out} in the most significant bit (MSB) of the product to complete the multiplication. Note that partial products are implicitly shifted as they are added using appropriate bit addressing in the bit-serial operation and no explicit shift is performed. Division is conducted similarly by implicit shifting and subtraction from a partial result. Floating point arithmetic is implemented using repeated integer add/sub/mult/div with predication. Fig. 14.2.3 provides a list of supported computations and their performance, demonstrating both the versatility of CRAM and its high performance due to bit-line parallelism.

Figure 14.2.4 shows measurement results from the prototype chip fabricated in 28nm CMOS that contains 8 CRAM banks (128KB memory with 2048 computing rows) and a Cortex-M0 processor. The figure shows measured frequency and energy efficiency of 8b addition and multiplication across supply voltage. At 1.1V the maximum frequency of 475MHz results in 122GOPS for 8b addition and 9.4GOPS for 8b multiplication. The best energy efficiency is achieved at 0.6V and 114MHz, resulting in 0.56TOPS/W for 8b multiplication and 5.27TOPS/W for 8b addition. Fig. 14.2.4 shows measured frequency and leakage power distributions for 21 measured dies.

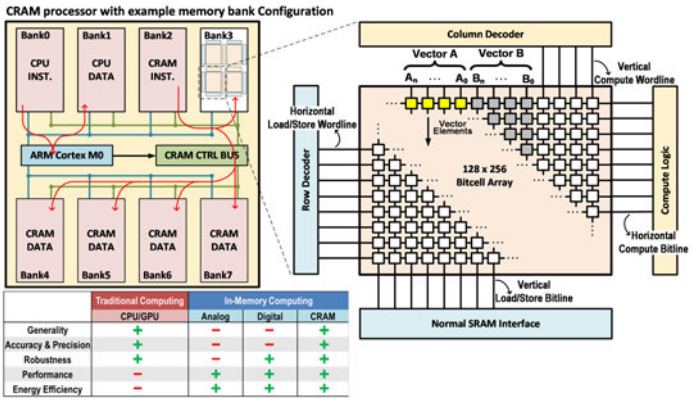
Figure 14.2.5 shows the performance of the test chip for diverse computationally intensive tasks ranging from neural networks to graph and signal processing. The total latency in cycles is compared with a baseline operation, where CRAMs are only used as data memories and the computation is entirely performed on the ARM CPU. The first benchmark is the 1st convolutional layer from Cuda-convnet and the second is the last fully connected layer from AlexNet. Due to their size, these layers must be executed in multiple smaller sub-sections. The third application consists of 512 simultaneous 32-tap FIR filters and the fourth application performs traversal of a directed graph represented by a 192×192 adjacency matrix. The workload breakdown shows the percentage of time spent on input loading and output loading vs. in-memory computation. Speedup, compared to executing the same workload with the ARM Cortex-M0, varies from 7.2-to-114×, with the greatest gains obtained when the operation is compute-heavy and low on input/output movement.

Figure 14.2.6 compares the proposed approach with other state-of-the-art in-memory accelerators. The proposed work is the only solution to provide a wide range of instructions and flexible bitwidth. It repurposes the memory storage already available in processors, thereby accelerating computation while maintaining programmability.

Acknowledgements:
We gratefully acknowledge TSMC University Shuttle Program for chip fabrication. This work was supported in part by ADA, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

References:
[1] J. Zhang, et al., "In-Memory Computation of a Machine Learning Classifier in a Standard 6T SRAM Array," *IEEE JSSC*, vol. 52, no. 4, pp. 915-924, 2017.
[2] A. Biswas, et al., "Conv-RAM: An Energy-Efficient SRAM with Embedded Convolution Computation for Low-Power CNN-based Machine Learning Applications," *ISSCC*, pp. 488-489, 2018.
[3] S. Gonugondla, et al., "A 42pJ/Decision 3.12TOPS/W Robust In-Memory Machine Learning Classifier with On-Chip Training," *ISSCC*, pp. 490-491, 2018
[4] W. Khwa, et al., "A 65nm 4Kb Algorithm-Dependent Computing-in-Memory SRAM Unit-Macro with 2.3ns and 55.8TOPS/W Fully Parallel Product-Sum Operation for Binary DNN Edge Processors," *ISSCC*, pp 496-497, 2018.
[5] Y. Zhang, et al., "Recryptor: A Reconfigurable In-Memory Cryptographic Cortex-M0 Processor for IoT," *IEEE Symp. VLSI Circuits*, 2017.
[6] K. Batcher, "Bit-Serial Parallel Processing Systems," *IEEE Trans. on Computers*, vol. 31, no. 5, pp. 377-384, 1982.
[7] C. Eckert, et al., "Neural Cache: Bit-Serial In-Cache Acceleration of Deep Neural Networks," *ACM/IEEE ISCA*, pp. 383-396, 2018.
[8] J. Seo, et al., "A 45nm CMOS Neuromorphic Chip with a Scalable Architecture for Learning in Networks of Spiking Neurons," *IEEE CICC*, 2011.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE



	Traditional Computing	In-Memory Computing		
	CPU/GPU	Analog	Digital	CRAM
Generality	+	−	−	+
Accuracy & Precision	+	−	−	+
Robustness	+	−	+	+
Performance	−	+	+	+
Energy Efficiency	−	+	+	+

Figure 14.2.1: Chip architecture and storage and computation of data in transposable memory array.

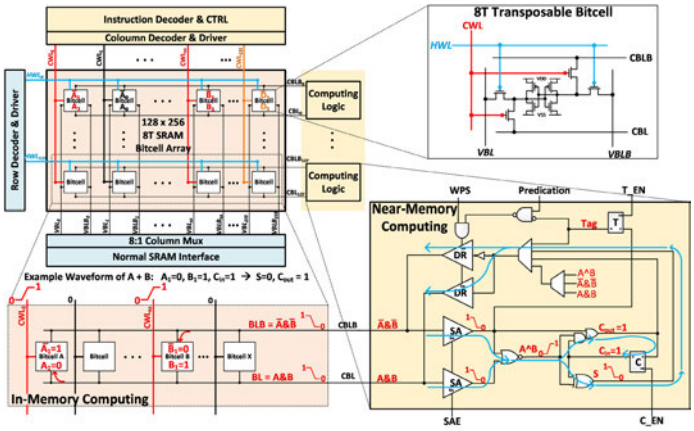
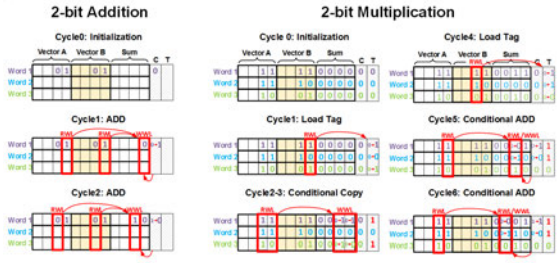


Figure 14.2.2: CRAM array architecture (top-left), 8T transposable bitcell (top-right), In-memory computing part (bottom-left) and near-memory computing part (bottom-right) of 1-bit addition. Addition of near-memory logic increased array size by 4.5%.



Type	Operation	# Cycles	32-bit Performance[2]	8-bit Performance[2]
Logic	AND	N[1]	30.4GOPS	122 GOPS
	NOR	N		
	XOR	N		
	NAND	N		
	OR	N		
	XNOR	N		
Integer	Add	N	30.4 GOPS	122 GOPS
	Sub	2N	15.2 GOPS	60.8 GOPS
	Mult	$N^2 + 5N - 2$	0.83 GOPS	9.40 GOPS
	UDiv	$1.5N^2 + 5.5N$	0.57 GOPS	6.97 GOPS
32-bit Float Point	Add/Sub	4978	0.20 GFLOPS	/
	Mult	679	1.43 GFLOPS	/
	Div	697	1.40 GFLOPS	/
Comparison	Equal	2N+1	14.9 GOPS	57 GOPS
	Greater/Less	2N+1		
	Search	N	30.4 GOPS	122 GOPS

1. N is the bit-width of data
2. Measured performance on 128KB CRAM test chip at 475MHz

Figure 14.2.3: 2-bit addition cycle-by-cycle demonstration (top-left), 2-bit multiplication cycle-by-cycle demonstration (top-mid & right), and list of CRAM instructions and its performance (bottom).

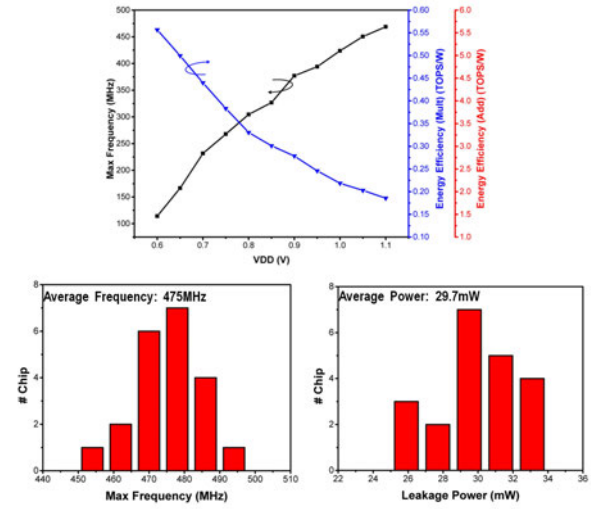


Figure 14.2.4: Frequency and energy efficiency of 8-bit multiplication and addition at different VDD (top), maximum frequency and leakage power distribution of 21 dies at 1.1V (bottom).

	CONV		FC		FIR		GRAPH	
Testbench	Cuda-Convnet 1st layer		AlexNet last layer		512 32-Tap Filter		Nearest Neighbor Traversal	
Input Size	24x24x3		24x1		32x10		192x192	
Parameter Size	5x5x3x64		1000x24		512x32		0	
Output Size	1x1x64		1000x1		512x10		192x192	
Bit Precision	8		8		4		1	
# Array used for compute	3		6		2		2	
# Wordline for computation	375		1000		512		192	
	cycle#	percentage	cycle#	percentage	cycle#	percentage	cycle#	percentage
Total Latency	39,628	100	33,434	100	251,290	100	1,572,628	100
Input loading	18,323	46.2	24	0.07	320	0.13	1,152	0.07
CRAM Compute	3,459	8.7	21,267	63.6	184,020	73.2	1,556,458	99.0
output readout	17,846	45.0	12,143	36.3	66,950	26.6	15,018	0.95
	cycle#	speedup	cycle#	speedup	cycle#	speedup	cycle#	speedup
Baseline	287,073	7.24x	1,174,032	35.1x	8,120,415	32.3x	164,456,448	114x

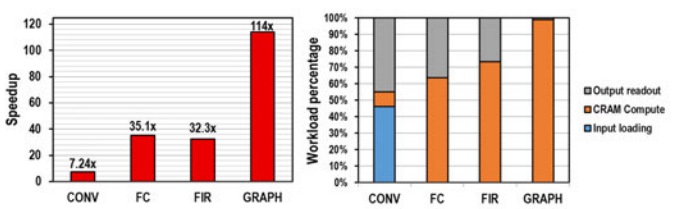
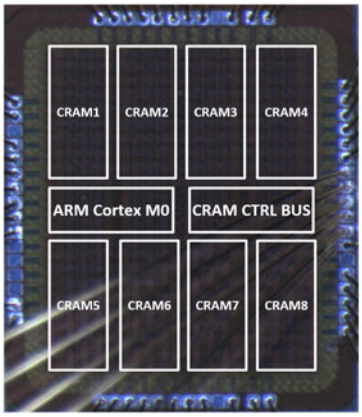


Figure 14.2.5: Performance comparison between CRAM and baseline scenario (top), workload breakdown (bottom).

	This Work	JSSC2017 [1]	ISSCC2018 [2]	ISSCC2018 [3]	ISSCC2018 [4]	VLSI2017 [5]
Technology	28nm	130nm	65nm	65nm	65nm	40nm
Supply Voltage	0.6~1.1V	1.2V	0.9~1.2V	0.65~1V	0.6~1V	0.65~0.9V
SRAM Macro size	16KB	2KB	2KB	16KB	0.5KB	8KB
SRAM bitcell	8T	6T	10T	6T	6T	10T
Method of Computing	Digital	Analog	Analog	Analog	Digital	Digital
Type of Supported Functions	Logic/Add/Sub/Mult/Div/FP	Add/Mult	Add/Mult	Add/Mult	Add/Mult	Logic
Bit precision	Arbitrary	5b (input) 1b (weight)	7b (input) 1b (weight)	16b (train) 8b (infer)	1b (input) 1b (weight)	Arbitrary
Die Area (mm2)	2.7	0.36	0.067	1.44	-	1.28
Max Freqency (MHz)	475	50	6.7	1000	435	90
Normalized Performance (GOPS)*	32.7	40.5	1.17	-	-	-
Performance per unit Area (GOPS/mm2)**	27.3	114	17.5	-	-	-
Normalized Energy Efficiency (TOPS/W) *	0.55 (mult) 5.27 (add)	0.16 (mult) 4.58 (add)	3.07	3.12	0.87	-

*All normalized to 8-by-8 bit Multiply and Add Operation
** Using SRAM area only

Figure 14.2.6: Comparison table.



Technology	28nm CMOS
Processor	ARM Cortex M0
Bitcell Size	0.405um x 1.93um (Logic-rule)
Chip Size	1.5 x 1.7 mm²
Supply Voltage	0.6 ~ 1.1V
Memory Capacity	128KB (8 x 16KB)
SRAM Macro Size	16KB
SRAM Sub Array	128 rows x 256 colums
Clock Frequency	475MHz @1.1V
Average power	105mW @1.1V

Figure 14.2.7: Die photo.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

14.3 A 43pJ/Cycle Non-Volatile Microcontroller with 4.7μs Shutdown/Wake-up Integrating 2.3-bit/Cell Resistive RAM and Resilience Techniques

Tony F. Wu[1], Binh Q. Le[1], Robert Radway[1], Andrew Bartolo[1], William Hwang[1], Seungbin Jeong[1], Haitong Li[1], Pulkit Tandon[1], Elisa Vianello[2], Pascal Vivet[2], Etienne Nowak[2], Mary K. Wootters[1], H.-S. Philip Wong[1], Mohamed M. Sabry Aly[3], Edith Beigne[2], Subhasish Mitra[1]

[1]Stanford University, Stanford, CA
[2]CEA-LETI-MINATEC, Grenoble, France
[3]Nanyang Technological University, Singapore, Singapore

Non-volatility is emerging as an essential on-chip memory characteristic across a wide range of application domains, from edge nodes for the Internet of Things (IoT) to large computing clusters. On-chip non-volatile memory (*NVM*) is critical for low-energy operation, real-time responses, privacy and security, operation in unpredictable environments, and fault-tolerance [1]. Existing on-chip NVMs (e.g., Flash, FRAM, EEPROM) suffer from high read/write energy/latency, density, and integration challenges [1]. For example, an ideal IoT edge system would employ *fine-grained temporal power gating* (i.e., shutdown) between active modes. However, existing on-chip Flash can have long latencies (> 23ms latency for erase followed by write), while inter-sample arrival times can be short (e.g., 2ms in [2]).

Our chip monolithically integrates two heterogeneous technologies: 18KB of on-chip resistive RAM (emerging on-chip NVM, technology details in Fig. 14.3.1) on top of commercial 130nm silicon CMOS (16b general-purpose microcontroller core with 8KB of SRAM). For various applications (in machine learning, control, and cryptography), we demonstrate active mode average energy of 43pJ/cycle (up to 5.7× lower vs. similar chips at similar speeds / technology nodes using on-chip Flash and FRAM), fine-grained temporal power gating (0.25μW during shutdown) with up to 8μs (average 4.7μs) transition from active to shutdown mode (up to 5,878× quicker vs. on-chip Flash), and 2-clock cycle (200ns) transition from shutdown to active mode. We also demonstrate a complete chip that stores multiple bits per on-chip RRAM cell (5 resistance values, i.e., 2.3b per cell) and processes stored information correctly (vs. previous demonstrations using standalone RRAM cells or few cells in standalone RRAM array). Such multi-bit storage improves the accuracy of neural network inference (2.3× for MNIST) on same hardware (vs. 1b per cell).

RRAM (like other emerging NVMs, such as phase change memory) exhibits write failures [1]. We overcome these challenges through the critical combination of two resilience techniques: 1) *dynamic address remapping*, which overcomes write failures during system operation with 0.5% active-mode energy increase and negligible execution time impact; 2) periodic *ENDUrance REsiliency using random Remapping (ENDURER* – Fig. 14.3.5*)* [3] – a new technique implemented here. This combination enables our chip to achieve a 10-year functional lifetime when running MNIST inference continuously.

To demonstrate fine-grained temporal power gating enabled by on-chip RRAM, our chip operates as follows (Fig. 14.3.1). During *active mode*, instructions are read from the on-chip 12KB instruction RRAM and executed by the microcontroller core (MSP430 instruction set). During this time, data is accessed from peripheral ports (e.g., off-chip sensors), on-chip 4KB data RRAM, or on-chip 8KB scratchpad SRAM (loop counters, temporary variables with repeated writes: memory-mapped using the compiler). After the data is processed, to transition to *shutdown mode*, results are written back to the 4KB on-chip data RRAM (consuming 168pJ over 5 clock cycles per 16b word, Fig. 14.3.2) and the hardware scheduler unit power-gates (i.e. turns off power) the core, memory controllers, and memory. Our chip performs this transition 5,878× quicker than those with on-chip Flash due to the low write latency of RRAM (500ns vs 23ms for Flash). The chip returns to active mode upon data arrival (e.g., from sensors).

We run 5 applications representing machine learning (logistic regression, support vector machine, convolutional neural network), control (Kalman filter) and cryptography (SHA256 hash) to demonstrate the effectiveness of our chip (Fig. 14.3.2). To put our results into perspective, we select a similar clock rate for our chip (10MHz, vs. industry chips with existing on-chip NVM such as FRAM and Flash) that is sufficient for fine-grained temporal power-gating, while avoiding excessive energy consumption. The active mode power of our chip varies between 407μW to 477μW (average active mode energy: 43pJ/cycle). We achieve average 4.7μs/1.6nJ transition from active to shutdown mode and a 200ns/152pJ transition from shutdown to active mode (Fig. 14.3.2). Although the industry chips might be engineered to include additional margins, the overall benefits demonstrated by our chip are expected to stay significant even after margins are taken into consideration.

We store multiple resistance levels (up to 5 in our chip) inside on-chip RRAM cells (e.g., neural network model weights, only read during inference) by special algorithms that change wordline voltage (V_{WL}) and bitline voltage (V_{BL}) in addition to modifying the pulse width (Fig. 14.3.3) and allocating larger resistance windows for levels with higher resistance values. With greater effective memory capacity (2.3b vs. 1b per RRAM cell) on the same hardware, higher-precision weights (e.g., 4b vs 8b) or larger neural network models (e.g., 6,490 vs. 9,402 weights) can be used (Fig. 14.3.3). Despite errors (cells with resistance values outside its intended resistance window) in 5 levels-per-cell storage, we achieve a 2.3× improvement in inference accuracy (i.e., 2.3× decrease in inference error) for neural networks (on the MNIST dataset, Fig. 14.3.3) when the weights are encoded as follows: two 5-level cells for magnitude and one 2-level cell for sign bit.

RRAM is subject to temporary write failures (*TWFs*) and permanent write failures (*PWFs*, resulting in limited *endurance*: maximum number of successful writes to a cell) [4] that degrade application accuracy over time (Fig. 14.3.4). Cell-level parameter adjustment to improve write failures is not sufficient [4]. To address TWFs, we employ a write-verify scheme with retries [4]. If a write to an RRAM address is unsuccessful after 4 retries, we map that address (during runtime) to another location in a separate backup RRAM array using *dynamic address remapping* (Figs. 14.3.1, 14.3.4). Our chip contains a backup RRAM array (256 16b words) for every 4KB of RRAM; 128 words of that backup array are used for this mapping. The mapping information is stored in a 128-entry volatile look-up table (*volatile LUT*, implemented using flip-flops, Fig. 14.3.1). During transition from active to shutdown mode, the contents of each volatile LUT are stored in the remaining 128 words of the corresponding backup array (*non-volatile LUT*). A write failure to a non-volatile LUT entry results in that entry marked invalid (majority vote over 5 RRAM bits decides entry validity). When the chip boots, the contents of the volatile LUTs are loaded from the corresponding non-volatile LUT. We use dynamic address remapping for our data RRAM, incurring 0.5% energy and negligible (0.005%) execution time costs; our data RRAM tolerates TWFs and PWFs in 17.3% and 2% of words, respectively (Fig. 14.3.4). We use stronger programming conditions (higher voltage, more retries) to mitigate TWFs and insert dummy instructions to avoid PWFs in instruction memory (as writes occur only during programming).

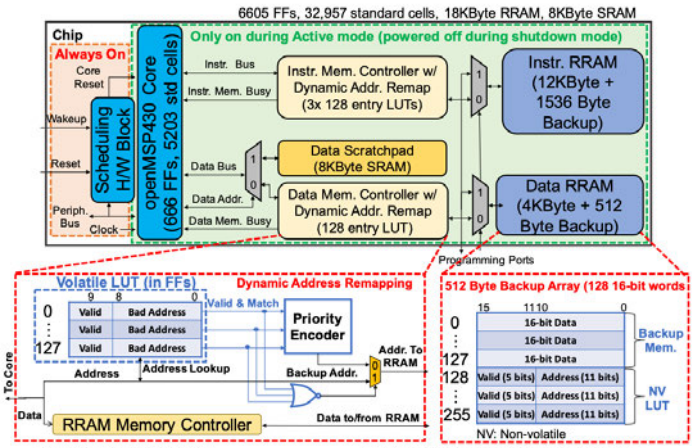
Despite limited write endurance of the 4KB data RRAM, we achieve 10-year lifetime using ENDURER (Fig. 14.3.5, software on FPGA + our chip) combined with dynamic address remapping, when running our neural network application (MNIST dataset) continuously (Fig. 14.3.6). We accelerate our tests to account for 10 years of running an application by first obtaining a sequence of all writes to RRAM (which account for 258 out of 617,669 total memory operations for a single inference) for the application. Then, we repeatedly perform the sequence of writes, through the ENDURER module on the FPGA, on the RRAM (skipping any read operations, writes to non-RRAM, and computation to save time). In our implementation of ENDURER, remapping is performed every 30 minutes and we use an SRAM buffer of 8 16b words.

On-chip RRAM NVM enables significantly lower energy during active mode (vs. existing on-chip NVM such as Flash and FRAM), fine-grained temporal power gating, and multiple bits per RRAM cell. Correct computation using multi-bit RRAM cells in a complete chip successfully improves neural network inference accuracy. Effective resilience techniques enable chips with on-chip RRAM to achieve 10-year lifetime (for neural network inference applications) despite write failures in the underlying RRAM. Our results can be further enhanced through domain-specific accelerators, bit-cost scalable 3D Vertical RRAM [5], and monolithic 3D integration of multiple RRAM layers [5]. The presented techniques (fine-grained temporal power gating, resilience) may be used for other emerging on-chip NVM (e.g., phase change) technologies as well.

Acknowledgements:
Work supported in part by DARPA, NSF/NRI/GRC E2CDA, and the Stanford SystemX Alliance.

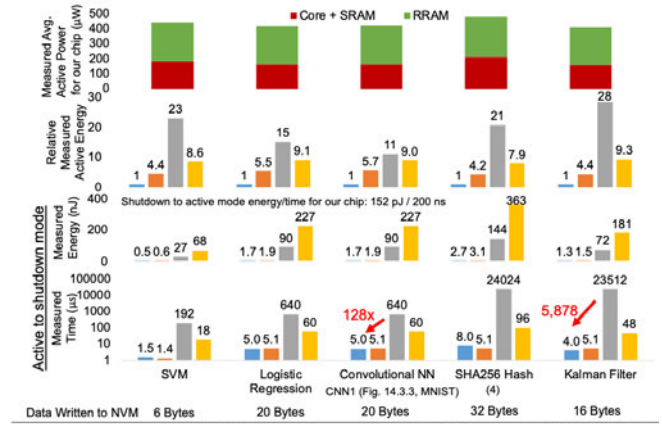
References:
[1] A. Chen, "A Review of Emerging Non-Volatile Memory (NVM) Technologies and Applications," *Solid-State Electronics*, vol. 25, pp. 25-38, 2016.
[2] R. Braojos, et al., "Nano-Engineered Architectures for Ultra-Low Power Wireless Body Sensor Nodes," *CODES+ISSS*, 2016.
[3] M. M. S. Aly, et al., "The N3XT Approach to Energy-Efficient Abundant-Data Computing," *Proc. IEEE*, 2019.
[4] A. Grossi, et al. "Fundamental Variability Limits of Filament-based RRAM," *IEDM*, pp. 4.7.1-4.7.4, 2016.
[5] H.-S. P. Wong, et al., "Memory Leads Way to Better Computing," *Nat. Nanotech.*, vol. 10, pp 191-194, 2015.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE



Chip Operating Conditions		RRAM Device (TiN/HfOx/Ti/TiN) Parameters			
Minimum VDD	0.71V	Set Voltage	1.55V	Reset Wordline Voltage	3.2V
Minimum RRAM Peripheral Logic VDD	3.2V	Reset Voltage	2.5V	Set/Reset pulse time	50ns
Clock Frequency	10MHz	Set Wordline Voltage	1.6V	Read Voltage	0.2V

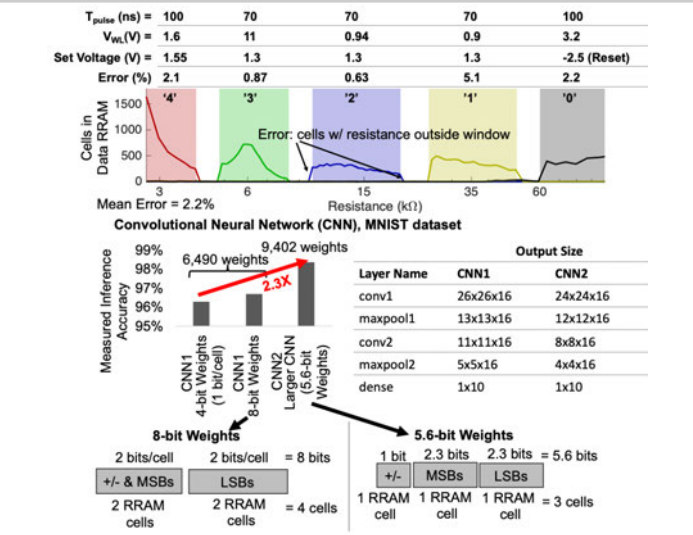
Figure 14.3.1: Block diagram of our chip.



	Our Chip (RRAM) (130nm)	130nm FRAM (1)	130nm Flash (2)	180nm Industry RRAM(3)
Operating Frequency (MHz)	10	8	8	10
Amount of NVM (KBytes)	18	16	128	64
Read/Write latency (ns)	23/50 (500 w/ write-verify(5))	50/50	140/64μs (Prgm)/23ms (Erase)	Not Reported
Read/Write energy (pJ/bit)	1.76/10.9 (Set) / 10.1 (Reset)	10.5/10.5	30.5/562.5	Not Reported
Shutdown mode power (μW)	0.25	0.64	0.54	0.42

(1) MSP430FR5739, (2) MSP430F5529, (3) MN101L (4) http://www.ti.com/tool/SHA-256 (5) Write-verify: read → set → verify set → reset → verify reset

Figure 14.3.2: Benefits of using our chip with on-chip RRAM.



T_{pulse} (ns) =	100	70	70	70	100
V_{WL}(V) =	1.6	11	0.94	0.9	3.2
Set Voltage (V) =	1.55	1.3	1.3	1.3	-2.5 (Reset)
Error (%)	2.1	0.87	0.63	5.1	2.2

Layer Name	CNN1	CNN2
conv1	26x26x16	24x24x16
maxpool1	13x13x16	12x12x16
conv2	11x11x16	8x8x16
maxpool2	5x5x16	4x4x16
dense	1x10	1x10

Figure 14.3.3: Using 2.3 bits per RRAM cell for convolutional neural network applications.

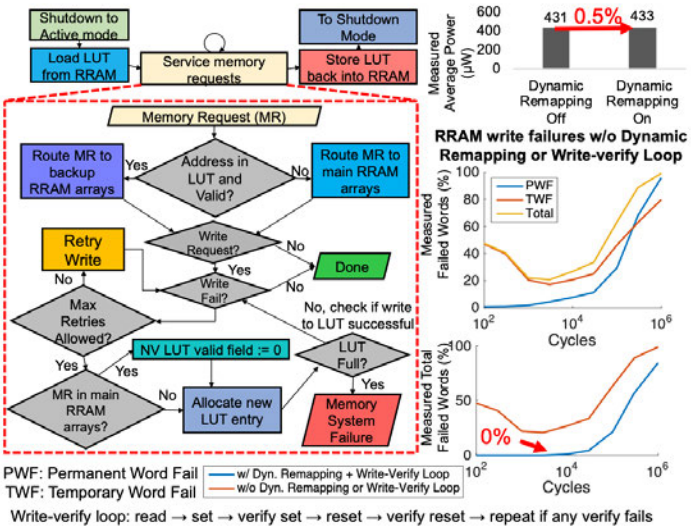


Figure 14.3.4: Dynamic address remapping with write-verify loop.

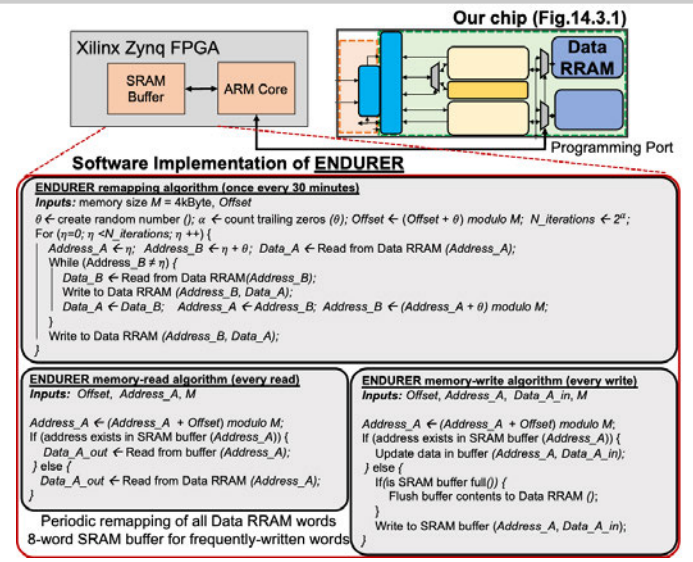


Figure 14.3.5: ENDURER test setup and remapping, read, and write algorithms.

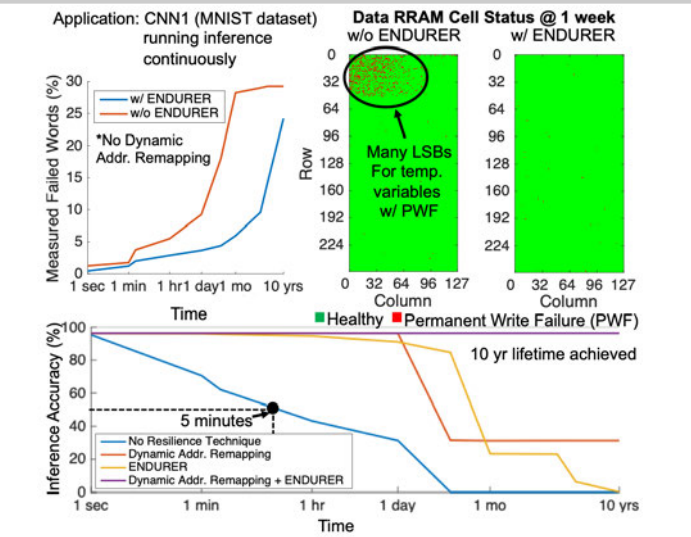
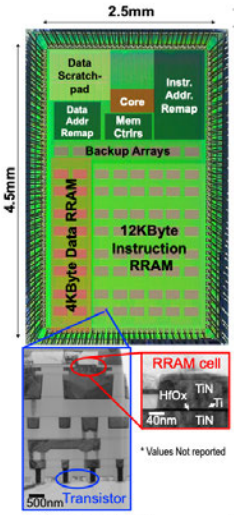


Figure 14.3.6: ENDURER and dynamic address remapping improves chip lifetime to 10 years.



	This work	Liu, et al. [6]	Su, et al. [7]	Chen, et al. [8]
Year	2019	2016	2017	2018
Supply Voltage (V)	0.71 - 1.2	0.8	0.8	1
Technology node (nm)	130	65	150	65
Clock Frequency (MHz)	10	100	20	64
Memory Technology	RRAM	RRAM	RRAM	RRAM
Amount of NVM(1) (KBytes)	18	12.2	1.3	128
# of bits/cell demonstrated	2.3	1	1	1
Type	16-bit microcontroller	8-bit NV(2) Processor	8-bit NV Proc. + Accelerator	In-memory compute macro
Applications (energy, pJ/cycle) / (Active to shutdown mode time) / (Active to shutdown mode energy):				
CNN (5x5 images, 4 class)	Dataset Not Avail.	No	110/0.1ms/0.5 µJ	Dataset Not Avail.
CNN (MNIST, 28x28 images)	42/5 µs/1.68 nJ	No	No	Yes*
SVM	44/1.5 µs/0.5 nJ	No	No	No
Linear Regression	42/5 µs/1.68 nJ	No	No	No
Kalman Filter	41/4 µs/1.34 nJ	No	No	No
SHA256 Hash	48/8 µs/2.69 nJ	No	No	No
Counter	24.2/0.5 µs/0.3 nJ	33/4µs - 1.02 ms/400 nJ	No	No
Active Power @ 25C (mW)	0.47 @ 0.71V	3.3	22	Not Reported
Fine-grained Temporal Power-gating demonstrated	Yes	Yes	Yes	N/A
Shutdown to Active mode time	200 ns	130 ns	50 ns	N/A
Shutdown to Active mode energy	152 pJ	450 pJ	510 pJ	N/A
RRAM Read/Write Latency (ns)	23/50	Not Reported	Not Reported	5/(Not Reported)
RRAM Read/Write energy (pJ/bit)	1.76/10.9 (Set) / 10.1 (Reset)	Not Reported /99	Not Reported /46.1	Not Reported
Resilience addressed by	Dynamic Addr. Remapping & ENDURER	None	None	None
Lifetime (years)	10	Not Reported	Not Reported	Not Reported

[6] Y. Liu et al., ISSCC, 2016. [7] F. Su et al., VLSI Circuits, 2017. [8] W. Chen et al., ISSCC, 2018.

* Values Not reported

(1) NVM: Non-volatile memory
(2) NV: Non-volatile

Figure 14.3.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

14.4 All-Digital Time-Domain CNN Engine Using Bidirectional Memory Delay Lines for Energy-Efficient Edge Computing

Aseem Sayal, Shirin Fathima, S. S. Teja Nibhanupudi, Jaydeep P. Kulkarni

University of Texas, Austin, TX

Convolutional Neural Networks (CNN) provide superior classification accuracy in a variety of machine learning applications, such as image/speech/sensor data processing. However, CNNs require intensive compute and memory resources making it challenging to employ in energy-constrained edge-computing devices. Specifically, Multiply-and-Accumulate (MAC) operations consume a significant portion of the total CNN energy [1].

Various analog compute techniques using charge manipulation schemes and A/D converters, as well as frequency-modulation-based approaches have been proposed to realize efficient MAC computations in a CNN accelerator (Fig. 14.4.6) [1-2, 5-6]. However, finite voltage headroom is required in analog MAC designs, whereas accurate frequency control is necessary for prior frequency-domain MAC approaches. This limits the voltage scalability of analog approaches and performance scalability of prior frequency-domain techniques degrading the MAC energy efficiency – critical for CNNs used in edge-compute devices. In this paper, we demonstrate an energy-efficient CNN engine implemented in a 40nm CMOS (Figs. 14.4.1, 14.4.7) featuring: 1) Bi-directional Memory Delay Lines (MDL) performing time-domain MAC operations; 2) multi-precision filter weight support (signed/unsigned 1-8b); 3) 16 filters each supporting 2×2 sub-sampling (max. pooling) and averaging; 4) all-digital, technology scalable design without requiring any capacitors, A/D converters, and/or frequency generators/modulators; and 5) near-threshold voltage operation supporting 16× speed-up with 4 input encoding modes.

The CNN data flow (Fig. 14.4.1) consists of: 1) training input data using TensorFlow/Keras software framework; 2) feeding trained filter weights and test input data to the test-chip using LabVIEW-PXI data acquisition instruments; 3) performing on-chip MAC, averaging, and pooling operations for the test input data; and 4) fully connected layers and soft-max computation in software (TensorFlow). The time-domain MAC computations are realized using the proposed MDL which accumulates the dot product of a weight bit and the time encoded input pulse-width (Fig. 14.4.2). It is derived from the concept of a 'time-register' used in high precision time-to-digital converters, which can perform time addition and time storage using a string of delay cells controlled by an enable (EN) signal [7]. Each MDL unit comprises two cross-coupled inverter pairs, S1-S4 switches, and a reset control (RST). The time-encoded dot product of input (X_i) and 1b filter weight (w_i) acts as an EN pulse and controls the MDL operating mode. During the accumulation phase, EN=1 and the MDL acts either as a forward delay line (for +ve dot product, S1 and S2 are ON) or a backward delay line (for -ve dot product, S2, S3 and S4 are ON), enabling bi-directional data flow emulating signed dot products. When EN goes low, the MDL acts as a memory storage line and retains the MDL state vector using cross-coupled inverters (S1, S3 and S4 are ON). The metastability risk during an EN falling transition is resolved by the next incoming EN pulse, as the MDL is transformed into a chain of cascaded delay cells. When the MDL state vector string progresses towards the either end of the MDL (node A or node E), an up-down counter is triggered, which translates time-domain dot product accumulation information into digital bits acting as a time-to-digital converter. If the accumulated dot product pulse-width exceeds the full-scale MDL delay, an overflow condition is detected, and the propagating edge is inverted (using S5-S6) and applied at the beginning node A of MDL (or trailing end node E). Thus, a finite length MDL can be used to perform long-duration time-domain accumulation using an up-down counter. The calibration unit consists of additional delay cells which can be added to the MDL to mitigate delay mismatch in the presence of process variations.

The bi-directional MDL forms the core of the CNN engine implementing 16 filters. Each filter consists of 4 bi-directional MDLs, a weight shift register, a shared pulse generator/selector, an up/down counter, a bi-directional barrel shifter, and pooling comparators (Fig. 14.4.3). 8b input data (X) is represented in the time domain as a Pulse-Width Modulated (PWM) signal as multiples of input clock period ($2t_0$). A pulse generator/selector circuit is designed to generate 0-$255t_0$ PWM signals using a two-stage approach [1]. As 4 MSBs ($X[7\text{-}4]$) correspond to a maximum of value of 240 (out of the 255 full-scale value of an 8b input), an MSB_EN signal is asserted for $240t_0$ duration in the first stage generating T0-T15 output pulses in increments of $16t_0$ duration. In the second step, the MSB_EN signal is de-asserted and 4 LSBs ($X[3\text{-}0]$) PWM signals are generated as outputs T0-T15. Four precision modes are implemented to support 1-to-16× speed-up in the input PWMs by quantizing 4 LSBs. As shown in the pulse generator timing diagram (Fig. 14.4.3), varying duration T0-T15 pulses are generated based on the precision mode and concatenated with $X[7\text{-}4]$ MSB pulses generated in the first step. The LSB pulse-width quantization steps are chosen to limit the quantization error to ±0.5*speed-up ratio. As the pulse generator operates continuously in every MAC_CLK period, pulse gating AND logic is implemented to ensure that only a single PWM input pulse is applied to the MDL in each dot product computation. Next, the single PWM input is multiplied with a 1b filter weight (w_i) stored in a 25b shift register (for a 5×5 filter size) producing the $X_i \bullet w_i$ dot product. The bi-directional MDL then performs signed accumulation, time-to-digital conversion (20b-up/down counter) and averaging/scaling (20b-bi-directional barrel shifter performing up to 7b shifts). For multi-bit filter weights, multiple instances of MDLs with each weight shift register initialized with one bit of the weight vector can be used. A sub-sampling operation using max pooling across a 2×2 window is implemented to reduce the convolution layer output data size by 75%. This is achieved by 4 concurrent MDL operations and feeding the MDL counter outputs to 8b MAX comparators. The pooled output from each filter is stored off-chip and reused as the input to the next convolution layer.

Figure 14.4.4 shows oscilloscope-captured waveforms from the 40nm CMOS test-chip (Fig. 14.4.7) confirming MDL functionality with delay phase and storage phase for different MDL lengths. The pulse generator and the pulse-gating logic functionality is verified with the correct toggling of MSB_EN, T15, T8, T4 outputs and the pulse-gating control signals. 1-to-16× speed-up in the PWM input representation is validated with multiple precision modes for a test-case input of 214. The measured classification accuracy (on 100 images) for LeNet-5 using the MINIST dataset is ~2% lower relative to software counterpart (Fig. 14.4.5). 16× speed-up mode resulted in lower accuracy because of input quantization and increased sensitivity of MDL residue. Simulation results using the proposed MDL-based CNN for AlexNet with 2-class ImageNet dataset (cats *vs.* dogs) with signed 8b weights, shows 13% lower classification accuracy compared to software (16b floating-point weights). The CNN engine is operable down to 375mV with more than 90% MNIST classification accuracy. For the LeNet-5 case, both C1- and C3-layer throughput increases with the higher speed-up mode and with the increasing supply voltage achieving a peak throughput of 0.38 (0.128) GOPS for the C3 (C1) layer at 585mV. The energy efficiency peaks with increasing supply voltage scaling and reaches a maximum of 13.46 (4.61) TOPS/W for the C3 (C1) layer at 496mV. Figure 14.4.6 tabulates LeNet-5 parameters, test-chip characterization results (at optimal voltage of 537mV), and compares with earlier approaches [1-6].

Acknowledgements:
Authors thank Vignesh Radhakrishnan and Jacob Rohan for helping with the test-chip measurements. Authors also like to thank TSMC University shuttle program for the test-chip fabrication support.

References:
[1] A. Biswas, et al., "Conv-RAM: An Energy-Efficient SRAM With Embedded Convolution Computation For Low-Power CNN-Based Machine Learning Applications," *ISSCC*, pp. 488-490, 2018.
[2] S. K. Gonugondla, et al., "A 42pJ/decision 3.12TOPS/W Robust In-Memory Machine Learning Classifier With On-Chip Training," *ISSCC*, pp. 490-492, 2018
[3] J. Sim, et al., "A 1.42TOPS/W Deep Convolutional Neural Network Recognition Processor for Intelligent IoE Systems," *ISSCC*, pp. 264-265, 2016.
[4] B. Moons, et al., "A 0.3–2.6 TOPS/W Precision-Scalable Processor for Real-Time Large-Scale ConvNets," *IEEE Symp. VLSI Circuits*, pp. 178-179, 2016.
[5] M. Liu, et al., "A scalable time-based integrate-and-fire neuromorphic core with brain-inspired leak and local lateral inhibition capabilities," *IEEE CICC*, pp. 1-4, 2017.
[6] A. Amravati, et al., "A 55nm Time-Domain Mixed-Signal Neuromorphic Accelerator With Stochastic Synapses and Embedded Reinforcement Learning For Autonomous Micro-Robots," *ISSCC*, pp. 124-126, 2018.
[7] K. Kim, et al., "A 9 bit, 1.12ps Resolution 2.5b/Stage Pipelined Time-to-Digital Converter in 65 nm CMOS Using Time-Register," *IEEE JSSC*, vol. 49, no. 4, pp. 1007-1016, 2014.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

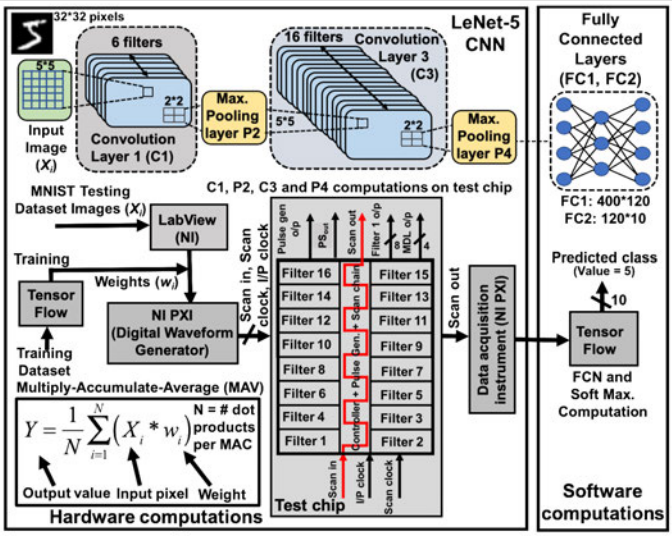


Figure 14.4.1: LeNet-5 CNN engine data flow and its interface with the test-chip which implements MAC, averaging, and pooling operations.

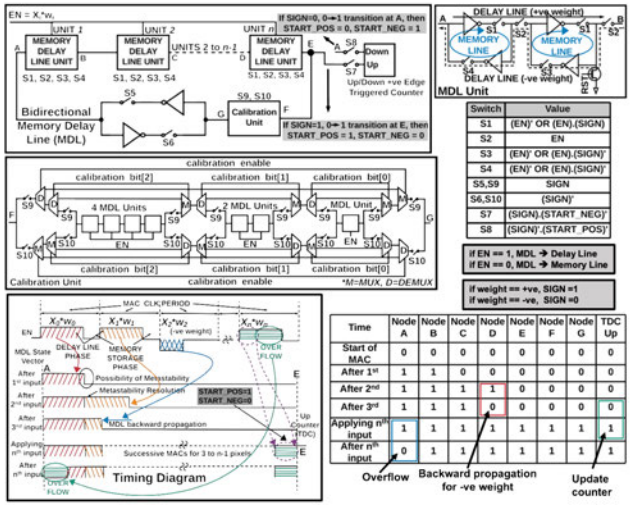
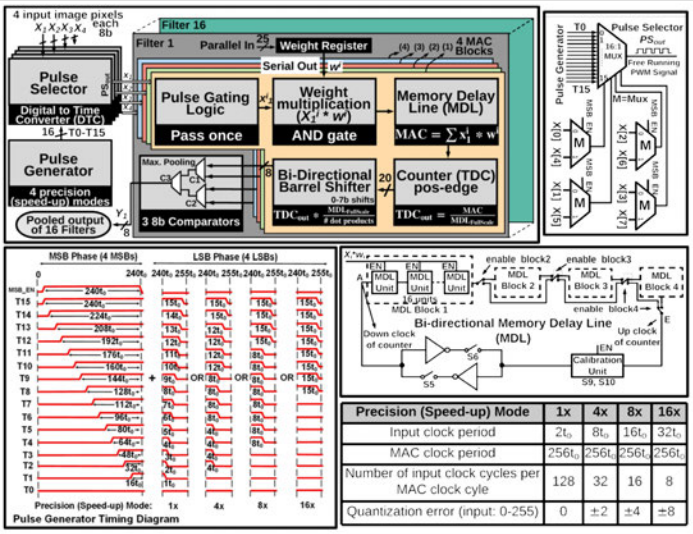


Figure 14.4.2: Proposed bi-directional Memory Delay Line concept (MDL) for time-domain, signed MAC computation: circuit topology, switch-selection logic, calibration unit, and timing diagram.



Precision (Speed-up) Mode	1x	4x	8x	16x
Input clock period	$2t_0$	$8t_0$	$16t_0$	$32t_0$
MAC clock period	$256t_0$	$256t_0$	$256t_0$	$256t_0$
Number of input clock cycles per MAC clock cyle	128	32	16	8
Quantization error (input: 0-255)	0	±2	±4	±8

Figure 14.4.3: Overall architecture of the proposed MD- based CNN engine with input PWM pulse generator supporting 1-to-16× speed-up modes.

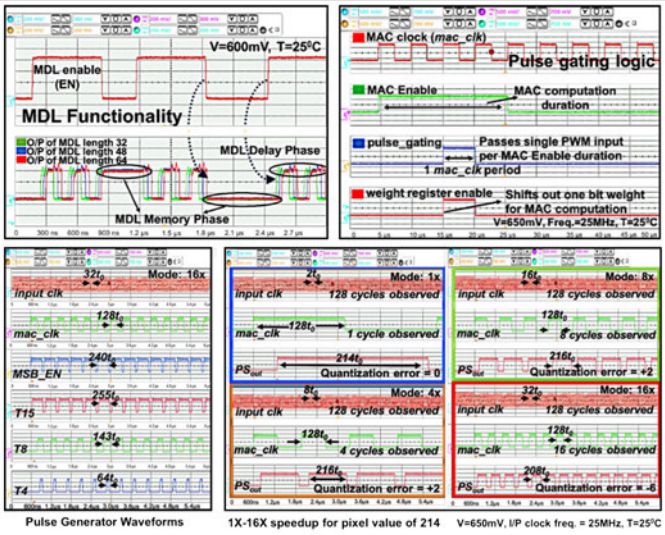


Figure 14.4.4: Experimental demonstration of the MDL functionality, pulse generator/pulse gating logic, and 1-to-16× speed-up with input quantization.

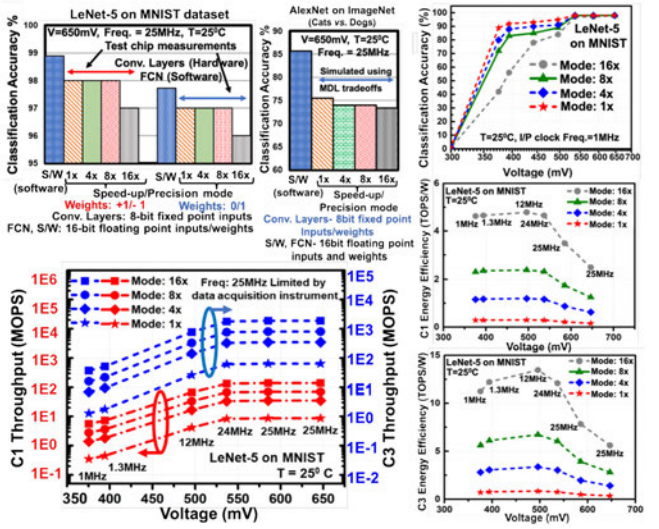


Figure 14.4.5: Measured throughput, energy efficiency, and classification accuracy for 1-to-16× speed-up modes on LeNet-5 using MNIST dataset.

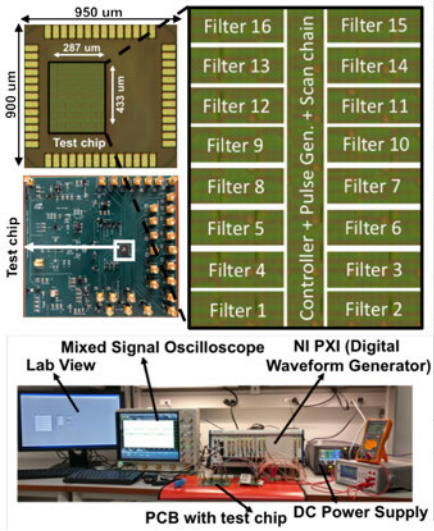
Parameters for LeNet-5	C1+P2	C3+P4
Filter Size	5*5*1*6	5*5*6*16
Input/Filter Size	8bits/ 1bit	8bits/ 1bit
Input Size	32*32*1	14*14*6
Output Size	14*14*6	5*5*16
#Filters	6	16
#Operations/ convolution*	(25*4*6) *2	(150*4* 16)*2
#MAC clock cycles/conv	28	158

LeNet-5 Results/Metrics	Convolution Layer C1 and Pooling Layer P2				Convolution Layer C3 and Pooling Layer P4			
Precision (Speed-up) Mode	1x	4x	8x	16x	1x	4x	8x	16x
Input clock frequency (MHz)	24.0	24.0	24.0	24.0	24.0	24.0	24.0	24.0
MAC clock frequency (MHz)	0.19	0.75	1.50	3.00	0.19	0.75	1.50	3.00
Convolution Cycle Time (us)	149.3	37.33	18.67	9.33	842.67	210.67	105.33	52.67
Operating Voltage (V)	0.537	0.537	0.537	0.537	0.537	0.537	0.537	0.537
Power (uW)	28.67	28.67	28.67	28.67	30.17	30.17	30.17	30.17
Throughput (GOPS)	0.008	0.032	0.064	0.128	0.023	0.091	0.183	0.365
Energy Efficiency (TOPS/W)	0.29	1.16	2.33	4.65	0.76	3.02	6.04	12.08

*Assuming 1 Multiply-Accumulate-Average, and 1 Pooling = 2 operations
**Scalable to multi-bit weights

Reference	Tech. (nm)	Circuit Type	Input/ Weight Size	Chip Size (mm²)	Pooling	Low Vcc Support	Capaci-tors or ADCs	Classifi-cation Accuracy	Throughput (GOPS)	Power (uW)	Energy Efficiency (TOPS/W)
ISSCC'18 [1]	65	Analog	6bits/1bit	0.067	No	No	Yes	96%	10.70	380.7	28.10
ISSCC'18 [2]	65	Analog	8bits	1.440	No	No	Yes	96%	-	-	3.125
ISSCC'16 [3]	65	Digital	16bits	16.000	Yes	No	No	98.3%	64	4.51E+4	1.42
VLSI'16 [4]	40	Digital	6bits/4bits	2.400	No	Yes	No	98%	102	3.9E+4	2.60
CICC'17 [5]	65	Time	8bits/3bits	0.24	No	Yes	Yes	91%	0.396	2.05E+4	0.019
ISSCC'18 [6]	55	Time	6bits/6bits	3.125	No	Yes	No	-	2.152	690	3.12
This work (MDL CNN)	**40**	**Time**	**8bits/ 1bit****	**0.124**	**Yes**	**Yes**	**No**	**97%**	**0.365**	**30.17**	**12.08**

Figure 14.4.6: Performance summary for C1+P2 and C3+P4 convolutional + pooling layers; proposed MDL comparison with prior approaches.



Technology	40nm CMOS
Area	0.124mm^2
# standard cells	~40,000
# of filters supported	16
Data type supported	Signed/ unsigned
Voltage Range	0.375-1.1V
Input clock frequency	25MHz
MAC clock frequency	0.2-3.1MHz
Energy Efficie-ncy @537mV	12.08 TOPS/W
Throughput @537mV	0.365 GOPS
Power @537mV	30.17μW
LeNet-5 MNIST classification accuracy	97%

Figure 14.4.7: Test-chip die micrograph, characterization setup, and measurements summary table.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

14.5 A 0.6-to-1.1V Computationally Regulated Digital LDO with 2.79-Cycle Mean Settling Time and Autonomous Runtime Gain Tracking in 65nm CMOS

Xun Sun, Akshat Boora, Wenbing Zhang, Venkata Rajesh Pamula, Visvesh Sathe

University of Washington, Seattle, WA

Low-Dropout Regulators (LDOs) play an important role in enabling fine-grained supply-voltage domains for energy-efficient SoC design [1]. Digital LDOs are of particular interest due to integration and scalability advantages, but their transient response is slowed down by intrinsic limitations in sampled feedback systems. Design margins to ensure stability across worst-case PVT conditions further degrade transient response. Meanwhile, voltage domains continue to shrink in size, thus mandating a faster LDO response to compensate for reduced available decoupling capacitance (decap).

Recently reported non-linear control and event-driven architectures offer fast recovery times [2] [3]. However, non-linear approaches face the challenge of ensuring stable mode transitions under random load current (I_L) conditions. Event-driven LDOs trigger logic to control MOS devices based on threshold crossings made by the regulated voltage (V_{out}). However, typical digital systems exhibit constant load fluctuation, which can result in prohibitive switching losses. To address the impact of worst-case margining, adaptive LDO designs have also been proposed, but they largely focus on suppressing V_{out} ripple [4], and compensating for load current variation [5].

This paper presents computational regulation, a technique for fast and stable transient response across PVT. This concept is demonstrated in a Digital LDO that drives a Cortex-M0 processor with an integrated linear algebra accelerator (Fig. 14.5.1). The key idea is to: 1) derive time-domain models that are more accurate than those obtained from the traditional discrete-time transfer function and 2) evaluate the resulting state equations at runtime for rapid regulator response. We also introduce Autonomous Gain Tracking (AGT), a low-overhead, low-complexity technique that examines V_{out} statistics for runtime loop-gain tuning to enable rapid LDO response across PVT. In any cycle *n*, the LDO samples and quantizes the error voltage $\Delta V[n] = V_{out}[n] - V_{ref}$. A *Solver* then uses $\Delta V[n]$ to evaluate LDO state equations to determine $k[n]$, the number of PMOS devices to be turned-on in the same cycle. Ideally, $k[n]$ is selected to meet I_L requirements and restore V_{out} within one LDO cycle. To determine $k[n]$, the *Solver* requires a well-tuned gain term, $G_S = C_L/(I_0T)$ (Fig. 14.5.1) that captures important design parameters. These parameters include: 1) I_0, the current delivered by a unit LDO PMOS device, 2) C_L, the V_{out} decap, and 3) T, the LDO clock period. G_S is therefore sensitive to PVT variation and must be tracked for rapid, robust response. AGT monitors the LDO loop gain and subsequently adjusts G_S to maintain speed and stability.

Figure 14.5.2 details the LDO architecture. The error between V_{out} and V_{ref} is first quantized by 13 clocked comparators. Non-uniform quantization offers a more effective trade-off between resolution and range [6]. To achieve sub-cycle loop delay, the LDO updates the number of conducting PMOS devices immediately after the *Solver* determines $k[n]$. Both comparator and *Solver* outputs are latched using suitably delayed clocks to provide sufficient evaluation time for each module. Although this approach suppresses loop delay to a fraction of the LDO clock (αT_{LDO}), loop delay is not eliminated. Consequently, the effective number of conducting PMOS devices in cycle *n* depends on both $k[n]$ and $k[n-1]$, an effect that needs to be modeled in fast LDOs. The implemented LDO *Solver* (Fig. 14.5.2) incorporates this dependence for improved droop response and stability.

Rapid, stable LDO operation requires good matching of G_S between that modeled by the solver ($G_{S,model}$) and the actual PVT-dependent G_S ($G_{S,actual}$). Any mismatch degrades either LDO stability or speed. The AGT loop (Fig. 14.5.3) acts as a low-bandwidth digital servo that uses the lag-1 autocorrelation of V_{out}, R[1], to characterize the LDO response and suitably update $G_{S,model}$ to track $G_{S,actual}$. Fig. 14.5.3 shows how the sign of R[1] reflects LDO response: a negative (positive) value of $R[1]$ suggests an under-damped (over-damped) I_L step response, indicating that $G_{S,model} > G_{S,actual}$ ($G_{S,model} < G_{S,actual}$). $R[1] = 0$ indicates an ideal response with no G_S mismatch. This relationship can be proven to extend beyond a unit-step load to random runtime current loads. Because only the sign of R[1] is needed, the AGT performs a bit-wise XOR of the MSBs of successive $\Delta V[n]$ values. An up-down counter accumulates the XOR result over 40 samples to account for statistical variation. Accumulated totals with sufficient magnitude reflect a systematic positive or negative trend in $R[1]$, causing the loop to update G_S.

The proposed LDO was fabricated in a low-power 65-nm CMOS process. All modules, except the PMOS header devices and comparators, were synthesized. The load current is provided by processor/accelerator execution and an on-chip I_L step generator. The test-chip die photograph is shown in Fig. 14.5.7. 250pF of decap was added to V_{out}. Both processor and accelerator operate at the same frequency.

Figure 14.5.4 shows the measured LDO response to a 5.6mA/100ps I_L load step. Also shown is the response obtained by separately disabling loop-delay modeling and AGT. The responses were obtained under nominal conditions (V_{in}=1.1V, V_{out}=1.0V) after allowing the AGT loop to track G_S during processor execution—no manual tuning of *Solver* parameters was performed. The LDO requires very few cycles to settle, achieving a mean settling time of 2.79 cycles. Single-cycle settling requires addressing variation in the alignment between the load step and LDO clock edges, as well as non-uniform quantization. Without loop-delay modeling, both droop response and stability are degraded. With V_{in}=1.1V, V_{out} was then scaled by 150mV to 0.85V under two scenarios: with and without enabling AGT. AGT allows a near-ideal step response at V_{out} = 0.85V. Without AGT, however, the LDO is unstable. The 150mV V_{out} reduction requires reducing G_S by 50% to stabilize the loop which in turn doubles the measured voltage droop at V_{out} = 1.0V.

Figure 14.5.5 shows the AGT loop modifying $G_{S,model}$ to track $G_{S,actual}$ by examining V_{out} statistics during processor execution. The experiment was repeated using three distinct initial $G_{S,model}$ values. Each point on the curve represents a G_S update based on a 40-sample R[1] accumulation. All three curves converge to a common level that lies within 10% of the optimal G_S. Measurements indicate that the AGT successfully performs the broad gain adjustments needed for consistently fast transient response across variation in V_{in}, V_{out} and temperature. These broad adjustments demonstrate the impact of margins on transient response.

Measured load regulation for V_{in} in the 0.65-to-1.0V range and regulator current efficiency are shown in Fig. 14.5.6. Also included is a comparison to related prior work. Regulator switching losses are dominated by *Solver* dynamic power and effectively amortized in designs with higher I_L. AGT losses make up only 5% of total LDO switching losses.

Computational Regulation presents a new approach to exploiting advances in computing performance and efficiency to realize LDOs with stable and fast transient response across PVT. The proposed autocorrelation-based autonomous gain tracking architecture tracks PVT-induced gain variations, reducing required stability margins and consistently enabling rapid transient response. Computational regulation is expected to offer enhanced efficiency and performance in advanced CMOS nodes.

Acknowledgments:
The authors thank Nasser Kurd, Praveen Mosalikanti, Carlos Tokunaga, Alvin Loke and Sanjay Pant for their valuable input. Funding for this work was provided by grants from Intel, and from SRC (TxACE) under grant No. 2712.006.

References:
[1] Z. Toprak-Deniz, et al., "Distributed System of digitally Controlled Microregulators Enabling Per-Core DVFS for the POWER8TM Microprocessor," *ISSCC,* pp. 98-99, 2014.
[2] T. Mahajan, et al., "Digitally Controlled Voltage Regulator Using Oscillator-Based ADC with Fast-Transient-Response and Wide Dropout Range in 14nm CMOS," *CICC*, pp. 1-4, 2017.
[3] D. Kim, et al., "A 0.5V-VIN1.44mA-class Event-Driven Digital LDO with a Fully Integrated 100pF Output Capacitor," *ISSCC*, pp. 346-347, 2017.
[4] W. Tsou, et al., "Digital Low-Dropout Regulator with Anti PVT-Variation Technique for Dynamic Voltage Scaling and Adaptive Voltage Scaling Multicore Processor," *ISSCC,* pp. 338-339, 2017.
[5] S. B. Nasir, et al., "A 0.13μm Fully Digital Low-Dropout Regulator with Adaptive Control and Reduced Dynamic Stability for Ultra-Wide Dynamic Range," *ISSCC*, pp. 1-3, 2015.
[6] H. Hu, et al., "Nonuniform A/D Quantization for Improved Dynamic Responses of Digitally Controlled DC–DC Converters," *IEEE Transactions on Power Electronics,* vol. 23, no. 4, pp. 1998-2005, 2008.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

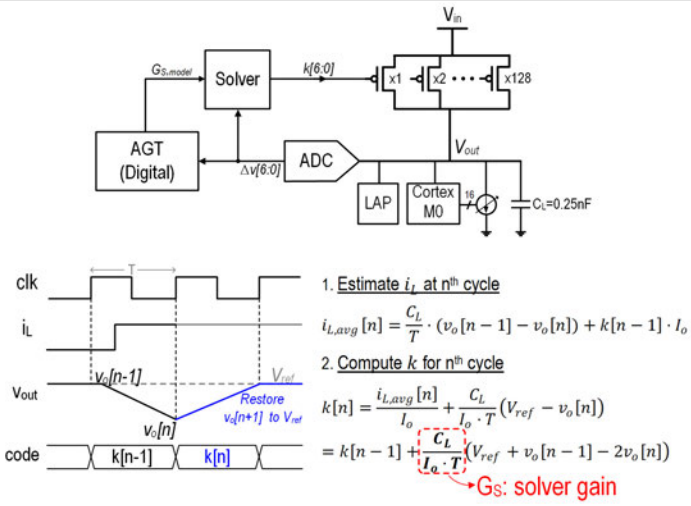


Figure 14.5.1: A computationally regulated digital LDO with low-bandwidth Autonomous Gain Tracking (AGT) loop. Intended operation and simplified equations.

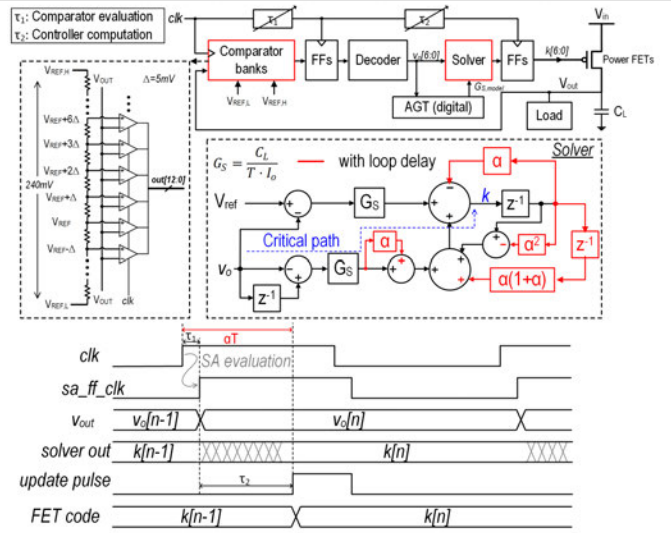


Figure 14.5.2: Detailed LDO block diagram and timing diagram of same-cycle update using sub-cycle (αT) loop delay.

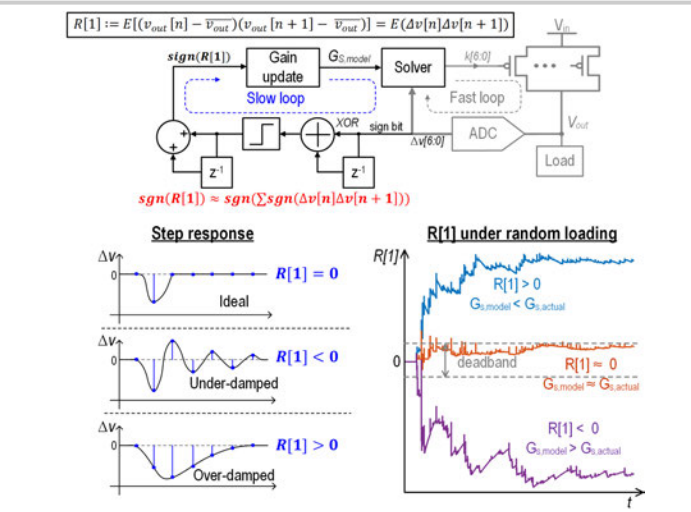


Figure 14.5.3: AGT tracks loop gain across PVT. The lag-1 auto-correlation (R[1]) sign effectively characterizes loop gain for $G_{S,model}$ update under load I_L step and random load conditions.

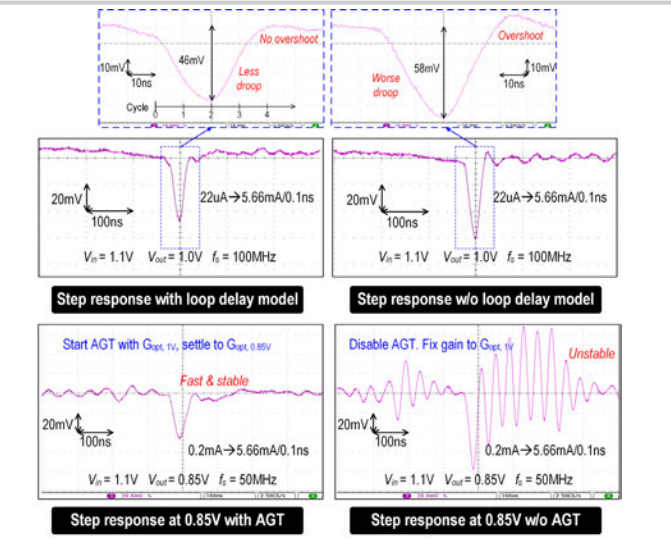


Figure 14.5.4: Measured LDO step-response under regular operating configuration, and with loop-delay modeling and AGT separately disabled.

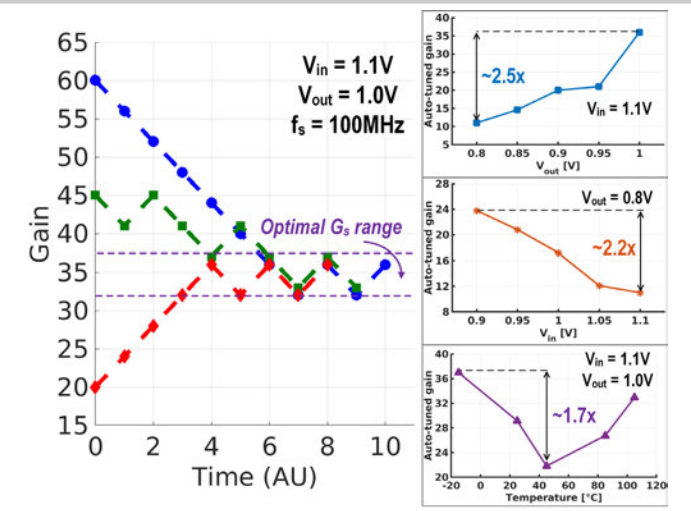
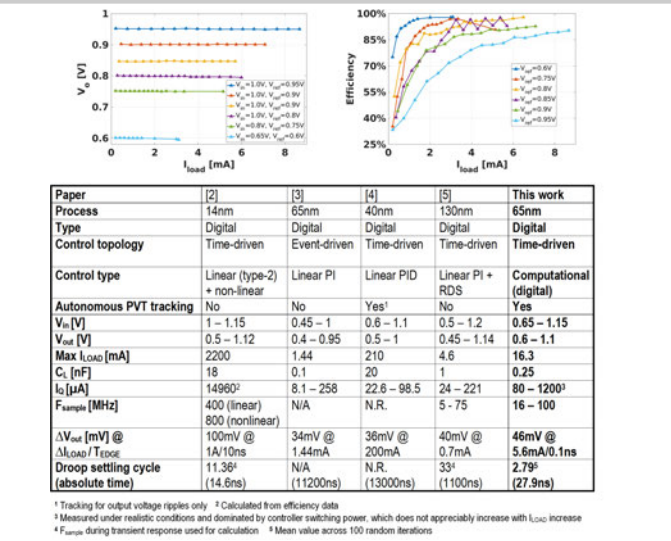


Figure 14.5.5: (Left) Runtime $G_{S,model}$ traces converging to $G_{S,actual}$ from distinct incorrect initial values. (Right) Measured $G_{S,model}$ across different V_{out}, V_{in} and temperature indicate the need for significant adjustment.



Paper	[2]	[3]	[4]	[5]	This work
Process	14nm	65nm	40nm	130nm	65nm
Type	Digital	Digital	Digital	Digital	Digital
Control topology	Time-driven	Event-driven	Time-driven	Time-driven	Time-driven
Control type	Linear (type-2) + non-linear	Linear PI	Linear PID	Linear PI + RDS	Computational (digital)
Autonomous PVT tracking	No	No	Yes[1]	No	Yes
V_{in} [V]	1 – 1.15	0.45 – 1	0.6 – 1.1	0.5 – 1.2	0.65 – 1.15
V_{out} [V]	0.5 – 1.12	0.4 – 0.95	0.5 – 1	0.45 – 1.14	0.6 – 1.1
Max I_{LOAD} [mA]	2200	1.44	210	4.6	16.3
C_L [nF]	18	0.1	20	1	0.25
I_Q [μA]	14960[2]	8.1 – 258	22.6 – 98.5	24 – 221	80 – 1200[3]
F_{sample} [MHz]	400 (linear) 800 (nonlinear)	N/A	N.R.	5 - 75	16 – 100
ΔV_{out} [mV] @ ΔI_{LOAD} / T_{EDGE}	100mV @ 1A/10ns	34mV @ 1.44mA	36mV @ 200mA	40mV @ 0.7mA	46mV @ 5.6mA/0.1ns
Droop settling cycle (absolute time)	11.36[4] (14.6ns)	N/A (11200ns)	N.R. (13000ns)	33[4] (1100ns)	2.79[5] (27.9ns)

[1] Tracking for output voltage ripples only [2] Calculated from efficiency data
[3] Measured under realistic conditions and dominated by controller switching power, which does not appreciably increase with I_{LOAD} increase
[4] F_{sample} during transient response used for calculation [5] Mean value across 100 random iterations

Figure 14.5.6: (Top) Load regulation and Current efficiency. (Bottom) Comparison with related works.

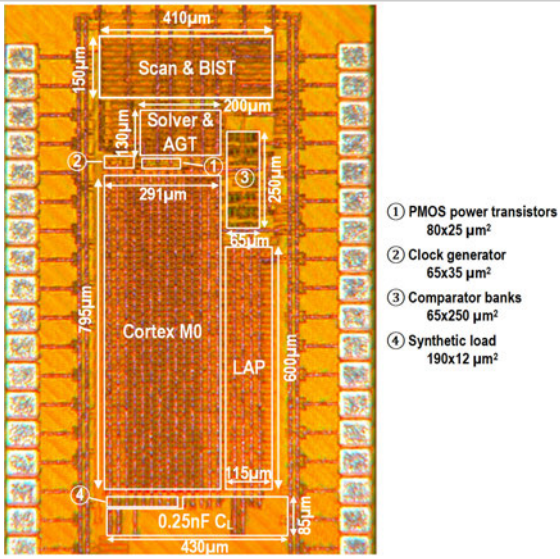


Figure 14.5.7: Die photograph of the LDO, implemented in 65nm CMOS.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

14.6 A 745pA Hybrid Asynchronous Binary-Searching and Synchronous Linear-Searching Digital LDO with 3.8×10^5 Dynamic Load Range, 99.99% Current Efficiency, and 2mV Output Voltage Ripple

Shuo Li, Benton H. Calhoun

University of Virginia, Charlottesville, VA

Voltage regulators for emerging nW-to-μW Internet-of-Things (IoT) systems-on-chip (SoCs) require ultra-low quiescent power to enhance lifetime, a large dynamic load range to support multiple components and applications, and fast transient performance to support duty-cycled loads. Switching regulators [1,2] can achieve sub-nW operation, but suffer from slow settling time and large output voltage ripple that deteriorate analog and RF performance. Digital low-dropout regulators (DLDOs) offer faster transient response, low-voltage operation, and flexible control schemes [3-6]. However, conventional synchronous DLDOs that use a high frequency clock for fast response consume too much power [3], and asynchronous DLDOs can achieve lower current consumption and fast response, but suffer from larger steady-state error [4]. Also, previous DLDOs do not target load current in the <10μA range. To achieve ultra-low-current consumption, a wide dynamic load range down to nA loads, and good transient and steady-state performance, we propose a hybrid asynchronous binary-searching (ABS) and synchronous linear-searching (SLS) DLDO that achieves 3.8×10^5 dynamic load range, 99.99% current efficiency, and 2mV output voltage ripple, while consuming 745pA quiescent current.

Figure 14.6.1 shows the architecture of the proposed hybrid DLDO, which comprises two asynchronous (async) comparators, one synchronous (sync) comparator, a hybrid ABS/SLS controller, a 14b binary-weighted PMOS switch array, a common-mode voltage (VCM) generator, and a clock management block. The two async comparators set the upper and lower thresholds of the dead zone (DZ) around V_{REF}, monitor V_{OUT}, and trigger ABS after a load current change. The StrongARM-based sync comparator helps V_{OUT} converge to V_{REF} during SLS and toggles the SRL bit during steady-state operation. The operation timing diagram (Fig. 14.6.1) shows how a load step change that moves V_{OUT} outside the DZ toggles VOH or VOL and triggers ABS operation. A 14b binary-weighted PMOS array is adopted to provide wide dynamic load range and low output ripple. After ABS settles V_{OUT} back in the DZ, the DLDO enters SLS operation to settle into steady state, where only M_{SRL} toggles. The hybrid control facilitates ultra-low current consumption for several reasons. First, ABS/SLS provides good transient and steady-state performance without requiring a fast clock. Second, steady-state SLS mode only toggles the M_{SRL} instead of driving the whole PMOS array, which lowers dynamic power. Finally, the low-speed sync path uses IO FETs to reduce leakage, and the async path uses HVT FETs for higher speed.

Figure 14.6.2 shows the details of the proposed hybrid ABS/SLS control scheme. The async ABS controller conducts binary searching with the 14b PMOS array, guided by VOH and VOL from the async comparators. For example, in the i-th comparison cycle, if V_{OUT} is below the lower threshold, VOL goes low and the controller turns on the whole PMOS array to drive V_{OUT} up into the DZ, then, when VOL goes high, it sets SR[14-i] to 0 and presets SR[14-i-1] to 0 before starting the next comparison cycle. If a SAR search cycle trips VOH, then SRS asserts to pull V_{OUT} downward back into the DZ before the next cycle, compensating for an otherwise slow response to the comparison due to light loads. The ABS-to-SLS transition is triggered by either finishing the full 14b binary search or waiting for a fixed number (3 by default) of clock cycles once V_{OUT} is settled inside the DZ. The transient measurement result in Fig. 14.6.2 shows a case when V_{OUT} settles in the DZ after searching the first 4b, so the DLDO switches to SLS after a timeout of three clock cycles. During SLS, V_{OUT} will linearly converge to the reference voltage and then toggle SRL only, which minimizes steady-state error at a very low power cost. During SLS operation, the VLK control bit effectively adds a small load current to offset the leakage of the PMOS array, ensuring that the steady-state operation can still work properly even during ultra-light loads when SR[13:0] = 14'b1, which increases the dynamic load range of the DLDO.

The async comparator determines the transient performance of the hybrid DLDO, so it must tradeoff a small propagation delay for rapid response with ultra-low current for low power. Fig. 14.6.3 shows the async comparator circuit, which adopts an inverter-based and self-biased switched-capacitor topology. This comparator is highly current efficient and can achieve a well-balanced trade-off between power and delay. Measurement results show that the CMPL has a falling-edge delay of 25.7μs and rising-edge delay of 15.4μs, while only consuming 111pA bias current. The offset of the comparators (tunable by 3b) sets the DLDO DZ thresholds, eliminating the need for multiple voltage references and keeping the system power low. The refresh operation for the comparator's capacitors is only enabled during SLS, so it does not affect the ABS control. A load current step occurring in the refresh phase is handled by the ABS after the ~1μs-long refresh completes, imposing minimal penalty on the response and settling time of the DLDO. A leakage-based current-starved ring oscillator generates the low-power clock, and the ultra-low-leakage switches allow the refresh frequency to reduce to a quarter of the clock frequency.

The hybrid DLDO is fabricated in 65nm. Fig. 14.6.4 shows the transient response with a load current step from 710pA to 270μA. V_{OUT} settles inside the DZ in the first SAR search cycle for 270μA, and the ABS executes the full 14b search during its transition back to 710pA current load. The transient waveform in Fig. 14.6.4 (bottom-left) shows how the ABS, SLS, and SRL toggling work for a load step change. When working at V_{IN} = 500mV, V_{OUT} = 450mV, and F_{CLK} = 464Hz, the DLDO has a total current consumption of 745pA, and Fig. 14.6.5 shows the measured current breakdown. Due to ultra-low current consumption, the DLDO still can keep a high current efficiency even under ultra-light load current conditions. The peak current efficiency is 99.99%, and the DLDO maintains a current efficiency >90% from 8nA to 270μA, which provides over four orders of magnitude of load current range.

Figure 14.6.6 compares the proposed hybrid ABS/SLS DLDO to other DLDOs, which have not previously approached the pA current consumption range. The hybrid DLDO maintains a high current efficiency (η > 90%) over a load range that is >100× larger than prior-art and consumes >100× less quiescent current. Thanks to the proposed hybrid control scheme, highly power-efficient comparators, and 14b PMOS array, our design achieves the lowest current consumption, the highest current efficiency, the largest dynamic load range, the lowest load current, and the smallest output voltage ripple for DLDOs. Fig. 14.6.7 shows the chip micrograph and a comparison with non-LDO, pW-consumption voltage regulators. Compared with other pW voltage regulators, the hybrid DLDO has >20× faster settling time, >7× smaller output ripple, and >20× better load regulation while maintaining ultra-low power consumption and a wide dynamic load range. These attributes make the design well suited for the power management on ultra-low-power IoT SoCs.

Acknowledgements:
This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Award Number DE-EE0008225 and by the NSF ASSIST Center (EEC-1160483).

References:
[1] S. Bandyopadhyay, et al., "A 1.1 nW Energy-Harvesting System with 544 pW Quiescent Power for Next-Generation Implants," *IEEE JSSC*, vol. 49, no. 12, pp. 2812-2824, 2014.
[2] A. Paidimarri and A. P. Chandrakasan, "A Wide Dynamic Range Buck Converter With Sub-nW Quiescent Power," *IEEE JSSC*, vol. 52, no. 12, pp. 3119-3131, 2017.
[3] L. G. Salem, et al., "A 100nA-to-2mA Successive-Approximation Digital LDO with PD Compensation and Sub-LSB Duty Control Achieving a 15.1ns Response Time at 0.5V," *ISSCC*, pp. 340-341, 2017.
[4] D. Kim, et al., "A 0.5V-VIN 1.44mA-Class Event-Driven Digital LDO with a Fully Integrated 100pF Output Capacitor," *ISSCC*, pp. 346-347, 2017.
[5] L. G. Salem and P. P. Mercier, "A Sub-1.55mV-Accuracy 36.9ps-FOM Digital-Low-Dropout Regulator Employing Switched-Capacitor Resistance," *ISSCC*, pp. 312-314, 2018.
[6] J. Lin, et al., "A High-Efficiency and Fast-Transient Digital-Low-Dropout Regulator with the Burst Mode Corresponding to the Power-Saving Modes of DC-DC Switching Converters," *ISSCC*, pp. 314-316, 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

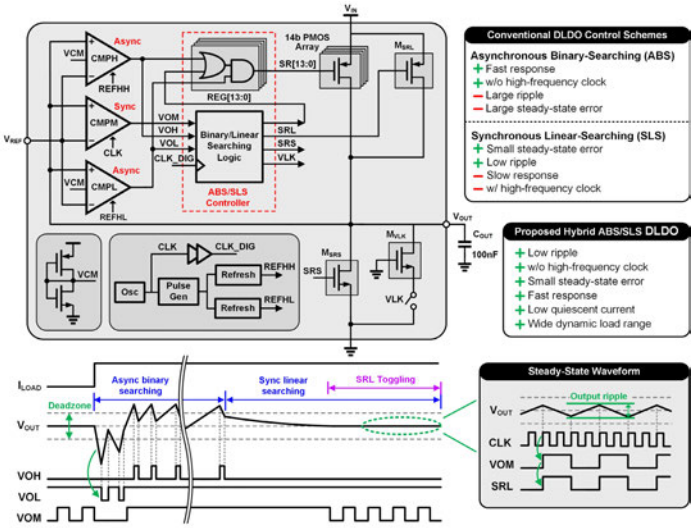


Figure 14.6.1: Architecture of the proposed hybrid ABS/SLS DLDO and its operation timing diagram.

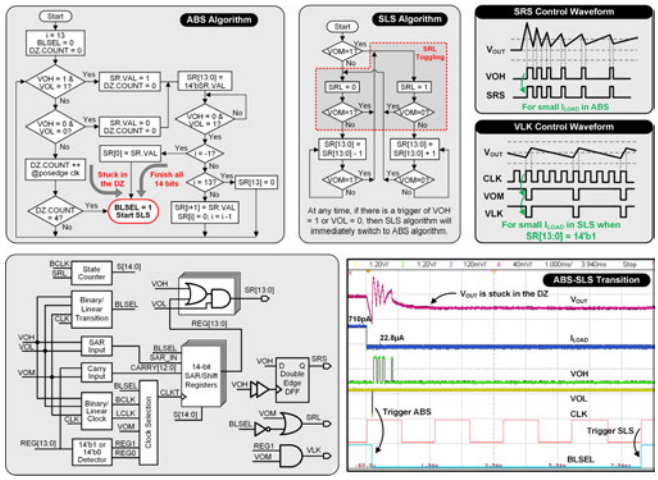


Figure 14.6.2: Flow chart of the ABS and SLS algorithm (top-left/middle); operation diagram of the SRS and VLK control bits (top-right); block diagram of the ABS/SLS controller and measured V_{OUT} transient waveform illustrating the ABS-SLS transition mechanism (bottom).

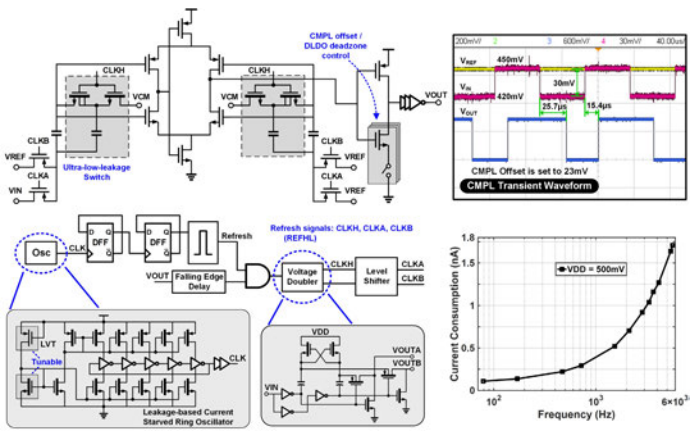


Figure 14.6.3: Structure of the async comparator and its measured transient waveform (top); block diagram of the clock and refresh block and their measured total current consumption vs. frequency (bottom).

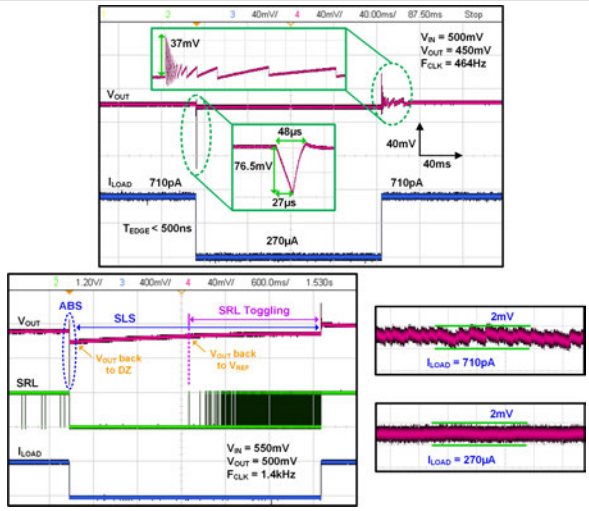


Figure 14.6.4: Measured transient response with a load change from 710pA to 270μA (top); measured load transient response illustrating ABS, SLS and SRL toggling, and measured steady-state waveform (bottom).

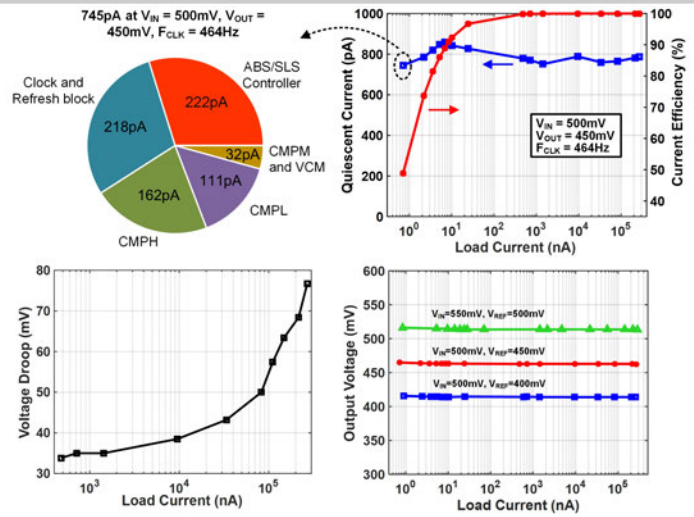


Figure 14.6.5: Measured quiescent current and current efficiency versus load current (top); measured voltage droop vs. load current and measured load regulation for different V_{REF} (bottom).

	[3] ISSCC'17	[4] ISSCC'17	[5] ISSCC'18	[6] ISSCC'18	This work
Process	65nm	65nm	65nm	40nm	**65nm**
Area [mm²]	0.0023*	0.03*	0.0014*	0.18	**0.048**
Loop control	SAR/PD/PWM	Event-Driven	Hyst. ROSC	NLSC	**Hybrid ABS/SLS**
V_{IN} [V]	0.5 - 1	0.45 - 1	0.5 - 0.9	0.6 - 1.1	**0.5 - 1**
V_{OUT} [V]	0.3 - 0.45	0.4 - 0.95	0.3 - 0.8	0.5 - 1	**0.4 - 0.95**
I_Q [μA]	14	8.1 - 258	48.4	9.2 - 19.6	**0.000745**
Dynamic load range ($I_{LOAD,MAX}/I_{LOAD,MIN}$)	100nA - 2mA (2×10^4)	14μA - 3.36mA (2.4×10^2)	10μA - 3mA (3×10^2)	2μA - 20mA (1×10^4)	**710pA - 0.27mA (3.8×10^5)**
Dynamic load range with η >90%	33.6μA - 2mA (60)	90μA - 3.36mA (37)	10μA - 1.75mA (175)	83μA - 20mA (241)	**8nA - 0.27mA (3.38×10^4)**
Peak current efficiency η [%]	99.8	99.2	99.3	99.95	**99.99**
Output voltage ripple [mV]	20 - 40	N/R	21.7	3 - 6	**2**
Steady-state error [mV]	< 5.2	< 10	< 1.55	N/R	**< 2.7**
ΔV_{OUT} @ ΔI_{LOAD}	40mV @ 1.06mA	34mV @ 1.44mA	20.5mV @ 3.25mA	40mV @ 19mA	**38mV @ 9.5μA 76.5mV @ 0.27mA**
FOM [pF]**	0.47	0.064	0.16	0.19	**0.049**

N/R = Not Reported
* Active area
** FOM = $(I_Q/I_{LOAD})\cdot(\Delta V_{OUT}/V_{OUT})\cdot C_{OUT}$

Figure 14.6.6: Comparison of the proposed hybrid DLDO with state-of-the-art DLDOs.

	[1] JSSC'14	[2] JSSC'17	This work
Process	180nm	65nm	65nm
Topology	Boost	Buck	DLDO
$P_{Quiescent}$ [pW]	544	240	373 @ V_{IN} = 0.5V
I_{OUT}	544pA - 4nA @ V_{OUT} = 1V	500pA - 1mA @ V_{OUT} = 0.8V	710pA - 0.27mA @ V_{OUT} = 0.45V
Dynamic load range	7.4	2×10^6	3.8×10^5
Settling time @ ΔI_{LOAD}	> 40s @ N/R*	> 1ms @ 0.4mA*	48µs @ 0.27mA
Output voltage ripple [mV]	N/R	15 - 40*	2
Load regulation [mV/mA]	N/A	> 60*	< 3
C_{OUT} [nF]	200	350	100

N/R = Not Reported; N/A = Not Applicable
* Observed from transient measurement

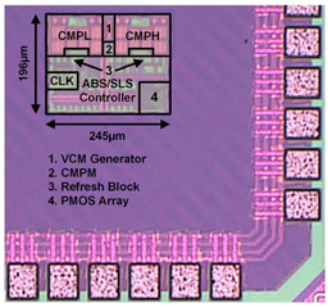


Figure 14.6.7: Comparison with state-of-the-art non-LDO, pW voltage regulators and chip micrograph of the proposed hybrid DLDO.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

14.7 A Modular Hybrid LDO with Fast Load-Transient Response and Programmable PSRR in 14nm CMOS Featuring Dynamic Clamp Tuning and Time-Constant Compensation

Xiaosen Liu, Harish K. Krishnamurthy, Taesik Na, Sheldon Weng, Khondker Z. Ahmed, Krishnan Ravichandran, James Tschanz, Vivek De

Intel, Hillsboro, OR

Complex SoCs in scaled CMOS processes integrate a large variety of digital, SRAM and noise-sensitive mixed-signal/analog circuit blocks such as PLLs, wireline/wireless/RF transceivers, sensor front-ends, etc. The on-die low-dropout regulators (LDO) used for fine-grain DVFS of digital and memory blocks must respond fast to large load transients to minimize voltage droops/overshoots without using large decoupling caps, while minimizing power overheads over a wide operating range. The noise-sensitive analog circuits, on the other hand, need LDOs that provide sufficiently high power supply rejection (PSR) and minimal output voltage ripple, while maximizing current efficiency. Fast and scalable digital (DLDO) [1], analog-assisted digital (AA-DLDO) [2-3] and hybrid (HLDO) [4] analog-digital LDOs targeted for digital and memory blocks, as well as high-PSRR analog LDOs (ALDO) [5] optimized specifically for analog circuits have been recently reported.

In this paper, we present a *universal* Modular Hybrid (MHLDO) analog-digital LDO in 14nm CMOS (Fig. 14.7.7) that can be efficiently configured to provide any desired combination of PSRR, output ripple, load transient response and operating range while minimizing losses and decoupling cap usage. The design-time configurability and robust runtime self-adjustments across different operating points offered by MHLDO via a scalable architecture alleviate the major complexities of implementing a large variety of ALDO, DLDO or AA-DLDO circuits to meet the varying requirements of the different digital, memory and mixed-signal/analog circuit blocks for SoCs in scaled CMOS.

MHLDO uses N identical PMOS DLDO and N+1 identical ALDO tiles where either can independently support the maximum load current I_L (Fig. 14.7.1). An OTA-based small dynamic range (I_L/N) ALDO leverages a dominant output pole architecture for improved bandwidth without any Miller capacitor compensation, and uses an R-C feed-forward from the input to the ALDO gate to improve PSR. An event-driven asynchronous linear digital controller, triggered by two output comparators at the low (V_L) and high (V_H) regulation thresholds, governs the combination of active ALDO/DLDO tiles at runtime to enable optimal load current sharing among them across different operating points that meet the PSR targets, while minimizing I_Q. A non-linear controller (NLC) turns all DLDO tiles ON (OFF) in response to output voltage droops (overshoots) below (above) V_L (V_H) caused by large load (unload) transients, until the output voltage goes above (below) the V_L (V_H) threshold, thus ensuring that the worst case voltage droop/overshoot targets are always met.

An MHLDO instantiation can be configured at design-time (Fig. 14.7.2) as a pure ALDO with the number of tiles selected to meet the required PSR of an analog circuit block with high I_Q, or as a pure DLDO supporting a digital or memory domain with minimal power overheads. One ALDO tile is always kept active to supply fractional current less than a single DLDO tile, thus minimizing output voltage ripple. Simulations illustrate MHLDO effectiveness in meeting PSR requirements of analog circuits in the pure ALDO mode, and also mitigating droop/overshoots due to large load transients in any mode via the global NLC loop. The optimal ALDO/DLDO-tiles combination for a MHLDO instance is dictated by the operating requirements of the load circuit block.

A key challenge for implementing proper runtime control of the optimal combination and number of ALDO/DLDO tiles that provide the right balance of PSR and I_Q across a range of operating points is posed by the slow start-up time of an ALDO tile, relative to the fast digital control loop. This speed mismatch can cause, for example, more than one ALDO tile to be turned ON by the controller when the output voltage hits the lower threshold V_L. A time-constant compensation scheme to mitigate this issue is implemented using a digital filter with programmable corner frequency. The corner frequency is set via the configurable register set_{tc}[2:0] to match the ALDO bandwidth, resulting in activation of an optimal number of ALDO tiles (Fig. 14.7.3). Limiting the DC gain of the ALDO tile amplifier also helps prevent overcurrent scenarios (per tile) for slow load current changes common in analog loads.

The high bandwidth NLC loop must also operate robustly in the presence of large parasitic package and die inductances at the input rail that can slow down the rate of change of input current through the DLDO tiles. As the NLC turns off all DLDO tiles when the output overshoots V_H, an uncontrolled oscillation can be triggered, if the output fall rate, dictated by the load current, is faster than the feedback loop bandwidth. A non-return-to-zero (NRZ) guardband reset feature is implemented to prevent such uncontrolled oscillations (Fig. 14.7.3), where a programmable minimum current is initialized after a few cycles of oscillation to slow down the output fall rate. A non-linear update to the linear shift counter, where it is reset to '0' for sudden unloading events to quickly bring the output back in regulation, is also implemented (Fig. 14.7.3).

Large load current transitions and PVT variations in digital domains can trigger undesirably large ripples due to repeated overshoots/undershoots and persistent clamping/unclamping via the DLDO tiles. Two DLDO clamp strength adjustment modes are implemented (Fig. 14.7.4): i) dynamic clamp strength (DCS) adjustment based on a configurable pattern detector, and ii) static clamp strength setting based on the steady state value of the controller. During a large unloading event, once a predefined set of clamp/unclamp signals are detected by a 5-flop pattern detector, the DCS is shifted right in Gray code. When the output voltage is within the V_L-V_H regulation range, the NLC is deactivated and retains the adjusted DCS. Since a heavy step loading event can trigger a large droop if the DCS is not at maximum, a third-level comparator at threshold V_R is used to detect the droop and restore DCS to full strength.

Measurements are performed on two different MHLDO configurations with 15 ALDO (+1 Always ON tile) and DLDO modular tiles, separately implemented in 14nm CMOS – i) high current (530mA max load, 45mA per ALDO/DLDO tile) for digital and memory blocks, and ii) high PSRR (15mA max load, 1.5mA per ALDO/DLDO tile) for analog circuits. The low current instance has an output decap of 200pF implemented using devices, while the high current instance has 4nF of output decap, 2.8nF of which was instantiated using high-density MIM cap and the remaining 1.2nF instantiated using device cap. Measurements were done with V_{IN} set to 1.1V. Measurements demonstrate that the "always ON" ALDO tile reduces output ripple and eliminates self-switching (Fig. 14.7.5), thus improving efficiency. Constant PSR is achieved for 1.5-to-9mA I_L if the current per ALDO tile is maintained at 1.5mA such that the number of ON ALDO tiles varies from 1 to 6 across the I_L range thereby saving power. For a 4.5mA I_L, increasing the number of ALDO tiles from 1 to 15 increases the PSR from -18dB to -42dB, with total I_Q increasing linearly from 6.4μA to 27μA. Power overhead of the high current MHLDO configuration scales linearly from 30μA to 52μA as the number of active tiles range from 1 to 15 for light load to full load. Voltage droop for a load transient of 21mA to 530mA with a >1A/ns slew rate for the high current instance (instantiated by controlling an on-die load) is measured to be 133mV, with a single ALDO tile ON (Fig. 14.7.5). DCS operation is demonstrated for an unloading event where the initial mod 4 and subsequent mod 8 clamp/unclamp patterns are detected, and the clamp strength is reduced to 14/30. The pattern detector is automatically disabled with consecutive clamp pairs indicating no significant output ripple.

Effects of two different time-constant compensation settings set_{tc} of 7 and 0 on output voltage responses to identical load transients show 9 and 15 ALDO tiles turning ON, respectively (Fig. 14.7.6). Thus, the turn-ON time of the ALDO tile must be carefully matched to the time constant setting to optimally balance PSR and I_Q. Comparisons with recently reported AA-DLDO, HLDO and ALDO show a better FOM, especially when a modified FOM calculation that explicitly includes the effect of di/dt slew rate is considered (Fig. 14.7.6).

Acknowledgements:
We would like to thank Tim Nolen, Changgeng Xi, and Eduardo Quijano Centeno in Intel Labs for their technical support.

References:
[1] D. Kim, et al., "A 0.5V-VIN1.44mA-Class Event-Driven Digital LDO with a Fully Integrated 100pF Output Capacitor," *ISSCC*, pp. 346-347, 2017.
[2] Y. Lu, et al., "A 500mA Analog-Assisted Digital-LDO-Based On-Chip Distributed Power Delivery Grid with Cooperative Regulation and IR-drop Reduction in 65nm CMOS," *ISSCC*, pp. 310-312, 2018.
[3] M. Huang, et al., "An Output-Capacitor-Free Analog-Assisted Digital Low-Dropout Regulator with Tri-Loop Control," *ISSCC*, pp. 342-343, 2017.
[4] S. Nasir, et al., "Switched-Mode-Control Based Hybrid LDO for Fine-Grain Power Management of Digital Load Circuits," *IEEE JSSC*, vol. 53, no. 2, pp. 569-581, 2018.
[5] Y. Lu, et al., "A 0.65ns-Response-Time 3.01ps FOM Fully-Integrated Low-Dropout Regulator with Full-Spectrum Power-Supply-Rejection for Wideband Communication Systems," *ISSCC*, pp. 306-307, 2014.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

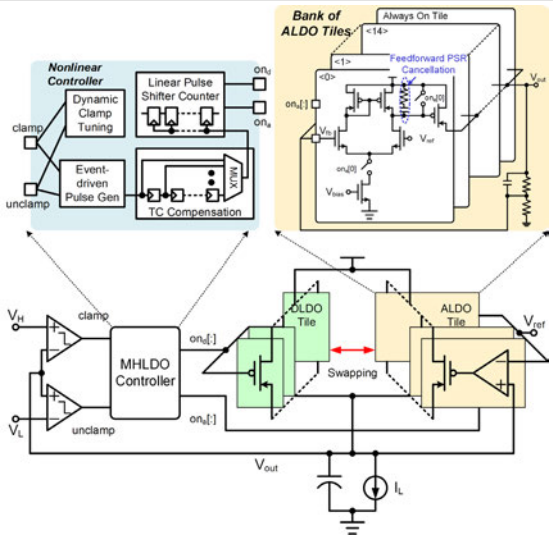


Figure 14.7.1: Architecture of the Modular Hybrid LDO with DLDO and feed-forward PSR cancellation ALDO tiles.

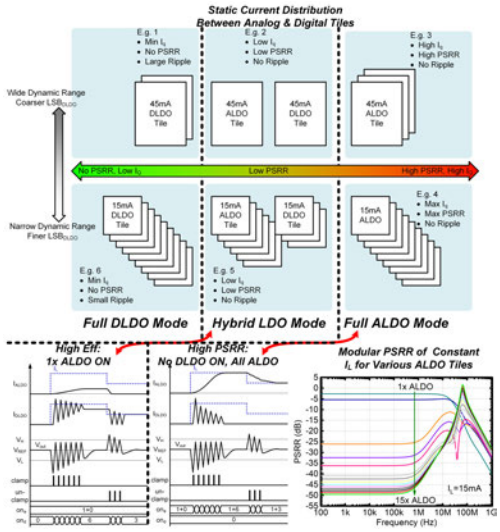


Figure 14.7.2: Self-adjusted operating status depends on modular PSRR and power consumption.

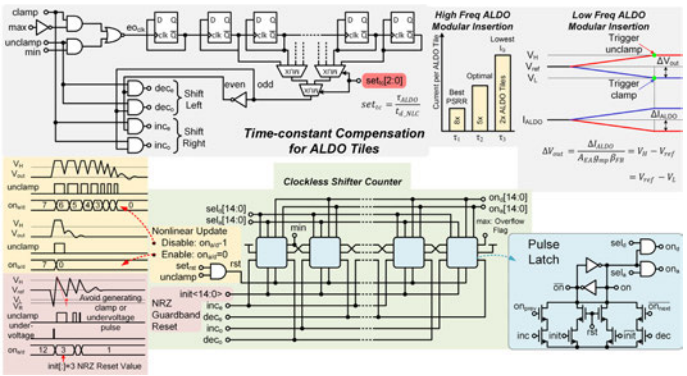


Figure 14.7.3: Time-constant compensation and high/low frequency ALDO modular insertion for proper ALDO tile number, and non-linear update functions and NRZ guardband reset for fast settling.

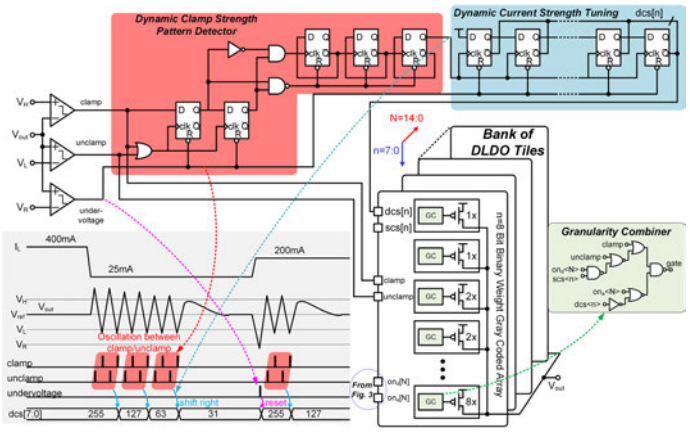


Figure 14.7.4: Dynamic clamp strength tuning triggered by the clamp/unclamp oscillation pattern detector, and the guard band reset mechanism.

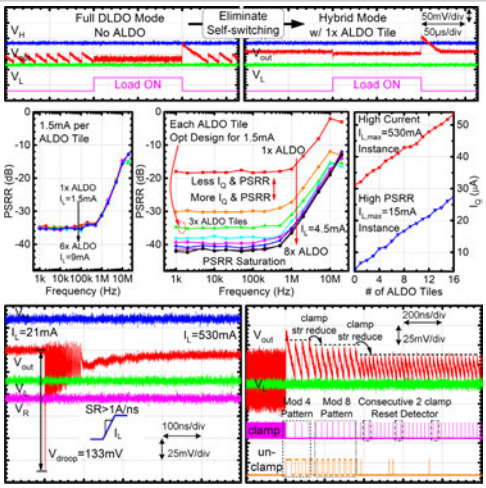
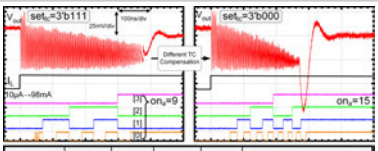


Figure 14.7.5: Load transients of full DLDO mode and hybrid mode with high efficiency, modular tradeoff between PSRR and power consumption, droop in the hybrid mode for fast SR $I_{L,max}$, and dynamic clamp tuning during unloading event in the full DLDO mode.



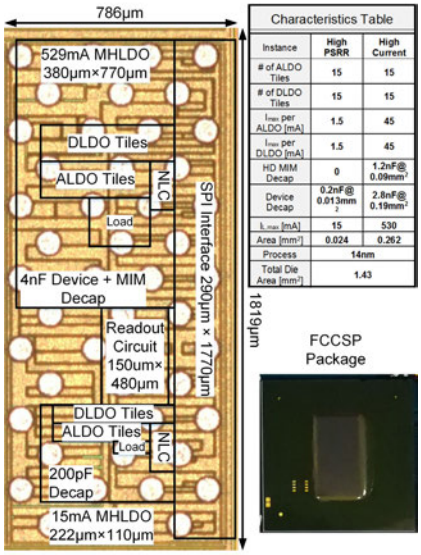
	[2]	[3]	[4]	[5]	This Work	
Process	65nm	65nm	130nm	65nm	14nm	
Architecture	AA-DLDO	AA-DLDO	HLDO	ALDO	MHLDO	
					High PSRR	High Current
$I_{L,max}$ [mA]	500	12	12	10	15	530
Area [mm²]	0.7758	0.03	0.082	0.0234	0.024	0.262
Density [mA/mm²]	644	400	146	427	625	2023
V_{IN} [V]	0.6-1.2	0.5-1	0.6, 1.1-1.2	1.2	1-1.2	
V_{OUT} [V]	0.55-1.15	0.45-0.95	0.5-0.55, 0.8-1.1	1	0.7-0.85	
PSRR [dB]	NA	NA	-9 ~ -12	-15.5	-2 ~ -42	-2 ~ -21
I_Q [μA]	500	>3.2	163.2	50	4.35 ~ 27.4	31.1 ~ 53.5
C_L [nF]	0.9	0.1	0.5	0.14	0.2	4
Droop [mV]	125@ 450mA	105@ 10mA	240@ 10.3mA	43@ 10mA	83@ 15mA	133@ 508mA
I_L Slew Rate [mA/ns]	22.5	10	42	50	>1000	
FOM* [ps]	0.28	0.34	185	3.01	0.47†	0.067†
New FOM‡ [ps]	11.4	0.5	187	3.51	0.48‡	0.083‡

* $FOM = \frac{C_L \cdot V_{droop}}{\Delta I_{L,max}} \times I_Q$, P. Hazucha et al., JSSC'05.

† New $FOM = I_Q \times \left(\frac{C_L \cdot V_{droop}}{\Delta I_{L,max}} + \frac{1}{SR}\right)$ including the slew rate effect of the load current with 1st-order model, and compare LDOs with different testing loads in a fair way.

‡ Calculated for the hybrid mode with I_Q of 1 ALDO tile ON to eliminate ripple and render PSR.

Figure 14.7.6: Waveforms depicting impact of two different time-constant compensation settings and impact on actuation of ALDO tiles, and comparison table of the MHLDO with prior-art LDOs.

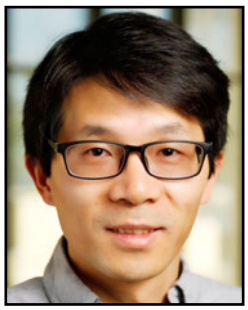


Characteristics Table		
Instance	High PSRR	High Current
# of ALDO Tiles	15	15
# of DLDO Tiles	15	15
I_{max} per ALDO [mA]	1.5	45
I_{max} per DLDO [mA]	1.5	45
HD MIM Decap	0	1.2nF@ 0.09mm²
Device Decap	0.2nF@ 0.013mm²	2.8nF@ 0.19mm²
$I_{L,max}$ [mA]	15	530
Area [mm²]	0.024	0.262
Process	14nm	
Total Die Area [mm²]	1.43	

Figure 14.7.7: Die micrograph with 14nm CMOS technology.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 15 Overview:
Power for 5G, Wireless Power, and GaN Converters
POWER MANAGEMENT SUBCOMMITTEE



Session Chair: ***Min Chen***
Analog Devices, Milpitas, CA

Associate Chair: ***Johan Janssens***
ON Semiconductor, Mechelen, Belgium

Subcommittee Chair: Yogesh Ramadass, Texas Instruments, Santa Clara

Power converters are widely used in many emerging applications like 5G communication, high-power wireless transfer, galvanically isolated systems and high-voltage automotive and offline systems. Efficiency, bandwidth, high frequency, low EMI, and high reliability are all key to these high-performance power converters. This session presents recent advances in high-bandwidth supply modulators for 5G systems, high-efficiency wireless power transfer, single-chip isolated power converters with on-chip transformers, and low-EMI high-reliability GaN power converters.

1:30 PM

15.1 An 88%-Efficiency Supply Modulator Achieving 1.08μs/V Fast Transition and 100MHz Envelope-Tracking Bandwidth for 5G New Radio RF Power Amplifier
J-S. Paek, Samsung Electronics, Hwaseong, Korea
In Paper 15.1, Samsung describes a supply modulator IC supporting 100MHz-envelope-tracking bandwidth for 5G sub-6GHz new radio transmitters. The proposed control scheme with a feed-forward fast-switching buck enables 88% of maximum efficiency with 5.8dB PAPR LTE modulated signal while achieving low output noise and output impedance.

2:00 PM

15.2 A 90ns/V Fast-Transition Symbol-Power-Tracking Buck Converter for 5G mm-Wave Phased-Array Transceiver
J-S. Paek, Samsung Electronics, Hwaseong, Korea
In Paper 15.2, Samsung presents a power management IC for a 5G mm-wave new radio (NR) phased-array transceiver. A novel symbol-power-tracking system is proposed featuring a fast transition of 90ns/V and 110ns/V for up- and down-tracking.

2:30 PM

15.3 A 100W and 91% GaN-Based Class-E Wireless-Power-Transfer Transmitter with Differential-Impedance-Matching Control for Charging Multiple Devices
S-H. Yang, National Chiao Tung University, Hsinchu, Taiwan
In Paper 15.3, National Chiao Tung University describes a 13.56MHz GaN-based differential Class-E wireless power-transfer system delivering up to 100W of output power. A differential impedance-matching technique and reverse-conduction calibration technique provides a peak efficiency of 91%.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

3:15 PM

15.4 A 52% Peak-Efficiency >1W Isolated Power Transfer System Using Fully Integrated Magnetic-Core Transformer

Z. Yue, Analog Devices, Beijing, China

In Paper 15.4, Analog Devices presents an isolated power-transfer system using fully integrated transformer with magnetic core. A fully symmetric power stage is utilized to achieve a peak efficiency of 52%, output power of 1.1W and passes CISPR 22 EMI limit by 5.8dB.

3:45 PM

15.5 An 800mW Fully Integrated Galvanic Isolated Power Transfer System Meeting CISPR 22 Class-B Emission Levels with 6dB Margin

W. Qin, Analog Devices, Shanghai, China

In Paper 15.5, Analog Devices presents an 800mW isolated power-transfer system using an on-chip coreless transformer. Various circuit techniques reduce emissions to meet EMI standard CISPR 22 Class B with a 6dB margin.

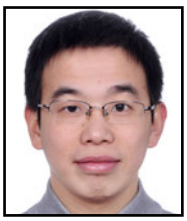
4:15 PM

15.6 A 10MHz *i*-Collapse Failure Self-Prognostic GaN Power Converter with T_J-Independent In-Situ Condition Monitoring and Proactive Temperature Frequency Scaling

Y. Chen, University of Texas at Dallas, Richardson, TX

In Paper 15.6, the University of Texas, Dallas, presents a failure self-prognosis GaN power converter achieving in-situ aging condition monitoring using switch R_{ON} as precursor. To eliminate the effect of junction temperature, a gate-leakage-based, aging-independent junction temperature sensor is developed.

4:45 PM



15.7 An 8.3MHz GaN Power Converter Using Markov Continuous RSSM for 35dBμV Conducted EMI Attenuation and One-Cycle T_{ON} Rebalancing for 27.6dB V_O Jittering Suppression

Y. Chen, University of Texas at Dallas, Richardson, TX

In Paper 15.7, the University of Texas, Dallas, presents an 8.3MHz GaN power converter implementing Markov continuous spread-spectrum modulation, achieving the lowest EMI noise floor of 27dBμV. One-cycle T_{ON} rebalancing suppresses output voltage jittering by 27.6dB.

5:00 PM

15.8 A 4.5V/ns Active Slew-Rate-Controlling Gate Driver with Robust Discrete-Time Feedback Technique for 600V Superjunction MOSFETs

S. Kawai, Toshiba, Kawasaki, Japan

In Paper 15.8, Toshiba presents an active slew-rate-controlling gate driver with discrete-time background feedback. In combination with an initial boost pulse, the technique enables dV/dt maximization for robust minimization of switching loss and turn-on delay.

15.1 An 88%-Efficiency Supply Modulator Achieving 1.08μs/V Fast Transition and 100MHz Envelope-Tracking Bandwidth for 5G New Radio RF Power Amplifier

Ji-Seon Paek, Dongsu Kim, Jun-Suk Bang, Jongbeom Baek, Jeonghyun Choi, Takahiro Nomiyama, Jaeyeol Han, Younghwan Choo, Yongsik Youn, Euiyoung Park, Sungjun Lee, Ik-Hwan Kim, Jongwoo Lee, Thomas Byunghak Cho, Inyup Kang

Samsung Electronics, Hwaseong, Korea

In the recent 3GPP standard, 5G New Radio (NR) in the sub-6GHz frequency band supports up to 100MHz bandwidth (BW) for uplink channel and requires fast on/off transition below 10μs to support short transition-time interval (TTI). Thus, it forces the supply modulator (SM) to track the NR-100MHz envelope signal for power saving of an RF power amplifier (RF-PA) and requires a fast dynamic voltage-scaling (DVS) transition. For the 3G/LTE uplink, the SM has to satisfy low receiver-band noise (RxBN) for LTE-FDD bands [1] and a boosted output for Power Class 2 high-power user equipment (HPUE) in TD-LTE [2,3]. In order to support the 100MHz envelope tracking (ET) BW, the SM must have a 3dB bandwidth of at least 125MHz while keeping sufficient phase margin in the presence of few-hundred-pF load capacitance. Furthermore, a typical linear-assisted SM topology has maximum efficiency limitation due to the AB bias current, sinking- and sourcing-current of the linear amplifier. In this work, a voltage-buffer-compensated (VBC) Class-AB linear amplifier (LA) and a feed-forward fast-switching (FFFS) buck converter are proposed to achieve the 100MHz ET BW and high system-power efficiency.

Figure 15.1.1 shows the proposed SM architecture consisting of a buck-boost (BB) converter with an auxiliary LDO, a dual-supply slow-switching buck converter with the return-to-battery switching scheme [3], and an AC amplifier. In a 2G APT mode, the BB supports only a step-down dynamic-voltage-scaling (DVS) voltages and the auxiliary LDO enables the 35dBm GSM PA output with a 3A maximum load current in typical 3.0-to-5.0V battery range. In 3G/LTE APT mode, the BB generates both step-down and step–up voltages and supplies up to 1.0A of load current even in the worst case of a 3.0V battery voltage with the boosted 5V output for HPUE PA. To meet the transition time of a short 5G NR TTI without increasing the output ripple voltage and maximum inductor current, both an adaptive load capacitor and a fast charging/discharging LDO are employed in the BB design. In the ET mode, the BB operates as a power source for the LA and the dual-supply buck. The BB output voltage is controlled by the output power level and is reduced by the offset voltage across the AC-coupling capacitor (V_{off_AC}) for better efficiency [1]. The dual-supply buck and the AC amplifier operate jointly to provide the ET supply voltage to the RF-PA. The AC amplifier supports three operating modes according to the signal-BW and radio-access-technology (RAT) change: 1.) the FFFS buck only for low BW; 2.) the FFFS buck assisted by the LA for medium BW; and 3.) the VBC Class-AB LA only for large BW.

The minimum bandwidth requirement of the VBC Class-AB amplifier in Fig. 15.1.2 is 125MHz for the NR-100MHz envelope tracking. However, the second pole (p_2), which is generated by an internal decoupling capacitor (C_{PA}) and an equivalent load resistor (R_L) of the RF-PA, is typically located below 100MHz. In order to compensate p_2, two compensation schemes are employed. First, the voltage buffer between the 1st OTA and the 2nd Class-AB output stage pushes the dominant pole, which is generated by the gate parasitic capacitance of the Class-AB output stage, to frequency higher than the GBW. Second, the p_2 is cancelled out by a zero ($1/R_zC_z$) generated by a passive-zero network in front of the voltage buffer. By employing the two schemes, the open-loop GBW ($(G_{m1}\cdot G_{m2}\cdot R_L)/C_z$) can be wider than p_2 as shown in the right-bottom plot of Fig. 15.1.2. In addition, a constant-G_m bias circuit minimizes the GBW variation due to PVT variations to enhance phase margin, and a capacitor in the feedback path (C_f) freezes the 3dB signal BW, although G_{m2} varies by Class-AB push-pull operation. Thus, the SM-IC achieves a 3dB signal BW above 125MHz. Although the voltage-buffer consumes additional bias current, the SM-IC reaches 82% peak efficiency at 1.2W of output power while tracking the NR-100M envelope signal.

A hybrid SM of the fast-switching buck and the slow-switching buck is one of strong candidates, which can surpass the maximum efficiency of the linear-assisted hybrid SM [4]. However, achieving wider bandwidth and low output noise requires fast switching frequency resulting in large switching loss [5]. Moreover, the typical linear-assisted hybrid SM has challenges to increase efficiency due to the AB bias, sinking and sourcing currents. Figure 15.1.3 compares the generated current waveforms between the typical linear-assisted structure and the proposed FFFS buck with the LA. In the proposed structure, most of output current is provided by the FFFS buck and the slow buck ($I_{SBK}+I_{FBK}$), and the remaining very-high-frequency component ($I_{PA}-I_{SBK}-I_{FBK}=I_{LA}$) is provided by the LA. To realize the proposed current waveforms, a feed-forward envelope signal is applied to the FFFS buck, and the LA receives the input signal delayed by the group delay difference between the FFFS buck and the LA. This delay can be adjusted by controlling the group delay of the lowpass filter in front of the LA. Due to the feed-forward input to the FFFS buck, the feedback loop of the FFFS buck can respond to the SM output variation immediately and the feedback loop of the LA starts to operate only at very high frequencies. Thus, the FFFS buck with the LA can achieve both low output impedance and low output noise while minimizing the power loss. However, if the two different loops operate independently, I_{FBK} and I_{LA} can run in the opposite direction while satisfying $I_{FBK}+I_{LA}=I_{PA}-I_{SBK}$. To prevent this scenario, another feedback signal (FB3) is applied to the FFFS buck input from the linear output, so that the two different loops can operate jointly and be stable. At low BW (<LTE 6RB), the FFFS buck provides all AC components solely, so that the SM-IC reaches maximum 88% efficiency by eliminating the LA bias current and maintains high efficiency at 82.5% until 300mW of back-off power region. The implemented 4th-order output filter with two small-size inductors and a fixed switching-frequency controller (SFC), which has a simple frequency-locked loop, guarantee the noise and efficiency performance for various LTE FDD bands in the presence of PVT variations.

Figure 15.1.4 shows the BB design and its measured output-transition waveforms. In a wide battery range, the BB provides 5V of the maximum output and a 1A of the average load current in a steady state. For a 1.08μs/V dynamic voltage-scaled fast transition, load capacitor is controlled adaptively in boost region and the assist-LDO is implemented for fast charging/discharging in a buck region. Figure 15.1.5 shows the SM efficiency with a fixed 3.2Ω-resistor load according to its output power for various input-signal BWs by using 5.8dB PAPR LTE and 6dB PAPR DFT-s-OFDM NR modulated signals. The SM achieves 88% of maximum efficiency over 0.8W of output power, and meets the RxBN requirement over the wide range of output power levels and all operating FDD bands. LTE ET performance is measured with a commercial low-band power amplifier module integrated duplexer (PAMiD) and NR ET performance is measured with an n78 PAMiD. Compared to the APT mode, the NR-100M ET-PA saves 400mW at 23dBm PAMiD output power. By applying memory digital pre-distortion (DPD), which compensates 4dB in-band gain variation in the 100MHz of channel frequency, E-UTRA ACLR of NR-100M ET PA output is measured as -39.6/-38.1dBc at 23dBm PAMiD output power. The overall performance summary and the comparison with prior works are shown in Fig. 15.1.6. The die size in Fig. 15.1.7 is 5.145mm^2 when implemented in a 90nm CMOS process and a 42-pin wafer-level chip-scale package.

References:
[1] J.-S. Paek et al., "An RF-PA Supply Modulator Achieving 83% Efficiency and -136dBm/Hz Noise for LTE-40MHz and GSM 35dBm Applications," *ISSCC*, pp. 354–355, Feb. 2016.
[2] C.-Y. Ho et al., "An 87.1% Efficiency RF-PA Envelope-Tracking Modulator for 80MHz LTE-Advanced Transmitter and 31dBm PA Output Power for HPUE in 0.153μm CMOS," *ISSCC*, pp. 432–433, Feb. 2018.
[3] T. Nomiyama et al., "A 2TX Supply Modulator for Envelope-Tracking Power Amplifier Supporting Intra- and Inter-Band Uplink Carrier Aggregation and Power Class-2 High-Power User Equipment" *ISSCC,* pp. 434-435, Feb. 2018.
[4] P. Arno et al, "Envelope Modulator for Multimode Transmitters with AC-Coupled Multilevel Regulators" *ISSCC*, pp. 296–297, Feb. 2014.
[5] V. Pinon et al., "A Single-Chip WCDMA Envelope Reconstruction LDMOS PA with 130MHz Switched-Mode Power Supply" *ISSCC*, pp. 564–565, Feb. 2008.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

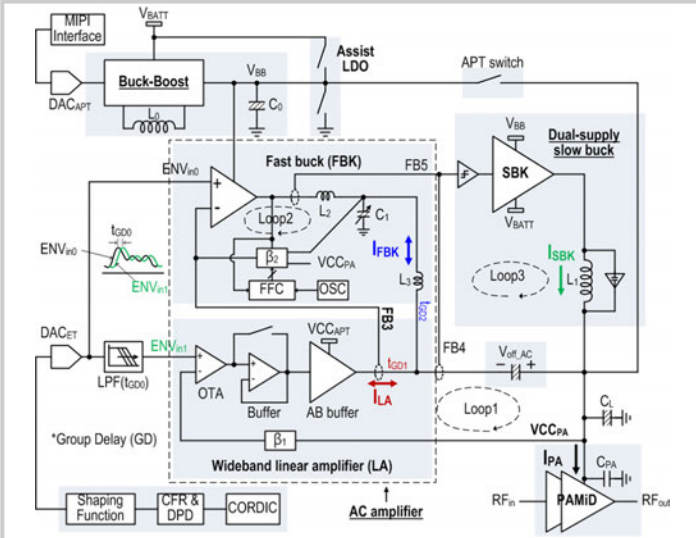


Figure 15.1.1: Proposed supply-modulator architecture supporting NR-100M ET and HPUE.

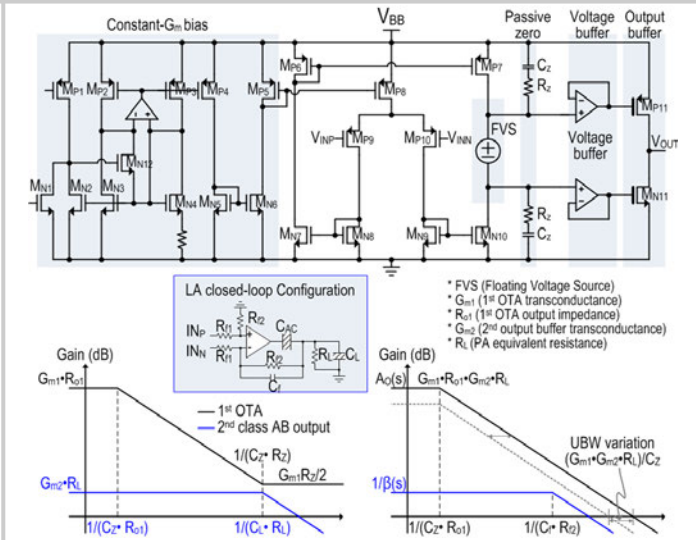


Figure 15.1.2: Wideband voltage-buffer-compensated linear amplifier with passive zero compensation.

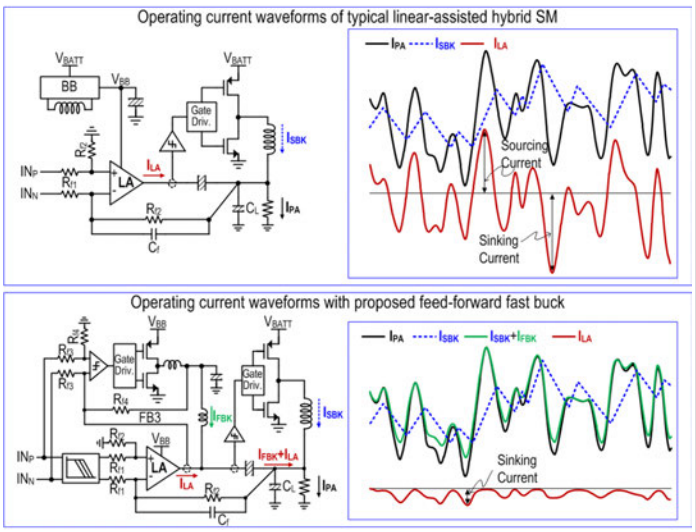


Figure 15.1.3: Presented hybrid SM with feed-forward fast switching buck converter.

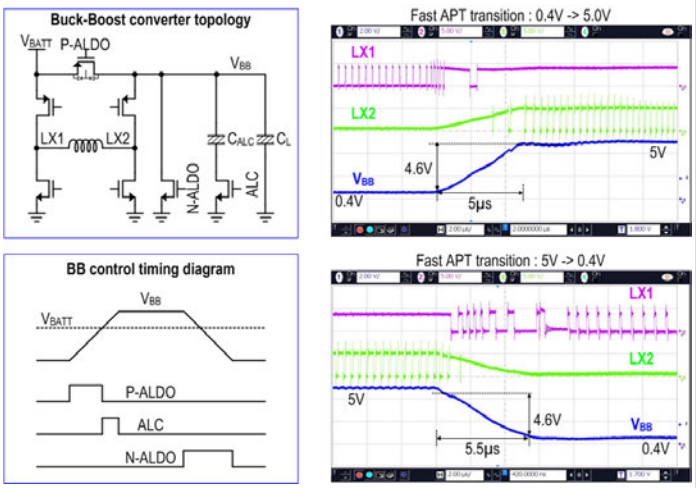


Figure 15.1.4: Measured APT transition waveforms of designed buck-boost converter.

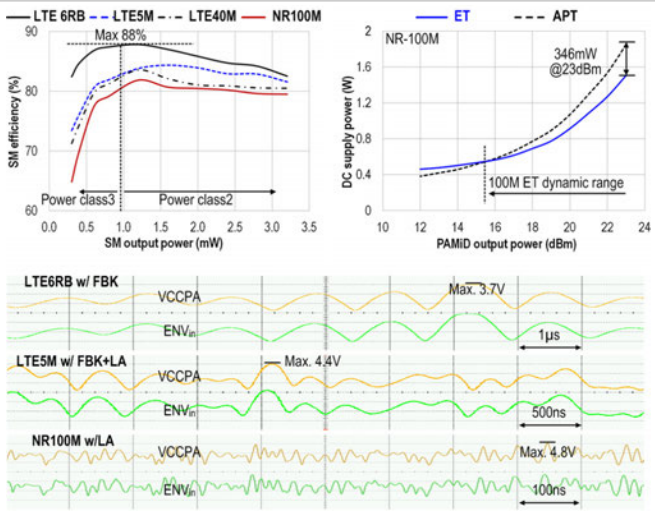
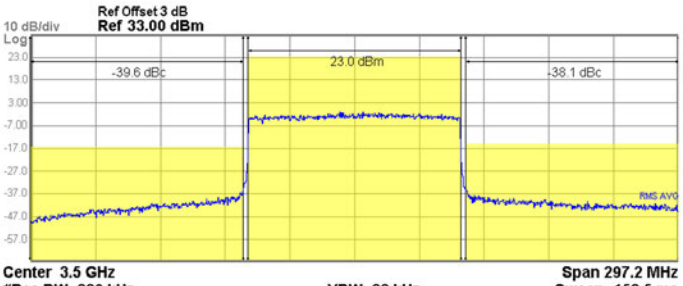


Figure 15.1.5: SM efficiency at ET mode and the DC power consumption comparison with n78 PAMiD between an APT-PA and ET-PA. SM output waveforms are measured with various BW signals.



Ref.	Supporting Protocol	Operation mode	RF-PA Power Class	Max. ET BW (MHz)	Max. SM eff. (%)		SM noise @30/95MHz (dBm/Hz)	APT Transition Time (μs)		RF-PA Band	DC power saving ET-PAMiD vs APT-PAMiD (mW)			Die area (mm²)	Tech. (nm)
											E-UTRA ACLR (dBc)				
This work	NR, LTE, WCDMA, GSM	1TX, ET, APT	Class2	100	88	@LTE6RB, 800mW	-132/-141	5	0.4-5.0 V	B41	771	-	@LTE6RB, 26dBm	5.0	90
					83	@LTE10M, 800mW					1395	-39	@LTE25RB, 26dBm		
					77	@NR100M, 800mW				B42	720	-38	@NR100M, 23dBm		
[1]	GSM, LTE, WCDMA	1TX, ET, APT	Class3	40	83	@LTE10M, 800mW	-136/-141	20	0.4-4.0 V	B1	470	-41.2	@LTE10M, 24dBm	4.0	130
[2]	GSM, LTE, WCDMA	1TX, ET, APT	Class2	80	82	@LTE20M, 1.5W	-143/-	18	0.5-4.2 V	B41	1710	-	@LTE20M, 26dBm	5.133	153
					78	@LTE80M, 1.5W					-	-38	@LTE80M, 23dBm		
[3]	GSM, LTE, WCDMA	2TX, ET, APT	Class2	40	82	@LTE40M, 800mW	-133/-142	20	0.4-5.0 V	B41	800	-38	@LTE40M, 26dBm	6.25*	90
[4]	LTE	1TX, ET	NA	10	86	@LTE10M, 1.8W	NA	NA		B3	NA	-40	@LTE10M, 26dBm**	6.35	130

* The reference is a single-chip SM-IC supporting two independently controlled TX outputs

Figure 15.1.6: Measured 23dBm n78 PAMiD output spectrum for NR100MHz and the comparison to prior art.

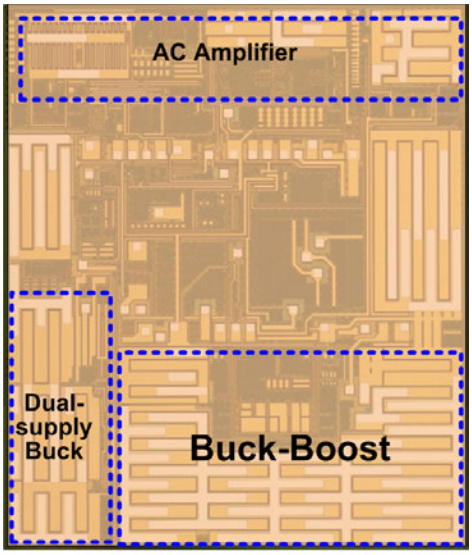


Figure 15.1.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

15.2 A 90ns/V Fast-Transition Symbol-Power-Tracking Buck Converter for 5G mm-Wave Phased-Array Transceiver

Ji-Seon Paek, Takahiro Nomiyama, Jaeyeol Han, Ik-Hwan Kim, Yumi Lee, Dongsu Kim, Euiyoung Park, Sungjun Lee, Jongwoo Lee, Thomas Byunghak Cho, Inyup Kang

Samsung Electronics, Hwaseong, Korea

The 5G-New-Radio (NR) standard in millimeter wave (mm-wave) bands requires a low-cost antenna module consisting of a phased-array transceiver with beamforming [1], an antenna array, and a power management IC (PMIC). Since a typical on-chip mm-wave CMOS power-amplifier (PA) arrays have poor power efficiency due to the lossy substrate and the low maximum available-gain frequency of CMOS processes, they cause serious thermal issues due to the high power density and reduce battery life in a mobile handset. Recently, supply-modulation (SM) techniques, such as an envelope tracking (ET) and an average-power tracking (APT) instead of a direct battery-connected supply, have been introduced to enhance PA efficiency [2]. However, these techniques have few challenges, such as limited ET tracking bandwidth and APT transition time, to support new requirements of the 5G NR standard. In order to manage the power effectively of a phased-array module, a special SM technique named symbol-power tracking (SPT) is proposed for the power management in this work. The SPT controls the supply voltage of mm-wave PA arrays every few micro-second (symbol to symbol), which is much faster than that of APT, which adjusts the output voltage every one millisecond (sub-frame to sub-frame).

Compared to a 4G LTE system, a 5G-NR system has several differences, such as the maximum channel bandwidth, various types of subcarrier spacings (SCS), symbol- and cyclic-prefix (CP) lengths, multiple slot formats, an on/off time mask, and a transmit-power update time. The 5G-NR-uplink (UL) channel bandwidth is increased up to 400MHz for the high throughput, and the SCS increases from 15 to 240kHz for low latency. Thus, the symbol and CP lengths are shortened to 4.16 and 290ns, respectively, at 240kHz of SCS. Main differences of the NR slot format are that the UL and downlink (DL) channel assignment and the transmit (TX) power level change at a symbol level. In order to achieve the best channel capacity and power efficiency in a 5G-NR system, a mm-wave phased-array module needs to support an accurate TX gain control and contain its own power-saving component. The SPT PA simplifies the TX-gain calibration and allows superior PA efficiency compared to a typical APT system. In order to generate an appropriate supply voltage (V_{SPT}) for every 4.16μs symbol interval, a DC-DC converter with a short transition time within 290ns is implemented.

Figure 15.2.1 shows the proposed mm-wave PMIC architecture with pseudo-single-inductor dual-output (SIDO) buck converter supporting the desired SPT. It consists of two buck (BK) converters, a 1.1V low-drop-out (LDO) linear regulator, an auxiliary LDO, a fast sourcing/sinking charger, four SPT switches ($SW_{L1,2}$, $SW_{SD1,2}$), and a SIDO switch (SW_{SIDO}). A microcontroller unit (MCU) in the mm-wave transceiver generates timing signal for every CP start time to synchronize the transmit-power update timing and SPT transition timing precisely. The two BKs (SPT and SIDO) generate the dynamic scalable supply voltage from 0.4 to 3V within 4μs with a 4.7μF load capacitor. For the symbol-power tracking, however, the required PA supply (V_{SPT}) transition time is only 290ns of 1CP duration, which is much faster than the 4μs transition time of each buck. In order to charge the 4.7μF capacitor to a certain level within one CP duration, a typical buck converter should have current drivability of over 10A, which also requires a large inductor and power transistors. To meet this requirement, two control schemes are implemented: 1). Capacitor swapping [3] between two load capacitors (C_{L1} and C_{L2}) and 2.) Fast charging/discharging for high-frequency load capacitor (C_{SPT}=220nF). Because other sub-blocks in the mm-wave transceiver consume large supply current from 1.1V, the SIDO buck generates the 1.3V supply voltage for efficient sub-regulation of the 1.1V LDO.

Figures 15.2.2 and 15.2.3 show the basic operation of the SPT with SIDO. To enable the symbol-power tracking, the SPT buck receives two data from a 5G modem. One is the power level of next symbol and the other is the CP start time. Based on the power level, the SIDO buck pre-charges the auxiliary capacitor (C_{L1} or C_{L2}) within the present symbol duration of 4.16μs. Since the 1.3V output cannot be regulated by the SIDO buck while the SIDO buck is operating for the SPT pre-charging, the auxiliary-LDO starts to regulate the 1.3V in a feedback loop seamlessly and provides the required output current. After finishing the SPT pre-charging, the SIDO buck returns to regulate the 1.3V output. After the SPT buck receives the CP start-time signal, C_{L1} (or C_{L2}) are disconnected from the V_{SPT} output, and the fast charger starts to charge or discharge C_{SPT}. When the voltage difference between C_{SPT} and C_{L1} (or C_{L2}) is within a threshold, the fast-charge detector generates a swap trigger signal (SWAP_EN), and then C_{L1} (or C_{L2}) is connected back to V_{SPT} and the transition ends within 290ns.

Figure 15.2.3 shows the detailed design of the two buck converters with a ripple-injected hysteresis control. To perform the SPT or SIDO operation, the two bucks have to regulate their target output with the loop stability and smooth transition even in the presence of the abrupt feedback-input change. Because a typical response time of the buck feedback loop is very slow to compensate the output LC double pole, the fast-changing SPT output or abruptly-changing feedback point of the SIDO buck, which is one of V_{C1}, V_{C2}, V_{13}, causes the loop stability issue resulting in a large overshoot or undershoot at V_{SPT} and V_{13}. In order to shorten the loop response time, a dynamic pre-charge (DPC) control scheme is employed at the hysteretic feedback loop, which allows direct feed-forward path from output to internal compensation node during only the SPT or SIDO transition time. By using the DPC, the loop response time is reduced from a few-μs to 250ns. According to the SPT and SIDO operating scenarios, the reference-voltage generator, which consists of two 7b-DACs, two buffers, and a track-and-hold circuit, makes proper reference voltages for each buck regulation. The hysteretic control causes the output switching frequency variation based on the input and output ratio, and this frequency variation results in unwanted spurious spectrum by modulating the PA transmit signal. With the selected LC filtering, to guarantee noise attenuation over PVT variations, the fixed switching frequency controller (SFC) is employed at the hysteretic controller.

With the wide 3-to-5V battery range, the two bucks provide dynamic voltage-scaled outputs form 0.4 to 3V within 4μs while providing a 1.5A maximum current. The measured symbol-power-tracking waveform with the SIDO is shown in Fig. 15.2.4. The V_{SPT} transits from 0.4 to 3.0V within 290ns. The up- and down-tracking speeds are 90ns/V and 110ns/V, respectively. Due to the auxiliary LDO, while the SIDO buck is pre-charging the auxiliary capacitor to the next symbol power, the 1.1V LDO output keeps constant without any interruption form the SIDO operation. The feedback loop of two buck converters achieves 250ns settling time by the DPC with less than 80mV of undershoot and overshoot voltages. Figure 15.2.5 shows the mm-wave-PA efficiency improvement by the SPT supply modulation as function of the PA supply voltage and compared to the efficiency for the PA voltage of 3V. The measured peak efficiency is 95.4%. Figure 15.2.6 compares the performance of the proposed converter with that of the prior-art works [2,4-6]. The die size in Fig. 15.2.7 is 4.07mm^2 when implemented a 90nm CMOS process and 30-pin wafer-level chip-scale package.

References:
[1] J. Dunworth et al., "A 28GHz Bulk-CMOS Dual-Polarization Phased-Array Transceiver with 24 Channels for 5G User and Basestation Equipment," *ISSCC*, pp. 70-71, Feb. 2018.
[2] J.-S. Paek et al., "An RF-PA Supply Modulator Achieving 83% Efficiency and -136dBm/Hz Noise for LTE-40MHz and GSM 35dBm Applications," *ISSCC*, pp. 354-355, Feb. 2016.
[3] T. Nomiyama et al., "A 2TX Supply Modulator for Envelope-Tracking Power Amplifier Supporting Intra- and Inter-Band Uplink Carrier Aggregation and Power Class-2 High-Power User Equipment," *ISSCC*, pp. 434-435, Feb. 2018.
[4] L. Cheng et al., "A 10/30MHz Wide-Duty-Cycle-Range Buck Converter with DDA-Based Type-III Compensator and Fast Reference-Tracking Responses for DVS Applications," *ISSCC*, pp. 84-85, Feb. 2014.
[5] S. Kudva and R. Harjani, "Fully-Integrated On-chip DC-DC Converter With a 450X Output Range," *IEEE JSSC*, vol. 46, no. 8, pp. 1940-1951, Aug. 2011.
[6] Y.-H. Lee et al., "Power-Tracking Embedded Buck-Boost Converter With Fast Dynamic Voltage Scaling for the SoC System," *IEEE Trans. Power Electronics*, vol. 27, no. 3, pp. 1271-1282, Mar. 2012.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

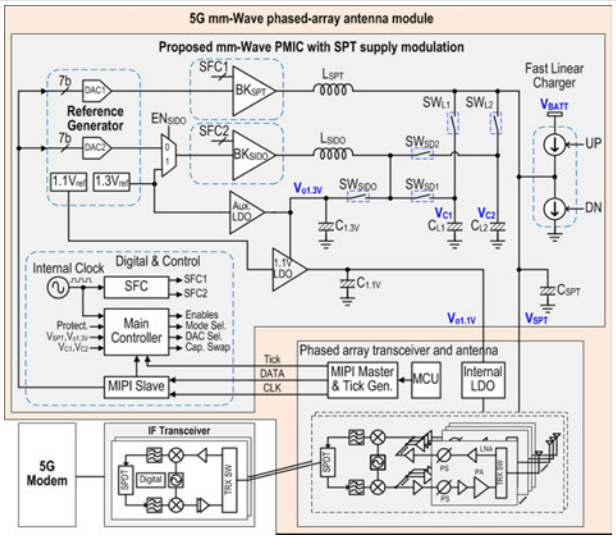


Figure 15.2.1: Proposed PMIC architecture enabling SPT supply modulation for mm-wave PA arrays.

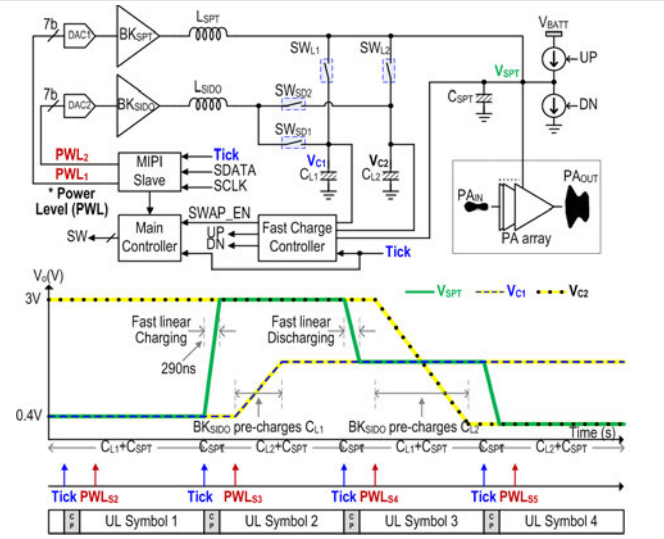


Figure 15.2.2: SPT-transition control diagram with the capacitor-swapping technique.

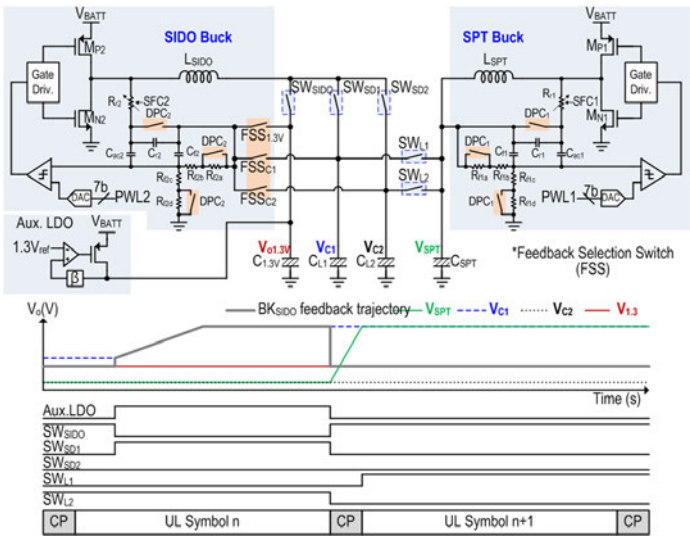


Figure 15.2.3: SIDO control diagram with the dynamic pre-charging technique.

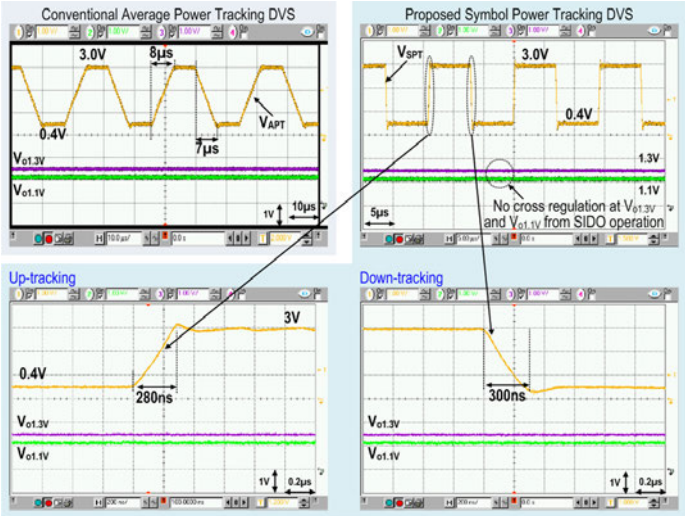


Figure 15.2.4: Measured waveforms showing the SPT transition compared to APT.

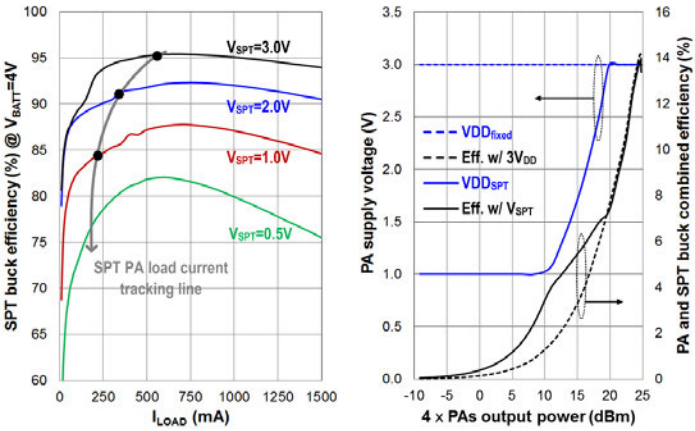


Figure 15.2.5: Stand-alone buck-converter efficiency versus load current and the power efficiency of mm-wave PA arrays versus PA output power level.

Ref.	[2]	[4]	[5]	[6]	This work
Supporting Protocol	WCDMA GSM, LTE	-	-	-	5G NR mmWave
Operating Mode	ET, APT	DVS	DVS	DVS	SPT
Topology	Hybrid SM	Buck	Buck	Buck-Boost	Buck
Control Scheme	Hysteresis & Linear-assist	Voltage Mode	Voltage Mode	Current Mode	Hysteresis & Cap. Swap
Maximum Output Power (W)	6	3.6	0.27	1.4	4.5
Inductor(μH)	4.7	0.33	2nH	1	0.47
Load Capacitor (μF)	4.7	1	5nF	0.88	4.7
Input Range (V)	3.0-5.0	3.3	1.2	2.5-4.5	3.0-5.0
DVS Range (V)	0.4-3.7	0.45-2.4	0.3-0.88	2.0-4.0	0.4-3.7
Peak Efficiency (%)	95	86.6	74.5	91	95.4
Up-tracking (μs/V)	3	0.67	2.6	20	0.09
Down-tracking (μs/V)	3	1.56	3.6	15	0.11
Technology (nm)	130	130	130	250	90

Figure 15.2.6: Performance summary and comparison with prior art.

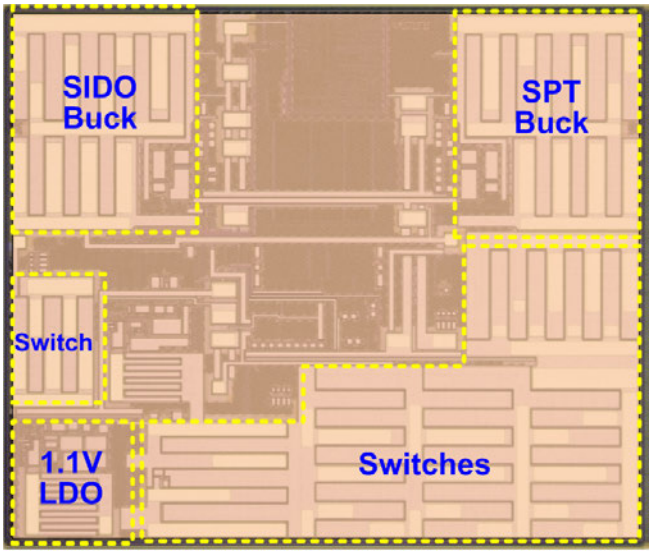


Figure 15.2.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

15.3 A 100W and 91% GaN-Based Class-E Wireless-Power-Transfer Transmitter with Differential-Impedance-Matching Control for Charging Multiple Devices

Cheng-Yu Xie[1], Shang-Hsien Yang[1], Shen-Fu Lu[1], Fa-Yi Lin[1], Yen-An Lin[1], You-Zheng Ou-Yang[1], Ke-Horng Chen[1], Kuo-Chi Liu[1], Yin-Hsi Lin[2]

[1]National Chiao Tung University, Hsinchu, Taiwan
[2]Realtek Semiconductor, Hsinchu, Taiwan

Having multiple devices charged by a wireless-power-transfer (WPT) system has become more common as illustrated in Fig. 15.3.1. A wide-power-range (no load ~100W), compact, and efficient WPT system needs to include the following features. First, an impedance-matching technique that achieves zero-voltage switching (ZVS) and zero voltage-derivative switching (ZVDS) on a Gallium Nitride (GaN) switch is needed to reduce the efficiency loss caused by hard switching (HS) and reverse conduction (RC). Second, under high-power conditions, it is desirable to reduce the voltage and current stress on each switch and passive components. Third, there is a need to reduce the number of external components for compact size. In [1] and [2], an external capacitor array for impedance matching occupies a large printed circuit board (PCB) area. If the output power is as high as 100W, the controlled switch needs to withstand high voltage stresses of up to 600V. The fractional-capacitance tuning technique in [3] achieves a wide range of equivalent capacitance, but there are still high-voltage-stress problems similar to [1,2]. Although [4] provides a high-power solution, the voltage-controlled-capacitance (VCC) technique needs to tune the internal parasitic capacitance, C_{OSS}, from 350 to 3500pF and thus requires more external components and large bias voltage V_{BIAS}.

In Fig. 15.3.2, this paper proposes a 13.56MHz GaN-based differential Class-E WPT system that uses a differential-impedance-matching (DIM) technique and a reverse-conduction-calibration (RCC) technique. The DIM technique ensures that both ZVS and ZVDS reduce HS and RC power losses, while the RCC technique compensates for inherent loop delays, further reducing the RC power losses. Owing to the differential topology, higher output-power transfer and lower current stress per component are achieved compared to a single-phase topology. Moreover, to further extend the impedance-matching tuning range, the proposed bilateral bias-capacitance (BBC) technique effectively reduces the voltage stress on each component. Consequently, the proposed differential Class-E WPT system can be more efficient at providing energy to several mobile phones, where a fixed reverse-L matching network comprising of inductors $L_{FC1,2}$ and capacitors $C_{FC1,2}$ is used. In the case of any impedance change at the multiple RXs, the DIM automatically adjusts the bias voltage (V_{BIAS}) in the BBC technique based on the single-ended switching node voltage V_{SW1} to ensure both ZVS and ZVDS for impedance matching. The sensed differential switching points $V_{SW1,2}$ are scaled down to $V_{SWD1,2}$ to the DIM controller input for the fine-control loop while V_S, which is generated by comparing V_{SWD1} with the reference V_{R1}, is also sent to the DIM controller for the coarse-control loop. The coarse and fine loops can provide fast and precise impedance matching, respectively. The differential BBC technique (bottom of Fig. 15.3.2) can reduce the V_{BIAS} voltage to half that of the single-ended BBC technique for the same MOSFET capacitance C_{OSS}. The large resistors $R_{DC1,2,3}$ reduce the large AC current flowing back to the bias generator, and the large capacitors $C_{BK1,2}$ block the DC component of the power path. Consequently, the $C_{COSS1,2}$ are controlled by a lower V_{BIAS} and thus high-voltage-stress problem in [4] can be alleviated. Due to the insertion of the equivalent $C_{COSS1,2}$, the impedance matching can be adjusted by monitoring V_{SW1}. The coarse loop decreases the $C_{COSS1,2}$ if V_{SW1} is higher than zero when the GaN switch turns on, and the HS power loss can be reduced. On the other hand, the $C_{COSS1,2}$ are increased if V_{SW1} is lower than zero when the GaN switch turns on, and the RC power loss can be reduced. At high power transmissions, any slight loop delay can cause severe RC power loss on the GaN power switch, ultimately reducing efficiency. The proposed RCC technique compensates for the loop delay by analyzing the delay-included information on V_S for achieving accurate impedance matching.

In the phase detector of the DIM controller (top of Fig. 15.3.3), signals V_{LD} and V_{LG} are used as a timing reference to lock V_S and generate the coarse-loop control signals V_{UC} and V_{DC}. In addition, a charge-area-calculation (CAC) circuit consisting of a pair of discrete-time integrators calculates the charge regions of the differential switching nodes V_{SW1} and V_{SW2} and produces corresponding signals V_{A1} and V_{A2}. The fine loop tracks the optimal charge region under a hysteresis control, which compares V_A with upper and lower references V_{R2} and V_{R3} to produce corresponding fine-loop control signals V_{UF} and V_{DF} (in the right middle of Fig. 15.3.3). Besides, the charge-area-lockout (CALO) circuit prevents the system from malfunctioning due to variations in component values. If the difference between V_{A1} and V_{A2} is greater than the reference V_{R4}, the controller automatically shuts down for safety. Besides, a matching bias generator (MBG) similar to a high-voltage low-dropout regulator produces a bias voltage V_{BIAS} to the BBC circuit. The coarse loop is enabled when the system is in startup or transient. The multiplexer selects V_{UC} and V_{DC} as control signals and applies 4× unit-size current in the MBG to accelerate the transient response. The fine loop is enabled next. Then, the multiplexer selects V_{UF} and V_{DF} as control signals and applies 1× unit-size current in the MBG. Finally, the MBG generates the appropriate V_{BIAS} for impedance-matching and improves efficiency. The differential dead-time (DDT) control circuit (bottom of Fig. 15.3.3) eliminates the overlapping switching of differential operation during high-power transfer and avoids input-current mismatch, overcurrent, and RF choke damage ($RFC_{1,2}$). GaN1 and GaN2 do not turn on at the same time due to the DDT control.

In the upper left of Fig. 15.3.4, severe loop delays not only cause RC power losses, but also increase the peak V_{SW} to approximately 6.5 times (in conventional designs) or 5 times that in [4] relative to the input voltage V_{IN}. Large V_{SW} causes damage to GaN devices. The proposed RCC technique (top right of Fig. 15.3.4) effectively eliminates the loop delays even due to PVT variations to restrict V_{SW} to only 3.6 times V_{IN}. The delay line selector (DLS) uses V_G (the gate of the GaN) as a reference to generate a sequence of stepping pulses V_{BD1} to V_{BDN}. Then, the RCC loop searches for the delay per period by shifting the stepping pulse V_{BD}. When the rising edge of V_S is between the rising edges of the two adjacent stepping pulses generated by the DLS, the RCC loop locks the clock V_{CK} of the counter. The RCC circuit outputs accurate timing references V_{LD} and V_{LG} for the DIM controller to improve impedance-matching accuracy. In addition, the analog fine loop can effectively remove the HS loss due to the limited resolution of the digital control. In the timing diagram (bottom of Fig. 15.3.4), first, during the startup transient, the RCC technique shifts the window generated by V_{LD} and V_{LG} (step 1) to find the corresponding delay-included reference V_S, then sets the previous window as the calibrated timing bounds (step 2) after the window finds V_S. Second, after the RCC calibration, the controller enters the coarse loop for the transient enhancement. Third, when V_S locks into the timing bound V_{LD} and V_{LG} again (step 3), the coarse loop switches to the fine loop. Finally, the GaN power switch eventually operates in ZVS and ZVDS when V_A locks to a preset hysteresis bound (step 4), which consists of V_{R2} and V_{R3}.

The test chip was fabricated in a CMOS SOI 0.5μm process. The proposed WPT system can achieve both ZVS and ZVDS for impedance matching (shown in region 3 of top of Fig. 15.3.5). Besides, with the RCC control loop (in region 1), the measured timing references V_{LD} and V_{LG} along with V_S demonstrate the compensation of loop delays for improving the accuracy. In addition, the coarse-fine loop control (in regions 2 and 3) shows fast and accurate impedance matching, respectively. Figure 15.3.6 shows the differential operation for high-power transfer. The statistic results (left top of Fig. 15.3.6) depict that the proposed TX achieves higher than 86% efficiency over a wide load range. The right top of Fig. 15.3.6 shows the power transfer level and efficiency of the proposed TX compared to prior arts. Relative to the prior art, the highest power level and efficiency are achieved by the proposed WPT system. The bottom of Fig. 15.3.6 shows the comparison table with prior arts. The proposed controller is capable of realizing impedance matching from no load to 100W load without using high-voltage-stress components. Measured results show the RC loss can be reduced from 5% in [4] to 1% corresponding to 2.5% efficiency improvement. The proposed WPT system provides an efficient (peak efficiency of 91%) and compact solution under high power application (100W). The die micrograph is shown in Fig. 15.3.7.

Acknowledgement:
The authors would like to thank Hann-Huei Tsai and Ying-Zong Juang for many useful discussions of high-voltage process. This work was supported and funded by the National Chip Implementation Center (CIC), Taiwan.

References:
[1] Y. Lim et al, "An Adaptive Impedance-Matching Network Based on a Novel Capacitor Matrix for Wireless Power Transfer," *IEEE Trans. on Power Electronics*, vol. 29, no. 8, pp. 4403-4413, Aug. 2014.
[2] F. Mao et al, "A Reconfigurable Cross-Connected Wireless-Power Transceiver for Bidirectional Device-to-Device Charging with 78.1% Total Efficiency," *ISSCC,* pp. 140-141, Feb. 2018.
[3] H. Kennedy et al., "A Self-Tuning Resonant Inductive Link Transmit Driver Using Quadrature-Symmetric Phase-Switched Fractional Capacitance," *ISSCC,* pp. 370-371, Feb. 2017.
[4] C.-H. Yeh et al., "A 70W and 90% GaN-Based Class-E Wireless-Power-Transfer System with Automatic-Matching-Point-Search Control for Zero-Voltage Switching and Zero-Voltage-Derivative Switching," *ISSCC,* pp. 138-139, Feb. 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

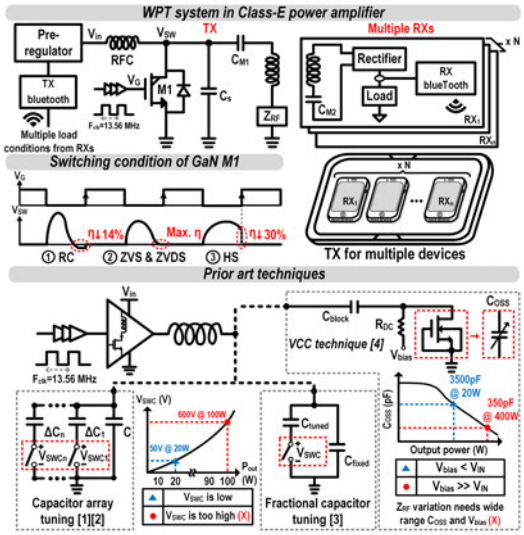


Figure 15.3.1: WPT system for charging multiple devices and prior-art impedance-matching techniques.

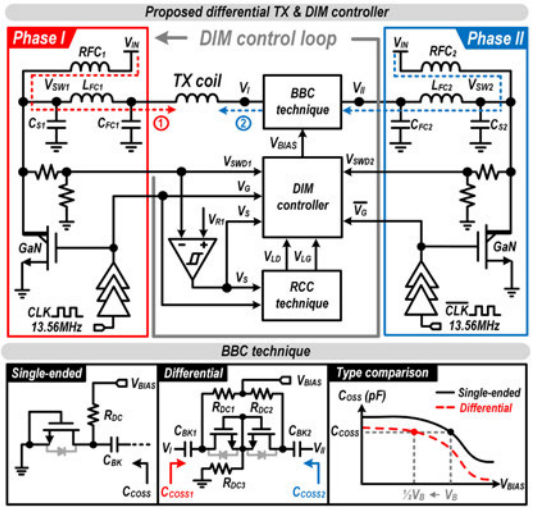


Figure 15.3.2: A high-power differential TX with the DIM controller, RCC technique, and BBC technique is proposed to improve impedance-matching accuracy.

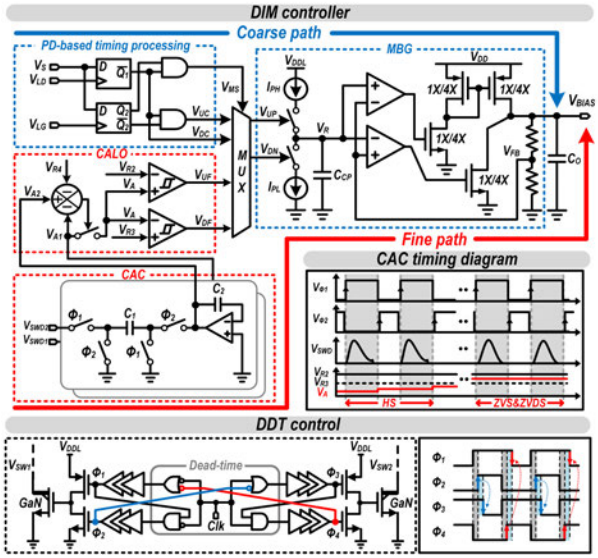


Figure 15.3.3: Circuit implementation for the DIM controller and DDT control.

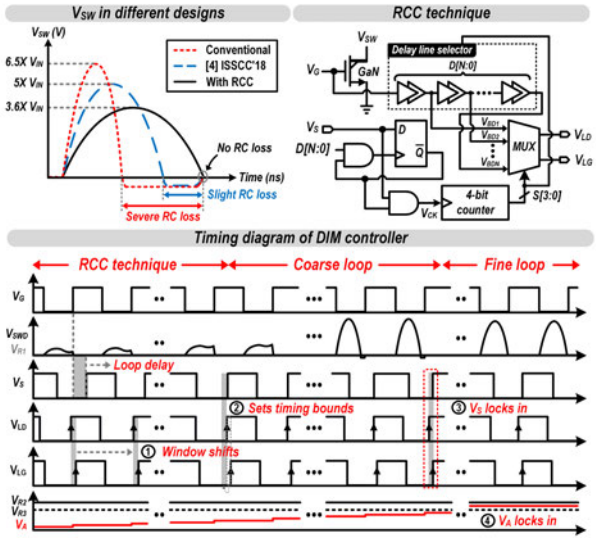
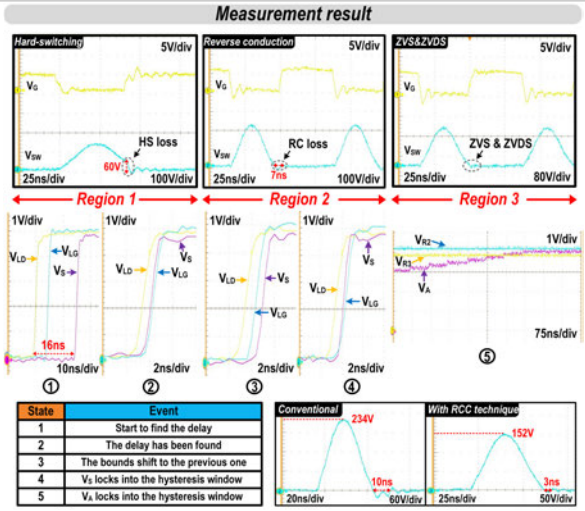
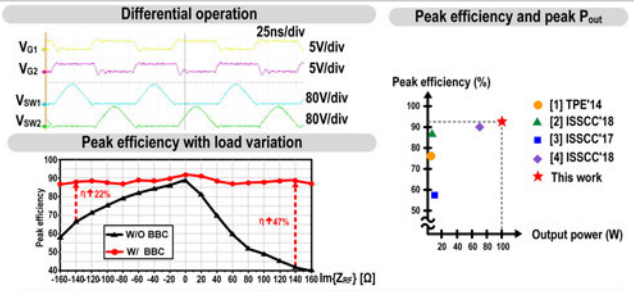


Figure 15.3.4: Circuit implementation of the RCC technique and the operation of DIM control loop.



State	Event
1	Start to find the delay
2	The delay has been found
3	The bounds shift to the previous one
4	VS locks into the hysteresis window
5	VA locks into the hysteresis window

Figure 15.3.5: ZVS and ZVDS techniques show correct impedance matching after fine tuning (top); key waveform of the coarse-fine control and the RCC technique (middle); and the RCC technique reduces the RC loss (bottom).



Comparison table

	[1] TPE'14	[2] ISSCC'18	[3] ISSCC'17	[4] ISSCC'18	This work
Process	CMOS 0.18μm	CMOS 0.35μm	CMOS 0.18μm	CMOS SOI 0.5μm	CMOS SOI 0.5μm
Operation frequency [MHz]	10.4~13.56	13.56	0.075~2	6.78	13.56
Control method	Capacitor matrix	Capacitor matrix	Fractional capacitor	Voltage controlled capacitor	Bilateral bias capacitor
Control loop	Single	Single	Single	Dual	Dual
Accuracy	Low	Low	Medium	Medium	High
Peak output power [W]	1	2.7	N/A	70	100
Peak efficiency [%]	88	78.1	N/A	90	91
Power switch	MOSFET	MOSFET	MOSFET	GaNFET	GaNFET

Figure 15.3.6: Peak efficiency versus load variations and the comparison table with prior arts.

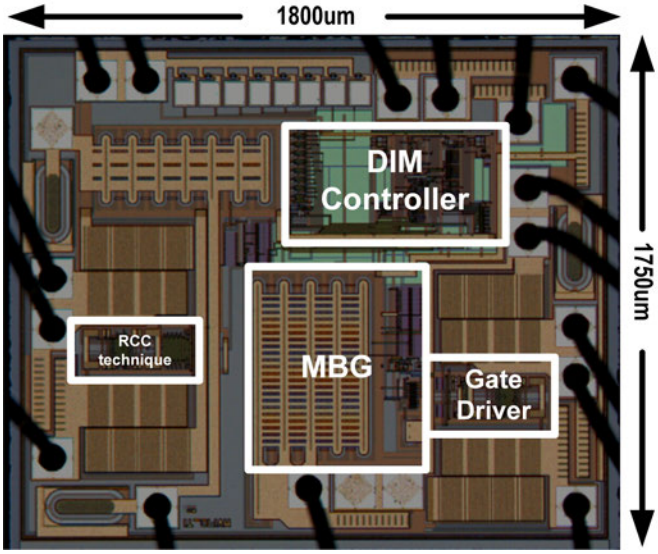


Figure 15.3.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

15.4 A 52% Peak-Efficiency >1W Isolated Power Transfer System Using Fully Integrated Magnetic-Core Transformer

Zhuo Yue[1], Ma Shaoyu[2], Zhao Tianting[1], Qin Wenhui[2], Zhao Yuanyuan[3], Guo Yingjie[1], Chen Baoxing[3]

[1]Analog Devices, Beijing, China; [2]Analog Devices, Shanghai, China
[3]Analog Devices, Wilmington, MA

Galvanic isolation is becoming increasingly important in a wide field of applications to guarantee human safety or better reliability in harsh industrial environments. The transfer of both data and power across an isolation barrier is often an essential requisite in these applications. While data transmission with galvanic isolation has been achieved through traditional optocouplers, capacitive coupling, or integrated transformers, transferring power across the isolation barrier presents several challenges regarding the size and the performance. Some of the state-of-the-art devices adopt post-processed micro-transformers [1], which use 6μm-thick plated Au for both the primary and secondary windings. The operating frequency of the integrated transformers is ~180MHz to achieve optimal quality factor of ~10, thus the switching power loss reduces the efficiency below 35%. At such a high frequency, the radiated emissions are a big concern for end users, especially for automotive and medical applications. In this work, a fully integrated isolated power transfer system is demonstrated using an integrated transformer with magnetic core. It represents a major improvement due to higher quality factor and lower operation frequency as a result of magnetic core integration on silicon.

Figure 15.4.1 shows the integrated solenoid transformer construction. The core is made of laminated magnetic material, which is a kind of high-permeability, high-resistivity material that provides good inductive coupling and low loss. Laminated layers further reduce eddy-current loss by isolating the magnetic layers. Gold windings are plated to an 8μm thickness to reduce metal loss and are co-wound to achieve optimal coupling. Polyimide provides isolation between primary and secondary windings, as well as between windings and the core. The transformer is processed on a high-resistivity silicon substrate. The inductance of the transformer reaches >100nH and the quality factor is higher than 13 at the 25MHz switching frequency, which make it possible to operate the power-stage driver at lower frequency than with other integrated isolated power converters. This benefits the reduction of the gate-drive loss and potential radiation.

Figure 15.4.2 illustrates the architecture of the proposed power transfer system, which consists of an isolated feedback link and an LLC full-bridge resonant power stage using an integrated magnetic core transformer as the power delivery device. The switching transient spikes are eliminated during MOSFET turning on thanks to MOSFET zero-voltage switching (ZVS) of the LLC topology, which benefits the EMI performance. On the other hand, the common-mode current across the isolation barrier results in dipole emissions and high EMI in isolated power transfer systems [2]. Various design techniques are adopted to minimize the common-mode current in [3]. Two external equal capacitors ($2\times C_R$) together with the transformer make up the LLC resonant converter, which is of a fully symmetric architecture. The voltage difference at the transformer terminals is fully differential and the common-mode voltage fluctuation is reduced significantly during the switching events. A multistage pre-driver with an internal latch makes the differential gate-drive signal more synchronous. The last stage of the pre-driver adopts the Class-AB topology with cross-coupled connections, which turns off power switches simultaneously. To improve the output power capability to higher than 1W, the parameters of the resonant network were redesigned with the resonant capacitance of 2nF and the transformer primary inductance of 130nH.

Figure 15.4.3 illustrates the control strategy to deliver >1W output power at the optimized efficiency. Traditionally, for an LLC resonant converter, frequency control is employed to adjust the resonant tank impedance and regulate the output voltage. However, using an on-chip transformer, the resonant converter operates at 10s of MHz. At light load, the switching frequency is higher, and the excessive power loss of gate drivers causes the efficiency to drop drastically. Moreover, the higher switching frequency implies higher emissions. In the prior art [3], a PWM control scheme is employed to achieve power regulation, where an LLC power stage operates at a fixed switching frequency, and a low-frequency PWM signal is used to turn on/off the power transferring action. However, to achieve >1W output power the PWM-controlled LLC converter suffers from low efficiency. On the other hand, optimal efficiency is obtained at the cost of lower output power of 0.7W. To address these challenges, a multimode control scheme is proposed in this paper, which combines the PWM control and frequency control to achieve optimal efficiency and output power. This approach essentially enables optimal efficiency over a wide output-power range of up to 1.1W.

Figure 15.4.4 shows the implementation of the multimode-feedback-control link. The transfer of the feedback signal is realized with a digital isolator using an on-chip coreless transformer [1]. The feedback link operates at 100s of kHz to avoid excessive power consumption in the isolated feedback channel, and a frequency multiplier (64×) at the primary side enables the LLC converter to operate close to the resonant frequency of 25MHz. The output-power regulation is achieved using an error amplifier, and the output voltage of the error amplifier is used to implement the multimode-control approach. For a light load, open-loop operation allows V_{OUT} to rise, resulting in a lower V_{EA}. For V_{EA} lower than a defined threshold, V_{PWM_TH}, the output is regulated by the duty cycle of the PWM-comparator output, and the loop is operated in the PWM mode. In this scenario, the PWM switching frequency is fixed at 400kHz, and the frequency multiplier at the primary side uses this PWM switching frequency to generate the drive signal of ~25MHz (400kHz × 64) for the LLC resonant converter. On the other hand, the duty cycle of the feedback signal is used to control the on/off time of the LLC power stage. For higher output power, V_{EA} is higher than V_{PWM_TH}, and the converter enters the frequency-control mode. In this scenario, the PWM-comparator output saturates to DC high, and the VCO takes over the loop control. V_{EA} modulates the frequency of the VCO from 400 to 250kHz, and the frequency is boosted (64×) to drive the power stage at the primary side. To make a smooth transition between the PWM mode and the frequency-control mode, the duty cycle of the VCO output is clamped at a defined maximum number with respect to the duty cycle of the PWM control. Using a single feedback link, the multimode approach is implemented to achieve optimal efficiency and output power.

Figure 15.4.5 shows the measurements of the proposed power transfer system. The switching signal was measured at one of the full-bridge driver output. When output load varies, the proposed power converter switches between the PWM mode and the frequency-control mode adaptively for optimal efficiency. Using a fully integrated transformer with a magnetic core and the multimode control scheme, the peak efficiency of 52% is achieved. Meanwhile, the efficiency is higher than 46% at the output power from 100mW to 1W, and the maximum output power of 1.1W is achieved at room temperature. At the worst conditions of low input voltage (4.5V) and high temperature (125°C), the converter is able to deliver output power up to 0.75W. Thanks to the fully symmetric LLC resonant topology, the emissions of the power transfer system are reduced significantly passing CISPR 22 Class B limits with a 1.9dB margin at 500mW output power using 2 ferrites at the power output. Employing a spread-spectrum switching technique, the emissions are reduced further to 5.8dB below CISPR 22 Class B limits at output power of 750mW. In Fig. 15.4.6, the measurement results are summarized and compared with prior art. Figure 15.4.7 shows the micrograph of the power transfer system in a SOIC package. The core of the power transformer with a magnetic core is 2mm×4mm. The isolated feedback transformer is made from two planar spiral coils of a 4μm-thick Au. The power driver die and the rectifier die are fabricated in a 0.35μm BCD process.

Acknowledgement:
The authors would like to thank the Analog Devices ISO group for design support as well as the Analog Devices IPD team for device fabrication.

References:
[1] B. Chen, "Fully Integrated Isolated DC-DC Converter Using Micro-Transformers," *IEEE Appl. Power Electron. Conf Expo. (APEC)*, pp. 335-338, Feb 2008.
[2] Analog Devices, AN-0971, "Recommendations for Control of Radiated Emissions with isoPower Devices." Rev. C, Accessed on Aug. 1, 2018, <https://www.analog.com/media/en/technical-documentation/application-notes/AN-0971.pdf>
[3] T. Zhao et al., "An Isolated DC-DC Converter with Fully Integrated Magnetic Core Transformer," *CICC*, pp.1-4, May. 2017.
[4] Analog Devices, Evaluation Kit for Adum6020 Low Emission, Isolated DC-to-DC Converters, Accessed on Aug. 1, 2018, <https://www.analog.com/en/products/adum6020.html#product-evaluationkit>
[5] N. Spina et al., "Current-Reuse Transformer-Coupled Oscillators with Output Power Combining for Galvanically Isolated Power Transfer Systems", *IEEE TCAS-I*, vol. 62, pp. 2490-2498. Dec. 2015.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

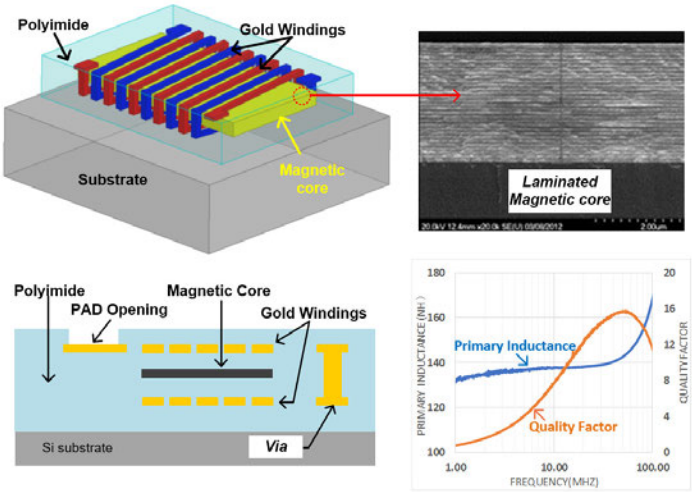


Figure 15.4.1: Integrated transformer construction with a laminated magnetic core, the device cross-section, and the key parameters: inductance and quality factor.

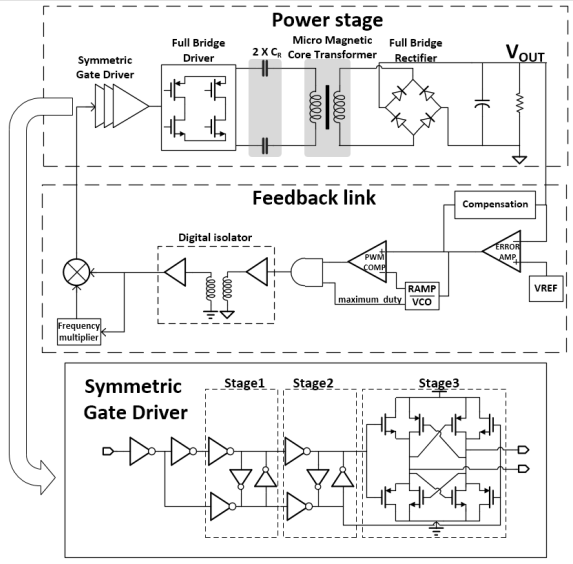


Figure 15.4.2: Architecture of the isolated power transfer system.

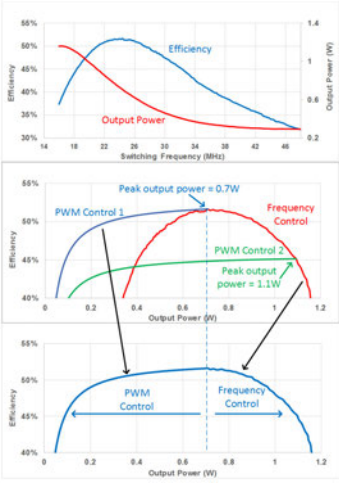


Figure 15.4.3: (top) Efficiency and output power for frequency-control mode as a function of switching frequency, (middle) optimized efficiency as a function of output power for the PWM control and frequency control, and (bottom) efficiency of the proposal multimode control approach.

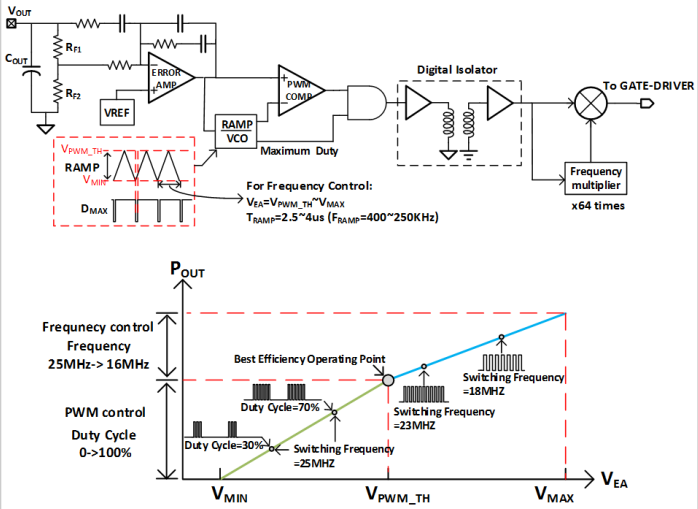


Figure 15.4.4: Implementation of the multimode-feedback-control link.

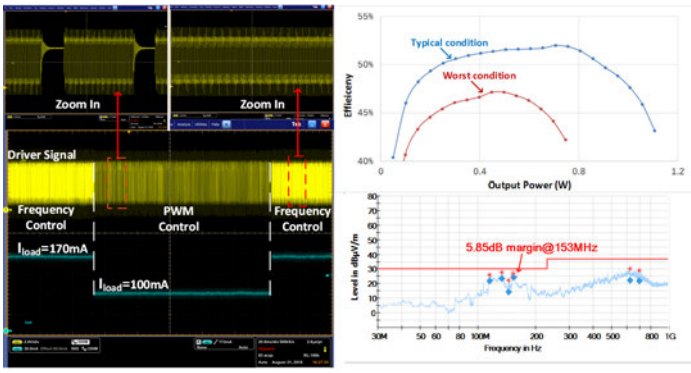


Figure 15.4.5: (left) Measured switching waveforms for output-current switching between 170 and 100mA, (right) the measured efficiency at different output-power levels and the measured radiation when using spread-spectrum techniques.

Parameters		[4]	[5]	[3]	This work
Transformer devices parameters	Quality factor	N/A	N/A	11	15.8
	Inductance	N/A	N/A	550nH	130nH
	Coupling factor	N/A	>0.6	0.92	0.92
	Isolation rating	5kV	5kV	5kV	4.5kV
Circuit topologies and parameters	Power stage architecture	N/A	Current-reuse transformer-coupled oscillator	Symmetric LLC resonant converter	Symmetric LLC resonant converter
	Control scheme	PWM control	N/A	PWM control	Multi-mode control
	Operation frequency	180MHz	240MHz	15.5MHz	16 to 25MHz
	Input voltage	5V	5V	4.5 to 5.5V	4.5 to 5.5V
	Output voltage	5V/3.3V	8V	5V	5V
	Output Power	500mW	200mW	200mW	1100mW
	Peak efficiency	33%	27%	46%	52%
	EMI Performance	Pass Class B limit with 4dB margin	N/A	Pass Class B limit with 10dB margin	Pass Class B limit with 5.8dB margin

*** EMI Performance is evaluated to CISPR 22 standard

Figure 15.4.6: Performance summary and comparison with previous works.

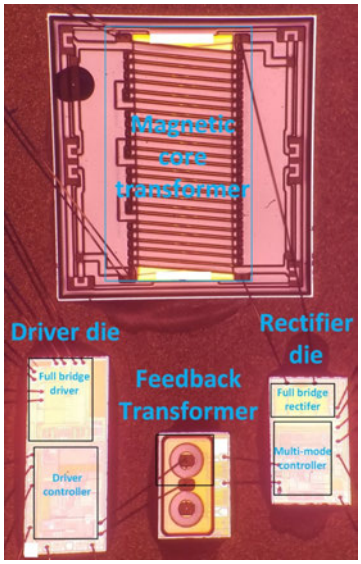


Figure 15.4.7: Micrograph of the overall system assembled in package.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

15.5 An 800mW Fully Integrated Galvanic Isolated Power Transfer System Meeting CISPR 22 Class-B Emission Levels with 6dB Margin

Wenhui Qin[1], Xin Yang[2], Shaoyu Ma[1], Fang Liu[2], Yuanyuan Zhao[3], Tianting Zhao[2], Baoxing Chen[3]

[1]Analog Devices, Shanghai, China
[2]Analog Devices, Beijing, China
[3]Analog Devices, Wilmington, MA

Industrial field networks play a key role in industrial automation systems. Employing galvanic isolation in the transceivers of such networks has become an essential requisite to guarantee safety and better reliability in harsh industrial environment. For industrial transceivers, isolated power levels between 100mW and 1W are required. Transciever size and cost are also constrained in many industrial applications. However, delivering 100s of mW is challenging for fully integrated solutions without bulky external transformers, and it is also important to control radiated emissions at such power levels. In [1,2], different ways of implementing fully integrated isolated power links have been proposed. However, they can only deliver 10s of mW, which limits them to low-power applications, e.g., sensor interfaces. Some of the state-of-the-art commercial devices use on-chip transformers made from thick gold traces to deliver 500mW at 33% efficiency [3], but a 4-layer PCB is required to implement a stitching capacitor across the isolation barrier to meet CISPR 22 Class B Electromagnetic Interference (EMI) standards [4]. In this work, a fully integrated galvanic isolated power transfer system is proposed. By adopting various circuit techniques, 800mW output capability and 34% efficiency is achieved, and component-level emissions are kept low, eliminating the need for costly EMI-mitigation techniques and simplifying the EMI-certification process.

As shown in Fig. 15.5.1, the key component is an on-chip transformer made from two co-axial planar spiral coils. The thickness of their gold windings (6μm) involves a trade-off between cost and parasitic resistance. Limited by the die area, a 900μm diameter coil with 3 turns of 60μm-wide windings achieves an optimal quality factor. A thick (>20μm) polyimide layer provides better than 5kV isolation between the coils, exploiting polyimide high dielectric-breakdown strength [3]. The transformer has a measured inductance of about 15nH, and its quality factor reaches the maximum of 6.8 at 200MHz.

The proposed system block diagram is shown in Fig. 15.5.2, which consists of a power-transmitter (Tx) die, a transformer die, and a power-receiver (Rx) die. To reduce switching losses at frequencies of 100s of MHz, a resonant switching scheme based on an LC tank is employed in the Tx. To maximize the output power, the tank oscillation amplitude is close to πV_{DD}, which is too high for the NMOS devices of a 5V CMOS process. Since using an HV NMOS [3] would require a costly process and large die area, and using a multi-NMOS stacking topology [2] would require complicated gate drivers, we use a low-cost N LDMOSs in a 0.35μm BCD process. Although the LDMOS maximum V_{DS} is 20V larger than the oscillation amplitude, its maximum V_{GS} is the same as that of a 5V CMOS process. An AC-coupled (ACC) structure is used to drive the LDMOS gate. The resulting V_{GS} is limited to 5V by the capacitor divider formed by the capacitor C_D and the LDMOS C_{gs}. Since the on-resistance of an LDMOS is lower than that of an HV MOS, the tank size is 1.6 times smaller than in [3]. A PWM controller in the Rx turns the LC-tank oscillation on/off rather than controlling each resonant switching event of the N LDMOS, separating energy regulation from energy conversion and allowing optimized power transfer and regulation. The PWM controller signal FBCTRL @600kHz is transmitted from the Rx to Tx via a data isolator channel. When the FBCTRL is at a logic high, the tank stops oscillating. At the falling edge of the FBCTRL, a kick-start circuit quickly restarts the oscillation to improve the output drive capability and efficiency. Finally, when the FBCTRL is at a logic low, the tank oscillates freely and delivers power to the Rx.

In the proposed system, the main source of EMI is dipole radiation, as shown in Fig. 15.5.3. TC1 and TC2 are the two end ports of the transformer primary coil, and their voltages are both half-sinusoids that are 180 degrees out of phase. They are, therefore, not differential signals, and so the common-mode (CM) voltage V_{com1} on the primary contains even harmonics of the oscillation frequency f_{osc}, while the CM voltage $V_{com2}=(V_{RC1}+V_{RC2})/2$ at the secondary is quiet. Due to the parasitic capacitance C_{par} between the two coils, the AC component v_{com1} of V_{com1} introduces a CM current ($=v_{com1}/C_{par}$) across the isolation barrier. This current is small and so does not cause safety issues, but it does generate dipole radiation, whose intensity is proportional to its amplitude.

As the dipole radiation is caused by v_{com1}, we can reduce the EMI by reducing the amplitude of v_{com1} with a frequency hopping technique, as shown in Fig. 15.5.4. With an N-bit hopping, the LC tank oscillates at 2^N frequencies in sequence, spreading the energy of v_{com1} over these frequencies. Compared to the result without hopping, $20\log(2^N)$ [dB] reduction of each harmonic peak values can be achieved. Frequency hopping reduces the harmonic peak values effectively only when the frequency-hopping step Δf_{osc} is large enough to ensure that there is no significant energy overlap between adjacent oscillation frequencies in the frequency domain; and its effectiveness degenerates when Δf_{osc} is so small that such overlap occurs, as shown in Fig. 15.5.4.

To maintain the system output capability and power efficiency, the LC-tank oscillation frequency is limited from 160 to 210MHz, corresponding to the transformer optimal quality factor. With the N-bit hopping, the maximum Δf_{osc} is $(210-160)/2^N$ [MHz]. Δf_{osc} decreases exponentially with the increasing of N, so the energy at 2nd harmonic overlaps each adjacent energy peak easily with large N. Finally, a 5b hopping with the FBCTRL frequency f_m=600kHz and Δf_{osc}=1.5MHz is adopted. As shown in Fig. 15.5.5, a 22dB reduction of all harmonic peak values is achieved, although f_m and Δf_{osc} vary with process, voltage, and temperature, more than 20dB reduction is achieved.

To realize the frequency-hopping scheme, there are 31 sub tanks in parallel to the main ACC LC tank, enabled respectively by the 31b outputs of a pseudo-random thermometer-code generator (PRTCG), as shown in Fig. 15.5.2. The total capacitance C in the LC resonant network changes according to the quantity of the enabled sub tanks, then the oscillation frequency $1/(2\pi\sqrt{LC})$ hops. The PRTCG output changes at the rising edge of the FBCTRL, then the oscillation frequency is constant during the oscillation-on time and changes during the oscillation-off time, as shown in Fig. 15.5.4.

As shown in Fig. 15.5.7, the die areas of the Rx die and the Tx die are, respectively, 1.35×2.8mm^2 and 1.2×2.3mm^2 in a 0.35μm BCD process, the transformer die is 1.56×3.33mm^2 in a post-passivation process, and they are assembled in an 8-lead wide-body SOIC package. Figure 15.5.6 shows the key performance summary and measured EMI performance of the proposed system, with the comparison to other prior art systems. Thanks to the frequency-hopping technique, with two external beads inserted between the system outputs and PCB secondary-side power/ground plane, the system can pass the CISPR22 Class-B certification with a 6.4dB margin at 500mW output power on a two-layers PCB without any stitching capacitor, which is more than 10× improvement compared with [3]. The ACC LC tank architecture with LDMOSs achieves a 0.8W maximum output power at 25°C and 34% peak efficiency, which is best among all the 5kV isolation rating designs [1,3]. Compared with [2], the system has 13× the maximum output power and 5× the isolation rating improvement.

References:

[1] P. Lombardo et al., "A Fully-Integrated Half-Duplex Data/Power Transfer System up to 40MB/s Data Rate, 23mW Output Power and On-Chip 5kV Galvanic Isolation," *ISSCC*, pp. 300-301, Feb. 2016.
[2] L. Chen et al., "A 50.7% Peak Efficiency Subharmonic Resonant Isolated Capacitive Power Transfer System with 62mW Output Power for Low-Power Industrial Sensor Interfaces," *ISSCC*, pp. 428-430, Feb. 2017.
[3] B. Chen, "Fully Integrated Isolated DC-DC Converter Using Micro-Transformers," *IEEE Appl. Power Electron. Conf Expo. (APEC)*, pp. 335-338, Feb 2008.
[4] Analog Devices, AN-0971, "Recommendations for Control of Radiated Emissions with isoPower Devices," Rev. C, Accessed on Jan. 2014, <https://www.analog.com/media/en/technical-documentation/application-notes/AN-0971.p

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

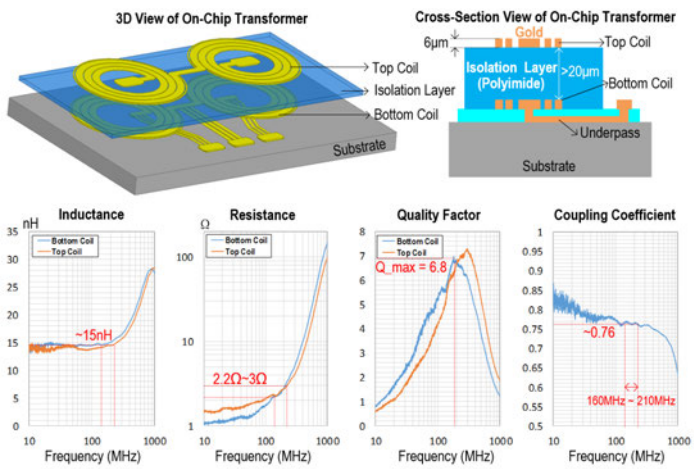


Figure 15.5.1: 3D view and measured results of the on-chip transformer.

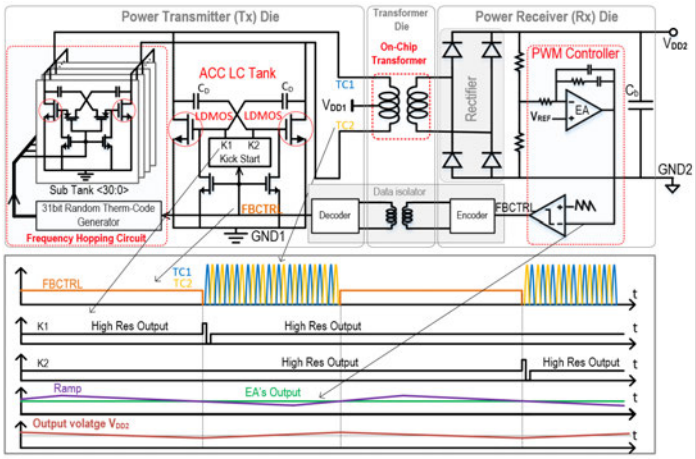


Figure 15.5.2: Block diagram of the proposed fully integrated galvanic isolated DC-DC converters.

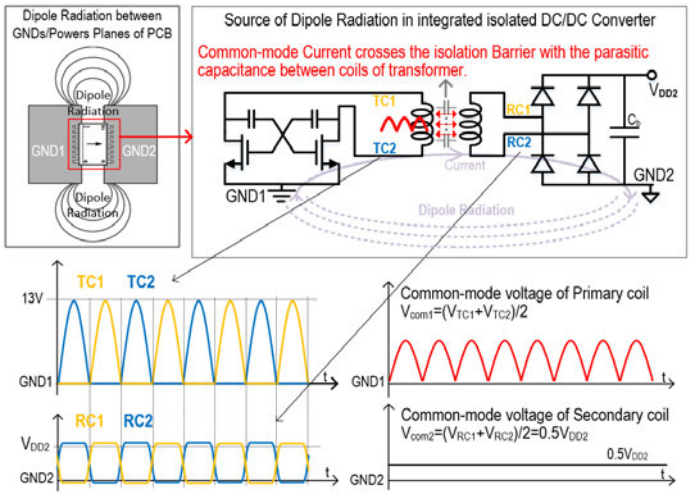


Figure 15.5.3: The source of dipole radiation in an integrated isolated DC-DC converter.

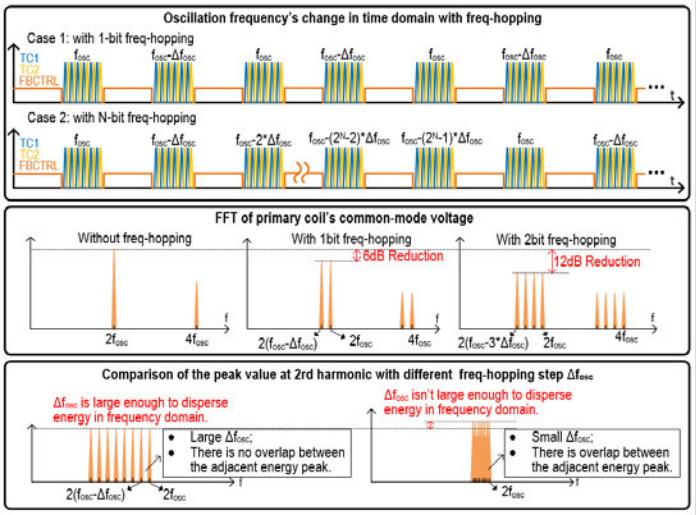


Figure 15.5.4: The mechanism of the frequency-hopping technique to reduce the peak values of harmonics and the limitation of the reduction.

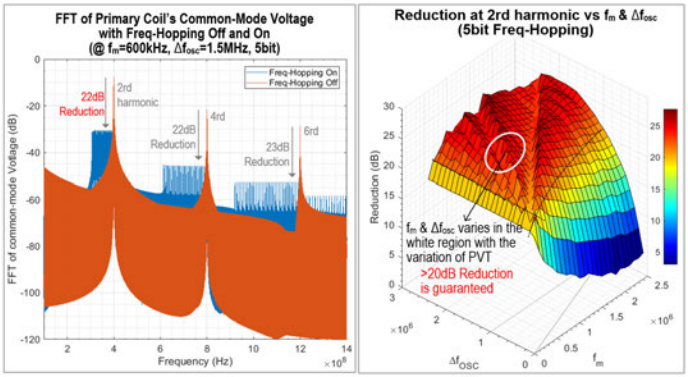
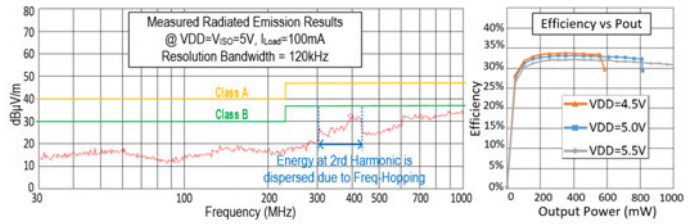


Figure 15.5.5: Simulation results of the reduction in the peak values of harmonics introduced by the 5b frequency hopping.



Reference	ISSCC'16 [1]	ISSCC'17 [2]	APEC'08 [3]	This work
Technology	0.35μM BCD	0.35μM BCD	0.6μM HV BiCMOS	0.35μM BCD
Isolation Type	magnetic	capacitive	magnetic	magnetic
Isolation Rating	5kV	<1kV	5kV	5kV
Input Voltage	3.3V	Up to 6V	5V	4.5~5.5V
Output Voltage	2V	3.3V	3.3V~5V	3.3V~5V
Max. Pout	20mW	62mW	0.5W	0.8W
Peak Efficiency	10.8%	50.7%	33%	34%
EMI Performance*	NA	NA	~20dB Over Class-B (without stitching Cap)	Pass Class-B with 6dB margin (without stitching Cap)

*Class-B means CISPR22 Class-B Certification

Figure 15.5.6: Measurement results of EMI and efficiency and performance comparison.

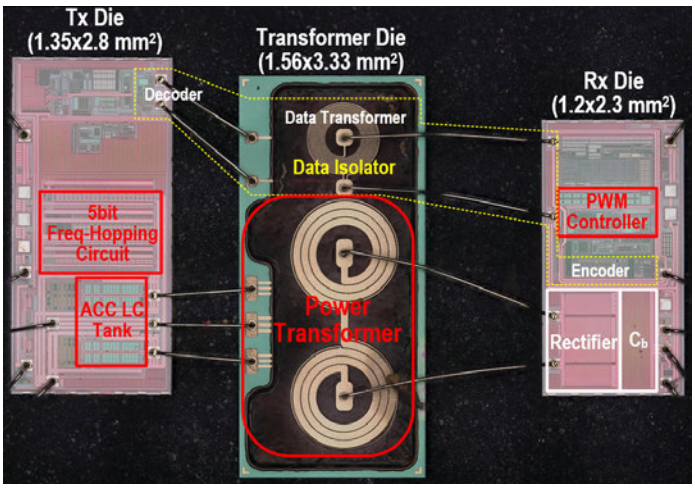


Figure 15.5.7: Micrograph of the whole system.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

15.6 A 10MHz *i*-Collapse Failure Self-Prognostic GaN Power Converter with T_J-Independent In-Situ Condition Monitoring and Proactive Temperature Frequency Scaling

Yingping Chen, D. Brian Ma

University of Texas at Dallas, Richardson, TX

With superior figure of merits, GaN switchs are highly anticipated to replace MOSFETs in high-performance power circuits [1,2]. However, GaN technology today still faces formidable reliability challenges [3]. While GaN device aging and failure mechanisms are not as well-studied as silicon counterparts, its unique structure and operation also induce new aging and failure problems. Use a GaN switch M_H in a buck converter of Fig. 15.6.1 as an example. As a high-side switch, it faces large-switching-current and high-input-voltage stress in each charge phase. After repetitive switching actions, a number of electron carriers can be injected into the AlGaN barrier and buffer layers, known as hot-electron injection. In discharge phase, M_H is off, but M_L becomes conductive, which shorts the source of M_H to ground, creating high V_{DS} stress on M_H. This induces charge traps in the insulator and buffer layers, known as charge trapping. As a joint effect of both mechanisms, trapped or injected electrons in the insulator, AlGaN barrier and buffer layers repel free electrons in the channel when M_H is on, weakening the 2-dimensional electron-gas (2DEG) layer and further elevating hot-electron injection. This effect, known as current collapse or *i*-collapse for short, degrades channel conductivity, increases the on-resistance R_{DS_ON}, and is a major cause of GaN-switch aging and failure [3]. On the other hand, another aging cause is thermal effect. To reduce manufacture costs and improve technology compatibility, it is common to fabricate GaN transistors on a silicon substrate. Accordingly, to reduce lattice mismatch, an AlGaN buffer layer is often inserted (Fig. 15.6.1). However, this increases the junction-to-ambient thermal resistance $R_{\theta JA}$, which, together with the increased R_{DS_ON} due to the *i*-collapse, causes higher power and heat generation, elevating the junction temperature, T_J. According to Arrhenius' Law, as T_J increases, the mean-time-to-failure (MTTF) drops exponentially [4]. Even worse, the elevated T_J deteriorates the *i*-collapse effect with even higher R_{DS_ON}, significantly reducing device lifetime.

To forecast the course of device failure and enhance circuit longevity, as shown in Fig. 15.6.2, this paper proposes three major techniques including: 1) R_{DS_ON} based aging condition monitoring, 2) gate-leakage I_{GSS} inspired, aging-independent, cross-die temperature sensing, and 3) proactive system temperature frequency scaling. To monitor *i*-collapse induced aging, R_{DS_ON} is highly desirable as the precursor because its change is directly related to the *i*-collapse effect. A simplified approach was accomplished offline in [5] for IGBTs. However, this can be very inaccurate and impractical since operation conditions between offline and online can largely differ, leading to high false monitoring rate (FMR). Hence, an online in-situ condition monitoring is highly preferable. A design in [6] reported an online condition monitoring by observing pole variation in the loop gain for power MOSFETs. However, the implementation of pole location precursor is highly sophisticated and noise-sensitive, resulting in a high FMR. Also, both pole location and R_{DS_ON} are T_J-sensitive precursors, meaning that self-heating and ambient temperature fluctuation can confuse the aging monitoring and raise FMR largely. Online T_J-effect calibration has to be implemented in the R_{DS_ON} condition monitoring. For such, an I_{GSS}-inspired T_J sensor is developed on the ground that I_{GSS} of a GaN switch is both T_J-sensitive and aging-independent. With sensed T_J, the in-situ condition monitor effectively removes the impact of T_J on R_{DS_ON} through the T_J-dependence remover. This also allows us to take more proactive measure to slow down the aging process through adaptive temperature frequency scaling (TFS). Overall, the three techniques require no interruption of conventional closed-loop voltage regulation and are completely non-intrusive. The on-chip implementation achieves online condition monitoring and failure self-prognosis with satisfying the FMR and T_J independence.

Figure 15.6.3 details the circuit implementation of the T_J-independent in-situ condition monitoring. To sense the precursor R_{DS_ON} online, both V_{DS_ON} and I_{DS_ON} are measured using respective detector circuits. A time synchronizer synchronizes the two detectors to ensure V_{DS_ON} and I_{DS_ON} are measured simultaneously. Meanwhile, as the fluctuation of T_J causes large variation on R_{DS_ON}, its influence must be removed. However, conventional temperature sensing circuits, such as PTAT, are not applicable in this scenario, because most GaN switches are not monolithically integrated with the controller and other sensing blocks. Hence, the chip temperature T_{IC} fundamentally differs from the T_J of the GaN switch. To overcome the problem, a cross-die sensing approach is proposed, which uses the leakage-current I_{GSS} inspired, aging-independent T_J sensor in Fig. 15.6.3. A device physics study proves that I_{GSS} is temperature-sensitive and aging independent, which becomes a perfect sensing parameter. As for the operation of the T_J sensor, as shown in the timing diagram of Fig. 15.6.3, at the beginning of each charge phase, the feedback control signal V_{PWM} triggers the gate driver to turn on M_H, which also turns on the sensing timer. After a time of t_{SETTLE}, the timer turns off the switch S_{SENSE} to initialize I_{GSS} sensing. At this instant, the gate of M_H is disconnected from the buffer driver. Due to I_{GSS}, its gate capacitor C_G is discharged at a rate of I_{GSS}/C_G, causing the gate-voltage drop of ΔV_{GSH}, which is AC-coupled and linearly amplified as V_{GSS}. At the end of t_{SENSE}, V_{GSS} is sampled as V_{TJ}. Meanwhile, the timer turns on S_{SENSE}. The gate-drive voltage V_{DR} is re-applied at the gate. Note that to minimize the impact on regular gate driving, it is important to keep t_{SENSE} very short and $\Delta V_{GSH} << V_{DR}$. With the sensed temperature (V_{TJ}), the T_J-dependence remover eliminates the temperature impact on R_{DS_ON} and calibrates it to R_{DS_ON}'. With such a technique, T_J-dependence of R_{DS_ON} drops from 40% to 2.1% between 0°C and 125°C, according to our measurement data. This reduces the FMR by 19 times.

The proposed in-situ condition monitoring and T_J sensing facilitate the design of a proactive measure of the device longevity improvement through a temperature-frequency-scaling (TFS) scheme. As depicted in Fig. 15.6.4, the operation speed and the heat generation can be modulated jointly by performance parameters (load current, transient speed, and efficiency) and T_J. At low level of T_J, aging stress is the least significant. The converter is regulated by the feedback regulation loop and optimized for performance. When T_J rises beyond a predefined threshold T_{SCL}, proactive thermal management is activated. Switch S_{LE} is turned on to enable the longevity-enhancement (LE) mode, in which the converter modulates f_{SW} inversely against T_J to reduce heat generation. To achieve optimal operation between performance and thermal aging, instant load condition is monitored. When the load current I_O changes, fast response is required to avoid large V_O overshoot/undershoot, which causes potential malfunctions of load modules. In this situation, V_{HP} is enabled to trigger the high-performance (HP) mode, minimizing the transient delay and voltage damping. At light load, f_{SW} decreases and discontinuous-conduction mode (DCM) is reinforced. On the opposite side, when T_J exceeds T_{MAX}, the converter is forced to turn off to protect the device from thermal failure.

The converter is implemented with a 0.18μm HV CMOS process, using two enhancement-mode GaN HEMTs (EPC8009) as M_H and M_L. It supports 5-to-40V input and 3.3-to-5V output at 10MHz and delivers a maximum power of 6W. The performance of the I_{GSS}-inspired T_J sensor is tested in Fig. 15.6.5. The on-chip sensed voltage V_{TJ} increases as T_J rises. The measurement data coincide well with the 3rd-order polynomial-fitted thermal model. The transient measurements on V_{GSH} clearly verify the sensing operation at 25°C and 75°C respectively. With the same t_{SENSE} of 20ns, ΔV_{GSH}s, caused by I_{GSS} discharge, are 35mV and 120mV, respectively, which are much lower than the gate drive voltage (5V) for the non-intrusive sensing in Fig. 15.6.3. On the other hand, the I_{GSS} versus thermal cycle test proves that the I_{GSS} aging dependence is below ±2.1%, much lower than conventional precursors. To validate the TFS scheme, the f_{SW}-T_J test is first given in Fig. 15.6.6. As T_J increases, f_{SW} scales across operation zones to reach an optimal operation between performance and aging. With the LE mode, P_{LOSS} is reduced along with f_{SW} scaling to suppress T_J, lowering the risk of aging greatly. By enabling the HP mode, I_O step-up and step-down transient responses are improved by 50% and 70%, respectively. During thermal stress tests, R_{DS_ON}' increases gradually after 900 thermal cycles, offering clear and accurate tracking on the course of aging. Figure 15.6.7 shows the die micrograph with an active die area of 1.56mm^2.

References:

[1] M. Guz et al., "Gallium Nitride: Identifying and Addressing Challenges to Realize the Full Potential of GaN in Power Conversion Applications," *IEEE Power Electronics Magazine*, vol. 5, no. 2, pp. 34-39, June. 2018.

[2] Y. Chen et al., "A 10MHz 5-to-40V EMI-Regulated GaN Power Driver with Closed-Loop Adaptive Miller Plateau Sensing," *IEEE Symp. VLSI Circuits*, pp. C120-C121, June 2017.

[3] S. Bahl et al., "Application Reliability Validation of GaN Power Devices," *IEDM*, pp. 20.5.1-20.5.4, Dec. 2016.

[4] B. Paine et al., "Lifetesting GaN HEMTs with Multiple Degradation Mechanisms," *IEEE Trans. Device and Materials Reliability*, vol. 15, no. 4, pp. 486-494, Dec. 2015.

[5] V. Smet et al., "Ageing and Failure Modes of IGBT Modules in High-Temperature Power Cycling," *IEEE Trans. Industrial Electron.*, vol. 58, no. 10, pp. 4931-4941, Oct. 2011.

[6] S. Dusmez et al., "*In Situ* Condition Monitoring of High-Voltage Discrete Power MOSFET in Boost Converter Through Software Frequency Response Analysis," *IEEE Trans. Industrial Electron.*, vol. 63, no. 12, pp. 7693-7702, Dec. 2016.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

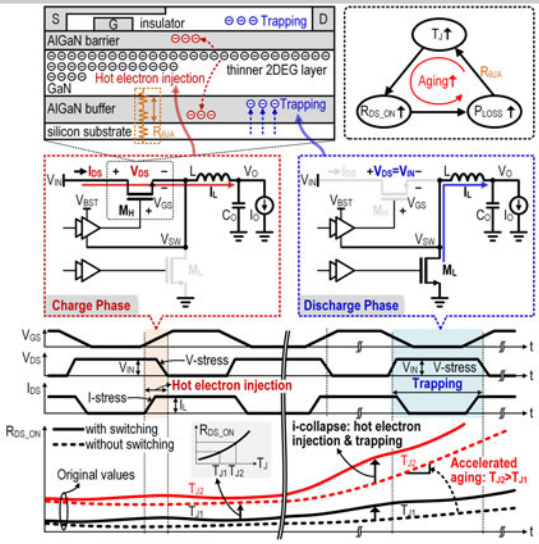


Figure 15.6.1: Device- and circuit-level *i*-collapse aging mechanisms in GaN power switches.

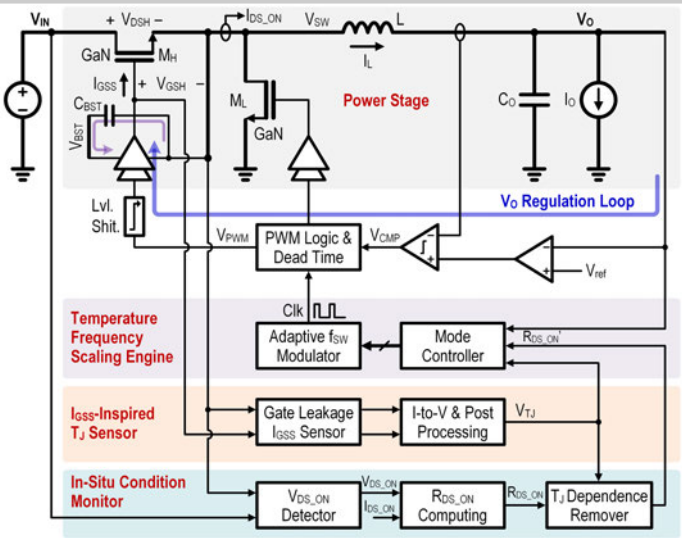


Figure 15.6.2: A GaN buck converter with proposed in-situ R_{DS_ON}-based aging-condition monitoring, I_{GSS}-inspired T_J sensing and the TFS engine.

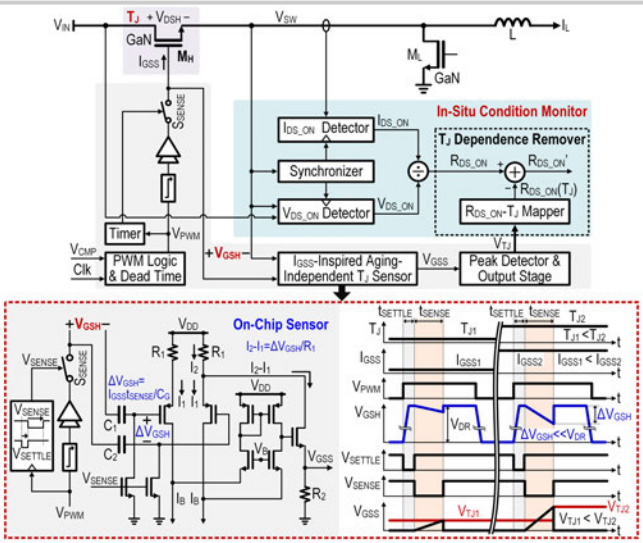


Figure 15.6.3: Detailed implementation of the T_J-independent in-situ condition monitoring with the I_{GSS}-inspired aging-independent T_J sensor.

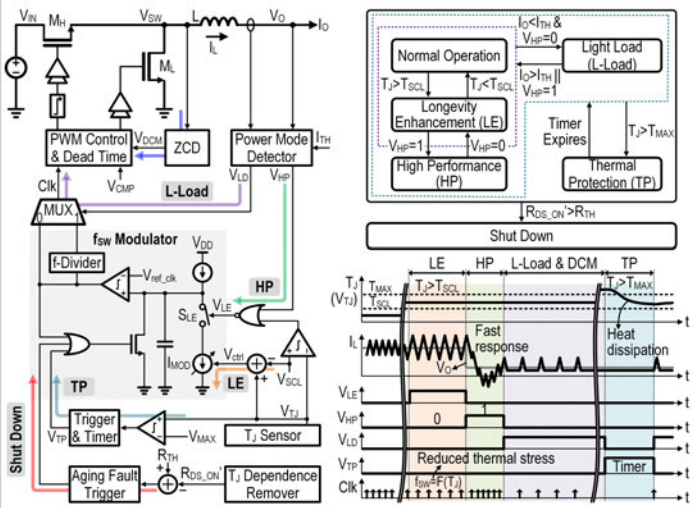


Figure 15.6.4: Block diagram and the operation principle of the temperature-frequency-scaling (TFS) engine.

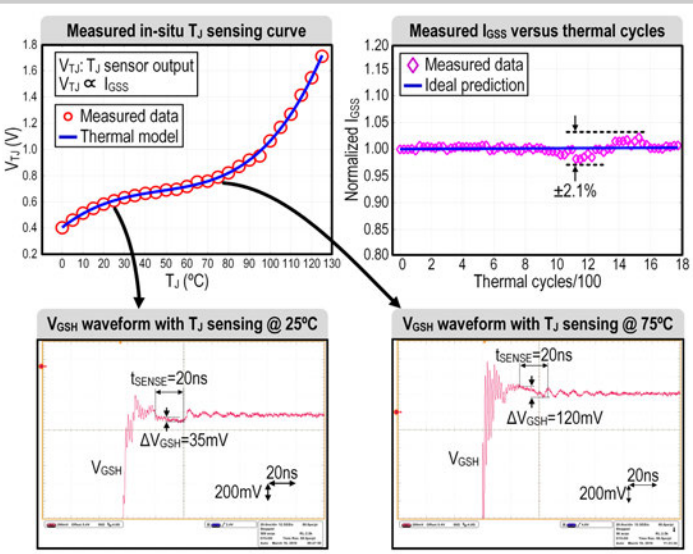


Figure 15.6.5: Measured results of the I_{GSS}-based T_J sensing.

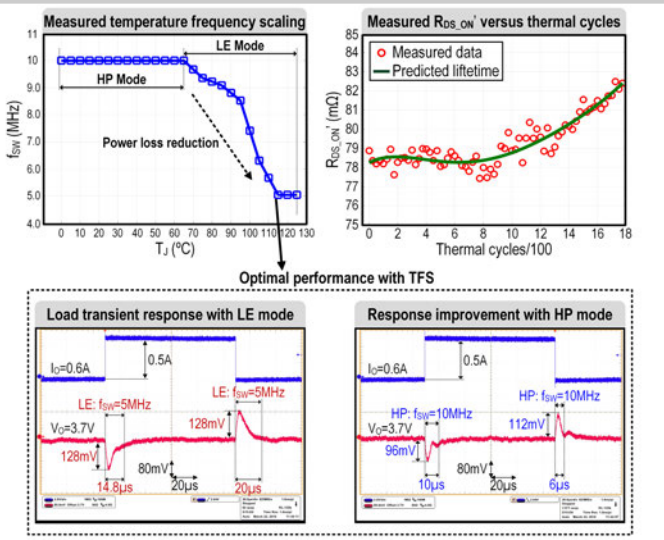


Figure 15.6.6: Temperature-frequency-scaling (TFS) measurement results proving adaptively balanced performance and thermal-aging operation.

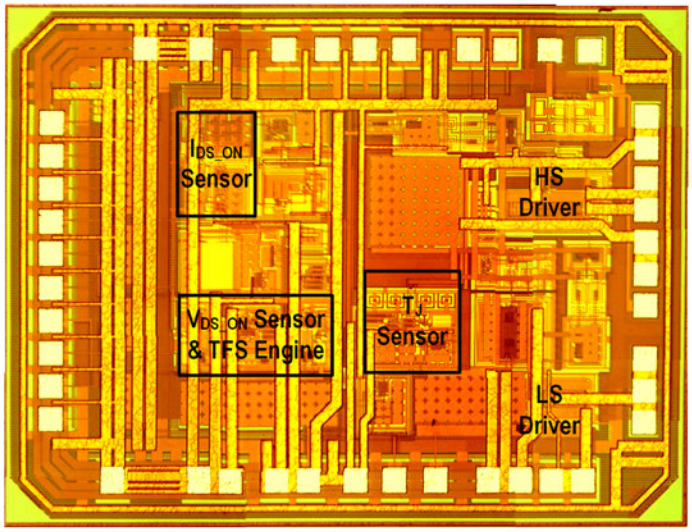


Figure 15.6.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

15.7 An 8.3MHz GaN Power Converter Using Markov Continuous RSSM for 35dBμV Conducted EMI Attenuation and One-Cycle T_{ON} Rebalancing for 27.6dB V_O Jittering Suppression

Yingping Chen, D. Brian Ma

University of Texas at Dallas, Richardson, TX

GaN power switches have gained fast-growing popularity in power electronics. With a similar R_{DS_ON} resistance, they boast 2-to-3-order lower gate capacitance than silicon counterparts, making them highly desirable in high-frequency (f_{SW}), high-performance power converters. However, at high f_{SW}, switching transitions have to be completed in much shorter times, creating much larger *di/dt* and *dv/dt* changes in power stage, which directly link to electromagnetic-interference (EMI) emissions [1]. To suppress EMI, spread-spectrum-modulation (SSM) techniques [2-5] have been proposed. As depicted in Fig. 15.7.1, a periodic SSM (PSSM) is straightforward and easy to implement. However, its EMI suppression is not effective [2]. A randomized SSM (RSSM) can outperform the PSSM, with lower peak EMI and near-uniform noise spreading, but its performance highly relies on the random clock design. In [3], an N-bit digital random clock was reported to achieve a discrete RSSM (D-RSSM). However, the bit number N has to be large in order to achieve satisfying EMI attenuation, significantly increasing circuit complexity, chip area, and power consumption. To overcome this, a thermal-noise-based random clock was proposed [4]. Unfortunately, thermal noise is very sensitive to temperature and is hard to predict. To apply this approach to a practical implementation requires additional signal processing with periodic signals to confine its range of randomization, which, in turn, reduces the benefits of the RSSM. To achieve a near ideal RSSM, a continuous RSSM (C-RSSM) with a cost-effective implementation is highly preferable. Meanwhile, another challenge of applying SSM schemes lies in the fact that the schemes deteriorate V_O voltage regulation. As shown in Fig. 15.7.1, as an SSM scheme continuously or periodically modulates f_{SW}, a converter switching period fluctuates cycle by cycle, causing random errors on the duty ratio and thus jittering effect on V_O. This is difficult to correct by a feedback control loop, as the duty-ratio error changes randomly between switching cycles. Due to a limited loop-gain bandwidth, the loop response usually lags far behind. Although a ramp compensation scheme was reported to resolve this [5], the improvement is very limited, and the scheme only works for voltage-mode converters.

To mitigate the above challenges, this paper presents a continuous RSSM (C-RSSM) scheme for EMI attenuation and a one-cycle T_{ON}-rebalancing (1TR) technique for V_O jittering suppression, which are implemented in the converter of Fig. 15.7.2. The Markov-chain-based random clock implements analog f_{SW} modulation within a sideband of Δf_C centered at f_{SW0}. Hence, it conducts SSM continuously and spreads spurious noise at f_{SW0} and its harmonics uniformly, achieving desirable C-RSSM. As illustrated in the spectrum plot, although conventional D-RSSM could achieve similar peak EMI reduction (ΔEMI) with a wider modulation range ($\Delta f_D > \Delta f_C$), it elevates noise floor significantly due to spectral overlap. In the proposed Markov C-RSSM, the peak EMI and noise floor are both effectively attenuated with minimal die-area overhead. On the other hand, the proposed 1TR technique performs a two-step operation in each charge phase. As shown in Fig. 15.7.2, during t_1, the 1TR module detects the instant duty ratio D, determined by the converter PWM controller. By analyzing the sensed information of D and the instant f_{SW} from the output of Markov-chain-based random clock, the 1TR adaptively adjusts the length of t_2. Hence, D and the ON-duty T_{ON} of the converter are thus dynamically rebalanced to instantly modulated f_{SW} within one cycle, preventing V_O from jittering.

Figure 15.7.3 depicts the implementation of the Markov C-RSSM. From simulated f_{SW} modulation, f_{SW} traverses arbitrarily within the sideband Δf_C, creating a continuously and randomly distributed modulation pattern. Thus, noise spectrum at center frequency (f_{SW0}) and its harmonics can be uniformly spread. The random clock is designed using a Markov chain described by the chaotic criteria below [6]:

$$V_{ran}[n+1]=\begin{cases} kV_{ran}[n]+[(1-k)V_M+\Delta V], & V_{ran}[n]<V_M \\ kV_{ran}[n]+[(1-k)V_M-\Delta V], & V_{ran}[n]\geq V_M. \end{cases}$$

Here, V_{ran} is the random output, V_M and ΔV represent the center reference and the variation range, respectively, and k is the slope that should be well designed to avoid divergence issues. In the circuit of Fig. 15.7.3, to compute the next random state $V_{ran}[n+1]$, the present state $V_{ran}[n]$ is fed into a G_m cell, which also drives the converted current I_{gm} into a summation node Σ. Meanwhile, $V_{ran}[n]$ is compared to V_M. The comparison result multiplies the current I_1 in binary to implement the Markov partition. Finally, a fixed current I_2 is subtracted from the node Σ to generate the random current I_{ran}, leading to $V_{ran}[n+1]=I_{ran}\times R_1$. Hence, the ratio of R_1/R_2 determines the slope k. By layout matching these two resistors, k could be well defined, ensuring the convergence of the Markov transitions. Triggered by the trailing edge of V_{CMP} from the feedback loop, $V_{ran}[n+1]$ is sampled and added to V_{DC}, adjusting the VCO input V_{ctrl} and hence the clock frequency. Thanks to Markov state transitions, V_{ctrl} varies randomly within ±ΔV centered at ($V_{DC}+V_M$) and f_{SW} is modulated accordingly.

The proposed 1TR scheme and circuit are detailed in Fig. 15.7.4. T_{ON} in a conventional PWM generator is initiated by the clock and reset by trailing edge of V_{CMP}. As f_{SW} is varied by the RSSM, T_{ON} cannot be corrected immediately due to a slow feedback path, resulting in V_O jittering. To avoid this, the proposed 1TR module creates a fast f_{SW} tracking path to modulate T_{ON} synchronously with f_{SW} dithering. As the trailing edge of V_{CMP} triggers f_{SW} transition, instant switching cycle of the converter is defined between two adjacent trailing edges of V_{CMP}. Two cycles, T[n] and T[n+1] (=α×T[n]), are shown in the waveforms. T_{ON} in each cycle is split equally into t_{on1} and t_{on2} ($t_{on1}=t_{on2}$) in steady state. The operation could be explained by focusing on $t_{on2}[n]$ and $t_{on1}[n+1]$. The leading edge of the clock initiates V_{PWM} and $t_{on2}[n]$. I_L then ramps up. Meanwhile, the phase detector sets V_{UP} to high, starting to charge C_{cp} with I_{UP}. Controlled by the VCO input $V_{ctrl}[n]$, I_{UP} is inversely proportional to T[n]. The $t_{on2}[n]$ is terminated until V_{CMP} triggers low by the feedback loop and then V_{UP} is reset to stop charging C_{cp}. Thus, $t_{on2}[n]/T[n]$ (=0.5×D[n]) is converted into V_{cp}, completing D sensing. T[n+1] begins after V_{CMP} flips to low. I_L continues to rise as V_{PWM} remains high. I_{DN} is enabled by the trailing edge V_{CMP} to discharge C_{cp}. As the trailing edge V_{CMP} updates $V_{ctrl}[n]$ to $V_{ctrl}[n+1]$, I_{DN} is inversely proportional to T[n+1]. When V_{cp} hits V_{RST}, the comparator CMP1 triggers to terminate $t_{on1}[n+1]$ by resetting V_{PWM}. The charge conservation of C_{cp} makes $t_{on1}[n+1]=\alpha\times t_{on2}[n]$. Hence, D is stabilized (D[n+1]=D[n]) cycle by cycle, eliminating V_O jittering.

The prototype chip was fabricated in a 0.18μm HV CMOS process. Two enhancement mode GaN FETs are used as power switches. The converter delivers a maximum I_O of 1.5A over wide input and output ranges of 3 to 40V and 1.2 to 5V, respectively. The C-RSSM modulates f_{SW} at the center frequency of 8.3MHz with ±10% modulation range. The measured switching node V_{SW} and EMI noise spectra are shown in Fig. 15.7.5. The Markov C-RSSM reduces the peak EMI noise by 31dBμV at the fundamental frequency and 35dBμV at the 3rd-order harmonic with the 27dBμV noise floor at the mid-frequency range. To validate the effectiveness of the 1TR scheme, Fig. 15.7.6 shows the measured V_O waveforms. With conventional PWM design, the peak-to-peak magnitude of V_O jittering is up to 240mV. With the proposed 1TR scheme, it drops below 10mV. In comparison to prior arts, this work achieves the lowest noise floor of 27dBμV. Implemented on a much higher technology node, the Markov-chain-based random source consumes 53% less silicon area than its digital counterpart [3]. With both using the C-RSSM, this work achieves 119% higher peak-EMI attenuation, and 3 times higher V_O-jittering suppression than [5]. The die micrograph is shown in Fig. 15.7.7, with the active area of 1.54mm^2.

Acknowledgement:
This work is in part sponsored by Semiconductor Research Corporation (SRC) under the research contract GRC 2810.006 and National Science Foundation (NSF) under the research contract NSF CCF 1702496.

References:
[1] Y. Chen et al., "A 10MHz 5-to-40V EMI-Regulated GaN Power Driver with Closed-Loop Adaptive Miller Plateau Sensing," *IEEE Symp. VLSI Circuits*, pp. C120-C121, June 2017.
[2] K. Tse et al., "A Comparative Study of Carrier-Frequency Modulation Techniques for Conducted EMI Suppression in PWM Converters," *IEEE Trans. Industrial Electron.*, vol. 49, no. 3, pp. 618-627, Jun. 2002.
[3] W. Yang et al., "An Enhanced-Security Buck DC-DC Converter with True-Random-Number-Based Pseudo Hysteresis Controller for Internet-of-Everything (IoE) Devices," *ISSCC*, pp. 126-128, Feb. 2018.
[4] X. Ke et al., "A 10MHz 3-to-4V V_{IN} Tri-Slope Gate Driving GaN DC-DC Converter with 40.5dBμV Spurious Noise Compression and 79.3% Ringing Suppression for Automotive Applications," *ISSCC*, pp. 430-431, Feb. 2017.
[5] E. Ho et al., "Ramp Signal Generation in Voltage Mode CCM Random Switching Frequency Buck Converter for Conductive EMI Reduction," *IEEE CICC*, pp. 1-4, Sept. 2010.
[6] T. Stojanovski et al., "Chaos-Based Random Number Generators-Part I: Analysis," *IEEE TCAS-I*, vol. 48, no. 3, pp. 281-288, Mar. 2001.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

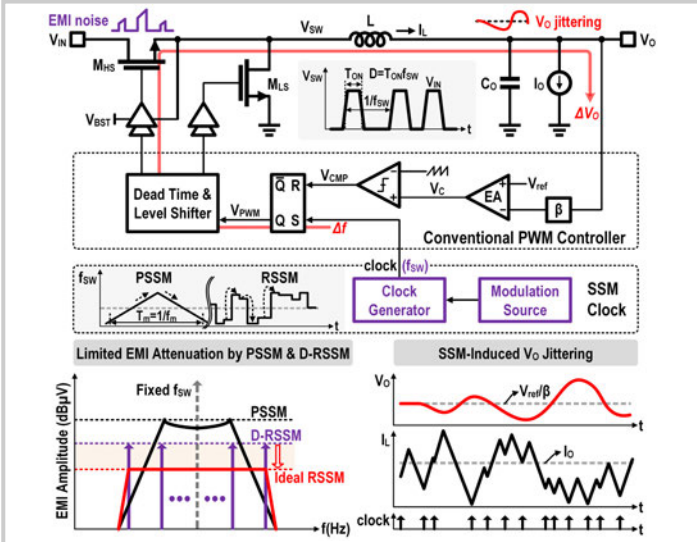


Figure 15.7.1: Challenges of implementing a spread-spectrum-modulation (SSM) scheme in power converters.

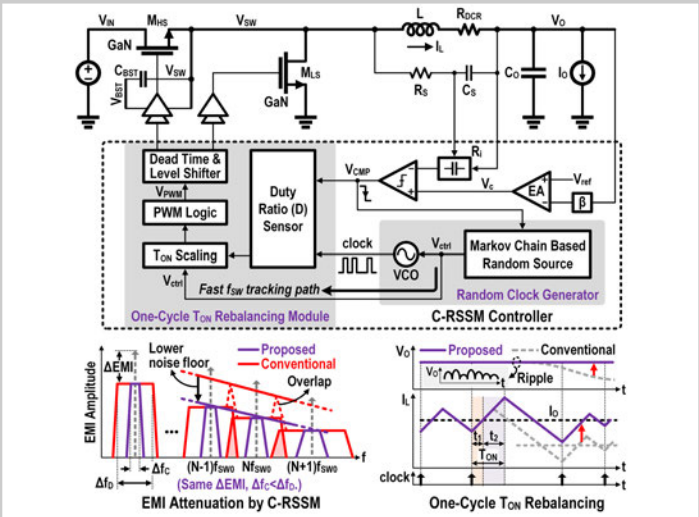


Figure 15.7.2: Proposed GaN power system with Markov continuous RSSM (C-RSSM) for EMI attenuation and one-cycle T_{ON} rebalancing to eliminate V_O jittering.

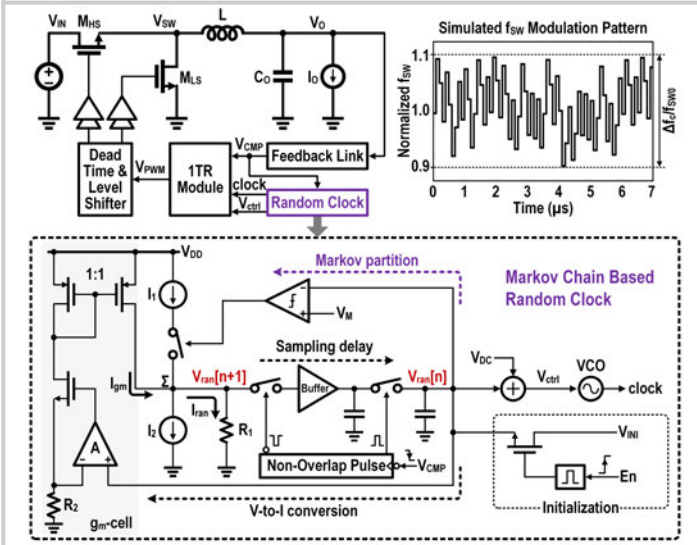


Figure 15.7.3: Implementation of the Markov continuous RSSM (C-RSSM) scheme.

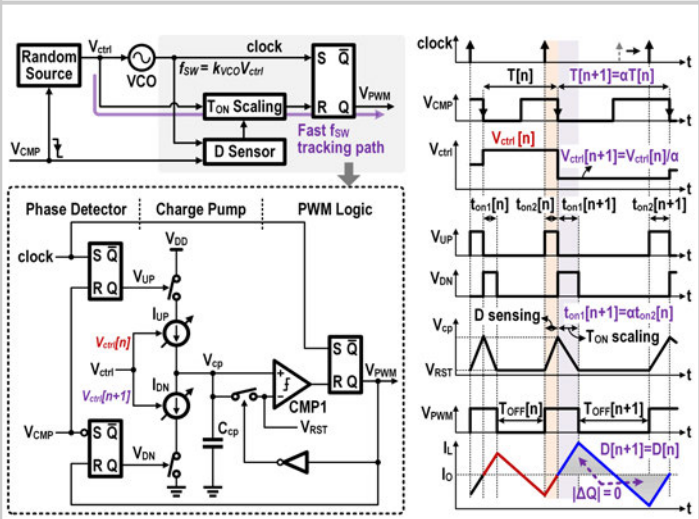


Figure 15.7.4: Scheme and circuit implementation of the proposed one-cycle T_{ON} rebalancing.

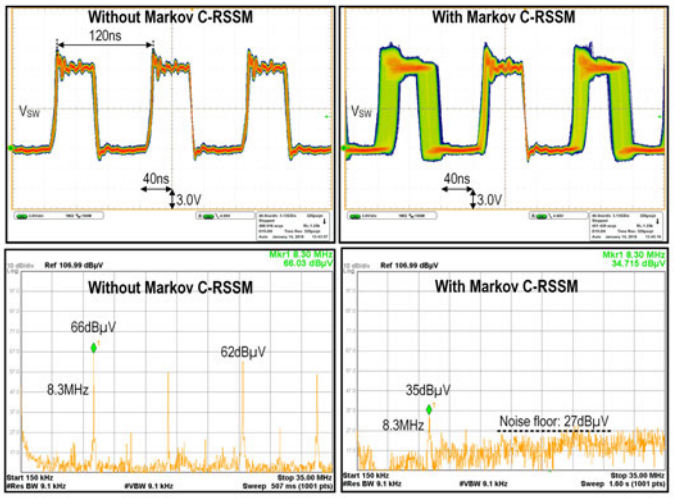
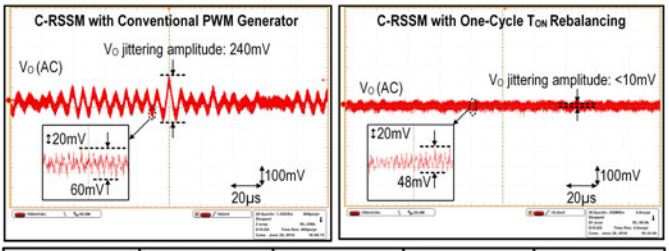


Figure 15.7.5: Measured switching-node transient waveforms and conducted-EMI noise attenuation with the C-RSSM.



Design		ISSCC2018 [3]	ISSCC 2017 [4]	CICC 2010 [5]	This Work
Process		55nm CMOS	0.35μm HV BCD	0.35μm CMOS	0.18μm HV CMOS
Power Switch		MOSFET	GaN FET	MOSFET	GaN FET
Control Mode		Hysteresis current mode	PWM voltage mode	PWM voltage mode	PWM current mode
Central f_{SW}		1MHz	10MHz	1MHz	8.3MHz
f_{SW} Modulation Way		Discrete RSSM	Triangular modulation	Continuous RSSM	Continuous RSSM
Random Source Area		0.17mm²	Not reported	Not reported	0.08mm²
f_{SW} Modulation Range		-21%/+22%	±9%	±20%	±10%
Peak EMI Attenuation		35dBμV	33dBμV	16dBμV	35dBμV
Peak Noise Floor		45dBμV	52dBμV	Not applied	27dBμV
V_O Jittering Suppression	Amplitude	Not reported	Not reported	18.1dB	>27.6dB
	Method	Pseudo hysteretic window	Not reported	Ramp compensation	1TR scheme

Figure 15.7.6: Measured V_O-jittering suppression by the one-cycle T_{ON} rebalancing and a comparison table.

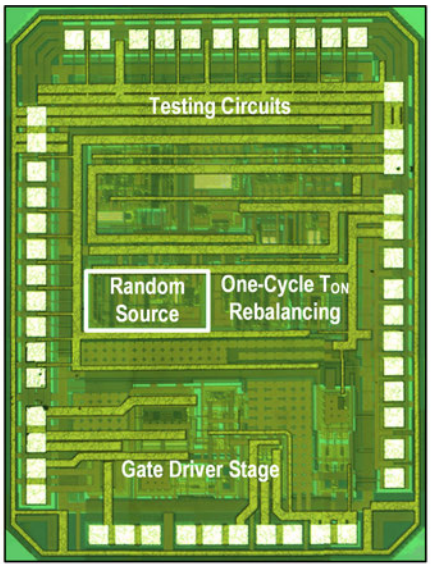


Figure 15.7.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

15.8 A 4.5V/ns Active Slew-Rate-Controlling Gate Driver with Robust Discrete-Time Feedback Technique for 600V Superjunction MOSFETs

Shusuke Kawai, Takeshi Ueno, Kohei Onizuka

Toshiba, Kawasaki, Japan

Active gate control is an emerging technique to minimize the switching loss of high-power converters facing noise-suppression challenges. In a conventional gate-driver design, a fixed value of gate resistance is chosen by the converter designers so that the slew rate (SR) of the drain voltage V_d, namely dV_d/dt, does not exceed noise-aware design guidelines in each application and use case. Minimizing the gate resistance leads to high dV_d/dt and the reduction in switching loss while shortening the turn-on delay for the overall converter performance. However, the impact is limited because of uncontrollable dV_d/dt drift caused by load-current, temperature, and V_{th} variations of the power transistors. Thus, in practice there is significant room for further loss and turn-on-delay minimization for the active gate control that adaptively modulates gate-driving ability within every switching cycle.

Several feedback techniques based on real-time SR monitoring have been studied with a view to optimizing gate-driving ability [1,2]. Although a digital-feedback approach for IGBTs controls SR maximum to around 0.0004V/ns [1], the technique is unrealistic for high-speed MOSFETs and GaN devices, whose SR usually exceeds 1V/ns [3]. Constraints with respect to the cost, area, and performance of the gate-driver process obstruct high-speed DAC and ADC implementations. A continuous-time analog-feedback controls 2V/ns IGBT SR [2] but still requires a 350MHz wide-BW op-amp, which is difficult to implement. In addition, one cannot control dV_d/dt when it is determined by the reverse-recovery current (I_{RR}) induced at the high-side diode and related dI_d/dt of the low-side device [4] because their transitions occur prior to the voltage transition. The dV_d/dt depends on the peak I_{RR} because the V_d falls in order to decrease the I_d after the I_{RR} reaches the peak. This is a fatal disadvantage for MOSFETs and other strongly I_{RR}-dependent devices. Furthermore, the turn-on delay reduction is also demanded for converter efficiency improvement.

The discrete-time background feedback proposed in Fig. 15.8.1 overcomes the challenges and satisfies all motivations summarized in (i) to (iii) in Fig. 15.8.1 simultaneously. The technique is based on the time-domain switching of two values of gate resistance, namely, R_1 and R_2, where R_2 is much smaller than R_1. R_2 is activated in the initial and the latter part of the switching cycle to boost the gate charging and shorten the turn-on delay. The duration of the intermediate high-gate-resistance state in which only R_1 feeds the gate current is controlled so that dV_d/dt reaches the maximum-allowable SR and minimizes the switching loss. dV_d/dt originating in I_{RR} can be controlled as well thanks to the background-feedback feature. Background here means that the loop samples the dV_d/dt in the previous cycle, reflecting it in duration of the high-gate-resistance state of the following cycles and achieving convergence in multiple switching cycles. In the loop, the transition period of V_d is translated into the voltage domain and sampled by the SR detector and held until the next switching cycle. The error voltage V_{err} indicates the integrated difference between the sampled voltage and the reference voltage V_{ref_dvdt}, which reflects the target dV_d/dt. The V_{err} modulates pulse width of V_2 at the next switching cycle. The dV_d/dt is gradually increased in each switching cycle and finally settles to the target value as shown at the bottom right of Fig. 15.8.1(b). The proposed SR-control (SRC) technique keeps the SR constant regardless of changes in the load current or temperature. This minimizes the switching loss, while in the original implementation without the SRC, dV_d/dt decreases and the loss increases as the load current or temperature decreases. The proposed feedback provides a short turn-on delay without violating the SR guideline thanks to the fast gate charge by R_2 and regulating I_{RR} by R_1.

The mechanism to control the SR and shorten the turn-on delay is described in more detail in Fig. 15.8.2, based on the waveform comparisons with and without the feedback. The voltage V_2, whose first pulse width is controlled by the SRC, rises at the same edge as that of V_1, which represents the original gate drive signal. At this point the gate-source capacitor of M_{PL} is mainly charged via R_2 that is smaller than R_1. The reverse recovery current (I_{RR1}) of the parasitic diode in M_{PH} and the load current I_L flow into the low-side device M_{PL} simultaneously before the drain voltage V_d starts to fall. The peak of the I_d is the summation of I_L and I_{RR1}. The V_2 falls to zero at t_2 and the gate voltage of the M_{PL} is charged via the large register R_1 and hence the speed of the I_d increment slows down. The I_d starts to decrease at t_3 and the V_d, whose SR depends on I_{RR1}, starts to decrease as well. When the falling edge of V_2 is moved from t_2 to t_1, the peak of I_d is decreased from (I_L+I_{RR1}) to (I_L+I_{RR2}). The speed of voltage transition that starts at t_4 slows down because I_{RR2} is smaller than I_{RR1}. The SR can be controlled by modulating the falling edge of V_2. The SRC feedback clearly shortens the turn-on delay thanks to the fast charge via R_2, when it is compared with the waveforms of the constant gate resistance shown on the right-hand side.

Detailed schematics of the proposed SRC-feedback driver and its waveforms are shown in Figs. 15.8.3 and 15.8.4, respectively. It consists of an SR-to-voltage-sampling (SRVS) block, an error amplifier, and a PWM block. A voltage-domain error amplifier is adopted to provide the target SR in the voltage domain. To accommodate this, the SR information is converted to the voltage domain at the SRVS block. As discussed below, an interleaving operation in the SRVS reduces the timing complexity of timing generation. The comparators CMP_1 and CMP_2 compare the divided drain voltage V_{d_fb} to the reference voltages V_{refH} and V_{refL}. The switch S1 turns-on when the divided voltage V_{d_fb} stays between the reference voltages, V_{refH} and V_{refL}, as shown in Fig. 15.8.4. The pulse width of T_p represents the duration of the drain-voltage transition and hence the SR. The T_p is converted to the voltage V_{SL1} by charging the capacitor C_1 with a current source I_1, as shown in Fig. 15.8.3. The charge $C_1 \times V_{SL1}$ is transferred to the capacitor C_{itg} at the timing of V_{SW3}. After a period of a half-switching cycle, V_{SL1} is reset.

As shown in Fig. 15.8.4, V_{S1}/V_{S2}, V_{SL1}/V_{SL2}, and V_{SW3}/V_{SW4} operate in an interleaving fashion, respectively. The interleaving operation reduces the complexity of timing generation, since all switching timing above can be generated based on V_{in}. The output voltage of the error amplifier V_{err} represents the integrated error between the target transition, indicated by reference voltage V_{ref_dvdt}, and the measured transition, indicated as V_{SL1} and V_{SL2}. The falling edge of V_2 is determined by the cross point of V_{err} and the saw-tooth waveform V_{saw}. Finally, the voltage information is converted into timing by using the comparator CMP_3. The pulse width of V_2 is increased when V_{err} is large, resulting in the increment of I_{RR} and the SR. V_2 rises again when V_{d_fb} reaches V_{ref_L} so that the gate is driven via the low-resistance R_2 after the power devices turns on.

To evaluate the continuous operation of the proposed SRC driver, it was implemented in a low-side switch of a boost converter. 600V-class superjunction MOSFETs (TK8A60W5) are utilized for both the low-side switch and the high-side diode, whose gate-source terminals are shortened. Figure 15.8.5 shows the measured transient waveforms at a low temperature of 45°C and a high temperature of 95°C. The peak of I_{RR} was kept constant by the SRC feedback against temperature change. The width of the boost-pulse V_2 is modulated to keep I_{RR} and dV_d/dt constant and the proposed SRC feedback is demonstrated. Figure 15.8.5 also shows the measured dV_d/dt with different drain currents and temperatures. As shown in the top right figure, dV_d/dt is decreased as the load current decreases without the SRC feedback. It is kept constant at 4.5V/ns or 3V/ns depending on the reference voltage of V_{ref_dvdt} by the SRC feedback. The dV_d/dt variation is compressed from 37 to 7% when the target is 4.5V/ns and from 31 to 10% when the target is 3.3V/ns. The bottom right figure shows the temperature characteristics of dV_d/dt. Thanks to the SRC feedback, dV_d/dt is kept constant. Figure 15.8.6 shows the loss-decrement ratio (1 - E_w/E_{wo}), where E_w and E_{wo} are the switching losses with/without the SRC feedback, respectively. The maximum loss reductions are 25% and 20% at I_d of 1A and temperature of 45°C, respectively. Figure 15.8.6 also shows the measured turn-on delays for different currents and temperatures. The turn-on delay t_{dl} is defined by the period between when V_{gs} reaches 10% of the maximum value and when V_{ds} reaches 10% of the maximum value. The t_{dl} reduction of 74% is achieved by the SRC feedback.

Figure 15.8.7 shows the micrograph of the die fabricated in an 18V 0.13μm HV-CMOS technology and the comparisons with the previous works.

Acknowledgement:
The authors would like to thank T. Tanaka, S. Kurushima, and W. Hatano from Toshiba Electronic Devices & Storage Corporation for their technical support.

Reference:
[1] M. Blank et al., "Digital Slew Rate and S-Shape Control for Smart Power Switches to Reduce EMI Generation," *IEEE Trans. Power Electron.*, vol. 30, no. 9, pp. 5170-5180, Sept. 2015.
[2] Y. Lobsiger and J. Kolar, "Closed-Loop di/dt and dv/dt IGBT Gate Driver," *IEEE Trans. Power Electron.*, vol. 30, no. 6, pp. 3402-3417, June 2015.
[3] L. Cong and H. Lee, "A 2MHz 150-to-400V Input Isolated DC-DC Bus Converter with Monolithic Slope-Sensing ZVS Detection Achieving 13ns Turn-On Delay and 1.6W Power Saving," *ISSCC*, pp 382-383, Feb, 2018.
[4] P. Haaf and J. Harper, "Understanding Diode Reverse Recovery and its Effect on Switching Losses," Fairchild Power Seminar, 2007. <https://www.fairchildsemi.com/technical-articles/Understanding-Diode-Reverse-Recovery-and-Its-Effect-on-Switching-Losses.pdf>
[5] K. Miyazaki et al., "General-Purpose Clocked Gate Driver IC With Programmable 63-Level Drivability to Optimize Overshoot and Energy Loss in Switching by a Simulated Annealing Algorithm," *IEEE Trans. Industrial Appl.*, vol. 53, no. 3, pp. 2350-2357, June 2017.
[6] L. Shu et al., "Active Current Source IGBT Gate Drive With Closed-Loop di/dt and dv/dt Control" *IEEE Trans. Power Electron.*, vol. 32, no. 5 pp. 3787-3796, May 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

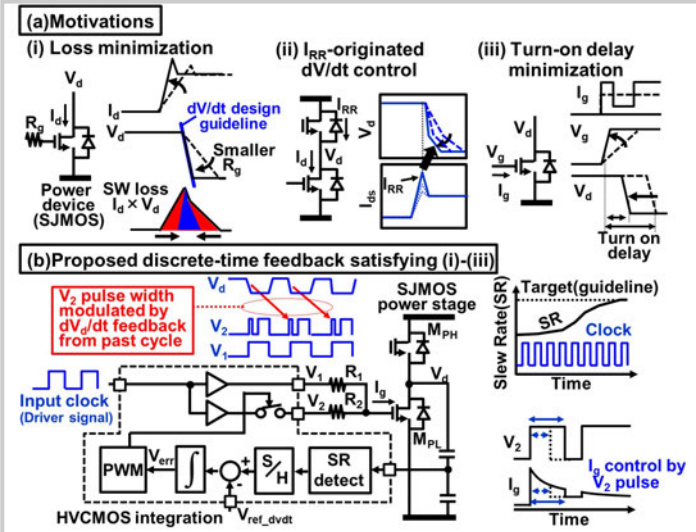


Figure 15.8.1: Conceptual diagram of the proposed slew-rate-control (SRC) feedback scheme.

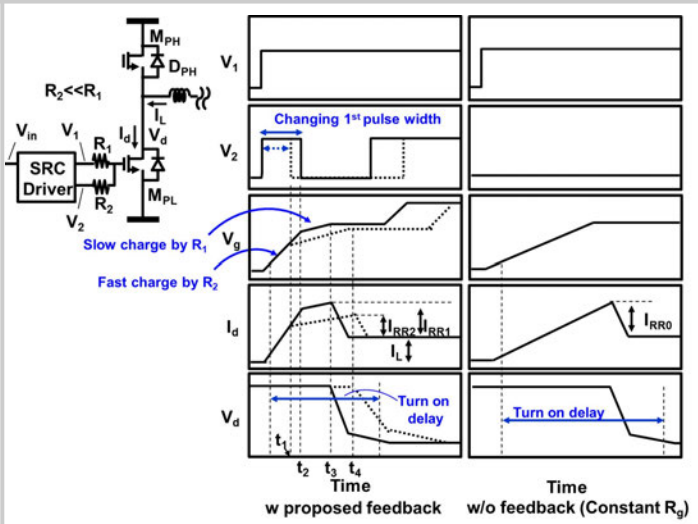


Figure 15.8.2: Time-domain waveforms and operation principle of the proposed SRC-feedback scheme.

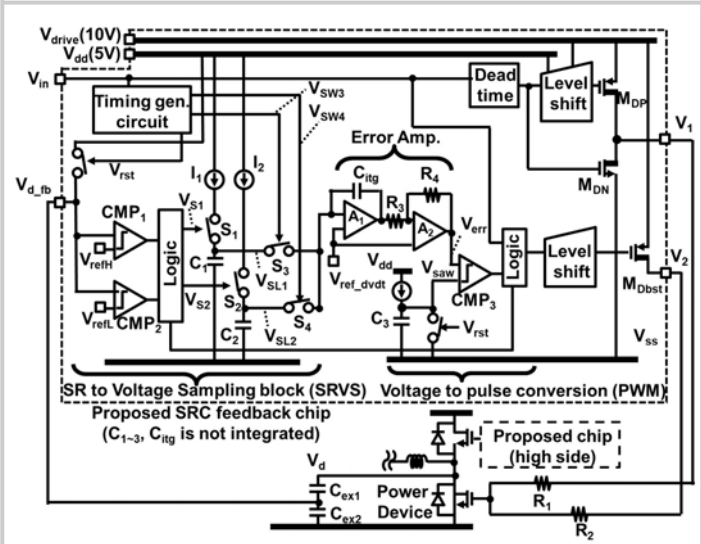


Figure 15.8.3: Circuit schematics of the proposed chip for SRC feedback.

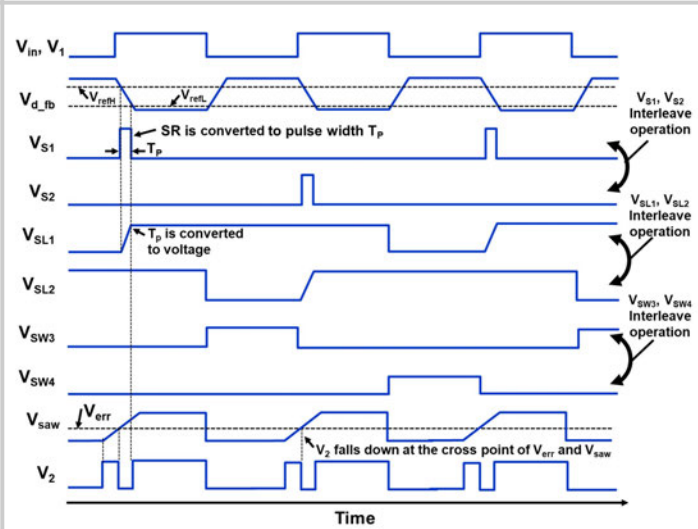


Figure 15.8.4: Transient waveform of the proposed circuit.

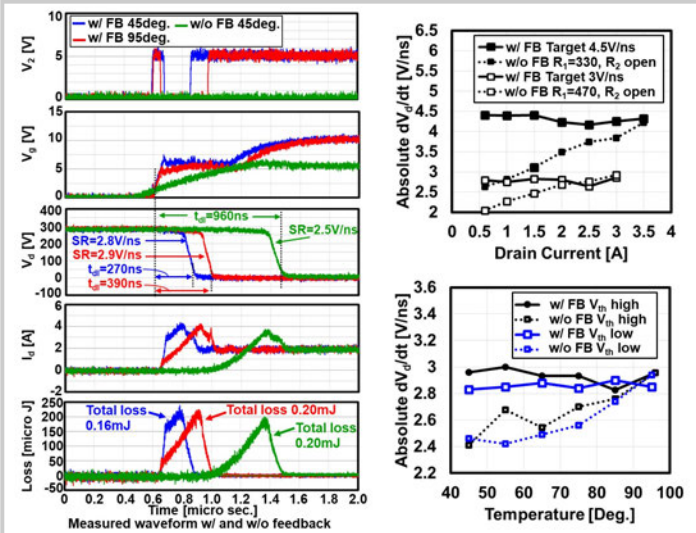


Figure 15.8.5: Measured transient waveform and dV_d/dt with the proposed feedback.

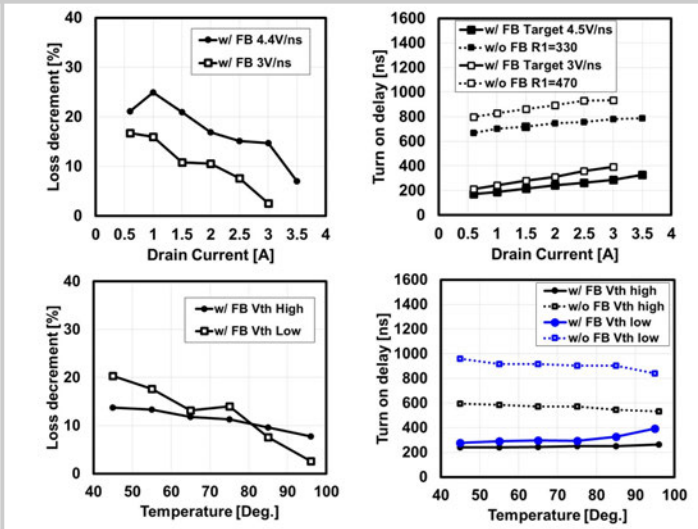
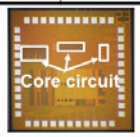


Figure 15.8.6: Measured turn-on delay and switching loss decrement under the temperature, drain current, and threshold voltage variations.

Performance comparison

	[1]	[2]	[5]	[6]	This work
Target dV_d/dt [V/ns]	0.0004	2	5	1.3	4.5
Feedback scheme	Digital	Continuous-time analog	Digital	Continuous-time analog	Discrete-time analog
Iteration	20 (estimated)	1	2400	1	20
Chip Integration	No	No	Driver only	No	Driver with feedback circuit
Process	NA	NA	0.18μm BiCD	NA	0.13μm HV-CMOS (18V)
Pro.	Precise control	Fast iteration	Precise control	Fast iteration	Easy integration
Con.	Slow operation	High-speed opamp	Large iteration number	High-speed opamp	Reverse recovery dependence



Die size
2.87mm x 2.87mm

Die micrograph

Figure 15.8.7: Die micrograph and performance comparison.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 16 Overview: *Frequency Synthesizers*

RF SUBCOMMITTEE

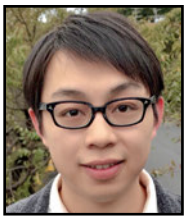
Session Chair: ***Xiang Gao***
Zhejiang University, Hangzhou, China

Associate Chair: ***Jaehyouk Choi***
Ulsan National Institute of Science Technology, Ulsan, Korea

Subcommittee Chair: Piet Wambacq, imec, Leuven, Belgium

This session presents the latest advances in digital and analog frequency synthesizers generating frequencies from 0.5 to 30GHz and covering diverse topics, such as ultra-low-power DPLLs, ultra-low-jitter sub-sampling PLLs, multi-loop BBPLL, FMCW synthesizers, and spurious tone mitigation techniques.

1:30 PM



16.1 A 265μW Fractional-N Digital PLL with Seamless Automatic Switching Subsampling/Sampling Feedback Path and Duty-Cycled Frequency-Locked Loop in 65nm CMOS

H. Liu, Tokyo Institute of Technology, Tokyo, Japan

In Paper 16.1, the Tokyo Institute of Technology describes a fractional-N DPLL with a mechanism to automatically switch between the subsampling and sampling modes. Implemented in 65nm CMOS, it achieved an ultra-low power of 265μW with a -237dB FoM and a -52dBc worst-case fractional spur.

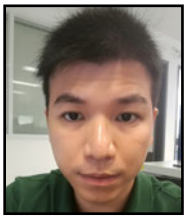
2:00 PM

16.2 A 76fs$_{rms}$ Jitter and –40dBc Integrated-Phase-Noise 28-to-31GHz Frequency Synthesizer Based on Digital Sub-Sampling PLL Using Optimally Spaced Voltage Comparators and Background Loop-Gain Optimization

J. Kim, Ulsan National Institute of Science and Technology, Ulsan, Korea

In Paper 16.2, UNIST presents a 28-to-31GHz frequency synthesizer cascading subsampling PLL and injection-locking frequency multiplier. This design achieved an ultra-low jitter of 76fs$_{RMS}$ in 65nm CMOS.

2:15 PM



16.3 A -246dB Jitter-FoM 2.4GHz Calibration-Free Ring-Oscillator PLL Achieving 9% Jitter Variation Over PVT

X. Yang, University of Macau, Macau, China

In Paper 16.3, the University of Macau demonstrates a calibration-free PLL with discrete-time phase noise cancellation technique. The 2.4GHz-PLL design consumed 4.1mW while achieving -63dBc reference spur and 248fs$_{RMS}$ integrated jitter.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

2:30 PM

16.4 A 0.5-to-2.5GHz Multi-Output Fractional Frequency Synthesizer with 90fs Jitter and –106dBc Spurious Tones Based on Digital Spur Cancellation

S-Y. Hung, University of California, Los Angeles, CA

In Paper 16.4, UCLA describes an open loop frequency synthesizer with a digital spur cancellation technique. Designed with 28nm CMOS, it achieved an integrated rms jitter <90fs and spurious tones <–106dBc at a 2.48GHz carrier frequency.

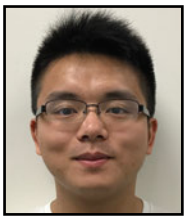
3:15 PM

16.5 A Fractional-N Synthesizer with 110fs$_{rms}$ Jitter and a Reference Quadrupler for Wideband 802.11ax

B. Razavi, University of California, Los Angeles, CA

In Paper 16.5, UCLA presents a 28nm-CMOS fractional-N synthesizer for 802.11ax utilizing a background-calibrated reference quadrupler. It achieved a jitter of 110fs$_{RMS}$, while consuming 22.8mW.

3:30 PM



16.6 A Calibration-Free Triple-Loop Bang-Bang PLL Achieving 131fs$_{rms}$ Jitter and -70dBc Fractional Spurs

D. Yang, University of California, Los Angeles, CA

In Paper 16.6, UCLA demonstrates a heterodyne fractional-N frequency synthesizer consisting of three loops: classical CP PLL, sampling PLL, and BBPLL. It achieved 131fs$_{RMS}$ jitter and -70dBc fractional spurs while consuming 13.4mW.

4:00 PM

16.7 A 30GHz Digital Sub-Sampling Fractional-N PLL with 198fs$_{rms}$ Jitter in 65nm LP CMOS

L. Grimaldi, Politecnico di Milano, Milan, Italy

In Paper 16.7, Politecnico di Milano describes a 30.6-to-34.2GHz fractional-N DTC-based PLL with sub-sampling bang-bang phase detection and a phase switching technique that halves the required DTC range. Fabricated in a 65nm LP process, it achieved 198fs$_{RMS}$ jitter at 35mW power.

4:30 PM

16.8 A 25.4-to-29.5GHz 10.2mW Isolated Sub-Sampling PLL Achieving -252.9dB Jitter-Power FoM and -63dBc Reference Spur

Z. Yang, University of Macau, Macau, China

In Paper 16.8, the University of Macau presents a sub-sampling PLL with an improved sub-sampling phase detector for better VCO isolation and low reference spur. Designed in 65nm CMOS, it achieved 71.16fs$_{RMS}$ jitter with 10.2mW and a PLL FoM of -252.9dB at 26.4GHz.

4:45 PM

16.9 4.48GHz 0.18µm SiGe BiCMOS Exact-Frequency Fractional-N Frequency Synthesizer with Spurious-Tone Suppression Yielding a -80dBc In-Band Fractional Spur

M. P. Kennedy, University College Dublin, Dublin, Ireland; Microelectronic Circuits Centre Ireland, Dublin, Ireland

In Paper 16.9, University College Dublin demonstrates a fractional-N frequency synthesizer utilizing a divider controller with a nested MASH digital delta-sigma modulator and a probability mass redistributor for low fractional spur. Implemented in 0.18µm SiGe BiCMOS, it achieves a -80dBc in-band fractional spur at 2.75kHz offset at 4.48GHz output.

16.1 A 265μW Fractional-N Digital PLL with Seamless Automatic Switching Subsampling/Sampling Feedback Path and Duty-Cycled Frequency-Locked Loop in 65nm CMOS

Hanli Liu, Zheng Sun, Hongye Huang, Wei Deng, Teerachot Siriburanon, Jian Pang, Yun Wang, Rui Wu, Teruki Someya, Atsushi Shirane, Kenichi Okada

Tokyo Institute of Technology, Tokyo, Japan

The demand for highly energy-efficient circuits and systems has exponentially increased for Systems on Chip (SoC). A fractional-N phase-locked loop (PLL) is one of the most important building blocks in SoCs for a variety of applications, such as frequency synthesis for wireless transceivers and system clock generation for processors, memories, and I/O interfaces. Recent developments in fractional-N digital PLLs (DPLLs) [1-3] have shown great potential for achieving low-power operation and small chip area. However, none of these works have achieved power consumption below 500μW due to the number of building blocks operating at the oscillator frequency. Furthermore, the digitally controlled oscillators (DCOs) in [1-3] consume more than 250μW of power to achieve a good phase noise and a high-enough amplitude for DPLL locking. A digital sub-sampling architecture [1,2,4] can potentially reduce the overall power consumption by bypassing these high-frequency building blocks. Unfortunately, the absence of frequency acquisition makes such architecture vulnerable to sudden or large frequency disturbances. Even though a background frequency-locked loop (FLL) [1,4] can be applied, it consumes large power due to the counter working at the DCO frequency. The typical solution to save power consumption is to turn off the FLL [2] after the PLL has been stabilized. Despite the benefit of the power reduction, a sub-sampling PLL has multiple frequency lock-in ranges near the integer multiple of the reference frequency, which could cause false locking if the frequency disturbances are within those ranges. To address the above issues, this work presents a fractional-N DPLL achieving a 265μW power consumption with robust phase and frequency acquisition with negligible power overhead in a 65nm CMOS technology. It also achieves an rms jitter of 2.8ps, which corresponds to an FoM of -236.8dB.

Figure 16.1.1 shows the overall DPLL architecture. To detect the phase/frequency disturbances, an out-of-dead-zone (ODZ) detector and a duty-cycled FLL (DC-FLL) are utilized and operate in the background. Most of the phase/frequency disturbances can be detected by the ODZ detection, if the phase error is larger than the DZ, and the state machine sets the EN signal to 1, which switches the feedback (FB) path to the sampling-PLL mode. After the phase is locked, the state machine produces the EN signal of 0 to switch the FB path to the sub-sampling mode for saving power from the power-hungry multi-modulus divider (MMD). Thanks to the delay unit after the DTC, the DPLL can remain locked when switching from the sampling mode to the sub-sampling mode. In the worst case, the frequency of the sub-sampling PLL may be falsely locked to the integer multiples of the reference frequency while the phase is still locked. Hence, the DZ detector cannot detect any phase errors. A duty-cycled FLL (DC-FLL) is used for assisting with the detection of this frequency error and sets EN to 1 to switch the PLL to the sampling mode for relocking. The power consumption can be improved from 415μW to 265μW with the proposed switching FB path, which corresponds to a 56% power reduction. A delta-sigma-modulator (DSM) based controller is used to generate the delayed sampling clock REFA by using a digital-to-time converter (DTC) for realizing the fractional-N operation. The phase error between CKV and REFB is quantized by a 4b TDC. Most of the building blocks are working at the 10MHz reference frequency or less. After the phase is locked, only the DCO, the DCO buffer, and the sampler are working at the DCO frequency, while the MMD and the coarse TDC are completely turned off. The high-frequency counter inside the DC-FLL operates with a duty cycle of 0.5%. This limits the power consumption of the FLL to only 3μW.

Figure 16.1.2 shows the circuit operation. During the phase/frequency locking, EN is 1 and the PLL is in the sampling mode. The DIV signal is selected at the sampler input where the output synchronizes to the CKV edge with a delay of τ_{SAM}. If the phase is locked, the FB signal is aligned with REFB. The EN automatically changes to 0, and the PLL enters the low-power sub-sampling mode, after the phase/frequency is locked. Now, the sampler samples the REFA by CKV. Due to the delay of τ_{FB}, the REFA always leads the REFB. Since REFB is aligned with the CKV edge, the sampled REFA by CKV is also aligned with REFB. As such, the phase-locked state remains uninterrupted and no meta-stability issue of the sampler occurs due to the switching operations. Note that τ_{FB} should be larger than τ_{SAM} and smaller than one CKV period. This could be well ensured by post-layout simulations. If a large frequency/phase disturbance appears, it introduces a large phase error (larger than the τ_{DZ}) between REFB and FB. This automatically sets EN to 1, and the MMD is turned on to assist with relocking. The DC-FLL is always operating in background, which doubly ensures the phase/frequency locking with a negligible power overhead.

For a low-power oscillator design, a small bias current I_{BIAS} is usually used to limit the power consumption of the DCO, which forces the oscillator to always work in the current limited region. The CMOS-type DCO [1-3] is advantageous for obtaining a larger output amplitude, V_{AMP}. However, V_{AMP} is mainly limited by the maximum available tank impedance R_{tank} and the required frequency-tuning range. Figure 16.1.3 shows the proposed DCO that minimizes power consumption while obtaining sufficient output amplitude. Two cross-coupled NMOS pairs share the same bias current through two center taps of the wings in the transformer. A large inductance L_S ensures the startup requirement of this DCO. The coupling between two oscillators through the transformer corresponds to a positive feedback loop for both cross-coupled oscillators. A coplanar transformer with minimum number of the cross sections is implemented to minimize the quality-factor degradation. In the presented design, V_{AMP} is enhanced by the transformer operation, which produces a gain of $G_{TF} \approx 4.35$. Hence, comparing with the conventional CMOS-type oscillator, a small I_{BIAS} can sustain the oscillation with enough amplitude. A frequency tuning range from 2.05 to 3.10GHz is also available by using a 11b coarse capacitor bank.

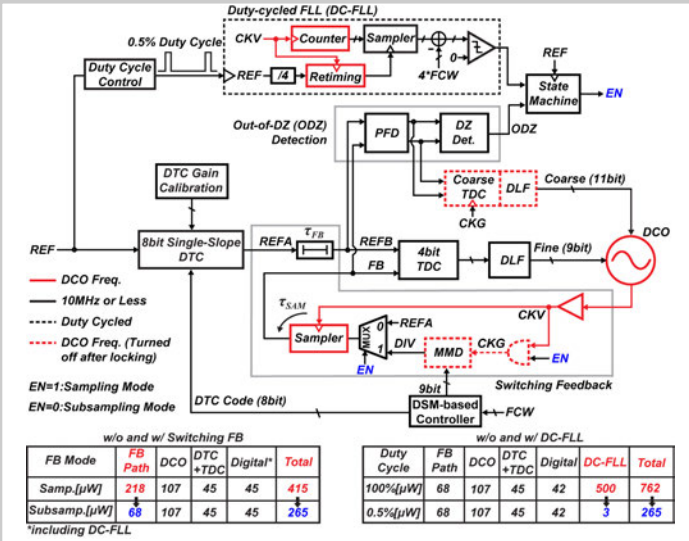
Figure 16.1.4 shows the details of the proposed low-power single-slope DTC based on the concept of the constant-slope DTC [5]. As widely known, the nonlinearity of the DTC significantly degrades the fractional-spur performance of DPLLs. The constant-slope DTC [5] achieves a fundamentally better linearity, while the power consumption is high from the current source and the digital-to-analog converter (DAC). Variable DTC delays could be achieved by varying the initial voltage on the capacitor V_P, which is pre-charged by the DAC. When input signal (IN) comes, SW3 is turned on for enabling the inverter and SW2 is turned off for stopping pre-charging C. After τ_1 delay, the current source is triggered by SW1 and starts charging C from V_{DAC1}. When the slope voltage reaches V_{TH} of the inverter, a rising edge is generated at the output (OUT). Different from the conventional operation, which keeps charging even after the comparison, the proposed DTC cuts off the current source when the comparison finished by using a feedback control, which minimizes the power consumption from the current source. By further holding τ_2 delay, the SW3 is then turned off to avoid the large leakage current of the inverter caused by $V_P \approx V_{TH}$, and the SW2 is turned on to let P to discharge to the new DAC voltage of V_{DAC2} for the next cycle of the DTC operation. The proposed DTC removes the unnecessary power consumed in the conventional architecture, which contributes to a 58% power reduction for the current source. For achieving a peak INL of 300fs and a resolution of 1.9ps/LSB, the proposed DTC consumes 20μW in the post-layout simulations. A 9b current bank is used for the DTC gain calibration.

The presented fractional-N DPLL occupies an area of 0.25mm^2 in 65nm CMOS, as shown in Fig. 16.1.7. Figure 16.1.5 shows the measurement results of the DPLL. The free-running DCO achieves a phase noise of -107dBc/Hz at a 1MHz offset frequency with a power consumption of 107μW at a 0.45V supply. The rest of the DPLL works at 0.85V. The automatic-switching FB path consumes a power of 68μW, which leads to a total power consumption of 265μW for the whole DPLL. An integrated jitter of 2.8ps is achieved at an FCW of 246.6. The measured worst-case in-band fractional spur is -52dBc. Figure 16.1.6 compares the prior-art low-power fractional-N PLLs. The presented work breaks the 500μW power consumption by almost twice with a PLL FoM of -236.8dB.

Acknowledgment:
This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References:
[1] V. Chillara et al., "An 860μW 2.1-to-2.7GHz All-Digital PLL-Based Frequency Modulator with a DTC-Assisted Snapshot TDC for WPAN (Bluetooth Smart and ZigBee) Applications," *ISSCC*, pp. 172-173, Feb. 2014.
[2] Y.-M. He et al., "A 673μW 1.8-to-2.5GHz Dividerless Fractional-N Digital PLL with an Inherent Frequency-Capture Capability and a Phase-Dithering Spur Mitigation for IoT Applications," *ISSCC*, pp. 420-421, Feb. 2017.
[3] H. Liu et al., "A 0.98mW Fractional-N ADPLL using 10b Isolated Constant-Slope DTC with FOM of –246dB for IoT Applications in 65nm CMOS," *ISSCC*, pp. 246-248, Feb. 2018.
[4] T. Siriburanon et al., "A 2.2GHz –242dB-FOM 4.2mW ADC-PLL using Digital Sub-sampling Architecture," *ISSCC*, pp. 440-441, Feb. 2015
[5] J. Ru et al., "A High-Linearity Digital-to-Time Converter Technique: Constant-Slope Charging," *IEEE JSSC*, vol. 50, no.6, June 2015.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE



w/o and w/ Switching FB

FB Mode	FB Path	DCO	DTC +TDC	Digital*	Total
Samp.[μW]	218	107	45	45	415
Subsamp.[μW]	68	107	45	45	265

*including DC-FLL

w/o and w/ DC-FLL

Duty Cycle	FB Path	DCO	DTC +TDC	Digital	DC-FLL	Total
100%[μW]	68	107	45	42	500	762
0.5%[μW]	68	107	45	42	3	265

Figure 16.1.1: Proposed fractional-N DPLL with a power consumption of 265μW.

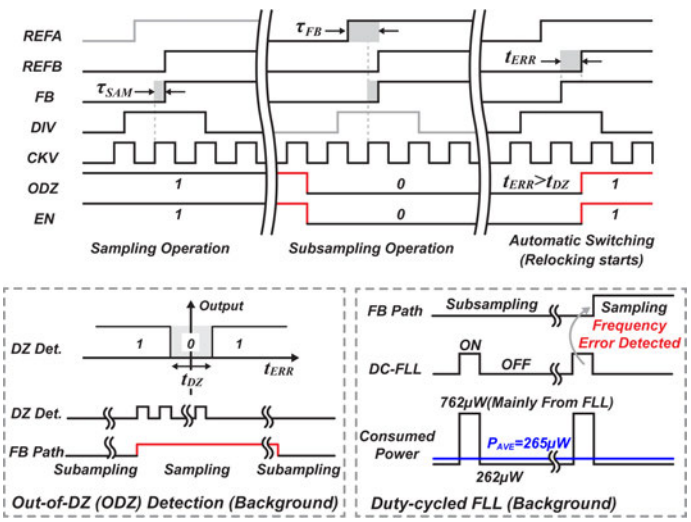
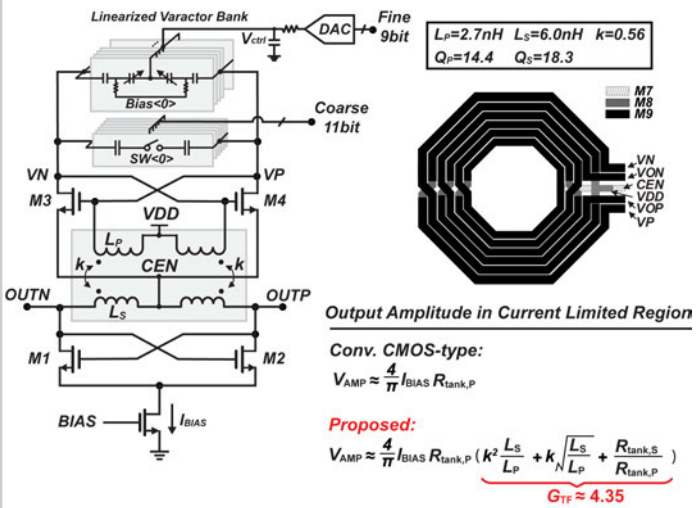


Figure 16.1.2: Operations of switching feedback path and duty-cycled FLL.



L_P=2.7nH L_S=6.0nH k=0.56
Q_P=14.4 Q_S=18.3

$V_{AMP} \approx \frac{4}{\pi} I_{BIAS} R_{tank,P}$

$V_{AMP} \approx \frac{4}{\pi} I_{BIAS} R_{tank,P} \left(k^2 \frac{L_S}{L_P} + k\sqrt{\frac{L_S}{L_P}} + \frac{R_{tank,S}}{R_{tank,P}}\right)$

$G_{TF} \approx 4.35$

Figure 16.1.3: 107μW DCO with ultra-low-power operation.

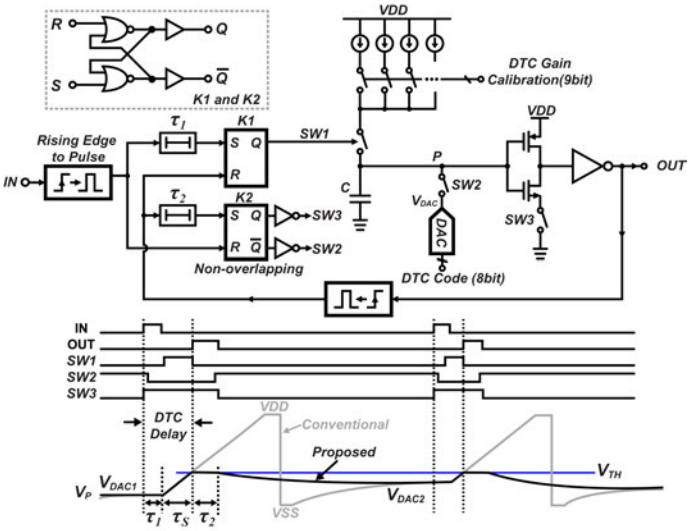
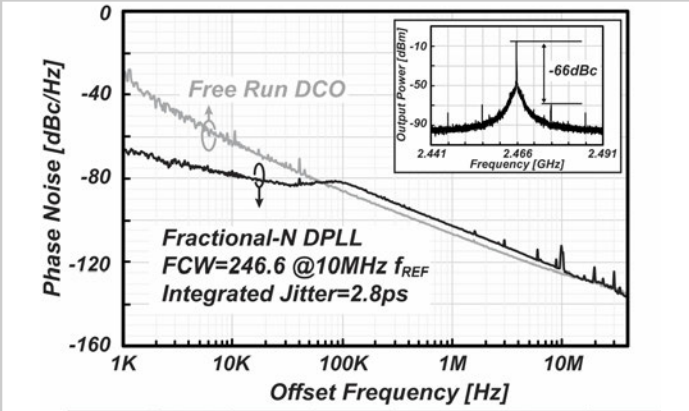


Figure 16.1.4: 8b low-power single-slope DTC with current-source gating and charge-reuse operations.



	DCO	DTC	TDC	FB Path*	Digital	Total
Power [μW]	107	23	22	68	45	265

*Including DCO Buffer

Figure 16.1.5: Measured phase noise and power breakdown.

	This Work	ISSCC'17[1]	ISSCC'14[2]	ISSCC'18[3]
Technology	65nm	40nm	40nm	65nm
Ref Freq (MHz)	10	N/A	32	52
Out Freq (GHz)	2.05 - 3.10	1.8 - 2.5	2.1 - 2.7	2.0 - 2.8
Power (μW)	**265**	673	860	980
Jitter (ps)	2.8 [10k-40M]	1.98 [10k-10M]	1.71 [1k-100M]	0.53 [10k-10M]
FOM (dB)	**-236.8**	-235.8	-236.0	-245.6
In-Band Frac Spur (dBc)	-52	-56	-38	-56
Power Eff. (μW/GHz)	107	312	358	271
Reference Spur (dBc)	-66	-62	-70	-68
Active Area (mm²)	0.25	0.18	0.20	0.23

Figure 16.1.6: Comparison table with prior-art fractional-N DPLLs.

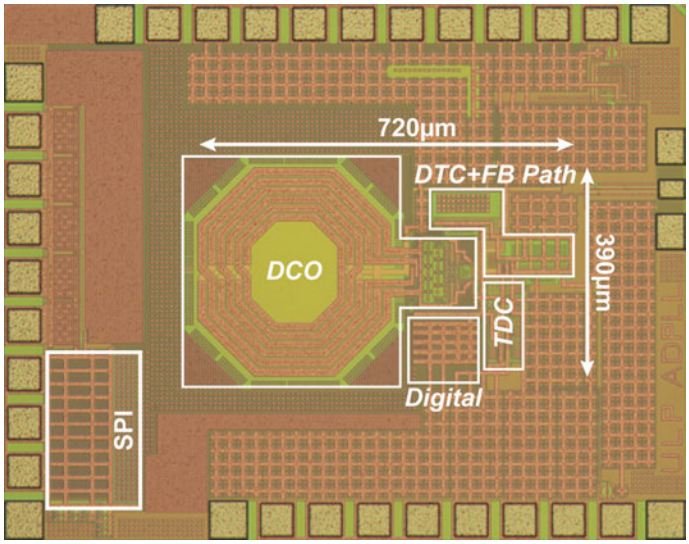


Figure 16.1.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

16.2 A 76fs$_{rms}$ Jitter and –40dBc Integrated-Phase-Noise 28-to-31GHz Frequency Synthesizer Based on Digital Sub-Sampling PLL Using Optimally Spaced Voltage Comparators and Background Loop-Gain Optimization

Juyeop Kim*, Heein Yoon*, Younghyun Lim*, Yongsun Lee, Yoonseo Cho, Taeho Seong, Jaehyouk Choi

Ulsan National Institute of Science and Technology, Ulsan, Korea
*Equally-Credited Authors (ECAs)

The generation of mm-wave (mmW) signals that have ultra-low phase noise (PN) is very important for the design of RF transceivers (TRXs) for high-data-rate 5G systems. Direct-RF-sampling TRXs also require high-frequency clock signals, having extremely low integrated PN (IPN) [1]. To satisfy such stringent noise requirements, the rms jitter of mmW-band signals must be reduced to sub-100fs. Recently, a charge-pump (CP) PLL in [1] achieved a very low rms jitter of less than 60fs at 14GHz. However, to suppress the in-band PN of PLL building blocks, that design used a reference clock that had an impractically high frequency, f_{REF}, of 500MHz. To avoid the use of such a high f_{REF} while minimizing in-band PN, sub-sampling PLLs (SSPLLs) are seen as a promising solution. However, conventional SSPLLs are not suitable for generating mmW-band signals directly, since, as the frequency increases, the capture range of their sampling operation is reduced rapidly, thereby hindering the reliable operation. To extend the capture range, a prescaler can be used after the VCO [2], but it increases the in-band PN and power consumption. Direct-mmW SSPLLs are limited even at suppressing out-of-band PN, since their PN skirt is determined by an mmW VCO that has a relatively low Q. To overcome the problems of analog SSPLLs, such as a large area and a PVT-sensitive loop gain, digital SSPLLs using ADCs to digitize the sampled voltage have been developed recently [3]. However, digital SSPLLs suffer from another problem in that, to reduce the quantization noise (Q-noise) and improve the overall IPN, they must use high-performance ADCs that concurrently have high-sampling frequencies, fine resolutions, and wide dynamic ranges. Thus, they demand high power and occupy larger area.

This work presents a digital SSPLL-based 28-to-31GHz frequency synthesizer (FS) that can achieve 76fs$_{rms}$ jitter and –40dBc IPN using a 100MHz f_{REF}. As shown at the top of Fig. 16.2.1, the digital SSPLL of the first stage operates at a GHz range so that it can obtain a wide capture range. For overcoming the Q-noise problem of the conventional digital SSPLLs, this work presents the optimally spaced voltage comparators (OSVC). To minimize the Q-noise, the proposed OSVC require only three simple 1b voltage comparators (VCs) instead of a high-resolution multi-bit ADC; thus, it can save significant power and silicon area. Since the subsequent injection-locked frequency multiplier (ILFM) has a bandwidth greater than 200MHz, the out-of-band PN of the mmW-band signal also is determined by a high-Q VCO of the GHz-range SSPLL. Consequently, this work can generate mmW-band output signals having ultra-low IPN and rms jitter.

The Q-noise is a critical problem also for the design of TDC-based digital PLLs. To reduce expenditures in power and area, a recent solution is to use a BBPD instead of a complex TDC and to optimize the loop gain of the PLL by controlling the error-correction gain, K, in the background [4]. Although this method is efficient in terms of power and area, it has a limited ability to minimize Q-noise, since the BBPD can obtain only the binary information of phase errors. To address this limit, [5] presented a method to use three parallel BBPDs and to optimize the spacing between the time thresholds of the BBPDs as well as K. This work found that this method of reducing the Q-noise of digital PLLs in the time domain can be applied for digital SSPLLs in the voltage domain. The bottom left of Fig. 16.2.1 shows that, by optimizing K in the background, digital SSPLLs using a 1b VC can reduce the Q-noise effectively, and, thus, achieve a low rms jitter despite PVT. It also shows that by optimizing V_{TH} and K together, digital SSPLLs with the OSVC using three 1b VCs can further reduce the rms jitter. The bottom right of Fig. 16.2.1 shows the conceptual diagram of an OSVC-based digital SSPLL. The level of S_{BUF} is sampled as V_{SH} by a sample-and-hold (SH) circuit. Then, V_{SH} is compared with V_{REF} by the three VCs having different input offsets, i.e. $+V_{TH}$, 0, and $-V_{TH}$, which had been inserted intentionally, so that the voltage error, V_{ERR}, can be represented more accurately as one of four values of D_{VC}. Finally, co-optimizing V_{TH} and K, the OSVC can remove instantaneous phase errors more precisely and minimize the Q-noise. Since ΔΣDACs can be used to calibrate V_{TH}, the concept of the OSVC is even more effective in the voltage domain.

Figure 16.2.2 shows the overall architecture of the mmW-band FS consisting of the digital SSPLL and the mmW ILFM. Since the mmW ILFM includes a divide-by-2 divider to generate injection pulses [6], the effective multiplication factor becomes 7.5. In the digital SSPLL, the differences of the differentially sampled voltages, $V_{SH,P}$ and $V_{SH,N}$, are quantized by the three VCs, i.e. VC_H, VC_M, and VC_L, generating D_H, D_M, and D_L, respectively. By adding the offset voltages, i.e. V_{TH+} and V_{TH-}, to the positive input of VC_H and the negative input of VC_L, respectively, the OSVC have four decision values in D_{VC}. To optimize V_{TH+} and V_{TH-} even in the presence of the intrinsic input offsets of the VCs, their values are adjusted separately using the three voltages from the V_{TH}-controller, V_H, V_M, and V_L. The V_{TH}-controller adjusts V_H and V_L by comparing the accumulated values of D_H and D_L with the optimal values, N_{TH+} and N_{TH-}, calculated as in [5]. Using a ΔΣDAC and the following lowpass filters, the values of V_H and V_L can be calibrated with a very high resolution, i.e. effectively 10fs. The value of V_M is fixed at the middle voltage of the DAC. As long as the output of the ΔΣDAC is monotonic, any effects of the intrinsic offsets of the VCs and the nonlinearity of the ΔΣDAC can be removed, since V_{TH+} and V_{TH-} are optimized continuously with a sufficiently fine resolution. The digital loop filter is designed such that the proportional (P) and the integral (I) paths control the VCO separately to avoid the use of any retiming circuits that could cause latency and degrade jitter. By ensuring zero autocorrelation of D_M [4], the loop-gain optimizer can keep adjusting the P-path gain, K_P, to be optimum. In the worst case, 750μs is required for the convergence of K_P. The LC VCO is based on a NMOS-type cross-coupled topology with a tail inductor for suppressing flicker noise.

Figure 16.2.3 shows that a cycle of the operation of the OSVC-based SSPLL starts with the sampling of $V_{SH,P}$ and $V_{SH,N}$ at the falling edge of S_{REF}. In *'V_{TH} update'* phase during ϕ_1, the V_{TH}-generator redefines V_{TH+} (or V_{TH-}) by charging C_{TH+} (or C_{TH-}) using the updated V_H and V_M (or V_M and V_L). In the next phase during ϕ_2, the connections of C_{TH+} (or C_{TH-}) return to $V_{SH,P}$ (or $V_{SH,N}$) and VC_H (or VC_L), thereby adding V_{TH+} (or V_{TH-}) to the input of VC_H (or VC_L). At the rising edge of ϕ_3, the three VCs make decisions and generate D_{VC}. In the design of the V_{TH}-generator, the sizes of the switches and the input transistors of the VCs are minimized to alleviate the charge-sharing effect, which would occur due to parasitic capacitance, thereby reducing the gain of the SH. The VCs were implemented using a double-tail regenerative topology.

The mmW FS of this work uses 42mW of power and 0.32mm^2 of area in 65nm CMOS. Figure 16.2.4 shows that, due to the SS operation and the Q-noise minimization, the proposed digital SSPLL achieved very low in-band PN, i.e. –115 dBc/Hz at a 10kHz offset (f_{REF} = 100MHz). Thus, both at 3.8 and 3.9GHz, it achieved ultra-low IPN and rms jitter, less than –58dBc and 72fs, respectively. Figure 16.2.5 shows that the mmW FS in this work achieved –40.3dBc IPN and 76fs$_{rms}$ jitter at 28.5GHz. It also shows that the PN of the FS output at 28.5GHz follows that of the SSPLL output at 3.8GHz with a constant gap of 17.5dB. In the first table of Fig. 16.2.6, the proposed digital SSPLL achieved the lowest rms jitter among prior-art SSPLLs. The second table shows that the mmW FS of this work achieved the best $FoM_{JIT,N}$ (normalizing FoM_{JIT} to f_{REF}) among prior-art mmW FSs.

Acknowledgement:
This work was supported by Samsung Research Funding Center of Samsung Electronics under Project Number SRFC IT1702-02.

References:
[1] D. Turker et al., "A 7.4-to-14GHz PLL with 54fs$_{rms}$ Jitter in 16nm FinFET for integrated RF-Data-Converter SoCs," *ISSCC*, pp. 378–380, Feb. 2018.
[2] V. Szortyka et al., "A 42mW 230fs-Jitter Sub-sampling 60GHz PLL in 40nm CMOS," *ISSCC*, pp. 366–368, Feb. 2014.
[3] Z. Chen et al., "A Sub-Sampling All-Digital Fractional-N Frequency Synthesizer with –111dBc/Hz In-Band Phase Noise and an FOM of –242dB," *ISSCC*, pp. 268–270, Feb. 2015.
[4] T. Kuan and S.I. Liu, "A Bang-Bang Phase-Locked Loop Using Automatic Loop Gain Control and Loop Latency Reduction Techniques," *IEEE JSSC*, vol. 51, no. 4, pp. 821–831, Apr. 2016.
[5] T. Seong et al., "A –242dB FOM and –75dBc-reference-spur ring-DCO-based all-digital PLL using a fast phase-error correction technique and a low-power optimal-threshold TDC," *ISSCC*, pp. 396–398, Feb. 2018.
[6] H. Yoon et al., "A –31dBc integrated-phase-noise 29GHz fractional-N frequency synthesizer supporting multiple frequency bands for backward-compatible 5G using a frequency doubler and injection-locked frequency multipliers," *ISSCC*, pp. 366–368, Feb. 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

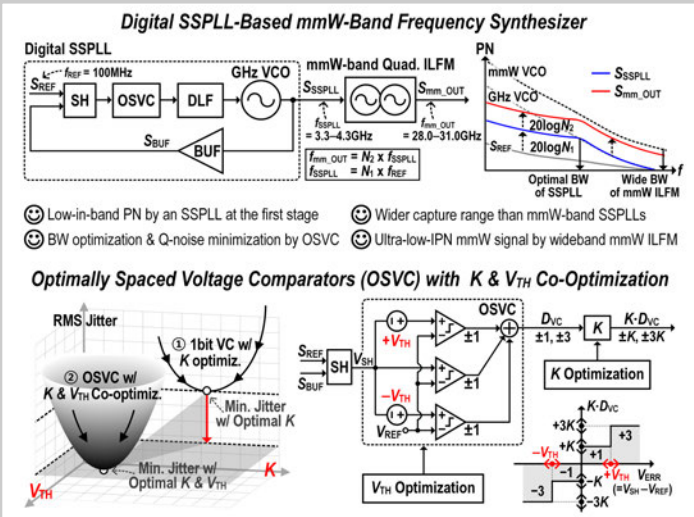


Figure 16.2.1: mmW-band frequency synthesizer based on a digital SSPLL using an OSVC (top), and the concept of an OSVC-based digital SSPLL with V_{TH} and K co-optimization (bottom).

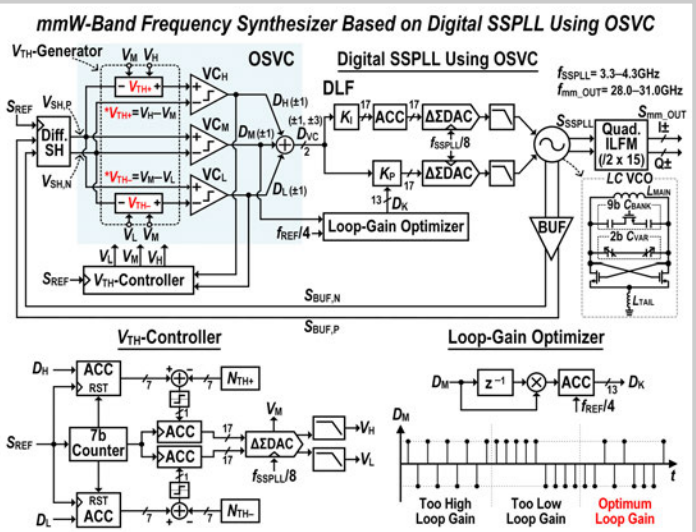


Figure 16.2.2: Overall architecture of the mmW-band frequency synthesizer (top), and the schematics of the V_{TH}-controller and loop-gain optimizer (bottom).

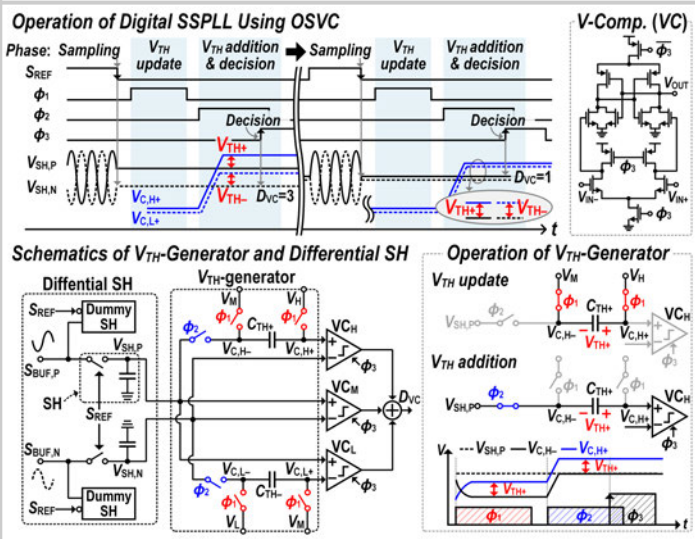


Figure 16.2.3: Operation of the proposed digital SSPLL using the OSVC (top), and the schematics and operation of the OSVC (bottom).

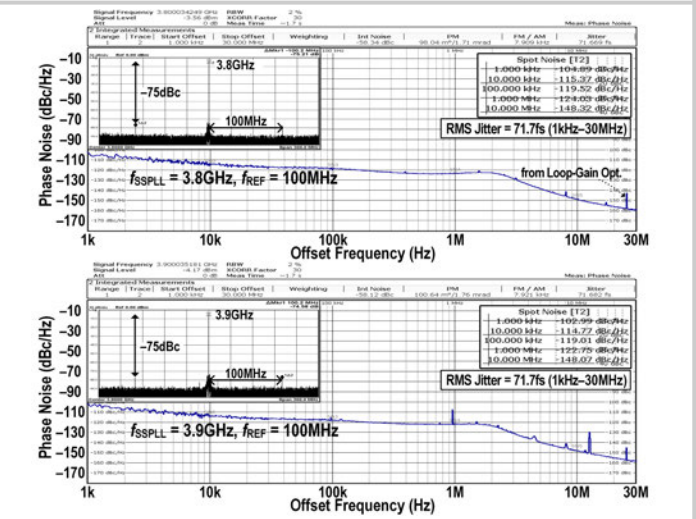


Figure 16.2.4: Measured phase noises and spectra of 3.8 and 3.9GHz-output signals of the OSVC-based digital SSPLL, when the reference frequency, f_{REF}, is 100MHz.

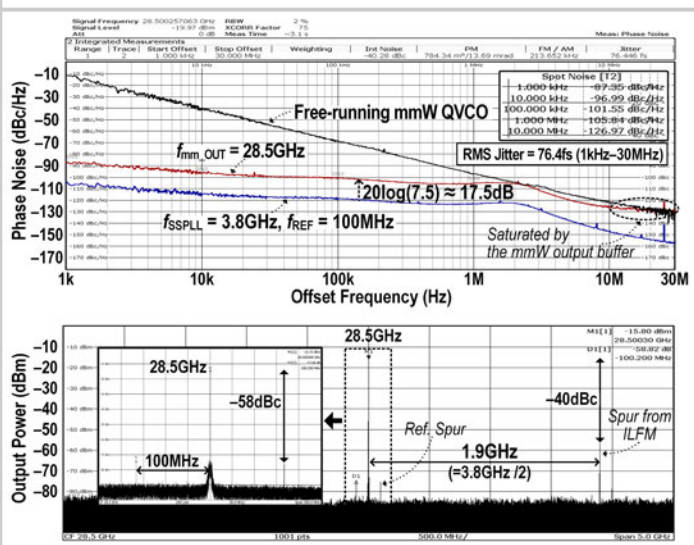


Figure 16.2.5: Measured PN (top) and spectrum (bottom) of the 28.5GHz signal of the proposed mmW-frequency synthesizer.

Comparison with state-of-the-art SSPLLs

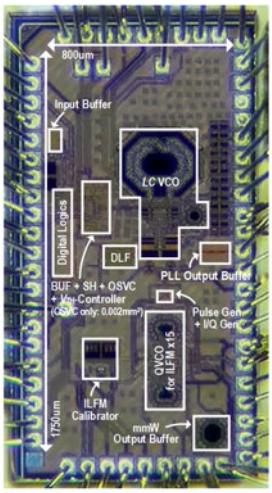
	This work	JSSC'18 D. Liao	ISSCC'15 [3] Z. Chen	ISSCC'15 T. Siriburanon	ISSCC'18 A. Sharkia
Process	65nm CMOS	130nm CMOS	65nm CMOS	65nm CMOS	65nm CMOS
Architecture	Digital SSPLL	Analog SSPLL	Digital SSPLL	Digital SSPLL	Analog SSPLL
Topology	OSVC-based	SS-PD based	ADC-based	ADC-based	SS-PD based
Type	Integer-*N*	Fractional-*N*	Fractional-*N*	Integer-*N*	Integer-*N*
f_{SSPLL} (GHz) / f_{REF} (MHz)	3.3–4.3 / 100	2.39–2.46 / 50	2.6–3.9 / 49.15	2.2 / 100	4.6–5.6 / 100
$Jitter_{RMS}$ (fs) @f_0 (GHz) (Integ. Range)	72 @3.8 (1k–30MHz)	169 @2.397 (10k–30MHz)	226 @2.68 (1k–100MHz)	380 @2.2 (10k–40MHz)	185 @5.0 (10k–50MHz)
Ref. spur (dBc) @f_0 (GHz)	–75 @3.8	–72 @2.397	–60 @2.68	–74 @2.2	–64 @5.0
Power Cons. (P_{DC}) (mW)	19.1	21.0	11.5	4.2	1.1
Active Area (mm²)	0.21	0.43	0.23	0.15	0.01
FoM_{JIT} (dB)**	–250.1	–242.2	–242.3	–242.2	–254.2

Comparison with state-of-the-art mmW-band frequency synthesizers

	This work	ISSCC'18 [1] D. Turker	ISSCC'14 [2] V. Szortyka	ISSCC'17 A. Hussein	ISSCC'18 [6] H. Yoon	JSSC'16 T. Siriburanon
Process	65nm CMOS	16nm FinFET	40nm CMOS	65nm CMOS	65nm CMOS	65nm CMOS
Architecture	GHz-Digital SSPLL + ILFM	CP based PLL	60GHz SS QPLL	All-Digital PLL	RFD + GHz-PLL + ILFMs	20GHz SSPLL + 60GHz QILO
Type	Integer-*N*	Integer-*N*	Integer-*N*	Fractional-*N*	Fractional-*N*	Integer-*N*
Quadrature	YES	NO	YES	NO	YES	YES
f_{OUT} (GHz) / f_{REF} (MHz)	28.0–31.0 / 100	7.4–14.0 / 500	53.8–63.3 / 40	50.2–66.5 / 100	25.0–30.0 / 120	55.6–65.2 / 40
$Jitter_{RMS}$ (fs) @f_0 (GHz) (Integ. Range)	76 @28.5 (1k–30MHz)	54 @12.5 (10k–10MHz)	230 @62.64 (1k–100MHz)	258 @65.35 (1k–40MHz)	206 @29.22 (1k–100MHz)	290 @60.5 (10k–40MHz)
IPN (dBc) Norm. to 28GHz (Integ. Range)	–40.4 (1k–30MHz)	–43.5 (10k–10MHz)	–30.8 (1k–100MHz)	–29.9 (1k–40MHz)	–31.8 (1k–100MHz)	–28.8* (10k–40MHz)
In-band noise (@10kHz) (dBc/Hz) Norm. to 28GHz	–97.1	–100.7	–88.1*	–86.1	–89.0	–85.2
Ref. spur (dBc) @f_0 (GHz)	–58 @28.5	–75 @12.5	–40 @62.64	NA	–83 @29.22	–73 @60.5
Power Cons. (P_{DC}) (mW)	41.8	45.0	42.0	46.0	36.4	32.0
Active Area (mm²)	0.32	0.35	0.16	0.45	0.95	1.08 w/ pads
FoM_{JIT} (dB)*/ $FoM_{JIT,IN}$ (dB)**	–246.1 / –270.7	–248.8 / –262.8	–236.5 / –268.5	–235.1 / –263.3	–238.1 / –262.0	–235.7 / –267.5

*Calculated from measurements **$FoM_{JIT}=10\log(\sigma_t^2 \cdot P_{DC})$ dB ***$FoM_{JIT,IN}=10\log(\sigma_t^2 \cdot P_{DC}/(f_{OUT}/f_{REF}))$ dB (K. M. Megawer, ISSCC'18)

Figure 16.2.6: Performance comparisons with prior-art SSPLLs and mmW-band frequency synthesizers.



Power Consumption (mW)		
PLL (19.1)	GHz *LC* VCO	8.0
	BUF (Before SH) + SH	7.0
	Digital Logics (Loop-gain Opt., etc.) + V_{TH}-Controller + DLF	4.0
	OSVC	0.1
ILFM /2 x 15 (22.7)	Pulse Gen. + I/Q Gen.	11.4
	mmW-Band QVCO	10.4
	ILFM Calibrator	0.9
Total		41.8

Figure 16.2.7: Die micrograph and power-breakdown table.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

16.3 A -246dB Jitter-FoM 2.4GHz Calibration-Free Ring-Oscillator PLL Achieving 9% Jitter Variation Over PVT

Xiaofeng Yang[1], Chi-Hang Chan[1], Yan Zhu[1], Rui P. Martins[1,2]

[1]University of Macau, Macau, China
[2]Instituto Superior Tecnico/University of Lisboa, Lisbon, Portugal

Low-jitter phase-locked loops (PLLs) are critical building blocks in various systems, including wireless and wireline communications and ADCs. LC oscillators exhibit low phase noise (PN) but suffer from large area, magnetic coupling, and small tuning range. While ring voltage-controlled oscillators (RVCO) are free from the above issues, their inferior PN performance restricts their applicability. The injection-locked clock multiplier (ILCM) attains a low PN through phase realignment. However, both the PN and the reference spur deteriorate when the injection instant drifts over PVT. Although the frequency-tracking loop (FTL) can calibrate the center frequency, its power consumption, converging speed, and accuracy limit the performance of the ILCM [1,2]. The Type-I PLL performs wideband filtering [3], but its efficiency is 3dB lower than that of the ILCM [4]. A PLL with fast phase-error correction (FPEC) [4] realizes a PN filtering close to the ILCM and keeps a small reference spur. Nevertheless, both Type-I and FPEC PLLs rely on the closed-loop wideband filtering to achieve a low PN. The drifted loop gain over PVT significantly degrades jitter performance and even causes instability, implying a trade-off between wideband filtering and stability. Consequently, a loop-gain calibration [4] is necessary to maintain the stability and low PN.

In this work, we present an open-loop discrete-time phase-noise cancellation (OPDTPNC), which realizes wideband filtering, as effective as with ILCMs, and simultaneously maintains a low reference spur over PVT without calibration. Figure 16.3.1 shows the concept of the OPDTPNC and compares it with other PLLs with low PN. The basic idea of our method is to sample the noise information from the PLL output with a reference clock (CLK_{REF}) and cancel it through the adjustment of the Δt timing of a delay cell in real time. Rather than using an ILCM that requires an additional FTL, the proposed scheme cascaded at the PLL output is naturally locked. Besides, the adopted Type-II PLL is inherently stable due to the low loop bandwidth (~1/20 F_{ref}). This work utilizes a common structure with PVT-tracking bias both in the sampler and the corrector of the OPDTPNC to maintain the correction efficiency under variations. Eventually, the prototype PLL only experiences a 9% jitter degradation under PVT and loop gain variations in the measurement. Unlike the feedforward-noise-cancellation PLL [5], the OPDTPNC accomplishes an effective phase realignment, like with the ILCM, and the discrete-time (DT) operations allow it to be free from the reference-spur penalty. The PLL and OPDTPNC circuits consume 4.1mW from a 1V supply and exhibit a reference spur and integrated jitter of -63dB and 248fs_{rms}, respectively, yielding a jitter-FoM of -246dB at 2.4GHz.

Figure 16.3.2 illustrates the block diagram of our Type-II sub-sampling (SS) PLL with the proposed OPDTPNC scheme. The SS PLL comprises a charge-pump (CP), a lowpass filter (LF), a frequency-locked loop (FLL), a RVCO, and an SS phase detector (PD). The OPDTPNC includes a constant delay cell (τ), SSPD, PVT-tracking bias, a DC compensator, and a DT corrector. The SS PLL works similarly as in [6] with its PN suppressed by the OPDTPNC that operates as follows. After the Type-II PLL settles, V_{SAM}=V_{REF} and the PLL output (OUT_{VCO}) contains not only the timing jitter (Δt_{VCO}) but also the DC information (V_{REF}). Then, the SSPD2 samples OUT_{VCO} and converts it into the voltage domain with a time-to-voltage conversion gain G_{SAM}. The constant-delay cell τ ensures that the output edges of RVCO arrive at the DT corrector later than the completion of the SSPD2 sampling over PVT. With both the DC and the PN sampled, the DC compensator rejects the undesired DC information. A common and PVT-tracked bias (V_{bp}) is provided for SSPD1, SSPD2, and the DT corrector. This DT corrector further uses the PN-modulated and PVT-tracked basing voltage (V_{SC}) to suppress the PN of the delayed OUT_{VCO}, leading to a low PN output (OUT_{COR}). After cancellation, the timing jitter (Δt_{COR}) of the OUT_{COR} can be expressed as $(1+G_{SAM}\cdot G_{COR})\cdot\Delta t_{VCO}$, where G_{COR} is the voltage-to-time conversion gain of the DT corrector. The corrector time adjustment cancels out the VCO jitter (Δt_{VCO}), when the product of G_{SAM} and G_{COR} is equal to -1, which is achieved by designing the loading capacitor ratio of SSPD2 and DT corrector.

Figure 16.3.3 shows the circuit schematic of the SSPD1, SSPD2, DC compensator, and DT corrector with the corresponding signal behaviors. During the reset phase (ϕ_{RST}), the SSPD1 output V_{SAM} is precharged to the supply voltage V_{DD}, with the SSPD2 output (V_{SC}) and the bottom plate of C_d of the DC compensator connected to V_{bp} and ground, respectively. After ϕ_{RST}, the bottom plate of C_d switches to VDD, redistributing the charge on the top-plate of C_d to C_{SC}, which generates a compensation voltage equal to V_{REF}. Then, OUT_{VCO} is sampled by Ref_Pul through an AND gate, where Ref_Pul is the pulse-modified CLK_{REF}. The sampled DC (V_{REF}) and PN data appear at the AND gate output (Samp_Pul) in the time domain with ϕ_{SAM}, which controls the discharging time in both samplers (SSPD1 and 2). The bias current (I_{bias}) is mirrored to both SSPD1 and 2 through M1 and M2, respectively. While it is further copied to the DT corrector through the diode connected transistor M3, the current in the SSPD1, SSPD2, and DT corrector can be tracked under PVT variations. Consequently, V_{SC} incorporates the bias voltage V_{bp} and the noise-modulated voltage ΔV after ϕ_{SAM}. Since ϕ_{RST} and ϕ_{SAM} finish within half cycle of the input clock (IN_{COR}), V_{SC} settles within the edges of IN_{COR}, which avoids CLK_{REF} interference. Reset and sampling finish in 50ps; therefore, a low reference spur can be maintained with a 2.4GHz output clock.

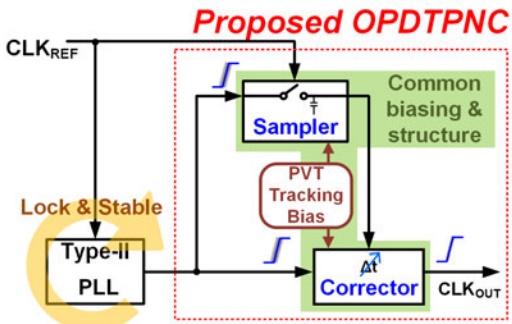
To reduce the reference spurs from the conventional SSPD circuit [6] and maintain a high-speed sampling, this work adopts a domino-based sampler. Besides, OUT_{VCO} is isolated from the time-varying sampler load by several inverters, thus leading to a 17dB improvement in the reference spur compared with [6]. Based on the simulation of a typical corner, normal supply, and 27°C temperature, the phase noise of the corrector is -146dBc at a 1MHz offset, implying that it only contributes to the PN as an inverter. The gains of the corrector (G_{COR}) and sampler (G_{SAM}) depend both on their biasing current and capacitance load. The PVT tracking bias maintains the biasing current of the samplers (I_{SAM}) and the corrector (I_{COR}) with the same value. This work also adopts MoM capacitors to implement C_d, C_{SC}, and C_{COR}; therefore, their relative change is constant over PVT. Nevertheless, the mismatch from the charge pump and between I_{SAM} and I_{COR} and C_d, C_{SC}, and C_{COR} can induce less than 2% jitter degradation. C_{bp} is a 10pF MOS capacitor designed for low noise, and the proposed OPDTPNC converges within one reference clock period once the PLL settles, thereby not degrading the overall settling time of the PLL.

The proposed PLL fabricated in 28nm CMOS consumes a total of 4.1mW, where the ring-VCO, the PLL-loop components, and the OPDTPNC consume 3.3, 0.5, and 0.3mW, respectively. Figure 16.3.4 shows the integrated rms jitter of the PLL (10kHz to 100MHz) that is ~248fs and 686fs while the measured phase noise at a 1MHz frequency offset is -125.4 and -109.3dBc/Hz with and without OPDTPNC, respectively. CLK_{REF} dominates the phase noise performance at the low and high frequency, which is similar to the FPEC and ILCM topologies. The closed-in spurs are caused by the supply ripple of the single-ended RVCO and suppressed by the OPDTPNC. The reference spur is -63dBc and the tuning range is 2.08 to 2.72GHz. Figure 16.3.5 plots jitter degradations over loop-gain variations, temperature, chip samples, and supply. When the temperature, supply, and loop gain vary from 0 to 100°C, ±5%, and 6dB, respectively, the jitter performance only degrades by less than 9%. Figure 16.3.6 summarizes the PLL performance and compares it with the prior art. Figure 16.3.7 shows the die micrograph that occupies an area of 0.023mm² including the OPDTPNC.

Acknowledgement:
This work was financially supported by the Macao Science & Technology Development Fund (FDCT) with Ref no: 0068/2016/A2 and NSFC Grant 61604180.

References:
[1] K. Megawer et al., "A 5GHz 370fs_{rms} 6.5mW Clock Multiplier Using a Crystal-Oscillator Frequency Quadrupler in 65nm CMOS," *ISSCC*, pp. 392-394, Feb. 2018.
[2] D. Coombs et al., "A 2.5-to-5.75 GHz 5 mW 0.3 ps_{rms}-jitter Cascaded Ring-Based Digital Injection-Locked Clock Multiplier in 65nm CMOS," *ISSCC,* pp. 152-153, Feb. 2017.
[3] L. Kong and B. Razavi, "A 2.4GHz 4mW Inductorless RF Synthesizer," *IEEE JSSC,* vol. 51, no. 3, pp. 626-635, Mar. 2016.
[4] Y. Lee et al., "A Low-Jitter and Low-Reference-Spur Ring-VCO Based Switched-Loop Filter PLL Using a Fast Phase-Error Correction Technique," *IEEE JSSC*, vol. 53, no. 4, pp. 1192-1202, Apr. 2018.
[5] S. Nagam and P. Kinget, "A –236.3dB FoM Sub-Sampling Low-Jitter Supply-Robust Ring-Oscillator PLL for Clocking Applications with Feed-Forward Noise-Cancellation," *IEEE CICC,* pp. 1-4, July 2017
[6] X. Gao, et al., "A Low Noise Sub-Sampling PLL in Which Divider Noise is Eliminated and PD/CP Noise is Not Multiplied by N^2," *IEEE JSSC,* vol. 44, no. 12, pp. 3253-3263, Dec. 2009.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE



Items / Topology	Stability	Random Jitter	Ref Spur	Calibration Required
ILCM	☺	☺	☹	☹
Type-I	☹	☹	☺	☹
FPEC	☹	☺	☺	☹
OPDTPNC	☺	☺	☺	☺

Figure 16.3.1: Comparison PLL-architecture features and the proposed OPDTPNC.

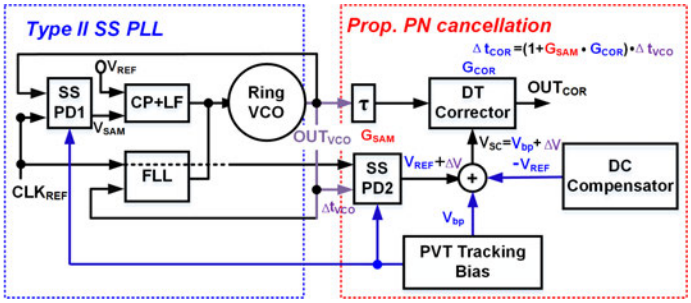


Figure 16.3.2: Block diagram with the OPDTPNC concept.

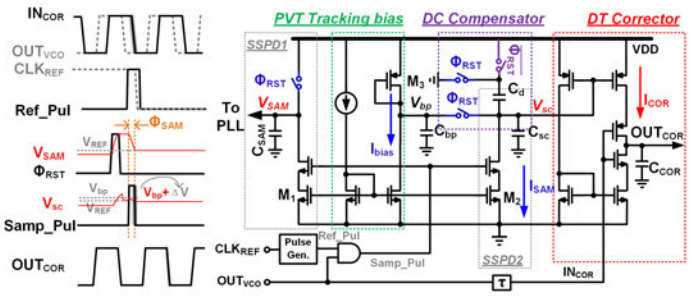


Figure 16.3.3: Circuit implementation of the domino-based sampler, corrector, and DC compensator. Timing diagram is also shown.

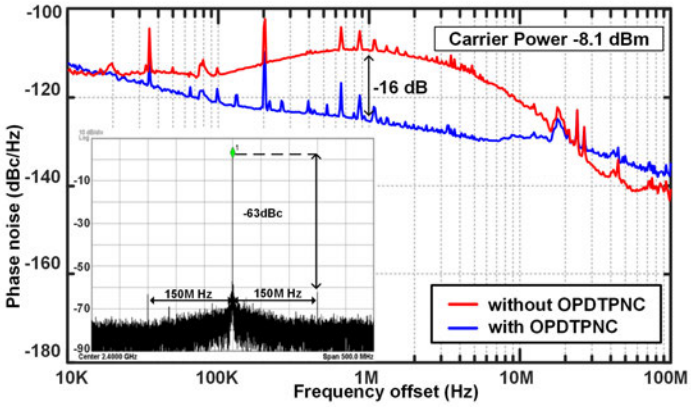


Figure 16.3.4: Measured phase noise w/wo OPDTPNC and reference spur.

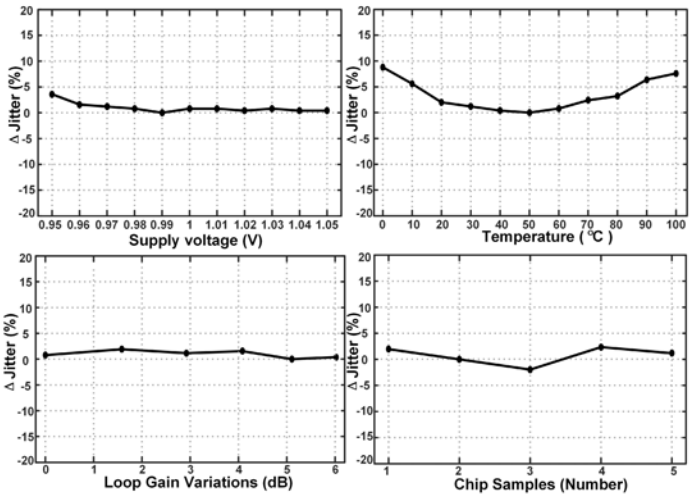


Figure 16.3.5: Measured integrated-jitter variations over PVT.

	ISSCC18 [1]	ISSCC17 [2]	JSSC16 [3]	JSSC18 [4]	CICC17 [5]	This work
Oscillator Topology	Ring	Ring	Ring	Ring	Ring	Ring
Synthesizer Topology	ILCM	ILCM	Type-I	FEPC	FFNC	OPDTPNC
Reference Freq. (MHz)	54	125	22.6	47	49.15	150
Output Freq. (GHz)	4.75	5	2.4	3.008	2.36	2.4
Phase noise @ 1MHz offset(dBc/Hz)	-114	-116	-113.8	-121.6	-119.1	-125.4
RMS jitter (fs) Integ. Range (MHz)	370 (0.01~30)	340 (0.01~40)	970 (0.001~200)	357 (0.001~80)	630 (0.001~100)	248 (0.01~100)
Ref. Spur (dBc)	-53	-45	-65	-71	-55.2	-63
Power(mW)	6.5	9.5	4	4.6	5.86	4.1
Area(mm²)	0.16	0.09	0.015	0.047	0.022	0.023
Technology (nm)	65	65	45	65	65	28
FOM_1 (dB)	-240.5	-242.4	-234.1	-242.3	-236.3	-246
FOM_2 (dB)	179.3	180.2	175.4	184.5	178.9	187

$FOM_1 = 10\log_{10}\left[\left(\frac{jitter_{rms}}{1s}\right)^2\left(\frac{Power}{1mW}\right)\right]$ $FOM_2 = 10\log_{10}\left[\left(\frac{f_{osc}}{\Delta f}\right)^2\left(\frac{1mW}{Power}\right)\right] - \text{Phase Noise(dBc/Hz)}$

Figure 16.3.6: Measured performance summary and comparison with existing integer-N synthesizers.

16

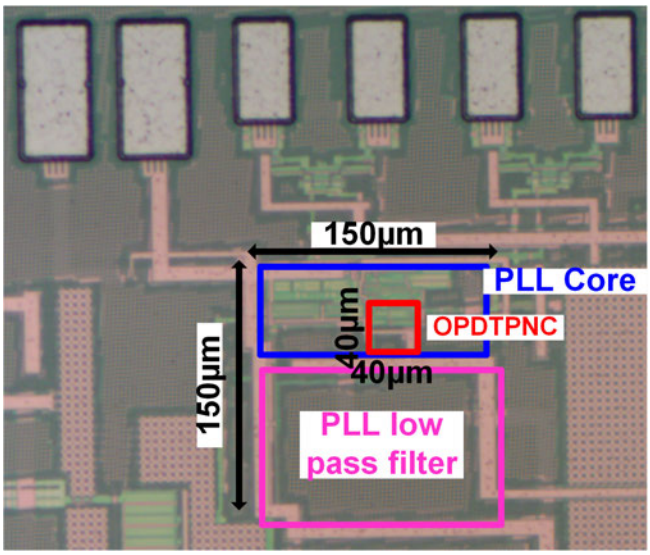


Figure 16.3.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

16.4 A 0.5-to-2.5GHz Multi-Output Fractional Frequency Synthesizer with 90fs Jitter and -106dBc Spurious Tones Based on Digital Spur Cancellation

Szu-Yao Hung, Sudhakar Pamarti

University of California, Los Angeles, CA

There is need for a low-power, compact, means of generating multiple, low-jitter, spectrally pure, clock signals at different frequencies using a single reference oscillator, both in wireline and wireless applications, owing to circuit size, cost, and complexity considerations. PLL-based frequency synthesis – analog or digital – can achieve very low jitter and spur levels, but the VCO size and/or power consumption concerns preclude its duplication for multiple on-chip clock generators [1,2]. Open-loop digital methods based on digital-to-phase or digital-to-time converters (DPCs) are compact and employ no power-hungry oscillators, but their circuit errors and mismatches result in strong spurious tones.

This paper describes a purely digital spur-cancellation technique that uses a bang-bang phase detector (BBPD) and a special noisy, but mostly spur-free, auxiliary clock generator to extract spurious tones precisely and digitally correct for them. When applied to a DPC-based frequency synthesizer, this technique achieves <90fs integrated jitter and spurious tones of <–106dBc for a 2.48GHz carrier.

Figure 16.4.1 shows a block diagram of the multi-clock generator based on the proposed technique. At the heart of each proposed clock generator are two 12b DPCs: *main* and *auxiliary*. The main DPC performs fractional division by $(N+\alpha)$ upon octant phases of frequency f_0 that are derived from an external low-noise PLL. A 2nd-order digital delta-sigma modulator (DDSM) and a phase-select-logic (PSL) block achieve the fractional division as described in Fig. 16.4.2. Even with a high-resolution DPC, inevitable errors in the DPC result in strong spurious tones. The auxiliary DPC performs the same fractional division, but uses a "spur-free" PSL that theoretically guarantees spur-free fractional division, at the expense of excessive phase noise. The BBPD extracts the instantaneous phase difference between divided versions of the low-noise main-DPC output and the noisy, but mostly spur-free, auxiliary-DPC output, both of which have the same frequency, $f_0/(N+\alpha)$. The one-bit BBPD output stream contains the spurious tones in the main DPC output along with the phase noise of the auxiliary DPC output. Digital comb filters with appropriately chosen center frequencies suppress the phase noise and extract just the spurs, including their harmonics [2]. The resultant phase error, *ph_err[n]*, is compensated for the effective bang-bang PD gain, β, (described below), converted to an effective frequency error, *freq_err[n]*, by digital differentiation, negated, and fed as a frequency correction signal to the main-DPC input fractional value, α. The BBPD, the DPCs, the DDSM, and the entire digital spur-cancellation logic run synchronized with the output clock, at a frequency of $f_{dig} = f_0/(N+\alpha)/4$.

Note that the technique assumes approximate knowledge of the spur frequencies. While fractional-spur frequencies can be predicted *a priori*, non-fractional-spur frequencies may have to be determined using frequency estimation techniques. Besides, in an industrial setting, they are generally known from preliminary lab measurements.

Figure 16.4.2 illustrates how fractional division is performed using a DPC and PSL and highlights the specific modifications of the "spur-free" PSL. As shown in Fig. 16.4.1, the fractional value, α, is quantized to a 12b sequence, v[n], using a 2nd-order DDSM. Once every cycle of f_{dig}, v[n] is integrated to generate the next 12b DPC-control code, e[n]. The 3 MSBs of e[n] choose an octant, whereas the 9 LSBs control an analog waveform interpolator, whose output frequency is divided by N. The generation of e[n] can be modeled as the 1st-order DDSM shown, where the feedback represents wrapping around a cycle of f_0. Consequently, adjacent low-to-high transitions of the DPC output are separated by $(N+v[n]/2^{12})/f_0$ seconds realizing fractional division by $(N+\alpha)$. The "spur-free" phase-select logic is different in one key respect. Large highpass-shaped phase dither (almost 0.5 cycles of f_0), generated using a one-bit random sequence $\eta[n]$ filtered by a highpass dither filter, G[z], is added to the integrated v[n]. It is theoretically established [4] that such filtered dither prevents memoryless DPC errors in the auxiliary DPC from causing spurious tones. Simulated baseband spectra shown in the figure confirm the same. Note that the auxiliary DPC is not useful for fractional synthesis because of the large amount of dither.

Figure 16.4.3 illustrates the linearized operation of the BBPD. This highly nonlinear block quantizes its input phase difference to just one bit. However, the large phase noise in its inputs renders its operation linear, modeled by a gain, β, and additive quantization noise, provided its inputs have a non-zero average phase difference. As shown in Fig. 16.4.1, any non-zero phase difference is estimated and fed back to the main DPC. This shifts its output phase just enough to remove the non-zero phase difference. In other words, it aligns the BBPD input edges on average. Furthermore, digital LMS algorithms, shown in Fig. 16.4.1, estimate the effective BBPD gain, β, and accordingly compensate the spur-cancellation path gain. Since the BBPD gain, β, can depend on the spurious-tone frequency location, each spur-cancellation path employs a dedicated gain-estimation LMS loop, as shown in Fig. 16.4.1.

A prototype IC was designed in UMC 28nm CMOS to demonstrate the proposed spur-cancellation technique. Two clock generators were designed to work in parallel. The 12b DPCs, frequency dividers, BBPD, and clock buffers were the only analog circuit blocks. Provisions were made to externally couple spurs into the buffers on the main-DPC output for measurement purposes, as shown in Fig. 16.4.1. The DPCs can operate up to 3GHz at nominal V_{DD} = 0.9V using standard mixed-signal circuit techniques. The rest of the circuitry was purely digital and was designed using a synthesis and place-and-route flow. A die micrograph is shown in Fig. 16.4.7. The core circuit consumes 2.6mW with area of 0.15mm^2.

Phase noise, jitter, and spurious-tone measurements were performed without and with spur-cancellation technique enabled on 3 different chips for multiple settings of α. Figure 16.4.4 shows measured power spectra for 2.48GHz carrier (for α = 1/40) without and with the spur-cancellation technique enabled. As is evident, >30dB spur cancellation is achieved, with measured spur levels limited only by the instrument noise floor. Similar results were obtained on all 3 chips measured. Note that the technique cancels both fractional spurs caused by the DPC and other circuit errors and "external spurs" introduced into the main-DPC output via intentional coupling into the supply of down-stream buffers. The external spurs are an attempt to emulate non-fractional spurs that are commonly observed in clock-generation circuits usually via supply, bias, or substrate coupling. It is important to note that similar spur cancellation was observed even as the fractional spur location was swept by using different values of α. Figure 16.4.5 shows example measured phase-noise plots for the same carrier frequency. As can be seen, 87fs rms jitter, integrated from 10kHz to 40MHz bandwidth was observed. Over the three measured chips and multiple carrier frequencies, the worst-case integrated jitter of 90fs was observed.

Figure 16.4.6 compares the performance achieved by proposed technique with recent prior art. As is evident, the technique achieves at least a 29dB-lower spurious tone and 8× lower integrated rms jitter for lower power consumption, *albeit* at a finer technology node.

Acknowledgement:
This work is supported by the National Science Foundation under grant CNS-1329755.

References:
[1] H. Kim et al., "A 5GHz -95dBc-Reference-Spur 9.5mW Digital Fractional-N PLL Using Reference-Multiplied Time-to-Digital Converter and Reference-Spur Cancellation in 65nm CMOS," *ISSCC*, pp. 258-259, Feb. 2015.
[2] C.-R. Ho and M. Chen, "A Digital PLL with Feedforward Multi-Tone Spur Cancelation Loop Achieving <-73dBc Fractional Spur and <-110dBc Reference Spur in 65nm CMOS," *ISSCC*, pp. 190-191, Feb. 2016.
[3] A. Elkholy et al., "A 20-to-1000MHz ±14ps Peak-to-Peak Jitter Reconfigurable Multi-Output All-Digital Clock Generator Using Open-Loop Fractional Dividers in 65nm CMOS," *ISSCC*, pp. 272-273, Feb. 2014.
[4] S. Pamarti and S. Delshadpour, "Spur Elimination Technique for Phase Interpolation-Based Fractional-N PLLs," *IEEE TCAS-I*, vol. 65, no. 6, pp. 1639-1647, June 2008.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

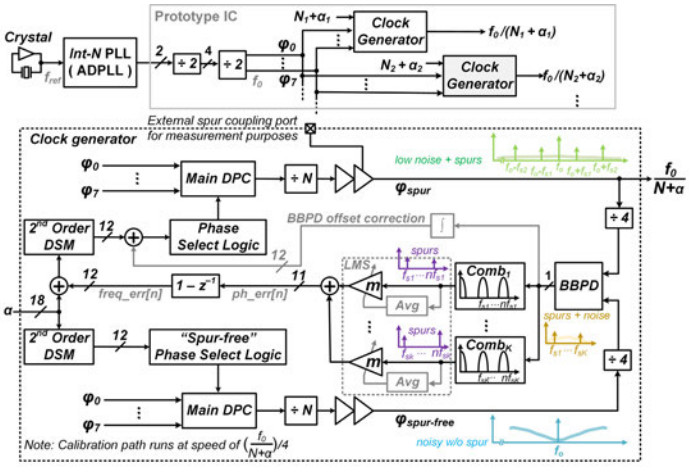


Figure 16.4.1: Proposed multi-output fractional frequency synthesizer based on digital spur cancellation.

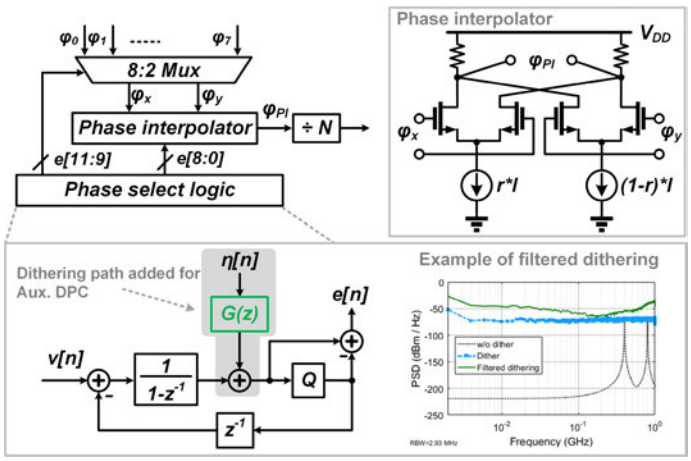


Figure 16.4.2: Implementation of 12b phase-interpolator-based DPCs. For the auxiliary DPC, a shaped filtering path is added to the phase-select logic.

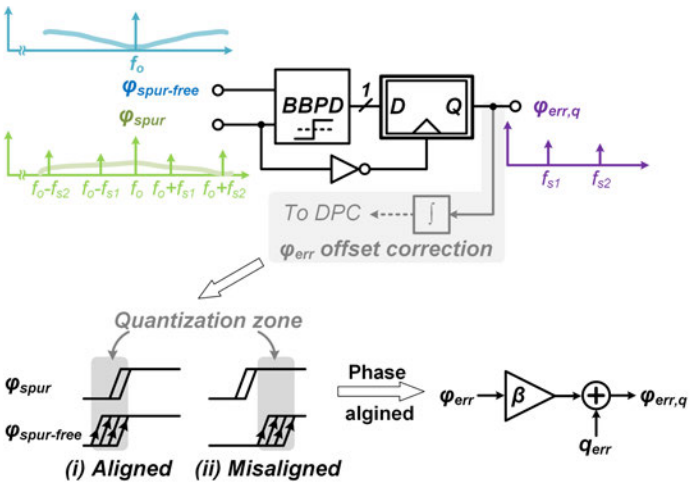


Figure 16.4.3: Linearized operation of a bang-bang PD for spurious phase-error quantization.

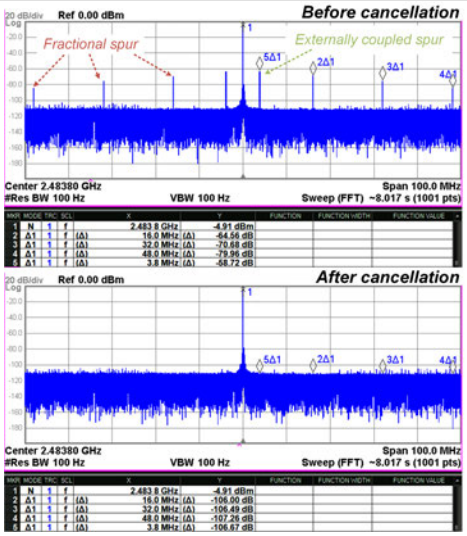


Figure 16.4.4: Measured cancellation of fractional spur and externally coupled spur.

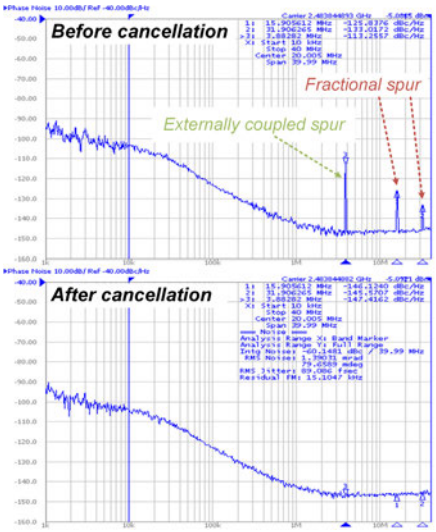
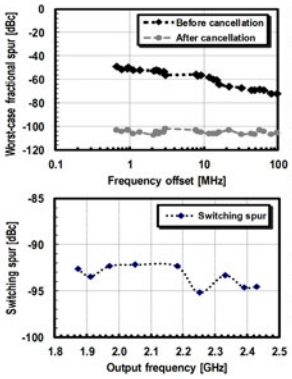


Figure 16.4.5: Measured phase-noise performance with and without spur cancellation.



	Ref [1] Kim	Ref [2] Ho	Ref [3] Elkholy	This work
Architecture	Frac-N TDC	Frac-N TDC	Frac-N Divider	Frac-N DPC
Technology	65nm	65nm	65nm	28nm
Supply [V]	1.0	1.0	0.9	0.9
Freq. range [GHz]	5.0 - 5.4	3.0 - 5.0	0.02 - 1.0	0.5 - 2.5
Fractional spur [dBc]	-42	-74	-55	<-103
External spur rej	No	Yes	No	Yes
RMS jitter [fs] (integartion range) [Hz]	701 (100~40M)	557 (10k~40M)	1440* (10k~40M)	90* (10k~40M)
Ref. / switching spur [dBc]	-95	-110	N/A	<-92***
Instantaneous switching	No	No	Yes	Yes
Core power [mW]	9.52	20.9	3.2**	2.6**
Active area [mm²]	0.22	0.77	0.017**	0.15**

* Measured using external input
** Reference PLL excluded
*** Spurs, located at $f_0/(N+\alpha)/4$, due to switching activities of digital blocks

Figure 16.4.6: Performance summary and comparison with prior art.

16

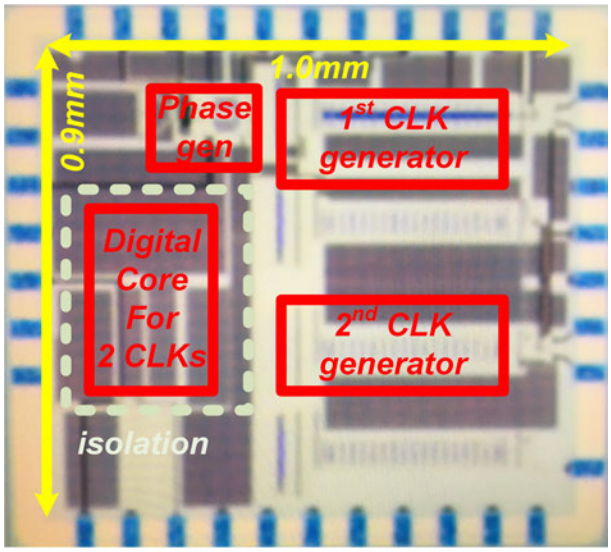


Figure 16.4.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

16.5 A Fractional-N Synthesizer with 110fs$_{rms}$ Jitter and a Reference Quadrupler for Wideband 802.11ax

Fei Song[1], Yu Zhao[1], Bart Wu[1], Litian Tang[1], Leon Lin[1], Behzad Razavi[2]

[1]Ubilinx Technology, San Jose, CA
[2]University of California, Los Angeles, CA

The next-generation 802.11ax WLAN standard improves the throughput by supporting 1024-QAM in a channel bandwidth of 160MHz, demanding extremely low jitter values for the transmitter (Tx) and the receiver (Rx) synthesizers. Recent work has achieved rms jitter below 200fs [1-4] with the latest rms jitter reported at 75fs [5]. The work in [5] faces a number of challenges if applied to 802.11ax: (1) with a VCO frequency of 5.7 to 7.3GHz, the circuit is prone to pulling by the PA, especially because it incorporates a single spiral inductor (rather than an 8-shaped inductor); (2) it does not provide quadrature outputs, and (3) it relies on a 10b digital-to-time converter (DTC) without calibration of its nonlinearity, which folds high-frequency ΔΣ-modulator (ΔΣM) noise unless extremely tight matching is guaranteed. Similarly, the design in [2] is prone to pulling as well.

This paper presents a fractional-N synthesizer that not only achieves a low jitter but also robustly lends itself for use in transceivers and for SoC integration. This level of performance is afforded through the use of three techniques: (1) a background-calibrated reference quadrupler, (2) a power-efficient frequency tripler, and (3) a twin-T notch filter, which is adjusted by the Rx-filter RC-calibration circuitry. Shown in Fig. 16.5.1, the proposed synthesizer is architected to address two critical issues, namely pulling by the PA and the peaking of the third-order ΔΣM noise. The former is resolved by the VCO frequency of 6 to 8GHz and a div-by-2/tripler cascade, and the latter is resolved by a reference quadrupler. In addition to standard techniques, such as a charge pump with an offset current to avoid noise folding and an 8-shaped inductor to suppress coupled noise in an SoC, the loop also incorporates a 40MHz notch filter to reduce both the out-of-band ΔΣM noise and the 40MHz spur. This filter negligibly affects the phase margin because the loop bandwidth must be small enough to limit the amplified phase noise of the crystal oscillator. The VCO includes an amplitude-control loop for optimum phase-noise performance with PVT.

The reference quadrupler plays a critical role in the performance as it lowers the ΔΣM quantization noise contribution to the synthesizer output by 15dB (from 94fs to 16.9fs) compared to a doubler. However, the quadrupler must deal with several issues: the duty cycle error of the 40MHz reference (ε_{40}), the duty cycle error of the doubled frequency (ε_{80}), and the phase noise generated by its constituent stages. This work proposes a three-point calibration method to eliminate the first two and deals with the phase noise through the use of fast edges.

As conceptually illustrated in Fig. 16.5.2, we employ a low-spur frequency-quadrupling calibration PLL (CalPLL) to create transitions that are free from deterministic modulation and compare the quadrupler output edges with these transitions. Even though designed for low power consumption, CalPLL does not contribute phase noise as the calibration routine averages its results over 8192 reference cycles. Note that the doublers must contain delay stages to generate new edges, adding their own phase noise. Taking into account both duty cycle errors, ε_{40} and ε_{80}, and the static phase offset of CalPLL, $\Delta\phi$, we can identify three types of rising edges in y(t): (1) E_{y1} incurs a constant phase error of $\Delta\phi$ with respect to the corresponding edge in x(t), E_{x1}; (2) E_{y2} inherits the duty cycle error of Doubler 1 plus $\Delta\phi$; and (3) E_{y3} inherits ε_{40} plus $\Delta\phi$. As explained below, these edges are utilized for calibration. Note that E_{y1} and E_{y3} are as clean as f_{REF} in terms of phase noise, whereas E_{y2} and E_{y4} inherit the phase noise of the delay stage in Doubler 1.

Figure 16.5.2 also illustrates that the standard delay-plus-XOR doubler topology suffers from two different effects in its output: a pulse-width error due to unequal input rise and fall times; and a pulse-position error arising from ε_{40}. Since it is difficult to calibrate both of these effects, this work introduces a doubler that operates with only rising edges and is free from the pulse-width error.

Figure 16.5.3 shows the detailed realization of the quadrupler. The 40MHz reference drives a duty-cycle-correction (DCC) circuit that generates complementary outputs A and $\overline{A}$. These signals are applied to the cascade of Doubler 1 and Doubler 2. Doubler 1 operates based on the principle that the frequency can be doubled if A and $\overline{A}$ are multiplexed by a selection command, SEL, that is 90° out of phase. This principle is realized by a self-sustained loop consisting of a delay stage, ΔT_1=6.25ns, a div-by-2 circuit, and the MUX. Doubler 1 thus processes only the rising edges of A and $\overline{A}$ as it generates g(t).

The calibration of $\Delta\phi$, ε_{40}, and ε_{80} is greatly simplified by observing that the edges E_{y2} and E_{y3} in Fig. 16.5.2 provide unique and complete information about ε_{80} and ε_{40}, respectively, if $\Delta\phi$ (the error between E_{y1} and E_{x1}) is zeroed. As shown in Fig. 16.5.3, the quadrupled output, y(t), and the CalPLL output are applied to a bang-bang phase detector (BBPD), which measures and keeps track of their phase errors on consecutive edges. The error E_{y1}-E_{x1} is averaged and applied to null $\Delta\phi$, E_{y2}-E_{x2} adjusts ΔT_1, and E_{y3}-E_{x3} adjusts the DCC circuit. After the loops settle, h(t) has a precise 50% duty cycle and very small spurs. The value of ΔT_2 is not critical, and Doubler 2 is based on the conventional topology as the PFD responds to only rising edges. It is important to note that the BBPD phase offset appears in all three measurements and is therefore absorbed by $\Delta\phi$, the DCC circuit, and ΔT_1. This background-calibration method ensures low spurs under all PVT conditions.

To avoid LO pulling, prior work has used cascaded mixing [6]. However, as shown in Fig. 16.5.4, if the waveforms are mixed in a circuit with hard switching, then the output frequency is not tripled, exhibiting a strong fundamental. The tripler proposed here utilizes rail-to-rail swings for high efficiency and employs phase interpolation and edge combing to directly generate the desired frequency while suppressing the first harmonic. Illustrated in Fig. 16.5.4, our method first generates I and Q phases by a divider, from which three phases 120° apart are created. The resulting edges are then combined to triple the frequency. For convenient device ratios, the phase interpolator (PI) approximates theses phases by three phasors equal to +6I+6Q, -8I+2Q and +2I-8Q (and their complements), incurring a maximum error of 3°. Figure 16.5.4 also shows the edge combiner circuit, which drives a negative resistance and a programmable tank to further suppress the mismatch-induced spurs.

The overall synthesizer has been fabricated in 28nm CMOS and occupies an active area of 0.47mm^2. The circuit draws 22.8mW. For ease of testing, the phase noise is measured after an additional div-by-2 stage. All of the spur values reported here are normalized to the LO frequency (f_{LO}). Figure 16.5.5 shows the measured phase noise at 2912.5MHz (f_{LO}=5825MHz) with the reference frequency doubled or quadrupled. The synthesizer achieves an integrated jitter of 110fs$_{rms}$ from 10kHz to 10MHz including all of the spurs in the quadrupler mode, which translates to an integrated phase error (IPE) of 0.23° (-48dBc) and exceeds the 802.11ax specification. By virtue of the 40MHz notch in Fig. 16.5.1, the jitter rises by only 2.6fs if integrated from 10kHz to 100MHz, showing promise for wideband 802.11ax. The measured reference spurs are -76.4dB at 40MHz, -73.4dB at 80MHz, and below -85dBc at 160MHz. The spur level at 40MHz falls by 3-to-5dB after the notch-filter calibration. All fractional spurs are below -60dBc. Table in Fig. 16.5.6 compares the performance of our prototype to that of the prior art.

References:

[1] T.-M. Chen et al, "An 802.11ac Dual-Band Reconfigurable Transceiver Supporting up to Four VHT80 Spatial Streams with 116fs$_{rms}$-Jitter Frequency Synthesizer and Integrated LNA/PA Delivering 256QAM 19dBm per Stream Achieving 1.733Gb/s PHY Rate," *ISSCC*, pp. 126-127, Feb. 2017.
[2] C.-W. Yao et al, "A 14nm Fractional-N Digital PLL with 0.14ps$_{rms}$ Jitter and -78dBc Fractional Spur for Cellular RFICs," *ISSCC*, pp. 422-423, Feb. 2017.
[3] X. Gao et al, "A 2.7-to-4.3GHz 0.16ps$_{rms}$ Jitter -246.8dB FOM Digital Fractional-N Sampling PLL in 28nm CMOS," *ISSCC*, pp. 174-175, Feb. 2016.
[4] Y.-L. Hsueh et al, "A 0.29mm^2 Frequency Synthesizer in 40nm CMOS with 0.19ps$_{rms}$ Jitter and <-100dBc Reference Spur for 802.11ac," *ISSCC*, pp. 472-473, Feb. 2014.
[5] W. Wu et al, "A 5.5-7.3GHz Analog Fractional-N Sampling PLL in 28nm CMOS with 75fs$_{rms}$ Jitter and -249.7dB FoM," *RFIC*, pp. 52-55, June 2018.
[6] X. Gao et al, "A 28nm CMOS Digital Fractional-N PLL with -245.5dB FOM and a Frequency Tripler for 802.11abgn/ac Radio," *ISSCC*, pp. 166-167, Feb. 2015.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

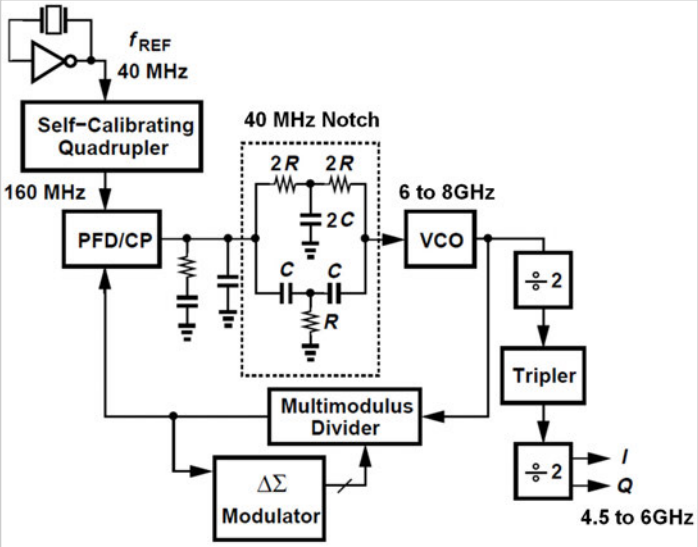


Figure 16.5.1: Proposed synthesizer architecture.

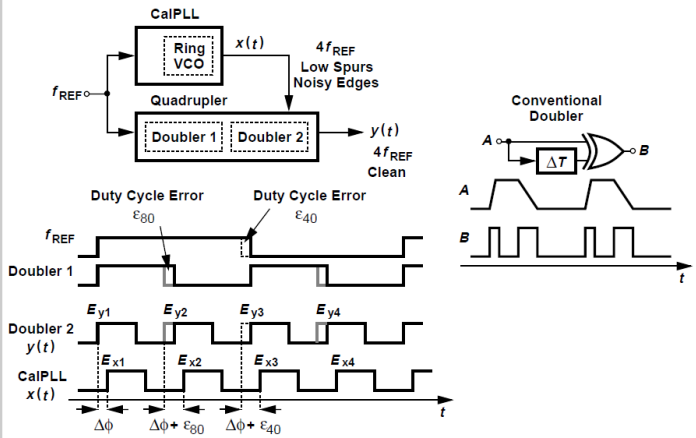


Figure 16.5.2: Quadrupler calibration concept and issues in conventional doubler.

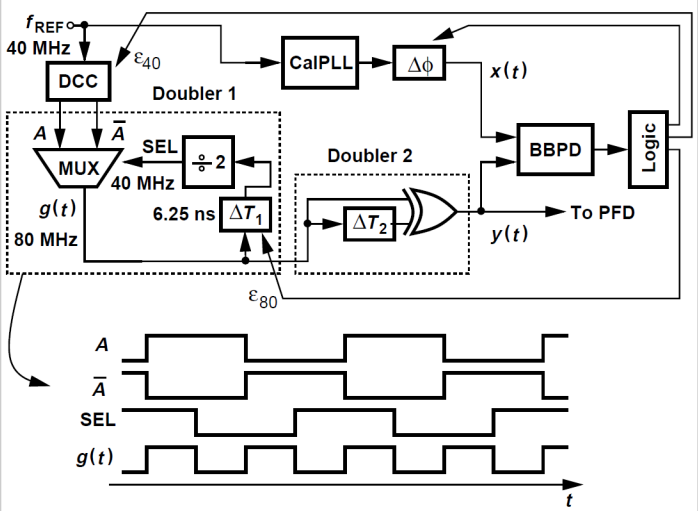


Figure 16.5.3: Detailed realization of the quadrupler.

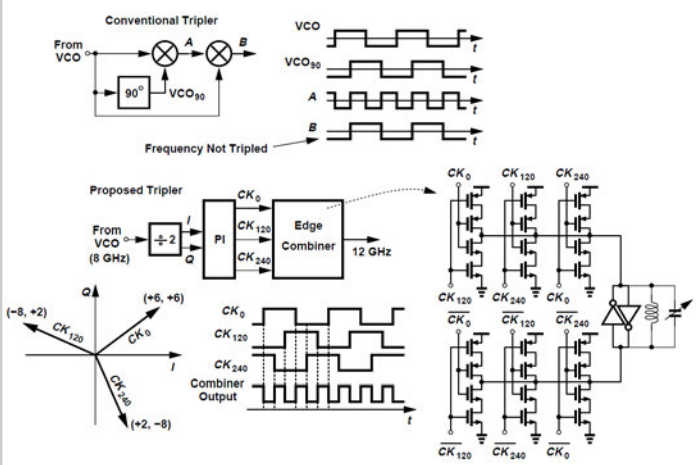


Figure 16.5.4: Conventional and proposed frequency triplers.

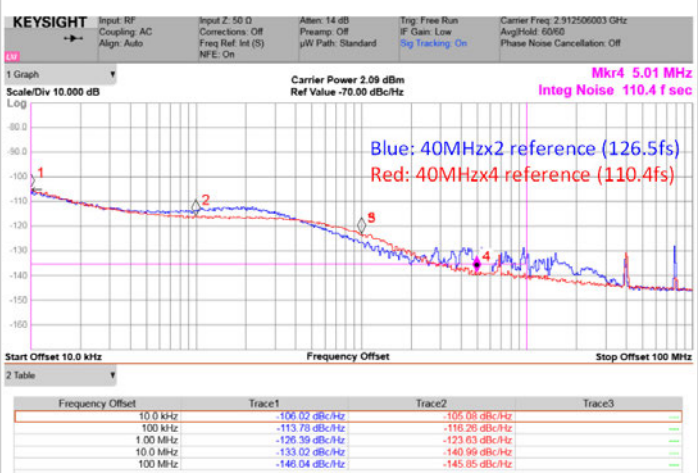


Figure 16.5.5: Measured output phase noise (after a divide-by-2 stage) at 2912.5MHz. The LO frequency is 2×2912.5MHz. Two profiles for reference quadrupling or doubling are shown.

	This Work @40MHzx4	This Work @80MHzx2	Chen. ISSCC'2017	Yao. ISSCC'2017	Gao. ISSCC'2016	Hsueh. ISSCC'2014	Wu. RFIC'2018
Process	28nm	28nm	40nm	14nm	28nm	40nm	28nm
Architecture	Analog	Analog	Analog	Digital	Digital	Analog	Analog
Reference (MHz)	40	80	40	26	40	26	52
RMS jitter (fs)	110	82	116	137	159	191	75
In-band PN Normalized to 5.825GHz (dBc/Hz)	-110.3 @100KHz	-112.7 @100KHz	-107.8 @100KHz	-106.9 @100KHz	-105.5 @100KHz	-102.0 @100KHz	-113.3 @100KHz
Reference Spur (dBc)	-73.4	-66.4	-	-87.6	-78.0	-100.0	-70.1
Power (mW)	22.8	14.7	20.0	13.4	8.2	17.5	18.9
FoM (dB)	-245.6	-250.0	-245.0	-246.0	-246.8	-242.0	-249.7
Area (mm^2)	0.47	0.47	0.43**	0.26	0.30	0.29	0.50

** Estimated from die micrograph

$$FoM = 20\log_{10}\frac{\sigma}{1s} + 10\log_{10}\frac{Power}{1mW}$$

Figure 16.5.6: Performance summary and comparison table.

16

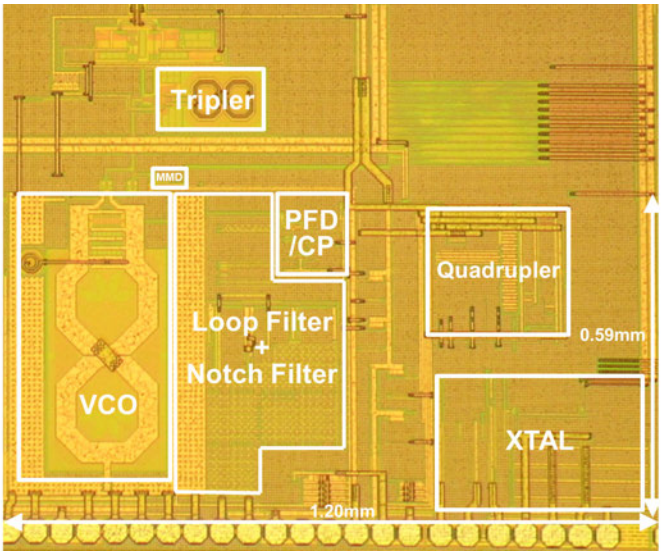


Figure 16.5.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

16.6 A Calibration-Free Triple-Loop Bang-Bang PLL Achieving 131fs$_{rms}$ Jitter and -70dBc Fractional Spurs

Dihang Yang[1], Asad Abidi[1], Hooman Darabi[2], Hao Xu[1], David Murphy[2], Hao Wu[2], Zhaowen Wang[3]

[1]University of California, Los Angeles, CA
[2]Broadcom, Irvine, CA
[3]Columbia University, New York, NY

To deliver a good EVM performance, modern communication standards, such as WiFi 802.11ax with a 1024-QAM mode, require RF clocks with extremely low integrated phase error and low spurs. Because of their good scalability, digital phase-locked loops (DPLLs) have been widely studied [1-3]. However, they face two problems: nonlinearity and quantization noise of time-to-digital converters (TDCs). High-performance DPLLs require complicated TDC structures and calibrations. By contrast, a bang-bang phase detector (PD), i.e. a one-bit TDC, can be linear and low noise [4]. However, in the fractional mode, the PD gain is lowered by the significant phase fluctuations from the fractional divider. The gain can be restored by calibration, but the nonlinearity of the calibration circuit creates large spurs in an otherwise good integrated rms noise [5]. A sub-sampling PLL [6] avoids amplifying the PD noise and eliminates the loop divider, but still needs calibration for fractional operation. This work describes a calibration-free fractional BBDPLL. With the assistance of two auxiliary PLLs, this triple-loop PLL architecture combines the advantages of the sub-sampling PLL and bang-bang PD to achieve 131fs$_{rms}$ jitter and lower than -70dBc fractional spurs.

Our design revives the heterodyne PLL concept implemented in high-performance instruments [7]. As shown in Fig. 16.6.1, two auxiliary PLLs generate a fractional auxiliary clock f_{FAC} for a high resolution and an integer auxiliary clock f_{IAC} with coarse tuning steps. In the main PLL, the mixer with a cascaded lowpass filter passes the frequency difference. The main LC-oscillator (LCO) spectrum is frequency shifted to IF without change in the SNR. By replacing the conventional frequency divider with a mixer, this main PLL achieves a unity closed-loop transfer function. The PD senses the phase error of the two inputs and generates the corresponding control signal for the main LC oscillator. Then, when the main PLL is locked, the output frequency is the sum of two auxiliary frequencies. Because of the unity closed-loop transfer function, the fractional PLL noise and spurs, which have been lowered by the divider N_{FAC}, appear unaltered at the main PLL output. This divider, when cooperating with the main PLL, is the key to lowering the spurs and noise of a fractional PLL. The triple-loop PLL used in instruments [7] has been historically realized by shielded discrete modules, thus enjoying small signal crosstalk. When implemented on the same chip, the integer LCO and the main LCO can suffer from strong mutual pulling (Fig. 16.6.2).

Sub-sampling is used to avoid the potential pulling while retaining the benefit of heterodyne PLLs. As shown in Fig. 16.6.2, instead of using a mixer to achieve the fundamental harmonic mixing, the integer auxiliary clock drives a sample-and-hold (SH) and uses its 5th image to downconvert the main LCO output. The LPF and the SH intrinsic sinc() frequency response ensures a low-spur IF. To prevent pulling further, the integer LCO is divided by 8 or 9 before sampling the main LCO. In this way, the ratio of the two LCO frequencies is a rational fraction, avoiding the mutual pulling from the harmonics of two LCOs. Now, the heterodyne PLL can be integrated.

As shown in Fig. 16.6.3, the integer LCO PLL is a Type-I sampling PLL. The current of the 12-to-14GHz Class-A LCO is controlled by tail resistors, which have no flicker noise. The four transistors of the LCO are kept in the saturation region to maintain a high loaded Q. The PD, a simple inverter, introduces negligible phase noise. The phase noise of the loop divider is designed to be less than the reference phase noise. The reference frequency and multiples lie at the nulls of the sample-and-hold sinc() frequency response [8], allowing the loop bandwidth to be widened to suppress the LCO noise. Thus, this LCO PLL achieves a 174fs$_{rms}$ jitter and a -251dB FoM. The fractional frequency is generated by a Type-II charge-pump (CP) PLL with a 2.7-to-3.2GHz ring oscillator (RO). There is negligible pulling of the two LCOs by this RO. The fractional divider is controlled by a 2nd-order $\Delta\Sigma$ modulator. After optimizing the fractional PLL bandwidth, at the cascaded divider output, the in-band noise and spurs are less than -130dBc/Hz and -70dBc, respectively. Due to the unity closed-loop transfer function, the divided noise and spurs of the fractional PLL are not amplified at the main loop output. Thus, with no calibration, the fractional PLL noise and spurs are efficiently suppressed. In the main bang-bang DPLL (Fig. 16.6.4), the 7-to-9GHz LCO is a complementary circuit with a tail inductor to lower phase noise. A coarse TDC is implemented to widen the PLL locking range. When the PLL is locked, the loop automatically switches to a low hysteresis Strongarm-latch bang-bang PD for fine control. The digital-to-analog converter (DAC) is a 10b resistor ladder. To avoid main LCO being pulled by the external clock, the clock for the digital circuits is generated from the LCO. The sample-and-hold circuit is similar to the one in the integer PLL.

The triple-loop PLL is fabricated in the TSMC 16nm process (Fig. 16.6.7). An on-chip 60MHz crystal oscillator serves as the reference. Fig. 16.6.5 shows the PLL noise performance at around 7.7 and 8.7GHz. At both frequencies, with the noise being integrated from 10kHz to 10MHz, the PLL achieves a 131fs$_{rms}$ jitter (spurs location >10MHz). After a 2/3 divider, which centers the frequency to the 5.805GHz 802.11a channel, this jitter corresponds to 0.27° rms integrated phase noise. The SH circuit parasitic capacitance was higher than expected, limiting the integer PLL bandwidth and raising upconverted flicker noise in the output. A better control of the SH parasitic will result in an estimated rms jitter of 121fs.

Figure 16.6.5 plots worst-case spurious tones, when sweeping fractional dividing ratio of the fractional PLL. The worst fractional spur is -70dBc. Since the main-PLL reference frequency is determined by the fractional auxiliary clock, when sweeping its frequency from 50 to 87MHz the corresponding offset frequency and amplitude of the main-loop reference spur are changed accordingly. The worst-case main-PLL reference spur is -66dBc. Because of strong filtering from the three PLLs, no spur is observed at the reference-crystal frequency. The supply of the PLL is 0.9V. The fractional PLL with the cascaded divider consumes 1.9mA; the integer PLL consumes 2.8mA; the divider cascaded with the integer PLL consumes 2.1mA due to its high input frequency and long routing wires; and the main PLL consumes 6.6mA of which the main LCO takes 4.9mA, leading to a total power consumption of 13.4mA. The FoM of the prototype is -246.4dB. The performance summary and comparison is shown in Fig. 16.6.6.

Acknowledgement:
The authors thank Wenlong Jiang for DPLL discussions, Long Kong for the advice on the integer PLL, and other UCLA peers for insightful discussions. We also thank Arya Behzad and Dandan Li for resource supports, Stephen Au and Richard Oh for technical support, Jessie Xu and Ningyi Wang for layout suggestions, Ryan Mascarenhas and Badri Amudachary for digital PnR, Guy Geshvindman and Ken Wong for PCB advise, and other Broadcom engineers for valuable discussions.

References:
[1] C. Yao et al., "A 14nm Fractional-N Digital PLL with 0.14ps$_{rms}$ Jitter and –78dBc Fractional Spur for Cellular RFICs," *ISSCC*, pp. 422-423, Feb 2017.
[2] X. Gao et al., "A 2.7-to-4.3GHz, 0.16ps$_{rms}$ Jitter, –246.8dB-FoM, Digital Fractional-N sampling PLL in 28nm CMOS," *ISSCC*, pp. 174-175, Feb. 2016.
[3] X. Gao et al., "A 28nm CMOS digital fractional-N PLL with –245.5dB FoM and a Frequency Tripler for 802.11abgn/ac Radio," *ISSCC,* pp. 166-168, Feb. 2015.
[4] H. Xu and A. Abidi, "Design Methodology for Phase-Locked Loops Using Binary (Bang-Bang) Phase Detectors," *IEEE TCAS-I*, vol. 64, no. 7, pp. 1637-1650, July 2017.
[5] L. Bertulessi et al., "A Low-Phase-Noise Digital Bang-Bang PLL with Fast Lock over a Wide Lock Range," *ISSCC*, pp. 252-254, Feb. 2018.
[6] Z. Chen et al., "Sub-Sampling All-Digital Fractional-N Frequency Synthesizer with –111dBc/Hz In-Band Phase Noise and an FoM of –242dB," *ISSCC*, pp. 268-270, Feb. 2015.
[7] D.Scherer et al., "Low-Noise RF Signal Generator Design," *Hewlett Packard Journal,* vol. 32, no. 2, Feb. 1981.
[8] L. Kong and B. Razavi, "A 2.4GHz 4mW Inductorless RF Synthesizer," *ISSCC*, pp. 450-451, Feb. 2015.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

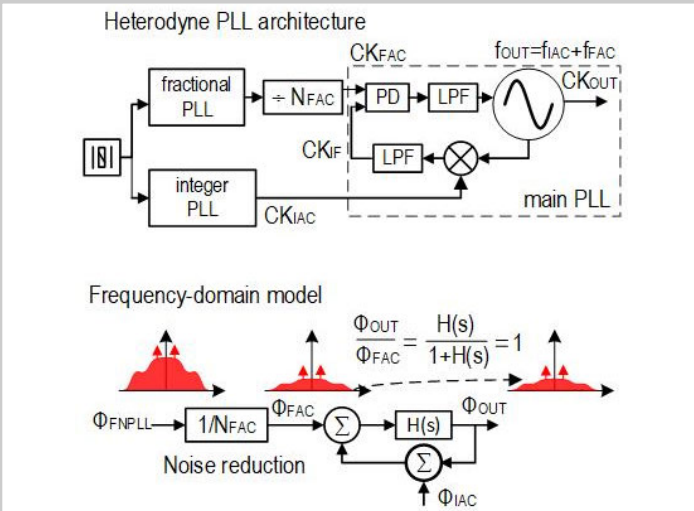


Figure 16.6.1: The concept of heterodyne synthesizer, showing how it lowers phase noise in output by unity closed-loop transfer function and frequency division.

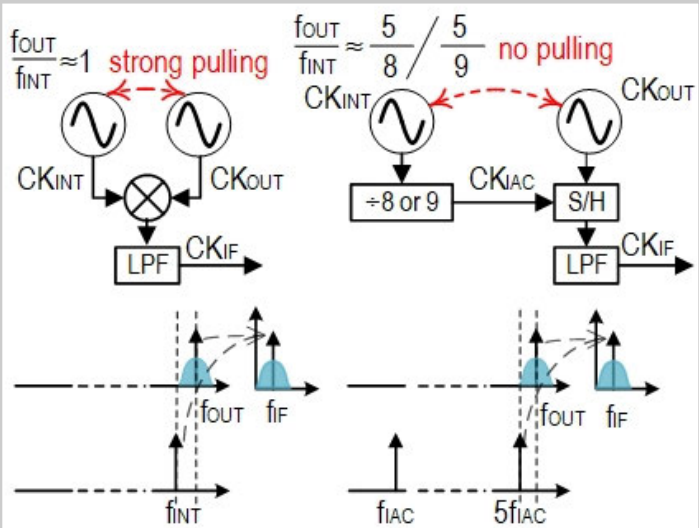


Figure 16.6.2: Avoiding parasitic frequency pulling between two oscillators by using a sample-and-hold, which allows sub-sampling, instead of a mixer.

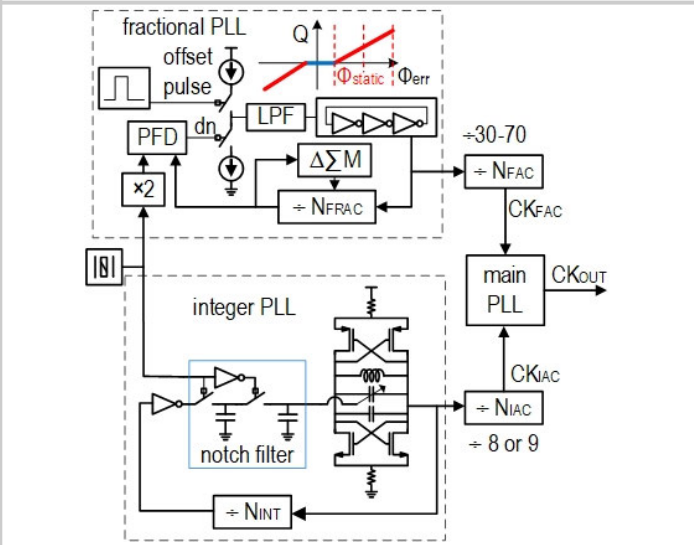


Figure 16.6.3: Details of the integer PLL using sample-and-hold phase detector and fractional PLL.

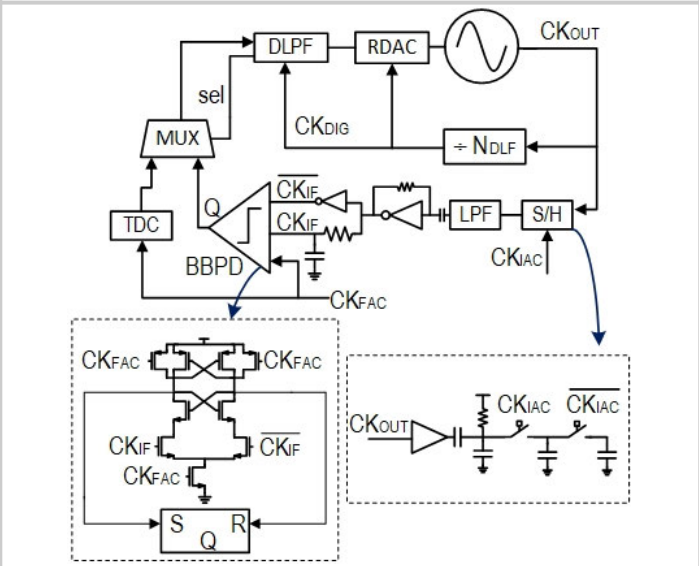


Figure 16.6.4: Circuit details of the main PLL.

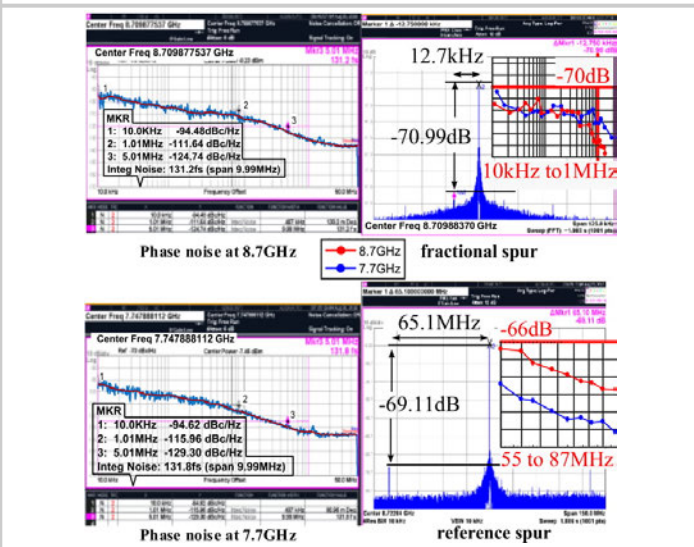


Figure 16.6.5: Measured phase noise and spur levels at around 8.7 and 7.7GHz output frequencies.

	Our work	Yao [1] ISSCC17	Gao [2] ISSCC16	Gao [3] ISSCC15	Bertulessi [5] ISSCC18	Chen [6] ISSCC15
PLL architecture	Triple-Loop BBDPLL	DPLL	DPLL	DPLL	BBDPLL	Subsampling DPLL
Calibration	no	yes	yes	yes	yes	yes
Reference (MHz)	60	26	40	40	52	49.15
Output (GHz)	7.0-9.0	2.7	2.7 - 4.33	3.2 - 3.8	3.7 - 4.1	2.6 - 3.9
In-band PN (dBc/Hz)[1]	-108.5	-103.4	-104	-103.5	-99.5	-100.3
Ref.Spur (dBc)[1]	<-66	-77.4	-74.5	-72	N/A	-50
Frac.Spur (dBc)[1]	<-70	-68.4	-50.5	N/A	-43	-52
RMS jitter (fs)	131 (121 with SH improved)	137	159	173	183	226
Power	13.4	13.4	8.2	9.5	5.28	11.5
FoM (dB)[2]	-246.4 (-247 with SH improved)	-245.9	-246.8	-245.5	-247.5	-241.8
Area (mm)	0.25[3]	0.257	0.3	0.3	0.61	0.23
Process (nm)	16	14	28	28	65	65

1 the phase noise and spurs are normalized to 8.7GHz for fair comparison

2 FoM = $10\log_{10}[(\sigma^2_{jitter})$*Power/1mw]

3 circuit core area

Figure 16.6.6: Comparison table.

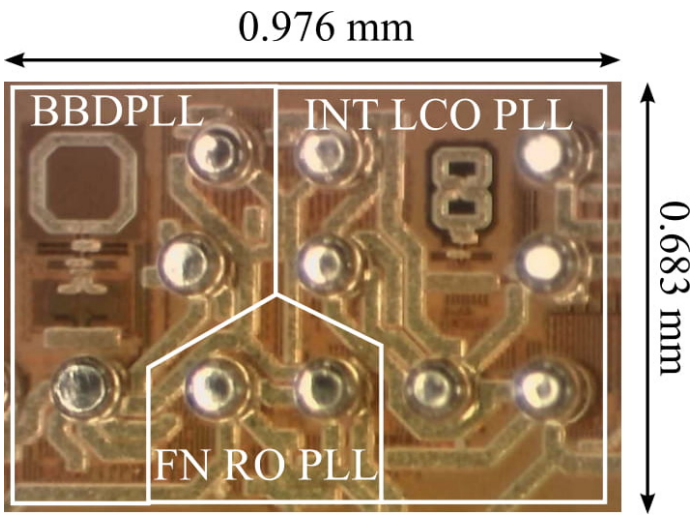


Figure 16.6.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

16.7 A 30GHz Digital Sub-Sampling Fractional-N PLL with 198fs$_{rms}$ Jitter in 65nm LP CMOS

Luigi Grimaldi, Luca Bertulessi, Saleh Karman, Dmytro Cherniak, Alessandro Garghetti, Carlo Samori, Andrea L. Lacaita, Salvatore Levantino

Politecnico di Milano, Milan, Italy

Digital phase-locked loops (DPLLs) have been demonstrated to achieve excellent performance as fractional-N frequency synthesizers in the sub-6GHz range, but, when used in the mm-wave range, at 30GHz and above, they typically feature poor jitter-power trade-off. This paper presents a 30GHz fractional-N DPLL in 65nm low-power (LP) CMOS achieving 198fs integrated jitter at 35mW power. The DPLL based on a digital-to-time converter (DTC) combines: (i) sub-sampling bang-bang phase detection to reduce power consumption, (ii) a digital error correction to reduce DTC noise contribution, and (iii) a low-power wide-lock-range divide-by-six circuit to allow operation up to 34GHz.

Figure 16.7.1 shows the simplified block diagram of the implemented DPLL. Unlike the fractional-N sub-sampling DPLL introduced in [1], this scheme avoids a multi-bit ADC, and, unlike the phase-domain DPLL (PD-DPLL) in [2], it avoids a multi-bit time-to-digital converter (TDC) and a synchronous counter. The output of the 30GHz digitally controlled oscillator (DCO) is frequency divided by six to 5GHz, via an injection-locked frequency divider (ILFD), and then is directly fed to the sub-sampling bang-bang (SSBB) phase detector, realized as a D-type flip-flop. The 5GHz *fb* signal drives a multi-modulus divider (MMD) in a coarse frequency acquisition loop, which remains idle as long as the time error is within the dead zone of a 3-level TDC. A DTC, whose control word is calibrated via LMS, is placed along the reference path of both loops to cancel the quantization error of the $\Delta\Sigma$ modulator (DSM). In DTC-based PLLs, the order of DSM higher than one is mandatory to allow fast LMS convergence, especially for near-integer frequencies. For instance, switching from a first- to a second-order DSM, the LMS settling time reduces from about 35ms to about 25μs for a channel 2kHz off the integer one. Unfortunately, the DTC time range should increase from T_{fb} to $2T_{fb}$ to accommodate the doubled DSM quantization error, where T_{fb} is the period of the *fb* signal. This would lead to larger DTC area and phase noise, which is typically the dominant contribution to the in-band region of the DPLL output spectrum. To avoid this extra-noise penalty, while maintaining the fast convergence of a second-order DSM, the output time range of the DTC is halved, via the multiplication of the accumulated DSM quantization error by 2^{-1} and the multiplication of the input and output of the DSM by 2 and 2^{-1}. After this operation, the modulus control MC[k] is no longer an integer number, but has the form (N+0.5F), with integer N and F. Hence, the coarse loop leverages a fractional MMD capable of adding time shifts multiple of $T_{fb}/2$. The block "Q", which is a modulo-2 accumulator, separates MC[k] into an integer word MC'[k] and a single bit p[k], that is equal to one when the DSM induces a delay of an odd multiple of $T_{fb}/2$. MC'[k] is used as modulus control of the integer MMD in the coarse loop, while p[k] is used to switch between the div0 and div1 outputs of the MMD, which are time offset by $T_{fb}/2$ [3].

The same approach cannot be applied to the sub-sampling loop because of the lack of an MMD. In principle, an analog multiplexer (MUX) switching between 0/180° phases of the 5GHz *fb* signal could be employed. To save power, instead of acting on the phase of *fb*, we leveraged a digital error correction at the output of the SSBB detector. The operation principle of the digital correction technique, based on changing the sign of e[k] on the basis of p[k], is illustrated in Fig. 16.7.2. Unlike a conventional divider-based bang-bang PLL, the output-vs-time-error characteristic of the SSBB detector is T_{fb}-periodic. This periodicity makes the sub-sampling PLL insensitive to time errors multiple of T_{fb}, because the detector still operates linearly on average though on a shifted threshold. Instead, for odd multiples of $T_{fb}/2$, the detector operating point is moved from around a positive-edge threshold to a negative one, thus changing the sign of its output. To avoid the resulting momentary inversion of the loop gain, which would significantly worsen phase noise, e[k] is sign-inverted each time a $T_{fb}/2$ delay is introduced by the fractional-N operation of the SSBB loop, i.e. each time p[k] = 1. Such a costless operation is enabled by the use of the SSBB detector and it ultimately allows using a second-order DSM together with a reduced-range DTC. In a practical situation, *fb* departs from the ideal 50% duty cycle, producing a degradation of phase noise. To solve this issue, an LMS-based calibration is introduced, which computes the correlation between the corrected SSBB detector output, e'[k], and p[k] and estimates a digital correction of the duty-cycle error, which is added to the DTC control word.

The RF building blocks are shown in Fig. 16.7.3. The 30GHz DCO is fine-tuned by a pair of NMOS varactors, biased by a 6b DAC (including a first-order DSM to refine its resolution and a two-pole lowpass filter to reduce the DSM quantization noise). Three thermometric-coded capacitor banks are employed for coarse tuning. A divide-by-six ILFD, AC-coupled to the DCO output, is introduced to allow low-power operation of both the 5GHz SSBB detector and MMD in a 65nm LP process. The divider is based on the cascade of a divide-by-two and a divide-by-three LC ILFD stages. The second stage exploits the divide-by-two secondary-locking technique introduced in [4]. However, unlike [4], it features an additional injection path by picking the second-harmonic signal at the common-mode floating sources, and AC-coupling it to the tail current generator [5]. This extra injection path (the arrow in Fig. 16.7.3) reinforces injection and widens the lock range (from a simulated 17 to 24% of the center frequency) without power penalty. Thanks to the introduction of the ILFD, the CML-logic SSBB detector and the CMOS-logic MMD dissipate 1.3 and 0.5mW, respectively, while the ILFD itself consumes 6mW.

The PLL has an input reference frequency of 100MHz, generated by an external oscillator, and is capable of fractional-N synthesis in the 30.6-to-34.2GHz range with a 191Hz resolution. The chip micrograph is shown in Fig. 16.7.7, where the active area is 0.55 mm^2. Figure 16.7.4 shows the output phase-noise spectrum, measured at the 5GHz *fb* output. The PLL achieves -103.4 and -115dBc/Hz at 0.1 and 1MHz offsets, respectively, which corresponds to -87.8 and -99.4dBc/Hz at the same offsets, when referred to the 30GHz output. The integrated output jitter is 197.6 and 180fs for fractional-N and integer-N channels, respectively. The measured in-band fractional spur of -57.8dBc at a 304kHz offset from the 5.1GHz carrier is presented in Fig. 16.7.5, which corresponds to -42.2dBc when referred to the 30GHz carrier. The total measured power consumption (excluding the pad driver) is 35mW from a 1.2V voltage supply. The jitter-power figure-of-merit (FoM) is equal to -238.6dB, which is about at par with the analog PLL in [6]. Figure 16.7.6 provides a comparison table with prior-art digital PLLs in the 24-to-66GHz range.

Acknowledgement:
The authors would like to thank Integrand Software Inc. for donating a license of EMX simulator and Xinyan Tang for helping with the design of the DCO.

References:
[1] Z.-Z. Chen et al., "A Sub-Sampling All-Digital Fractional-N Frequency Synthesizer with -111dBc/Hz In-Band Phase Noise and an FoM of -242dB," *ISSCC*, pp. 268-269, Feb. 2016.
[2] A. Hussein et al., "A 50–66-GHz Phase-Domain Digital Frequency Synthesizer With Low Phase Noise and Low Fractional Spurs," in *IEEE Journal of Solid-State Circuits*, vol. 52, no. 12, pp. 3329-3347, Dec. 2017.
[3] D. Tasca et al., "A 2.9-to-4.0GHz Fractional-N Digital PLL with Bang-Bang Phase Detector and 560fs$_{rms}$ Integrated Jitter at 4.5mW Power," *2011 ISSCC*, pp. 88-90, Feb. 2011.
[4] J. Wu et al., "Divide-by-Three Injection-Locked Frequency Divider Combined With Divide-by-Two Locking," *IEEE Microwave and Wireless Components Letters*, vol. 23, no. 11, pp. 590-592, Nov. 2013.
[5] A. Garghetti et al., "A Novel Single-Inductor Injection-Locked Frequency Divider by Three With Dual-Injection Secondary Locking," *IEEE TCAS-I*, in press.
[6] H. Yoon et al., "A –31dBc Integrated-Phase-Noise 29GHz Fractional-N Frequency Synthesizer Supporting Multiple Frequency Bands for Backward-Compatible 5G using a Frequency Doubler and Injection-Locked Frequency Multipliers," *ISSCC*, pp. 366-368, Feb. 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

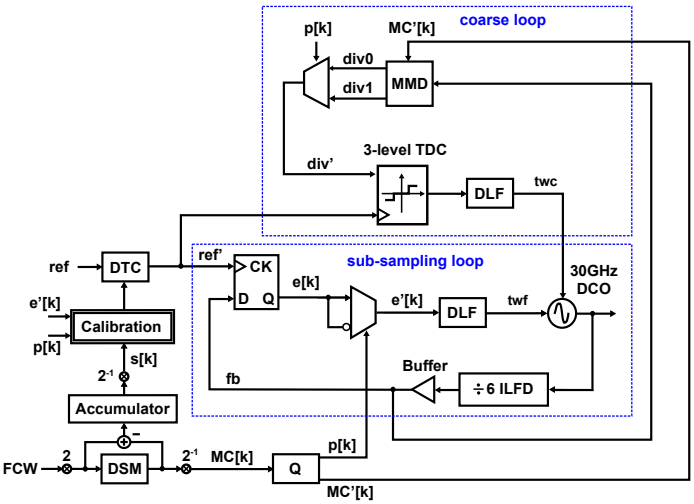


Figure 16.7.1: Simplified block diagram of the 30GHz digital sub-sampling bang-bang PLL.

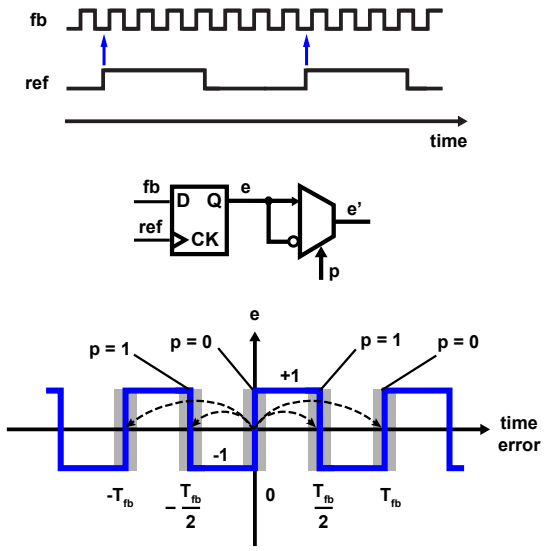


Figure 16.7.2: Characteristic of the SSBB detector and digital correction.

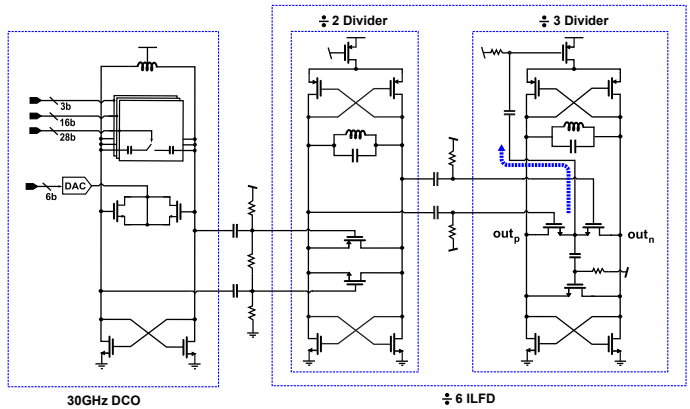


Figure 16.7.3: Circuit implementation of the 30GHz oscillator and the divide-by-six ILFD.

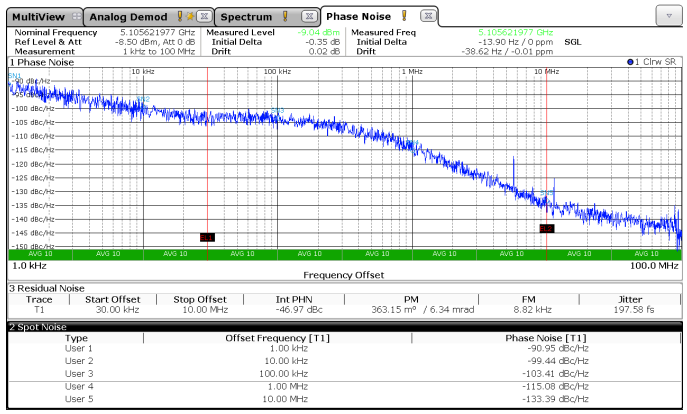


Figure 16.7.4: Measured phase noise in the fractional-N mode at the divided-by-six output.

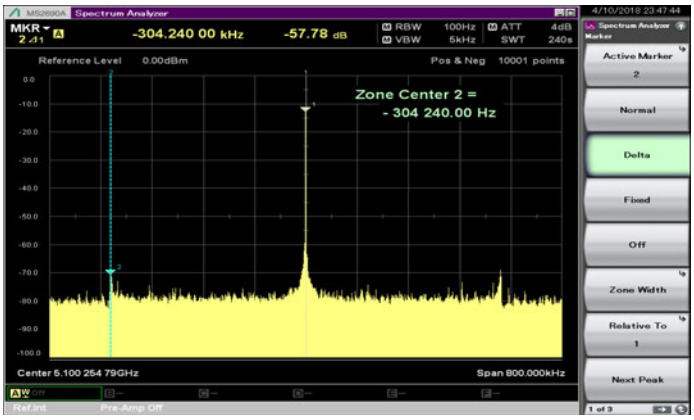


Figure 16.7.5: Measured in-band fractional spur at the divided-by-six output.

	This work	Wanghua Wu JSSC14	Hussein JSSC17	Dayanik ESSCIRC15	Hekmat JSSC15	Ferriss ISSCC15	Weyer ISSCC18
Architecture	SSBB-DPLL	PD-DPLL	PD-DPLL	DPLL	BBDPLL	Hybrid PLL	DPLL
Type	Fractional	Fractional	Fractional	Fractional	Integer	Fractional	Fractional
Reference (MHz)	100	100	100	270	390	104	120
Output (GHz)	30.6-34.2	56.4-63.4	50.2-66.4	28.5-33.5	21.4-25.1	13.1-28	36.3-38.2
Tuning Range (%)	13	11.6	28	16	16	73	5.1
BW (MHz)	0.5	0.3-0.4	5	1	8	2	1
PN*(dBc/Hz) 10kHz	-83.9	-81	n.a.	n.a.	n.a.	-67.4	-88.1
PN*(dBc/Hz) 100kHz	-87.8	-78	-85.7	-87	-93.4	-77.4	-87.4
PN*(dBc/Hz) 1MHz	-99.5	-96	-94.9	n.a.	-101	-79.6	-75.7
PN*(dBc/Hz) 10MHz	-117.8	-117	-122.9	n.a.	-95.4	-114.2	-114.3
RMS jitter (fs)	197.6[a]	590.2	962/1540[a]	545	394	1027**[a]	1150
Jitter integr. BW	30kHz-10MHz	n.a.	n.a.	10kHz-1MHz	100kHz-100MHz	20kHz-50MHz	10kHz-100MHz
Power (mW)	35	48	46	34.8	64	31	68
FoM(dB)***	-238.6	-227.8	-223.7/-219.6**	-230	-230**	-225**	-220.5**
In-band fractional spur (dBc)	-42.2*	n.a.	-54.6/-57.3*	n.a.	n.a.	n.a.	n.a.
Voltage Supply	1.2	1.2	1	1.2	n.a.	1	0.9/1.1/1.2
Area (mm^2)	0.55	0.48	0.45	0.12	0.1	0.24	0.18
CMOS process	65 nm LP	65 nm LP	65 nm GP	65 nm LP	40 nm	32 nm SOI	40 nm

* Scaled to 30 GHz carrier
** Extracted from published results
*** FoM = 10*log(σ^2* P/(1mW))
[a] Fractional mode

Figure 16.7.6: Comparison table.

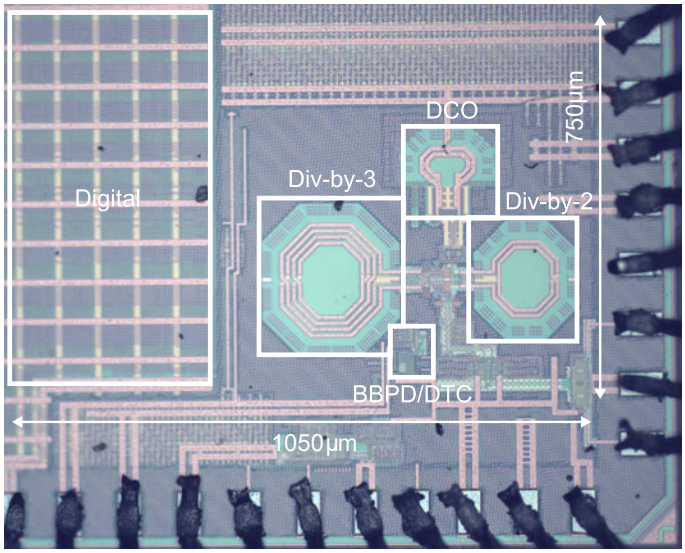


Figure 16.7.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

16.8 A 25.4-to-29.5GHz 10.2mW Isolated Sub-Sampling PLL Achieving -252.9dB Jitter-Power FoM and -63dBc Reference Spur

Zunsong Yang[1], Yong Chen[1], Shiheng Yang[1], Pui-In Mak[1], Rui P. Martins[1,2]

[1]University of Macau, Macau, China
[2]Instituto Superior Tecnico/University of Lisboa, Lisbon, Portugal

Recent mm-wave PLLs have explored different architectures to enhance their jitter performance at low power. Without noisy loop components, the injection-locked PLL in [1] using a GHz reference (REF=2.25GHz) can effectively suppress the integrated jitter (86fs$_{rms}$), resulting in a better jitter-power FoM (-247.2dB). Yet, high-frequency REF injection leads to large spur (-32dBc), entailing continuous frequency tracking to withstand the PVT variations. Also, at the system level, the GHz REF has to be generated on-chip (i.e. cascaded PLLs). The power overhead, e.g., additional 20mW in [2], and unwanted coupling between the two VCOs become inevitable. To this end, direct-synthesis mm-wave PLLs using a MHz REF are of higher interest, despite the challenge of a large division ratio (N). An example is a Type-II mm-wave PLL reported in [3] that achieves 115fs$_{rms}$ integrated jitter, but the involved divider, charge pump (CP), and VCO totally draw 31mW to suppress the in-band and out-of-band phase noise (PN).

The sub-sampling PLL in [4] using a sub-sampling phase detector (SSPD) also holds promise for improving the jitter-power FoM, as it eliminates the divider noise and removes the N^2 noise amplification of the SSPD, CP, and loop filter (LF). Still, with the direct sub-sampling and pulse-controlled gain-reduction techniques, both [4] at RF and [5] at mm-wave suffer from a tight trade-off between low spurs and low in-band PN.

This paper reports an isolated-sub-sampling PLL (iSS-PLL) dissipating only 10.2mW at 26.4GHz. When locked to a 103MHz REF, low integrated jitter (71fs$_{rms}$) is achieved without compromising the spur (-63dBc). The jitter-power FoM (-252.9dB) is at least 5.7dB better than the recent art [1-3,5]. These results are enabled by: 1) minimizing the disturbances at the sensing and control sides of the VCO by using a master-slave isolated SSPD (iSSPD) plus a V/I converter, and 2) optimizing the inter-stage sampling capacitor and amplitude of the iSSPD to suppress the in-band PN.

For a conventional SSPD (Fig. 16.8.1, upper), the PMOS-based track-and-hold (T&H) has to interface directly with the VCO tank, causing a spur due to: 1) periodic tank-capacitance variation induced by the sampling capacitance (C_s), alike the binary-frequency-shift-keying (BFSK) modulation; 2) charge injection to the VCO tank from the switch at the sampling edge (SE), and 3) charge sharing between the VCO tank and C_s at the tracking edge (TE). The BFSK effect leads to an amplitude/frequency-modulated spur proportional to 20log(C_s), opposing to the in-band PN ($L_{in\text{-}band,SSPD}$) that is inversely proportional to C_s.

The proposed T&H (Fig. 16.8.1, lower) manages to prevent the BFSK modulation and suppress the charges kickback to the VCO tank. Specifically, the NMOS (M_x: 2/0.1μm) mainly allows the VCO tank to see its gate capacitance (C_g) as a constant load, eliminating the BFSK effect during each T&H operation. Our T&H only has a modulation effect at the drain node of M_x, resulting in ~180mV voltage variation together with the charges due to injection and sharing in the current branch. Such voltage variation at the VCO tank is attenuated by 34dB (from simulation), since the coupling can only happen through the small parasitic capacitance (C_{gd}=1.6fF) of M_x. Thus, the maximum voltage variation (ΔV_2 in Fig. 16.8.1) is <3.5mV at F_{REF}, much smaller than the 190mV ΔV_1 in [5]. As a result, the VCO tank sees little disturbance from the sensing side.

The REF spur can also be induced from the control side of the VCO (Fig. 16.8.2, upper). In a feedforward master-slave configuration for sub-sampling and holding [4, 5], the pulser-embedded CP with ripples on I_{CP} induces the spur. The pulser supports gain reduction by tuning the duty ratio (DR_{pul}<1). Ideally, the amplitude mismatch between the up/down current sources (I_{UP}/I_{DN}) can be eliminated at the locked point if the on-time is constant. In practice, timing alignment can hardly be achieved due to the variable on-time. Together with the possible mismatch between the PMOS and NMOS switches, it is hard to use a wide PLL bandwidth (BW) to suppress the in-band PN without causing large ripples on I_{CP}.

Our iSSPD cascades two isolated T&Hs to address the above trade-off (Fig. 16.8.2, lower). Controlled by two non-overlapping clocks ($\phi_{1,2}$), the two T&Hs operate in a master-slave manner and finally output a highly constant voltage (V_S). The V/I converter transforms two differential constant voltages from the two iSSPDs into a constant current (I_{VI}). During ϕ_2, the charge injection and sharing result in a current disturbance to I_{UP}/I_{DN}, but they can be mostly cancelled by each other in a push-pull fashion. This disturbance can also be weakened by narrowing the duty ratio of $\phi_{1,2}$, meanwhile reducing the static current consumption. By properly sizing the iSSPD, the entire gain (K_{iSSPD}) of the iSSPD, realized to be <1, avoids the use of a pulser in the CP and reduces the capacitor area in the LF. The master T&H attenuates the sampled amplitude (V_M) operating at the VCO frequency (F_{VCO}) with the gain K_M. The slave T&H centered at the REF frequency (F_{REF}) is for DC-to-DC transformation with the gain K_S, suppressing the ripples on I_{VI}.

Figure 16.8.3 depicts the block diagram of our PLL, composed of a frequency-locking loop (FLL) for coarse tuning and a Type-II iSS-PLL for fine tuning. Since the master T&H operates up to 30GHz with a rise time of ~10ps for REF, we choose C_{s1}=10fF for an output swing (V_{Buffer}=0.56V) of the VCO buffer, which uses a compact (10×10μm^2) stacked inductor to balance the tracking BW and in-band PN of the master T&H. Additionally, the output swing at V_M is set at 0.12V to reduce the disturbance kickback to the VCO tank. Without penalizing the spur, C_{s2} can be safely upsized (70fF) to compensate the shrinkage of the output swing at V_M, resulting in a lower in-band noise contribution of the slave T&H. By setting K_S≈1.6, we can realize K_{iSSPD}≈0.35 for gain reduction. A 3rd-order LF and a VCO buffer further reduce the disturbances at the control and sensing sides of the VCO, respectively.

The mm-wave LC VCO is based on a robust Class-F topology using 2nd-to-4th harmonic resonance to reduce the impulse-sensitivity-function rms value [6]. It measures an $FoM_{@1MHz}$ of 189dBc/Hz and a tuning range of 14.9% (from 25.4 to 29.5GHz). Its output swing is adequate to directly drive the two cascaded divider-by-2 circuits with no extra buffers. For compactness and robustness, the divider-by-2 circuit, based on load-modulated dynamic latches, exhibits a >100% locking range at ~25GHz.

The iSS-PLL prototyped in 65nm CMOS operates at 1V and 0.55V (for the VCO). An off-chip crystal oscillator provides the 103MHz REF. Figure 16.8.4 shows the measured PN after an on-chip divider-by-4 with the PLL running at 26.368GHz. The restored in-band PN is -112.8dBc/Hz at a 1MHz offset. The integrated jitter is 71.16fs$_{rms}$, which is insensitive to the loop BW, and its variation is <10fs$_{rms}$ among the 6 chips measured. The in-band PN is dominated by the iSSPD + V/I converter. The loop BW is regulated at ~4MHz for an optimal jitter performance, i.e. roughly equal noise contribution by the iSSPD and VCO. The REF spur (-63dBc) is insensitive to the PLL loop BW, and its variation is <3dB among the 6 chips measured (Fig. 16.8.5).

Benchmarking with the recent mm-wave PLLs (Figs. 16.8.6 and 16.8.7), our iSS-PLL shows a 5.7dB improvement of the jitter-power FoM, and an 11.8dB improvement of the FoM_t normalized to the 103MHz REF. The iSS-PLL occupies an active area of 0.24mm^2 (Fig. 16.8.7).

Acknowledgements:
The work is funded by the University of Macau - MYRG2017-00167-AMSV, and Macau FDCT - SKL Fund.

References

[1] S. Yoo et al., "A PVT-robust -39dBc 1kHz-to-100MHz Integrated-Phase-Noise 29GHz Injection-Locked Frequency Multiplier with a 600μW Frequency-Tracking Loop using the Averages of Phase Deviations for mm-Band 5G Transceivers," *ISSCC*, pp. 324-325, Feb. 2017.
[2] H. Yoon et al., "A -31dBc Integrated-Phase-Noise 29GHz Fractional-N Frequency Synthesizer Supporting Multiple Frequency Bands for Backward-Compatible 5G using a Frequency Doubler and Injection-Locked Frequency Multipliers," *ISSCC*, pp. 366-367, Feb. 2018.
[3] S. Ek et al., "A 28-nm FD-SOI 115-fs Jitter PLL-based LO system for 24-30-GHz sliding-IF 5G transceivers," *IEEE JSSC*, vol. 53, no. 7, pp. 1988-2000, July 2018.
[4] X. Gao et al., "A 2.2GHz 7.6mW Sub-Sampling PLL with -126dBc/Hz In-Band Phase Noise and 0.15ps$_{rms}$ Jitter in 0.18μm CMOS," *ISSCC*, pp. 392-393, Feb. 2009.
[5] V. Szortyka et al, "A 42mW 230fs-Jitter Sub-sampling 60GHz PLL in 40nm CMOS," *ISSCC*, pp. 366-367, Feb. 2017.
[6] H. Guo et al., "A 0.083-mm^2 25.2-to-29.5 GHz multi-LC-tank class-F_{234} VCO with a 189.6-dBc/Hz FOM," *IEEE SSCL*, vol. 1, no. 4, pp. 86-89, Apr. 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

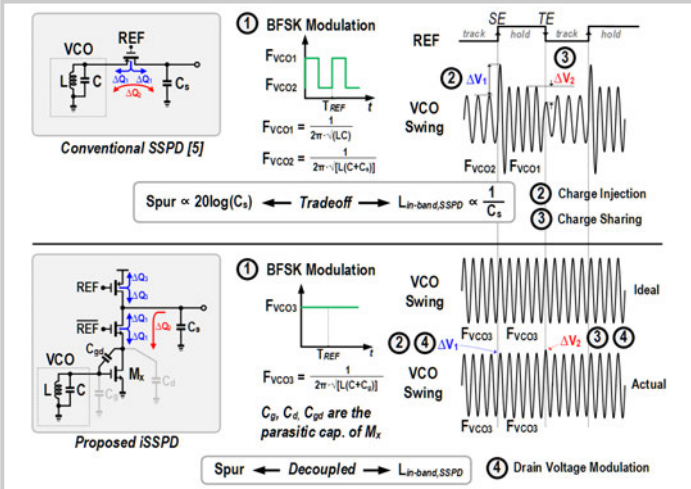


Figure 16.8.1: *Upper*: Conventional SSPD suffers from large spur due to 1) BFSK modulation, 2) charge injection and 3) charge sharing. *Lower*: Proposed iSSPD lowers the kickback charges and indirectly samples the VCO to decouple the spur with the in-band PN.

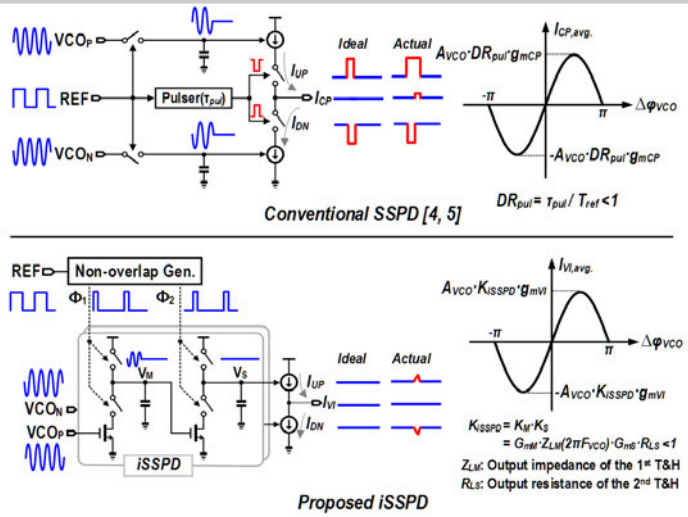


Figure 16.8.2: *Upper*: Conventional SSPD with a pulser-embedded CP induces spur due to the duty-ratio and switch mismatches. *Lower*: Proposed iSSPD with inherent gain reduction offers a highly constant V_s for steady current conduction.

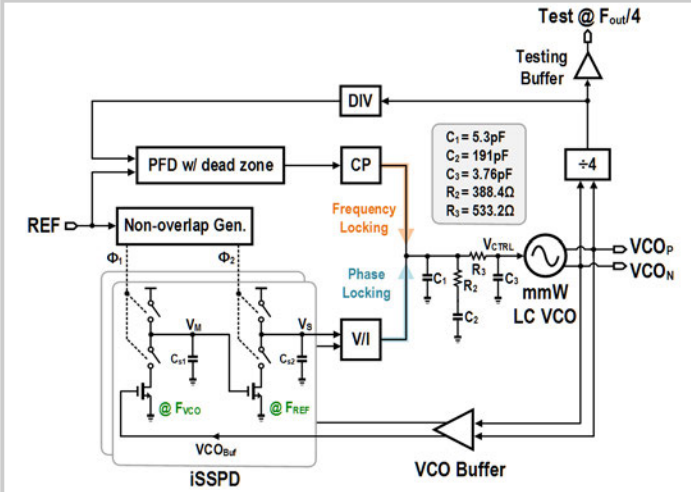


Figure 16.8.3: Block diagram of the proposed iSS-PLL. The iSSPD and V/I converter reduce the disturbance from the control side of the VCO, alleviating the trade-off between spur and PN. The VCO buffer enhances the sensing-side isolation between iSSPD and VCO.

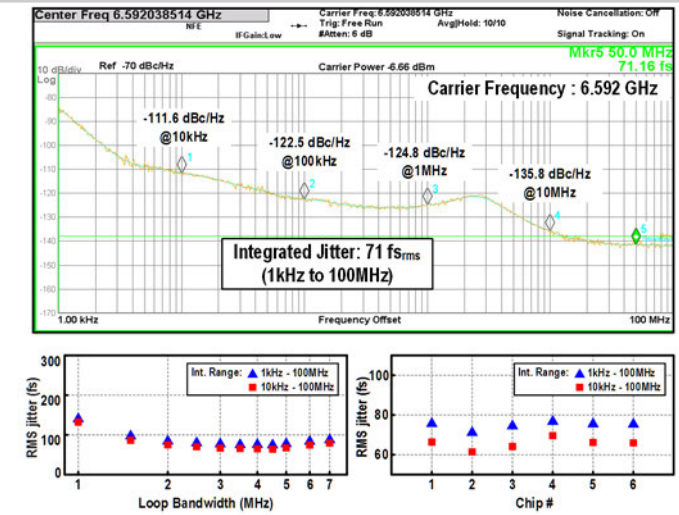


Figure 16.8.4: Measured PN after an on-chip divider-by-4 (i.e. 26.368GHz PLL frequency). The rms jitter is insensitive to the loop BW and its variation is <10fs$_{rms}$ among 6 chips.

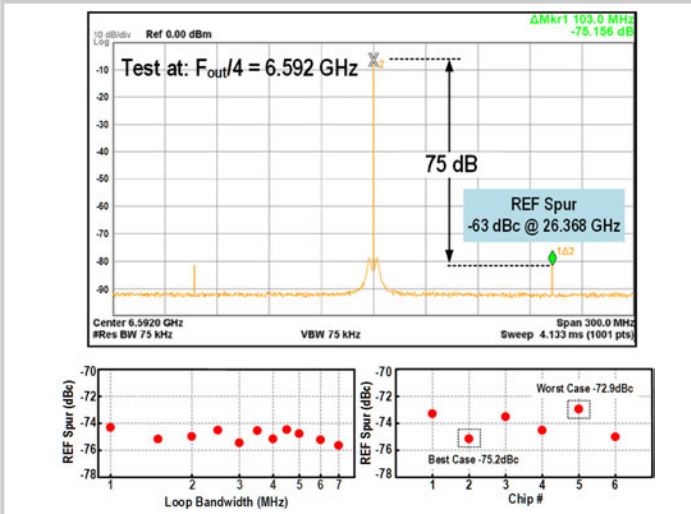


Figure 16.8.5: Measured spur after an on-chip divider-by-4 (i.e. 75dB-12dB = 63dB). The REF spur is insensitive to the loop BW and its variation is <3dB among 6 chips.

	This Work	[3] JSCC'18	[2] ISSCC'18	[1] ISSCC'17	[5] ISSCC'14
CMOS Technology (nm)	65	28	65	65	40
Key PLL Techniques	iSSPD PLL	Analog Type-II PLL	Cascaded + ILCM PLL	Injection-Locked PLL	SSPD PLL
Supply Voltage (V)	1 (0.55 for VCO)	1.2	N/A	N/A	0.9/1
Freq. Range (GHz)	25.4 to 29.5 (14.9%)	23.3 to 30.2 (25.8%)	25 to 30 (18.2%)	27.4 to 30.8 (11.7%)	53.8 to 63.3 (16.2%)
Ref Freq. (MHz)	103	491.5	120	2250	40
Output Integrated Jitter ($f_{s_{rms}}$)	71 * @26.368GHz (1k-100MHz)	114 @25.95GHz (10k-40MHz)	206 @29.22GHz (1k-100MHz)	86 @29.25GHz (1k-100MHz)	214 @62.64GHz (1k-100MHz)
PN @ 1MHz (dBc/Hz)	-112.8 *	-104	-101.9	-115.6	-90
Total Power (mW)	10.2 ^	31	36.4	24.3	42
FOM (dB)	-252.9	-244	-235.1	-247.2	-237
FOM$_r$ (dB)	-252.9	-237.2	-232.1	-233.8	-241.1
Power Eff. (mW/GHz)	0.39	1.19	0.70	0.83	0.67
Ref. Spur (dBc)	-63 *	-65	-83 #	-32.6	-40.2
Active Area (mm^2)	0.24	0.11	0.95	0.11	0.16

$\mathrm{FOM} = 10\log\left[\left(\frac{\sigma_{\mathrm{rms}}}{1\,\mathrm{sec}}\right)^2 \cdot \frac{\mathrm{Power}}{1\,\mathrm{mW}}\right]$ $\mathrm{FOM_r} = 10\log\left[\left(\frac{\sigma_{\mathrm{rms}}}{1\,\mathrm{sec}}\right)^2 \cdot \frac{\mathrm{Power}}{1\,\mathrm{mW}} \cdot \frac{f_{\mathrm{REF}}}{103\,\mathrm{MHz}}\right]$, normalized to $f_{\mathrm{REF}} = 103$ MHz

* Restored from the measured output at 6.592GHz with on-chip divider-by-4 # No reported injection spur

^ Power breakdown (VCO: 6.9mW, VCO Buffer: 0.56mW, FLL: 2.15mW, iSSPD + CLK Gen. + V/I: 0.58mW), excluding the Reference Buffer of 5.08mW.

Figure 16.8.6: Chip summary and comparison with the prior art.

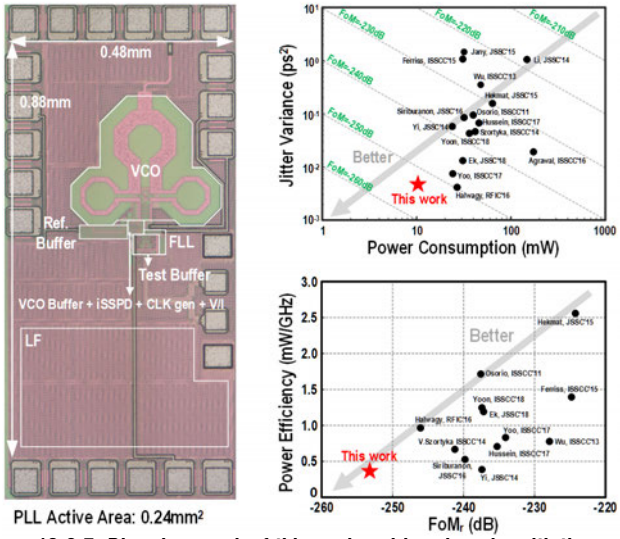


Figure 16.8.7: Die micrograph of this work and benchmarks with the recent PLLs operating at ≥25GHz.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

16.9 4.48GHz 0.18μm SiGe BiCMOS Exact-Frequency Fractional-N Frequency Synthesizer with Spurious-Tone Suppression Yielding a -80dBc In-Band Fractional Spur

Michael Peter Kennedy[1,2], Yann Donnelly[1,3], James Breslin[4], Stefano Tulisi[4], Sanganagouda Patil[4], Ciarán Curtin[4], Stephen Brookes[4], Brian Shelly[4], Patrick Griffin[4], Michael Keaveney[4]

[1]University College Dublin, Dublin, Ireland
[2]Microelectronic Circuits Centre Ireland, Dublin, Ireland
[3]Microelectronic Circuits Centre Ireland, Cork, Ireland
[4]Analog Devices, Limerick, Ireland

The instantaneous divide value of the multimodulus divider in the feedback path of a fractional-N PLL is determined by a divider controller, which is usually implemented as a digital delta-sigma modulator (DΔΣM). A disadvantage of the fractional-N PLL is the presence of fractional spurs, which result from interaction between the signal introduced by the DΔΣM and nonlinearities in the loop. When fractional spurs at frequencies close to integer boundaries lie inside the loop bandwidth, they cannot be attenuated by filtering. The Successive Requantizer (SR) is an alternative to the DΔΣM-based divider controller, which randomizes the quantization process more effectively than a DΔΣM [1]. Wang et al. reported a worst-case in-band fractional spur of -64dBc in a 2.4GHz charge-pump PLL [1]. Liang and Wang reported a -70dBc worst-case fractional spur in a 2GHz analog PLL with a hybrid VCO and a DΔΣM-based divider controller [2]. Familier and Galton improved the performance of the SR by implementing higher-order noise shaping. They achieved a worst-case fractional spur of -72dBc in a 3.3GHz analog PLL with a third-order SR [3]. The SR quantizes the frequency-control word one bit at a time, and, therefore, requires n stages in the case of an n-bit modulus. Thirunarayanan et al. implemented a hybrid MASH-SR divider-controller structure using four SR quantization blocks [4]. The divider-controller architecture described in this paper enables a 4.48GHz analog PLL to exhibit an in-band fractional spur of -80dBc and a -145dBc reference spur. It comprises a conventional MASH DΔΣM followed by a programmable Probability Mass Redistributor (PMR). The PMR requantizes the output of the DΔΣM and redistributes its samples in such a way that the in-band spurs are reduced by 7dB compared to the DΔΣM alone.

The test vehicle for evaluating the performance of the MASH-PMR hybrid divider controller is an otherwise conventional Type-II charge-pump fractional-N frequency synthesizer implemented in a 0.18μm SiGe BiCMOS technology. The tri-state PFD, CP, loop filter, VCO, and divider of the fractional-*N* frequency synthesizer shown in Fig. 16.9.1 are similar to those described in [5]. The multi-core VCO covers the 4-to-8GHz range. With 2× and 4× frequency multipliers, dividers by factors 2 to 64, and a mixed-radix MASH 1-1-1 DΔΣM divider controller, the synthesizer is capable of exact frequency synthesis from 62.5MHz to 32GHz. Bipolar bleed current can be used to bias the operating point of the PFD-CP loop away from zero, thereby improving the linearity of the PFD/CP. Two variants of the divider controller have been implemented: (i) a conventional mixed-radix MASH 1-1-1 DΔΣM and (ii) a mixed-radix MASH-PMR hybrid. Figure 16.9.2 shows the detail of the divider controller. The output frequency is $f_{out}=(N_{int}+(x_1+x_2/M_2)/M_1)f_{PFD}$, where M_1 and M_2 are the moduli of the main and auxiliary modulators (labelled MASH 1-1-1 and DDSM). The EN signal switches between two cases. The first variant is a conventional MASH 1-1-1 DΔΣM with a 25b word ($M_1=2^{25}$). The auxiliary first-order modulator has a 14b programmable modulus M_2. A one-bit LFSR signal is added to the input of the second error feedback modulator to provide first-order shaped dither. The MASH is initiated from an odd initial condition to maximize its cycle lengths. The input word *u* is applied directly to the MASH and its output *v* is fed directly to *y*. In the second variant, the bottom 24 bits of *u* are multiplied by 2, fed to the input of the MASH and then added to the MSB of *u* to form *w*. This signal *w* is subsequently processed by the PMR, whose redistributed output *z* is passed to *y*.

In the PMR, a random signal *s* is added to *w* and the result divided by 2 to yield *z*. As in an SR quantization block, $z=(w+s)/2$. A lossy mapping transforms a 4b pseudo-random input stream into a signal with a tailored probability mass function and a flat frequency spectrum. The mapping is dictated by a fully programmable 5×4 lookup table, which allows for considerable flexibility in the statistics of the modulator output. In one configuration, the PMR can emulate a single SR quantization block [3]. The spectral qualities of the MASH are preserved by imposing noise shaping on the mapped signal in order to produce a random signal, *s*, which is noise-shaped. Noise shaping is carried out using digital difference blocks. Forcing the accumulated sums $r[k]=s[0]+\ldots+s[k]$, $q[k]=r[0]+\ldots+r[k]$, $p[k]=q[0]+\ldots+q[k]$ to be bounded causes *s* to be first-, second-, or third-order shaped, respectively. Previous values of the mapped output influence the current mapping. This accounts for the effects of noise shaping on the statistics of the mapped output.

Figure 16.9.3 shows the representative measured spectra of the synthesizer with MASH and MASH-PMR divider controllers. The statistical properties of the output *y* are different for the MASH-PMR hybrid and the MASH, yielding different spur and noise-floor amplitudes for each modulator. In the case shown, the in-band noise increases marginally from -98dBc/Hz to -96dBc/Hz; the difference is negligible out of band. In the fractional-N mode, the synthesizer exhibits in-band spurs. These can be reduced by biasing the CP in a more linear region of operation. Figure 16.9.4 shows the spectrum at $N_{int}=73$, $x_1=1500$, $x_2=1$, $M_2=8191$, $M_1=2^{25}$. The reference frequency is 61.44MHz and the loop bandwidth is 100kHz. The in-band fractional spur with the MASH alone is at -73dBc. When the MASH-PMR divider controller is used, the amplitude of the in-band spur is reduced by 7dB to -80dBc. Figure 16.9.5 shows plots of the largest measured fractional spurs for a sweep of 67 PLL frequency offsets in the range from 1kHz to 10MHz. Manual adjustment of the operating point by applying bleed current reduces the worst-case spur from -47 to -67dBc. The second- and third-order MASH-PMR hybrids further reduce the worst-case in-band spurs to -70dBc and -72dBc, respectively. Note that the MASH-PMR does not need the bleed current. Indeed, worst-case in-band spurs of -71dBc are recorded with no bleed current. This is because the MASH-PMR successfully suppresses in-band spurs in the region of even the worst PFD/CP nonlinearity.

Figure 16.9.6 compares the performance of the MASH-PMR-based synthesizer with prior art. The test chip, shown in Fig. 16.9.7, is relatively large and power hungry due to features that are not directly relevant to this work, namely the oscillators, multipliers, dividers, and output stages, which can source 4dBm. We focus on the MASH 1-1-1, which occupies 0.097mm². The PMR requires a further 0.021mm². This is negligible compared to the area of the VCO and multipliers. Familier and Galton have reported a worst-case in-band spur of -72dBc at 3.35GHz and a -79dBc reference spur with a third-order 16-stage SR divider controller. This work reports a -72dBc fractional spur at 4.485GHz and -110dBc reference with mixed-radix exact-frequency capability. Normalizing for frequency, this represents an improvement of 2.5dB in the in-band spur performance. While [1-3] present sweeps of worst-case in-band fractional spurs, Yao et al. reported a single measurement of a -78.6dBc fractional spur and a -87.6dBc reference spur in an ADPLL with a DΔΣM-based divider controller operating at 2.69GHz [6]. By comparison, this work demonstrates a -80dBc fractional spur at 4.485GHz, which is 5.8dB better than [6] when normalized to the VCO frequency.

By using a PMR to redistribute the output samples of a conventional MASH1-1-1 DΔΣM, this work demonstrates an in-band fractional spur of -80dBc in a 4.485GHz CP-PLL. The MASH-PMR hybrid divider controller can be used as a drop-in replacement for a conventional MASH. The extra area requirement for the hybrid is 22% of the MASH alone. The MASH-PMR can be used in addition to or instead of conventional linearization strategies, such as bleed current.

Acknowledgements:
This work has been funded in part by Science Foundation Ireland under grants 13/IA/I979 and 13/RC/2077 and by the Irish Research Council under grant GOIPG/2014/14222.

References:
[1] K. Wang et al., "Spurious-Tone Suppression Techniques Applied to a Wide-Bandwidth 2.4GHz Fractional-N PLL," *ISSCC*, pp. 342-343, Feb. 2008.
[2] C.-F. Liang and P.-Y. Wang, "A Wideband Fractional-N Ring PLL Using a Near-Ground Pre-Distorted Switched-Capacitor Loop Filter," *ISSCC*, pp. 190-191, Feb. 2015.
[3] E. Familier and I. Galton, "Second and Third-Order Successive Requantizers for Spurious Tone Reduction in Low-Noise Fractional-N PLLs," *IEEE CICC*, pp.1-4, May 2017.
[4] R. Thirunarayanan et al., "A 51.4Mb/s FSK Transmitter Employing a Phase Domain Digital Synthesizer with 1.5μs Start-up for Energy Efficient Duty Cycling," *ESSCIRC*, pp. 129-132, Sept. 2008.
[5] M. Kennedy et al., "A 0.3-4.3GHz Frequency-Accurate Fractional-N Frequency Synthesizer with Integrated VCO and Nested Mixed-Radix Digital Δ-Σ Modulator-Based Divider Controller," *IEEE JSSC*, vol. 49, no. 7, pp. 1595-1605, July 2014.
[6] C.-W. Yao et al., "A 14nm Fractional-N Digital PLL with 0.14ps$_{rms}$ Jitter and -78dBc Fractional Spur for Cellular RFICs," *ISSCC*, pp. 422-423, Feb. 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

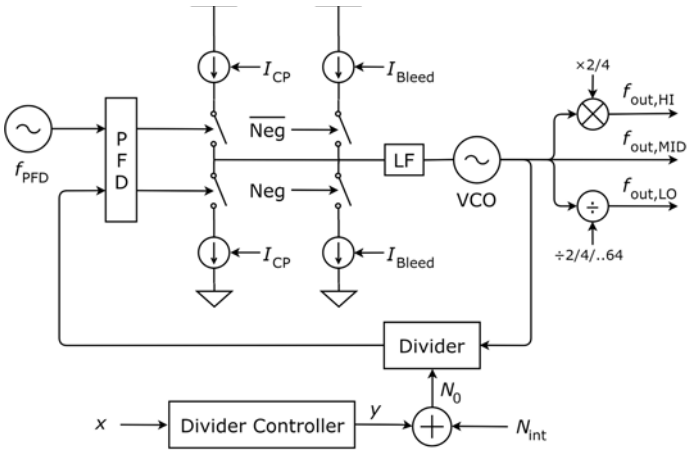


Figure 16.9.1: Block diagram of fractional-N frequency synthesizer.

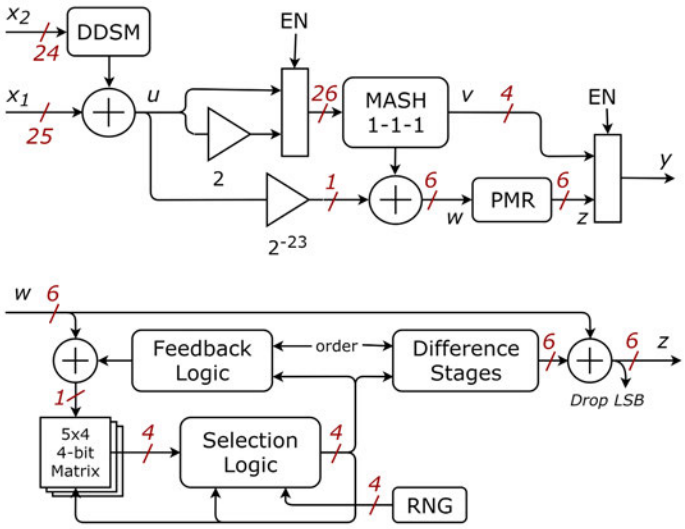


Figure 16.9.2: Block diagram of the exact frequency-divider controller and the detail of PMR.

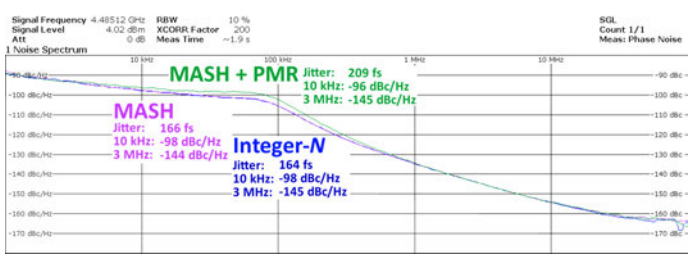


Figure 16.9.3: Representative measured phase noise of the synthesizer output signal.

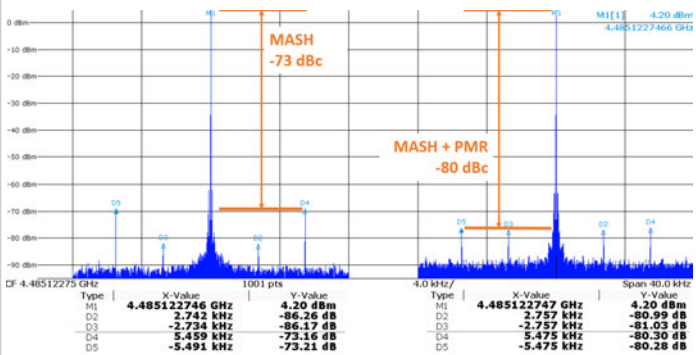


Figure 16.9.4: Measured in-band fractional spurs with MASH and MASH-PMR divider controllers.

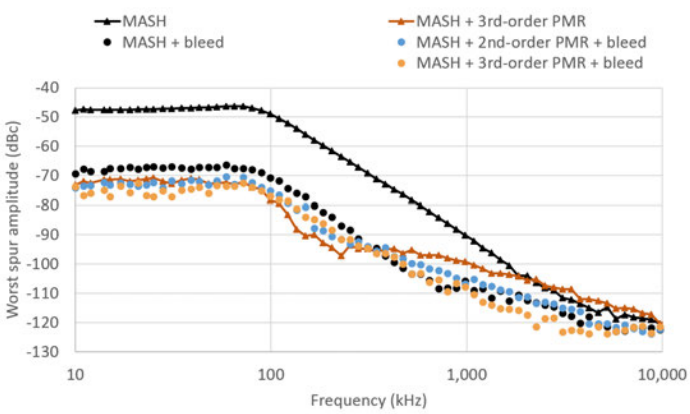


Figure 16.9.5: Largest measured fractional spurs for 67 frequency offsets with MASH and MASH-PMR controllers, with and without manual bleed current calibration.

Reference	[1]	[2]	[6]	[3]	This work
Technology [nm]	180	65	14	65	180
Supply (V)	1.8	1.0/2.0	N/A	1.0/2.0	1.2/3.3/5.0
Power Consumption (mW)	36.9	9.1	13.4	19.52	118
Area (mm^2)	4.84	0.046	0.257	0.341	13.2
Exact frequency capability	No	No	No	No	Yes
f_{vco} [MHz]	2436	2002	2690	3350	4485
f_{PFD} [MHz]	12	26	26	26	61.44
Loop Bandwidth [kHz]	975	1500	N/A	48	100
In-band Phase Noise [dBc/Hz]	-98 @ 100kHz	-91 @ 5kHz	**-113.6 @ 100kHz**	-87.5 @ 10kHz	-98.4 @ 50kHz
In-band Phase Noise FoM* [dBc/Hz]	-92.7	-84.0	**-109.2**	-85.0	-98.4
Out-of-band PN [dBc/Hz @ 3MHz]	-121	-105	-138	-137	**-145**
Largest in-band fractional spur (sweep) [dBc]	-64	-70	**N/A -78.6****	-72	**-72 -80****
Largest in-band fractional spur (sweep) FoM* [dBc]	-58.7	-63	**N/A -74.2****	-69.5	**-72 -80****
Reference Spur [dBc]	-70	-87	-87.6	-79	**-110**
FOM	-214.9	-202.9	**-228**	-203.9	-211.2
Architecture	CP-PLL	CP-PLL	ADPLL	CP-PLL	CP-PLL

Figure 16.9.6: Performance summary. These values are derived from other measurements. *Normalized to 4.485GHz output frequency. **One-off measurement.

16

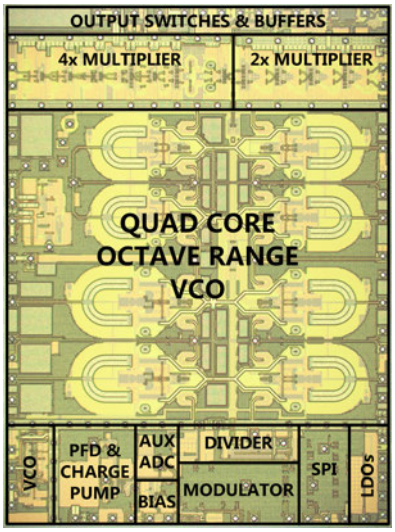


Figure 16.9.7: Die micrograph of the fractional-N frequency synthesizer. Die size: 13.2mm^2.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 17 Overview:
Technologies for Human Interaction & Health
TECHNOLOGY DIRECTIONS SUBCOMMITTEE

Session Chair: ***Patrick Mercier***
University of California,
San Diego, La Jolla, CA

Associate Chair: ***Shuichi Nagai***
Panasonic, Moriguchi, Osaka, Japan

Subcommittee Chair: Makoto Nagata, Kobe University, Kobe, Japan

Innovations in sensors, actuators, computation, and communication circuits are enabling new and exciting applications in human interaction and healthcare. This session begins with an invited system paper that describes how advances in artificial intelligence and processors, networks, and actuators have enabled new classes of human-interacting robots. The next three papers then describe innovations in voice-activity detection, large-scale tactile sensing, and low-power embedded motion detection systems for human-interactive applications. Subsequent papers introduce ultra-low-power body-area networking circuits, dual optical/electrical tomographic imaging circuits, and ultra-low-power PPG sensing for health-related applications.

INVITED PAPER **1:30 PM**

17.1 AI x Robotics: Technology Challenges and Opportunities in Sensors,Actuators, and Integrated Circuits
M. Fujita, Sony, Tokyo, Japan
In Invited Paper 17.1, Sony presents key technological developments in processors, sensors, actuators, and wireless communications towards innovative new applications in AI and robotics. These are combined in a demonstration robotic platform called, aibo, which is a fully autonomous entertainment robot embedded with AI-based behavior control.

2:00 PM

17.2 A 142nW Voice and Acoustic Activity Detection Chip for mm-Scale Sensor Nodes Using Time-Interleaved Mixer-Based Frequency Scanning
M. Cho, University of Michigan, Ann Arbor, MI
In Paper 17.2, the University of Michigan demonstrates a voice and acoustic activity detector using mixer-based architecture and an ultra-low power neural network-based classifier. Due to 4× and 12× power reduction in feature extraction and neural processing, a 91.5%/90% speech/non-speech hit rate at 10dB SNR with power consumption of only 142nW is achieved.

2:30 PM

17.3 Hybrid System for Efficient LAE-CMOS Interfacing in Large-Scale Tactile-Sensing Skins via TFT-Based Compressed Sensing
L. E. Aygun, Princeton University, Princeton, NJ
In Paper 17.3, Princeton University introduces an architecture for high-spatial-resolution tactile-sensing skins based on LAE-CMOS hybrid systems with compressed sensing. By readout and control with a resistive force-sensing array using ZnO TFTs, an acquisition error of $0.7k\Omega_{rms}$ over a 100kΩ-to-15kΩ sensing range, at energy and rate of 1.2μJ/frame and 31fps is realized.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

3:15 PM

17.4 16MHz FRAM Micro-Controller with a Low-Cost Sub-1μA Embedded Piezo-Electric Strain Sensor for ULP Motion Detection

S. Khanna, Texas Instruments, Dallas, TX

In Paper 17.4, Texas Instruments presents a 0.13μm 16MHz FRAM micro-controller with an embedded piezoelectric strain sensor using an array of thin film Lead Zirconium Titanate (PZT) capacitors for motion detection. The sensor requires only 650nA and occupies an area of 0.6mm^2.

3:30 PM

17.5 A 0.8mm^3 Ultrasonic Implantable Wireless Neural Recording System with Linear AM Backscattering

M. M. Ghanbari, University of California, Berkeley, CA

In Paper 17.5, the University of California, Berkeley describes a 0.8mm^3 neural recording system that is powered by and communicates via ultrasound. A linear AM backscatter technique is introduced that enables operation at a depth of up to 5cm while operating a small power-efficient LNA with –44dB THD over a 20mV$_{pp}$ input range.

3:45 PM

17.6 A Sub-40μW 5Mb/s Magnetic Human Body Communication Transceiver Demonstrating Trans-Body Delivery of High-Fidelity Audio to a Wearable In-Ear Headphone

J. Park, University of California, San Diego, La Jolla, CA

In Paper 17.6, the University of California, San Diego presents a transceiver design to exploit the low path loss enabled by magnetic human body communication (mHBC). A test chip demonstrates reliable trans-body communication at 5Mb/s at TX and RX power consumptions of 18.56 and 6.3μW, respectively.

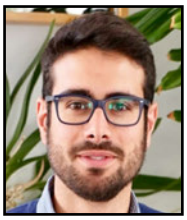
4:15 PM

17.7 A 7.0fps Optical and Electrical Dual Tomographic Imaging SoC for Skin-Disease Diagnosis System

Y. Lee, KAIST, Daejeon, Korea

In Paper 17.7, KAIST describes a portable lightweight (3g) system that can simultaneously image the skin in the electrical and optical domains to help improve skin disease diagnostics. Incorporation of diffused optical and electrical impedance tomography enables achieving 0.5mm sensitivity and 2mm selectivity.

4:45 PM



17.8 A 2.6μW Monolithic CMOS Photoplethysmographic Sensor Operating with 2μW LED Power

A. Caizzone, EPFL, Neuchâtel, Switzerland

In Paper 17.8, EPFL presents a photophlethysmography (PPG) sensor that utilizes an array of dedicated pinned-photodiodes in a 0.18μm CIS process. Thanks to improved sensitivity offered by the diodes, the total power of the PPG sensor is 4.6μW.

17.1 AI x Robotics: Technology Challenges and Opportunities in Sensors, Actuators, and Integrated Circuits

Masahiro Fujita

Sony, Tokyo, Japan

In 1956 at the Dartmouth conference, the terminology of artificial intelligence (AI) was first used. In those days it was also called as symbolic AI (Symbolic-AI) [1]. For example, let us assume a block world problem in Fig 17.1.1 (right), where a block is represented as a symbol, "Block1", and it can be operated by operators such as "PICKUP(Block1). In order to apply the operator "PICKUP" to the target object, "Block1", the AI system has to check the pre-condition such as "CLEAR Block1", which means there is no object on "Block1". The Fig.17.1.1 (right) shows an example of a task from State-A to State-B. The system has to search the possible operators and the pre-conditions so that State-B is achieved. There are many basic algorithms developed in Symbolic-AI era, which are often used today including the A*-search algorithm. Shakey is the representative example of intelligent robots based on Symbolic-AI. It was a wheel-based movable robot equipped with a TV-camera, Laser-Range-Finder, etc. It can move blocks in the real world using Symbolic-AI technologies. Its behavior control architecture is shown in Fig 17.1.1 (left). It has three steps, SENSE, PLAN, and ACT. Therefore, it is known as the SENSE-PLAN-ACT architecture. It is computationally intensive especially in the PLAN computation, therefore it is difficult if the environment is dynamically changing.

To solve the issue of Symbolic-AI, in the 1980s, Embodied-AI, or Behavior-Based-AI, was developed [2]. Its behavior control architecture is shown in Fig 17.1.2 (left). Compared with Symbolic-AI, it is composed of multiple layers, and basically there is no PLAN step. Genghis was the representative of Embodied-AI, which looked like a 6-legged insect, and had two feelers [2], It didn't symbolize the environment, nor search in the symbolic world, but sensed the environment and reacted according to sensing information. Thus, it was also called SENSE-ACT architecture compared with SENSE-PLAN-ACT architecture. For example, in the Avoiding Action layer, if the left feeler becomes on, then turn-right action is activated so that it can avoid the obstacles on the left. In the Wondering Action layer, if there are no stimuli of the feelers, then the go-forward action is activated and after a while the random-turn action is activated. In the Exploring Action layer, using the distance sensor such as an ultra-sonic sensor, it activates the wall-following action so that it explores the world by move along walls, which is considered a good strategy for exploration in a maze. Avoiding Action's priority is higher than Wondering Actions, and Wondering Action's priority is higher than Exploring Action. Thus, according to the complex environment, Genghis shows complex behaviors without planning. However, it is difficult to achieve intelligent behaviors using such as planning.

In the 1990s, Hybrid-AI was developed that integrated both of Symbolic-AI and Embodied-AI [2]. Hybrid-AI's behavior control architecture is shown in Fig 17.1.2 (right). There are three pathways in the behavior control architecture. One is the symbol representation pathway to the deliberative behavior layer, the second is the situated base pathway to situated behavior layer, and the third is a quick response pathway to reflexive behavior layer. In the deliberative layer, the Symbolic-AI such as planner is executed, and in the situated behavior layer, many behavior modules are evaluated and coordinated to execute proper behavior according to the situation. In the reflexive behavior layer the action is directly activated as a reaction of external stimuli in an emergent situation such as avoiding a collision. In the situated behavior layer the coordination of behaviors is often a selection of the most proper one, or mixed if the robot can do those in parallel. The behaviors from deliberative behavior layer are usually to give bias signals to the behaviors in the situated behavior layers.

The behavior architecture of an entertainment robot AIBO released in 1999 was based on Hybrid-AI, but in addition it has internal states so-called pseudo instinct [4]. AIBO's behavior control architecture is shown in Fig 17.1.3. In AIBO it evaluates the behaviors based on both the external states and the internal states. The dynamics of the internal states is 1) decreasing as time proceed, 2) increasing when related behavior to the internal state is executed successfully. The evaluation score of the behavior is high when the related internal state is out of a set range. For example, if the internal state named exercise-instinct is bellow the low limitation of the range, the evaluation score of behaviors that can increase the exercise-instinct such as ball-kicking behavior becomes high. In this case, if a ball is found near AIBO, which means the external states also support the evaluation score of ball-kicking behavior. Then, if ball-kicking behavior is executed to increase the exercise-instinct and to be in the range. Thus, it is a homeostasis system and it is a base mechanism that AIBO can execute spontaneous behaviors without any commands from users.

Let us compare AIBO with a new aibo released in 2018. Hereafter we distinguish the two robots using AIBO released in 1999 and aibo released in 2018. The hardware architectures of AIBO and aibo are shown in Fig 17.1.4 (left) and Fig 17.1.4 (right). Regarding the processor, AIBO used a RISC architecture, a MIPS R400x with about 300MHz clock. There is a companion chip with functions, 1) image filter bank for pyramidal image generator, 2) object detection based on 8 colors, 3) Inner product engine, and 4) Serial bus manager for devices. The combination of the R400x and the companion chip makes it possible to process images and other sensors in real time. On the other hand, aibo uses a 2.4GHz Snapdragon SoC found in smartphones. The SoC has 4 ARM CPU cores, DSPs, GPUs, etc, and using only the SoC it can process multiple imagers and other sensors in real time. Regarding the wireless LAN (WLAN), AIBO uses 802.11b (22.2Mbps max), but aibo has 802.11n (77.2Mpbs max), and cell-phone network LTE(150Mbps max). One of the big advantages of aibo is in the progress of the wireless communication environment enabling software updates that can be done through the cloud. In the AIBO case, it should be done using a physical memory card. In addition, the power of cloud computing and storage is very useful for the embedded system to off-load computation tasks such as learning, to keep important state information of aibo, etc. In addition, the wide availability of smartphone helps the user to control and monitor the aibo using a graphical user interface (GUI).

More details of sensors, actuators, effectors, and AI processing of aibo are shown in Fig 17.1.5. and Fig 17.1.6. Two back-illuminated(BI) -CMOS sensors are used in aibo. One CMOS sensor is at the top of the nose, and its image signal is used to detect human faces, hands, legs, and torsos using an embedded deep neural network. In addition, aibo detects the objects that designed for aibo to play with or use, such as the pink ball, the pink born like object named ai-born, the charge station, AIBO, and aibo. Furthermore, aibo can detect general objects such as table, chair, TV, which are usually found in a living room. Another CMOS sensor is in its back. It is used for so-called simultaneous localization and mapping (SLAM), that is to build a map of a room by moving and to infer a location and posture of aibo itself.In the front face and breast, there equipped a time-of-flight (ToF) sensor for obstacle detection, pyroelectric-IR sensor for live human detection (not human in photo), and position sensing detection (PSD) sensor for near-located-obstacle and a cliff such as table edge. There are 4 microphones to estimate the sound source direction, and to reduce noise using beamforming for improving speech recognition performance. Speech recognition is processed in the SoC using a deep neural network for acoustic modeling. There are three tactile sensors in the head, chin, and back to engage the physical interaction with users such as patting for reward. Two inertia meter unit (IMU) are equipped at the head and the body for utilizing with visual SLAM, and for detecting when held up by users. There are 22 actuators to realize real dog-like motions including mouth and ears. Those actuators are controlled by a servo-control MCU in the body. Its eyes are displayed using two OLED devices. It is important for aibo to show emotions using the eye patterns such as happy, angry, stressed, sleepy, and surprise, etc.

As CPUs, networks, and wireless communications progress, so will AI and Robotics progress. Three examples are mentioned here. First one is about imager. Imagers were originally developed for taking a scene like photos and movies that are watched by humans. Now, imagers are developed for other applications such as perception and recognition by machines. For example, a BI stacked CMOS image sensor can take a video with 1000 frame/sec and can track objects with logic circuits on the back-side[6]. It is more capable in ability than human perception. This direction will be more enhanced with cloud computing. Second one is about force control. In order to realize a human-robot collaborative society, safety physical interaction is necessary. Force sensors and tactile sensors with force control is important [7]. Thirdly, multiple robot systems integrated with a cloud system are important in the creation of a human-robot collaborative society. Finally, wireless communication with wide bandwidth, low latency, and secure protocol is an important technology.

References:

[1] A.Russel and P.Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 1995.

[3] R. Murphy and R. Arkin, *Introduction to AI Robotics*, MIT Press, 2000.

[4] M. Fujita, et al., "An open architecture for robot entertainment", *Autonomous Agents*, pp. 435-442, 1997.

[5] M. Fujita, "Intelligence Dynamics: a concept and preliminary experiments for open-ended learning agents", *Journal of AAMAS*, vol. 19, no. 3, pp. 248-271, Dec. 2009.

[6] T. Yamazaki, et al., " A 1ms High-Speed Vision Chip with 3D-Stacked 140GOPS Column-Parallel PEs for Spatio-Temporal Image Processing" *ISSCC Dig. Tech Papers,* pp. 82-83, Feb. 2017.

[7] K.Nagasaka, et al.,"Whole-body Cooperative Force Control for a Two-Armed and Two-Wheeled Mobile Robot Using Generalized Inverse Dynamics and Idealized Joint Units", ICRA2010, pp. 3377-3383, 2010.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

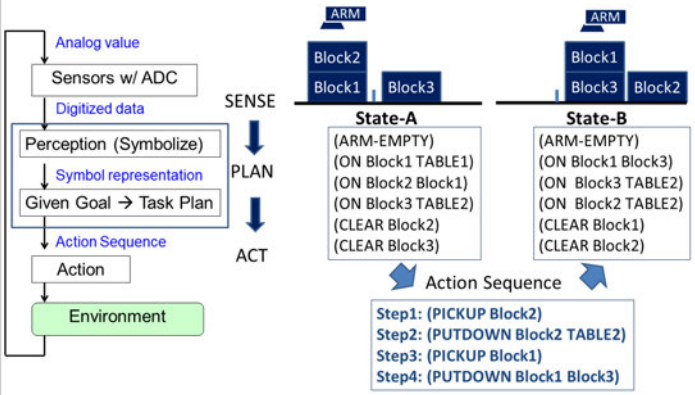


Figure 17.1.1: Symbolic-AI behavior control architecture (left) and typical representation of block world (right). Symbolic-AI is often called Sense-Plan-Act, and the Plan part is computationally intensive.

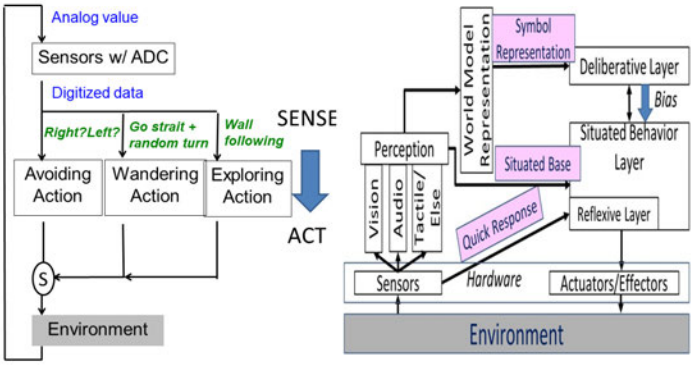


Figure 17.1.2: Embodied-AI behavior control architecture (left) and Hybrid-AI behavior control architecture (right). The notation of "S in a circle" means that actions from multiple layers are properly selected.

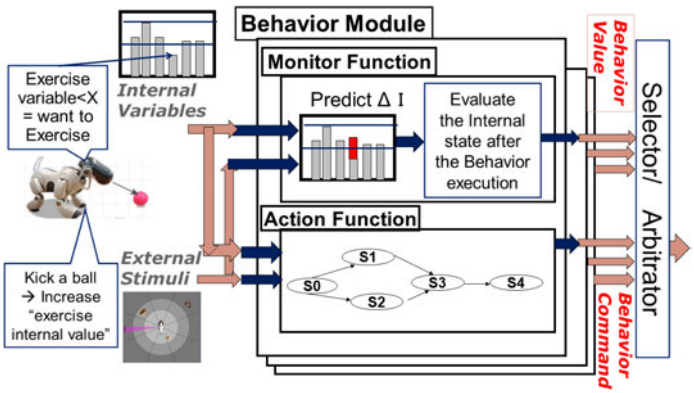


Figure 17.1.3: AIBO behavior control architecture based on Hybrid-AI (left) and action evaluation of AIBO using both external stimuli and internal stimuli (pseudo instinct) (right).

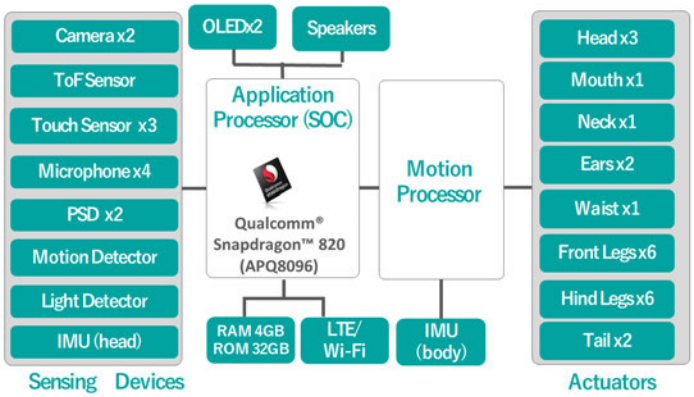


Figure 17.1.4: Hardware architecture of aibo. AIBO needed a dedicated LSI for visual processing, Inner Product Engine, etc., and multiple servo-controller for actuators. In the other hand, aibo needs one SoC and Motion Controller for all actuators.

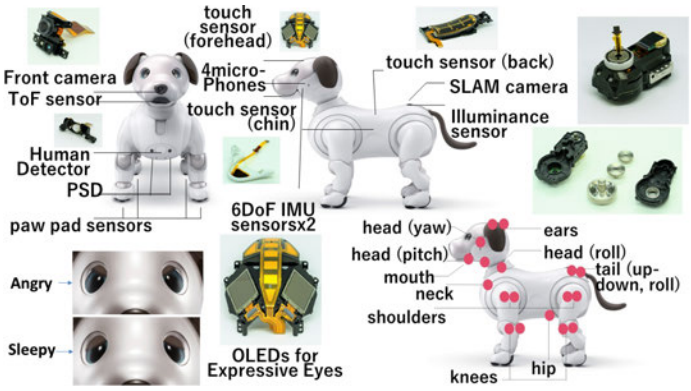


Figure 17.1.5: Major sensors and effectors of aibo.

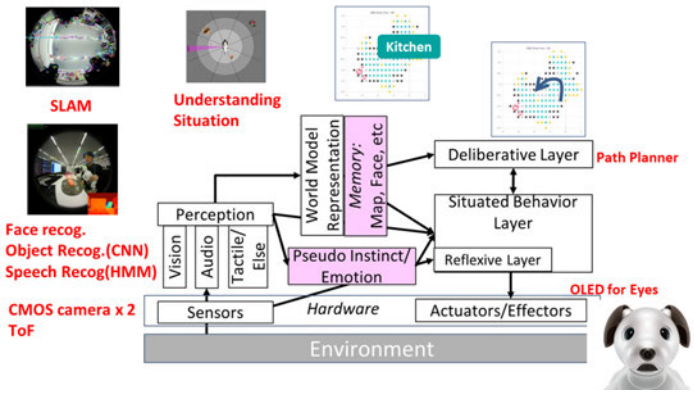


Figure 17.1.6: aibo's behavior control architecture. Left bottom picture is the output of the deep neural network for face and object detection. The top pictures from left to right are, camera output in back for SLAM, memory of map in the room with label "Kitchen" and path planner result from the current position to "Kitchen".

17.2 A 142nW Voice and Acoustic Activity Detection Chip for mm-Scale Sensor Nodes Using Time-Interleaved Mixer-Based Frequency Scanning

Minchang Cho*, Sechang Oh*, Zhan Shi, Jongyup Lim, Yejoong Kim, Seokhyeon Jeong, Yu Chen, David Blaauw, Hun-Seok Kim, Dennis Sylvester

University of Michigan, Ann Arbor, MI
*Equally-Credited Authors (ECAs)

Acoustic sensing is one of the most widely used sensing modalities to intelligently assess the environment. In particular, ultra-low power (ULP) always-on voice activity detection (VAD) is gaining attention as an enabling technology for IoT platforms. In many practical applications, acoustic events-of-interest occur infrequently. Therefore, the system power consumption is typically dominated by the always-on acoustic wakeup detector, while the remainder of the system is power-gated the vast majority of the time. A previous acoustic wakeup detector [1] consumed just 12nW but could not process voice signals (up to 4kHz bandwidth) or handle non-stationary events, which are essential qualities for a VAD. Prior VAD ICs [2,3] demonstrated reliable performance but consumed significant power (>20μW) and lacked an analog frontend (AFE), which further increases power. Recent analog-domain feature extraction-based VADs [4,5] also reported μW-level power consumption, and their simple decision tree [4] or fixed neural network-based approach [5] limited broader use for various acoustic event targets. In summary, no sub-μW VAD has been reported to date, preventing the use of VADs in unobtrusive mm-scale sensor nodes.

This work presents a 142nW programmable, neural network-based acoustic sensing system for both VAD and non-voice event detection. We use a *time-interleaved mixer-based architecture* that sequentially scans and down-converts the 4kHz bandwidth signal to a ≤500 Hz passband, reducing amplifier, ADC, and DSP power by 4×. The neural network (NN) processor employs computational sprinting, which minimizes static energy dominance in low frequency/voltage regime, providing 12× power reduction in the digital domain. The architecture (Fig. 17.2.1, top) has two signal chains: an ULP chain with 142nW consumption that is always on and a 18μW high performance (HP) chain that wakes upon event detection by the ULP chain. Unlike the ULP chain, the HP chain has a full 4kHz bandwidth AFE while sharing the same digital backend with the ULP chain. In addition to VAD, the system features an inaudible acoustic signature detection mode to enable remote silent system wakeup. With always-on VAD, the system has a 4.5-year lifetime with a 5mm mini coin-cell battery (2mAh) and achieves 91.5% voice detection accuracy.

Figure 17.2.1 shows the *time-interleaved mixer-based architecture* that reduces power consumption of AFE and DSP by lowering their bandwidth and sampling rate to 500Hz and 1kHz, respectively. The incoming signal from the microphone is amplified by an LNA with the full 4kHz bandwidth. At this point the mixer, switched by a binary discrete cosine transform (DCT) sequence, immediately down-converts the frequency of a desired feature to a programmable intermediate frequency (IF) of <500Hz. The digital binary sequence generator supports an arbitrary DCT frequency for the mixer switch control; for example, the 4kHz band can be divided into 31.25Hz frequency bins using a 128-pt DCT, and the energy content of 32 bands is sequentially extracted by sweeping DCT frequencies (F_1, … F_{32}). The 32 bands are chosen during NN training for each target event. The IF down-converted signal is further amplified and low-pass filtered with 500Hz bandwidth (via a PGA) and digitized at 1kS/s. Finally, the digital IF mixer down-converts the signal to DC, and the feature power is measured. With a DCT length of 16ms per feature (128-pt DCT with 8kHz binary mixing), 32-feature extraction requires a 512ms frame. The mixer-based structure reduces bandwidth, sampling rate, and clock frequency of AFE and DSP after the mixer; thus, the feature extraction power consumption is decreased from 225nW (simulation) to 60nW. IF is set to ~250Hz to avoid PGA 1/f noise, while the image aliasing issue of non-quadrature mixing is mitigated by a NN trained with image-aliased signals.

Figure 17.2.2 shows the circuit diagram of the AFE with ULP and HP chains. Each chain consists of an LNA, PGA, ADC driver, and ADC. Both chains share a single MEMS microphone and charge pump. For high sensitivity, a three-stage Dickson pump is used to bias the microphone at 10V. The microphone switches between the chains by controlling *ULP_CH_EN* and *HP_CH_EN*, which are level-shifted to 10V. As coupling capacitors for the level-shifting may suffer from leakage between infrequent mode switches, we refresh their charge periodically. Capacitive feedback and pseudo-resistor dc-servo loops are used for the ULP LNA (18dB gain). The LNA OTA adopts an inverter-based cascode amplifier for noise efficiency. Its common-mode feedback is composed of two loops. Coupling capacitors provides fast loop response, and the DDA output sets the DC voltage. The auxiliary amplifiers (Aux-amp) in the dc-servo loop shift common-mode voltages for high dynamic range. In addition, Aux-amp attenuates large LNA output due to diode-connected nature, which reduces the maximum amplitude seen by the pseudo-resistors, reducing their amplitude-dependent drift. The mixer is composed of transmission gates switched by the DCT sequence generator. Unlike the LNA OTA, PGA OTA uses only a PMOS input pair for the maximum output range. By tuning cap C_{I2}, the gain is adjustable between 4.5 and 31.2dB, and C_L sets 500Hz BW for ULP mode. The ADC driver is followed by an 8b SAR ADC. The HP blocks are similar to the ULP counterparts except that they are scaled for low noise and full 4kHz bandwidth. We minimize the ULP-HP transition time by temporarily turning on the fast settling switches during the transition. This helps to set the common-mode voltage very quickly (100ms vs 6s, measured).

Figure 17.2.3 (top) depicts the digital backend architecture. Always-on modules are implemented with thick oxide I/O devices to suppress leakage, while power-gated modules (NN processor, FIFO, and audio compressor [6]) are designed with standard devices. The NN processor uses a 16kB custom ultra-low retention leakage SRAM for the 4-bit weights storage, and its ISA includes matrix-vector multiplication, non-linear activation, FFT, conditional branch, element-wise vector operation, and min/max/averaging. In the ULP mode, the processor sprints at a relatively high frequency clock (700kHz, Fig. 17.2.3, bot right) when the sequential feature extraction is complete (every 512ms) and then power-gates to minimize the leakage power, resulting in 12× power reduction (sprint/sleep ratio of 0.008). In HP mode, the processor computes a full (non-sequential) FFT and a larger NN without duty cycling for the improved latency (32ms) or hit rate at the cost of higher power consumption (14μW). The binary mixer sequence generator (Fig. 17.2.3, bot left) is programmable for different DCT sizes, feature frequency resolutions, and number of features. Due to the mixer-based architecture, digital processing runs at 1kHz (vs. 8kHz Nyquist rate), yielding 41% reduction of feature extraction power.

The system also features inaudible acoustic signature detection to enable silent remote system wakeup (Fig. 17.2.4). The binary mixing sequence is replaced with a maximal length sequence (MLS) signature generated by a 1kHz programmable LFSR. Correlation between the incoming wakeup signal and the local sequence is performed through the ULP mixer and PGA. To synchronize the wakeup and local sequence, we employ a time-drift synchronization scheme that uses intentional frequency mismatch between the two so that they naturally time-synchronize periodically. This inaudible (–10dB SNR) signature detection consumes only 66nW with 4s worst case latency.

The chip is fabricated in 180nm CMOS and integrated with a MEMS microphone (Fig. 17.2.7). The ULP and HP chain amplifiers consume 31nW and 370nW with 16μVrms and 9.1μVrms input-referred noise, respectively. Figure 17.2.5 (top left) shows the measured mixer-based frequency scanning operation and input referred noise spectrum. Figure 17.2.5 (bot left) shows the measured ULP chain power breakdown. For VAD evaluation, speech from the LibriSpeech dataset is mixed with babble noise from the NOISEX-92 dataset. NN training and evaluation use exclusive datasets. Figure 17.2.6 compares the system with prior work. The system achieves 91.5%/90% speech/non-speech hit rates at 10dB SNR with babble noise (electrical test, Fig. 17.2.5 top right) in ULP mode when programmed with a NN of size 32-32-16-2 neurons, exhibiting ~7.5% better hit rate at 7× less power consumption than prior state-of-the-art. Unlike prior-art, we also report acoustic VAD test results measured in a sound chamber, showing >83%/85% speech/non-speech hit rates with a signal level down to 50dBA SPL (Fig. 17.2.5, bot right).

Acknowledgements:
This work was supported by Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References:
[1] S. Jeong, et al., "A 12nW Always-on Acoustic Sensing and Object Recognition Microsystem Using Frequency-Domain Feature Extraction and SVM Classification," *ISSCC Dig. Tech. Papers*, pp. 362-363, Feb. 2017.
[2] A. Raychowdhury, et al., "A 2.3 nJ/Frame Voice Activity Detector-Based Audio Front-End for Context-Aware System-On-Chip Applications in 32nm CMOS," *IEEE J. Solid-State Circuits*, vol. 48, no. 8, pp. 1963-1969, Aug. 2013.
[3] M. Price, et al., "A Scalable Speech Recognizer with Deep-Neural-Network Acoustic Models and Voice-Activated Power Gating," *ISSCC Dig. Tech. Papers*, pp. 244-245, Feb. 2017.
[4] K. Badami, et al., "Context-Aware Hierarchical Information-Sensing in a 6μW 90nm CMOS Voice Activity Detector," *ISSCC Dig. Tech. Papers*, pp. 430-431, Feb. 2015.
[5] M. Yang, et al., "A 1μW Voice Activity Detector Using Analog Feature Extraction and Digital Deep Neural Network," *ISSCC Dig. Tech. Papers*, pp. 346-347, Feb. 2018.
[6] M. Cho, et al., "A 6×5×4mm^3 General Purpose Audio Sensor Node with a 4.7μW Audio Processing IC," *IEEE Symp. VLSI Circuits*, pp. 312-313, June 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

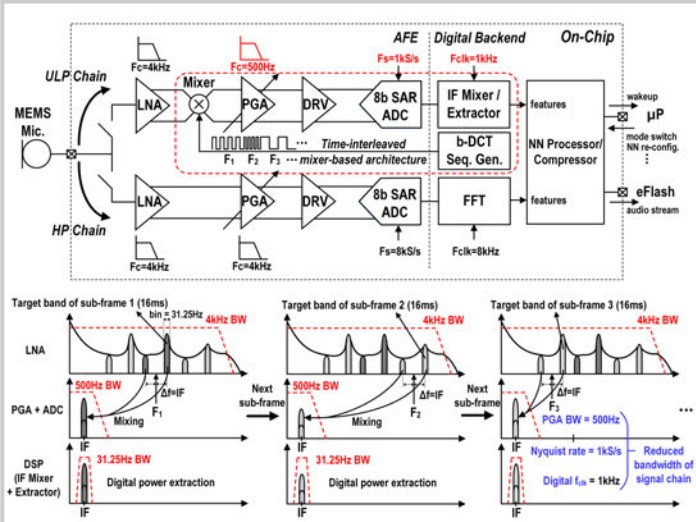


Figure 17.2.1: Acoustic sensing system architecture (top), and operation principle of time-interleaved mixer-based frequency scanning (bottom).

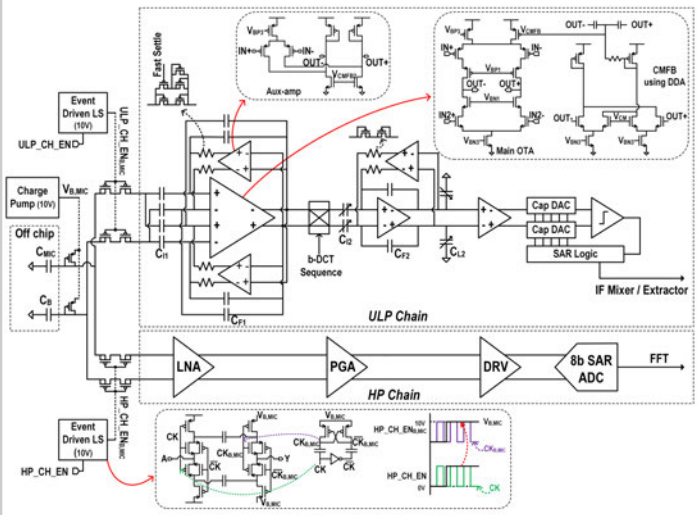


Figure 17.2.2: Circuit diagram of the analog front-end with ULP and HP chains.

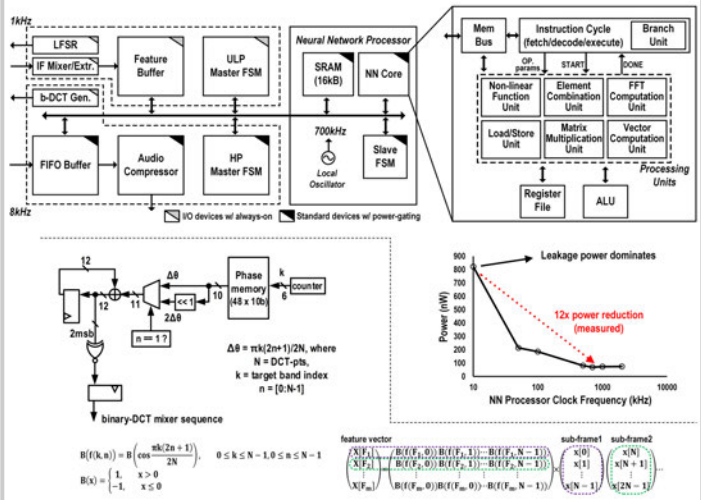


Figure 17.2.3: Digital backend architecture including neural network processor (top), measured power reduction from computational sprinting (bottom right), and binary DCT mixer sequence generator (bottom left).

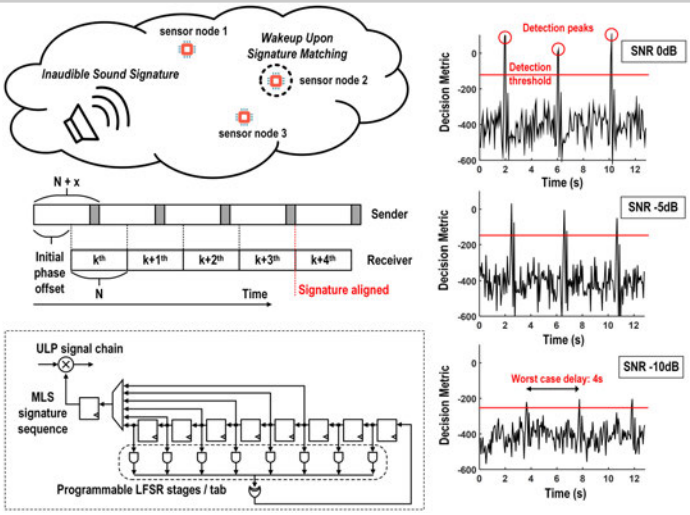


Figure 17.2.4: Acoustic signature wakeup detection (left), and measurement results with 6 stages, 63-length sequence at various SNRs (right), showing detection down to -10dB SNR.

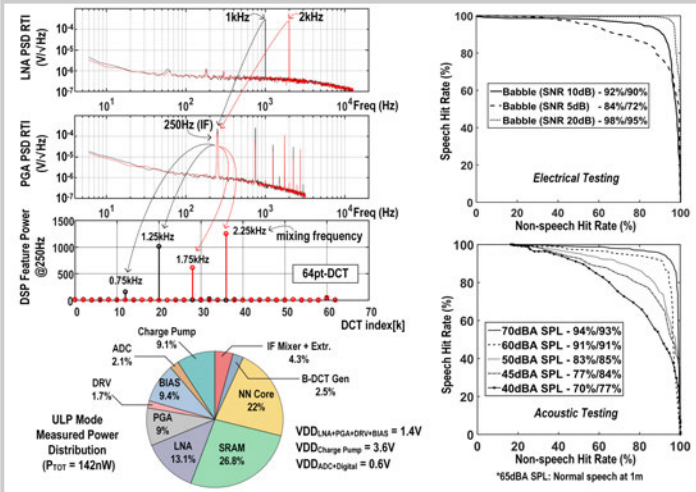


Figure 17.2.5: Chip measurement results. Power spectral density for LNA, PGA, and DSP (top left). Two different applied tones are mixed down to 250Hz in IF and extracted by DSP at two mixing frequencies each. ULP mode power distribution (bot left), and ROC curves for VAD (right).

Feature Extractor	This Work	ISSCC'18 [5]	ISSCC'16 [1]	ISSCC'15 [4]
Technology (nm)	180	180	180	90
Feature Extraction Type	Mixed-signal	Analog to events	Digital	Analog
Channel Number	16 – 48	16	4 – 16	16
Frequency Range (Hz)	75 – 4k	100 – 5k	0.2 – 470	75 – 5k
Power (nW)	60	380	10	6000
Normalized Power* (nW)	5	71	34	1186
Dynamic Range (dB)	47	40	N/A	40
Building Blocks	LNA, Mixer, LPF, DSP	LNA, BPF, FWR, IAF	DSP	LNA, BPF, FWR, LPF

ULP AFE(LNA+PGA) Summary	
Gain(dB)	22.5-49.2
Out Noise (mVrms)	0.83@min gain 4.6@max gain
Max Output (Vpp)	1.45(<0.4%THD)
Dynamic Range(dB)	55.8@min gain 40.9@max gain
Power (nW)	31
Bandwidth (Hz)	500
VDD(V)	1.4

Voice Activity Detector	This Work	ISSCC'18 [5]	ISSCC'17 [3]	ISSCC'15 [4]	JSSC'13 [2]
Technology (nm)	180	180	65	90	32
Acoustic Input	Analog/passive mic. w/ gain stage	Analog mic. w/ gain stage	Assume digitized	Analog mic. w/ gain stage	Assume digitized
Classifier	Neural network	Neural network	Neural network	Decision tree	Energy-based
Classifier Topology Programmability	Yes	No	Yes	No	No
Dataset**	LibriSpeech + NOISEX-92	AURORA4 + DEMAND	AURORA2	NOISEUS	N/A
Latency (ms)	512	10	10	<100	10
Power (μW)	0.142	1	22.3	6	~300
Accuracy SP/Non-SP hit rate (electrical test)	91.5%/90%*** @ babble 10dB SNR	84%/85% @ restaurant 10dB SNR	90%/90%**** @ unspecified context 7dB SNR	89%/85% @ babble 10dB SNR	97% accuracy @ unspecified context unspecified SNR
Acoustic Testing Performed	Yes	No	No	No	No

*calculated according to the equation in [5], normalized to 4kHz; **all datasets are similar in speech quality;
***measured at ULP mode with 128pt-DCT, 32 feature channels, and 250Hz IF. ****converted from EER in [3].

Figure 17.2.6: Comparison table for feature extractor (top left), VAD (bottom), and performance summary of ULP AFE (top right).

17

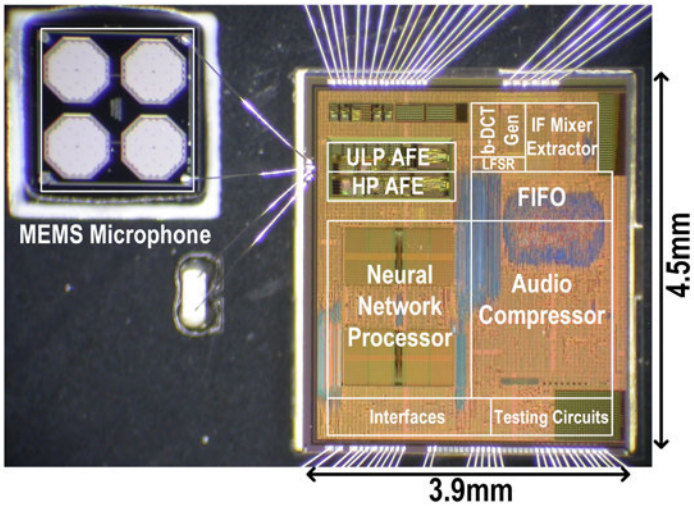


Figure 17.2.7: Die micrograph and system intgration with MEMS microphone.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

17.3 Hybrid System for Efficient LAE-CMOS Interfacing in Large-Scale Tactile-Sensing Skins via TFT-Based Compressed Sensing

Levent E. Aygun*, Prakhar Kumar*, Zhiwu Zheng*, Ting-Sheng Chen*, Sigurd Wagner, James C. Sturm, Naveen Verma

Princeton University, Princeton, NJ
*Equally-Credited Authors (ECAs)

Tactile sensing has wide-ranging applications, from intelligent surfaces to advanced robotics. Large-Area Electronics (LAE), based on low-temp. fabrication (<200°C) of thin films, presents distinct capabilities, due to compatibility with a broad range of materials (enabling diverse transducers), as well as large and flexible substrates and materials-deposition methods (enabling expansive and formfitting sensing arrays). However, low performance/energy-efficiency of LAE thin-film transistors (TFTs) necessitates hybrid systems, integrating Si-CMOS ICs for system functions (sensor readout/control, processing, etc.). Initial work shows that a primary challenge in hybrid systems is the large number of interfaces required between LAE and CMOS, particularly as the number of sensors scales [1,2]. This paper presents a force-sensing system that exploits signal sparsity exhibited in many large-area tactile-sensing applications (e.g., detecting point damage/stress in structures [3]), to reduce interfacing complexity to the level of sparsity, rather than a level related to the number of sensors (e.g., [1]). This is achieved via compressed sensing (CS), enabling sensor-acquisition by simple switches, readily implemented using TFTs. While CS has previously been leveraged in a hybrid-system architecture targeting signal sampling-rate requirements [2], this system applies it for high spatial resolution in tactile sensing.

Figure 17.3.1 summarizes CS and the system-level motivations. As shown, CS involves sampling different superpositions $\vec{y}$ of input data $\vec{x}$ (having dimensionalities M and N, respectively), via a measurement matrix ϕ. CS theory says that: (1) if $\vec{x}$ has at most K non-zero elements (K sparse), only $M>K\times\log_2(N/K)$ different superpositions are needed to reconstruct $\vec{x}$; and (2) a random and simple matrix ϕ, with elements 0/1, can be found for acquiring such superpositions. This makes TFT implementation possible, and Fig. 17.3.1 compares how system complexity (vs. number of sensors N) would scale compared to traditional TFT architectures for hybrid-system interfacing (sequential scanning, active matrix), based on the following key metrics: (1) number of acquisition cycles (operational complexity); (2) number of interfaces (system-assembly complexity); (3) dynamic range (DR) of each interface (CMOS/readout complexity); (4) number of TFTs (LAE complexity). If M,K<<N (i.e., high sparsity), CS substantially benefits cycles and interfaces, which together with DR represent total data bandwidth of transfer, and can be traded-off. But, modest increase in DR (~K) is preferred, as it pushes complexity to the high-efficiency CMOS domain and because sensor proximity to embedded signals in tactile sensing typically yields large responses [3]. A drawback with CS is the number of TFTs; but, as described below, the statistical nature of acquisition with CS enables high tolerance to typical TFT faults.

Figure 17.3.2 shows the CS hybrid-system architecture, demonstrated for an array of resistive force sensors (applicable to other resistive sensors, e.g., temp./strain/pressure/light). While the transfer function of $R_{SNS,i}$, shown for 10 sensors, exhibits variation, the aim of tactile sensing is often coarse readout from each sensor, but with high spatial resolution. Each sensor feeds an access TFT, controlled to implement the 0/1 elements of ϕ for one column. A TFT Matrix-Logic Control block controls the access TFT, via a Matrix-Row-Selection code R/Rb[4:0] from CMOS. The control-block details are shown, consisting of matrix TFTs arranged in 32 branches with 5 TFTs/branch. The Matrix-Control signal C/Cb[4:0] enables one branch at a time, driving the access-TFT gate high/low (V_{ON}/V_{OFF}). To simplify design, the Matrix-Logic Control block is the same for each sensor, but the connections between R/Rb[4:0] and C/Cb[4:0] are different, so that the access TFTs are enabled differently as R/Rb[4:0] is cycled through 32 codes (e.g., truth-table shown for 20th sensor). By superimposing the currents from accessed sensors via the CMOS TIA, which ensures constant sensor biasing of 0.4V, a 32-row ϕ matrix is realized. The architecture thus employs a single interface and M=32 cycles, one for each conceptual row of ϕ (same effective data bandwidth as Fig. 17.3.1); but, an additional 5 differential interfaces are required for R/Rb[4:0]. M=32 supports up to N=120 sensors at target sparsity of K=3, but 20 sensors are implemented for demonstration (requiring 20×(5×32)=3200 matrix TFTs). The W/L's for matrix/access TFTs are shown, designed for access-TFT on-resistance $R_{ACC}\approx1.5k\Omega$, well below the target range for force sensors.

Figure 17.3.3 analyzes TFT fault tolerance. While all interfacing architectures in Fig. 17.3.1 require an access TFT, the 160 matrix TFTs per sensor pose notable overhead (though, this increases slowly with number of sensors N, due to $\log_2$-dependence of M on N). Figure 17.3.3 shows substantial system tolerance to matrix TFT faults. Typical faults include: (1) S-D open; (2) S-D leakage/short; (3) G-S/D leakage/short. Probability of G-S/D shorts must be controlled, as it impacts the global control signal R/Rb[4:0]; this can be done via gate-dielectric processing (oxide quality/thickness, minimizing gate-electrode hillocks). Other faults can be analyzed by considering their effective impact on ϕ. As shown, S-D opens cause element values to be set by charge dynamically held at the access-TFT gate from the previous state, corresponding to previous row. On the other hand, S-D leakage/shorts and G-S/D leakage cause contention at the access-TFT gate, making elements take intermediate values. Monte Carlo simulations, applying such fault models, show high reconstruction SNR (RSNR) is maintained to high fault rates (much higher than typically allowed for access TFTs), and RSNR improves with even higher N. This is due to use of statistical-optimization methods and many superpositions in the CS reconstruction process.

Figure 17.3.4 shows a block diagram of the CMOS IC, having 8 channels of a TIA, offset-correction, and 10b ADC readout chain, as well as digital control and R/Rb[4:0] generation. Offset correction consists of a 32-word register file and 7b current DAC (I-DAC). This enables an offset-correction current to be applied to the TIA for each matrix-row readout, via a code in the register file determined at start-up (with no force applied). The I-DAC consists of 6 binary-weighted N/PMOS current sources. The TIA employs a 2-stage op-amp, designed for stability with 500pF of input capacitance, supporting >200 access-TFT load. The ADC is a 10b SAR with 5b main/sub cap-DACs. The 8 channels enable further scaling in sensor number, via parallel arrays.

The force-sensing system is implemented with in-house-fabricated ZnO TFTs, commercial force sensors, and custom 130nm CMOS IC (Fig. 17.3.7). To ease testing, 3 PCBs are used, consisting of: (1) 20 force sensors; (2) 20 wire-bonded matrix-control/access-TFT die; (3) CMOS IC. The TFTs are fabricated on glass for dicing and wire bonding, but are fully flex compatible (process temp. <200°C). Figure 17.3.5 (left) shows overlaid waveforms from 32 TFT die from one sample, showing proper matrix-TFT operation, and proper pull up/down by access TFT (with test 10kΩ load resistor connected to 1V). Also shown are the I_D-V_{DS}'s of the access TFTs, showing $1k\Omega<R_{ACC,i}<1.9k\Omega$ (in V_{DS} range of interest for sensor resistance). Figure 17.3.5 (right) shows CMOS IC measurements, including ADC and full-system (TIA+ADC) DNL/INL, as well as the ADC code vs. I-DAC code (no input current). A summary table is at the bottom.

Figure 17.3.6 shows the demonstration setup, as well as full-system sensor acquisition/reconstruction. A sample heat map is shown for force-sensor resistance, both directly measured and reconstructed by the system, with <2.9% error (after correcting for nominal $R_{ACC,i}$=1.5kΩ). The scatter plot is derived from many such measurements, showing error of $0.7k\Omega_{RMS}$. R/Rb[4:0] codes switch at 1kHz, giving frame rate of 1kHz/32=31fps, and total energy of 1.2μJ/frame.

Acknowledgements:
This work was supported by the Semiconductor Research Corporation (SRC), Princeton Program in Plasma Science and Technology, DARPA and Princeton University. IC fab provided by MOSIS.

References:
[1] Y. Afsar, et al., "Large-Scale Acquisition of Large-Area Sensors Using an Array of Frequency-Hopping ZnO Thin-Film-Transistor Oscillators," *ISSCC Dig. Tech. Papers*, pp. 256-257, Feb. 2017.
[2] T. Moy, et al., "A Flexible EEG Acquisition and Biomarker Extraction System Based on Thin-Film Electronics," *ISSCC Dig. Tech. Papers*, pp. 294-295, Feb. 2016.
[3] B. Glisic, et al., "Strain Sensing Sheets for Structural Health Monitoring Based on Large-area Electronics and Integrated Circuits," *Proc. of IEEE*, vol. 104, no. 8, pp. 1513-1528, June 2016.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

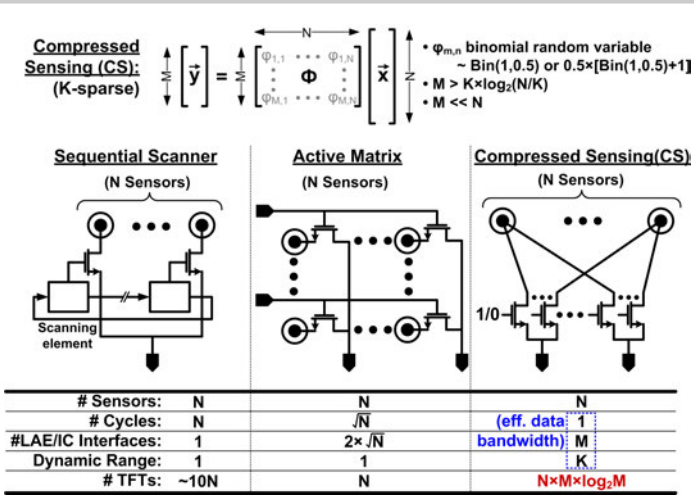


Figure 17.3.1: Summary of CS (M measurements for N sensors), and comparative analysis of systems with traditional hybrid-system-interfacing architectures, as N scales.

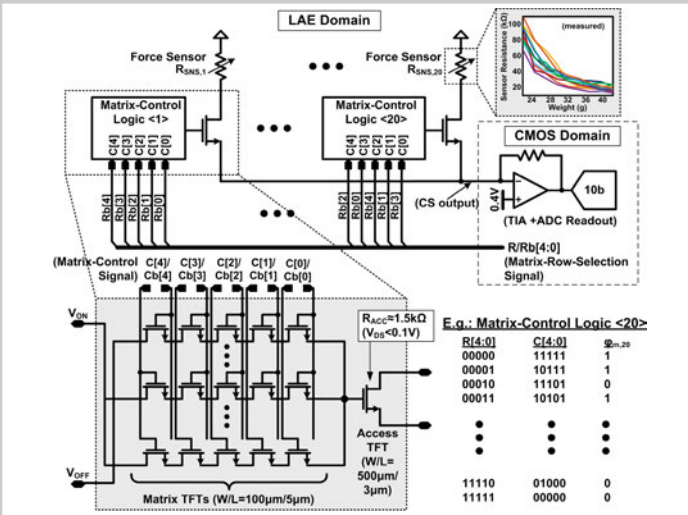


Figure 17.3.2: System for performing readout of sensor-current superpositions, controlled by access and matrix TFTs, using Matrix-Row-Selection signal R/Rb[4:0].

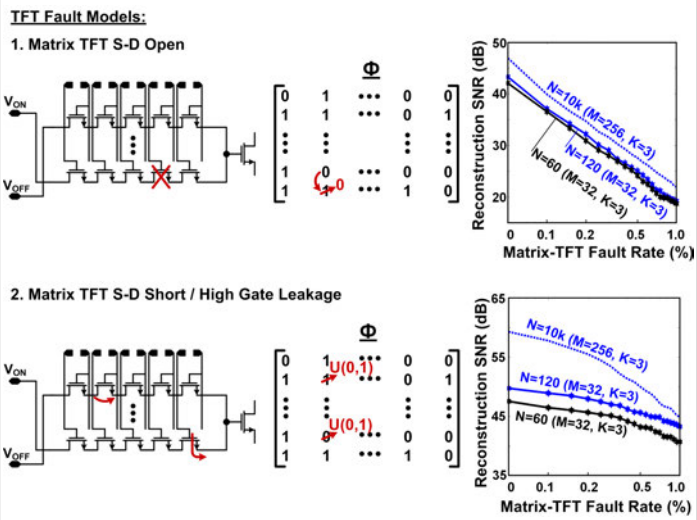


Figure 17.3.3: Analysis of reconstruction-performance tolerance to typical faults in matrix TFTs.

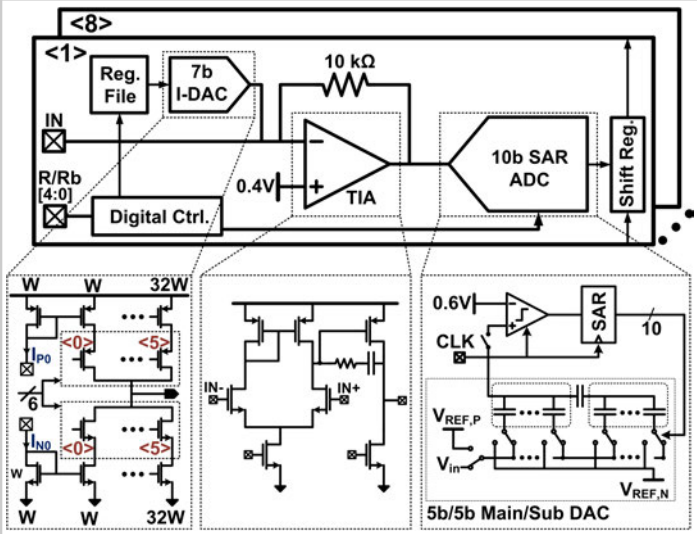
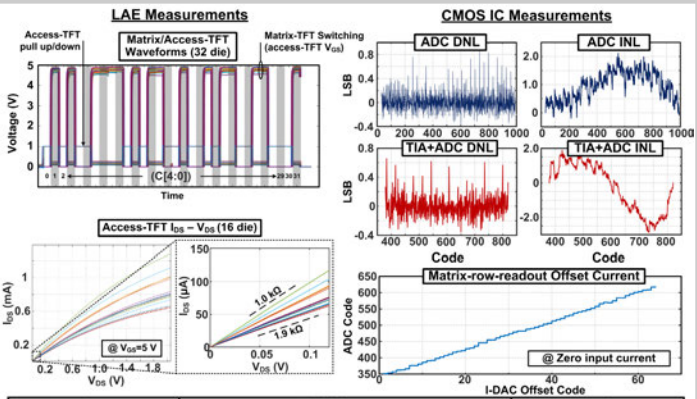


Figure 17.3.4: Block diagram of CMOS IC for matrix-row readout, showing 8 readout channels comprised of offset-canceling I-DAC, TIA, and 10b ADC.



LAE Summary		CMOS Summary				Full-system Summary	
No. sensors	20	Technology	130nm	ADC resolution	10 b	Sensor F range	24 g-40 g
TFT technology	ZnO	V_{DD}	1.2V	ADC energy	212 pJ/conv.	Sensor R range	100 kΩ-15 kΩ
Process temp.	<200°C	Area	4mm²	ADC DNL/INL	0.85/2.0	Sparsity	3
TFTs/sensor	161	No. channels	8	ADC rate	100 kS/s	Recon. error	0.7 $k\Omega_{RMS}$
Matrix-TFT speed	1 kHz	TIA power/ch.	37μW	ADC ENOB	8.7 b @50kHz	Frame rate	31 fps
Access-TFT res.	1.5 kΩ	I-DAC res.	6b	TIA+ADC DNL/INL	0.65/2.0	E/frame	1.2 μJ

Figure 17.3.5: Matrix-Control block and access TFTs (left), CMOS ADC DNL/INL, full-system (ADC+TIA) DNL/INL, and I-DAC transfer function (right), with summary table.

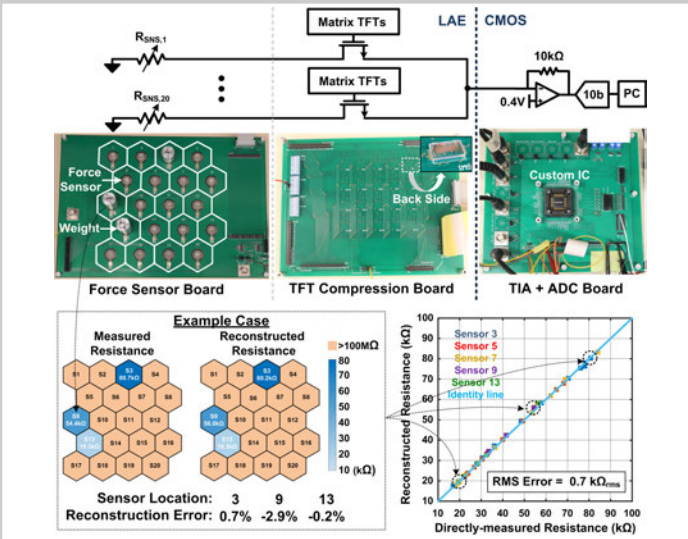


Figure 17.3.6: Measurement and demonstration setup as well as characterization of reconstruction accuracy.

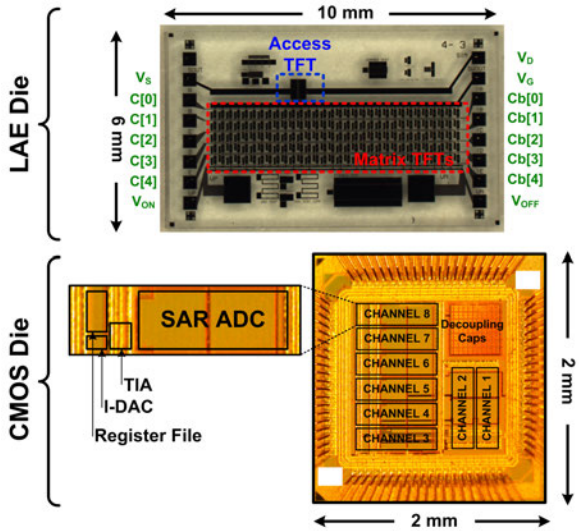


Figure 17.3.7: TFT die micrograph (showing matrix/access TFTs) and CMOS IC die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

17.4 16MHz FRAM Micro-Controller with a Low-Cost Sub-1μA Embedded Piezo-Electric Strain Sensor for ULP Motion Detection

Sudhanshu Khanna[1], Michael Zwerg[1], Brian Elies[1], Juergen Luebbe[1], Nagaraj Krishnasawamy[1], Hadi Najar[2], Suman Bellary[2], Wei-Yan Shih[1], Scott Summerfelt[1], Steven Bartling[1]

[1]Texas Instruments, Dallas, TX
[2]Texas Instruments, Santa Clara, CA

Ultra-low Power Microcontrollers (MCUs) [1-4] have played a central role in embedded IoT systems providing programmability, analog and digital processing and control, A/D interfaces, and power management. As IoT applications expand, efficient sensing is increasingly becoming part of MCUs. In this paper we present a 130nm 16MHz Ferro-electric RAM (FRAM) based MCU with a sub-1uA embedded piezo-electric strain sensor and AFE for ULP motion detection (Fig. 17.4.1). To our knowledge, this is the first reported MCU with an embedded motion detection strain sensor. Existing applications that would benefit from such a MCU are applications like toys and remote controls that can turn off while not in use. Motion detection in a key fob improves security by preventing a "man in the middle" attack while the key fob lies stationary at home. Tamper detection and strain gauges are other potential applications. When used in "wake-on-motion" applications, the sensor IP is powered-on at all times waiting for a motion event. Hence minimizing its power consumption is crucial. Also, considering ULP MCUs have many cost-sensitive applications, the sensor must be small in area, and not require any additional masks or special processing steps. A single-chip solution allows reuse of power management, programmability and control circuits already existing in the MCU for use in the strain sensor IP, reducing the system level cost vs a 2-chip solution.

The strain sensor in this work is built using thin film Lead Zirconium Titnate (PZT), which was observed to have a high piezoelectric sensitivity. Other CMOS compatible elements like NWELL resistors also showed piezoelectric behavior but with much lower sensitivity. FRAM, the main memory in this MCU, also uses PZT capacitors (fecaps) as the non-volatile element. The motion detection strain sensor can be manufactured using the same processing steps and masks as FRAM, resulting in no additional processing cost for the sensor. Further, the sensor and AFE together consume only 0.6mm^2, making this a low cost addition to a 130nm based MCU. The IP consumes just 650nA sampling at 100Hz, enabling long battery life.

Figure 17.4.2 shows the sensor IP architecture. The core of the IP is a fecap-based sensor and a 3-stage AFE. The sensor is an array of capacitors and is thus high impedance. Polarized fecaps contain crystal domains that have a polarized dipole associated with them. When strain is applied to a polarized piezo-electric material, the dipoles are displaced resulting in a net surface charge at the terminals of the fecaps. Parasitic leakage from transistors connected to the PZT capacitor terminals prevents them from maintaining the evoked charge indefinitely, so PZT sensors are only suitable for sensing dynamic strain. Gain stages amplify this surface voltage (100's of uV) before it disappears due to leakage. Surface charge that develops when strain is applied is proportional to the magnitude of the fecap polarization. The FSM is used to polarize each fecap upon power-up, setting up the sensor for motion detection. To connect the FSM to the sensor, we must use ultra-low leakage switches since minimizing leakage is critical in enabling detection of low frequency human movement. Sensor, AFE and clock parameters such as sensor size, gain, bandwidth, sample-hold clock duty cycles, clock frequency and strain detection threshold are software programmable. The AFE is sampled to achieve the low power budget. Human motion is slow, and hence a sampling rate of ~200Hz is sufficient, though the AFE is capable of sampling at rates of 1Hz-to-1 kHz. The LNA is a sampled resistive load PMOS-differential amplifier. LNA bandwidth is configurable using capacitors connected to its output. During the long off-time of the sampling period, the drain, source and body of the LNA input pair must be biased at VREF (sensor DC bias) to prevent transistor leakage from draining away the evoked charge on the high impedance sensor or disturbing the sensor's DC bias. AFE Stages 2-3 are switch-capacitor amps with configurable gain. Note that the LNA is an open loop amplifier and thus does not have well controlled offset and gain. LNA offset correction is implemented by trimming current from its resistive load. A current-DAC provides a flexible level of current trim based on software configuration offset control registers. Offset and gain are initially measured by applying a known input voltage (from the REF block) and then offset and gain control registers are set to achieve zero offset and the required gain (using the stage 2-3 PGAs). This procedure is repeated periodically as temperature can change over time. A DAC-subtractor subtracts a user-defined threshold from the AFE output and provides it to a comparator block. If the sensor output exceeds this threshold, the comparator output triggers a wake-up interrupt for the MCU core.

Figure 17.4.3 shows the sensor circuit design. When strain (defined as ΔL/L) is applied, an individual fecap develops only uVs of signal/ustrain. To generate a voltage large enough to be detectable above the noise floor of the LNA, fecaps must be stacked in series. Note that stacking capacitors reduces the effective capacitance of the sensor output. To avoid large charge sharing losses with the LNA input capacitance, eight stacks of fecaps are connected in parallel increasing the capacitance of the sensor output (resulting in an 8 wide x "N" tall fecap stack). A discussion of parasitic effects is now due in order to explain why voltage gain rolls off as fecap stacking height increases. In this process technology, the fecap PZT layer is deposited between active/poly and MET1. Thus, parasitic capacitors exist on the fecap from its terminals to metal layers, poly, active and substrate. Transistors connected to the fecap terminals (low leakage switches) also add parasitic drain capacitance. Parasitic capacitance is lowered by not allowing any metal/poly or active above/below the sensor. Effective sensor capacitance decreases as one traverses upwards from the bottom of a stack of fecaps, but this decreasing effective capacitance charge shares with a parasitic capacitance that remains relatively constant at every level in the stack. Thus, stacking fecaps has diminishing returns. In the absence of parasitics, a 64-series 8-parallel single ended sensor would generate the same signal as a 32-series 8-parallel differential (2-leg) sensor while using the same number of total fecaps (512). However, in the presence of parasitics signal from a single ended 64-stack sensor is degraded by 36%. Using a 32-stack differential sensor cuts the stack height from 64 to 32, reducing parasitic loss to 14% while also providing higher supply and reference noise rejection. Low leakage FSM-Sensor interface switches are built using isolated p-well high-voltage thick-gate-oxide NMOS. By biasing the p-well and source to the same DC voltage as the sensor (drain) we can drastically reduce reverse diode and sub-threshold leakage on the fecap terminals allowing signal detection up-to 1Hz.

Motion in any application produces strain on the PCB that is transferred to the sensor through the package. Thus, the strain sensor characteristics as observed by the integrated sensor are also a function of the PCB mechanical system characteristics (resonant frequency, strain transfer efficiency, etc.) in addition to piezoelectric material properties. Our measurement/evaluation-module (EVM) setup uses conventional PCB material and thickness (FR4, 2 layer, 31 mils), conventional die thickness and an off-the-shelf QFN package. The setup is extremely sensitive and detects very light taps/knocks (picture in Fig. 17.4.6), walking or waving. Figure 17.4.4 shows the sensor output for different motion types along with accelerometer output (located on the same PCB).

Figure 17.4.5 shows the sensor o/p and also shows how the output is modulated by PCB characteristics. The lower frequency range (before PCB resonance) is used to show the PZT sensor's relative temperature independence. Sensitivity is 80mV/g at 60× gain (1.33mV/g at input). Reducing PCB/package/die thickness or designing the PCB resonance to match the target motion increases the sensor output even further. For example, a second smaller and thinner (15mil) PCB with just supply, ground and LED pins connectors designed as an add-on to a primary system board helped achieve even higher sensitivity. Measured total noise is 800uV (input referred, dominated by sensor), and varies <15% across -40C to 85C. Measured sensor IP current is 650nA at 100Hz, and 400nA at 33Hz sampling rate (dominated by analog current). The IP is 0.6mm^2 and needs no additional masks or special processing steps resulting in low cost. This IC is the only reported MCU with an embedded motion detection strain sensor. The closest possible comparisons are accelerometer-MCU SIP/2-chip solutions [5-6], which cost more and/or consume more power due to the use of MEMS and higher power AFEs (essential for high precision applications, but not motion detection). Figure 17.4.6 shows the measurement results and comparison table and Fig. 17.4.7 shows the die micrograph.

References:
[1] J. Kwong, et al., "A 65nm Sub-Vt Microcontroller with Integrated SRAM and Switched-Capacitor DC-DC Converter," *ISSCC Dig. Tech. Papers*, pp. 317-318, Feb. 2008.
[2] M. Zwerg, et al., "An 82μA/MHz microcontroller with embedded FeRAM for energy-harvesting applications," *ISSCC Dig. Tech. Papers*, pp. 333-334, Feb. 2011.
[3] Y. Lee, et al., "A Modular 1mm^3 Die-Stacked Sensing Platform With Optical Communication and Multi-Modal Energy Harvesting," *ISSCC Dig. Tech. Papers*, pp. 402-403, Feb. 2012.
[4] A. Klinefelter, et al., "A 6.45 μW Self-Powered IoT SoC with Integrated Energy-Harvesting Power Management and ULP Asymmetric Radios," *ISSCC Dig. Tech. Papers*, pp. 384-385, Feb. 2015.
[5] ST Microelectronics LIS331DLH Datasheet, https://www.st.com/resource/en/datasheet/lis331dlh.pdf
[6] Analog Devices ADXL362 Datasheet, https://www.analog.com/media/en/technical-documentation/data-sheets/ADXL362.pdf

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

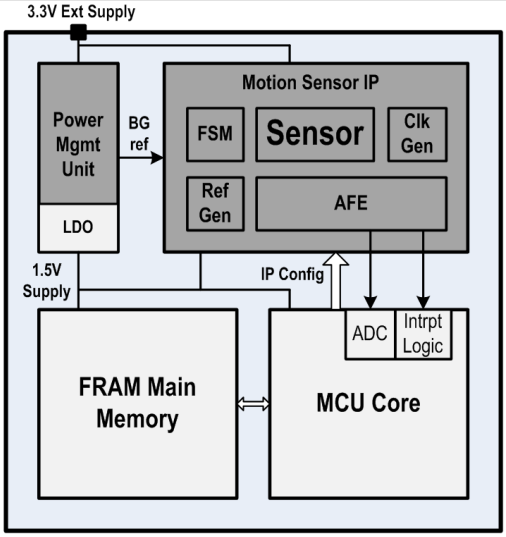


Figure 17.4.1: MCU SoC architecture. The SoC features the sensor IP, power management, memory, and digital and analog peripherals.

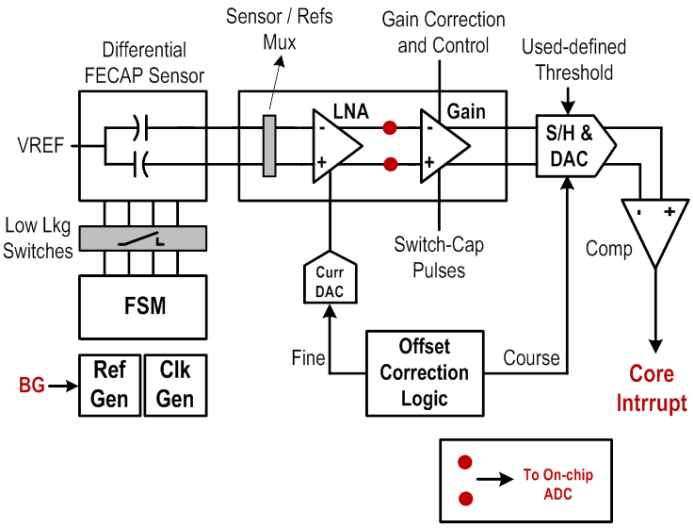
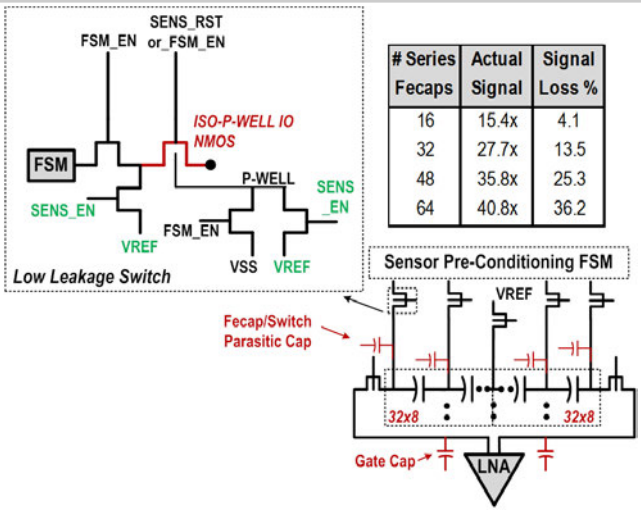


Figure 17.4.2: Sensor IP architecture.



# Series Fecaps	Actual Signal	Signal Loss %
16	15.4x	4.1
32	27.7x	13.5
48	35.8x	25.3
64	40.8x	36.2

Figure 17.4.3: Sensor and switch circuit design. The switch is used in three modes: FSM, sensor, and sensor reset. The table shows how signal loss due to parasitics increases with the sensor series stack height.

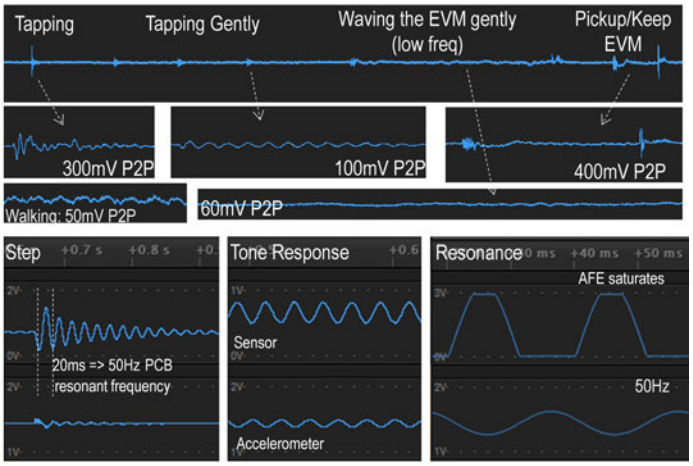


Figure 17.4.4: Measured sensor response to human motion and shaker table based step and sine functions.

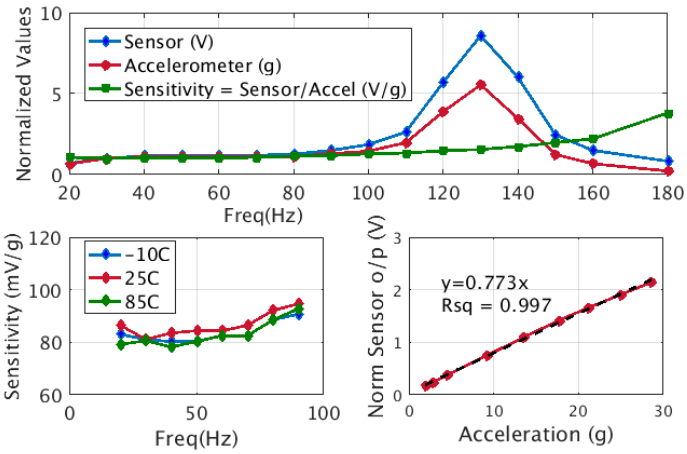


Figure 17.4.5: Measured sensor o/p, accel o/p and sensitivity vs freq and temp (shaker table driven at constant amplitude) and sensor o/p vs amplitude (linearity). Freq<15Hz limited by shaker table.

Technology	130nm
Supply	2.5V-3.6V
Area (Sensor + AFE)	0.6mm2
Sensor IP Power	
Sampling @ 33Hz	400nA
Sampling @ 100Hz	650nA
Total SoC Standby Power	1.4uA
Sampling Range	1Hz - 1kHz

Sensitivity (i/p ref)*	1.33mV/g
Noise (i/p ref)	
AFE	300uV
Sensor+AFE	800uV
LNA Bandwidth	20kHz
Sensor Bandwidth**	>>200Hz
Resolution	9-bit
Range	30g

* g-values measured by on-board accelerometer. Sensitivity = sensor/accel
** Sensor BW measurement limited by PCB bandwidth.

Metric	This work	[5]	[6]
MEMS processing	No	Yes	Yes
Cost Estimate	Low	Medium	High
Power	0.65uA	10uA	3.7uA
Sampling Rate	100Hz	10Hz	100Hz
Suitable for precision appls?	No	Yes	Yes

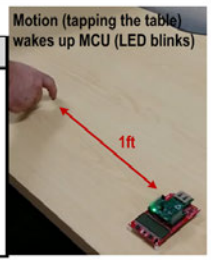


Figure 17.4.6: Measurement results and comparison table. Demo shows MCU waking-up in response to a tap. Power measured at 25C.

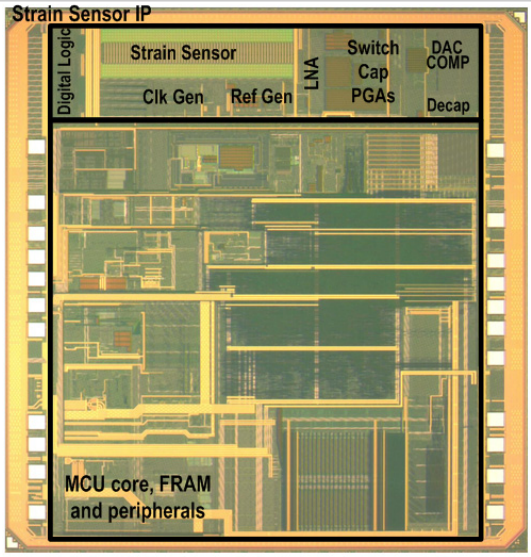


Figure 17.4.7: Die micrograph showing the MCU SoC with the sensor IP.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

17.5 A 0.8mm³ Ultrasonic Implantable Wireless Neural Recording System with Linear AM Backscattering

Mohammad Meraj Ghanbari, David K. Piech, Konlin Shen, Sina Faraji Alamouti, Cem Yalcin, Benjamin C. Johnson, Jose M. Carmena, Michel M. Maharbiz, Rikky Muller

University of California, Berkeley, CA

Miniaturization of implantable neural recording systems to micron-scale volumes will enable minimally invasive implantation and alleviate cortical scarring, gliosis, and resulting signal degradation. Ultrasound (US) power transmission has been demonstrated to have high efficiency and low tissue attenuation for mm-scale implants at depth in tissue [1,2,3], but has not been demonstrated with precision recording circuitry. We present an US implantable wireless neural recording system scaled to 0.8mm³, verified to safely operate at 5cm depth with state of the art neural recording performance an average circuit power dissipation of 13μW, and 28.8μW including power conversion efficiency. Sub-mm scale is achieved through single-link power and communication on a single piezocrystal (Lead Zirconate Titanate, PZT) utilizing linear analog backscattering, small die area, and eliminating all other off-chip components.

A simplified block diagram of the system is shown in Fig. 17.5.1, which consists of an external interrogator and a free-floating mote comprised of only a 0.42mm³ PZT, a 0.25mm² IC, and a flex PCB interposer. The mote is encapsulated with a 10μm layer of parylene-C. The interrogator (Cephasonics, cQuest Cicada) sends a pulse in transmit (TX) mode, which arrives at the mote after one time of flight (ToF). The interrogator then switches to receive (RX) mode to receive the reflected amplitude-modulated (AM) wave from the mote after another ToF. The RX chain then amplifies, filters, and digitizes the received signal for demodulation. Interrogation is performed in a pulse-echo fashion to relax the dynamic range requirements on the RX chain imposed by TX-to-RX acoustic leakage and eliminate the need for a circulator. Thus, to avoid TX/RX interference, the pulse width is set to <2 ToF. Since implant depth can vary, a mote placed close to the interrogator (e.g. ~1cm) would have 2xToF~13μs, too short for wake-up and robust multi-bit digital modulation in a single pulse. To overcome this limitation, the IC performs analog AM backscattering which carries higher information per cycle. The analog signal may be transmitted even if only a few cycles are available per pulse. The addition of a unique subcarrier frequency in each mote can enable simultaneous multi-mote interrogation.

The PZT starts harvesting energy upon the arrival of an incident US wave. To eliminate the need for a large off-chip storage capacitor, the mote wakes up when each pulse arrives and sleeps in absence of a pulse. AC voltage generated across the PZT is rectified and regulated by the IC. At its series resonance frequency (f_r=1.78MHz), the PZT can be modeled as an AC source (V_s) and a series resistance (R_s~4kΩ). The acoustic reflection coefficient (Γ) at the boundary of the mote PZT and tissue can be modulated by tuning the impedance in shunt with the PZT (R_m), as shown in Fig. 17.5.2. However, $\Gamma \propto R_m/(R_m+R_s)$ and is therefore nonlinear with respect to R_m making this method impractical for analog backscattering. Linear backscatter modulation is possible as long as the peak voltage across the piezo terminals (V_{PZ}) is modulated linearly. This can be achieved by connecting synchronously upconverted baseband message current (I_m) to the PZT terminals (Fig. 17.5.2). To save area and complexity, the upconversion is implemented by reusing the available active rectifier on-chip, which automatically synchronizes the conduction period to f_r. During each half cycle, I_m generates a voltage drop across R_s linearly modulating $V_{PZ}=V_{PZ}^+-V_{PZ}^-$. The modulation depth of V_{PZ} is designed such that the rectified voltage is always higher than 1.1V, allowing the LDO to continuously regulate the 1V supply ($V_{DD,1V}$). A small on-chip capacitor (C_{store}=130pF) reduces voltage ripple on the rectified voltage (V_{RECT}). At the highest signal frequency of interest (55kHz subcarrier+5kHz neural signal), the impedance of C_{store} is ~5× larger than the impedance of the PZT, which results in I_m being supplied by the PZT rather than C_{store}. Therefore, in steady state, when the active rectifier pass transistor turns on, the peak voltage across V_{PZ} is, to first order, linearly modulated by I_m.

The IC block diagram is shown in Fig. 17.5.3. As V_{RECT} rises, a power-on reset (POR) is triggered and $V_{DD,1V}$ is generated within 10 cycles (~5μs) after the arrival of the incident wave. Control and timing signals are synchronized by a clock, which is extracted from the PZT voltage ($T_{CK}=1/f_r$). The low-noise amplifier is implemented by an auto-zeroed current-reuse capacitive feedback OTA (Gain 24dB, 250kHz BW required for passing a chopper/subcarrier (~55kHz)). Fast reset switches (ϕ_1, ϕ_2 and ϕ_3) and a power-gated low impedance voltage source are used to quickly charge the input terminals of the OTA to V_{CM}=0.5V in $4T_{CK}$ at the start of the sample. Chopper switches upmodulate the signal to simultaneously reduce 1/f noise of the amplifier and to act as a subcarrier frequency for communication. The kT/C noise sampled on C_s by the reset switches is modulated out of band when the subcarrier is downmodulated on the Rx side in digital. The chopper/subcarrier frequency (f_{CH}) is divided down from the main clock, while its phase is maintained from sample-to-sample by a reset triggered by the POR signal. The amplifier initialization completes in $12T_{CK}$ and the output of the AFE is valid after $22T_{CK}$ (~11μs). 11μs of wakeup time allows backscatter modulation for ~83% of the power-on time at a depth of 5cm. The LNA differential output is forced across a PMOS in triode by means of super source followers generating a linear current over ±200mV of LNA output voltage range. In absence of incident US pulse, the mote PZT is no longer actuated and the circuit turns off gradually as C_{store} discharges (in <10μs), and POR returns to the off state.

The IC was fabricated in a 65nm LP CMOS process. Figure 17.5.4 shows five measured interrogation events of modulated V_{PZ} in response to a 20mV$_{PP}$ input sine wave. The 11μs start-up time and subcarrier modulation at f_{CH} is observable on each pulse. The demodulated output signal and its corresponding spectrum are plotted for a 300Hz, 20mV$_{PP}$ input sine wave whose measured SFDR and THD are -50dB and -44dB, respectively. No harmonic tones are observed in the spectrum for a 10mV$_{PP}$ signal.

The measured static input-output characteristic is shown in Fig. 17.5.5 (output defined as the envelope of V_{PZ}). A 23dB voltage gain over 20mV$_{PP}$ of input range with <1.2% static nonlinearity is achieved. PEDOT:PSS coated electrodes (200x200μm², electroless nickel immersion gold plating) present <10mV DC offset, which is within the V_{IN} range. The demodulated input-referred noise spectral density with and without subcarrier modulation is plotted in Fig. 17.5.5. The total input-referred noise spectral density is 328nV/√Hz, and is dominated by carrier noise, which contributes 319nV/√Hz, therefore the noise contributed by the chip is 76nV/√Hz.

An *in vitro* measurement is shown in Fig. 17.5.6, where the assembled mote is suspended at a distance of 5cm away from the interrogator in water. Backscatter modulation and demodulated signal are shown for a 300Hz, 20mV$_{pp}$ sine wave applied at the input. A comparison with recently published fully integrated sub-mm³ neural recording implants is shown in the table in Fig. 17.5.7. We introduce a linearization technique for AM backscattering in ultrasonic implants. This design occupies the same volume as the smallest neural recording implant, but improves performance by combining low-noise, power-efficient, and linear neural recording with a PEF 2× lower, a depth 2.5×, and a static nonlinearity 40× lower than state of the art [1,4,5].

Acknowledgments:
The authors thank DARPA BTO, the sponsors of BWRC and TSMC for chip fabrication. Thanks to Ka Yiu Lee and Burak Eminoglu for technical discussion.

References:
[1] D. Seo, et al., "Wireless recording in the peripheral nervous system with ultrasonic neural dust". *Neuron*, vol. 91, pp. 529-539, 2016.
[2] T. C. Chang, et al., "27.7 A 30.5mm3 fully packaged implantable device with duplex ultrasonic data and power links achieving 95kb/s with <10E–4 BER at 8.5cm depth," IEEE *ISSCC Dig. Tech. Papers*, pp. 460-461, Feb. 2017.
[3] B. C. Johnson, et al., "StimDust: A 6.5mm3, wireless ultrasonic peripheral nerve stimulator with 82% peak chip efficiency," *IEEE CICC*, pp. 1-4, 2018.
[4] S. Lee, et al., "A 330μm×90μm opto-electronically integrated wireless system-on-chip for recording of neural activities," *ISSCC Dig. Tech. Papers*, pp. 292-294, Feb. 2018.
[5] P. Yeon, et al., "Towards a 1.1 mm2 free-floating wireless implantable neural recording SoC," *IEEE CICC*, pp. 1-4, 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

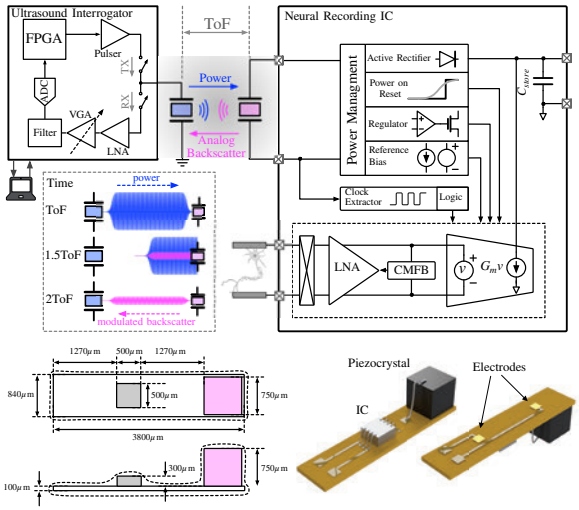


Figure 17.5.1: Simplified system block diagram, time-space conceptual diagram of power transmission and modulated backscattered waves. Assembled mote diagram.

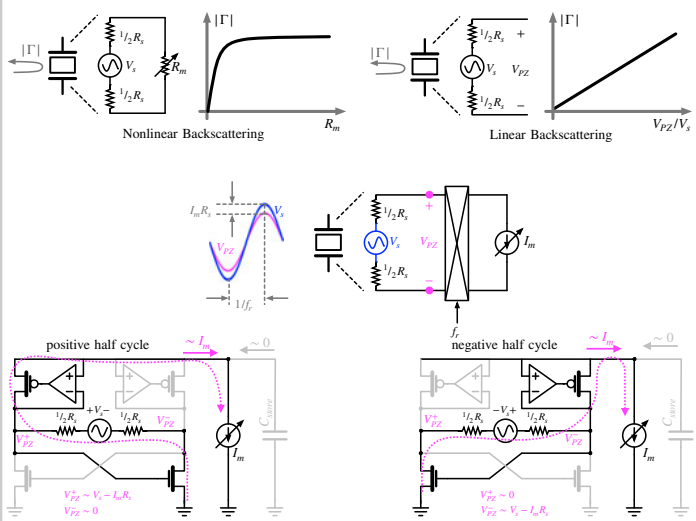


Figure 17.5.2: Backscatter uplink: nonlinear Γ vs. Rm. Linear Γ vs. V_{pz}. Reusing available active rectifier as a synchronous base-band I_m up-converter.

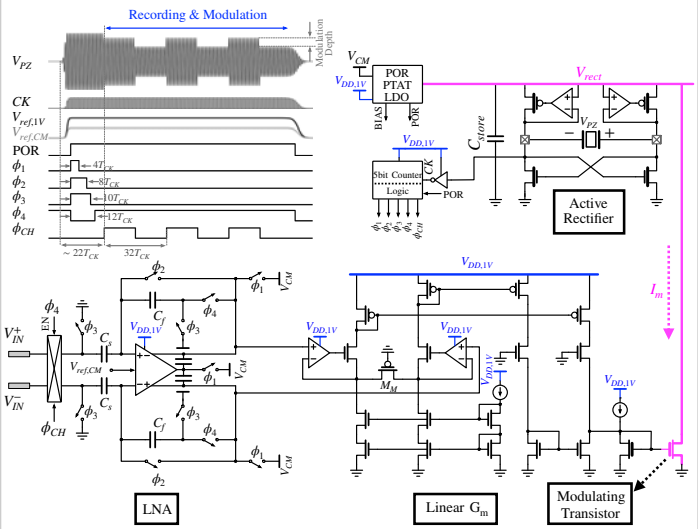


Figure 17.5.3: IC circuit schematics and timing diagram for a single sample pulse. Front-end LNA, subcarrier generation, and linear modulation are shown.

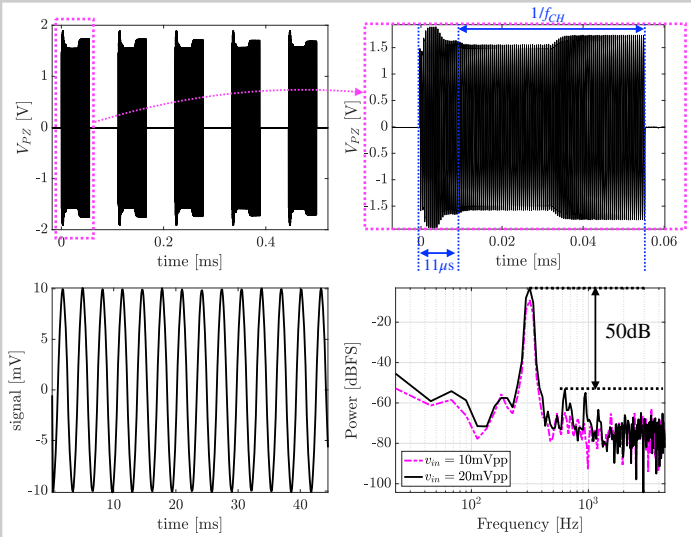


Figure 17.5.4: Five consecutive sample pulses for a 300Hz, 20mVpp input signal, the demodulated input-referred signal, and its corresponding spectrum.

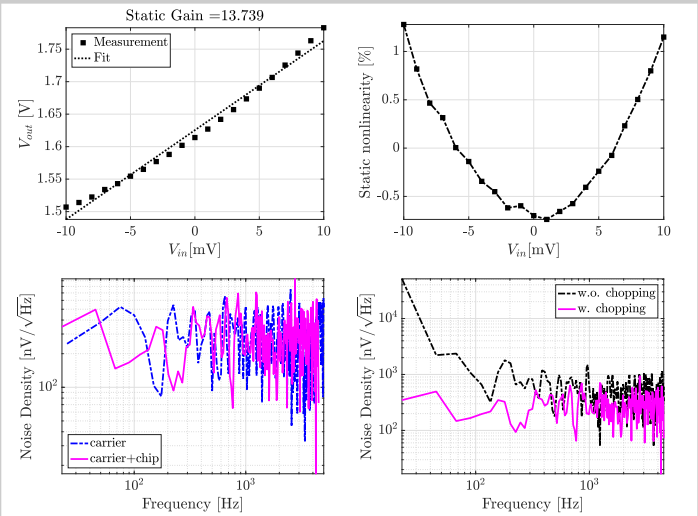


Figure 17.5.5: Static gain and nonlinearity (top); input-referred noise compared with the noise of the carrier alone, with and without chopping/subcarrier (bottom).

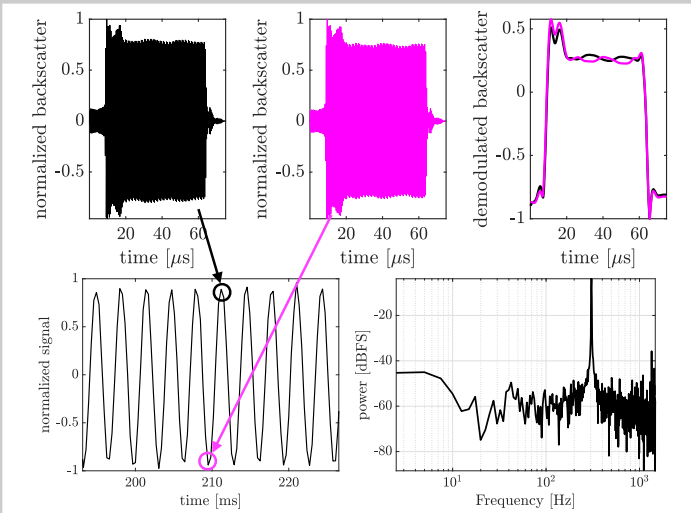
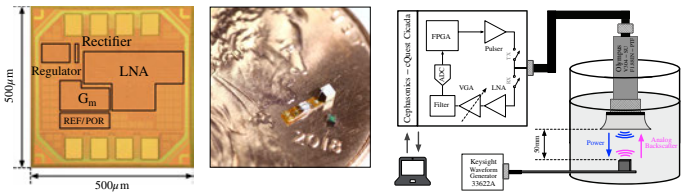


Figure 17.5.6: In vitro measurement results at 5cm depth: two individual backscatter signals, their corresponding AM demodulation and reconstructed 300Hz signal.



		Neuron'16	ISSCC'18	CICC'18	This work
Link		Ultrasound	Optical	RF	Ultrasound
Back telemetry		AM Backscatter	Analog PPM	IR-UWB	AM Backscatter
Implant depth (mm)		8.8	–	20	50
Total Volume (mm^3)		0.8	–	<1	0.8
Fully packaged		Yes	–	Yes	Yes
Depth/Volume (mm^2)		11	–	30.3	61.7
Technology (nm)		65	180	350	65
Total power (μW)		–	0.9	296*	28.8
IC area (mm^2)		500×450	90×330	1050×1050	500×500
Wake-up time (μs)		3.3	1000	110	11
THD (dB) @10mV$_{pp}$/ @20mV$_{pp}$		–	–	–	-52/-44
Static nonlinearity (%)		–	49† @6mV$_{PP}$	–	1.2@20mV$_{PP}$
LNA	noise floor (μV$_{rms}$)	180	42	3.78	5.3
	Bandwidth (kHz)	5	10	12	5
	power (μW)	–	0.52	42.9	4
	NEF/ PEF	–	12.3/136	6.6/78.4	5.87/34.55
	Gain (dB)	0	24	53	24

†estimated
*after AC-to-DC

Figure 17.5.7: Chip and assembled mote micrograph. In vitro measurement setup and performance comparison table.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

17.6 A Sub-40μW 5Mb/s Magnetic Human Body Communication Transceiver Demonstrating Trans-Body Delivery of High-Fidelity Audio to a Wearable In-Ear Headphone

Jiwoong Park, Patrick P. Mercier

University of California, San Diego, La Jolla, CA

Emerging wearable devices such as wireless headphones, smart glasses, and medical monitors require increasingly high-throughput wireless communications at ultra-low-power. Since far-field RF has significant path loss around the human body (e.g., up to 70dB at 2.4GHz), most RF body-area-network (BAN) systems such as Bluetooth Low Energy (BLE) have significant energy-expensive amplification requirements consuming milliwatts of power, and thus do not meet the energy demands of emerging small devices. Popular wireless earbuds, for example, only achieve a battery life of a few hours. Exasperatingly, emerging high-fidelity streaming audio and video content requires higher data rates than what BLE can currently accommodate. Human body communication (HBC) systems, for example ones based on electric fields (eHBC), in theory have lower path loss and can thus potentially offer more efficient links [1-2]. However, measurements from form-factor-accurate prototypes reveal path loss that is still rather large (e.g., 30-to-45dB across 20cm [3]), with unfortunately severe variation with posture and environments that requires energy-expensive compensation. Since the human body is magnetically inert, magnetic HBC (mHBC) systems, illustrated in Fig. 17.6.1, offer much lower path loss (e.g., 5-to-30dB over 1m [4]) without severe variation, and can thus theoretically achieve lower communication energy. However, to date there has not been any mHBC transceiver (TRX) developed to exploit this inherently efficient communication channel, and if there were, the high Q of employed coils would limit data rate of a straightforward approach to <800kb/s.

This paper presents the design of an mHBC TRX that consumes <40μW at a data rate of 5Mb/s, for the lowest energy/bit amongst pragmatic prior art wireless meter-range BAN transceivers. Efficient operation is achieved by: 1) exploiting the intrinsically low path loss between high-Q mHBC coils to reduce TX output power and RX gain/noise requirements; 2) employing an mHBC coil as a high-Q resonant element in an energy-efficient directly-OOK-modulated power oscillator (PO); 3) communicating at a high data rate (5Mb/s) despite the high-Q TX coil, which nominally limits bandwidth to <800kHz, via an all-digital frequency-locked-loop (ADFLL)-based synchronous injection-locked accelerated quenching and kick-start circuit, which also serves to dynamically tune the TX resonant capacitor to maintain center frequency lock between TX and RX while compensating the impedance variation of body-worn coils; and 4) biasing dynamic-threshold MOS (DTMOS) RX amplifiers and envelope detector (ED) in deep subthreshold, all at link-budget-appropriate noise levels.

A block diagram of the proposed mHBC TRX is shown in Fig. 17.6.2. The TRX utilizes a power oscillator to directly generate the 40MHz carrier by using the coil itself as the resonant element in the oscillator [5], as shown in Fig. 17.6.3. However, since the Q of the TX coil can be large (e.g., 50), driving the PO on and off for OOK modulation via the tail transistor cannot be normally accomplished at a data rate faster than 800kb/s. This is one of the reasons why conventional near-field communication (NFC) protocols implementing low-frequency (<13.5MHz) near-field coupling (as opposed to simultaneous near- and far-field coupling at a higher frequency in mHBC [4]) cannot support sufficiently high data rates. Here, to increase data rate, transistors M_{s1} and M_{s2} are deployed to rapidly shunt the inductor current during transmissions of logic '0's. On the other hand, to rapidly kick-start oscillations when transmitting logic '1's, which would normally take more cycles at 40MHz than are available in a 5Mb/s symbol period, injection transistors M_{i1} and M_{i2} are deployed, improving oscillation startup time by 20×. To maximize the effect of injection-locked kick-start, the resonant frequency of the PO, which can change slightly with posture or environmental variation (though not as severely as in eHBC), is precisely set to the same frequency as the crystal-based injection source for synchronous injection via a 4×4 unit-capacitor array that is dynamically tuned via an ADFLL, also shown in Fig. 17.6.3. This also helps to minimize any frequency mismatching between TX and RX. The fast control bit output (2.5Moutput/s) of the ADFLL enables frequency tuning within 10-to-20μs every 1-to-10ms to minimize wait time during calibration.

The RX, shown in Fig. 17.6.3, is designed using a common-source (CS) preamplifier, an active CS envelope detector (ED), a single-ended-to-differential (S-to-D) buffer, and variable gain amplifier (VGA). All RX circuits are biased in subthreshold and utilize DTMOS inputs to improve transconductance or conversion gain by up to 32% iso-current. The ED output drives a S-to-D buffer designed with low-V_t devices to minimize input capacitance to help ensure 2.5MHz bandwidth at the output of the low-current ED; the S-to-D then drives a differential 3b VGA. The VGA output is bit-sliced via a regenerative comparator at the appropriate common-mode voltage thanks to the VGA's common-mode feedback (CMFB) circuit.

The mHBC transceiver is fabricated in 0.12mm^2 of core area in 65nm CMOS. Figure 17.6.4 (top left) shows the output waveform of the TX under worst-case modulation of alternating 1s and 0s at 5Mb/s. Thanks to the accelerated quenching and kick-start circuit, the ASK-like modulation index improves from 15.4% to 84.7% compared to a conventional power oscillator, thereby enabling the desired high data rate in a manner that is compatible with the low-complexity RX. The spectrum of a signal received by an mHBC coil across a 10cm channel is shown in Fig. 17.6.4 (top right). Here, it can be seen that the ADFLL automatically re-tunes the intentionally-mis-tuned PO resonance frequency (~40.4MHz) of TX to the correct carrier frequency (40MHz), while also demonstrating that injection locking further improves total TX output power by 17.4dB while providing the rapid kick-start behavior described above. The inset plot illustrates the frequency-tracking capability of the TX, while validating the fast frequency settling (<10μs) of the ADFLL. BER testing in Fig. 17.6.4 (bottom left) under asynchronous operating conditions with 2× oversampling demonstrates receiver sensitivities of -63.5dBm to -56dBm at data rates from 1.25-to-5Mb/s, respectively, and from -57dBm to -54dBm according to VGA gain setting at 5Mb/s. When operating over a distance of ~1m across a human body, a pair of mHBC TRXs, with -24.8dBm TX output power and the -56dBm RX sensitivity setting, achieved a TX-to-RX latency at 5Mb/s of 388ns with error-free operation, as shown in Fig. 17.6.4 (bottom right).

To demonstrate feasibility of a practical application, two-channel analog audio data was generated by a smartphone and delivered via a two-channel 24b ADC and AES3 encoder to an mHBC TX employing a 2-turn 1mm width coil printed on the outline of 11×5.5cm^2 1oz copper PCB mounted on the smartphone. An mHBC RX, employing a 32mm-diameter coil mounted within an in-ear headphone prototype, wirelessly received data, playing audio data through a speaker via an AES3 decoder and DAC. A block diagram of the test setup is shown in Fig. 17.6.2 (bottom), a photograph of the test in progress is shown in Fig. 17.6.5, and the developed components are shown in Fig. 17.6.6 (top). Here, audio is shown to be delivered from the smartphone case to the headphone prototype across the body with no bit errors and <90μs latency (dominated by the 88μs of latency of the AES3 decoder).

As summarized in Fig. 17.6.6 (bottom), the TX consumes 18.6μW at -24.8dBm of output power, resulting in an efficiency of 3.7pJ/bit at 5Mb/s. At -56dBm sensitivity, the RX consumes 6.3μW for an efficiency of 1.3pJ/bit. When including the power of a 17.2μW crystal oscillator, the TX and RX consume 35.8μW (7.2pJ/bit) and 23.5μW (4.7pJ/bit), respectively, all while reliably covering a ~1.5m BAN. This is the lowest power and most efficient TRX amongst prior-art that reliably communicate over a BAN and that includes the power consumption of frequency synthesis. A die micrograph is shown in Fig. 17.6.7.

Acknowledgement:
This work was supported in part by the National Science Foundation under Grant No. 1751293.

References:

[1] S. Maity, et al., "A 6.3pJ/b 30Mbps -30dB SIR-tolerant Broadband Interference-Robust Human Body Communication Transceiver using Time Domain Signal-Interference Separation," *IEEE CICC*, pp. 1–4, Apr. 2018.
[2] J. Jang, et al., "4-Camera VGA-resolution capsule endoscope with 80Mb/s body-channel communication transceiver and Sub-cm range capsule localization," *ISSCC Dig. Tech. Papers*, pp. 282–283, Feb. 2018.
[3] J. Park, et al., "Channel Modeling of Miniaturized Battery-Powered Capacitive Human Body Communication Systems," *IEEE. Trans. Biomed. Eng.*, vol. 64, no. 2, pp. 452–462, Feb. 2017.
[4] J. Park, et al., "Magnetic Human Body Communication," *IEEE EMBC*, pp. 1841–1844, Aug. 2015.
[5] P. P. Mercier, et al., "A Sub-nW 2.4 GHz Transmitter for Low Data-Rate Sensing Applications," *IEEE J. Solid-State Circuits*, vol. 49, no. 7, pp.1463-1474, July 2014.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

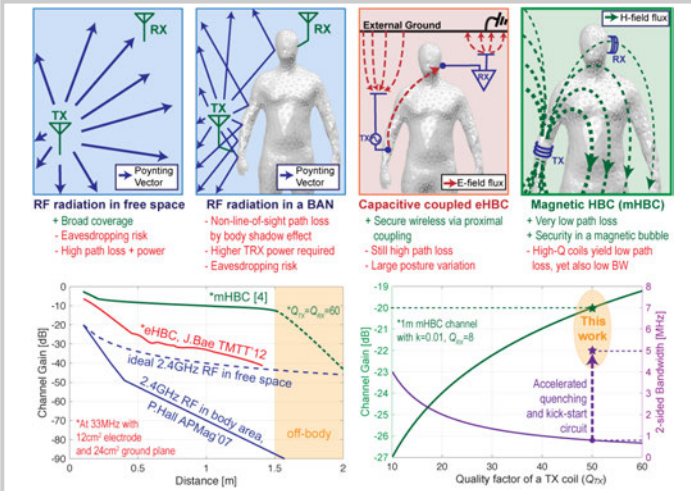


Figure 17.6.1: The path loss of far-field RF, eHBC, and mHBC compared (top and bottom left); increasing the Q of the TX coil decreases path loss, but also nominally decreases bandwidth without the proposed accelerated quenching and kick-start circuit (bottom right).

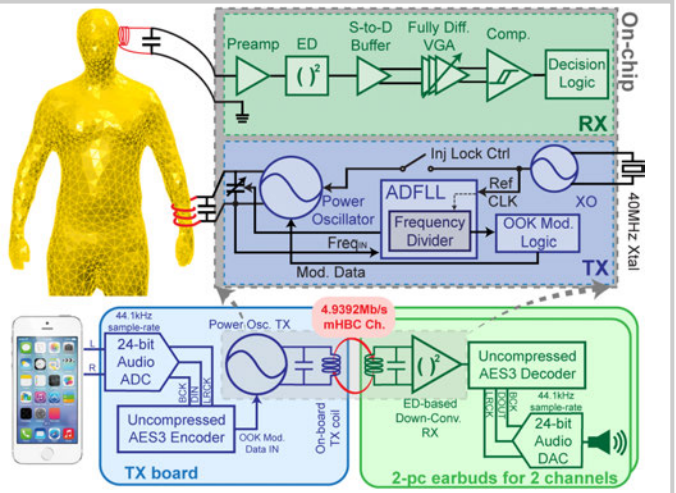


Figure 17.6.2: Block diagram of proposed mHBC transceiver (top); block diagram of the audio streaming demonstration test setup delivering two channels of uncompressed 44.1kHz 24b audio to demonstrate the high-throughput capabilities of the mHBC link (bottom).

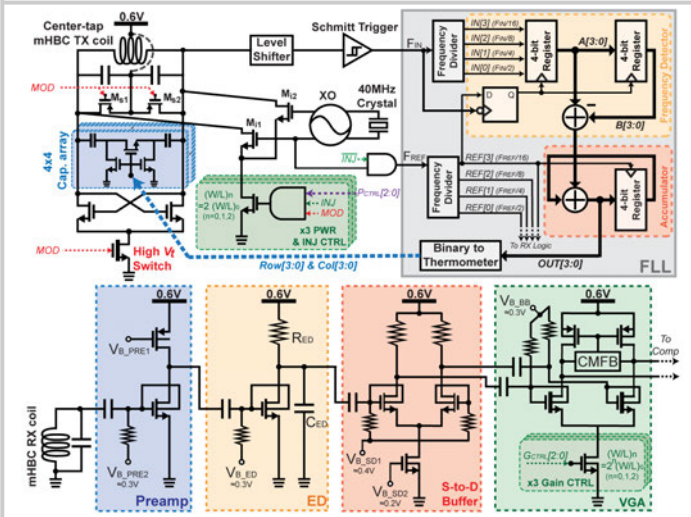


Figure 17.6.3: Schematics of the TX power oscillator and injection locking circuit for the faster-than-Q data-rate modulation (top); the sub-Vt DTMOS envelope-detector based RX (bottom).

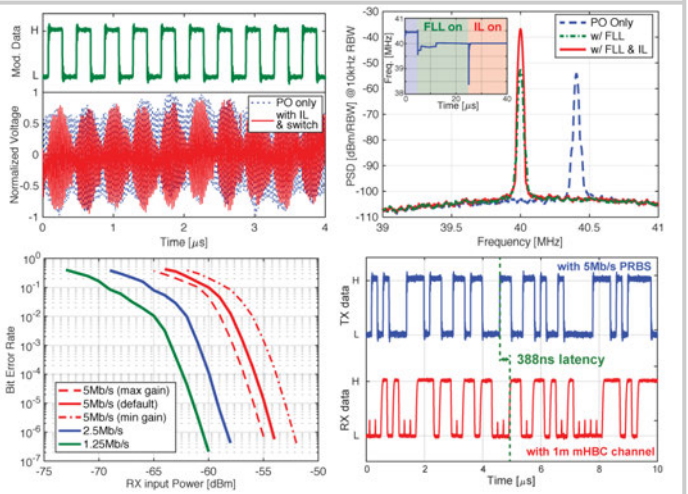


Figure 17.6.4: Measured 5.5× improved worst-case OOK-modulated TX output waveform (top-left), measured spectrum and frequency transient (inset) of dynamic resonance tuning by the ADFLL (top-right), measured RX BER curves (bottom-left), and measured cross-body waveform results (bottom-right).

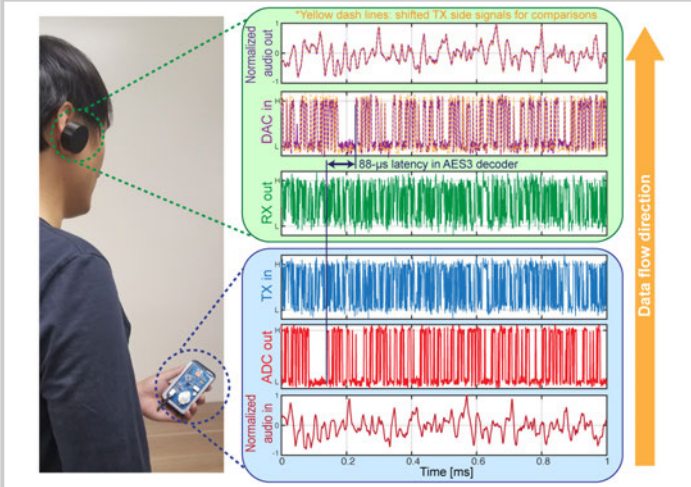
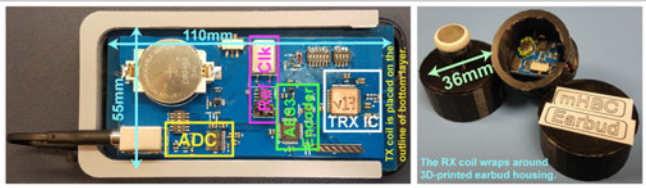
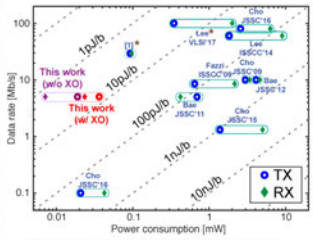


Figure 17.6.5: System demonstration results of high-fidelity audio data streaming (2-channel 24b 44.1kHz audio) via an mHBC link supporting 4.9392Mb/s AES3 format to demonstrate the high-throughput capabilities of the mHBC TRX.



Setting			Min.	Default	Max.
XO		Power [μW]		17.2	
TX (w/ 50% OOK Mod.)		Output power [dBm]	-34.1	-24.8	-22.3
	Power [μW]	PO		5.1	
		Inj. Lock	2	13.1	20.9
		ADFLL*		0.02	
		Ctrl Logic		0.34	
		Total	**7.46**	**18.56**	**26.36**
		E/bit [pJ/bit]	**1.5**	**3.7**	**5.3**
RX		Sensitivity [dBm] (0.1% BER)	-54	-56	-57
	Power [μW]	Preamp		1.8	
		ED		0.53	
		S-D buffer		2.6	
		BB VGA	0.94	1.3	1.9
		Comp. + Logic		0.07	
		Total	**5.94**	**6.3**	**6.9**
		E/bit [pJ/bit]	**1.2**	**1.3**	**1.4**

* Note : Operates for 20-μs duration every 1 ms.



*Note: Not including the power of frequency synthesis.

Figure 17.6.6: Photograph of the battery-operated audio streaming prototypes: the TX unit fits within a smartphone case (top-left) and the earbud-based RX prototypes were 3D printed for demonstration purposes (top-right); table of specifications (bottom-left), and the state-of-the-art comparison chart (bottom-right).

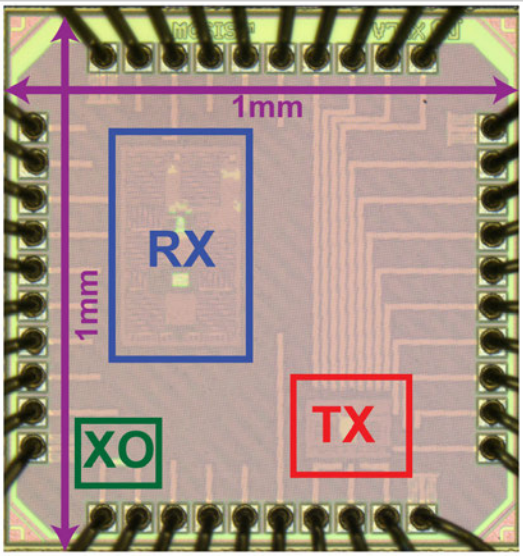


Figure 17.6.7: Die micrograph of the proposed mHBC transceiver.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

17.7 A 7.0fps Optical and Electrical Dual Tomographic Imaging SoC for Skin-Disease Diagnosis System

Yongsu Lee, Kwantae Kim, Jiwon Lee, Kyoung-Rog Lee, Surin Gweon, Minseo Kim, Hoi-Jun Yoo

KAIST, Daejeon, Korea

Recently, skin disease patients have been rising rapidly, and more than 5.4 million cases of skin cancer were treated in over 3.3 million people in the U.S. in 2012, and more people are diagnosed with skin cancer each year than all other cancers combined in the U.S. [1]. Because a malignant lesion is difficult to distinguish from pigmented nevus, elaborate detection devices are needed for accurate diagnosis [2]. Optical Coherence Tomography and Laminar Optical Tomography are mainly used for skin disease diagnosis, however, equipment for those tomography are too bulky to be used in house. The Electrical Impedance Spectroscopy is also applied to analyze skin disorders [3], but its equipment requires assistance from experts.

In this paper, a coin-sized light-weight (3g) and portable system that simultaneously takes optical and electrical domain skin images is proposed for accurate skin disease diagnosis. It uses Diffused Optical Tomography (DOT) and Electrical Impedance Tomography (EIT) concurrently for hemodynamic and electrical characteristics measurements of the skin.

Figure 17.7.1 shows the proposed skin disease diagnosis system. The circular system has only a 29mm diameter and 3g weight enabling a seamless measurement on any part of the human body. The system is composed of two modules: Micro-sensor module and chip module. Micro-sensor module is implemented on flexible PCB substrate containing 19 micro-cells. The micro-cell consists of 3 micro-LEDs with different wavelengths (622, 528, and 470 nm), a photodiode (PD) for DOT, and a surrounding electrode for EIT. Three wavelengths are selected because the number of main absorbers in human skin is three with different absorption peak wavelengths, which is melanin, Hb, and HbO_2. Hemodynamic characteristics is extracted from DOT for diagnosis of Angioma. Frequency difference EIT (FD-EIT) is adopted for acquiring electrical information of lesion, which can be used for diagnosis of Melanoma.

Figure 17.7.2 shows the overall block diagram of the SoC. It consists of: 1) Power management unit (PMU) for generating stable two voltage domains (7V for V_{LED}, 1.8V for V_{DD}), 2) Current driver for 3 wavelengths CDMA DOT and 2 frequencies combined generator for FD-EIT, 3) DOT receiver with fast settled DC rejection scheme, 4) EIT receiver implemented with weaver demodulator and IQ phase calibrator, and 5) Digital module for data acquisition and communication with external devices.

Figure 17.7.3 describes the dual calibration loop between boost converter and DOT current driver. The switched capacitor type is adopted because higher efficiency can be achieved in <10mW output power range compared to inductive boost converter [4]. The boost converter consists of 5 cells of 1:2 bi-phasic voltage doubler. The switch configuration controlled by 5-bit control signal (S_C), is determined according to comparison results between battery voltage (V_{BAT}) and reference voltage. S_C switches the connection of unit cells to boost V_{LED} by an initial coarse step voltage increase. Then, V_{LED} calibration loop controls the output impedance of boost converter with negative feedback. The VCO at the next stage of analog comparator controls operation frequency (f_{CLK}) of boost converter, and stabilizes V_{LED} to 7V with fine voltage step. To avoid breakdown from high V_{GS} (>5 V), double-step switch is proposed. The (A_0 – 1) V is generated from a charge pump and connected to the V_{SS} terminal of the level shifter circuits, and then V_{GS} of M_0 and M_1 is always < 5V. Due to process mismatch and aging, the value of Extinction Ratio (ER), $\Delta P/\Delta I$, may vary LED by LED. The PD (mPD) from inner micro-cell monitors ER, and 4-bit controlled current steering DAC (I_{EDAC}) increases current level until ΔI equals to I_{ER}.

Figure 17.6.4 shows the circuit schematics and the measurement results of the DOT receiver. To measure accurate hemodynamic characteristics from optical sensing, fast imaging speed (3-to-100 fps) and wide dynamic range (>80 dB) are required. The settling time of analog DC rejection scheme is too long (>1s) to measure constant data, and digital DC rejection scheme has limited dynamic range due to quantized rejection resolution [5]. To achieve short settling time and high resolution, duty-cycled resistor based DC rejection loop is proposed. In the feedback loop, the cut-off frequency is decided with R_E of which value is amplified by 1/D with the help of duty-cycled resistor. The "D" is derived from the ratio of pulse-width to period of clock. When the clock period sweeps from small to large values, the cut-off frequency decreases while reducing the settling time with the same DC rejection. When there is 1μA input current fluctuation, 100 times lower settling time (50μs) is measured compared to the fixed period case as shown in the measurement results. To reduce the effect of parasitic capacitance (C_P) between switching MOSs, PCIS is proposed as shown in the figure. In addition, input capacitance (C_{IN}) cancellation scheme is adopted to increase the dynamic range of TIA. C_{IN} is the main cause of input referred noise, and positive feedback loop with source follower [6] compensates for it. Therefore, 100.4dB dynamic range of TIA is measured with 9.57pA_{rms} [50Hz] input referred noise.

Figure 17.7.5 shows the architecture and schematics of the EIT receiver. The weaver architecture is commonly used as the image rejection scheme in RF circuit [7]. The proposed modified weaver demodulator utilizes the image rejection channel as f_2 channel for FD-EIT. This architecture is effective for narrow bandwidth between f_1 (100kHz) and f_2 (80kHz) of FD-EIT without necessity of high order low-pass filter. In the FD-EIT, since only 1° phase error leads to 2.8% demodulation error, phase detector and phase interpolator are used to compensate for I/Q clock phase mismatch.

Figure 17.7.6 shows the measurement results of an in-vitro and in-vivo environment. 1.2% w/w agar phantom is used instead of human skin. A 5mm diameter of India-ink-mixed agar is modeled as pigment nevus, and 5mm × 5mm square-cut ink-dyed carrot is utilized as a malignant lesion. From the DOT image, two regions are unable to be distinguished due to similar absorbance parameter. However, an EIT image differentiates two regions due to different electrical characteristics with a blurred boundary image. By combining DOT and EIT together, images with both advantages, high sensitivity (0.5mm) and high selectivity (2mm) can be obtained. In addition, even a 3-D image can be reconstructed, and 80% thresholding shows the image with positions of only malignant lesion. Two kinds of letter (N, S) tattoo are tested for the in-vivo measurements, and the DOT image shows a similar shape to the real image. 3-wavelength measurements can extract the change of HbO_2 while pressure is applied to skin.

Figure 17.7.7 shows the chip micrograph, performance summary of SoC, and comparison table between DC rejected optical sensors. The proposed SoC is fabricated in 180nm high voltage CMOS process. The die area is 3.3mm × 5.0mm. The average power consumption is 7.54mW, and 2.8 ~ 4.2 V input voltage convert to 1.8V for V_{DD} and to 7V for V_{LED}. CDMA DOT current driver operates with 16MHz clock frequency and 3.9kHz chip rate to achieve 7.0fps hemodynamics imaging speed. Current range is wide from 0.05mA to 6.4mA. The DOT receiver achieves a fast settling time withq DC rejection (50s) and wide dynamic range (100.4dB). Weaver demodulator based EIT receiver operates with I/Q mismatch calibrator. As a result, 0.5mm sensitivity and 2mm selectivity dual tomographic imaging can be obtained with the proposed SoC for skin disease diagnosis system.

References:

[1] American Cancer Society, "Cancer Facts and Figures 2018."

[2] T. J. Muldoon, et al., "Analysis of skin lesions using laminar optical tomography," *BIOMEDICAL OPTICS EXPRESS*, pp. 1701-1712, June 2012.

[3] J. Malvehy, et al., "Clinical performance of the Nevisense system in cutaneous melanoma detection: an international, multicentre, prospective and blinded clinical trial on efficacy and safety," *British Journal of Dermatology*, pp. 1099-1107, Oct. 2014.

[4] G. Villar-Pique, et al., "Survey and Benchmark of Fully Integrated Switching Power Converters: Switched-Capacitor Versus Inductive Approach," *IEEE Trans. Power Electronics*, vol. 28, no. 9, pp. 4156-4167, Sept. 2013.

[5] E. S. Winokur, et al., "A Low-Power, Dual-Wavelength Photoplethysmogram (PPG) SoC With Static and Time-Varying Interferer Removal," *IEEE TBCAS*, vol. 9, no. 4, pp. 581-589, Aug. 2015.

[6] C.-M. Tsai, "A 40mW 3Gb/s Self-Compensated Differential Transimpedance Amplifier With Enlarged Input Capacitance Tolerance in 0.18μm CMOS Technology," *IEEE JSSC*, vol. 44, no. 10, pp. 2671-2677, Oct. 2009.

[7] S.-C. Hwu and B. Razavi, "An RF Receiver for Intra-Band Carrier Aggregation," *IEEE JSSC*, vol. 50, no. 4, pp. 946-961, Apr. 2015.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

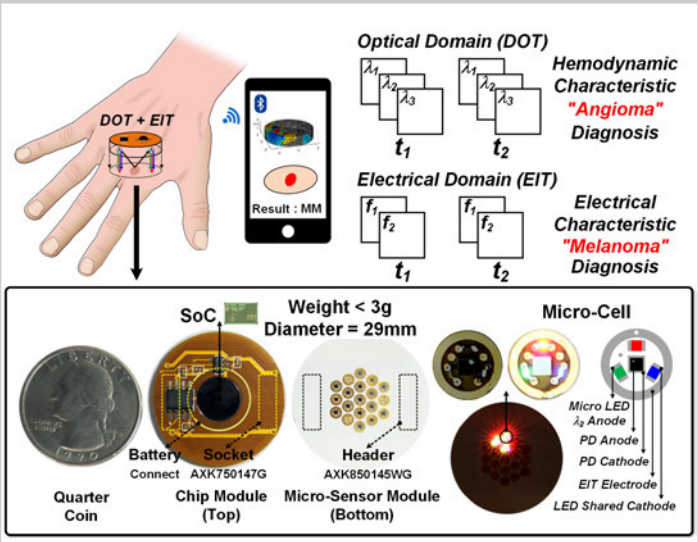


Figure 17.7.1: Proposed skin disease diagnosis system.

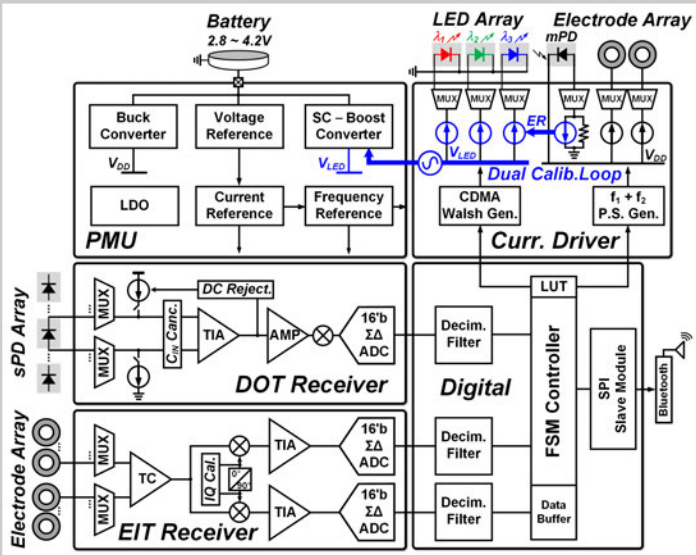


Figure 17.7.2: Overall block diagram of SoC.

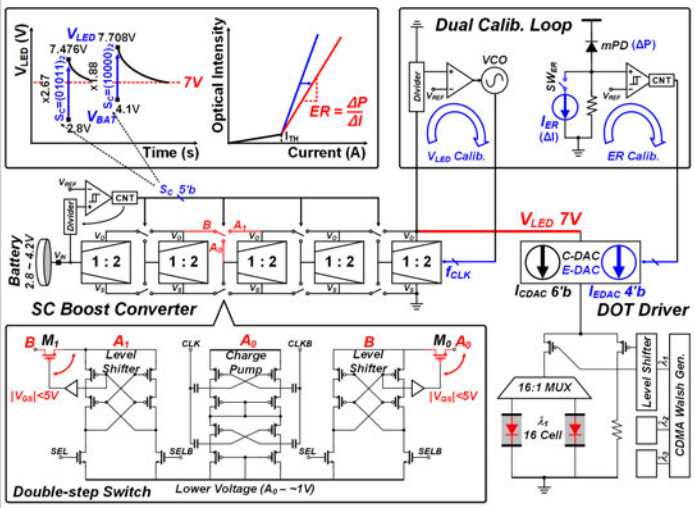


Figure 17.7.3: The dual calibration loop between boost converter and DOT current driver.

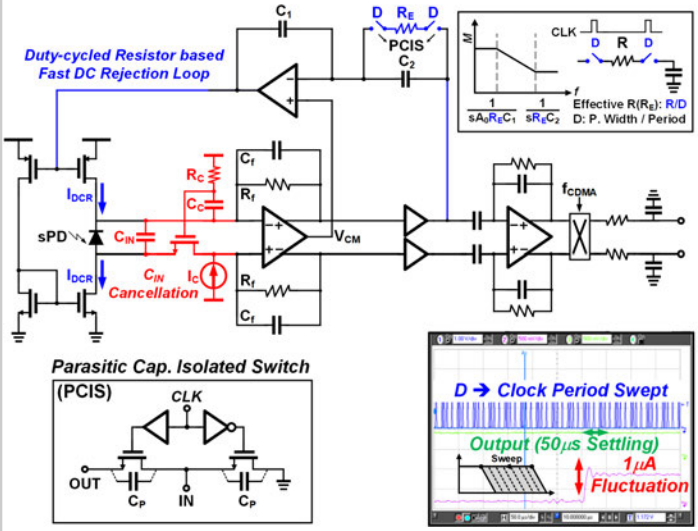


Figure 17.7.4: The circuit schematics and the measurement results of DOT receiver.

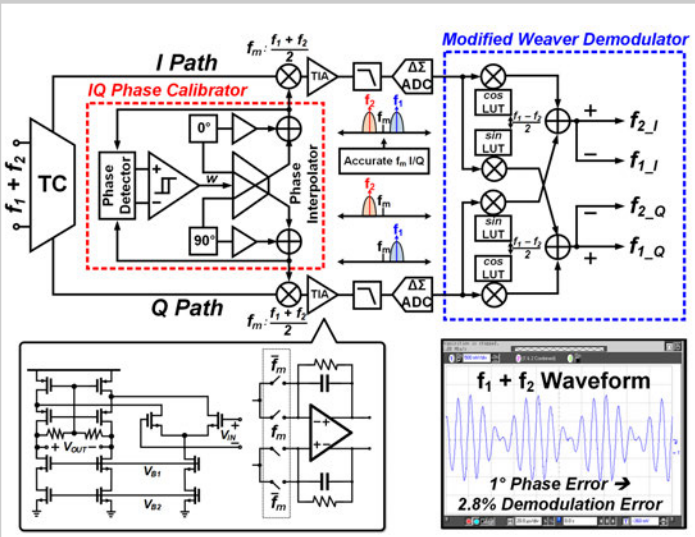


Figure 17.7.5: The architecture and schematics of EIT receiver.

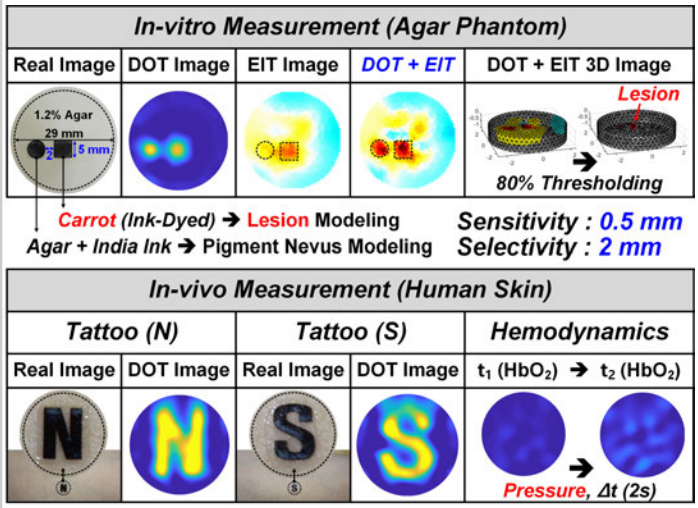
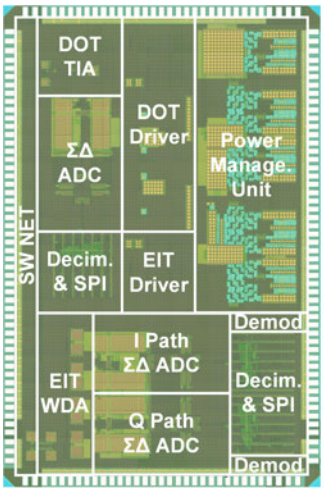


Figure 17.7.6: The measurement results of in-vitro and in-vivo environment.

17



Process	180 nm High Voltage CMOS	
Die Size	3.3 x 5.0 mm (Including Pads)	
Supply Voltage	1.8 V for V_{DD}, 7 V for V_{LED}	
Power Consump.	7.54 mW for Average	
Power Management Unit	Input Voltage	2.8 V ~ 4.2 V
	Peak Efficiency	Worst 51 %, Best 74 %
	Calibration Scheme	Coarse(5'b), Fine 2 Steps
CDMA DOT Current Driver	Current Range	0.05 ~ 6.4 mA
	Operation Freq.	Clk: 16 MHz, Chip Rate: 3.9 kHz
	Calibration Scheme	4'b Extinction Ratio Calib.
DOT Receiver	Frame Rate	7.0 fps
	Dynamic Range	100.4 dB (TIA DR)
	DCR Settling Time	50 μs
EIT Receiver / Current Driver	Frequency	f_C: 10 k ~ 80 kHz, f_2: 20 k ~ 100 kHz
	Calibration Scheme	I/Q Mismatch Calibrator
	Current Range	0.1 ~ 1 mA

	AFE 4403	TBCAS 08	TCAS 16	TBCAS 15	JSSC 16	This Work
Application	PPG, SpO_2	PPG	Optical Comm.	PPG	PPG	DOT
TIA Dynamic Range (dB)	105	29.8	45	57	89	100.4
DC Rejection (DCR) Type	Digital DCR	Analog DCR	Digital DCR	Digital DCR	Digital DCR	Fast Settled Analog DCR
DCR Automation	X	O	X	O	O	O
Settling Time (s)	N/A	1.68	N/A	100 m	N/A	50 μ
CMDR (dB)	20 dB (1~10μA)	54.5 dB (0.1~53.4μA)	N/A	N/A	N/A	> 50 dB

Figure 17.7.7: Chip micrograph, performance summary and TIA comparison table.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

17.8 A 2.6μW Monolithic CMOS Photoplethysmographic Sensor Operating with 2μW LED Power

Antonino Caizzone, Assim Boukhayma, Christian Enz

EPFL, Neuchâtel, Switzerland

Photoplethysmography (PPG) is a key technology allowing non-invasive monitoring of vital indicators such as heart rate (HR) and oxygen saturation (SpO_2). Today, the total PPG sensor power consumption is dominated by the few tens of mA of the LEDs driving current. Different solutions have been proposed to solve this bottle-neck, either by reducing the LED duty cycle [1-2] or by non-uniform sub-sampling [3]. A heart-beat-locked loop system that significantly reduces the LED power has been recently demonstrated [4]. However, this power reduction comes at the cost of more complexity since it requires a non-trivial heart beat prediction scheme and intrinsically hinders the full PPG wave representation.

In this work, we address the power issue by combining a high sensitivity photo-detector with an ultra-low noise and low power readout chain achieving the same signal-to-noise ratio (SNR) compared to conventional solutions, but at significantly lower LED power. We present a PPG sensor integrating an array of pinned-photodiodes (PPD), commonly used in CMOS imagers for achieving sub-electron noise [5]. The PPDs array is implemented on the same chip with the full analog front-end (AFE) including the analog-to-digital conversion (ADC). The full CMOS integration allows one to dramatically reduce the parasitic capacitance leading to a larger conversion gain and a lower noise. This approach also provides higher miniaturization and lower cost compared to traditional solutions with off-chip photodiodes. The use of an array additionally enables spatial averaging leading to further noise reduction. Consequently, the LED power can be reduced dramatically. A prototype has been implemented in a 0.18μm CMOS Image Sensor (CIS) process, achieving 4.6μW total power consumption, including 1.97μW LED power, at 1.38bpm HR average error.

The block diagram of the PPG sensor is shown in Fig. 17.8.1. It consists of a fully integrated chip embedding an array of PPDs, a passive averaging block, a switched-cap (SC) amplifier and an ADC. The photosensitive area consists of four clusters of 50 rows and 256 columns. The 50 pixels along the i-column are assembled in a macro-pixel (MP). The pixels of one MP share the same source-follower (SF) saving power while maintaining a reasonably low parasitic capacitance at the shared sense node (SN). The spatial averaging is done in two steps: first, a charge averaging on the shared SN within the same column and, second, a voltage averaging among the 1024 columns at the SFs output. This leads to a considerable shot and read noise reduction, together with a full pre-filtering of the PPG signal itself. Next, the PPG signal is amplified by a low noise programmable gain SC amplifier. In addition to the noise optimization, the chosen pixel achieves between 66% and 75% quantum efficiency (QE) for the selected LED wavelengths (Fig. 17.8.4) further reducing the LED power needed for achieving the target SNR [1].

Figure 17.8.2 shows the circuit implementation together with the timing diagram. The sense nodes, shared by 50 PPDs and corresponding to the SFs inputs, are first reset. In parallel, each individual PPD starts integrating the impinging light corresponding to the ambient light (AL). In order to precisely control the light-induced charge integration, the PPDs are first emptied by the sink switch TGs. At the end of the first integration phase, the generated photoelectrons are transferred, via the transfer gate TGt, to the SNs. The capacitance of each SN, shared by the 50 pixels along the same column, converts the integrated photoelectrons into a voltage and performs the charge averaging across the 50 rows. After the transfer is completed the SF output voltages are sampled on capacitors $C1_i$, via $S1_i$. Next, the SNs are reset again and the LED is pulsed on. As above, the PPDs precisely integrate the LED light superimposed to AL, and the related voltage level is sampled on $C2_i$, via $S2_i$. The power switch S_{SF} is only closed during this charge transfer and sampling phase for minimal power consumption. By closing $S3_i$ and $S4_i$ all the capacitors $C1_i$ and $C2_i$ related to the same sample are connected in parallel and share their charge, resulting into a voltage equal to the passive spatial average of the array pixel output samples. In addition, the large capacitor resulting from the parallel connection of multiple column-level capacitors acts as a large hold capacitor for the following stage. The full averaging operation comes with a shot noise variance reduction of 50×1024 and a read noise variance reduction of 1024. In order to extract the difference of the two averaged samples, corresponding to the AL and the AL plus LED light, a SC amplifier is operated as shown in Fig. 17.8.2: first, the averaged AL sample is stored in C3, via S5 and SAZ. Then, SAZ is opened and the charge stored in C3 is transferred to C4. At the closing of S6, the amplified difference between the two average values is obtained at the amplifier's output, leading to an efficient AL cancellation. The amplifier embeds a programmable gain, from 1 to 32 to adapt to different operating conditions. It then drives a 14b incremental ADC, through SF2. The power switches S_{Amp} and S_{ADC} are only closed at the beginning of the array averaging and opened after the digitization.

A die micrograph of the 180nm chip is shown in Fig. 17.8.7. The linearity of the light to digital conversion and the total noise measurements are shown in Fig. 17.8.3. The sensor is exposed to an LED shining at increasing driving current and the resulting output digital number is acquired. The programmable amplifier gain has been set to 8, which enables a wide range of emitting light conditions without saturation. With the exception of sub-mA LED operations, the chip shows ±3% non-linearity in the light to digital conversion. Note that this includes all the sources of non-linearity from the LED, the readout chain to the ADC. Figure 17.8.3 also shows that the output noise standard-deviation (STD) remains constant across the dynamic range, demonstrating the effectiveness of the shot noise reduction by spatial averaging. The total noise measured at the output of an off-chip 11 taps FIR low-pass filter is 3.1 DN_{RMS} in average corresponding to an input-referred noise as low as 0.68 e^-_{RMS} per PPD, thanks to the noise shaping introduced by the incremental ADC. Indeed, the noise measured directly at the output of the ADC corresponds to 9.43 DN_{RMS}. This accounts for all the noise components including the readout noise, quantization noise, shot noise and LED flicker noise.

An in-vivo acquisition of PPG has been performed in reflection mode on the index finger. Figure 17.8.4 shows the LEDs and sensor module used for the experiment, as well as samples of the PPG signal obtained using different LED wavelengths at a record-low LED average power of 1.97μW. Figure 17.8.5 shows a comparison between the HR directly extracted from the sensor output and a commercial ECG chest strap featuring a HR average error of 1.38bpm only measured on three healthy male subjects. These measurements have been obtained with a sampling frequency of 40Hz, an average LED driving power of 1.97μW, at a duty cycle of 0.07%, and a readout (AFE+ADC) average power consumption of 2.63μW only. This is obtained, referring to Figure 17.8.2, by closing S_{SF} for 80μs and S_{Amp} with S_{ADC} for 0.5ms.

Compared to recent state-of-the-art (SOA), shown in Fig. 17.8.6, this work achieves the lowest power consumption. Moreover, among the most relevant SOA works, this is the only solution featuring a full integration of the photosensitive area together with the AFE and the ADC. This last feature, in addition to the extremely low power consumption, makes this solution a serious candidate for the future of wearable digital healthcare.

References:
[1] K. N. Glaros, et al., "A Sub-mW Fully-Integrated Pulse Oximeter Front-End," *IEEE TBioCAS,* vol. 7, no. 3, pp. 363-375, June 2013.
[2] E. S. Winokur, et al., "A Low-Power, Dual-Wavelength Photoplethysmogram (PPG) SoC With Static and Time-Varying Interferer Removal," *IEEE TBioCAS,* vol. 9, no. 4, pp. 581-589, Aug. 2015.
[3] P. V. Rajesh, et al., "A 172μW Compressive Sampling Photoplethysmographic Readout with Embedded Direct Heart-Rate and Variability Extraction from Compressively Sampled Data," *ISSCC Dig. Tech. Papers,* pp. 386-387, Feb. 2016.
[4] D. H. Jang, et al., "A 43.4μW Photoplethysmogram-Based Heart-Rate Sensor Using Heart-Beat-Locked Loop," *ISSCC Dig. Tech. Papers,* pp. 474-476, Feb. 2018.
[5] A. Boukhayma, et al., "A Sub-0.5 Electron Read Noise VGA Image Sensor in a Standard CMOS Process," *IEEE JSSCC*, vol. 51, no. 9, pp. 2180-2191, Sept. 2016.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

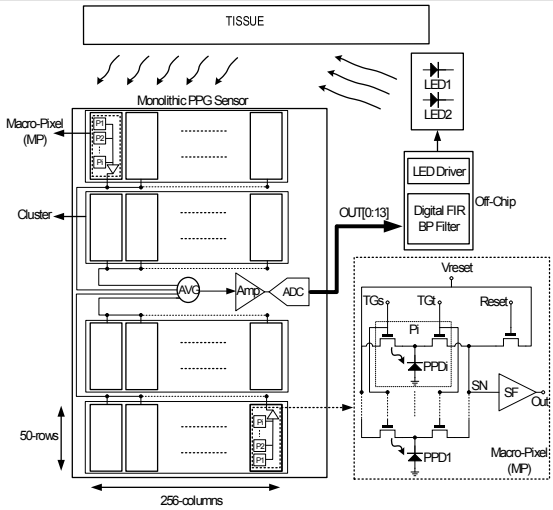


Figure 17.8.1: Block diagram of the monolithic PPG sensor embedding an array of macro-pixels, as photosensitive area, an averaging block, an amplifier and an ADC.

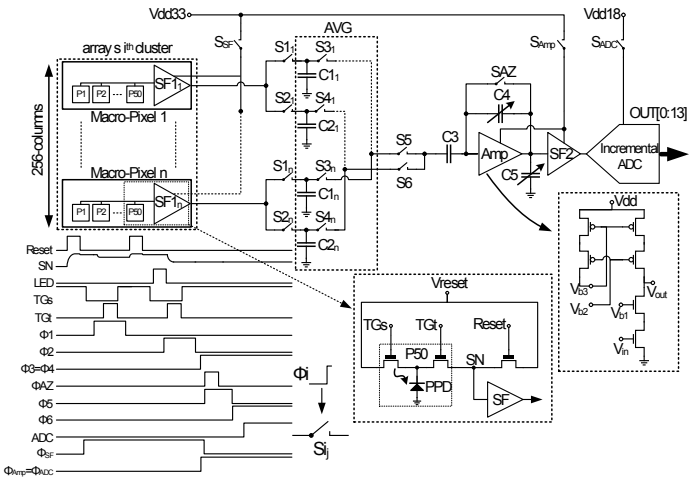


Figure 17.8.2: Architecture of the monolithic PPG sensor and detailed schematic of the OTA used in the amplifier, the averaging block, with the timing diagram.

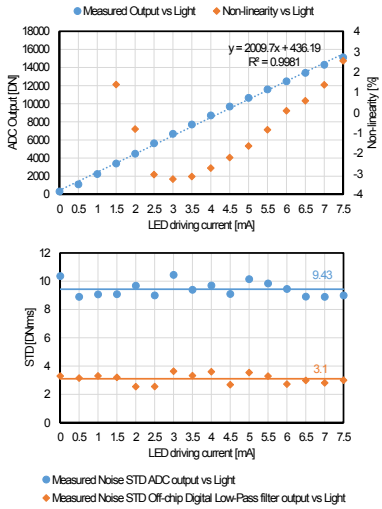


Figure 17.8.3: Measured light-to-digital linearity and total noise standard-deviation (STD), both at the ADC output and an off-chip FIR digital low-pass filter.

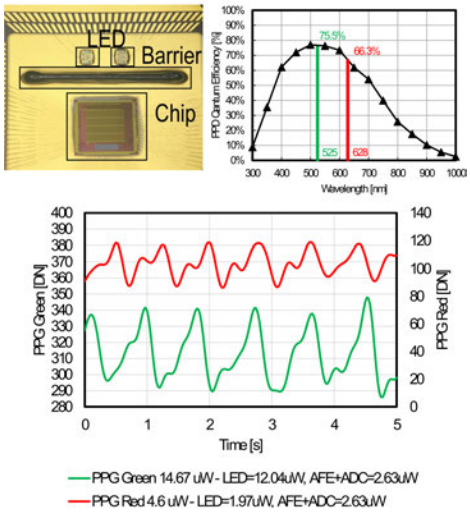


Figure 17.8.4: LEDs, chip and optical barrier. QE of the photodetector. PPG signals at two LEDs emitting power levels and colors and 2.63μW at the readout level.

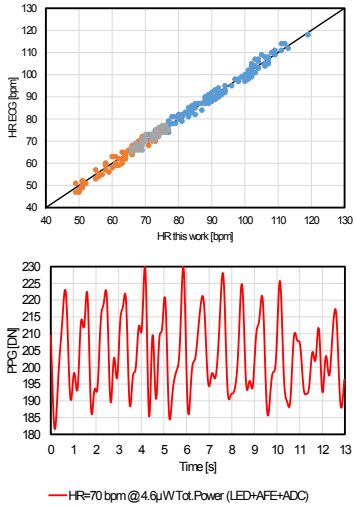


Figure 17.8.5: HR correlation with a commercial ECG chest strap. PPG signal of a subject. The experiment has been performed at 4.6μW total power (LED+AFE+ADC).

	This work	[2]	[3]	[4]
Year	2018	2015	2016	2018
Technology	180 nm	180 nm	180 nm	180 nm
Supply Voltage	3.3/1.8[1]	1.8	1.2	3.3
Sampling Frequency	40 Hz	165 Hz	4 Hz	100 Hz
LED Duty Cycle	0.07%	0.7%	0.0125%	0.0175%
Fully Integration[2]	Yes	No	No	No
Feature Extraction	HR/SpO_2/RR[3]	HR	HR	HR
Avg HR error	1.38 bpm[4]	NA	NA	NA
Max HR error	3 bpm[4]	NA	10 bpm	2.1 bpm
Power LED	1.97 μW	120 μW[5]	43 μW[5]	16 μW[5]
Power AFE+ADC	2.63 μW	216 μW	172 μW	27.4 μW
Power TOT[6]	4.6 μW	336 μW[5]	215 μW[5]	43.4 μW[5]

1: ADC supplied at 1.8V.
2: Photosensitive area+AFE+ADC.
3: HR (heart rate), SpO_2 (oxygen saturation), RR (respiration rate).
4: Measured at 4.6 μW total power vs a commercial ECG chest strap on 222 separate measurements on three healthy individuals, corresponding to data from Fig.5.
5: Best case from the given data.
6: HR mode.

Figure 17.8.6: Comparison table.

17

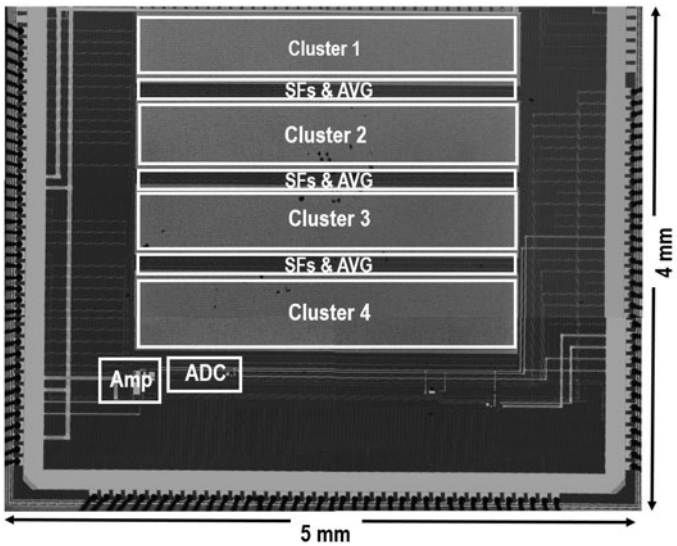


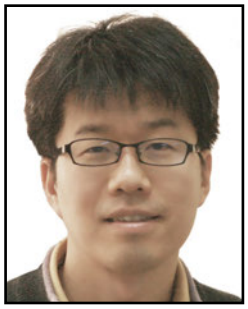
Figure 17.8.7: Micrograph of the fully integrated PPG sensor (4 mm × 5 mm).

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 18 Overview: *Analog Techniques*

ANALOG SUBCOMMITTEE

Session Chair: ***David Blaauw***
University of Michigan, Ann Arbor, MI



Associate Chair: ***Byungsub Kim***
POSTECH, Pohang, Gyeongbuk, Korea

Subcommittee Chair: Kofi Makinwa, Delft University of Technology, Delft, The Netherlands

This session presents advances in audio amplifiers, auto-zeroed buffers and capacitance-to-digital converters, followed by three papers describing low-power and fast-startup crystal oscillators and two papers describing new voltage references. The audio amplifier (Paper 18.1) achieves -105dBc THD+N. The auto-zeroed buffer (Paper 18.2) incurs a √2 noise penalty, which approaches the theoretical limit. The capacitance-to-digital converter (Paper 18.3) achieves a resolution of 0.29fF resolution with 16fJ/conversion-step. Two fast startup crystal oscillators (Papers 18.4 and 18.5) use dynamic frequency and phase locking, respectively, to achieve ~20μs startup time. A 32kHz crystal oscillator (Paper 18.6) consumes only 0.55nW. The final two papers (Paper 18.7 and 18.8) present advances in low-power and sub-1V voltage references.

1:30 PM

18.1 A -105dBc THD+N (-114dBc HD2) at 2.8V_{PP} Swing and 120dB DR Audio Decoder with Sample-and-Hold Noise Filtering and Poly Resistor Linearization Schemes

S-H. Wen, MediaTek, Hsinchu, Taiwan

In Paper 18.1, MediaTek describes an audio decoder that shifts reference noise to the sub-audio band and a new frequency compensation scheme that allows its load capacitance to be increased from 390pF to 22nF. The decoder achieves -105dBc THD+N (-114dBc HD2) at 2.8V peak-peak swing and 120dB dynamic range.

2:00 PM

18.2 A 16fJ/Conversion-Step Time-Domain Two-Step Capacitance-to-Digital Converter

X. Tang, University of Texas, Austin, TX

In Paper 18.2, the University of Texas, Austin describes a two-step capacitance-to-digital converter that combines a coarse SAR ADC and a fine time-domain-based Delta-Sigma Modulator. It achieves a resolution of 0.29fF and energy consumption of 16fJ per conversion-step.

2:30 PM

18.3 An Auto-Zero Stabilized Voltage Buffer with a Quiet Chopping Scheme and Constant Input Current

T. Rooijers, Delft University of Technology, Delft, The Netherlands

In Paper 18.3, Delft University of Technology describes an auto-zeroed voltage buffer with a novel chopped stabilization loop. This allows its low-frequency noise density to be lowered from √(5)× white noise to nearly √(2)×. It also achieves an offset of 0.4μV and 0.8pA input current which is nearly constant over input voltage.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

3:15 PM

18.4 A 0.55nW/0.5V 32kHz Crystal Oscillator Based on a DC-Only Sustaining Amplifier for IoT

H. Esmaeelzadeh, University of California, Los Angeles, CA

In Paper 18.4, the University of California at Los Angeles describes a 32kHz crystal oscillator that uses an up/down-conversion technique that allows its sustaining amplifiers to operate near DC. The oscillator achieves 0.55nW at 0.5V supply and an Allan deviation floor of 14ppb.

3:45 PM

18.5 A 54MHz Crystal Oscillator with 30× Start-Up Time Reduction Using 2-Step Injection in 65nm CMOS

K. M. Megawer, University of Illinois, Urbana-Champaign, IL

In Paper 18.5, the University of Illinois at Urbana-Champaign describes a fast crystal start-up technique that uses a two-step signal-injection technique based on a ring oscillator that is subsequently precisely tuned by a PLL. The 54MHz oscillator achieves 20μs startup time across a temperature range of -40ºC to 85ºC.

4:15 PM

18.6 A 32MHz Crystal Oscillator with Fast Startup Using Synchronized Signal Injection

B. Verhoef, Dialog Semiconductor, Hengelo, The Netherlands

In Paper 18.6, Dialog Semiconductor describes a fast crystal startup technique that uses dynamic phase alignment to allow quasi-synchronized signal injection. The 32MHz oscillator achieves 24μs startup time and 20.2nJ startup energy and demonstrates operation across -40ºC to 140ºC.

4:30 PM

18.7 A 0.7V, 2.35% 3σ-Accuracy Bandgap Reference in 12nm CMOS

Y-W. Chen, TSMC, Hsinchu, Taiwan

In Paper 18.7, Taiwan Semiconductor Manufacturing Company describes a sub-1V bandgap voltage reference that uses a current-controlled feedback loop to improve current-mirror matching. It achieves an average temperature coefficient of 40.5ppm at 0.7V and a 3σ accuracy of 2.35% at 25ºC across 0.7V-to-0.9V supply.

4:45 PM

18.8 A 192pW Hybrid Bandgap-V_{th} Reference with Process Dependence Compensated by a Dimension-Induced Side-Effect

Y. Ji, Pohang University of Science and Technology, Pohang, Korea

In Paper 18.8, Pohang University of Science and Technology describes a hybrid bandgap-V_{th} reference using a process dependence compensation scheme. With a 0.69V reference voltage, it consumes 192pW and achieves a standard deviation of 0.52% over process corners, a temperature coefficient of 33ppm/°C and a line sensitivity of 0.02%/V.

18.1 A -105dBc THD+N (-114dBc HD2) at 2.8V_{PP} Swing and 120dB DR Audio Decoder with Sample-and-Hold Noise Filtering and Poly Resistor Linearization Schemes

Shon-Hang Wen, Kuan-Dar Chen, Chuan-Hung Hsiao, Ya-Chi Chen

MediaTek, Hsinchu, Taiwan

Three major design issues that arise for high-fidelity audio decoders are: 1) DAC reference noise limiting achievable SNR [1,2]; 2) THD+N degradation at large output swing [3,4]; and 3) Distortion arising from limited amplifier loop gain as a consequence of high output load capacitance (C_L) [5]. In the first issue, reference noise along with individual DAC cell noise generally limits SNR for a full-scale signal. The use of large device sizes [1], source degeneration [2] and chopping can mitigate 1/f noise, but none are effective for reducing thermal noise. Consequently, either more power or an external bypass capacitor for noise filtering is necessary for reducing DAC reference noise. In the second issue, THD+N of high-output-swing amplifiers degrades proportionally as the output swing increases above 1.6V_{PP}, even with a 4.5V supply [3,4]. The primary cause of this severe 2nd-order harmonic distortion (HD2) is due to the depletion effect of poly resistors [6]. Lastly, for adequate stability margin, the UGB and loop gain of the conventional nested Miller compensation (NMC) amplifier is restricted by an output limiting pole (ω_{limit}) and C_L. In [5], a frequency compensation scheme is proposed to push the UGB close to ω_{limit} and enhance the loop gain over the audio band (20Hz to 20kHz) while handling a C_L up to 10nF. However, with a C_L of 22nF, the amplifier begins to ring for a transient step. In this work, three solutions are presented to solve the aforementioned issues: 1) an area- and power-efficient sample-and-hold (S&H) noise filtering technique is introduced to shape the 1/f and thermal noise of the reference to frequencies below the audio band, thus greatly improving SNR for a full-scale signal; 2) a poly resistor linearization scheme is presented to improve HD2 by mitigating the depletion effect of resistors; and 3) a frequency compensation method for multistage amplifiers is introduced that boosts loop gain and thus enhances amplifier linearity without being limited by large C_L. Combining these techniques, the decoder and amplifier achieve -105dBc THD+N (-114dBc HD2) and 120dB DR, and can support a C_L up to 22nF.

Figure 18.1.1 shows the block diagram of the audio decoder with the proposed S&H noise filter. Although the tri-level current-mode DAC in [1] is used to significantly suppress the DAC reference noise for small output signals, there is no efficient way to suppress reference noise for full scale signals [1,2]. In simulation, reference noise accounts for 91% of total decoder output noise. An S&H filter with a long hold period (T_H) and a short turn-on sample period (T_S) can reduce not only the power consumption [7] but also all noise sources of the reference circuit in the audio band. For S&H circuits, the total noise power is kT/C and becomes concentrated at lower frequencies as T_H increases. For audio applications, a T_H which is longer than 100ms shifts high-frequency thermal noise below 10Hz. In simulation, the sampled-and-held noise is reduced to 0.9μVrms with T_H=100ms. During the off time, a switch with a high finite open resistance ($R_{SW,open}$) along with a proper hold capacitance (C_H) is required for noise filtering of the input and the thermal noise of $R_{SW,open}$. To reduce the noise leakage, the switch is designed with a small aspect ratio, and its bulk is connected to the S&H filter output via a unity-gain buffer as in [7]. By a proper selection of $R_{SW,open}$ and C_H, the leakage noise can be reduced to 1.3μVrms. To avoid noise coupling from DAC switching, cascode current-source cells are adopted. This technique can improve the total decoder output noise for full-scale DC output by 10dB to 3.25μVrms, which is close to the 3.17μVrms contributed by the DAC cells and amplifier. Thus, the DAC reference noise is almost eliminated within the audio band. The simulated SNR for a full-scale sinusoidal output can be enhanced to 112.7dB, compared to 102.1dB without the S&H noise filtering technique.

The linearization scheme to alleviate the depletion effect of poly resistors is illustrated in Fig. 18.1.2. The depletion effect causes the resistance to vary proportionally with the voltage across the resistor, resulting in 1st-order nonlinearity ($V_{CR1}\neq 0$) when one node of the resistor and its bulk are connected to global ground. This leads to serious HD2 for a single-ended output. In contrast to the feedback resistors, the distortion from the CMFB resistors cannot be rejected by the CMFB loop gain. In simulation, the HD2 degrades to -90dBc for a 2.8V_{PP} output swing. The proposed scheme utilizes two parallel resistors with inverse depletion effects by connecting their bulks to the opposite nodes to cancel out the depletion effect. The HD2 improvement depends on the mismatch between the two resistors in parallel ($\sigma_{MIS,AB}$) and between the CMFB resistors ($\sigma_{MIS,12}$). With the proposed scheme, the simulated HD2 can be improved by 25dB, down to -115dBc at a 2.8V_{PP} output swing with 0.5% mismatch.

In Fig. 18.1.3, the proposed frequency compensation scheme adds an additional damping circuit at the 1st-stage output (G_{m1}), which allows a higher loop gain over the audio band and improves the ability to handle large C_L. A high amplifier loop gain is required to suppress output distortion. However, the maximum loop gain of NMC amplifiers at an operating frequency (ω_{work}) is limited by $G_{m3}/(4C_L\omega_{work})$, where G_{m3} is the transconductance of the output stage. Increasing G_{m3} can improve the loop gain, but consumes more current. The proposed scheme introduces a 2-pole/1-zero frequency response for the 3-stage amplifier to boost the loop gain at audio frequencies but reduce the loop gain near the UGB. At higher frequencies outside the audio band, the first damping circuit at G_{m1} output realizes a large capacitance C_{eq1}, and the second damping circuit at the 2nd-stage output provides a small resistance $1/G_{d2}$ to decrease the equivalent Miller capacitance of C_{m1} to be smaller than C_{eq1} by lowering the 2nd-stage gain. These two circuits work together to conduct most of the high-frequency signal current to C_{eq1} instead of C_{m1}, effectively decreasing the gain around the UGB. By keeping the UGB below ω_{limit}, amplifier stability can be guaranteed. At audio frequencies, the maximum loop gain can be improved by a factor of $1+G_{d2}(1+C_{eq1}/C_{m1})/(G_{m2}G_{m3}R_z)$. To reduce the compensation capacitor (C_{d1}) size of the first damping circuit, a highpass filter is utilized to remove audio-band signals and increase the operation range for high-frequency signals. Around UGB, the input impedance of the damping circuit can be modeled as the equivalent damping resistance (R_{eq1}) in series with the C_{eq1}. To effectively decrease R_{eq1} but increase C_{eq1}, a power-efficient g_m-enhanced circuit ($M_{5,6}$) is adopted to boost the g_{m2}. The circuit achieves an equivalent C_{eq1} of 53pF using only a small physical C_{d1} of 3pF. In this work, the loop gain is improved by 20dB at 1kHz.

The audio decoder with a negative charge-pump that provides the negative voltage for the ground-referenced audio output is realized in a 0.153μm CMOS process. Unless otherwise stated, all tests are done with a single-ended load of 16Ω||1nF. The FFT spectra of the decoder, with the noise filter enabled and disabled, are shown in Fig. 18.1.4. The THD+N is significantly improved by 11dB when the noise filter is applied with T_H=160ms. If T_H is not long enough, e.g., T_H=1.25ms, the DAC reference noise is folded back into the audio band and degrades the SNR. Figure 18.1.5 (left) shows the measured HD2 vs. frequency plot with and without the proposed poly resistor linearization scheme. The proposed scheme greatly improves the HD2 to -114dBc from -91dBc. For 20 measured samples, the mean HD2 is improved by 20dB. The measured THD+N and DR at 2.8V_{PP} output are -105dBc and 120dB (A-weighted) respectively. The comparisons of amplifier transient responses under different C_L, with and without the damping circuit of C_{eq1} enabled, is shown in the Fig. 18.1.5 (right). With the proposed compensation method, no ringing in the rising or falling steps occurs until 47nF. The ability to handle a large C_L is substantially improved by 56 times, from 390pF to 22nF, compared with the damping circuit disabled. Figure 18.1.6 summarizes the measured results and Fig. 18.1.7 shows the die photo. Among high-output-swing (>2V_{PP}) amplifiers, this work achieves 21dB better linearity and 12dB higher SNR than prior state of the art, resulting in a 5.6× higher linearity FoM and 3.5× higher SNR FoM.

Acknowledgements:
We would like to thank Stacy Ho, Wen-Chieh Wang, and Yen-Hsun Hsu for helpful technical discussions, Li-Ting Ku for layout help, and Kuan-Hung Chen and Chien-Ming Chen for integration assistance.

References:

[1] K. Nguyen et al., "A 108 dB SNR, 1.1 mW Oversampling Audio DAC with A Three-level DEM Technique," *IEEE JSSC*, vol. 43, no.12, pp. 2592-2600, Dec. 2008.
[2] X. Jiang et al., "A Low-power, High-fidelity Stereo Audio Codec in 0.13 μm CMOS," *IEEE JSSC*, vol. 47, no. 5, pp. 1221-1231, May 2012.
[3] J. Chen et al., "A 62mW Stereo Class-G Headphone Driver with 108dB Dynamic Range and 600μA/Channel Quiescent Current," *ISSCC*, pp. 182-183, Feb. 2013.
[4] K. Abdelfattah et al., "A 40 nm Fully Integrated 82 mW Stereo Headphone Module for Mobile Applications," *IEEE JSSC*, vol. 49, no. 8, pp. 1702-1714, Aug. 2014.
[5] N. Mehta et al., "A 1mW -101dB THD+N Class-AB High-Fidelity Headphone Driver in 65nm CMOS," *IEEE Symp. VLSI Circuits*, pp. 235-236, June 2018.
[6] G. Gildenblat, *Compact Modeling: Principles, Techniques and Applications*. Springer, 2010, pp. 282-284.
[7] V. Ivanov et al., "An Ultra Low Power Bandgap Operational at Supply from 0.75 V," *IEEE JSSC*, vol. 47, no. 7, pp. 1515-1523, July 2012.
[8] C. Mohan and P. M. Furth, "A 16-Ω Audio Amplifier with 93.8-mW Peak Load Power and 1.43-mW Quiescent Power Consumption," *IEEE TCAS-II*, vol. 59, no. 3, pp. 133-137, Mar. 2012.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

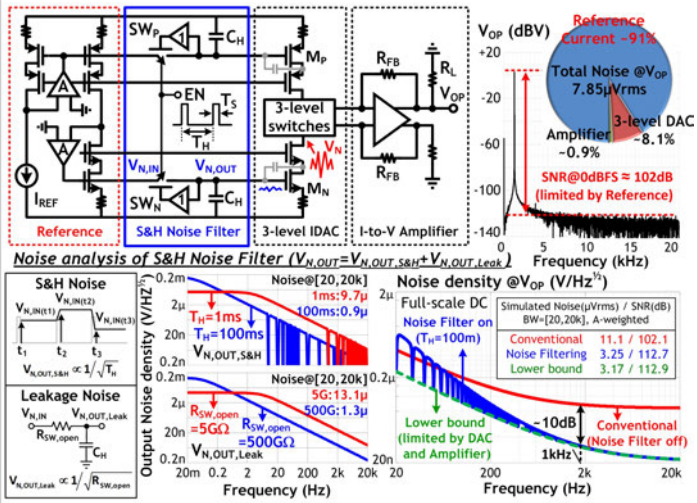


Figure 18.1.1: The audio decoder with the proposed S&H noise filter.

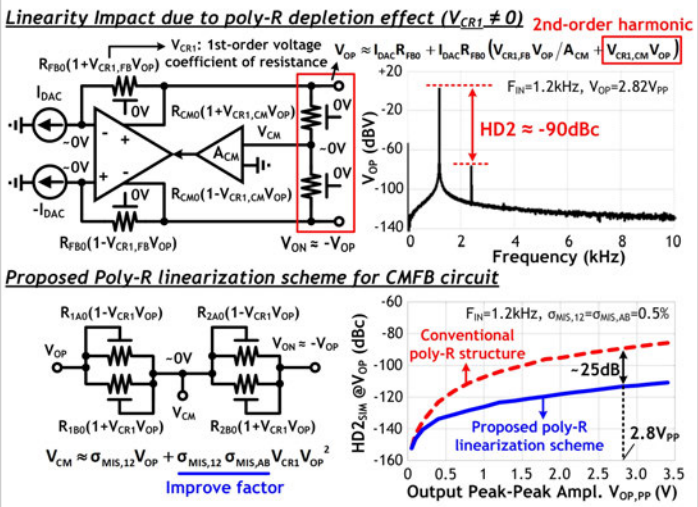


Figure 18.1.2: The poly resistor linearization scheme for headphone amplifiers.

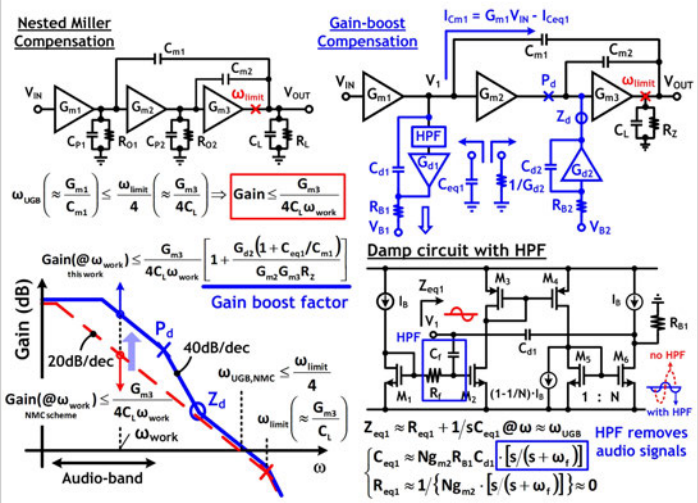


Figure 18.1.3: The gain-boost compensation scheme and its damping circuit.

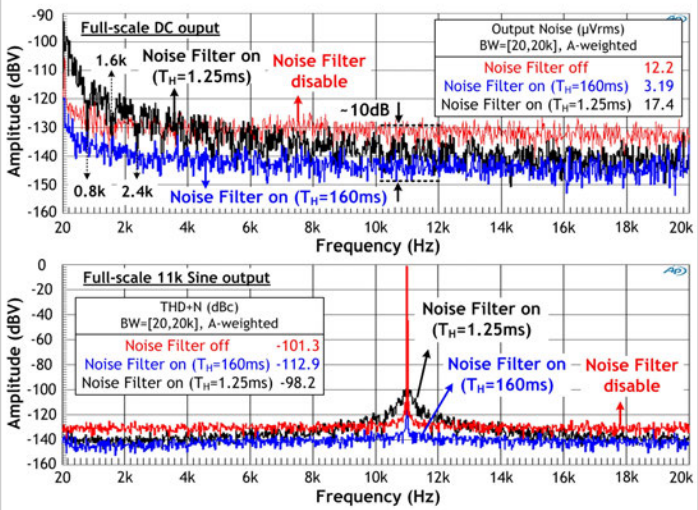


Figure 18.1.4: FFT comparisons that enable/disable the S&H noise filter.

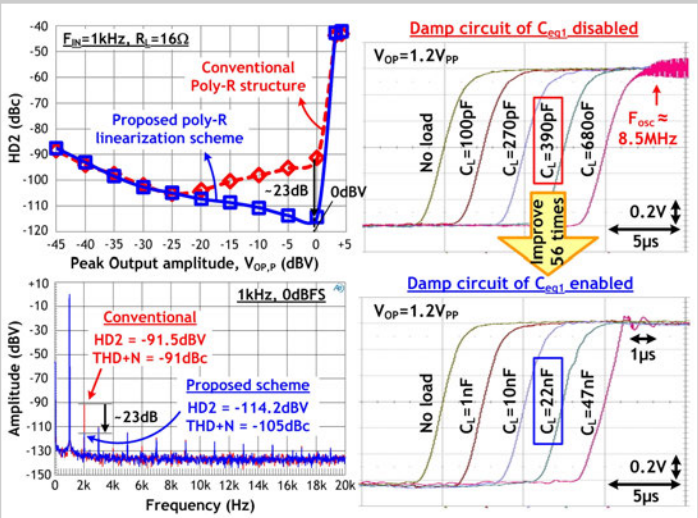


Figure 18.1.5: Measured HD2 vs. frequency plot, FFT spectrum (left), and transient responses of the amplifier for various C_L values (right).

Parameters	This work	[8] TCASII'12	[4] JSSC'14	[3] ISSCC'13	[2] JSSC'12
Process (nm)	153	500	40	180	130
Supply (V)	1.8	3.0	3.1-4.5	2.95-4.5	1.5
Load (R_L‖C_L)	16Ω ‖ 0–22nF	16Ω ‖ 10p–5nF	16Ω – 600Ω**	16Ω**	16/32Ω**
@ 1kHz, Max. output power without clipping					
Output voltage (Vpp)	2.82	3.46	3.20	2.53	2.5
Output power (mW)	62	93.8	80*	50*	49
THD+N (dB)	-105	-77.9	-82*	-84*	-72
@ 1kHz, lowest THD+N					
Output voltage (Vpp)	2.82	n/a	1.13	1.13	n/a
Load power (mW)	62	n/a	10	10	n/a
THD+N (dB)	-105	n/a	-86*	-90*	n/a
SNR (A-weighted) (dB)	120	79.9	100	108	98
Quiescent power (mW) (Driver only, /channel)	3.8***	1.43	3.78	2.7	1.5
FOM (THD+N) †	29,014	5,151	2,650	2,935	1,300
FOM (SNR) ‡	16,316	648	2,105	4,652	2,595

* Extracted from plot ** no information about load capacitance (n/a)
*** The quiescent current of the driver and the decoder (includes DACs, drivers and negative charge-pump) are 2.1mA@1.8V/channel and 3.1mA@1.8V/channel respectively
† FOM(THD+N)=Pout,max/[Quiescent power*THD+N(%)] @no clipping case
‡ FOM(SNR)=Pout,max/[Quiescent power*SNR*1000] @no clipping case

Figure 18.1.6: Performance Summary.

18

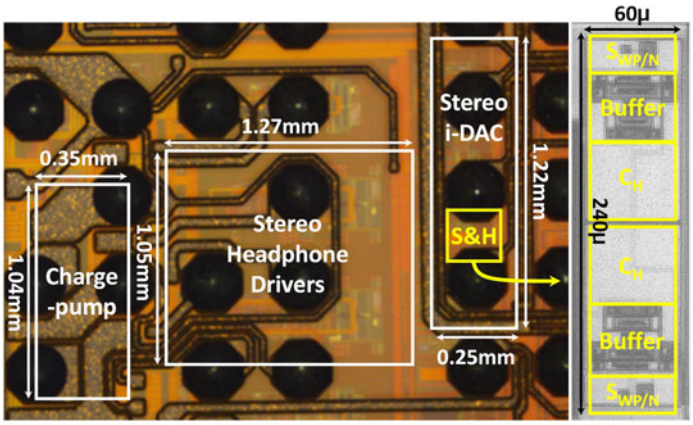


Figure 18.1.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

18.2 A 16fJ/Conversion-Step Time-Domain Two-Step Capacitance-to-Digital Converter

Xiyuan Tang[1], Shaolan Li[1], Linxiao Shen[1], Wenda Zhao[1], Xiangxing Yang[1], Randy Williams[1], Jiaxin Liu[2], Zhichao Tan[3], Neal Hall[1], Nan Sun[1]

[1]University of Texas, Austin, TX
[2]University of Electronic Science and Technology of China, Chengdu, China
[3]Analog Devices, Boston, MA

Capacitive sensors are widely used to measure various physical quantities, including pressure, humidity [1], and displacement [2]. Ultra-low-power capacitance-to-digital converters (CDCs) are required for sensors with limited battery capacity or powered by energy harvesters. A SAR CDC is simple to design and well-suited for low-to-medium resolution applications. However, to reach high resolution, it requires a low-noise comparator [3] or OTA-based active charge transfer [4], resulting in degraded power efficiency. The ΔΣ CDC [1] is suitable for high-resolution applications, but it requires OTAs and repeated charging of the sensing capacitor, leading to high power consumption. The Zoom CDC in [5] achieves high resolution with only one-time charging, but its energy efficiency is still limited by power-hungry OTAs. The open-loop SAR-VCO CDC in [6] achieves low power consumption by eliminating the OTA; however, the VCO gain variation causes inter-stage gain error and requires background calibration, which increases the design complexity and makes it unsuitable for single-shot measurement in sensor node applications due to the long convergence time.

This paper presents an incremental two-step CDC with a time-domain ΔΣM (TDΔΣM) that achieves a resolution of 0.29fF while dissipating only 0.083nJ per conversion, which improves the energy efficiency by >2× comparing to state-of-the-art CDCs. This is achieved by performing ΔΣ modulation in the time domain. Unlike the classic zoom CDC [5], this work replaces the OTA-based active-RC integrator by a VCO-based integrator, which brings several merits: 1) the VCO is mostly digital and scaling friendly; it works well under low supply voltage and consumes low power; 2) it provides infinite DC gain in the phase domain, and thus is well-suited for high-precision applications that demand high DC loop gain; 3) it has intrinsic spatial phase quantization, and thus, enables a simple 4b quantization using only minimum-size DFFs; it obviates the need for an array of low-offset comparators. With a 4b quantizer, 85dB SQNR can be achieved with a 1st-order loop and a low OSR of 15, which reduces the ΔΣM energy. Comparing to the open-loop VCO-based CDC [6], the proposed closed-loop TDΔΣM CDC obviates the need for background calibration. Its closed-loop gain is set by capacitor ratio, which is precisely matched by merging the ΔΣ feedback DAC with the SAR DAC. Therefore, the VCO gain variation cannot change the feedback factor, and thus, has negligible impact on CDC performance (±20% VCO gain variation results in only 2dB of SQNR change). By reusing the SAR comparator as the G_m stage of the VCO-based integrator, offsets in the SAR and the ΔΣM are inherently matched, which obviates the need for offset mismatch calibration.

Figure 18.2.1 shows the block and timing diagrams of the proposed CDC. During ϕ_0, a signal charge proportional to the difference between the sensor capacitor C_X and the offset capacitor C_{OS} is transferred to the merged DAC. It is first quantized by a coarse 8b SAR in ϕ_1. After that, a 4b TDΔΣM performs a fine quantization of the SAR conversion residue. A simple digital integrator is used as the decimation filter for the incremental ΔΣM. With notches at integer multiples of $f_s/15$, it provides suppression of periodic interferences. To reduce the offset and flicker noise, the CDC is chopped at the system level: the overall conversion is performed twice with swapped input polarities and the two conversion results are averaged.

Figure 18.2.2 shows the simplified schematic. The circuit consists of two halves that are excited differentially. Only a single-ended circuit is shown in the figure for illustration. The negative circuit includes another pair of C_X and C_{OS} that are connected through PADs to the replica of SAR and TDΔΣM DAC arrays on chip. During ϕ_0, the CDC is reset. When ϕ_0 ends, by switching the bottom-plate voltage of C_X and C_{OS}between V_{refp} and V_{refn}, a differential voltage V_X proportional to (C_X - C_{OS}) is created at the comparator input. The SAR performs an 8b conversion of V_X. After that, the comparator is reconfigured as a G_m stage and drives 2 current-controlled oscillators (CCOs) to perform the phase-domain integration. The output of the 14-level phase quantizer is fed back to the DAC array to realize the ΔΣ modulation. With the 1st-order loop and 25% clock cycle retiming delay, this ΔΣM does not require any excess loop delay compensation. The tradeoff between resolution and redundancy is considered when choosing the unit capacitor size of the ΔΣM DAC. In this work, a tri-level feedback DAC with C/2 unit capacitance is chosen to provide an inter-stage gain of 4 and ±1.25 SAR LSB of redundancy between the SAR and the ΔΣM, which helps absorb offset mismatch and SAR conversion error if any. The linearity of the SAR conversion relies on matching. As in any multi-bit ΔΣM, the DAC mismatch can cause nonlinearity. This issue is typically addressed by having an explicit DEM circuit to scramble the DAC element selection pattern; however, it incurs additional power/area cost. In this work, the dual-VCO-based integrator brings intrinsic CLA (ICLA) capability [7] as shown in Fig. 18.2.2. The DAC elements are naturally rotated at twice the VCO center frequency. Hence, the mismatch errors are up-modulated and inherently suppressed by the decimation filter. As a result, >85dB linearity is ensured without an explicit DEM.

Figure 18.2.3 shows the VCO schematic. During ϕ_0 and ϕ_1, the two 7-stage CCOs are disabled. When ϕ_2 arrives, the CCO rings are closed and start to oscillate with the same initial phase. The comparator input pair $M_{1p,n}$ and tail M_b are reused as a G_m stage that converts V_X into current to drive the CCOs. As a result, the offset mismatch between the SAR and the VCO-based ΔΣM is minimized. The CCO cell is made differential to improve power supply rejection. Only PMOS cross-coupling is used to reduce the capacitive load and increase the VCO tuning gain. A phase-frequency detector (PFD) is used to provide tri-level phase quantization. Comparing to an XOR gate that is widely used in VCO-based quantizers [6], the PFD doubles the quantizer resolution and quadruples the input phase range. The PFD output is retimed after 1/4 clock cycle delay. An encoder maps the PFD outputs (U,D) to a tri-level signal (*−1,0, +1*) that naturally drives a tri-level DAC (V_{refn}, V_{cm}, V_{refp}). Note that V_{cm}does not need to be precise as it does not affect differential DAC outputs, which are ±(V_{refp} - V_{refn}) and 0. The OSR is set to 15 to achieve 25dB boost in SQNR.

The prototype CDC was fabricated in 40nm CMOS with a 20fF SAR unit capacitor and 10fF TDΔΣM unit capacitor. It occupies an area of 0.06mm^2 (Fig. 18.2.7). The analog supply is 1.1V, while the digital supply is reduced to 0.6V to save power. The CDC consumes 0.083nJ per conversion, in which 0.044nJ is consumed by the reference that charges the capacitors, 0.023nJ is consumed by the VCO, and 0.016nJ is consumed by digital logic. Figure 18.2.4 shows the static and dynamic performances of the CDC in the test mode, where a fixed ceramic capacitor C_X of 5pF is connected to the CDC and a sine wave of 1kHz is applied to emulate the C_X change. The measured SNDR and SFDR are 75.8dB and 88.9dB, respectively. The corresponding CDC resolution is 0.29fF. The measured DNL and INL are within ±0.11 LSB and ±0.15 LSB, respectively.

Figure 18.2.5 shows the amplitude sweep. The measured DR is 79dB. Results with the actual pressure sensor [8] are also shown. During this test, two identical pressure sensors are connected to the differential sensing network while two fixed offset capacitors of 8pF are applied to cancel the DC capacitance of the sensors.

Figure 18.2.6 summarizes the performance of this work and compares it with prior publications. This work is OTA-free; it uses a VCO to realize the TDΔΣM and also obviates the need for background calibration by operating the VCO in closed-loop. Overall, it achieves a CDC FoM of 16fJ/conversion-step, which represents a >2× energy efficiency improvement over the state-of-the-art.

Acknowledgement:
The authors would like to thank the TSMC University Shuttle Program for chip fabrication, and Arindam Sanyal and Chen-Kai Hsu for technical discussions.

References:
[1] Z. Tan et al., "A 1.2-V 8.3-nJ CMOS Humidity Sensor for RFID Applications," *IEEE JSSC*, vol. 48, no. 10, pp. 2469–2477, Oct. 2013.
[2] S. Xia et al., "A Capacitance-to-Digital Converter for Displacement Sensing with 17b Resolution and 20us Conversion Time," *ISSCC*, pp. 198–200, Feb. 2012.
[3] H. Omran et al., "A 35fJ/Step Differential Successive Approximation Capacitive Sensor Readout Circuit with Quasi-Dynamic Operation," *IEEE Symp. VLSI Circuits*, pp. 1–2, 2016.
[4] H. Ha et al., "12.6 A 160nW 63.9fJ/Conversion-Step Capacitance-to-Digital Converter for Ultra-Low-Power Wireless Sensor Nodes," *ISSCC*, pp. 220–221, Feb. 2014.
[5] S. Oh et al., "15.4b Incremental Sigma-Delta Capacitance-to-Digital Converter with Zoom-In 9b Asynchronous SAR," *IEEE Symp. VLSI Circuits,* pp. 1–2, 2014.
[6] A. Sanyal and N. Sun, "An Energy-Efficient Hybrid SAR-VCO Delta Sigma Capacitance-to-Digital Converter in 40-nm CMOS," *IEEE JSSC*, vol. 52, no. 7, pp. 1966–1976, July 2017.
[7] S. Li and N. Sun, "A 174.3dB FoM VCO-Based CT ΔΣ Modulator with a Fully Digital Phase Extended Quantizer and Tri-Level Resistor DAC in 130nm CMOS," *ESSCIRC*, pp. 241–244, 2016.
[8] "Murata SCB10H datasheet," Murata Manufacturing Co., www.murata.com.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

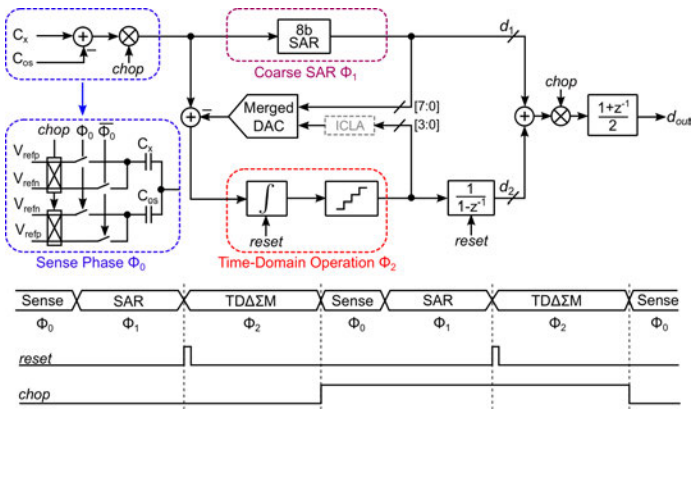


Figure 18.2.1: Architectural diagram of the proposed CDC, and the timing diagram.

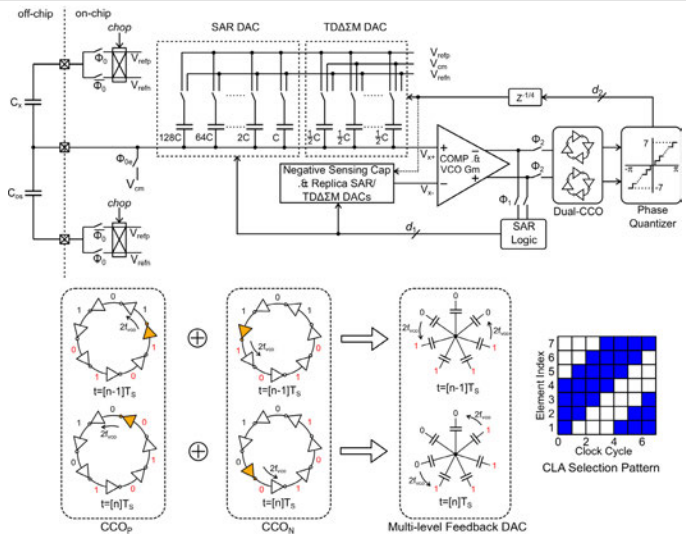


Figure 18.2.2: The simplified CDC schematic, and the illustration of Intrinsic CLA in the VCO-based ΔΣM.

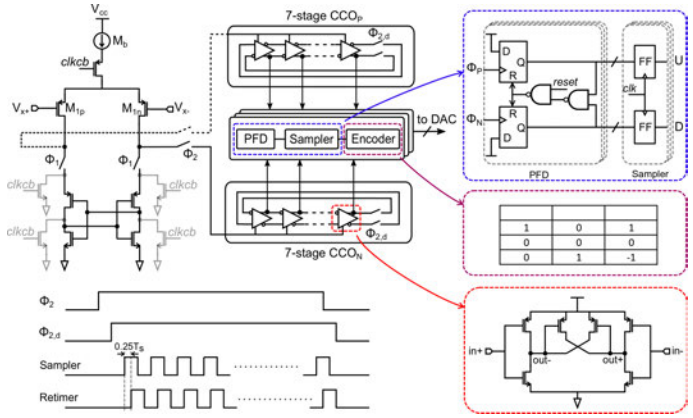


Figure 18.2.3: Circuit schematic of the proposed VCO design, reusing the comparator as a Gm stage, and the timing diagram of the TDΔΣM operation.

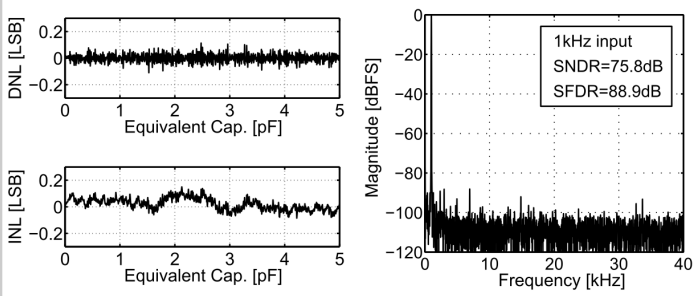


Figure 18.2.4: Measured DNL/INL plots, and the spectrum measured in test mode.

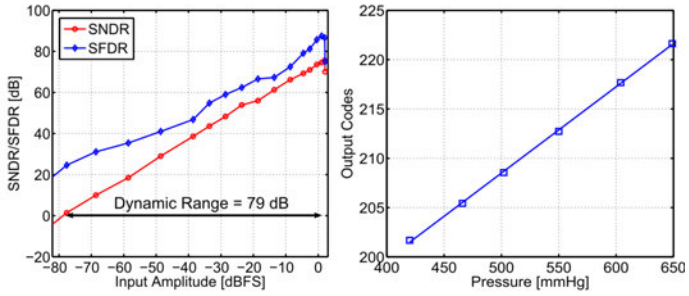


Figure 18.2.5: Measured dynamic range plot, and pressure sensor test with CDC.

	VLSI-16 [3]	ISSCC-14 [4]	VLSI-14 [5]	JSSC-17 [6]	**This work**
Process [nm]	180	180	180	40	**40**
Architecture	SAR	SAR	Zoom SAR+ ΔΣM	SAR+VCO	**Zoom SAR+ TDΔΣM**
Measured Sensor	Dual	Single	Single	Single	**Dual**
Need Back. Cal.	NO	NO	NO	YES	**NO**
OTA-Free	NO	YES	NO	YES	**YES**
Input Range [pF]	0-12.66	2.5-75.3	0-24	0-5	**0-5**
Meas. Time [μs]	16	4000	230	1	**12.5**
Resolution [fF]	1.1	6	0.16	1.1	**0.29**
Energy [nJ]	0.12	0.64	7.75	0.075	**0.083**
SNDR [dB]	70.6	81.8	94.7	64.2	**75.8**
FoM[1] [fJ/conv-step]	35	64	175	55	**16**

[1] $FoM = \frac{Energy}{2^{ENOB}}$

Figure 18.2.6: Performance summary and comparison with state-of-the-art CDCs.

18

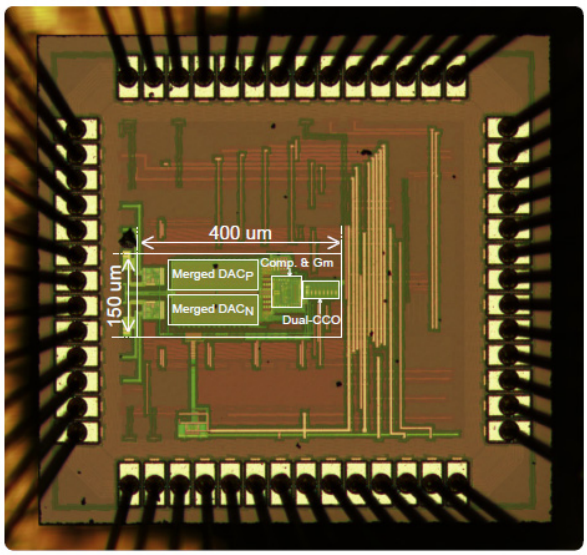


Figure 18.2.7: Die micrograph of the CDC.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

18.3 An Auto-Zero Stabilized Voltage Buffer with a Quiet Chopping Scheme and Constant Input Current

Thije Rooijers, Johan H. Huijsing, Kofi A. A. Makinwa

Delft University of Technology, Delft, The Netherlands

The readout of high-impedance sensors and sampled voltage references [1] requires amplifiers that can achieve both low offset and low input current. Recently, it has been shown that this unique combination can be achieved by an auto-zero (AZ) stabilized buffer [2]. However, its low-frequency noise density is $\sqrt{5}$ times higher than the buffer's own white-noise voltage spectral density e_n. Furthermore, its input current is not constant, but varies significantly with the input voltage. To overcome the first issue, a chopped AZ stabilization loop with an optimized duty-cycle is proposed to bring the low-frequency noise density close to $\sqrt{2}\cdot e_n$, the fundamental limit of an AZ stabilized amplifier. The second issue is solved by replacing the transmission-gate input switches used in [2] with NMOS switches and a constant Vgs drive. This keeps their charge injection constant over a wide input voltage range, and results in a constant input current.

As shown in Fig. 18.3.1, the AZ stabilized buffer consists of a buffer (BUF), whose offset (V_{os1}) and 1/*f* noise are periodically cancelled by a first AZ stabilization loop, which consists of an OTA (AZ1), an integrator (INT and $C_{int11\text{-}int12}$, 10pF each) and an OTA (AZ3) [2]. AZ1 is itself auto-zeroed by a second AZ loop that consists of another integrator (INT and $C_{int21\text{-}int22}$, 10pF each) and another OTA (AZ2). To prevent noise folding, the correction range of both AZ loops is quite small (~500μV). So a digitally-assisted AZ loop [2] is used to trim the mV offsets of AZ1 and BUF at startup (Fig. 18.3.2).

The minimum low-frequency noise density of the AZ stabilized buffer is $\sqrt{5}\cdot e_n$ [2], where e_n is the white-noise voltage spectral density of AZ1 and BUF (Fig. 18.3.1). This can be intuitively understood as follows. During phase ϕ_1 the first AZ loop cancels the low-frequency noise of BUF, while AZ1 contributes noise both directly and via the previously stored noise at nodes EF, leading to a total contribution of $\sqrt{2}\cdot e_n$. During phase ϕ_2 the noise of BUF, AZ1 *and* the noise of AZ1 held on nodes EF is stored on nodes GH, leading to a total contribution of $2\cdot e_n$. Assuming a 50% AZ duty cycle, the noise over the two phases would be $\sqrt{3}\cdot e_n$. However, the correlation of the sampled AZ1 noise at nodes EF in both phases leads to a low-frequency noise of $\sqrt{5}\cdot e_n$.

In order to remove this correlation, the AZ stabilization loop is chopped (Fig. 18.3.1). Choppers around AZ1 periodically invert the polarity of the noise held at nodes EF on its way to nodes GH. Over one chopping period, this should cancel its noise contribution, reducing the low-frequency noise floor to $\sqrt{3}\cdot e_n$. Since the second AZ loop is still active, however, the noise held at nodes EF will slowly change during chopping and so will not be perfectly cancelled. The amount of change depends on the duration of the chopping period, and hence on the AZ frequency f_{AZ}. Setting f_{AZ} = 15kHz, which is enough to cancel the 1/*f* noise of the buffer, only reduces the low-frequency noise floor slightly (Fig. 18.3.1). To approach the $\sqrt{3}\cdot e_n$ limit, simulations show that f_{AZ} would have to be some 3× higher, resulting in significantly more input switching, and hence, more input current.

To preserve the noise on nodes EF during a chopping period, extra integration capacitors $C_{int31\text{-}int32}$ (1pF each) are used to implement a modified ϕ_2 phase, denoted by ϕ_3 (Fig. 18.3.1). During ϕ_3, nodes EF are not connected to the output of INT. Also, resistors $R_{1\text{-}2}$ (2MΩ) limit the gain of INT, preventing it from clipping due to the residual output current of AZ1 (Fig. 18.3.1). The noise at nodes EF can then be perfectly chopped without changing f_{AZ}. The chopping transitions are arranged to occur in phase ϕ_2, when the input chopper is disconnected from the input by the AZ switches. With these two measures, the AZ stabilization loop can be chopped without introducing extra spikes or input current.

Another challenge is the offset of the integrator, which, via the chopper at its input, appears as a square-wave ripple voltage at the output of AZ1. This will be translated into a ripple voltage at the input of the buffer via the transconductance and finite output impedance of AZ1. With an expected integrator offset of about 2mV, the output impedance of AZ1 should be quite high (>150MΩ) to ensure that the resulting ripple is negligible (<100nV). This requirement is met by implementing AZ1 as a folded-cascode OTA.

The buffer noise level can be further reduced by adjusting the ϕ_1: ϕ_2 duty-cycle, noting that the noise in phase ϕ_1 ($\sqrt{2}\cdot e_n$) is less than the noise in phase ϕ_2 ($2\cdot e_n$). So the more time the buffer spends in phase ϕ_1, the closer its noise gets to $\sqrt{2}\cdot e_n$. Since some settling time is required for ϕ_2, the resulting noise level will then be somewhere between $\sqrt{3}\cdot e_n$ and $\sqrt{2}\cdot e_n$.

In [2], the input switches are implemented as transmission gates. Depending on the input voltage, this means that either the NMOS, PMOS or both devices will be turned "off" by clock transitions. Their charge injection, and thus, the input current will then be dependent on the input voltage (Fig. 18.3.3), typically changing by about 0.9pA over the input voltage range of the buffer (0 to 1.3V). In this design, the input switches are implemented as NMOS switches. Half-sized dummy transistors are used to partially cancel their charge injection. The NMOS switches are driven at constant Vgs by a circuit consisting of two latching transistors (MN1,2) and two boosting capacitors (CB1,2), which ensure that the clock waveform is superimposed on the input voltage. A simple buffer drives the capacitors, ensuring that their charging current does not load the input of the main buffer. Measurements show that the resulting input current is quite constant, varying by less than 0.2pA with input voltage (Fig. 18.3.3). As in [2], the input current is measured by observing the discharge of the hold capacitor C_H (= 36pF) of an on-chip low-leakage S&H circuit.

The buffer was realized in a 0.18μm CMOS process (Fig. 18.3.7). It draws 210μA from a 1.8V supply, and has an active area of 0.55mm², 0.12mm² of which is occupied by the S&H circuit. With a 1V input and f_{AZ} = 15kHz, measurements show that its input current is below 0.8pA (15 samples), and that its offset does not exceed 0.4μV (Fig. 18.3.4). In Fig. 18.3.5, the voltage noise density of the buffer is shown. With auto-zeroing alone, a low-frequency noise density of 31nV/√Hz is achieved, which equals the $\sqrt{5}$× noise limit. When chopped with the extra ϕ_3 phase enabled, the low-frequency noise drops to 25nV/√Hz, which equals the $\sqrt{3}$× noise limit. Finally when a 75% duty-cycle is used, a low-frequency noise of 20nV/√Hz is achieved, which is very close to the $\sqrt{2}$× noise limit. This reduction from $\sqrt{5}$× to $\sqrt{2}$× noise corresponds to a 2.5× power saving in the input stages. No tones at f_{AZ} or f_{CH} can be seen, demonstrating that the use of chopping does not add spikes, while the 75% duty-cycle does not change the offset or input current.

In Fig. 18.3.5, the input current vs input voltage of a typical sample is shown. With and without chopping, no significant difference in input current can be seen. The voltage drift across C_H is also shown (typical sample, 1V input). With AZ off, no drift can be seen, illustrating that the leakage of the S&H is indeed negligible. Furthermore, a measurement of the input current vs f_{AZ} shows a linear relationship, which indicates that charge injection is the dominant source of input current. In Fig. 18.3.6 the performance of the auto-zeroed voltage buffer is summarized and compared with the state-of-the-art. A special chopping scheme allows it to achieve a low-frequency voltage noise of 20nV/√Hz, while also achieving a constant and low input current (0.8pA), as well as state-of-the-art offset (0.4μV).

References:

[1] V. Ivanov et al., "An Ultra Low Power Bandgap Operational at Supply From 0.75 V," *IEEE JSSC*, vol. 47, no. 7, pp. 1515-1523, July 2012.
[2] T. Rooijers et al., "A Quiet Digitally Assisted Auto-Zero-Stabilized Voltage Buffer with 0.6pA Input Current and 0.6μV offset," *ISSCC*, pp. 50-52, Feb. 2018.
[3] Y. Kusuda "A 5.9nV/√Hz Chopper Operational Amplifier with 0.78μV Maximum Offset and 28.3nV/°C Offset Drift," *ISSCC*, pp. 242-244, Feb. 2011.
[4] A. T. K. Tang, "A 3μV-Offset Operational Amplifier with 20 nV/√Hz Input Noise PSD at DC Employing both Chopping and Autozeroing," *ISSCC*, pp. 386-387, Feb. 2001.
[5] M. A. P. Pertijs and W. J. Kindt, "A 140 dB-CMRR Current-Feedback Instrumentation Amplifier Employing Ping-Pong Auto-Zeroing and Chopping," *IEEE JSSC*, vol. 45, no. 10, pp. 2044-2056, Oct. 2010.
[6] Analog Devices Inc., "AD8551 data sheet", June 2015, <http://www.analog.com/media/en/technical-documentation/data-sheets/AD8551_8552_8554.pdf>
[7] S. Sakunia et al., "A Ping-Pong-Pang Current-Feedback Instrumentation Amplifier with 0.04% Gain Error," *IEEE Symp. VLSI Circuits*, pp. 60-61, June 2011.
[8] Q. Fan et al., "A 21 nV/√Hz Chopper-Stabilized Multi-Path Current-Feedback Instrumentation Amplifier With 2μV Offset," *IEEE JSSC*, vol. 47, no. 2, pp. 464-475, Feb. 2012.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

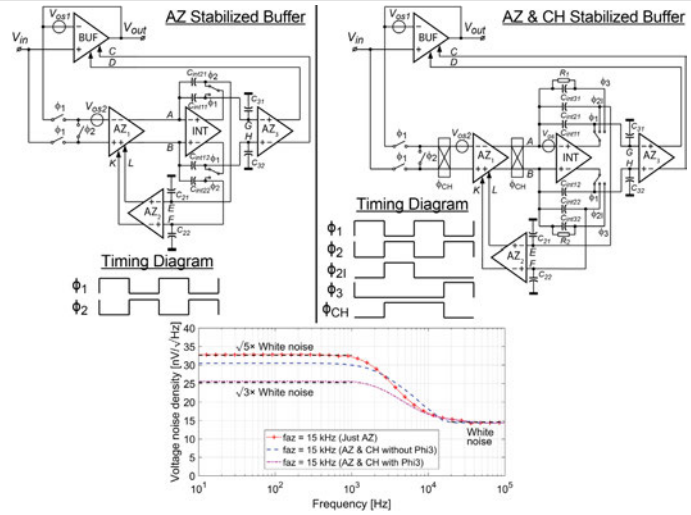


Figure 18.3.1: An auto-zero stabilized buffer without and with a chopped stabilization loop and additional Φ3 phase. Voltage noise density simulation for different configurations (AZ, AZ&CH without and with Φ3).

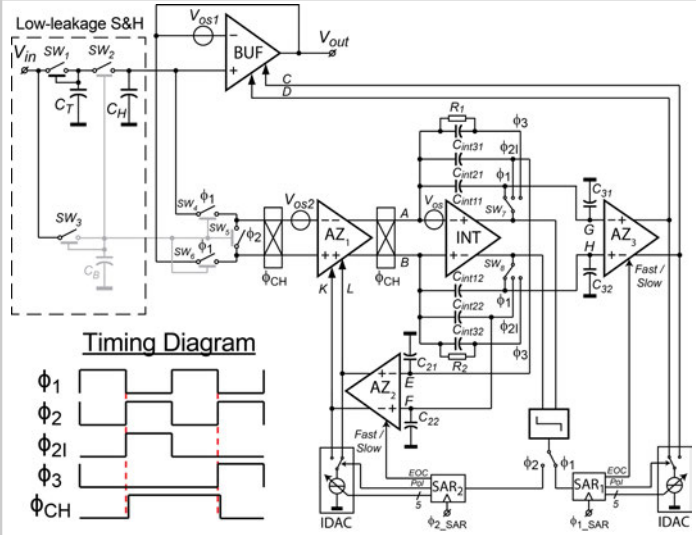


Figure 18.3.2: Block diagram of the auto-zero stabilized voltage buffer with chopped stabilization loop and digital assistance.

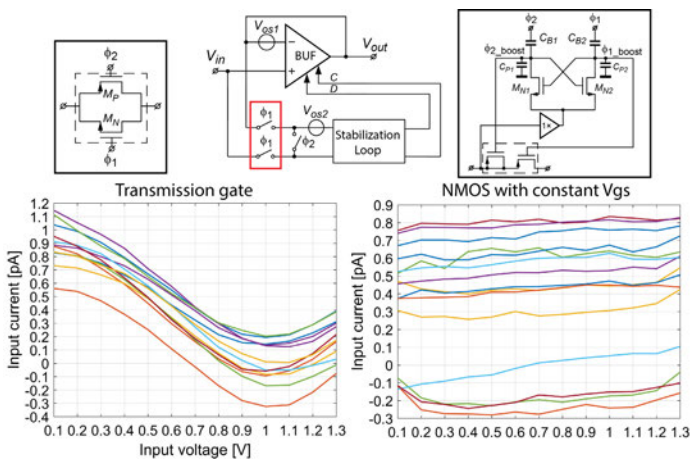


Figure 18.3.3: Measured input current vs input voltage for different type of input switches: transmission gates (Left) and NMOS with constant Vgs drive (Right).

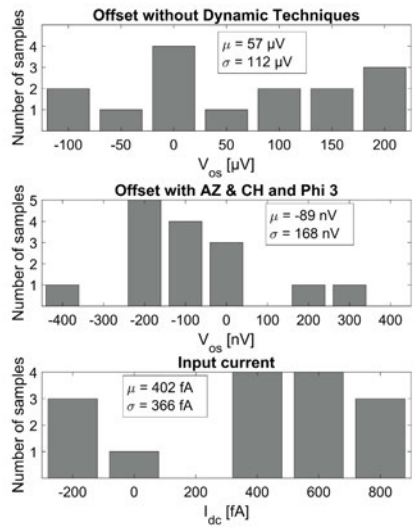


Figure 18.3.4: Histograms (15 samples) of the measured offset (without and with AZ) and input current with AZ&CH and Φ3 (fAZ = 15 kHz, Vin = 1V).

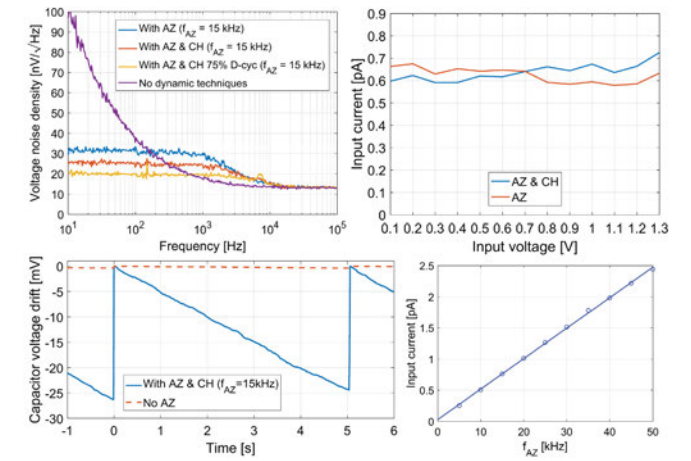


Figure 18.3.5: Measured voltage noise PSD: No AZ, with AZ and with AZ & CH (Top left). The input current vs the input voltage for a typical sample for AZ and AZ&CH (Top right). The capacitor voltage drift of a typical sample with and without AZ & CH (Bottom Left). The input current vs AZ frequency with chopped stabilization loop (Bottom Right).

	This work	[2]	[3]	[4]	[5]	[6]	[7]
Dynamic technique(s)	**Auto-zeroing and Chopped Stabilization**	Auto-zeroing	Chopping	Chopping and Auto-zeroing	Chopping and Auto-zeroing	Auto-zeroing	Chopping and Auto-zeroing
Input current (Max)	**0.8 pA**	1.2 pA*	72 pA	40 pA	-	50 pA	-
Offset (Max)	**0.4 μV**	0.6 μV	0.78 μV	3 μV	2.8 μV	5 μV	4 μV
Voltage noise (nV/√Hz)	**31 (AZ) 25 (AZ&CH) 20 (75% AZ&CH)**	29	5.9	20	38 (AZ) 27 (CH&AZ)	75	140 (AZ) 28 (CH&AZ)
NEF	**7.3**	7.4	8.7**	21.8**	43.5**	-	-
GBW (MHz)	**1.45**	1.45	4	2.5	0.8	1	-
PSRR (dB)	**125**	125	142	-	138	130	128
Frequency (kHz)	**15 (f_{AZ}) / 7.5 (f_{CH})**	15	200	15 (f_{CH}) / 7.5 (f_{AZ})	28 (f_{AZ}) / 14 (f_{CH})	4	11 (f_{CH}) / 7.33 (f_{AZ})
Supply current	**210 μA**	210 μA	1.47 mA	800 μA	1.7 mA	750 μA	480 μA
Supply voltage	**1.8 V**	1.8V	2.5 - 5.5 V	5 V	2.7 - 5.5 V	2.7 V	3.3 - 5.5 V
Technology	**0.18 μm**	0.18 μm	0.35 μm	0.6 μm	0.5 μm	-	0.5 μm
Die area (including pads)	**1.4 mm²**	1.4 mm²	1.26 mm²	0.67 mm²	2.5 mm²	-	1.48 mm²

* Input current over the full input voltage range (0 to 1.3V) ** Estimated value [8]

Figure 18.3.6: Performance summary and comparison with previous works.

18

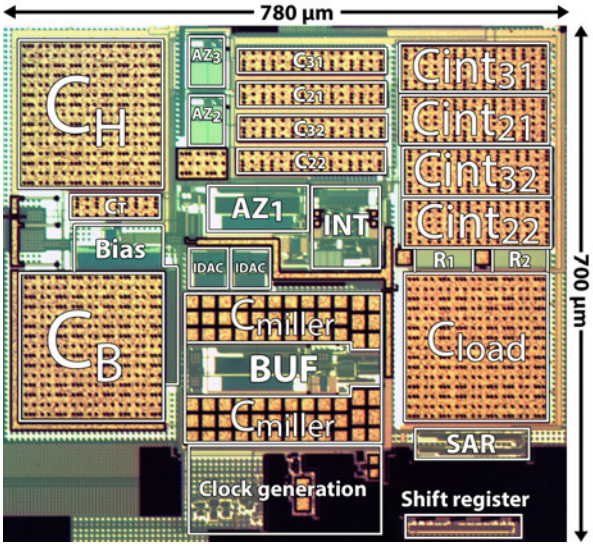


Figure 18.3.7: Die micrograph of the active area.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

18.4 A 0.55nW/0.5V 32kHz Crystal Oscillator Based on a DC-Only Sustaining Amplifier for IoT

Hani Esmaeelzadeh, Sudhakar Pamarti

University of California, Los Angeles, CA

An always-ON, stable, 32kHz crystal oscillator (XO) provides key time-keeping, synchronization, and sleep-timer functions in most electronic systems. Ultra-low power (ULP) consumption of the XO is critical in highly duty-cycled, energy-constrained systems such as Internet-of-things (IoT). Conventional XOs, typically implemented in the Pierce configuration, consume 10 to 100nW: a large enough transconductance, $g_m > \omega_0^2 \cdot C_L^2 \cdot R_m$, is needed to compensate for loss in the crystal's motional resistance, R_m, and sustain oscillations at frequency ω_0, across load capacitors (C_L > 15pF in Fig. 18.4.1) that are inevitably large to ensure frequency stability over process, voltage, and temperature (PVT) variations. Recent 32kHz XO designs report sub-10nW power consumption [1-4] by either duty-cycling the sustaining amplifier [4] or by providing energy to compensate the crystal losses only via narrow pulses at the peaks of the XO waveform [1-3]. However, they require high-power components [2], complicated designs with multiple power domains [1-2] to properly time the pulses, or significant calibration and off-chip components [4] to reduce PVT sensitivity.

This paper presents a new ULP XO architecture that achieves the lowest reported power consumption of 0.55nW from a 0.5V supply, among state-of-the-art publications as shown later in Fig. 18.4.6. As illustrated in Fig. 18.4.1, the proposed architecture, instead of amplifying the 32kHz signal from the crystal, downconverts the crystal's motional current to DC, amplifies it at DC, then upconverts the amplified signal back to 32kHz. By processing the signal at DC, the technique eliminates the aforementioned minimum g_m requirement of the Pierce XO, and only requires the DC amplifier's feedback resistor to be greater than the motional resistance of the crystal, i.e., $R_f > R_m$. As shown later, the DC amplifier can consume as low as 50pW. Note that the frequency conversions (to DC and back) can result in a non-zero phase shift across the resonator (due to the delay in the mixers' gate-drivers, etc.), that varies over PVT, potentially degrading XO frequency stability. So, the proposed technique incorporates I/Q paths in the frequency conversions as shown in Fig. 18.4.1 to preserve the phase information and ensure nominally zero phase shift across the resonator. Consequently, the XO oscillates at its series-resonance frequency, which improves its frequency stability even without large load capacitors.

A ULP delay-locked loop (DLL) derives the binary I/Q signals from the oscillation signal and is described later in Fig. 18.4.3; any higher-order harmonic content resulting from mixing is greatly suppressed by the low-bandwidth DC amplifier and the high-Q crystal. Since the I/Q signals are unavailable at start-up, a dedicated start-up circuit is required. While a Pierce-based sustaining amplifier can be used for XO start-up [1-3], a ring-oscillator-based pre-energization technique [5] was employed to make the I/Q signals immediately available. During start-up, the DLL and the biasing circuits settle to their steady state to guarantee proper start-up and normal operation.

The block diagram of the proposed ULP XO is shown in Fig. 18.4.2. The key difference (from the conceptual block diagram in Fig. 18.4.1) is in how the amplified 32kHz crystal drive, $XO_2 = a \cdot I + b \cdot Q$ in Fig. 18.4.1, is generated. Instead of using a power-hungry 32kHz adder, XO_2 is generated by switching between four DC voltage levels, $V_1, \ldots, V_4$, corresponding to the values (±a±b). The switching between $V_1, \ldots, V_4$ is done using non-overlapping 25%-duty-cycle signals, D_1 through D_4, as shown in the timing diagram. Ignoring the TIA common-mode voltage for simplicity, the voltage levels (a+b) and (a–b) are generated by downconverting XO_1 using (I+Q) and (I–Q) instead of I and Q, respectively. This is achieved using additional passive current-commutating mixers feeding directly into the DC TIAs as shown in Fig. 18.4.2. The other voltage levels (–a–b) and (–a+b) are available simply from using differential TIAs. All four passive mixers are implemented using transmission gates. Small transistors suffice since their on-resistance is smaller than the high R_m of 32kHz crystals (>30kΩ). Oscillation amplitude control is achieved by comparing $\max(V_1,\ldots, V_4)$ with a reference voltage and adjusting the TIA feedback resistance accordingly, as shown. Computation of $\max(V_1,\ldots, V_4)$ is achieved using low switching activity voltage comparators and consumes negligible power; details are omitted here for brevity.

The TIAs use a low-voltage, ULP two-stage opamp shown in Fig. 18.4.2. To minimize leakage, thick-oxide devices are used. To reduce the threshold voltage of the input PMOS pair, their bulks are connected to their corresponding gates. Note that the source/drain junction diodes remain reverse-biased at all times as the devices are always operating in subthreshold region. Since the TIAs do not require a high bandwidth, large input devices are used to minimize the voltage offset and mismatch between the TIAs. Pseudo-resistors are used to define the output common-mode levels and provide high gain. Each opamp consumes only 50pW and is biased by a PTAT current source to ensure a constant transconductance over temperature. The ultra-low PTAT bias current is generated by a switched-capacitor technique shown in Fig. 18.4.2. Clocked by the accurate 32kHz XO signal, this circuit generates currents as low as 10pA without requiring any off-chip components.

Figure 18.4.3 shows the DLL implementation. The XO output is first converted to a differential signal, and then fed to a two-stage current-controlled delay line. The delay line is controlled by a charge pump such that it generates quadrature outputs. The simulated transient response of the DLL, plotted in Fig. 18.4.3, shows that the DLL settles to its steady-state in 13ms, well within the XO start-up time of 0.2s. To reduce the power consumption of the output buffer, the rail-to-rail swing at either I/Q signals can be taken as the XO final output.

A prototype IC was fabricated in TSMC 65nm CMOS technology with an active area of 0.027mm^2 (Fig. 18.4.7). Figure 18.4.4 shows the measured power consumption and frequency variation of the proposed ULP XO under the supply voltage variations from 0.4V to 0.9V (at 25°C), and temperature variations from -20°C to 80°C (at VDD=0.5V). The measurements were performed across 20 chips. As seen, the proposed XO consumes 0.55nW from a 0.5V supply at 25°C (averaged across 20 chips with a worst-case power of 0.7nW). Measured frequency variation is within 80ppm across temperature, with an average line sensitivity of ~13ppm/V. The top part of Fig. 18.4.5 plots the measured Allan deviation. The measured Allan floor shows that the long-term frequency stability is better than 14ppb. To verify the robustness of the proposed XO to mismatches between the I/Q paths, intentional errors were introduced between I and Q phases using an external clock generator. The bottom part of Fig. 18.4.5 shows the deviation of the oscillation frequency as a result of the intentional I/Q phase mismatch (from 90°). The frequency deviation is minimal (±0.6ppm within ±5°), as the high-Q resonator only has to be slightly pulled off resonance to compensate for the phase mismatch.

Figure 18.4.6 shows the power contribution of the main blocks; summarizes the performance of the proposed XO; and compares it with other state-of-the-art 32kHz XOs. Unlike the Pierce XO where the amplifier consumes >85% of the power, the DC opamps in the proposed XO only contribute ~14%. Most of the power consumption is in the DLL, Mixer drivers, and the 25%-duty-cycle clock generator that are mostly digital and scale well with CMOS technology. As can be seen, the proposed XO only consumes 0.55nW, which is ~2.7× lower than the best case in prior art shown in Fig. 18.4.6. The proposed architecture does not need calibration and multiple supplies, unlike most prior art [1,2,4].

Acknowledgement:
This work was supported by the National Science Foundation under NSF grant CNS-1329755.

References:
[1] Y. Zeng et al., "A 1.7nW PLL-Assisted Current Injected 32kHz Crystal Oscillator for IoT," *IEEE Symp. VLSI Circuits*, pp. C68-C69, 2017.
[2] D. Yoon et al., "A 5.58nW Crystal Oscillator Using Pulsed Driver for Real-Time Clocks," *IEEE JSSC*, vol. 51, no. 2, pp. 509-522, Feb. 2016.
[3] K. Hsiao "A 1.89nW/0.15V Self-Charged XO for Real-Time Clock Generation," *ISSCC*, pp. 298-299, 2014.
[4] A. Shrivastava et al., "A 1.5nW, 32.768KHz XTAL Oscillator Operational From a 0.3V Supply," *IEEE JSSC*, vol. 51, no. 3, pp. 686-696, March 2016.
[5] H. Esmaeelzadeh and S. Pamarti, "A Quick Startup Technique for High-Q Oscillators Using Precisely Timed Energy Injection," *IEEE JSSC*, vol. 53, no. 3, pp. 692-702, March 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

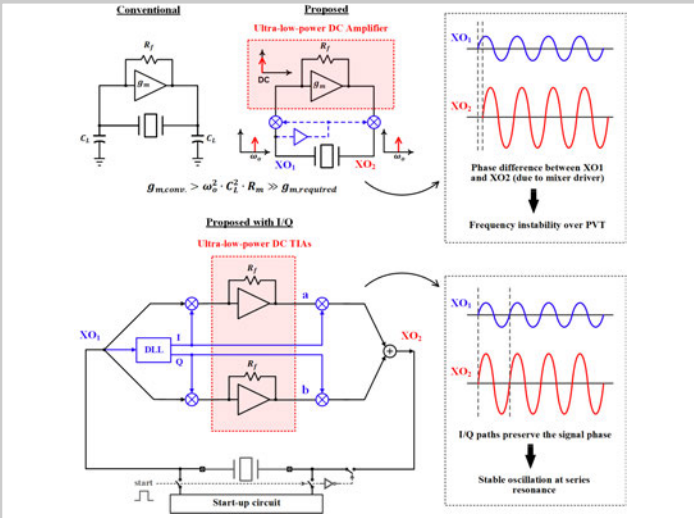


Figure 18.4.1: Conventional Pierce XO (top-left); basic concept of the proposed ULP XO (top-right); employing I/Q paths to maintain zero-phase shift across the resonator-series resonance (bottom).

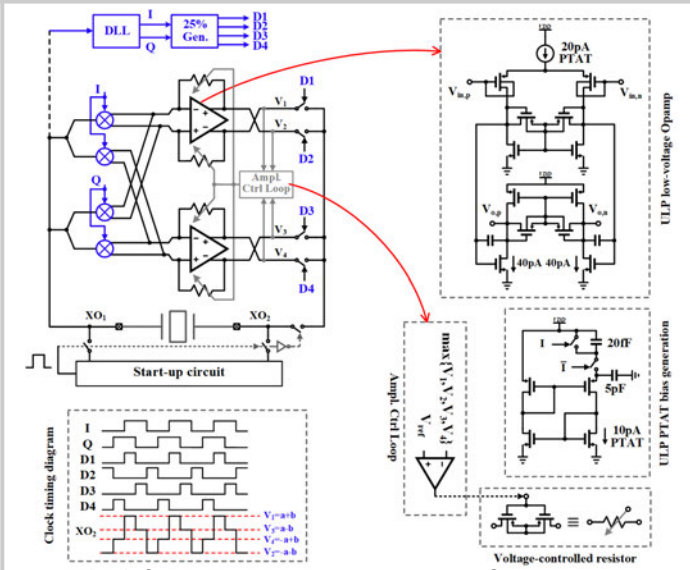


Figure 18.4.2: Simplified implementation of the ULP XO.

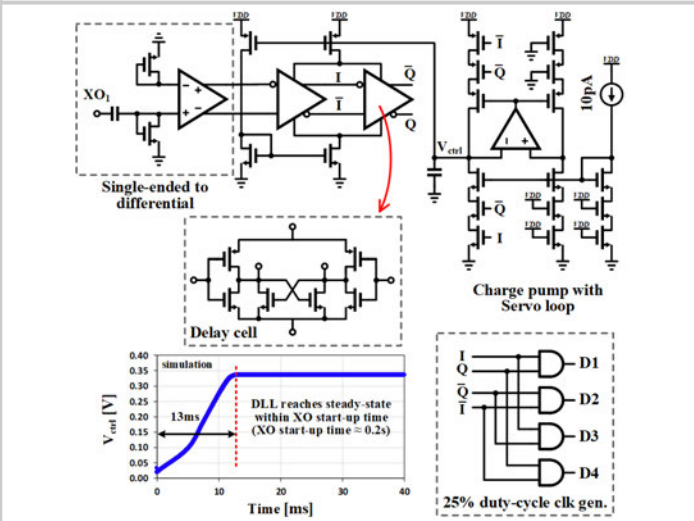


Figure 18.4.3: Implementation of the delay-locked loop to generate the I/Q signals from the XO oscillation signal (top), and 25%-duty-cycle clock generation (bottom-right).

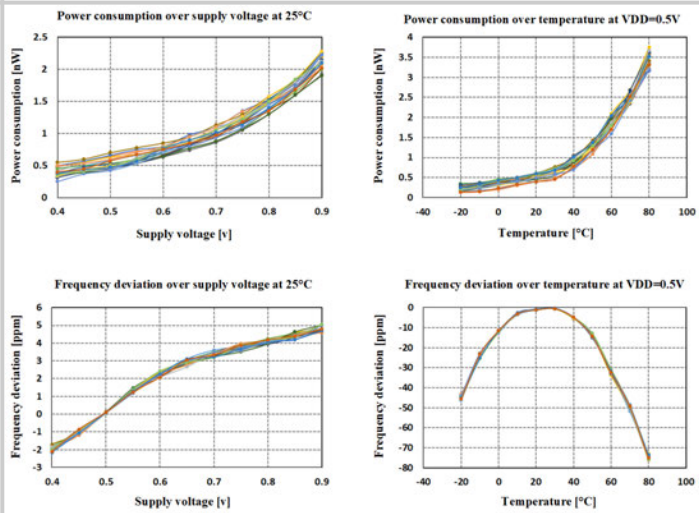


Figure 18.4.4: Measurement results for 20 dies: power consumption and frequency variations over VDD (left); power consumption and frequency variations over temperature (right).

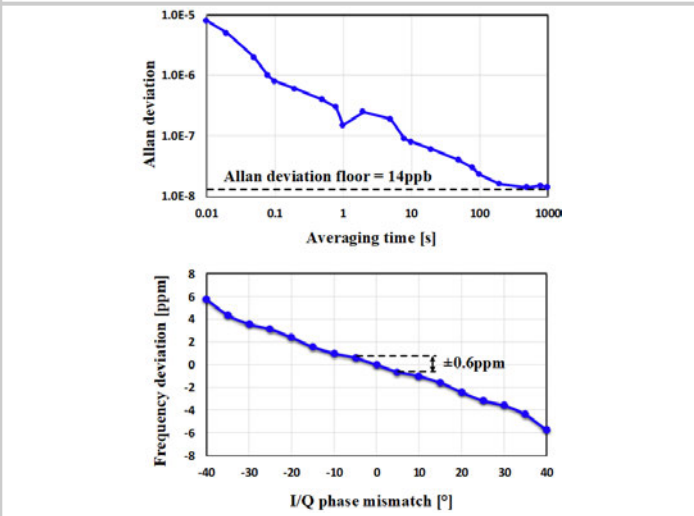
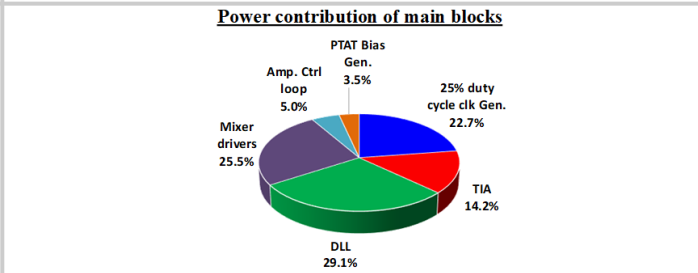


Figure 18.4.5: Measured Allan deviation for one chip (top); measured sensitivity of oscillation frequency to intentionally-introduced mismatch in I/Q phase using external signal generators (bottom).



	This work*	VLSI'17 [1]	JSSC'16 [2]	ISSCC'14 [3]	JSSC'16 [4]
Technology (nm)	65	55	180	28	180
Area (mm²)	0.027	0.16	0.3	0.03	0.062
Supply voltage (V)	0.5	0.5	1.2	0.2	0.3
Power at 25°C (nW)	0.55	1.7	5.58	5	1.5
Temperature stability (ppm)	80 (-20°C to 80°C)	109 (-20°C to 80°C)	133 (-20°C to 80°C)	50 (-20°C to 80°C)	150 (0°C to 80°C)
Line sensitivity (ppm/V)	13	6.7	30.3	85	7
Allan Deviation (ppb)	14	25	10	10	70
Calibration required	NO	YES	YES	NO	YES

*Averaged across 20 dies with worst-case power of 0.7nW.

Figure 18.4.6: Power contribution of main blocks (top); performance summary and comparison with state-of-the-art 32kHz XOs (bottom).

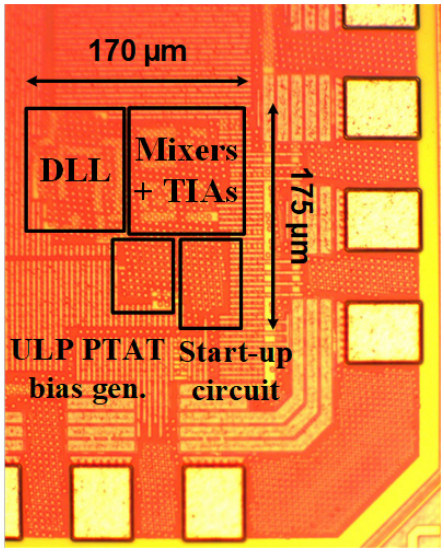


Figure 18.4.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

18.5 A 54MHz Crystal Oscillator with 30× Start-Up Time Reduction Using 2-Step Injection in 65nm CMOS

Karim M. Megawer[1], Nilanjan Pal[1], Ahmed Elkholy[1], Mostafa G. Ahmed[1], Amr Khashaba[1], Danielle Griffith[2], Pavan Kumar Hanumolu[1]

[1]University of Illinois, Urbana-Champaign, IL
[2]Texas Instruments, Dallas, TX

The start-up time of crystal oscillators (T_{START}) is a major bottleneck in reducing the average power of heavily duty-cycled wireless/wireline communication systems [1]. Among all the reported schemes to reduce T_{START}, techniques that increase initial noise amplitude by injecting a surge of energy into the crystal resonator are shown to be most effective [1-3]. These approaches are proven to be robust if the frequency of the injection signal is equal to the crystal oscillator (XO) frequency ($F_{INJ} = F_{XO}$), which is difficult to achieve across PVT with on-chip oscillators. Any mismatch ($\Delta F = F_{INJ} - F_{XO}$) even as small as a few 100 ppm can greatly increase T_{START}. Sweeping the injection frequency using a chirp oscillator [2] or dithering the injection frequency between two values [1] can partially alleviate this issue but because $\Delta F \neq 0$, this only reduces T_{START} to about 14× the theoretical minimum in [2]. On the other hand, it was shown in [3] that the use of a precise injection period $T_{INJ,OPT}$ can help reduce T_{START} even in the presence of large ΔF. $T_{INJ,OPT}$ must be chosen such that current in the motional branch of the resonator, $i_m(t)$, reaches its steady-state value, $I_{m,SS}$ ($i_m(T_{INJ,OPT}) = I_{m,SS}$) as shown in Fig. 18.5.1 [3]. However, small T_{START} and large tolerance to ΔF can be achieved only when $I_{m,SS}$ is very small, which translates to small XO output amplitude (V_{XO} < 200mV) and degraded phase noise. For example, as illustrated in Fig. 18.5.1, $T_{START} \approx T_{INJ,OPT}$ because $i_m(T_{INJ,OPT}) = I_{m,SS1}$ even when ΔF is as large as 1000ppm. However, for $I_{m,SS2} > I_{m,SS1}$, no T_{INJ} can ensure $i_m(T_{INJ}) = I_{m,SS2}$ if ΔF > 1000ppm, thus greatly increasing T_{START}. Therefore, ΔF must be small (<500ppm) to achieve a large V_{XO} even with precisely-timed injection, a condition that is difficult to meet in practice even with the best-reported temperature-compensated on-chip oscillators [4]. In view of these drawbacks, we present a robust 2-step injection technique that can tolerate large ΔF and achieves close to theoretical minimum T_{START}, large output swing and excellent phase noise. Fabricated in a 65nm CMOS process, the prototype XO achieves T_{START} of less than 20µs across the -40°C to 85°C temperature range, which is within 1.5× of the theoretical minimum and represents an over-30× reduction in T_{START} compared to that of a normal/un-injected XO.

The proposed T_{START} reduction technique uses a 3-step XO start-up process illustrated in Fig. 18.5.1. In the first step, a clock generated by a free-running ring oscillator (RO) is injected into the XO for a duration of $T_{INJ1} = 2L_m I_{m1}/V_{INJ} = 1/\lambda\Delta\omega_r$, where L_m is the motional inductance, ω_r is the resonance frequency and I_{m1} is the motional current (for V_{XO} = 200mV) of the crystal resonator, respectively, V_{INJ} is the injection amplitude ($V_{INJ} = \lambda V_{XO}$, λ is a constant) and $\Delta\omega_r = \omega_{XO} - \omega_r$ [3]. Assuming the RO frequency is within ±5000ppm of the XO resonance frequency, both the XO amplitude and frequency rapidly approach 200mV and the desired F_{XO} (to within 20ppm), respectively [3]. In the second step, using low-swing XO output as the reference clock, a fast-locking digital phase-locked loop (PLL) locks RO to the XO frequency in less than 6µs. In the final step, the PLL feedback loop is opened and free-running RO is injected into the XO for a duration of $T_{INJ2} \approx T_{INJ,OPT} - T_{INJ1}$, where $T_{INJ,OPT} = -2L_m/R_m \ln(1 - 0.35\omega_{XO}R_m(2C_L + 2C_o)$, R_m is the resonator's motional resistor, C_L is the resonator's load capacitance, and C_o is the resonator's parasitic capacitor, which indicates that $T_{INJ,OPT}$ is only a function of the crystal resonator's parameters and constant across temperature [2]. Thanks to the PLL, F_{INJ} will be within 20ppm of the desired F_{XO}, and the resonator current and therefore output amplitude reach the desired steady-state values immediately after releasing the injection (see Fig. 18.5.1). In our implementation, $T_{INJ} = T_{INJ1}+T_{INJ2}$ = 12µs, which is determined to be optimal from XO parameters and verified experimentally (see Fig. 18.5.4).

Figure 18.5.2 shows the complete architecture of the proposed XO. It consists of an inverter-based high-swing single-ended Pierce XO, switches/buffers for injection, a fast-locking digital PLL, and a finite-state machine that controls the start-up sequence described earlier. The PLL is composed of a phase-frequency detector (PFD), a 4-bit time-to-digital converter (TDC) that measures the input phase difference (Φ_E) with a resolution of 15ps, a proportional-integral digital loop filter (DLF), digitally controlled RO (DCRO), and auxiliary circuits used to speed up the phase-locking process. When the PLL is enabled, the RO starts oscillating at a frequency that is within ±5000ppm of F_{XO} in the worst case. However, this doesn't bound the initial input phase error (Φ_{E0}), which can be as large as one XO period (~20ns). Such a large Φ_{E0} can increase PLL lock time (T_{LOCK}) to more than 100µs, severely restricting the extent of T_{START} improvement. Because Φ_{E0} is typically much greater than phase error caused by the initial frequency error, we surmise that T_{LOCK} can be greatly reduced by presetting Φ_{E0} also to zero. To this end, the input reference clock (REF) and the feedback (FB) signal are synchronized so that they arrive at the same time to the input of the PFD when the PLL is enabled. As shown in Fig. 18.5.2, the enable signal, PLL_{EN}, is synchronized with the rising edge of REF such that the rising edges of REF_{SYNC} and PLL_{EN_SYNC} are aligned. Resetting the RO output phase by replacing the RO edge with that of PLL_{EN_SYNC} aligns REF_{SYNC} and FB signals, thus making $\Phi_{E0} \approx 0$. Delays associated with the oscillator phase reset and clock buffers that may cause the FB rising edge to arrive later than REF_{SYNC} are compensated by a digitally controlled delay line (DCDL) placed in the reference path (see Fig. 18.5.2). After the PLL is locked, the last 128 samples of the integral path accumulator are averaged and used to preset the RO frequency in the 3rd step of the XO start-up process.

Figure 18.5.3 shows the schematic of DCRO. It is implemented using three delay stages and a multiplexer for resetting the phase. Each delay stage consists of an inverter and an RC section and is designed such that the delay due to the RC portion (T_{RC}) is much larger than inverter delay ($T_{RC} >> T_{INV}$). As a result, delay sensitivity to temperature is dominated by the RC section, which is reduced by using a metal capacitor and implementing the resistor with a combination of resistors that have opposing temperature coefficients. Coarse delay tuning is performed by a 6-bit MOM capacitor DAC (LSB = 5fF) and 14-bit fine tuning (6.7ppm frequency resolution) with a PMOS varactor. Simulations of DCRO across PVT show a tuning range of ±5% and frequency variation less than ±5000ppm across the entire temperature range of -40°C to 85°C. The measured temperature sensitivity plot shown in Fig. 18.5.3 verifies this experimentally.

A prototype fast start-up XO was implemented in a 65nm CMOS process. Figure 18.5.4 shows XO output waveforms that illustrate the measured start-up behavior under different conditions. The proposed 2-step injection reduces T_{START} from 600µs (no injection) to 19µs (1µs (DCRO amplitude and frequency settling) + 2µs (T_{INJ1}) + 6µs (T_{LOCK}) + 10µs (T_{INJ2})), which represents >30× improvement. In order to compare how the achieved T_{START} = 19µs compares with the theoretical minimum, T_{START} was measured across different injection times while injecting with a precisely-calibrated signal ($F_{INJ} = F_{XO}$) and the best possible T_{START} was found to be $T_{INJ,OPT}$ = 13µs (see Fig. 18.5.4). Thus, the proposed approach reduces T_{START} to within 1.5× of the theoretical minimum. Figure 18.5.5 shows the measured temperature dependence of T_{START} for normal start-up (no injection), precisely-timed injection [3], and the proposed 2-step injection. The transient start-up behavior for the worst-case cold-start conditions is also shown. Thanks to the fast-locking PLL and the 2nd injection, T_{START} remains <20µs across the entire temperature range while it increases up to 380µs for the precisely-timed injection case. The buffered XO output achieves a phase noise of –149.7dBc/Hz at 10KHz, and –157.8dBc/Hz at 100kHz, and integrated jitter of $100fs_{rms}$ from 10kHz to 10MHz. The start-up energy is 34.9nJ of which injection buffers and oscillator consume 78% and 12.2%, respectively, indicating the penalty due to the PLL is minimal. A performance summary and comparison table is shown in Fig. 18.5.6. This work achieves the best T_{START} improvement of >30x in a robust manner while operating with large swing and achieving excellent phase noise. The die micrograph is shown in Fig. 18.5.7 with active area of 0.069mm².

References:
[1] D. Griffith et al., "A 24MHz Crystal Oscillator with Robust Fast Start-Up Using Dithered Injection," *ISSCC*, pp. 104-105, Feb. 2016.
[2] S. Iguchi et al., "Variation-Tolerant Quick-Start-Up CMOS Crystal Oscillator with Chirp Injection and Negative Resistance Booster," *IEEE JSSC*, vol. 51, no. 2, pp. 496-508, Feb. 2016.
[3] H. Esmaeelzadeh et al., "A Quick Startup Technique for High-Q Oscillators Using Precisely Timed Energy Injection," *IEEE JSSC*, vol. 53, no. 3, pp. 692-702, March 2018.
[4] Y.-K. Tsai et al., "A 51.3-MHz 21.8-ppm/°C CMOS Relaxation Oscillator with Temperature Compensation," *IEEE TCAS-II*, vol. 64, no. 5, pp. 490–494, May 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

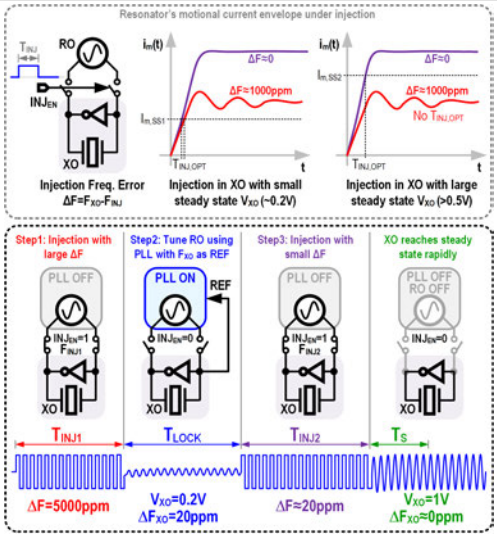


Figure 18.5.1: Illustration of the principle behind the proposed XO start-up time reduction technique.

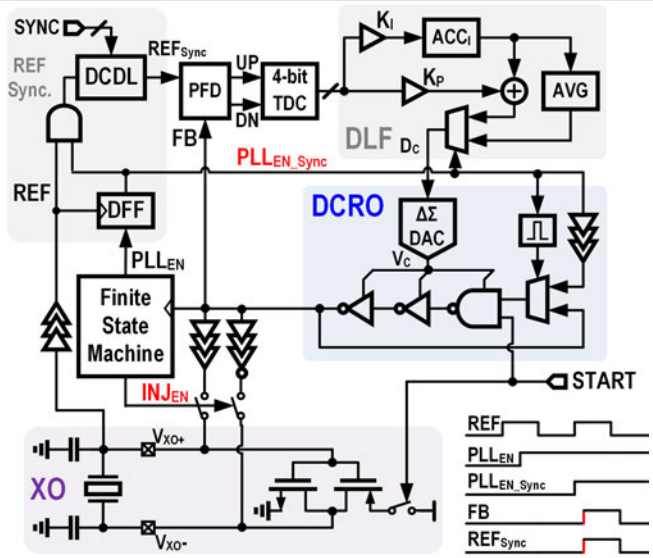


Figure 18.5.2: Proposed XO architecture.

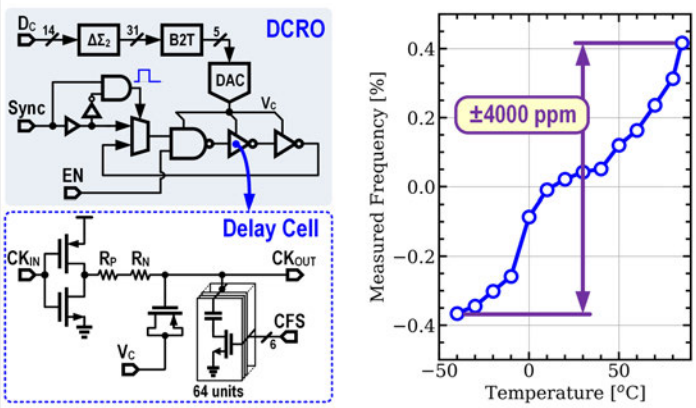


Figure 18.5.3: Details of digitally controlled ring oscillator and its measured temperature sensitivity.

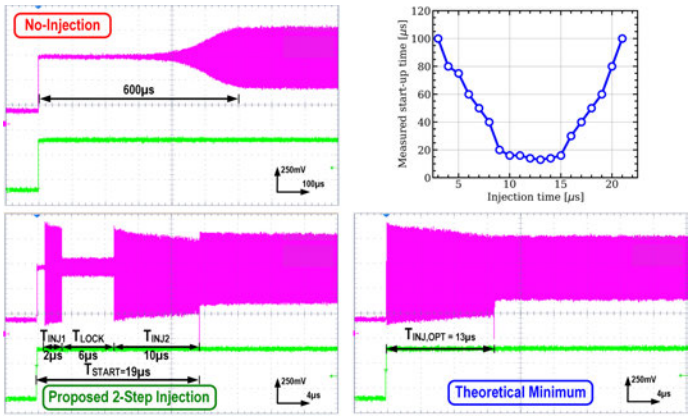


Figure 18.5.4: Measured XO start-up time waveforms at nominal temperature for normal start-up (no injection), theoretical minimum, and the proposed 2-step injection.

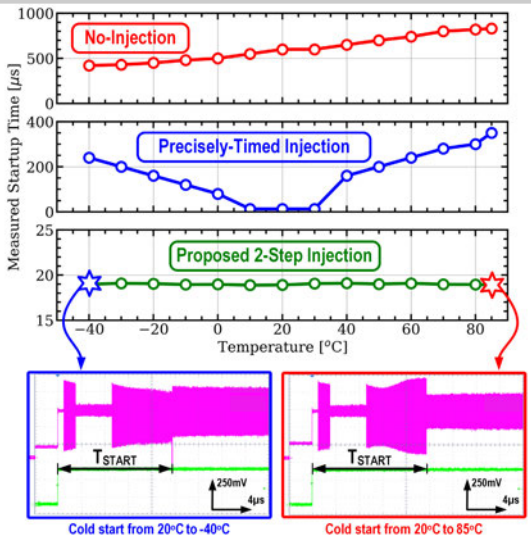


Figure 18.5.5: Measured XO start-up time across temperature for normal start-up (no injection), precisely-timed injection, and the proposed 2-step injection.

		Chang JSCC'12	Iguchi JSCC'16 [2]	Griffith ISSCC'16 [1]	Ding ISSCC'17	Lei ISSCC'18	Esmaeelzadeh JSSC'18 [3]		This Work
Technique		None	Chirp Injection+NRB	Dithered Injection	Dynamically-Adjusted Load + gm	ULV inductive 3-stage gm + SSCI	Precisely-Timed Injection		2-step Injection
Technology [nm]		65	180	65	90	65	65		65
Supply voltage [V]		1.8	1.5	1.68	1.0	0.35	1.0		1.0
Frequency [MHz]		26	39	24	24	24	50		54
Xtal manufacturer		N/A	Daishinku	N/A	N/A	N/A	Abracon	TXC	Connor-Winfield
Load capacitance [pF]		8	6	6	10	6	9		6
Steady state core power [uW]		2160	181	393*	95	31.8	74.2	195	198
Start-up time [usec]		3200	158	64	200	400	1.95	2.2	19
Start-up time [cycles]		83333	6162	1536	4800	9615	98	110	1026
Start-up reduction ratio		1x	13.3x	6.7x	13.3x	3.3x	N/A		31.5x
Steady state amplitude [V]		1.0	N/A	N/A	N/A	0.28	<0.25		0.7
PN (dBc/Hz)	@1KHz	-136	-147	N/A	N/A	-134	N/A		-139.5
	@10KHz	-149.1	-161			N/A			-149.7
	@100KHz	-153.4	-172.5						-157.8
	@1MHz	N/A	-175.5						-163.1
Output jitter [fs_{rms}]		420 [10k – 5MHz]	N/A	N/A	N/A	N/A	N/A		103 [10k – 10MHz]
Start-up energy [nJ]		N/A	349	N/A	36.7	14.2	9.4	13.3	34.9
Start-up time over temperature		N/A	±7% [-30 to 125°C]	±35% [-40 to 90°C]	27.5% [-40 to 90°C]	7.5% [-40 to 90°C]	10% [-40 to 85°C]	7%	±1.25% [-40 to 85°C]
Area [mm²]		0.15	0.12	0.08	0.072	0.023	0.09 per XO		0.069

Figure 18.5.6: Performance comparison with state-of-the art fast start-up XOs.

18

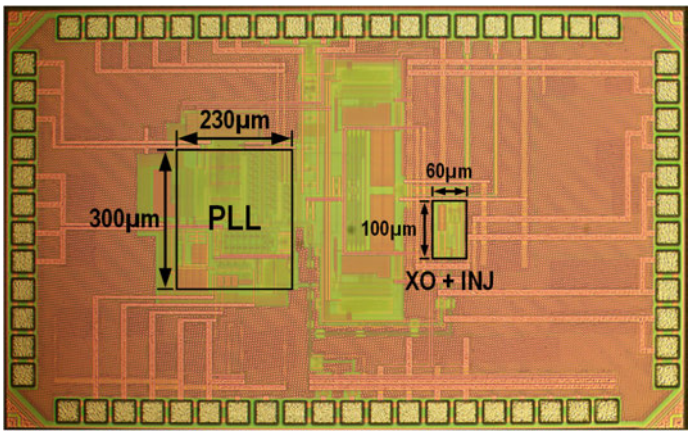


Figure 18.5.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

18.6 A 32MHz Crystal Oscillator with Fast Start-up Using Synchronized Signal Injection

Bram Verhoef[1], Jan Prummel[2], Wim Kruiskamp[2], Rein Post[2]

[1]Dialog Semiconductor, Hengelo, The Netherlands
[2]Dialog Semiconductor, 's-Hertogenbosch, The Netherlands

Low-power sensor nodes (e.g. Bluetooth Low Energy, BLE) use low duty-cycle transceivers to obtain an overall low power consumption. The system typically spends most of the time in a low-power sleep state and only briefly wakes up to transmit and receive data. Due to the high Q (>100k) of the crystal resonator the start-up time of a typical crystal oscillator is relatively long (≈ 0.5 to 5ms) which causes a substantial amount of energy to be consumed by the crystal oscillator in a transmit/receive event. As the system usually is in an active state when the crystal oscillator ramps up (multiple sub-systems are powered up) not only should the start-up energy of the crystal oscillator be optimized, but also the start-up time. As the crystal resonator is one of the most costly components in IoT sensor nodes, a wide range of crystal resonators should be supported by the circuit. This work proposes a synchronized signal injection (SSI) start-up mechanism that can drive the crystal up to any amplitude in a minimum amount of time, does not require a high-precision RC oscillator, is insensitive to PTV and is effective for a wide range of crystals: the technique does not require the shunt capacitance or the load capacitance to be small. A programmable capacitor-bank is included to support a range of crystals (load-capacitance) and allow for the customer to calibrate the frequency but is not required for the SSI technique.

The start-up of a crystal oscillator can be reduced in several ways. An increased GM (bias) may be used with the dynamic load-capacitance adjustment [2] technique. At start-up, not only the GM is boosted, but also the load-capacitance is reduced (programmable cap-bank). As shown by [1], this technique is only effective if the shunt capacitance of the crystal and PCB (C_0) is low, which tightens the requirements of the crystal and PCB design. Further techniques employ some sort of signal injection to drive the crystal resonator, providing some initial energy from which the oscillation starts. The amplitude that is obtained at start-up by the signal injection is proportional with the number of driving clock-cycles (1-to-3mV/clock-cycle) and is limited by the frequency accuracy and noise of the injected signal, specifically for smaller crystals (low equivalent resonance capacitance, C_m). Typically, approximately 300-to-600 clock cycles are required to obtain a peak-to-peak signal level of 750mV (drive level of 1.2V). In [3] an accurate (average) frequency is obtained by dithering the frequency setting of an RC oscillator and a maximum drive time of about 120 drive cycles is achieved over PVT. Most of the amplitude needs to be obtained by a conventional start-up, which means the total start-up is only fast when a crystal resonator with a low shunt capacitance (C_0) and a load-capacitance (C_L) is used. This problem is addressed by [4], in which a 3-stage GM reduces the effect of C_0 and reduces power consumption. The start-up time, however, is about 400µs.

The SSI start-up mechanism is shown in Fig. 18.6.1. An RC oscillator (RCOSC) is used that drives the crystal resonator in bursts. Additionally, a DRIVER, DELAY, and a control block (CTRL) are used in the start-up sequence of the oscillator. The comparator (CMP) is used by the start-up circuitry and is reused in steady-state operation to generate the output clock signal. After each burst (drive) the phase of the RC oscillator is re-aligned with the crystal resonance. By using this technique, the crystal can be driven indefinitely. Even if the frequency of the driving signal is not exactly correct, the phase error never accumulates to a point where it starts to counteract the crystal resonance. Therefore, only a one-time trim of the RCOSC is required.

A detailed view of the SSI driving sequence is shown in Fig. 18.6.2. The resonance voltage VL_M increases by $2V_{DD}$ after each clock edge (differential drive). After N clock-cycles the sync signal that is generated by the control block stops the RCOSC and the DRIVER. After a delay, in which the highpass filter in the comparator is allowed to settle, the control block clears the sync signal. The RCOSC and DRIVER are subsequently re-started on the following negative clock edge of the dxtal_d signal. This synchronizes the RCOSC signal to the crystal resonance. As the phase of the crystal resonance is 90° with respect to the drive signal, the dxtal signal is delayed by a quarter period. The accuracy of this delay is not critical because it introduces a one-time phase error that does not accumulate over clock cycles. A test mode has been added that allows for a one-time calibration of the delay. In each burst the amplitude of the resonance grows linearly and after 10 bursts a steady-state peak-to-peak differential resonance voltage level of 750mV has been reached. A one-time calibration uses a status signal of the amplitude regulator to determine the required number of drive cycles. When the drive sequence ends, the standard negative conductance driver (GM) and amplitude regulator [1] are enabled and a steady-state oscillation is maintained. During the crystal start-up phase, the programmable capacitor bank is set to its lowest setting (C_{CB}=9pF). After the initial burst of N=32 clock-cycles the driver is disabled and a peak-to-peak resonance voltage of approximately 100mV is obtained, which sufficiently exceeds the offset of the comparator (9mV, 3σ).

The crystal oscillator is part of a BLE SoC [5] and was implemented in 55nm LP CMOS. The active area of the start-up circuitry is 0.035mm^2 (not including control logic), the core oscillator area is 0.049mm^2 and a programmable capacitor bank is 0.08mm^2 (Fig. 18.6.7). The steady-state power consumption is 190uW (typical crystal) at a supply voltage of 1.2V. Figure 18.6.3 shows a start-up sequence without any fast-start-up technique and one with the SSI mechanism, with a crystal/PCB that has a low shunt capacitance (≈0.63pF). The start-up time is defined as the duration for the frequency to settle to within 20ppm (640Hz) and the amplitude to settle to within 90% of its final value. The amplitude regulator is set to control to a peak-to-peak differential voltage of 750mV. Without any fast-start-up technique the frequency settles in 400µs. With SSI, the start-up time reduces to 23µs. Furthermore, the amplitude is at the required steady-state level directly after the drive sequence finishes (after ≈16µs). The crystal is driven by the RCOSC for a total of 300 clock-cycles in N=32 cycle bursts (1µs); the synchronization phase takes 7 clock-cycles (200ns). Fig.ure 18.6.4 shows that the SSI start-up technique also improves the start-up performance when a high shunt- and load capacitance PCB/crystal is used (7pF, 12pF). The conventional design starts up (even with optimum GM [1]) in 2.9ms. When the SSI technique is enabled, the start-up time reduces to 32µs.

The obtained resonance amplitude and required start-up energy are shown as a function of the number of drive clock-cycles in Fig. 18.6.5. The output voltage can be driven up to any voltage level and is seen to increase with ≈2.4mV/clock-cycle, which is close to the maximum of 3.0mV/clock-cycle. The start-up time is shorter than 25µs (variation less than ±2.5µs) over a -40 to 140°C temperature range. Figure 18.6.6 summarizes the SSI performance and compares it to prior-art. The proposed SSI technique allows for a fast start-up of a crystal oscillator with low or high parasitic crystal resonators/PCB layouts. As the injection signal is re-synchronized, the frequency and noise performance requirements of the reference RCOSC are strongly reduced and the amplitude of the resonance can be driven to an arbitrary level.

References:

[1] E. A. Vittoz, M. G. R. Degrauwe and S. Bitz, "High-Performance Crystal Oscillator Circuits: Theory and Application," in *IEEE JSSC*, vol. 23, no. 3, pp. 774-783, June 1988.
[2] M. Ding et al., "A 95µW 24MHz Digitally Controlled Crystal Oscillator for IoT Applications with 36nJ Start-Up Energy and >13× Start-Up Time Reduction Using a Fully-Autonomous Dynamically-Adjusted Load," *ISSCC*, pp. 90-91, Feb. 2017.
[3] D. Griffith, J. Murdock and P. T. Røine, "A 24MHz Crystal Oscillator with Robust Fast Start-Up Using Dithered Injection," *ISSCC*, pp. 104-105, 2016.
[4] K. Lei, P. Mak, M. Law and R. P. Martins, "A Regulation-Free Sub-0.5V 16/24MHz Crystal Oscillator for Energy-Harvesting BLE Radios with 14.2nJ Start-Up Energy and 31.8pW Steady-State Power," *ISSCC,* pp. 52-54, 2018.
[5] J. Prummel et al., "A 10mW Bluetooth Low-Energy Transceiver with On-Chip Matching," *ISSCC,* pp. 1-3, Feb. 2015.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

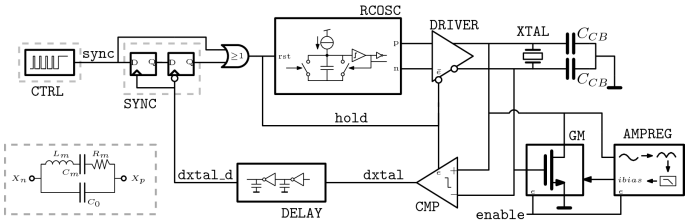


Figure 18.6.1: Block diagram of the fast-start-up crystal oscillator using synchronized signal injection (SSI) and lumped element model of crystal resonator. C_0 represents the crystal shunt/PCB capacitance and C_{CB} is implemented by a capacitance bank (on-chip).

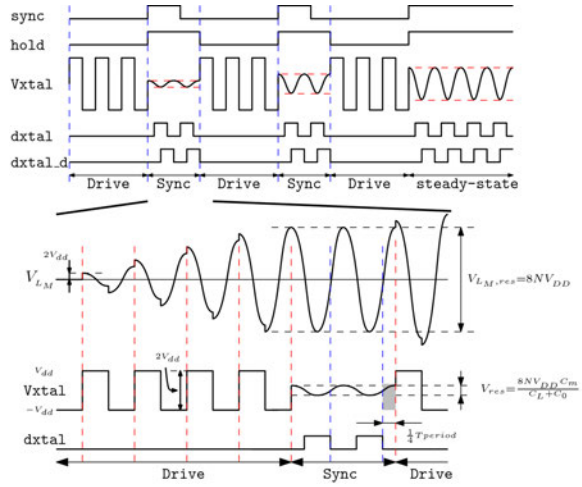


Figure 18.6.2: Detail of SSI drive sequence: the crystal is driven in bursts of N periods. After each burst the amplitude of the resonance increases with V_{res}, which is proportional with N, the supply voltage (V_{DD}) and crystal parameter C_m and inversely proportional to the load and shunt capacitance, C_L+C_0.

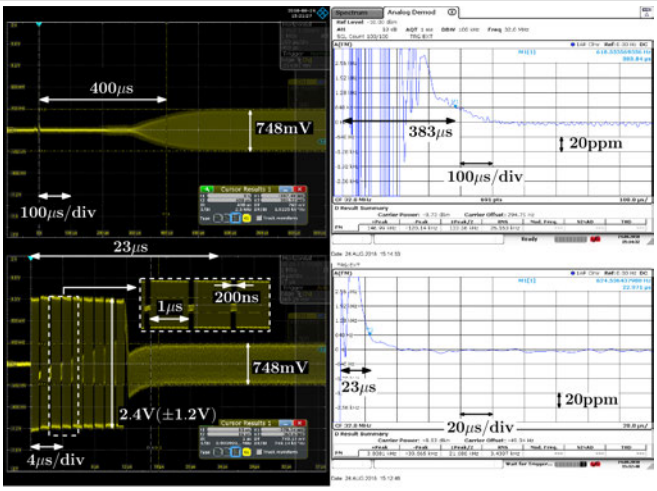


Figure 18.6.3: Measured start-up behavior (resonator voltage and frequency over time) with a typical crystal (C_L=6pF, C_{CB}=12pF). Top: conventional start-up, bottom: SSI start-up. The transient and frequency behavior are shown on the left and right side, respectively.

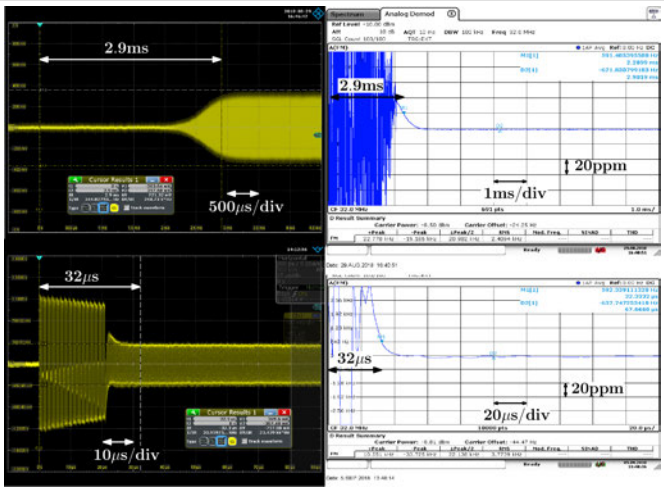


Figure 18.6.4: Measured start-up behavior with crystal that has a high shunt capacitance (C_0=7pF, on PCB) and high load capacitance (C_L=12pF). The capacitance bank is set to C_{CB}=10pF to obtain the required load-capacitance. Top: conventional start-up (optimum GM), bottom: SSI start-up.

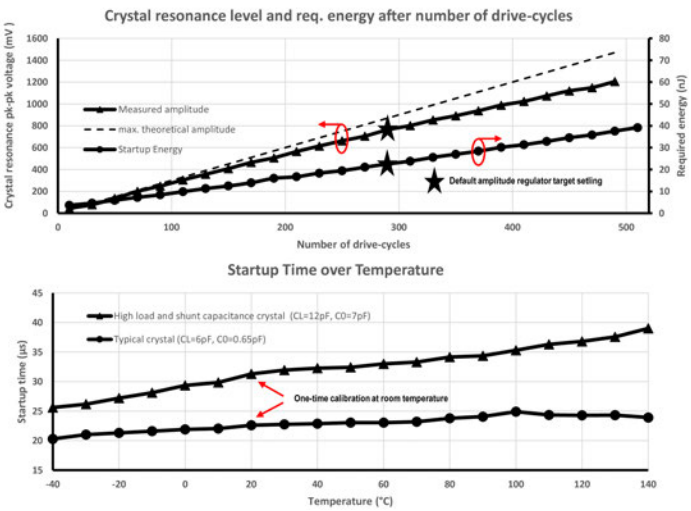


Figure 18.6.5: Resonance amplitude and start-up energy versus number of drive cycles and start-up time over temperature.

	This work		[2]		[3]		[4]
Fast startup technique	**SSI**		DAL	DAL + gm	Dithered inj.		3-stage GM
Technology (nm)	**55**		90		65		65
Core area (mm^2)	**0.049***		0.072		0.08		0.023
Supply (V)	**1.2**		1		1.68		0.35
Frequency (MHz)	**32**		24		24		24
Load capacitance , C_L (pF)	**6**	**12****	10		6	9	6
Startup time, T_S (μs)	**23**	**32**	375	200	64	435	400
Startup energy (nJ)	**20.2**	**44.2**	39	36.7	-		14.2
Area of startup circuits (mm^2)	**0.035**		0.0073		0.017		-
T_S variation with Temp ($\pm\mu s$).	**2.3**	**6.7**	97.5	55	22.4	87	30
Temperature range ($°C$)	**-40 to 140**		-40 to 90		-40 to 90		-40 to 90

*does not include capacitor bank area ($0.08mm^2$)
**7pF shunt-capacitance (capbank $C_{CB} = 10pF$)

Figure 18.6.6: Performance summary and comparison to prior art.

18

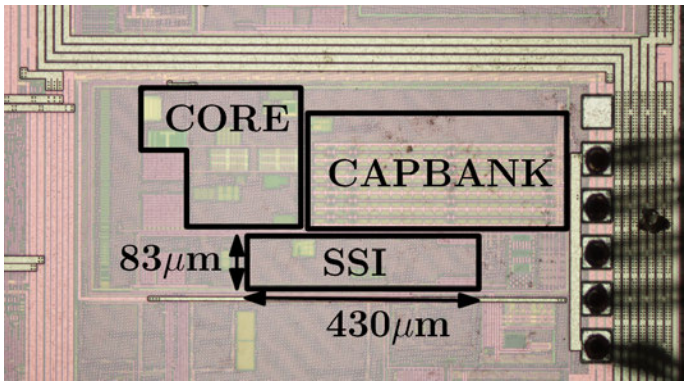


Figure 18.6.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

18.7 A 0.7V, 2.35% 3σ-Accuracy Bandgap Reference in 12nm CMOS

Yi-Wen Chen, Jaw-Juinn Horng, Chin-Ho Chang, Amit Kundu, Yung-Chow Peng, Mark Chen

TSMC, Hsinchu, Taiwan

Bandgap reference (BGR) circuits are widely used due to their stable output voltage over process, supply voltage and temperature variations. Reference voltage stability is critical for data-acquisition applications and lower supply voltages can reduce the power of mixed-signal systems. However BGR for analog circuits is one of the bottlenecks for sub-1V supply operation because BGR supply voltage is limited by $V_{EB}+V_{DS}$ [1]. V_{EB} refers to the emitter-base voltage of a pnp transistor which is limited to ~0.6 to 0.7V due to silicon junction cut-in voltage, while V_{DS} is the drain-source saturation voltage of a current-mirror. The BGR temperature dependence is decided by the weighted sum of proportional-to-absolute-temperature(PTAT) and complementary-to-absolute-temperature (CTAT) terms. An alternative PTAT generator can be implemented by dV_{GS} (gate-to-source voltage difference) of a MOS pair in subthreshold [2]. The CTAT generator can be implemented by special devices or using the gate-source voltage V_{GS} of subthreshold MOSFETs. Although the V_{GS} of a subthreshold MOSFET is smaller than emitter-base voltage of a pnp transistor, the MOSFET model inaccuracy in the subthreshold region and high process-dependent characteristic of MOSFET gate-source voltage induces high variation for voltage reference circuits.

A sampled BGR [3] uses a capacitive voltage divider to divide the output voltage of the bipolar transistor without resistors and it achieves sub-1V operation. A 2×-charge-pump-based BGR [4] further pumps the supply voltage to enable 0.5V operation. However both structures need switched-capacitor circuits, which creates output ripple and is hence a major design trade-off in high-accuracy data converters.

Figure 18.7.1 shows the measured values of V_{EB} and dV_{EB} (emitter-to-base voltage difference) at 25°C and -40°C at different bias currents. The data was obtained from 35 pairs of nominally identical BJTs (emitter areas of ~3um^2), which were biased at a current density ratio of 15:1. V_{EB} is decided by the design trade-off between bias current, minimum operating temperature and supply voltage. Lower supply voltage needs a lower and stable V_{EB} using small BJT bias current and also needs to consider the operating temperature range. Figure 18.7.1 also shows not only that the BJT has larger V_{EB} and dV_{EB} variation at lower bias current but also MOS current-mirror mismatch is increased due to subthreshold operation. If the V_{EB} at -40°C decreases from 0.9V to 0.6V by reducing the BJT bias current from 10uA to a few nanoamps for low-supply-voltage application, the current mirror area blows up by 16× to meet the current-mirror matching requirement of a 3% 3σ output accuracy. A large value of resistance is also needed to provide low bias current, which costs additional silicon area. Although BJT bias current density can be reduced by modifying the 1:N BJT array to M:N×M BJT arrays to relax current-mirror and resistance area, BJT array area will increase. As a result, there is an area trade-off among current-mirror, BJT array and resistor array choices.

Figure 18.7.2 shows the proposed Bandgap circuit diagram, which is composed of a current-mode Bandgap and an automatic current-controlled feedback loop. The automatic current-controlled feedback loop is composed of a current mirror, an opamp and a resistor. Through the opamp negative feed-back characteristic, the voltage drop across R5 is the difference between $V_{EB(Q1)}$ and $V_{EB(Q2)}$, defining the current as $dV_{EB}/R5$. Since voltage drop across R3 is also dV_{EB}, the current flowing into the feedback loop can be derived as $(R3/R5)\cdot I_{R3}$. Therefore, the BJT current is controlled well by resistor ratio so as to let the BJT mismatch be fixed even for a lower bias-current range. Meanwhile, the opamp in the feedback loop produces a feedback signal to both of its inputs. An overall negative feedback is ensured since M5 output resistance is $R_{ds(M5)}//(R5+R_{Q1}//R1//R_{ds(M1)})$ and M4 output resistance is $R_{ds(M4)}//(R_{Q2}//(R3+(R2//R_{ds(M2)})))$, and negative-feedback loop gain is larger than positive- feedback loop gain due to R5. Figure 18.7.3 shows the Bode plot of the automatic current-controlled feedback loop. The DC gain is 52.8dB with a phase margin of 45°. With the automatic current-controlled feedback loop, the current flowing into the BJT array is well controlled for low supply voltage operation and prohibits the current-mirror MOS to enter the subthreshold region without increasing the MOS size and thus ensuring good mismatch performance.

Figure 18.7.3 also shows the output-variation breakdown chart of the proposed Bandgap. The dominant variation sources are the two current mirrors (M1, M2, M3 and M4, M5). Both have mismatch values designed at 0.08%. However the contributed percentages of output variation are 62% and 13% respectively for the 2 aforementioned current mirrors since the current-mirror mismatch in the automatic current-controlled feedback loop is suppressed by the feedback loop gain. The sizes of the two current mirrors can be further optimized and their area can be shrunk further based on the mismatch impact level and current ratio. The current ratio between the two current mirrors is decided by the resistor ratio of R5/R3, the temperature sensitivity of PTAT, CTAT voltage and the minimum supply voltage. It also needs to be mentioned that the 1σ input-offset voltage variation of the opamp in the automatic current-controlled feedback loop only contributes 1% of total output variation. The opamp input-offset voltage (V_{OS}) changes the voltage across R5 from dV_{EB} to $dV_{EB}+V_{OS}$ and consequently current ratio I_{R3}/I_{M5} is derived as $R5/R3/(1-V_{OS}/dV_{EB})$. Since V_{OS} can be easily designed to be much smaller than dV_{EB}, the impact of the opamp input-offset voltage to the output variation is almost negligible.

The proposed Bandgap was implemented in 12nm FinFET CMOS technology. 10nA BJT bias current for a 1:15 BJT array was chosen to reduce V_{EB} and 100nA current in R3 (~900kΩ) is achieved by an R3/R5 ratio of 0.9. This ratio was selected to reduce total circuit current consumption. However for production purposes, a higher resistor ratio can be selected for increasing current in R3, which further mitigates Bandgap output variation and also reduces current-mirror area. Meanwhile, a current-mirror ratio (M1:M2:M3) of 1:1:1 is used for matching purposes. With the existing process model, simulation results show that current-mirror mismatch increases 47% and Bandgap output variation increases 23% when the automatic current-controlled feedback loop is removed and the current mirror is directly biased at 10nA.

Figure 18.7.4 shows the on-wafer measured reference output voltages across temperature from 35 dies of 1 wafer with a 0.7V supply voltage. The untrimmed measured output voltage from -20°C to 125°C varies from 204mV to 210mV. The measured temperature coefficient is from 4.9ppm/°C to 82.3ppm/°C and the average temperature coefficient is 40.51ppm. Figure 18.7.5 shows the measured distribution diagram of the output voltage at 25°C and 0.7V supply voltage. The average voltage is 207mV and the standard deviation is 1.63mV, showing a 3σ-accuracy of 2.35%. The measured 3σ-accuracy at 25°C also shows a constant value of approximately 2.35% with supply voltage changed from 0.7V to 0.9V.

Figure 18.7.6 compares this work with other state-of-the–art voltage references that achieve <3.5% 3σ-accuracy. With the proposed automatic-feedback-loop technique, this BJT-based reference achieves competitive accuracy with a low supply voltage, while providing a ripple-free reference output voltage. Figure 18.7.7 shows the die micrograph.

References:

[1] H. Banba et al., "A CMOS Bandgap Reference Circuit with Sub-1-V Operation," *IEEE JSSC*, vol. 34, no. 5, pp. 670-674, May 1999.

[2] Y. Osaki et al., "1.2-V Supply, 100-nW, 1.09-V Bandgap and 0.7-V Supply, 52.5-nW, 0.55-V Sub-bandgap Reference Circuits for Nanowatt CMOS LSIs," *IEEE JSSC*, vol. 48, no. 6, pp. 1530-1538, June 2013.

[3] V. Ivanov et al., "An Ultra Low Power Bandgap Operational at Supply from 0.75V," *IEEE JSSC*, vol. 47, no. 7, pp. 1515-1523, July 2012.

[4] A. Shrivastava et al., "A 32nW Bandgap Reference Voltage Operational from 0.5V Supply for Ultra-Low Power Systems," *ISSCC*, pp. 94-95, Feb. 2015.

[5] Y. Ji et al., "A 9.3nW All-in-One Bandgap Voltage and Current Reference Circuit," *ISSCC*, pp. 100-101, Feb. 2017.

[6] J. M. Lee et al., "A 29nW Bandgap Reference Circuit," *ISSCC*, pp. 100-101, Feb. 2015.

[7] U. Kamath et al., "A 1 V Bandgap Reference in 7-nm FinFET with a Programmable Temperature Coefficient and an Inaccuracy of ±0.2% from -45°C to 125°C," *ESSCIRC*, pp. 78-81, 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

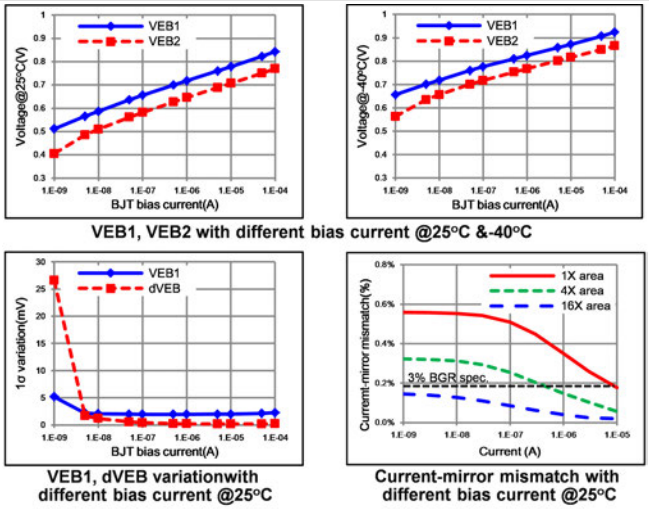
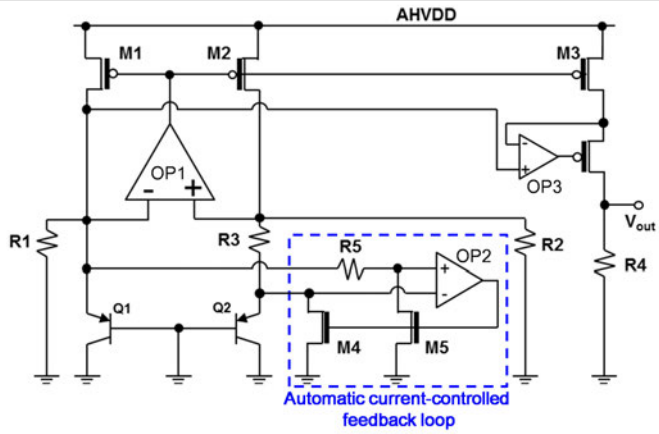


Figure 18.7.1: V_{EB}, dV_{EB} silicon results @25°C, -40°C and simulated current-mirror mismatch at 25°C. The current-density ratio between BJTs of V_{EB1} and V_{EB2} is 15:1.



$$V_{out} = I_{M3} \times R4 = (I_{R3} + I_{R2})R4 = (\frac{dV_{EB}}{R3} + \frac{V_{EB}}{R2})R4$$

$$\frac{I_{R3}}{I_{M5}} = \frac{dV_{EB}/R3}{dV_{EB}/R5} = \frac{R5}{R3}$$

Figure 18.7.2: Proposed Bandgap circuit diagram.

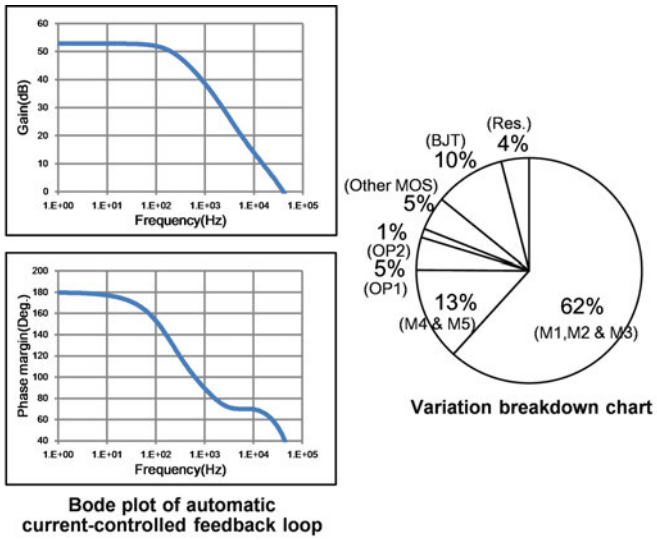


Figure 18.7.3: Stability and output variation sensitivity analysis of proposed Bandgap.

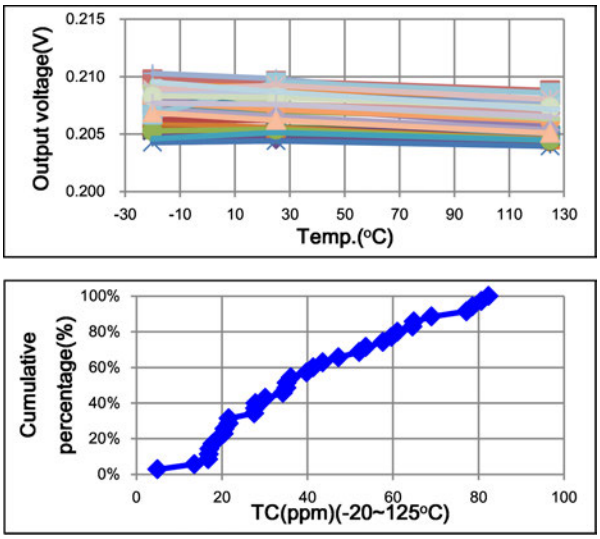


Figure 18.7.4: Measured output voltage from 35 dies and the cumulative distribution of output voltage from -20°C to 125°C.

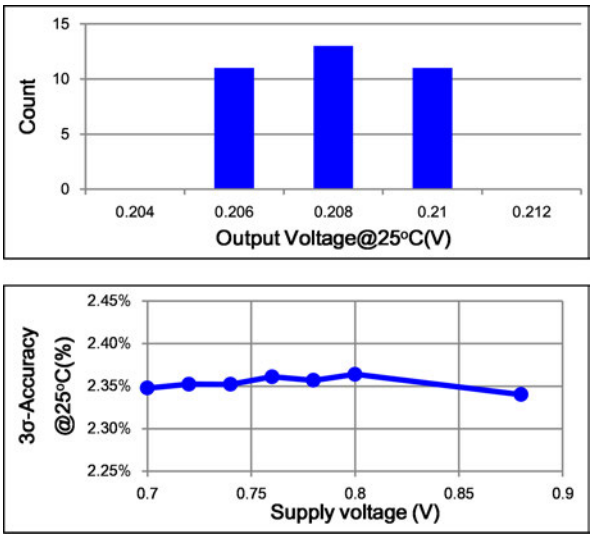


Figure 18.7.5: Measured distribution diagram of the output voltage at 25°C and 0.7V supply voltage and 3σ-accuracy at 25°C from 0.7V to 0.9V supply voltage.

	JSSC 2012 [3]	ISSCC 2015 [4]	JSSC 1999 [1]	JSSC 2013 [2]	ISSCC 2017 [5]	ISSCC 2015[6]	ESSCIRC 2018 [7]	This work
Process	0.13um	0.13um	0.4um	0.18um	0.18um	0.35um	7nm	12nm
Min. supply voltage(V)	0.75	0.5	0.84	0.7	1.2	1.2	1.375	0.7
Temp. Range(°C)	-20~85	0~80	27~125	-40~120	0~110	-10~110	-45~125	-20~125
$3\sigma_{meas.}$(%)	3	2	2.9	3.15	1.29	0.6	0.6	2.35
$TC_{meas.}$ (ppm)	40	75	119	114	26	12.75	6	40.51
# of chips	90000	6	23	9	10	10	7	35
Type	Switching cap	Switching cap	DC	DC	DC	DC	DC	DC
Output ripple	20mV	50uV	No	No	No	No	No	No
BJT only	Yes	Yes	Yes	No	No	No	Yes	Yes
Area (mm²)	0.07	0.0264	0.1	0.0246	0.055	0.48	0.078	0.065
Power	170nW	32nW	1.85uW	52.5nW	9.3nW	28.7nW	947uW	6.88uW

Figure 18.7.6: Measurement summary and comparison table.

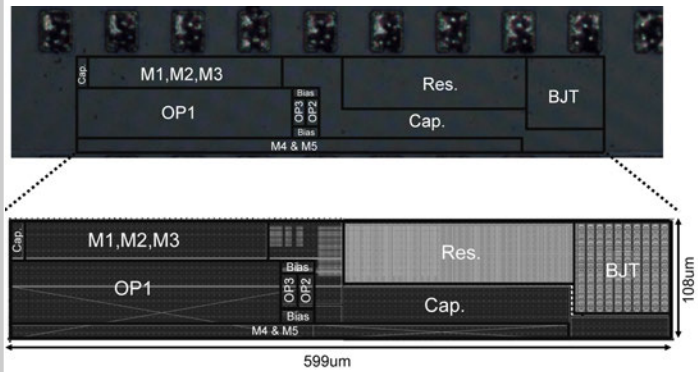


Figure 18.7.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

18.8 A 192pW Hybrid Bandgap-V_{th} Reference with Process Dependence Compensated by a Dimension-Induced Side-Effect

Youngwoo Ji, Jungho Lee, Byungsub Kim, Hong-June Park, Jae-Yoon Sim

Pohang University of Science and Technology, Pohang, Korea

A voltage reference circuit is an essential block of a system to generate various internal voltages. Since it consumes static power in standby modes, it plays an important role in energy management of battery-limited applications. The bandgap reference (BGR) has been a widely used approach since it provides a well-defined large value (~1.15V) with strong immunity to process, supply and temperature changes. Recently proposed BGR approaches achieved a great reduction of power consumption by taking only complementary-to-absolute-temperature (CTAT) quantity from a PN junction while they obtained proportional-to-absolute-temperature (PTAT) quantity from alternative CMOS circuits such as a CTAT divider [1] or leakage-based two diodes [2,3]. However, these BGR schemes are formed with multiple branches fed from a supply voltage above 1.4V and require power consumption of order larger than 10nW. To further reduce power consumption, threshold-based reference approaches with CMOS-only circuits have been proposed [4,5]. However, generation of a practical voltage level by up-scaling of a threshold-based reference also causes an amplification of the uncertainty by the same factor. In addition, though [4] has successfully achieved sub-nW power consumption, the threshold voltage eventually suffers from a large sensitivity to process variation because the threshold voltage is affected by process and design parameters. To reduce the effect of process variation, [5] proposed a PMOS-only circuit. However, it requires a different body biasing for a threshold difference that is needed to generate a non-zero reference.

This paper presents a hybrid circuit scheme that generates a reference voltage associated with both the bandgap (V_{BG}) and the threshold voltage (V_{th}). The circuit generates a nominal value of $V_{BG} - V_{th}$ with process dependence compensated by a dimension-induced side-effect. The proposed circuit, fabricated in a 0.18μm CMOS technology, generates a reference voltage of 0.6926V and consumes 192pW from a supply voltage of 1V. Measurements show a standard deviation of 0.52% over process corners with an average temperature coefficient of 33ppm/°C.

The threshold voltage of a MOS transistor in modern technology cannot be clearly modeled since it is a complicated function of various physical effects coming from geometry design and doping non-uniformity. For example, the narrow-channel effect conventionally causes V_{th} to increase as the channel width decreases. However, if a shallow-trench isolation (STI) is adopted in technology, V_{th} decreases as the channel width decreases, which is called the reverse narrow-channel effect. The V_{th} also varies as the channel length decreases. The traditional short-channel effects of drain-induced barrier lowering (DIBL) and velocity saturation are also mixed with the reverse short-channel effect caused by HALO implants. As a result, for given technology, the V_{th} turns out to be a function of dimension-induced side-effects. Figure 18.8.1 illustrates the concept of process-dependence compensation with a typical case of a CMOS technology. It utilizes the change of V_{th} as the channel width or the channel length decreases. The V_{th} varies with a stronger sensitivity to dimension for smaller devices. Therefore, the V_{th} of a device with a narrow channel (W_2) or with a short channel (L_2) has a larger process sensitivity than that with a wide channel (W_1) or with a long channel (L_1). The V_{th} difference between a small and a large device ($V_{th,W2} - V_{th,W1}$ or $V_{th,L2} - V_{th,L1}$) can be taken as a quantity to be complementary to faster-skewed processes (CTFP). To compensate CTFP, we take the difference between the silicon bandgap and the threshold voltage ($V_{BG} - V_{th1}$) as a quantity with proportionality to a faster-skewed process (PTFP). The CTFP coefficient can be controlled by choosing the size of smaller device (W_2 or L_2). As a result, the process-dependence compensation can be achieved by balancing CTFP and PTFP, generating a reference voltage. It should be noted that this reference voltage includes a large process-independent quantity, V_{BG} (~1.15V). Therefore, this hybrid scheme provides a great benefit in the generation of a practical reference since it does not need an up-scaling of the small V_{th} which also amplifies the uncertainty by the same factor. In this work, we fixed the channel length and formed a circuit which uses the width-induced V_{th} modulation for process-dependence compensation.

Figure 18.8.2 shows the core circuit ideas to compensate both the process and temperature dependences. Flowing leakage current to a PN diode and to a MOS diode results in different characteristics. The PMOS transistors ($M_1 \times N_1$) with V_{GS} of 0 flow leakage current as

$$I \cong \mu C_{ox} \frac{N_1 W_1}{L} (m_1 - 1) {V_T}^2 e^{\frac{-V_{th1}}{m_1 V_T}}$$

where W_1, N_1, m_1 and V_T represent the selected width of unit transistor, the number of unit transistors, the subthreshold slope factor and the thermal voltage, respectively. Flowing this through the PN diode (Q_1) leads to a voltage of

$$\left(V_{BG} - \frac{V_{th1}}{m_1}\right) - V_T ln(kT^2)$$

where k represents a proportionality factor. The former term provides a PTFP as explained in Fig. 18.8.1. As the PN diode has stronger dependence on temperature than the MOS, the overall voltage follows a CTAT characteristic. On the other hand, flowing this through the MOS diode ($M_2 \times N_2$) leads to a voltage of

$$(V_{th2} - V_{th1}) + m_2 V_T ln\left(\frac{N_1 W_1}{N_2 W_2}\right).$$

The former and the latter terms are a CTFP and a PTAT voltages, respectively. The design procedure to achieve both process and temperature compensations starts from choosing the unit widths, W_1 and W_2, to compensate process dependence. With given W_1 and W_2, the ratio between N_1 and N_2 is then optimized to compensate the temperature dependence. Since the process variation terms are determined only by dimensions of W_1 and W_2, the optimization of N_1 and N_2 does not affect the compensated process dependence. This idea of compensation can be generally applied to different technologies. The plot in Fig. 18.8.2 shows simulated threshold voltage differences with extreme process corners for three different CMOS technologies. This reveals that an appropriate pair for W_1 and W_2 can be chosen to obtain CTFP if technology is given.

Figure 18.8.3 shows the overall circuit diagram of the proposed hybrid bandgap-V_{th} reference. The PN and MOS diodes of the two circuits in Fig. 18.8.2 are stacked in series to form a single branch circuit. To avoid an excessive size ratio between M_1 and M_2 for temperature compensation, two PMOS diodes ($M_2 \times N_2$) are stacked on the PN diode (Q_1). The reference voltage, V_{REF}, can be derived as

$$V_{REF} \cong \left[V_{BG} - \frac{V_{th1}}{m_1}\right] + [2(V_{th2} - V_{th1})] + \left[2m_2 V_T ln\left(\frac{N_1 W_1}{N_2 W_2}\right)\right] + [-V_T ln(kT^2)].$$

The first-to-fourth terms are PTFP, CTFP, PTAT and CTAT, respectively. As a result, the V_{REF} becomes approximately $V_{BG} - V_{th1,0}$ where $V_{th1,0}$ represents the process and temperature independent term of V_{th1}. 500 runs of Monte-Carlo simulation results show the effect of the proposed process-dependence compensation. PTAT and CTAT voltages with process corners are also plotted to demonstrate how both the process and temperature dependences are compensated together. A native NMOS transistor (M_{N1}) is inserted on top of the leakage transistor [6]. This structure improves the supply regulation by a factor of the intrinsic gain of M_{N1} while the core operation is not affected.

The proposed hybrid bandgap-V_{th} reference was implemented in a 0.18μm CMOS technology with an active area of 4500μm². The circuit generates a reference voltage of 0.6926V with 1V supply voltage and dissipates 192pW at 25°C. Figure 18.8.4 shows the distribution of untrimmed V_{REF} obtained from 45 chips (15 chips from each of TT, SS and FF wafers). The average temperature coefficient for the total chips is 33ppm/°C in a temperature range from -20°C to 100°C. The measured V_{REF} from TT, SS and FF wafers show standard deviations of 0.21%, 0.38% and 0.22%, respectively. The standard deviation of the total 45 chips over three wafers is 0.52%. Figure 18.8.5 summarizes measured supply sensitivity, showing a line sensitivity of 0.020%/V and PSRR of -55dB at 100Hz without any use of a capacitor. Figure 18.8.6 compares performance with previously reported state-of-the-art voltage reference circuits. Figure 18.8.7 shows the die micrograph.

Acknowledgements:
This work was supported by Green Science Program of POSCO

References:
[1] M. Kim et al., "A 0.8V, 37nW, 42ppm/°C Sub-Bandgap Voltage Reference with PSRR of -81dB and Line Sensitivity of 51ppm/V in 0.18um CMOS," *IEEE Symp. VLSI Circuits*, pp. 144-145, June 2017.
[2] J. M. Lee et al., "A 29nW Bandgap Reference Circuit," *ISSCC*, pp. 100-101, Feb. 2015.
[3] Y. Ji et al., "A 9.3nW All-in-One Bandgap Voltage and Current Reference Circuit," *ISSCC*, pp. 100-101, Feb. 2017.
[4] I. Lee et al., "A Subthreshold Voltage Reference With Scalable Output Voltage for Low-Power IoT Systems," *IEEE JSSC*, vol. 52, no. 5, pp.1443-1449, May 2017.
[5] Q. Dong et al., "A 114-pW PMOS-Only, Trim-Free Voltage Reference with 0.26% within-Wafer Inaccuracy for nW Systems," *IEEE Symp. VLSI Circuits*, pp. 98-99, June 2016.
[6] M. Seok et al., "A Portable 2-Transistor Picowatt Temperature-Compensated Voltage Reference Operating at 0.5 V," *IEEE JSSC*, vol. 47, no. 10, pp. 2534-2545, Oct. 2012.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

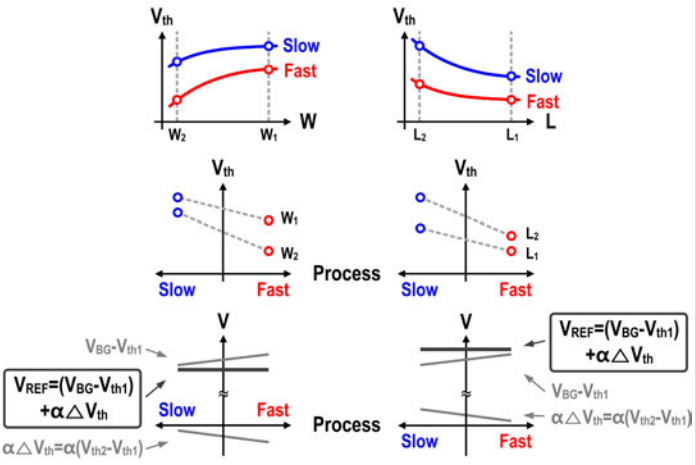


Figure 18.8.1: Concept of process-dependence compensation.

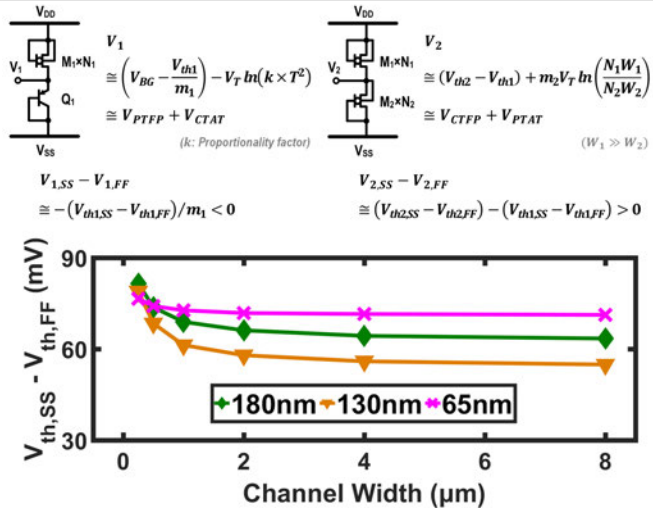


Figure 18.8.2: Core circuit ideas to compensate process and temperature dependences and simulated threshold differences ($V_{th,SS} - V_{th,FF}$) with different technologies.

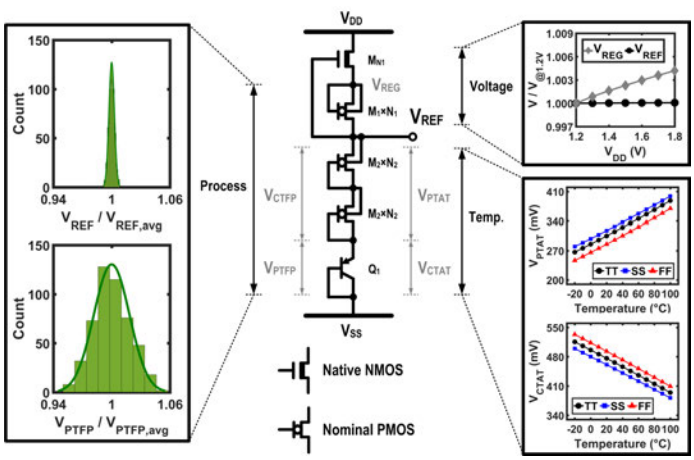


Figure 18.8.3: Overall circuit diagram of the proposed bandgap-V_{th} reference with simulated results over process/voltage/temperature variations.

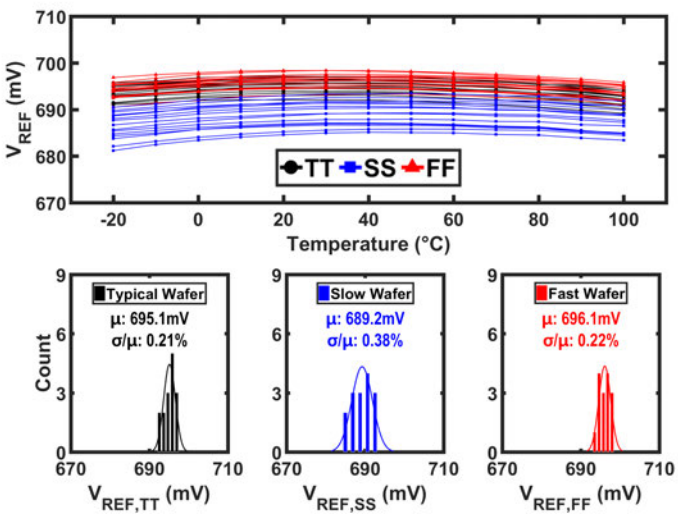


Figure 18.8.4: Measured untrimmed V_{REF} from 45 chips in three extreme-skewed wafers.

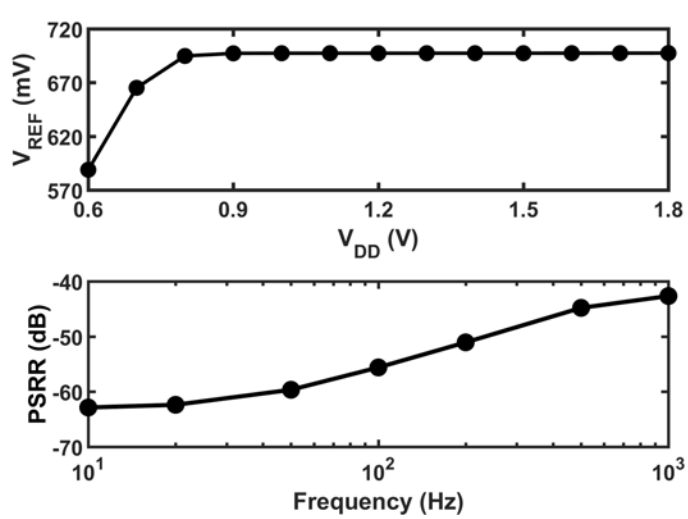


Figure 18.8.5: Measured line sensitivity and PSRR of V_{REF}.

	This work	[1] VLSI 17	[2] ISSCC 15	[3] ISSCC 17	[4] JSSC 17	[5] VLSI 16	[6] JSSC 12
Technology (μm)	**0.18**	0.18	0.35	0.18	0.18	0.18	0.18
Type	**Hybrid**	Hybrid	BJT	BJT	CMOS	CMOS	CMOS
Power (pW)	**192**	37000	28700	9300	34	114	6
Min. V_{DD} (V)	**1.0**	0.8	1.3	1.4	1.4	1.2	0.5
V_{REF} (V)	**0.6926**	0.477	1.176	1.238	1.25	0.862	0.3268
Measured Within-Wafer σ/μ (%)	**0.21**	0.89	0.2	0.43	0.8	0.26	0.8
Measured Wafer-to-Wafer σ/μ (%)	**0.52**	N/A	N/A	N/A	N/A	1.9	N/A
Line Sensitivity (%/V)	**0.020**	0.0451	0.198	0.08	0.31	0.38	0.044
PSRR @100Hz (dB)	**-55 w/o Cap**	-51 w/ N/A	N/A	-46 w/ N/A	-41 w/ 1.8pF	-42 w/ 1.8pF	-49 w/ N/A
Temp. Range (°C)	**-20~100**	-20~120	-10~110	0~110	0~100	-40~85	-20~80
T.C. (ppm/°C)	**33 (25~48)**	51 (7~90)	13 (N/A)	26 (N/A)	23 (8~53)	N/A (48~124)	N/A (54~176)
Trimming	**No**	Yes	N/A	No	No	No	No
Active Area (μm²)	**4500**	34000	480000	55000	2500	4880	1425
# of chips (wafer)	**45 (3)**	15 (1)	10 (1)	10 (1)	60 (3)	60 (3)	14 (1)

Figure 18.8.6: Comparison table.

18

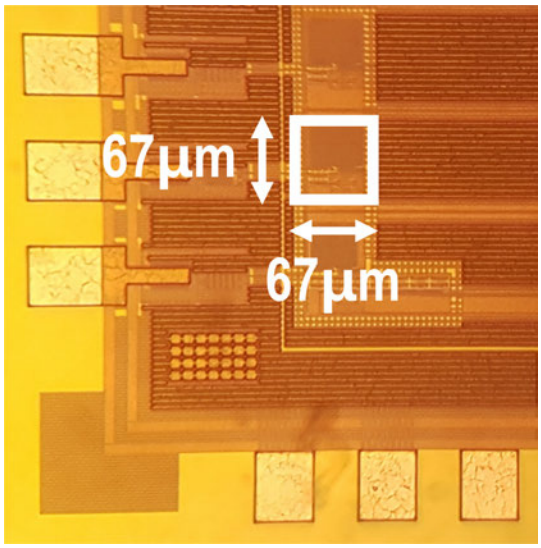


Figure 18.8.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 19 Overview:
Adaptive Digital and Clocking Techniques
DIGITAL CIRCUITS SUBCOMMITTEE

Session Chair: ***John Maneatis***
True Circuits, Los Altos, CA

Associate Chair: ***Alicia Klinefelter***
Nvidia, Durham, NC

Subcommittee Chair: Edith Beigné, Facebook, Menlo Park, CA

This session focuses on circuits that adapt to PVT, workload, and precision requirements, including novel techniques in clocking and digital regulators. The presented adaptive digital and clocking circuits demonstrate minimum energy/power-point tracking, adaptivity and unified clock and voltage control schemes, instruction-based dynamic clock scaling, digital leakage compensation for clocking circuits, adaptive forward/back biasing, and time-series processing for energy-efficient processing. The first two papers introduce novel techniques to track minimum energy and power in the system to adjust the body bias or supply voltage. The next paper unifies the output clock and the voltage supply control generation into a single loop to reduce voltage margins and decrease response times. The following paper exploits instruction-level dynamic timing slack to guide elastic pipeline timing redistribution guiding clock period adjustment. The last three papers cover clocking and biasing schemes to improve performance. The first of these papers focuses on reducing static phase offset and reference spurs that can degrade performance. The next paper explores back biasing in FDSOI to maximize power efficiency. The final paper uses a novel multi-bit time flip-flop to improve energy efficiency in time series classification.

8:30 AM

19.1 Computationally Enabled Total Energy Minimization Under Performance Requirements for a Voltage-Regulated 0.38-to-0.58V Microprocessor in 65nm CMOS

F. U. Rahman, University of Washington, Seattle, WA

In Paper 19.1, the University of Washington presents a fully integrated digital technique for total energy-per-cycle minimization under performance constraints demonstrated on a switched-capacitor-based 0.38-to-0.58V microprocessor in 65nm CMOS. Post-silicon measurements show minimum-energy-point (MEP) tracking within 5mV of the actual MEP across operating conditions, and seamless transition to performance-constrained operation.

9:00 AM

19.2 A 6.4pJ/Cycle Self-Tuning Cortex-M0 IoT Processor Based on Leakage-Ratio Measurement for Energy-Optimal Operation Across Wide-Range PVT Variation

J. Lee, University of Michigan, Ann Arbor, MI

In Paper 19.2, the University of Michigan makes a key observation that at the optimal energy operating point, the ratio of leakage to dynamic power is a constant value across temperature, process, and workload. A Cortex-M0 processor incorporates a control loop with a custom DC-DC converter to dynamically measure the leakage ratio and automatically adjust the body bias and supply voltage to achieve 6.4pJ/cycle at 0.55V and 200kHz.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

9:30 AM

19.3 A 7nm All-Digital Unified Voltage and Frequency Regulator Based on a High-Bandwidth 2-Phase Buck Converter with Package Inductors

F. I. Atallah, Qualcomm, Raleigh, NC

In Paper 19.3, Qualcomm presents a 7nm all-digital fast response 2-phase buck regulator with 3 turns and 6 metal-level package inductors that unifies the clock and voltage supply control generation into a single loop to reduce voltage margins arising from dynamic parameter variations with an area of 0.0065mm^2. Measurements show a fast response time of 60ns from a 177mA output current step with maximum FOUT measured at 3.0GHz, and VOUT is regulated from 0.45 to 0.9V with loads from 1 to 900mA.

10:15 AM

19.4 An Adaptive Clock Management Scheme Exploiting Instruction-Based Dynamic Timing Slack for a General-Purpose Graphics Processor Unit with Deep Pipeline and Out-of-Order Execution

T. Jia, Northwestern University, Evanston, IL

In Paper 19.4, Northwestern University introduces an instruction-based hierarchical adaptive clocking scheme to exploit instruction-level deterministic dynamic-timing slack for multicore deep-pipeline graphic processors. Critical-path messengers and elastic pipeline timing redistribution are used to guide the clock period adjustment with additional enhancement for machine learning applications and the proposed clocking scheme achieves 18.2% performance improvement or 30.4% energy savings in benchmark programs.

10:45 AM

19.5 Digital Leakage Compensation for a Low-Power and Low-Jitter 0.5-to-5GHz PLL in 10nm FinFET CMOS Technology

B. Xiang, Intel, Hillsboro, OR

In Paper 19.5, Intel presents a digital leakage compensation circuit for an analog PLL in 10nm FinFET CMOS technology to overcome the sensitivity to device or defect leakage. The PLL has a frequency range of 0.5 to 5GHz, a power dissipation of 0.32 to 2.65mA, and an area of 0.026mm^2, and the static phase error, random jitter, and reference spur are reduced from 230ps, 2.23ps, -44.61dB to -8ps, 1.87ps, -55.52dB, respectively, at 3.2GHz with 1.53mA current and an FoM of -233dB.

11:15 AM

19.6 A 40-to-80MHz Sub-4μW/MHz ULV Cortex-M0 MCU SoC in 28nm FDSOI with Dual-Loop Adaptive Back-Bias Generator for 20μs Wake-Up From Deep Fully Retentive Sleep Mode

D. Bol, UCLouvain, Louvain-la-Neuve, Belgium

In Paper 19.6, Université catholique de Louvain presents a complete ULP MCU SoC in 28nm FDSOI, including an on-chip dual-loop digital adaptive forward-back-bias generator. Unified frequency/back-bias regulation pushes the performance trade-off beyond the state of the art by full exploitation of the FDSOI back-biasing capability with peak efficiency of 280DMIPS/mW at 40DMIPS, an active power consumption down to 3μW/MHz at 48MHz, and an area of 0.7mm^2.

11:45 AM

19.7 A Scalable Pipelined Time-Domain DTW Engine for Time-Series Classification Using Multibit Time Flip-Flops with 140Giga-Cell-Updates/s Throughput

Z. Chen, Northwestern University, Evanston, IL

In Paper 19.7, Northwestern University proposes a scalable pipelined DTW engine in the time-domain for time-series classification using a novel multibit time flip-flop with an area of 1.67mm^2. It achieves the highest reported throughput of 140×10^9 cell-updates/s for DNA applications, leading to a 9× improvement compared to previous work, and can run at 0.7V with only 2.3% increase in error rate.

19.1 Computationally Enabled Total Energy Minimization Under Performance Requirements for a Voltage-Regulated 0.38-to-0.58V Microprocessor in 65nm CMOS

Fahim ur Rahman, Rajesh Pamula, Akshat Boora, Xun Sun, Visvesh Sathe

University of Washington, Seattle, WA

Integrated circuits for ultra-low-power applications strive to minimize total system energy, while satisfying performance requirements. The supply voltage (V_{dd}) can be set to a Minimum Energy Point (MEP) [1,2], where leakage and dynamic energy are suitably balanced. However, controlling operating frequency (f_{clk}), while concurrently tracking a MEP sensitive to PVT and switching activity is not possible. Meanwhile, the traditional approach of locking to the minimum required frequency (f_{targ}), and adjusting V_{dd} to maintain timing slack precludes the possibility of minimum-energy computing. Therefore, there exists a need for a minimum-energy computing architecture that meets performance requirements.

Prior work has demonstrated MEP tracking across PVT and switching activity variation [3]. The approach relies on large capacitors for sample-and-hold operation at sub-threshold frequencies, and cannot account for the significant regulator losses (often amounting to 10%–50% of total energy) necessary for total system energy minimization. Furthermore, clock generation using a free-running oscillator is a requirement, precluding any regulation of f_{clk}.

This paper presents a digital architecture for total system energy minimization subject to performance requirements (Fig. 19.1.1). The design supports two modes of operation – *MEP-lock* and *perf-lock* – and seamless, uninterrupted execution during transitions between them. In *MEP-lock*, the design tunes V_{dd} to first search for and then track the minimum total Energy Per Cycle (tEPC) point inclusive of regulator losses. The system clock is generated by a V_{dd}-powered Critical-path Replica Oscillator (CRO), whose clock period matches the system critical path delay across V_{dd}. In doing so, the CRO automatically adjusts f_{clk} to ensure timing slack. This resulting MEP frequency (f_{MEP}) is sampled and compared against f_{targ}. If $f_{targ} > f_{MEP}$, the system is determined to be performance-constrained and seamlessly transitions to *perf-lock*, a state corresponding to phase-locked operation at f_{targ}.

Although the proposed architecture is applicable across regulator types, we demonstrate it here for a 2:1 switched-capacitor (SC) converter powering a Cortex M0 processor and an FFT accelerator (Fig. 19.1.2). The FFT clock can be selectively gated to modulate switching activity. Both *MEP-* and *perf-lock* rely on elastic clocking provided by the CRO to minimize V_{dd} margins for supply-noise and PVT variation. Phase-lock is achieved during *perf-lock*, despite a V_{dd}-powered CRO, by unifying clock and power regulation into a single loop (UniCaP) [4]. The SC converter is subsumed into the PLL and modulates the CRO phase through V_{dd} control to achieve lock without affecting timing slack. V_{dd} is regulated by the SC converter through its output impedance by modulating its switching capacitance (C_{fly}); the frequency of the converter remains fixed [5].

The total energy delivered by the supply to the regulator and the load over one SC cycle (E_{tot}) can be modeled by a single equation that depends only on V_{in}, C_{fly} and V_{dd} (Fig. 19.1.2). Combining E_{tot} with the number of execution cycles that occur within this SC cycle (f_{clk}/f_{SC}), tEPC can be computationally derived:

$$\text{tEPC} = \frac{E_{tot}}{f_{clk}/f_{SC}} = \frac{E_{tot} f_{SC}}{N_{clk} f_{REFCLK}} = \frac{C_{fly}(V_{in}-2V_{dd})V_{in} \cdot f_{SC}}{N_{clk} f_{REFCLK}}, \quad (1)$$

where N_{clk} is the number of CRO cycles that occur within one REFCLK, and f_{REFCLK} is the PLL reference clock frequency.

Transition into *MEP-lock* initially involves a rapid MEP search using sign-gradient descent, followed by fine grained MEP tracking across PVT and load variation (Fig. 19.1.3). In each search step, the system is briefly operated at the search voltage to determine the V_{dd}-dependent variables, C_{fly} and N_{clk}, needed for tEPC comparison. *MEP-ctrl* uses a DAC to set V_{ref}, relying on closed loop operation to determine C_{fly} as needed to regulate V_{dd} to V_{ref}. In this duration, the PLL frequency divider, which also counts V_{dd}-powered CRO clock cycles during each *REFCLK* cycle, provides N_{clk}. Thus, the *MEP-ctrl* module computes $E_{tot}[n]$ and $N_{clk}[n]$ for search iteration *n*. The module exploits convexity in the *tEPC* search space to perform search in four phases, each with successively finer step-size for rapid search. Finding the MEP requires knowledge of relative (not absolute) values of *tEPC* between successive search points. This observation is exploited to avoid computationally expensive division. By noting that *tEPC* is proportional to the ratio of E_{tot} and N_{clk}, the module instead compares $E_{tot}[n] \cdot N_{clk}[n+1]$ vs. $E_{tot}[n+1] \cdot N_{clk}[n]$. Once the MEP is found, tracking is performed by continuously evaluating tEPC at adjacent V_{dd} points and moving to the lowest tEPC point.

The proposed design was fabricated in a low-power 65nm CMOS process (die photograph in Fig. 19.1.7). Computational MEP tracking incurs a 0.043mm^2 area overhead that is further amortized, either in larger designs or through shared use by multiple voltage domains. The approach also presents a 0.5% energy overhead at a nominal frequency of $0.1 f_{REFCLK}$ which can be further reduced in applications requiring less frequent MEP tracking. All measurements were qualified by correct processor and FFT module operation.

Transitioning between *perf-* and *MEP-lock*, and rapidly searching and tracking the MEP is critical to tEPC minimization under performance requirements. Fig. 19.1.4 shows measured V_{dd} waveforms during mode transitions between *perf-* and *MEP-lock*. The system initially operates in *perf-lock* with V_{dd} controlled to lock the CRO to REFCLK. Transitioning into *MEP-lock* first begins with a MEP search, performed using sign-gradient descent in four phases, each performing successively finer V_{dd} adjustments (step-size). Each phase begins with $V_{dd,MEP}$ obtained from the previous search with the larger step size. After finding the MEP with the smallest step size (5mV), the system tracks the MEP, evaluating relative *tEPC* at excursions of ±5mV. Subsequently as $f_{targ} > f_{MEP}$, the system transitions back to *perf-lock*.

Figure 19.1.5 shows measured V_{dd} waveforms during *MEP-lock* mode, while tracking switching activity changes due to intermittent FFT operation. FFT turn-off reduces dynamic energy dissipation, increasing the MEP voltage. The inset shows the resulting 50mV change in the MEP being tracked by the system after recovering from the initial regulator IR surge. Fig. 19.1.6 demonstrates the quality of the computational MEP tracking approach across multiple test chips and variation in temperature and dynamic loading. The computationally derived MEP was benchmarked against the actual MEP, which was found using direct bench measurement. Direct measurement involved sweeping C_{fly} to adjust V_{dd} (Fig. 19.1.3) with 2mV resolution, recording f_{clk}, and measuring the current drawn from the test supply to determine *tEPC* at each point. The figure shows the resulting *tEPC* curves obtained across V_{dd} for the case of varying dynamic load. The computationally derived MEP was found to be within 5mV of directly measured MEP. A similar approach was used to benchmark tracking performance against temperature variation and under nominal conditions across 20 test chips. Computational MEP was within 5mV of direct measurement, equivalent to the resolution of the finest MEP search step.

Figure 19.1.7 shows a die photograph of the test chip and comparison to prior related work. This work represents a low-overhead digital architecture for minimum energy computing while providing performance guarantees. Incorporating MEP tracking within a unified clock and power framework minimizes wasteful V_{dd} guardbands, and enables seamless transition between durations of minimum energy computing and performance constrained computing. This approach is especially conducive to advanced CMOS nodes.

Acknowledgments:
A Loke, T. Zhang, C. Tokunaga, M. Khbeis for valuable inputs. Funded by SRC (TxACE Task 2712.006) and Qualcomm Technologies Inc.

References:
[1] A. Wang and A. Chandrakasan, "A 180-mV Subthreshold FFT Processor Using a Minimum Energy Design Methodology," *IEEE JSSC*, vol. 40, no. 1, pp. 310-319, 2005.
[2] B. Calhoun, et al., "Modeling and Sizing for Minimum Energy Operation in Subthreshold Circuits," *IEEE JSSC*, pp.1778-1786, 2005.
[3] Y. Ramadass and A. Chandrakasan, "Minimum Energy Tracking Loop with Embedded DC-DC Converter Delivering Voltages Down to 250mV in 65nm CMOS," *ISSCC*, pp. 64–65, 2007.
[4] F. Rahman, et al., "An All-Digital Unified Clock Frequency and Switched-Capacitor Voltage Regulator for Variation Tolerance in a Sub-Threshold ARM Cortex M0 Processor," *IEEE Symp. VLSI Circuits,* pp. 65–66, 2018.
[5] Y. Ramadass, et al., "A Fully-Integrated Switched-Capacitor Step-Down DC-DC Converter with Digital Capacitance Modulation in 45 nm CMOS," *IEEE JSSC*, vol. 45, no. 12, pp. 2557-2565, 2010.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

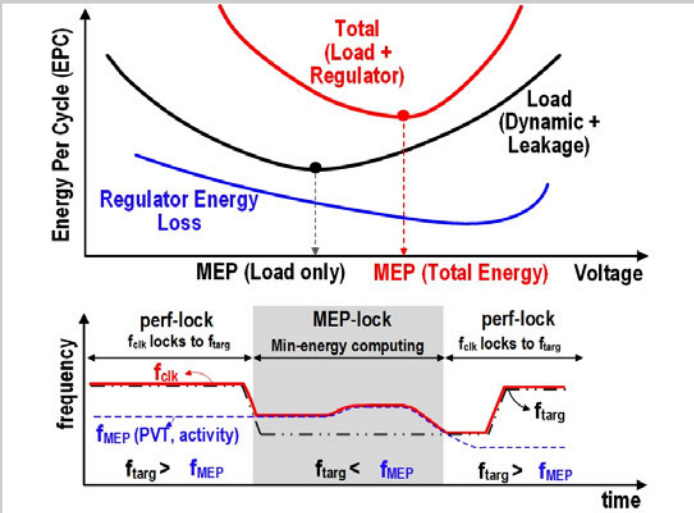


Figure 19.1.1: (Top) The proposed system minimizes total energy dissipation inclusive of VR losses. (Bottom) System-level operation: *perf-lock* mode when $f_{targ}>f_{MEP}$ and seamless transition to *MEP-lock* when $f_{targ}<f_{MEP}$.

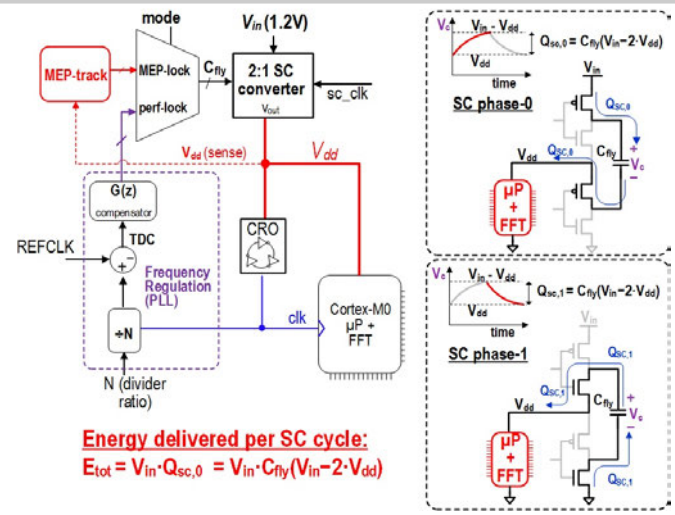


Figure 19.1.2: (Left) Performance-constrained MEP tracking architecture: V_{dd} is controlled by tuning the flying capacitance (C_{fly}). (Right) Variables V_{in}, V_{dd}, and C_{fly} determine E_{tot}.

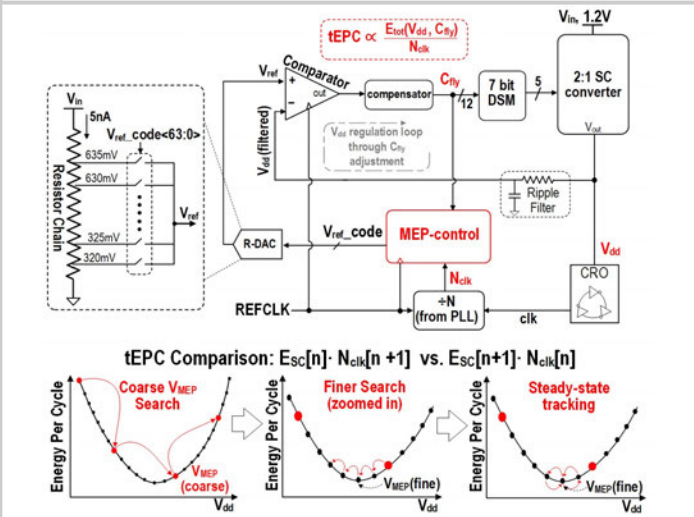


Figure 19.1.3: (Top) Circuit architecture: Closed-loop operation at given search-V_{dd} yields key variables, C_{fly} and N_{clk} for *tEPC* evaluation. (Bottom) MEP search methodology using successively finer search-steps.

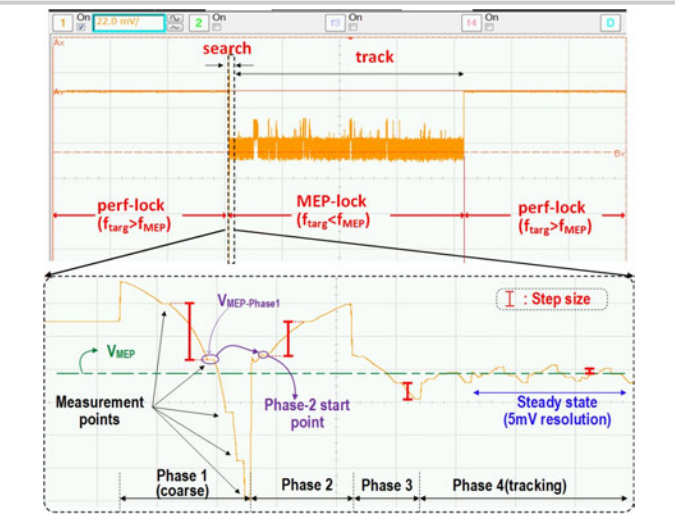


Figure 19.1.4: Measured V_{dd} waveforms during transition between *perf-lock* and *MEP-lock* modes. (Inset) Phases of MEP search, and corresponding V_{dd} transition.

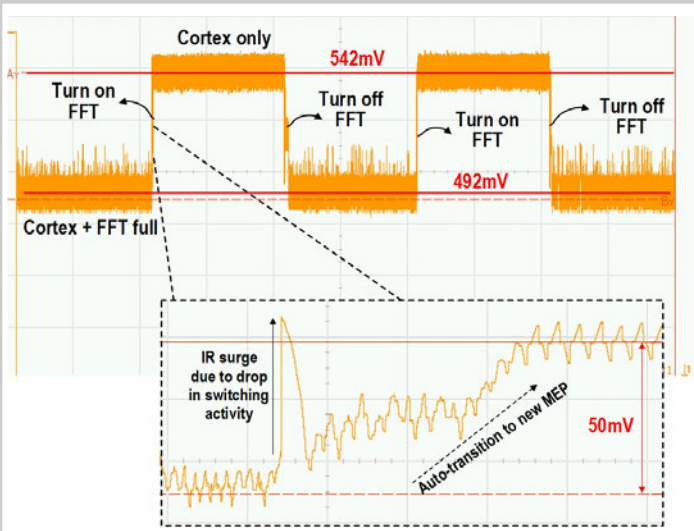


Figure 19.1.5: Measured V_{dd} waveform under MEP-lock during run-time changes in switching activity enabled by FFT on-off operation.

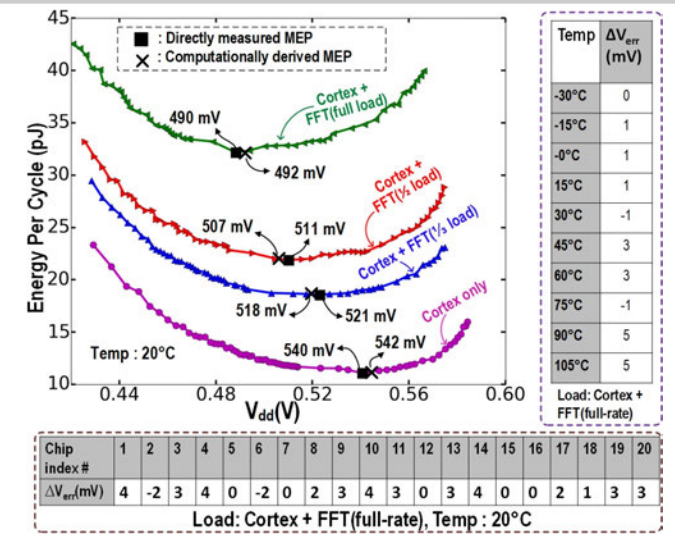
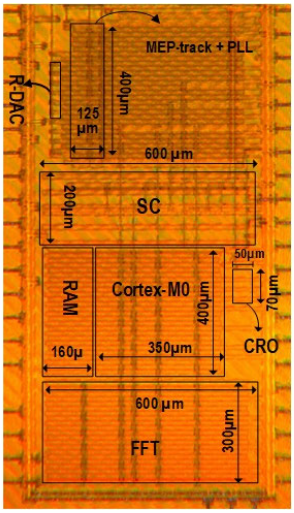


Figure 19.1.6: Computational MEP-tracking accuracy vs. direct MEP measurement across different workloads, temperature and 20 test-chips.

19



	This work	ISSCC '07 [3]
Process	65nm LP CMOS	65nm CMOS
Operating voltage	0.38-0.58	0.25-0.7V
Regulator type	Switched-Capacitor	Buck converter
Load	Cortex + FFT	FIR filter
EPC read-out	Computational	Mixed signal
Temp. tracking	Yes	Yes
Load variation tracking	Yes	Yes
Total energy minimization	Yes	No*
Freq. regulation	Yes	No
Perf.-constrained min-energy	Yes	No
MEP-tracking accuracy	≤ 5mV	Not reported
% Energy reduction (load variation only)	18%	Not reported

* Regulator loss not accounted for. However the method relies on nearly constant converter efficiency across V_{dd}

Figure 19.1.7: (Left) Die-photo of the test-chip and (Right) comparison with related work.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

19.2 A 6.4pJ/Cycle Self-Tuning Cortex-M0 IoT Processor Based on Leakage-Ratio Measurement for Energy-Optimal Operation Across Wide-Range PVT Variation

Jeongsup Lee[1], Yiqun Zhang[1], Qing Dong[1], Wooteak Lim[1], Mehdi Saligane[1], Yejoong Kim[1], Seokhyeon Jeong[1], Jongyup Lim[1], Makoto Yasuda[2], Satoru Miyoshi[3], Masaru Kawaminami[2,3], David Blaauw[1], Dennis Sylvester[1]

[1]University of Michigan, Ann Arbor, MI
[2]Mie Fujitsu Semiconductor Limited, Kuwana, Japan
[3]Fujitsu Electronics America, Inc., Sunnyvale, CA

Wireless sensors for IoT applications have become a prominent computing class and are typically severely power constrained. IoT devices are deployed in a wide range of environments and low power consumption must be guaranteed across a wide temperature range. The combination of dynamic voltage scaling (DVS) and adaptive body biasing (ABB) can achieve minimum total energy per cycle, where dynamic and leakage energy are at an optimal trade-off point [1]. However, due to the dependence of leakage on temperature and workload fluctuations over time, this optimal operating point requires runtime adjustment. Traditional approaches to track an optimal operating point involve repeated measurement of processor power inside a body-bias and supply voltage search loop. However, determining processor power consumption adds significant complexity and/or power overhead since it requires separate measurement of supply voltage and current followed by their multiplication.

In this paper, we make the key observation that at the optimal energy operating point, the ratio of leakage to dynamic power is a nearly constant value across temperature, process variation, and workload. We refer to this as the *optimal leakage ratio*. We derive the optimal leakage ratio value from basic MOSFET characteristics and show that it results in energy per cycle within 4.6% of optimal across PVT conditions. We then show how the leakage ratio can be efficiently and simply measured by counting the modulated frequency of a DC-DC converter and comparing the count obtained at two different clock frequencies. The proposed method was implemented in a Cortex-M0 processor in MIFS 55nm deeply depleted channel (DDC) CMOS with a custom DC-DC converter for supply voltage regulation and charge pumps for body bias generation. A control loop dynamically measures the leakage ratio and processor speed using a critical-path replica and automatically adjusts the body bias and supply voltage. Measurements demonstrate that the design can maintain the optimal ratio across a wide temperature range (–20°C to 125°C) and 5 process corners from split-wafer lots. The processor achieves 6.4pJ/cycle, making it the most energy efficient reported among recent low-power industry-standard processors and operates from 100kHz-6MHz for IoT applications.

Figure 19.2.1 provides a derivation of the optimal leakage ratio in a typical digital circuit. It is well known that total energy shows a minimum at a particular supply voltage (referred to as the minimum energy point, or MEP) due to the growing leakage energy associated with increased delay at reduced voltage. Theoretically, the MEP occurs when the slopes of dynamic and leakage energy become equal and opposite. Dynamic power can be modeled as a simple quadratic with supply voltage (Fig. 19.2.1, Eqn. 2), and we use a power model ($y=aV_{DD}^{-b}$, Eqn. 1) to fit leakage, similar to [1], to facilitate the derivation. Combining these two expressions and minimizing power (setting slopes equal) we find that the ratio of dynamic power to leakage power at MEP is half the parameter value of b, which is 9 in the targeted 55nm technology leading to the P_{leak}/P_{total} ratio of 18%. While this ratio varies with technology node (0.32 in 28nm and 0.10 in 40nm CMOS, simulation) it is nearly constant across temperature and process corners in a particular technology node, as confirmed in Fig. 19.2.1 (top right, simulation), where the 18% ratio yields power within 2% from the minimum.

The overall system architecture is shown in Fig. 19.2.2 and consists of a configurable DC-DC converter that supplies the processor and memory and two bias-voltage generators for compensating leakage across PVT. To tune the processor operation to the optimal leakage ratio (1/4.5) we need to measure both leakage and dynamic power of the processor which we accomplish by monitoring the frequency of the DC-DC converter. The DC-DC converter supplied current is determined by its flying cap, voltage drop (Δh), and frequency (Fig. 19.2.2). At runtime, the flying cap size is constant and Δh is regulated, implying that the DC-DC converter frequency is proportional to its load current (consisting of leakage and dynamic currents). Since dynamic current is proportional to clock frequency, two equations (1), (2) in Fig. 19.2.2 can be derived by modulating processor clock frequency occasionally to measure its leakage/dynamic current ratio. Since the desired optimal leakage ratio is 1/4.5 for this system, the ratio of CNT_{slow}/CNT_{norm} (R_{CNT}/FF_{16}) should be 0.59 (i.e., $96_{16}/FF_{16}$), as derived in Fig. 19.2.2. By tracking this optimal leakage ratio, the system determines if the system is at MEP with very low hardware and power overhead.

Figure 19.2.3 shows a detailed block diagram of the implemented system. The system has two power domains; 1.2V and the DC-DC output power domain. The 1.2V power domain supplies the DC-DC converter, charge pumps, clock generator, and controller. All blocks within this power domain use extremely low leakage (ELL) thick-gate-oxide devices to achieve low leakage power across all corners and temperatures. Due to the large V_{th} of ELL devices, we adopt a new body-controlled transmission gate (Fig. 19.2.3) for the configuration switches in the DC-DC converter to ensure small on-resistance and large off-resistance across all corners and wide temperature range, improving DC-DC converter efficiency. Charge pumps for the processor and SRAM were designed to avoid leakage paths through the power and/or body since the system supports both forward and reverse body bias to enable MEP operation across all process and temperature corners. The Nwell charge pump automatically adjusts its output voltage matching the V_{th} of PMOS and NMOS [2], hence the system is able to compensate for skewed corners as well. Charge pumps for ELL devices were designed for reverse body bias only. To prevent a current path from the body to power or ground when charge pumps are running, 3-transitor switches are used. By default, the charge pumps for the ELL devices are turned off and 3-transistor switches are turned on. At extremely high temperature, such as 125°C and at the FF corner, where the leakage of ELL devices become significant, this charge pump is enabled to limit leakage in ELL devices. A custom 8KB SRAM is implemented to support the Cortex-M0 processor. SRAM word line drivers are boosted using the 1.2V supply. Since the processor and SRAM share the same power domain, the SRAM is designed to match processor speed.

Figure 19.2.4 describes the control algorithm to find the MEP of the system. The target range of R_{CNT} is 90_{16}~98_{16}, within which the operating power is within 4.6% of optimal. For R_{CNT} smaller than 90_{16} the leakage power is too small, indicating that we should apply further forward body bias first before decreasing V_{DD} to prevent the processor from exiting its functional voltage window. A critical-path replica is used to detect how close to the boundary of the functional voltage window the processor is operating. Two thresholds for the number of errors detected by this replica circuit are used to tune the processor close to the maximum operating frequency, while maintaining functionality. Fig. 19.2.4 also includes oscilloscope traces demonstrating the fully automated control loop running on the test chip. The system finds the MEP at 60°C from a default non-optimal starting point. Then, as temperature gradually changes from 60°C to 10°C, the system automatically adjusts its V_{DD} and body bias in the direction of the MEP. As the temperature drops, leakage power reduces and forward body bias is automatically applied to maintain the optimal leakage ratio, while V_{DD} decreases to lower dynamic power. These changes are made while maintaining safe operation in the functional voltage window using the critical-path replica.

Figure 19.2.5 provides measurement results for the implemented system. The design is tested at TT/FF/SS/FS/SF corners, as well as across a wide temperature range from -20°C to 125°C. Various target frequencies are chosen, ranging from 100kHz to 6MHz. The proposed approach achieves power consumption within 4.6% of optimal at 1MHz across all mentioned process and temperatures (Fig. 19.2.5, top, labeled MEP). Fig. 19.2.5 also shows the optimal V_{DD} and body bias values across target frequencies and temperatures. At fixed temperature, optimal V_{DD} is nearly constant, while body bias adjusts as frequency varies (Fig. 19.2.5, bottom right). The Shmoo plot shows R_{CNT} gradually changes according to the V_{DD} and body bias and indicates the optimal value at the selected operating point. The pie chart gives the power breakdown of the system: the DC-DC converter supplies the core and 8KB SRAM for instructions and data, while support circuits such as interfaces, the ELL charge pump, clock, and the bus interface run at constant 1.2V and consume 10.3%. Figure 19.2.6 provides a comparison table to other small IoT processors. The proposed approach offers closed-loop MEP-tracking via clock frequency modulation embedded in the DC-DC converter, and lowest energy per cycle of the listed commercial processors. Figure 19.2.7 provides the die photo.

References:
[1] S. M. Martin, et al., "Combined Dynamic Voltage Scaling and Adaptive Body Biasing for Lower Power Microprocessors under Dynamic Workloads," *IEEE/ACM ICCAD*, pp. 721-725, 2002.
[2] G. Ono, et al., "Threshold-Voltage Balance for Minimum Supply Operation," *IEEE JSSC*, vol. 38, no. 5, pp. 830–833, 2003.
[3] J. Myers, et al., "An 80nW Retention 11.7pJ/Cycle Active Subthreshold ARM Cortex-M0+ Subsystem in 65nm CMOS for WSN Applications," *ISSCC*, pp. 144-145, 2015.
[4] D. Bol et al., "SleepWalker: A 25-MHz 0.4-V Sub-mm^2 7-μW/MHz Microcontroller in 65-nm LP/GP CMOS for Low-Carbon Wireless Sensor Nodes," *IEEE JSSC*, vol. 48, no. 1, pp. 20–32, 2013.
[5] G. Chen, et al., "Millimeter-Scale Nearly Perpetual Sensor System with Stacked Battery and Solar Cells," *ISSCC*, pp. 288-289, 2010.
[6] B. Zhai, et al., "A 2.60pJ/Inst Subthreshold Sensor Processor for Optimal Energy Efficiency," *IEEE Symp. VLSI Circuits*, pp. 154-155, 2006.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

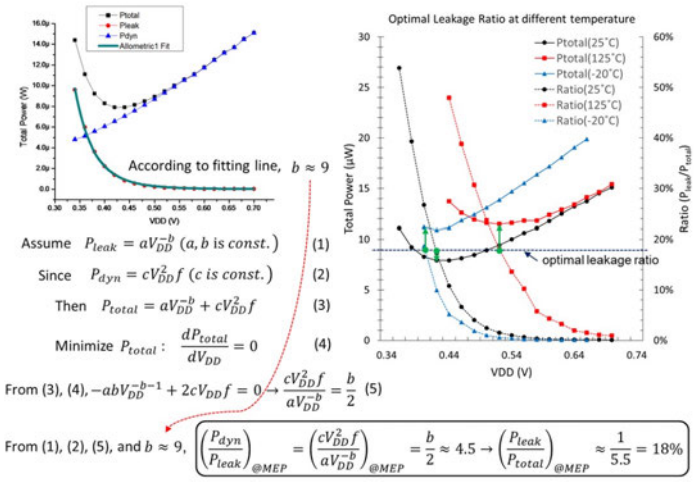


Figure 19.2.1: Derivation of optimal leakage ratio (left); Plot at right shows that total power is within 2% of optimal at derived 18% leakage ratio across temperatures from -20°C to 125°C.

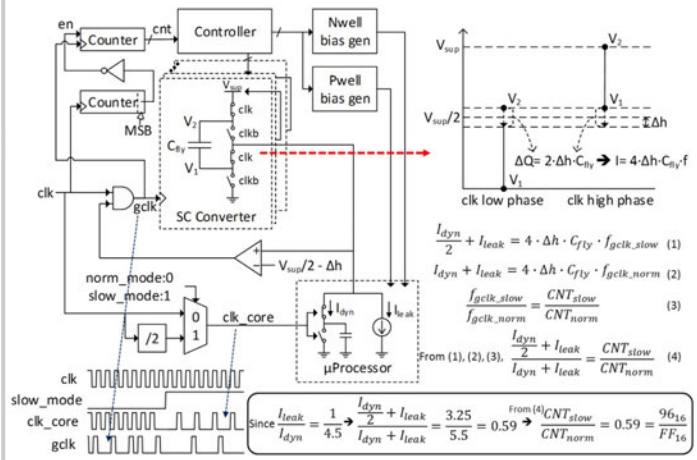


Figure 19.2.2: Proposed approach to measure the ratio of leakage to dynamic power and achieve the desired ratio of CNT_{slow} to CNT_{norm}.

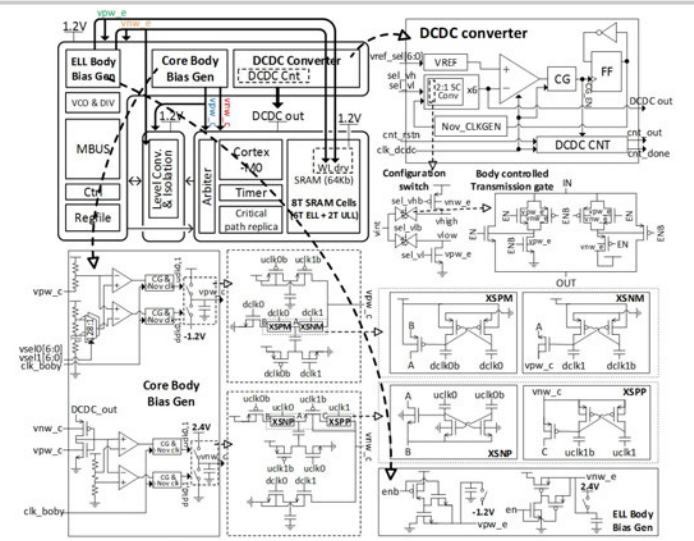


Figure 19.2.3: Top-level block diagram and detailed circuits of DC-DC converter and charge pumps.

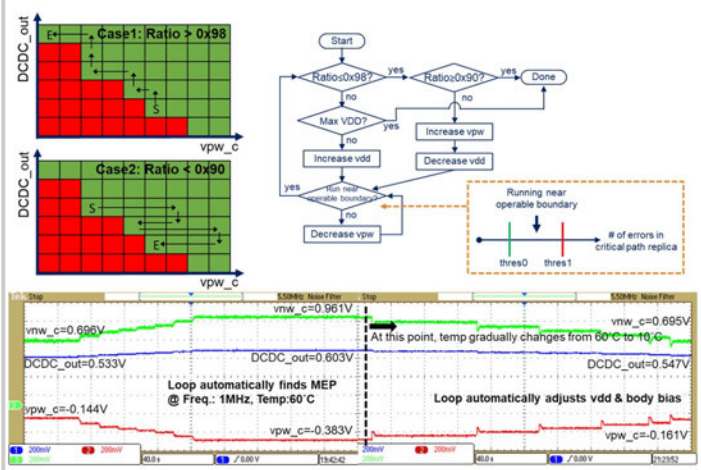


Figure 19.2.4: Algorithm describing system operating loop and measured automatic DC-DC output and body bias adjustment.

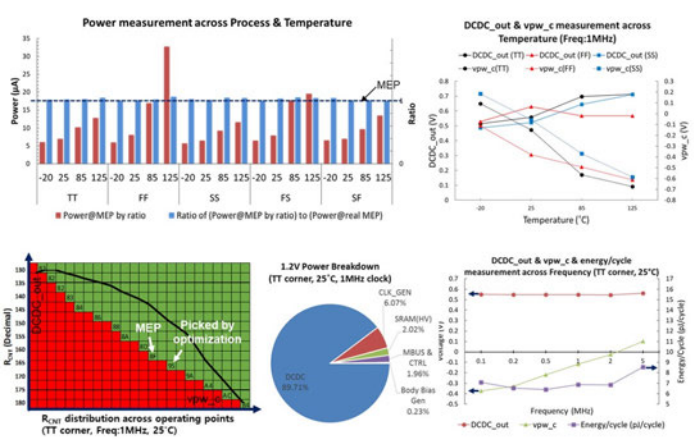
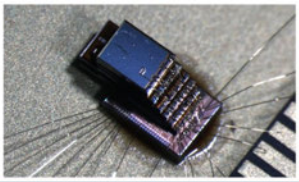


Figure 19.2.5: Measurement results.



	This work	[3] James Myers	[4] David Bol	[5] Gregory Chen	[6] Bo Zhai
Technology	55 nm	65 nm	0.13 μm	0.18 μm	0.13 μm
CPU	ARM Cortex M0	ARM Cortex M0+	MSP430 compatible processor (16b)	ARM Cortex M3	Custom 8-bit ISA
Dynamic power management	On-chip Closed-loop MEP-tracking	On-chip Open-loop Voltage scaling	On-chip Closed-loop Voltage scaling	On-chip Open-loop Voltage scaling	Off-chip Open-loop Voltage scaling
Low Voltage Memory	8Kbyte	8Kbyte	N/A	5Kbyte	256byte
Operating Voltage	0.48V ~ 0.75V	0.19V ~ 1.2V	0.32V ~ 0.48V	0.4V / 0.5V	0.2V ~ 1.2V
Operating Frequency	100kHz ~ 6MHz	29kHz ~ 66MHz	8MHz ~ 71MHz	73kHz / 1MHz	20kHz ~ 10MHz
Minimum Energy Per Operation	6.4pJ/cycle @ 0.55V, 500kHz	11.7 pJ/cycle @ 0.35V, 750kHz	7 pJ/cycle @ 0.375V, 23.6MHz	28.9 pJ/cycle @ 0.4V, 73kHz	2.6 pJ/inst @ 0.36V, 833kHz

Figure 19.2.6: Comparison table and photo of the proposed system.

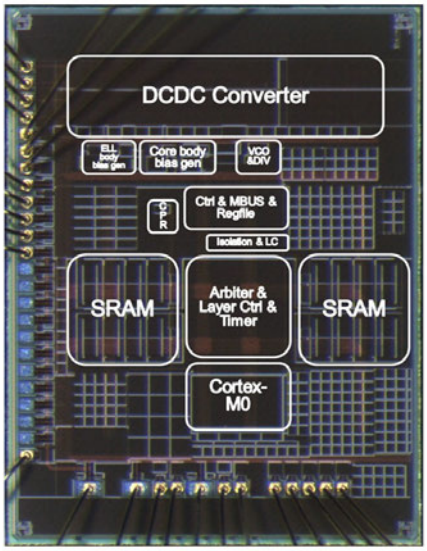


Figure 19.2.7: Die photo in 55nm CMOS.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

19.3 A 7nm All-Digital Unified Voltage and Frequency Regulator Based on a High-Bandwidth 2-Phase Buck Converter with Package Inductors

Francois Atallah[1], Keith Bowman[1], Hoan Nguyen[1], Jihoon Jeong[1], Daniel Yingling[1], Yu Sun[1], Brad Appel[1], Anthony Polomik[1], Mahesh Harinath[1], Joshua Morelli[1], Thomas Moore[1], Nathaniel Reeves[2], Amer Cassier[2], Arijit Raychowdhury[3]

[1]Qualcomm, Raleigh, NC
[2]Qualcomm, San Diego, CA
[3]Georgia Institute of Technology, Atlanta, GA

Conventional processors regulate the supply voltage (V_{DD}) and clock frequency (F_{CLK}) in two separate and independent control loops. A buck converter, switched-capacitor, or low-dropout (LDO) voltage regulator are example control loops for regulating V_{DD} based on a reference voltage (V_{REF}). Processors commonly integrate a phase-locked loop (PLL) to separately regulate F_{CLK} based on a reference clock frequency (F_{REF}), where the F_{CLK} control loop is unaware of the impact of dynamic parameter variations such as V_{DD} droops or temperature changes on the path-timing margin because the PLL voltage-controlled oscillator (VCO) operates on a separate analog voltage. For this reason, conventional processors require either V_{DD} or F_{CLK} guardbands or adaptive and resilient circuits to ensure correct functionality while in the presence of worst-case dynamic parameter variations [1].

Recent work [2-5] combines the regulation of V_{DD} and F_{CLK} in one control loop to reduce the guardbands for V_{DD} and temperature variations, while protecting the processors from path-timing violations. These techniques enable a tight relationship between V_{DD} and F_{CLK}. These previous single V_{DD} and F_{CLK} regulation designs employ either a buck converter [4], an LDO [3], or a switched-capacitor [2], [5] voltage regulator. Today's high-performance SoC processors require the high efficiency of a Buck voltage regulator with a fast-transient response. The design with a Buck converter in [4] uses a large external inductor of 10μH and a large capacitor of 10μF to provide a peak power efficiency of 96.3%, but at the expense of a 20μs settling time for a 90mA load step with a 1ns rise time. Although this design effectively demonstrates high efficiency and guardband reduction with the unified V_{DD} and F_{CLK} control loop, the long settling time during a V_{DD} droop is a major concern for a commercial processor because the V_{DD} droop magnitude may exceed the minimum operating V_{DD} for memory circuits, resulting in failures, and sustained performance degradation of 20μs is unacceptable for some applications. While the design in [4] operates the processor at a faster F_{CLK} after recovering from the V_{DD} droop, this performance loss may exceed the minimum requirements for applications with real-time deadlines. Recent advances of integrating inductors into the package [6] provide an opportunity to significantly improve the transient response for the unified V_{DD} and F_{CLK} control loop during a V_{DD} droop. This paper describes an all-digital unified voltage and frequency buck regulator (UVFBR) with in-package inductors in a 7nm [7] test chip to enable a fast-transient response to V_{DD} droops as required for high-performance processors.

The 7nm test chip (Figs. 19.3.1,19.3.7) contains the UVFBR, performance counters, and a programmable noise generator. The UVFBR design extends the unified voltage and frequency regulator in [3] with a Buck converter, consisting of two in-package inductors. The UVFBR controls both the output clock frequency (F_{OUT}) and output voltage (V_{OUT}) for the digital load in one loop. The UVFBR generates the clock from a tunable-replica circuit (TRC) oscillator and supplies the regulated V_{OUT} to power both the TRC oscillator and the digital load. In the UVFBR control loop, F_{OUT} is divided by a programmable F_{REF} multiplication factor (N) to produce F_{OUT}/N. The UVFBR continuously monitors F_{OUT}/N and adjusts V_{OUT} to lock F_{OUT}/N to a target F_{REF}, resulting in the optimum V_{OUT}. Since V_{OUT} is regulated to achieve a target F_{OUT}, a V_{REF} is not required. F_{OUT} dynamically adapts to V_{OUT} and temperature variations to compensate for delay changes in critical paths to maintain a nearly constant timing margin [1-5]. When a V_{OUT} droop occurs, the TRC oscillator slows down and provides a larger clock cycle time to the load until UVFBR raises V_{OUT} to satisfy the desired F_{OUT}. During a V_{OUT} overshoot, the critical-path delays become faster and compensate for the TRC oscillator speeding up and generating a higher F_{OUT} until UVFBR lowers V_{OUT} back to the nominal value.

The UVFBR includes two 4b Johnson counters with one counter clocked by F_{REF} and the other clocked by F_{OUT}/N. The Johnson counter produces 4 phases of each F_{REF} and F_{OUT}/N. The UVFBR is a 2-phase Buck regulator, with each phase separated by 180°. Each XNOR logic gate produces a signal representing the difference in phase between F_{REF} and F_{OUT}/N to drive the Buck converter output stage to supply the necessary output current load for a target F_{REF}. To avoid phase aliasing associated with the XNOR comparison, the Johnson counter uses an overrun protection (OP) scheme [3]. The OP design holds the Ri value if Li=Ri and propagates the previous stage value (Ri-1) to Ri if Li≠Ri. In addition, the OP holds the Li value if Li≠Ri and propagates the previous stage value (Li-1) to Li if Li=Ri. If F_{OUT} considerably slows down with respect to F_{REF} due to a large load current demand, the phase difference may saturate to 180°, thus increasing the duty cycle to ~100% and maximizing the time for enabling the high side of the output stage. On the contrary, if F_{OUT} speeds up with respect to F_{REF} due to a large V_{OUT} overshoot event, the duty cycle may reduce to ~0%, minimizing the time the high side of the output stage is enabled.

The TRC oscillator enables the interdependent relationship between V_{OUT} and F_{OUT} in the UVFBR. The TRC contains configurable delays to calibrate the oscillator clock period to match the critical-path delay of the digital load (i.e., processor). Each coarse tuning-bit (Coarse_s[3:0]) adjusts the TRC oscillator cycle time by ~30ps, and each small-tuning bit (S[0:12]) adjusts the TRC oscillator by ~4ps resulting in a worst-case timing inaccuracy of ~1.0% at 3.0Ghz. The performance counters allow an on-die measurement of F_{OUT}, which is scanned out of the chip to capture the TRC oscillator frequency. The UVFBR can operate in single-phase mode by either asserting phase1a_select or phase1b_select or in 2-phase mode by asserting both signals. Six programmable noise generators provide the ability to stress UVFBR with variant load currents.

The three-turn inductors (Fig. 19.3.2) are in package and each use six levels of copper metal, where the last three metals constitute the three turns. The simulated DC resistance at room temperature is ~50mΩ and the simulated inductor quality (Q) factor at the UVFBR operating frequencies varies between 17 and 30. Fig. 19.3.3 describes the UVFBR small signal s-domain model. Simulations indicate more than 100° of phase margin at the unity gain bandwidth. While there are two complex poles that sharply drop the phase at high frequencies, these poles are far from the unity gain bandwidth and below -20dB.

The UVFBR is implemented in a 7nm test chip (Fig. 19.3.7). It occupies 6,478μm². To distribute the UVFBR output to the regulated region, the UVFBR components are distributed along the height of the regulated region, thus increasing the effective area to 50,004μm². The area of the package inductors matches the area of the UVFBR regulated region of 639,000μm². Fig. 19.3.4 captures the UVFBR V_{OUT} transition from 0.9V to 0.55V (1.4GHz) in single-phase and 2-phase modes as an example of a dynamic voltage-frequency scaling (DVFS) transition. The settling time is measured at 510ns for the 1-phase mode and 250ns for the 2-phase mode. UVFBR demonstrates a settling time of 60ns and a voltage droop magnitude of 55mV at an F_{OUT} of 2GHz with a current load step from 1mA to 178mA at a rise time of 500ps. This fast-transient response is critical for high performance processors to limit the voltage-droop magnitude and the time while operating at a lower F_{OUT}. Fig. 19.3.5 describes the UVFBR F_{OUT} regulation vs. output load current from 1mA to 900mA with temperature ranging from -15°C to 105°C across 24 dies, demonstrating consistent and highly accurate F_{OUT} regulation for a target F_{OUT} from 1.0GHz to 3.0GHz in steps of 500MHz across a wide range of process, temperature, voltage, and load current. Fig. 19.3.6 plots the measured and simulated power efficiencies for loads between 10mA and 600mA with a 1.6nH inductor, indicating agreement between measurements and simulations. Since the measured peak power efficiency of 60% for a 1.6nH inductor is too low for a processor, simulations indicate that power efficiencies of 90% require an ~10× larger inductor to guide future designs. The small test-chip area allocated and regulated by the UVFBR limits the size of the inductor to 1.6nH and the number of phases to only two. From these simulations, the 1.6nH inductor size limits the measured power efficiency in the UVFFBR implementation. For a high-performance processor, the area is much larger than the test chip area allocated to the UVFBR, therefore allowing for a larger inductor size and/or a greater number of phases to provide acceptable power efficiencies. Fig. 19.3.6 also provides a comparison table with state-of-the-art designs, highlighting the UVFBR fast-transient response.

References:

[1] K. A. Bowman, "Adaptive and Resilient Circuits: A Tutorial on Improving Processor Performance, Energy Efficiency, and Yield via Dynamic Variation Tolerance," *IEEE Solid-State Circuits Magazine*, vol. 10, no. 3, pp. 16-25, 2018.

[2] D. Bol, et al., "SleepWalker: A 25-MHz 0.4-V Sub-mm² 7μW/MHz Microcontroller in 65-nm LP/GP CMOS for Low-Carbon Wireless Sensor Nodes," *IEEE JSSC*, vol. 48, no. 1, pp. 20-32, 2013.

[3] S. Gangopadhyay, et al., "UVFR: A Unified Voltage and Frequency Regulator with 500MHz/0.84V to 100KHz/0.27V Operating Range, 99.4% Current Efficiency and 27% Supply Guardband Reduction", *ESSCIRC*, pp. 321-324, 2016.

[4] X. Sun, et al., "A Combined All-Digital PLL-Buck Slack Regulation System with Autonomous CCM/DCM Transition Control and 82% Average Voltage-Margin Reduction in a 0.6-to-1.0V Cortex-M0 Processor," *ISSCC*, pp. 302-304, 2018.

[5] F. ur Rahman, et al., "An All-Digital Unified Clock Frequency and Switched-Capacitor Voltage Regulator for Variation Tolerance in a Sub-Threshold ARM Cortex M0 Processor," *IEEE Symp. VLSI Circuits*, pp. 65-66, 2018.

[6] E. A. Burton, et al., "FIVR — Fully Integrated Voltage Regulators on 4th Generation Intel® Core™ SoCs," *IEEE Applied Power Electronics Conference and Exposition*, pp. 432-439, 2014.

[7] S.-Y. Wu, et al., "A 7nm CMOS Platform Technology Featuring 4th Generation FinFET Transistors with a 0.027μm² High Density 6-T SRAM Cell for Mobile SoC Applications," *IEDM*, pp. 43-46, 2016.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

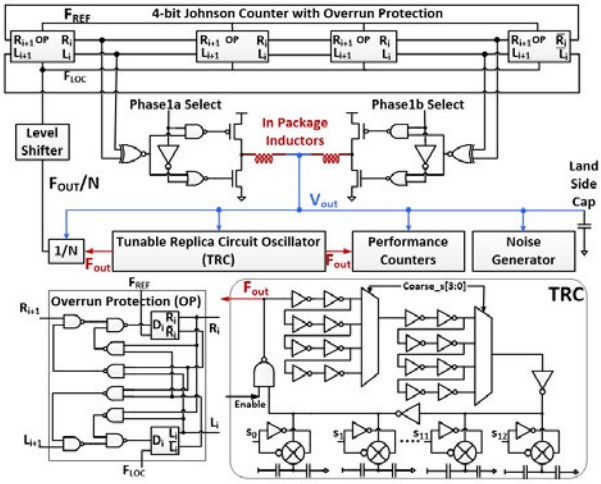


Figure 19.3.1: Test-chip block diagram of the UVFBR with two in-package inductors, performance counters, programmable noise generators, and schematics of the 4b Johnson Counter with overrun protection, and the tunable-replica circuit (TRC) oscillator.

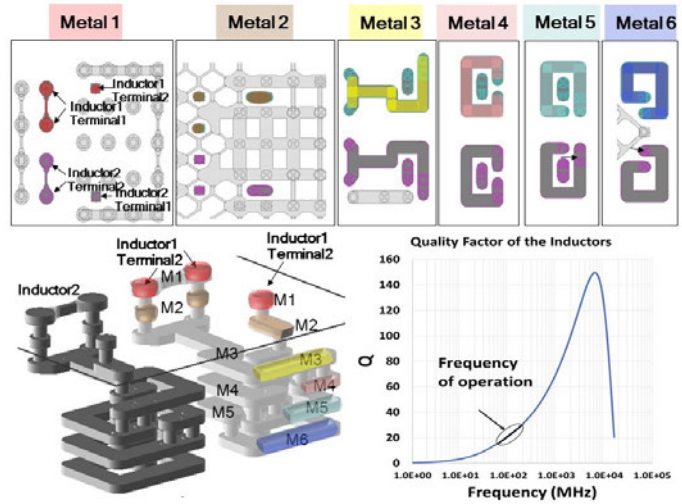


Figure 19.3.2: 3D view of the six metal levels for the two in-package inductors, a detailed view of each metal layer, and the simulated inductor quality (Q) factor. Q ranges from 17 to 30 in the region of operation.

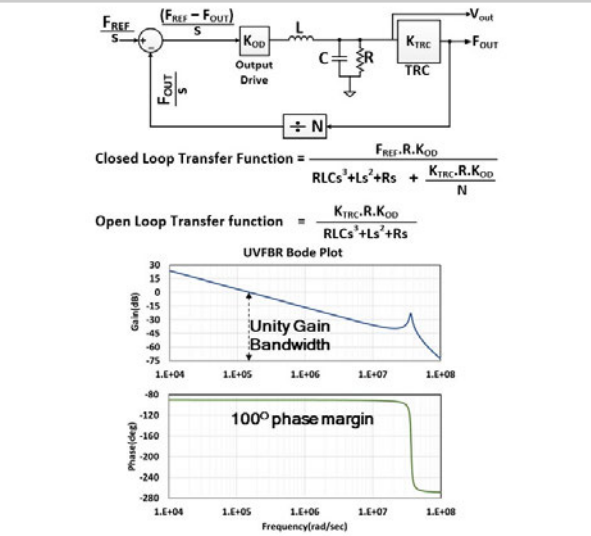


Figure 19.3.3: UVFBR small signal analysis and simulated UVFBR loop gain and phase.

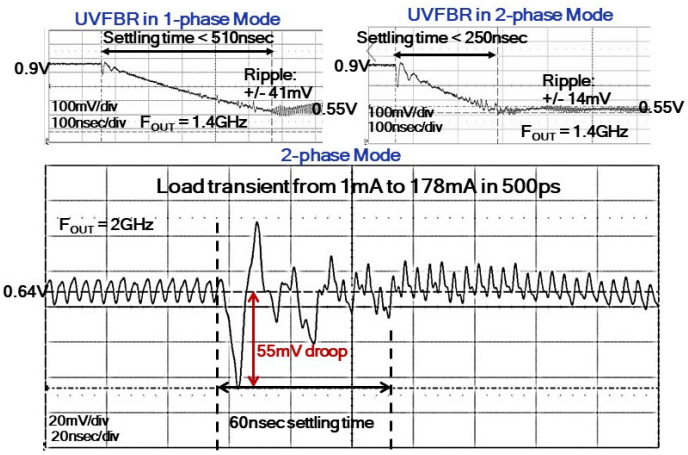


Figure 19.3.4: Measured oscilloscope captures the UVFBR 1-phase and 2-phase modes during a 0.9V to 0.55V (F_{OUT} =1.4GHz) transition and the UVFBR transient response (F_{OUT}=2GHz) in 2-phase with a current load step from 1mA to 178mA at a rise time of 500ps.

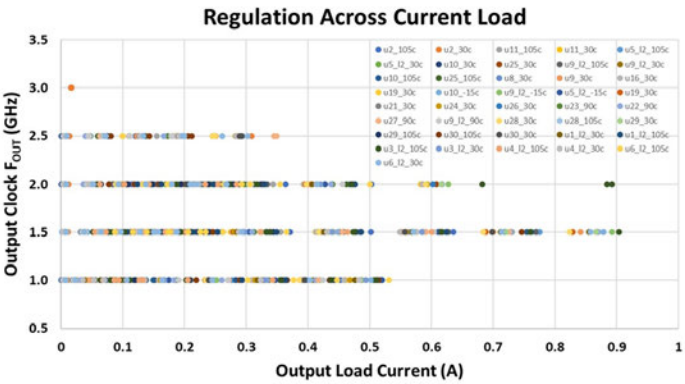
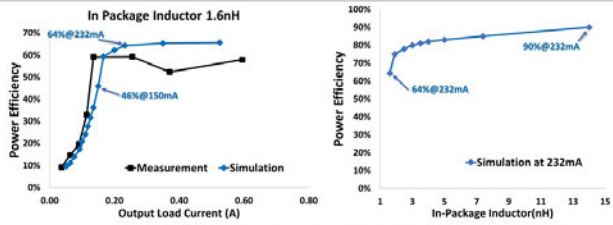


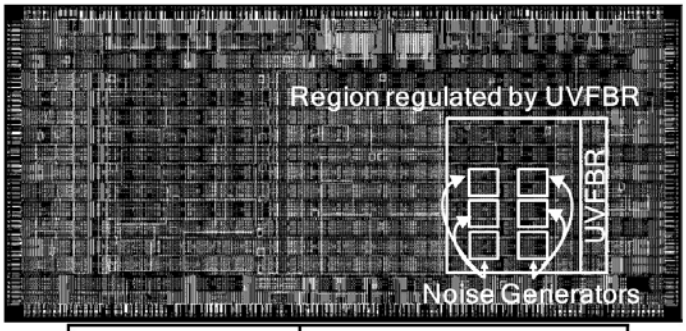
Figure 19.3.5: Measured UVFBR output clock frequency (F_{OUT}) regulation vs. output load current for temperature ranging from -15°C to 105°C across 24 dies.



	[3]	[4]	[5]	This Work
Process Technology	130nm CMOS	65nm	65nm CMOS LP	7nm FinFET CMOS
Input Voltage (V)	0.6 - 1.0	1.2	1.2	1.1-1.3
Topology	LDO	Buck	Switched capacitor	Buck
Output Voltage (V)	0.38 - 0.81	0.6 - 1.0	0.44 - 0.56	0.45 - 0.9
All Digital	Yes	Yes	Yes	Yes
Unified Voltage-Frequency Regulation	Yes	Yes	Yes	Yes
Settling Time (current step)	180ns (3mA/0.5ns)	20,000ns (90mA/ns)	N/A	60ns (177mA/0.5nsec)
Peak Efficiency	99.4% (current efficiency)	96.3%	82.4%	64% with 1.6nH/470nF (measured) 90% with 1nH/470nF (simulated)
I_{out}(mA)	0.1 - 6	1 - 100	N/A	1 - 900
Inductor Technology	N/A	External	N/A	In-Package
Inductor /Capacitor	N/A	10μH/10μF	N/A	1.6nH/470nF

Figure 19.3.6: Measured and simulated UVFBR power efficiency vs. output load current and in-package inductor size and comparison table with other unified voltage-frequency regulation designs.

19



Technology	7nm FinFET CMOS
Chip Area	8mm^2
TRC Area	48μm^2
UVFBR Area	6,478μm^2
UVFBR Distributed Area	50,004μm^2
Region Regulated by UVFBR	639,000μm^2
Area of Noise Generators	32,571μm^2 X 6

Figure 19.3.7: Test-chip die micrograph and characteristics.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

19.4 An Adaptive Clock Management Scheme Exploiting Instruction-Based Dynamic Timing Slack for a General-Purpose Graphics Processor Unit with Deep Pipeline and Out-of-Order Execution

Tianyu Jia, Russ Joseph, Jie Gu

Northwestern University, Evanston, IL

Cycle-by-cycle dynamic timing slack (DTS), which represents extra timing margin from the critical-path timing slack reported by the static timing analysis (STA), has been observed at both program level and instruction level. Conventional dynamic voltage and frequency scaling (DVFS) works at the program level and does not provide adequate frequency-scaling granularity for instruction-level timing management [1]. Razor-based techniques leverage error detection to exploit the DTS on a cycle-by-cycle basis [2]. However, it requires additional error-detection circuits and architecture-level co-design for error recovery [3]. Supply droop-based adaptive clocking was used to reduce timing margin under PVT variation, but does not address the instruction-level timing variation [4]. Recently, instruction-based adaptive clock schemes have been introduced to enhance a CPU's operation [5-6]. For example, instruction types at the execution stage were used to provide timing control for a simple pipeline structure. However, this scheme lacks adequate consideration for other pipeline stages whose timing may not be opcode dependent [5]. In [6], the instruction-execution sequence was evaluated at the compiler level with the timing encoded into the instruction code. The scheme considers all pipeline stages but relies on in-order execution of instructions for proper timing encoding from the compiler.

Several unaddressed critical issues limit the exploitation of DTS into more complex CPU or general-purpose graphics-processor unit (GPGPU) platforms, as shown in Fig. 19.4.1. First of all, all the existing work are based on a simple in-order pipeline structure, where the instruction execution sequence can be predicted ahead of time. For a more complex GPU architecture, there could be multiple vector SIMD units within one compute unit. The instructions issued to each SIMD and other execution units are dynamically scheduled across multiple wavefronts, leading to the difficulties predicting the runtime execution of instructions at the compiler stage. Secondly, existing works are based on shallow pipeline stages with a single core. Fig.19.4.1 shows the simulated dynamic timing slack distribution across 10 pipeline stages of a GPGPU core used in this work [7]. With a deep pipeline and multicore operation, every stage may become a timing bottleneck rendering difficulty in timing management tasks across pipeline stages and requiring clock synchronization across domains, e.g. multi-cores and memory. Thirdly, solutions are lacking when non-execution pipeline stages, e.g. the fetch stage, become the timing bottleneck, which limits the performance benefit by applying DTS techniques.

To overcome the challenges, we present an instruction-driven adaptive clock management scheme with the following features: 1) Use of on-chip critical-path (CP) "messengers", which are different from critical-path monitors, to help predict the appearance of the critical path one cycle earlier for dynamic clock scaling. The combination of real-time CP messengers and multi-stage timing arbiter circuits provide an in-situ cycle-by-cycle clock management. 2) Hierarchical clock circuitry including a global PLL, local delay-locked lock (DLL), and an asynchronous clock interface are used for multi-core clock management across clock domains. 3) An elastic pipeline clocking scheme, developed to mitigate timing bottlenecks within non-execution stages and enhance error-tolerant machine-learning (ML) applications.

Figure 19.4.2 shows the system diagram and the pipeline architecture of the implemented processor, which is a simplified design of an open-source GPGPU architecture following AMD Southern Islands ISA [7]. The chip includes two compute units (CUs), each with 10 pipeline stages including Fetch, Wavepool, Decode, Issue (two stages), Execution (four stages) and Writeback. Inside the execution stage, there are 4 SIMD vector modules with both integer and floating-point modules, and a scalar ALU. The two CUs share an L2 memory bank for the data communication through an asynchronous interface. Fig. 19.4.2 shows the critical-path messengers, which are developed to pessimistically predict the appearance of critical paths in the next clock cycle. Two types of messengers are used. For non-execution stages, e.g. Fetch, Wavepool, etc., the critical paths are traced at the roots of the operation, i.e. D pin of flip-flops. The messenger is formed by detecting transitions on critical signals that will trigger the critical paths, delivering a clock-cycle lead time for the clock controller to react. For the Execution stage, the real-time issued instruction opcodes are used as messengers. The selection of messengers is determined through an internally developed transitional static timing analysis approach, where cycle-by-cycle critical paths are identified at given pipeline conditions. Overall, less than 1% of registers are selected as messengers, leading to negligible area overhead. All the messengers are combined inside a timing arbiter, which issues the final clock period value based on the worst-case timing from all messengers.

Figure 19.4.3 shows the hierarchical clocking scheme enabling instruction-driven clock management. A global PLL delivers the clock to the local DLL for each CU and associated L1 cache. Compared with a PLL, the use of the DLL reduces area overhead needed for clocking each CU. All the phases on the DLL delay chain are equally delayed and carefully matched, with one of them dynamically selected through a glitch-free multiplexer for cycle-by-cycle clock period adjustment. The DLL can be locked with a programmable number of stages, from 30 to 60, providing wide locking range and fine phase selection resolution. At each CU, a timing arbiter combines "messengers" from all pipeline stages and arbitrates the clock period for the current cycle. The timing adjustment step for each messenger is provided by a small register file. For simplicity, only dynamic shrinking of the clock period is allowed. Due to the dynamic clock period of each CU, an asynchronous interface is used to establish data communication between two CUs and L2 caches. A double-triggered data buffer is used to ensure proper latching of the incoming data even with misalignment of clock phases.

As shown in Fig.19.4.4, an elastic clocking mode is created to allow additional timing margin for non-execution stages to mitigate critical timing bottlenecks. Essentially, the clock period for each pipeline stage can be redistributed. Tunable delay modules are inserted at the clock root of each pipeline stage rendering elasticity of timing margin for each stage. As a result, the DLL output clock period remains short due to redistributed pipeline clock. To enhance the performance of the GPGPU for ML applications, e.g. multiplication operations of neural networks, the elastic clocking scheme can apply more aggressive timing for the execution stage where timing errors are tolerable.

Figure 19.4.5 shows the measurement results. The clock for each CU core and selected critical-path messengers were routed out on PCB boards for probing. Measured waveforms confirmed the proper clock management based on the detected CP messengers. Data from RFs and caches were scanned out for verification of proper execution. The instruction-driven adaptive clock scheme has been tested using eight kernel programs, achieving up to 18.2% performance improvement or equivalently 30.4% energy savings using a lower supply voltage. Fig. 19.4.6 shows the additional energy benefit of using the elastic pipeline clocking for ML. The MNIST dataset is tested using floating point SIMD for inference. When allowing timing errors to happen, 4.6% additional energy savings is obtained without accuracy degradation. At 2% accuracy loss, an additional 8% energy saving can be achieved. The scheme was also tested from nominal 1V down to 0.4V, showing similar performance gains. Comparing with prior work [1], this work achieves fine-grained instruction-level clock management for a GPU processor and obtains significant benefits beyond program-level DVFS. Comparing with Razor techniques [2], there are no error detection flip-flops or pipeline recovery needed. The benefit can be maintained down to 0.4V as the improvement does not trade off hold timing. Compared with previous supply droop-based adaptive clocking [4], this work exploits deterministic instruction-level timing variation, and overcomes multicore, deep pipeline, multi-thread execution challenges, faced by previous simple CPU operations [5-6]. For each CU, the timing arbiter and messengers contribute to 1.7% area overhead. 1.6% additional area was introduced by the min-delay padding for the special elastic-clocking mode. The DLL design with elastic-tunable buffers consumes about 1.5% area of the CU. Fig. 19.4.7 shows the die micrograph.

Acknowledgements:
This work was supported in part by the National Science Foundation under grant numbers CCF-1116610 and CCF-1618065.

References:
[1] P. Meinerzhagen, et al., "An Energy-Efficient Graphics Processor Featuring Fine-Grain DVFS with Integrated Voltage Regulators, Execution-Unit Turbo, and Retentive Sleep in 14nm Tri-Gate CMOS," *ISSCC*, pp. 38-39, 2018.
[2] Y. Zhang, et al., "iRazor: 3-Transistor Current-based Error Detection and Correction in an ARM Cortex-R4 Processor," *ISSCC*, pp. 160-161, 2016.
[3] S. Kim, and M. Seok, "Variation-Tolerant, Ultra-Low-Voltage Microprocessor with a Low-Overhead, Within-a-Cycle In-Situ Timing-Error Detection and Correction Technique," *IEEE JSSC*, vol. 50, no. 6, pp. 1478-1490, 2015.
[4] K. Bowman, et al, "A 16nm Auto-Calibrating Dynamically Adaptive Clock Distribution for Maximizing Supply-Voltage-Droop Tolerance Across a Wide Operating Range," *ISSCC*, pp. 152-153, 2015.
[5] J. Constantin, et al., "DynOR: A 32-bit Microprocessor in 28 nm FD-SOI with Cycle-by-Cycle Dynamic Clock Adjustment," *ESSCIRC*, pp. 261-264, 2016.
[6] T. Jia, et al., "An Instruction Driven Adaptive Clock Phase Scaling with Timing Encoding and Online Instruction Calibration for a Low Power Microprocessor," *ESSCIRC*, pp. 94-97, 2018.
[7] R. Balasubramanian, et al., "Enabling GPGPU Low-Level Hardware Explorations with MIAOW: An Open-Source RTL Implementation of a GPGPU," *ACM TACO*, vol. 12, no. 2, article 21, 2015.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

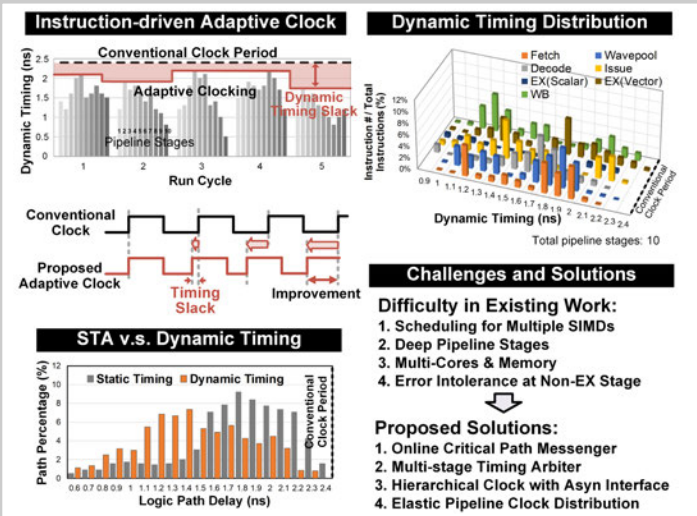


Figure 19.4.1: Simulated dynamic timing slack at instruction level for a 10-stage GPGPU architecture, and the adaptive clock challenges to exploit DTS for the deep pipeline, multicore GPU architectures.

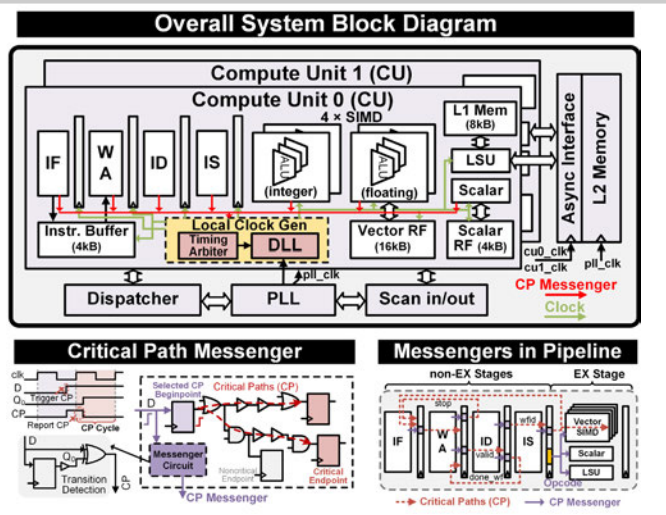


Figure 19.4.2: Block diagram of the implemented open-source GPU processor, and the critical-path messenger design, which predicts that the CP will be exercised one cycle in advance.

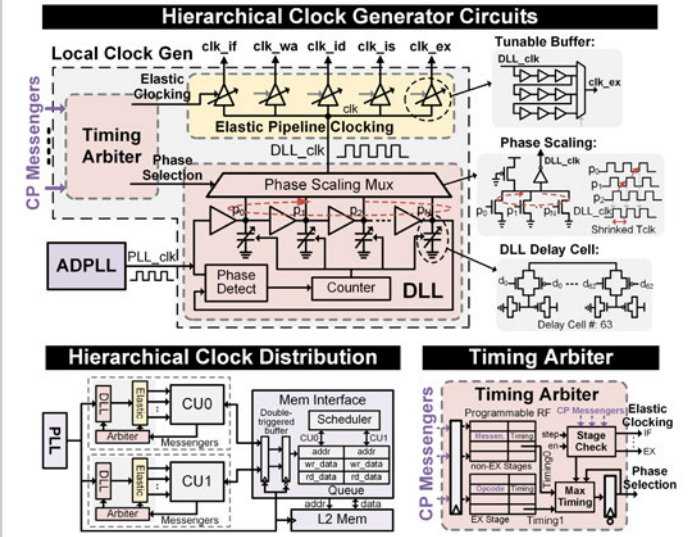


Figure 19.4.3: Design details of the proposed hierarchical clock-management scheme.

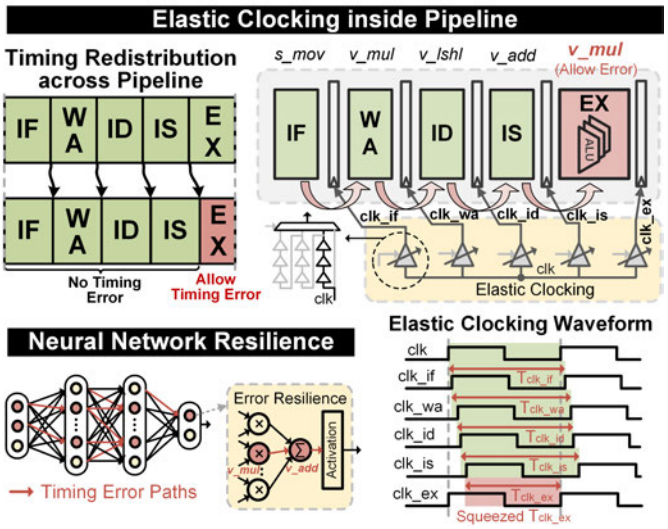


Figure 19.4.4: Elastic pipeline timing distribution across all stages to exploit machine-learning error resilience.

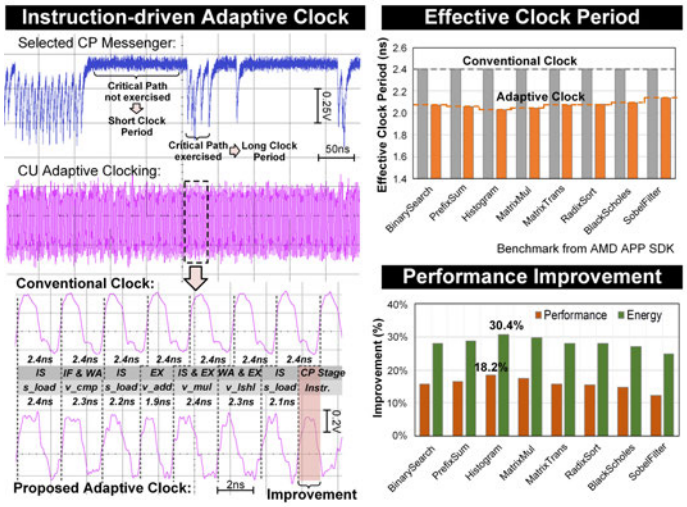
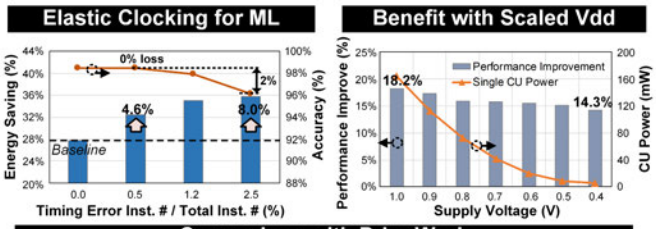


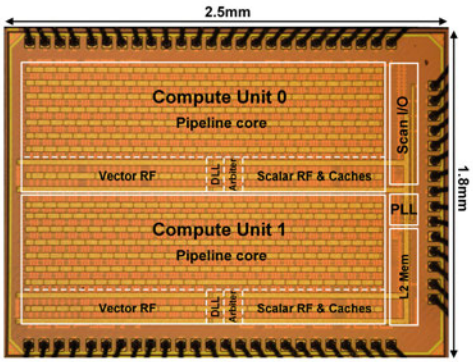
Figure 19.4.5: Measured instruction-driven adaptive clock guided by the messenger signals (waveform distortion due to impedance matching on PCB board) and the performance improvement for eight kernel programs.



Comparison with Prior Work

	[1]	[2]	[5]	[6]	This work
Process	14nm	40nm	28nm	65nm	65nm
Processor	GPU	CPU	CPU	CPU	GPU
Granularity	Program	Instruction	Instruction	Instruction	Instruction
Pipeline Stages	N/A	8	6	6	10
Core Number	18	1	1	1	2
Fsw (MHz)	400	843	509	625	417
Energy Saving	32%	41%	15%	28%	30.4%
Clock/Voltage Management	SCVR+ DLDO	PLL+ Razor	DCO+ LUT	PLL+ Compiler	PLL+DLL+ Messenger
ML Enhanced	No	No	No	No	Yes
Power (mW)	N/A	48@572M	112@1.3G	107@625M	468@417M
Area Overhead	N/A	13.6%	~7.3%	5.4%	1.7% (Arbiter) 1.6% (Elastic Hold)

Figure 19.4.6: Benefit from the elastic pipeline clocking for machine-learning applications, and the improvement with scaled voltage levels.



Technology	65nm CMOS
Die Area	4.5mm^2
CU Power	164mW (pipeline) 52mW (RF & Cache)
PLL Power	6.8mW
Total Power	468mW
Base Freq.	417MHz
Nom Vdd	1V
Vdd Range	0.4-1V

Figure 19.4.7: Die micrograph and design details.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

19.5 Digital Leakage Compensation for a Low-Power and Low-Jitter 0.5-to-5GHz PLL in 10nm FinFET CMOS Technology

Yongping Fan, Bo Xiang, Dan Zhang, James S. Ayers,
Kuan-Yueh James Shen, Andrey Mezhiba

Intel, Hillsboro, OR

As the push for low-power, low-jitter, and low-area phase-locked loop (PLL) design [1] continues in the 10nm node and beyond, the leakage from various sources could become an issue in the applications where the reference clock frequency is low and the static phase error is required to be close to zero at all process corners and temperatures. The impacts of leakage currents on analog PLLs are the increase of static phase error, higher jitter and reference spur. To mitigate the leakage sensitivity in analog PLL designs, digital PLLs have been designed [2-3], but they often exhibit higher power consumption and higher jitter due to quantization noise, and require high-performance voltage regulators or supply-noise cancellation techniques [4] due to their high sensitivity to power-supply noise. To compensate for the leakage current in analog PLLs, analog compensation circuits are proposed [5-6], but it is impossible to achieve zero or low phase error across PVT corners.

In this paper, a digital leakage compensation circuit is proposed to eliminate the static phase error by cancelling the leakage current using pulse width modulation (PWM). Since no current mirror circuits are needed to measure/compensate the leakage current, it is much easier to implement, saves area, and does not add additional leakage current when compared to previous analog solutions. The proposed technique can also be used to cancel the static phase error due to charge pump current mismatch, and makes an analog PLL a viable option as the power reduces to sub-mA levels and process scales down to 7nm and 5nm nodes.

In the proposed digital leakage compensation, voltage pulses are added to integral charge pump inputs, generating current pulses that cancel the leakage charge on the charge pump capacitor. The total pulse width is the same as static phase error of the PLL before the leakage compensation is turned on. For a given leakage current I_{leak}, the static phase error (due to leakage) after the PLL locks is proportional to the net leakage current and can be expressed by $T_e = (I_{leak}/I_{cpi})T_{ref}$, where I_{cpi} is the integral charge pump current, and T_{ref} is the reference clock period. When charge pump current or reference clock frequency is low, the static phase error increases, which reduces timing margin in synchronous clock-domain crossing in high-performance microprocessor design.

Figure 19.5.1 (top) shows the block diagram of the PLL with digital leakage compensation. The time-to-digital converter (TDC) converts the static phase error (after lock) to a 4b digital code, along with a single bit signal, which indicates the relative phase of the reference and feedback clocks. The digital leakage-compensation circuit generates PWM pulses which are used to cancel the leakage. Fig. 19.5.2 (top left) shows the schematic for generating the two PWM compensation pulses: one pulse from a falling edge of the reference clock, and another from a falling edge of (UP+DN). Each PWM pulse is generated by a resettable flip-flop and a digitally controlled delay line, where the delay of each stage can be adjusted (not shown). The PWM pulse width can be adjusted by the 3 control bits. The 4th control bit is used to select the number of pulses (one pulse or two pulses). The PWM pulses are added to UP_I or DN_I according to sign of the static phase error, and then fed to the integral charge pumps and sample-reset loop filter [7] as shown in Fig. 19.5.2 (top right). Fig. 19.5.2 (bottom) also shows the waveforms of the reference clock, feedback clock, UP/DN pulses, and leakage current. Leakage current is assumed to be from the supply to VctI, resulting in a lower VCO frequency and a lagging feedback clock. To cancel the leakage, two PWM compensation pulses are added to DN_I. The DN_I pulses turn on the pull-down legs of the integral charge pump for an additional time period that is equal to the measured static phase error before compensation is turned on, and therefore, neutralize the leakage charge, and align the reference and feedback clocks. Fig. 19.5.1 (middle) shows the simplified TDC schematic used to measure static phase error after the PLL locks and before the digital leakage compensation is turned on. It consists of 16 delay stages, where each delay stage is the same as the one used in the digital leakage compensation circuit. In the case where reference clock is leading (sign=1), DN is passed through 16 delay stages and each delayed version is sampled with UP to generate 16 thermometer bits. To find the static phase error, or average phase error, the phase-error measurements are done in many reference cycles, for example, 128 cycles. The 16 thermometer bits are accumulated in a 16b counter, the bit corresponding to the maximum accumulation represents the magnitude of the static phase error, and are converted to 4b binary code. Fig. 19.5.1 (bottom) shows the flow chart of digital leakage compensation. The phase error can also be measured periodically to ensure leakage is cancelled at all temperatures. Simulation shows that with the digital leakage compensation, ~700ps static phase error is reduced to almost zero, the control voltage ripple and period jitter are reduced by ~10×.

Silicon data on the units with high static phase error shows that, with the proposed digital leakage compensation, the static phase error is reduced from 230ps to -8ps, the integrated jitter is reduced from 2.23ps to 1.87ps, and reference spur is reduced from -44.61dB to -55.52dB as shown in Fig. 19.5.3, when operating at 3.2GHz with a 100MHz reference clock and 1.38mW power consumption at 0.9V supply. When operating at 1.6GHz with a 100MHz reference clock, the static phase error is reduced from 384ps to -17ps, the integrated jitter is reduced from 4.33ps to 3.69ps, and the reference spur is reduced from -50.96dB to -56.82dB as shown in Fig. 19.5.4. Power consumption is 0.59mW at 0.9V supply. Fig. 19.5.5 shows that the static phase error can be corrected to almost 0 with low VCO frequency 0.8GHz (low charge-pump current), and also with a low reference-clock frequency of 19.2MHz. Fig. 19.5.6 summarizes the measured performance and compares it with other state-of-the-art ring-based PLLs. The proposed PLL achieves wide operating range from 0.5-5GHz with very low power 0.29mW (0.5GHz, 0.9V), 2.39mW (5GHz, 0.9V), as well as low integrated jitter 9.3ps (0.5GHz), 1.6ps (5GHz). It shows the figure-of-merit (FOM) of -233dB at a VCO frequency of 3.2GHz. Fig. 19.5.7 shows the die photo of the PLL with digital leakage compensation in a 10nm FinFET technology. The PLL area is 195μm×135μm. The proposed digital leakage compensation circuit consumes only 20μA, and occupies an area of 1600μm^2 (6% of the PLL area). With the proposed digital leakage compensation, an analog PLL remains a viable option with continued future process scaling and power reduction.

Acknowledgements:
The authors acknowledge Syed Feruz Syed Farooq, B. Grossnickle, D. Koach, D. Baker, K. Nguyen, C. Liu, and members of analog design team in Intel Logic Technology Development for their contributions to this work.

References:
[1] K. Shen, et al., "A 0.17-to-3.5mW 0.15-to-5GHz SoC PLL with 15 dB Build-In Supply Noise Rejection and Self-Bandwidth Control in 14 nm CMOS", *ISSCC*, pp. 330-331, 2016.
[2] T. Jang, et al., "A 2.5ps 0.8-to-3.2GHz Bang-Bang Phase- and Frequency-Detector-Based All-Digital PLL with Noise-Self-Adjustment", *ISSCC*, pp.148-149, 2017.
[3] N. August, et al., "A TDC-Less ADPLL with 200-to-3200MHz Range and 3 mW Power Dissipation for Mobile SoC Clocking in 22nm CMOS", *ISSCC,* pp. 246-247, 2012.
[4] C. Yeh, et al., "A 3.2GHz Digital Phase-Locked Loop with Background Supply-Noise Cancellation", *ISSCC*, pp. 332-333, 2016.
[5] J. Chang, et al., "A Phase-Locked Loop With Background Leakage Current Compensation", *IEEE Trans. Circuits and Systems – II. Express Briefs,* vol. 57, no. 9, pp. 666-670, 2010.
[6] K. Kim, et al., "A 32nm and 0.9V CMOS Phase-Locked Loop with Leakage Current and Power Supply Noise Compensation", *Journal of Semiconductor Technology and Science*, vol. 7, no. 1, pp. 11-19, 2007.
[7] J. Maneatis, et al., "Self-Biased High-Bandwidth Low-Jitter 1-to-4096 Multiplier Clock Generator PLL", *IEEE JSSC*, vol. 38, no. 11, pp. 1795-1803, 2003.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

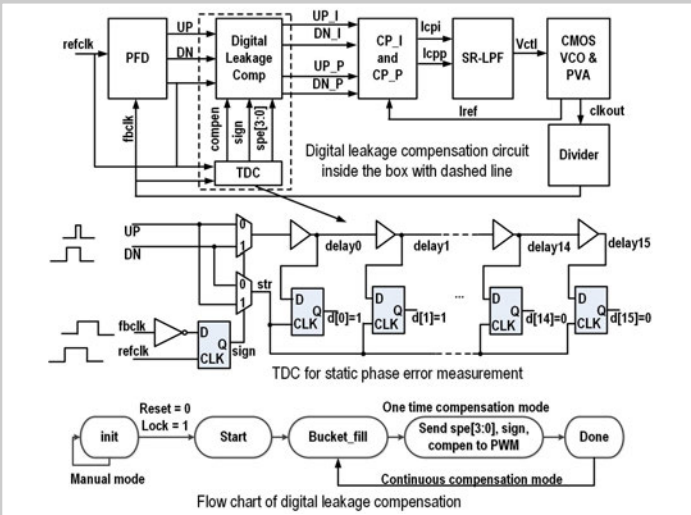


Figure 19.5.1: Block diagram of PLL with digital leakage compensation; TDC for static phase-error measurement and flow chart of digital leakage compensation.

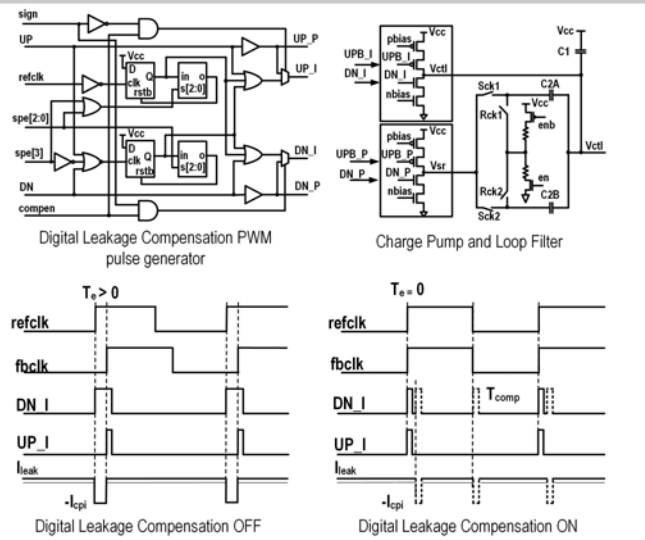


Figure 19.5.2: Schematic of PWM pulse generator and charge pump/loop filter (top); time domain illustration of digital leakage compensation (bottom).

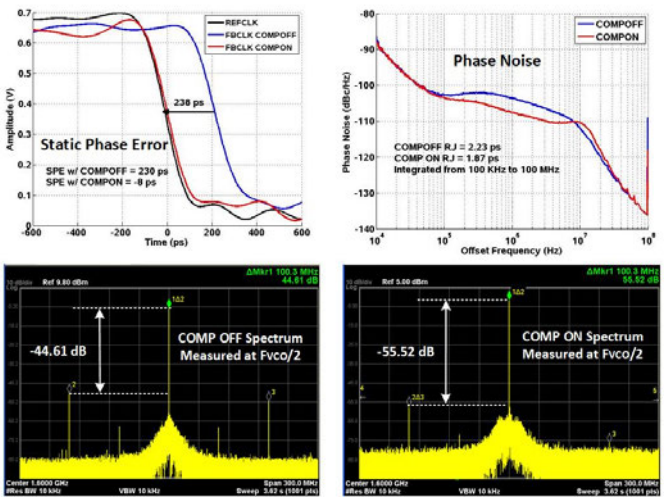


Figure 19.5.3: Measured static phase error, phase noise, reference spur with and without digital leakage compensation at 3.2GHz with 100MHz reference clock.

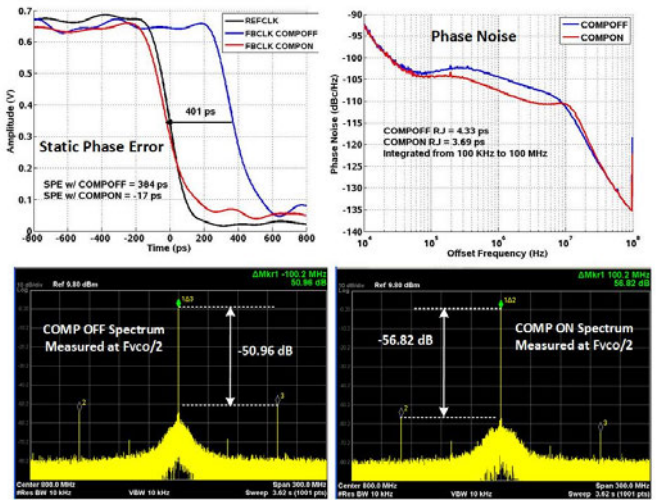


Figure 19.5.4: Measured static phase error, phase noise, reference spur with and without digital leakage compensation at 1.6GHz with 100MHz reference clock.

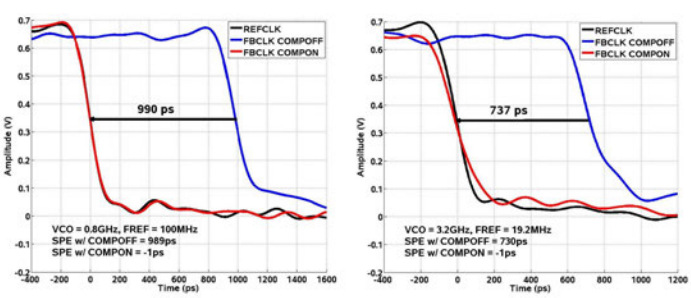


Figure 19.5.5: Static phase-error reduction with digital leakage compensation at low VCO frequency low charge-pump current, and low reference-clock frequency.

	This Work	K. Shen ISSCC'16 [1]	T. Jang ISSCC'17 [2]	N. August ISSCC'12 [3]	C. Yeh ISSCC'16 [4]
Topology	Ring	Ring	Ring	Ring	Ring
PLL	Analog	Analog	Digital	Digital	Digital
Technology	10 nm	14 nm	28 nm SOI	22 nm	40 nm
Core Area (mm^2)	0.026	0.021	0.049	0.017	0.0216
Power Supply (V)	0.9	0.95	-	1	1.1
Reference (MHz)	100	100	50	100	200
Output (GHz)	0.5~5	0.5~5	0.8~3.2	0.3~3.2	3.2
RMS Integrated Jitter (ps)	1.87ps @ 3.2GHz	1.264ps @ 4GHz	2.52ps @ 2.4GHz	3.1	3.54
Power (mW)	1.38	2.56	5	3.4	2.915
FoM (dB)	-233	-233.9	-226.5	-224.9	-224

FoM=10·log((integrated jitter/1s)^2 ·(power/1mW)); lower is better

Figure 19.5.6: Performance summary and comparison to prior works.

19

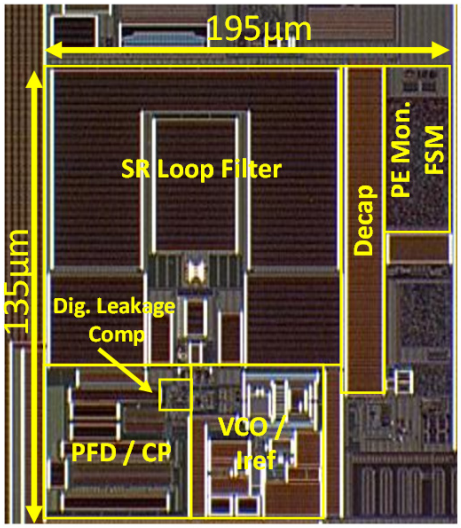


Figure 19.5.7: Die photo of PLL with digital leakage compensation in 10nm FinFET Technology.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

19.6 A 40-to-80MHz Sub-4μW/MHz ULV Cortex-M0 MCU SoC in 28nm FDSOI with Dual-Loop Adaptive Back-Bias Generator for 20μs Wake-Up From Deep Fully Retentive Sleep Mode

David Bol[1], Maxime Schramme[1], Ludovic Moreau[1], Thomas Haine[1], Pengcheng Xu[1], Charlotte Frenkel[1], Rémi Dekimpe[1], François Stas[2], Denis Flandre[1]

[1]UCLouvain, Louvain-la-Neuve, Belgium
[2]e-peas semiconductors, Louvain-la-Neuve, Belgium

Near-threshold circuits operating at ultra-low voltage (ULV) have matured with integration in commercial products, such as ultra-low-power (ULP) MCUs for the IoT [1]. In this market, MCU design faces the key performance tradeoff between speed, active power, deep-sleep retention power and wakeup time, with the challenge of preserving it over PVT corners. We present a ULP MCU SoC in 28nm FDSOI codenamed SleepRunner, exploiting back-biasing (BB) capability of FDSOI to push the performance tradeoff beyond the state-of-the-art.

As shown in Fig. 19.6.1, the MCU logic at 0.4V (V_{DDL}) includes a Cortex-M0 CPU, an FFT accelerator, a wakeup interrupt controller (WIC) and various interfaces. A custom 32KB ULP SRAM [2] is implemented as program memory (PMEM). It uses a custom single-ended 8T bitcell based on negative-differential-resistance (NDR) MOS structures for ultra-low leakage and a 16-bank 512-bitcell/column divided-wordline architecture for ultra-low read access energy at 0.5V (V_{DDS}) of 1.6pJ per 32b word, at the cost of a limited density of 125KB/mm^2. Compared to the PMEM, the Cortex-M0 DMEM contributes less to the active power because of less-frequent accesses. A high-density (HD) 32KB foundry SRAM supplied at 0.8V (V_{DDH}) is thus selected as DMEM to limit die area. The main system clock (MCLK) with register-programmable frequency is generated on-chip from an external 12MHz reference clock (REF_CLK). Embedded power management (ePM) includes switched-capacitor voltage regulators (SCVRs) to generate V_{DDL}, V_{DDS} and V_{DDH} internal supplies from the single 1.8V I/O voltage V_{DDIO}. They support ±5% output over/underdrive to expand the MCU frequency range from 48-72MHz (at nominal output voltage) to 40-80MHz.

To support 40-to-80MHz operation at ULV, LVT devices are used for logic and ULP SRAM. At ULV, upsizing their gate length (L_g) with poly biasing (PB) reduces the logic energy per cycle (E_{cycle}) at the minimum-energy point (MEP) thanks to improved subthreshold swing, lower DIBL and variability [3]. Fig. 19.6.2 shows that a 16nm PB (PB16) yields an E_{cycle} close to 1pJ/cycle for frequencies below 10MHz. Forward BB (FBB) applied to both the logic and ULP SRAM is used in active mode to shift the MEP with the PB16 library to the target frequency range. SCVRs generating the ULV supplies from V_{DDIO} can only achieve high power conversion efficiency (PCE) in a limited output voltage range, which depends on their topology. We thus aim for 0.4V operation because it leads to efficient conversion with the divide-by-4 SCVR topology (V_{DDL}). To reach 72MHz operation at 0.4V (Fig. 19.6.2), we use asymmetric FBB with a stronger PMOS BB (BBP) than NMOS BB (BBN), as PMOS are slower and have a lower back-bias effect, with nominal (i.e. TT corner at 25°C) BBN/BBP of +1V/-2V. However, this high FBB level leads to 30μW logic leakage, which is fine for active mode (30% of the logic power) but prohibitive for deep-sleep mode. Therefore, BBN and BBP are driven by the on-chip FBB drivers to 0V in deep-sleep mode to kill leakage power, while preserving full state retention in both logic and ULP SRAM, which avoids SW initialization at wakeup.

To avoid a conservative V_{DD} guardband for robust operation over PVT corners, we perform unified frequency/back-bias regulation (UFBR) with an adaptive BB generator that dynamically tunes the BB to the current PVT conditions. Moreover, independent BBN and BBP regulation by two distinct loops enables preserving balanced rise/fall transitions in skewed process corners (e.g. slow NMOS/fast PMOS) [4]. This dual-loop UFBR is a key ingredient of the SleepRunner MCU, as its locking time sets the MCU wakeup time and its in-lock power contributes to the MCU active power. Fig. 19.6.3 shows the dual-loop UFBR architecture. BBN is tuned in a frequency-locked loop (FLL) to enable fractional ratios between MCLK and REF_CLK frequencies at low power. A counter senses the frequency of a critical-path replica tunable ring oscillator (TRO) used as both MCLK generator and f_{max} sensor [3,5,6], thus performing UFBR tolerant to V_{dd} droops. The actuation on BBN is based on a PWM-modulated charge pump (CP) with proportional control for fast startup. BBP is tuned in a delay-locked loop (DLL) with digital N/PMOS process imbalance sensors to detect skewed Faster NMOS/Slower PMOS (NFPS) conditions or the opposite (NSPF). A fast inaccurate version is used at startup and a slow accurate one is used when the lock is detected. The actuation is performed by a switched-cap negative CP (as in [5], but with 2 stages to support wide range asymmetric FBB) and a discharge pump (DP), with bang-bang control. Both BBN and BBP drivers are designed with I/O devices for wide-range BBN (0 to 1.8V) and BBP (0 to -3V) to fully compensate for PVT corners. The challenge in the dual-loop design compared to AVS-based unified frequency/voltage regulation (UFVR) [3] is to ensure stable and fast lock despite their inter-dependency. Indeed, the large triple-well cap inside the logic and ULP SRAM tightly couples BBN and BBP nodes, which leads to instability. Fig. 19.6.3 shows that the use of an external decap on BBN ensures stable startup. A 2nF decap (5× the parasitic coupling cap) avoids BBN overshoot, which speeds up the lock and consequently slightly reduces the wakeup energy. Fig. 19.6.3 also shows that this architecture can generate independent asymmetric BBN/BBP voltages to compensate for skewed corners.

For BBP regulation, previous N/PMOS process imbalance sensors were based on analog current comparison, which leads to a hard trade-off between DC power and response time [5,6]. In our BBP DLL, we designed a digital N/PMOS process-imbalance sensor (Fig. 19.6.4) to overcome this challenge, while using only digital standard cells for good correlation to the logic to be calibrated. It is based on two delay lines with selective sensitivity to N/PMOS logic delay. Comparison of their delay provides binary information on the relative NFPS/NSPF conditions. The N/PMOS selective sensitivity is made by alternating strong (×38) and weak (×2) driving cells, such that the low current of the weak cells has to charge the large input cap of the strong cells. Transistor stacking in NAND2/NOR2 with shorted inputs as weak cells further improves the selective sensitivity. However, the weak cells suffer from a high local V_t mismatch, which can lead to strong die-to-die (D2D) variations in the resulting closed-loop BBP voltage and thus significant leakage overhead (Fig. 19.6.4). We limit this effect in the accurate sensor (used in lock) through averaging by increasing the number of delay stages at the cost of slower sensing. The fast and accurate sensor versions can be sampled up to 6 and 3 MS/s for fast startup and loop response, respectively. Bang-bang control in the BBP DLL can lead to spurious CP activity in lock and thus MCU active power overhead. To avoid this, we added a small deadband in the process-imbalance sensor with additional delays in the sensor comparison logic (Fig. 19.6.4). MCU timing closure was performed taking the BBP deadband and variability into account.

SleepRunner was manufactured in 28nm FDSOI (Fig. 19.6.7). Measurement results in Fig. 19.6.5 confirm the FBB capability to preserve the MEP over a frequency range, while reaching 3μW/MHz (at least 2.9× lower than previous MCUs with similar speed and memory). Deep-sleep retention power is dominated by the HD SRAM, which can further be power gated by the WIC at the cost of a SW initialization at wakeup. Compared to [4,6], SleepRunner closed-loop adaptive FBB generation achieves at least 10× faster startup (<20μs) with full clock generation and deep-sleep mode integration. Finally, the comparison in Fig. 19.6.6 shows that the SleepRunner MCU achieves best-in-class power efficiency (280DMIPS/mW), without compromising speed, deep-sleep full retention power nor wakeup time.

Acknowledgments:
This work was supported by the Fonds européen de développement régional (FEDER) and the Wallonia within the « Wallonie-2020.EU » program and the Plan Marshall, as well as by the FRS-FNRS of Belgium.

References:
[1] AmbiqMicro, Apollo2 datasheet, on-line.
[2] T. Haine, et al., "An 8-T ULV SRAM Macro in 28nm FDSOI with 7.4 pW/bit Retention Power and Back-Biased-Scalable Speed/Energy Trade-Off," *IEEE SOI-3D-Subthreshold Microelectronics. Tech. United Conf.*, 2018.
[3] D. Bol, et al., "SleepWalker: 25-MHz 0.4-V Sub-mm² 7-μW/MHz Microcontroller in 65-nm LP/GP CMOS," *IEEE JSSC*, vol. 48, no. 1, pp. 20-32, 2013.
[4] G. de Streel, et al., "SleepTalker: A ULV 802.15.4a IR-UWB Transmitter SoC in 28-nm FDSOI," *IEEE JSSC*, vol. 52, no. 4, pp. 1163-1177, 2017.
[5] M. Blagojević, et al., "A Fast, Flexible, Positive and Negative Adaptive Body-Bias Generator in 28nm FDSOI," *IEEE Symp. VLSI Circuits*, pp. 60-61, 2016.
[6] A. Quellen, et al., "A 2.5μW 0.0067mm² Automatic Back-Biasing Compensation Unit," *ISSCC*, pp. 304-305, 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

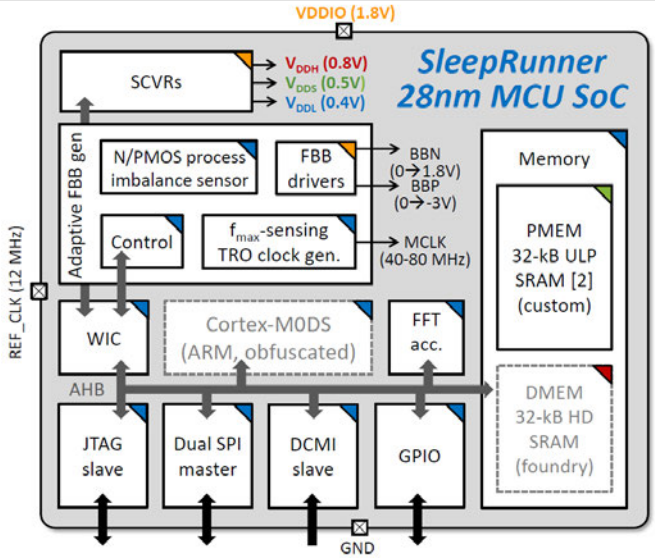
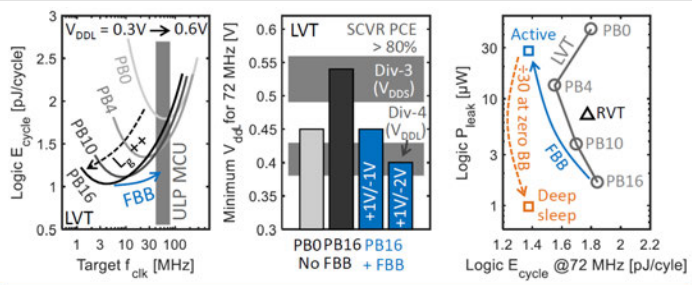


Figure 19.6.1: MCU SoC architecture showing the power domains.



Power mode	System clock MCLK	CPU clock	Logic & ULP SRAM	HD SRAM	SCVR range	Wake-up time
Active	40-80 MHz	40-80 MHz	Adapt. FBB	On	3×140µW	-
CPU sleep	40-80 MHz	Gated	Adapt. FBB	On	3×140µW	1 cycle
Deep sleep (retention: 64 kB + logic incl. CPU)	750 kHz (divided REF_CLK)	Gated	Zero BB	On	3×10µW	< 20 µs

Figure 19.6.2: Optimization of the trade-off for the logic between active energy per cycle (E_{cycle}) and deep-sleep leakage power (P_{leak}) under 72MHz timing constraints (SPICE simulation results at TT corner, 25°C, with timing/power model calibrated from post-layout place/route results).

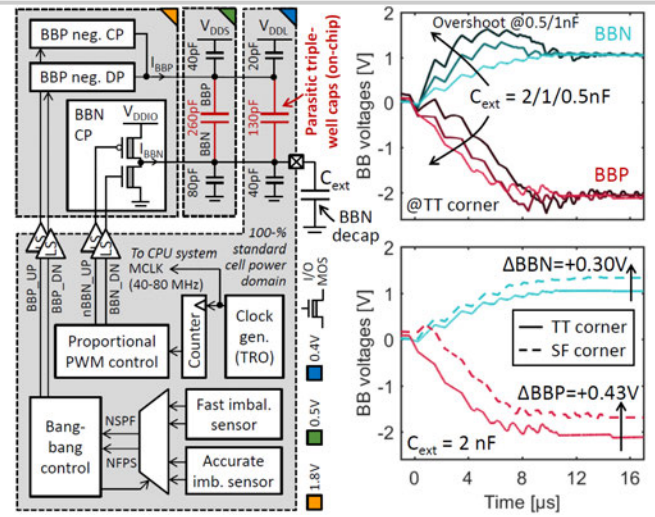


Figure 19.6.3: Dual-loop (FLL for BBN, DLL for BBP) UFBR architecture, startup showing the faster lock with a 2nF external BBN decap and startup response to a skewed process corner (mixed-signal HDL/SPICE simulations at 25°C).

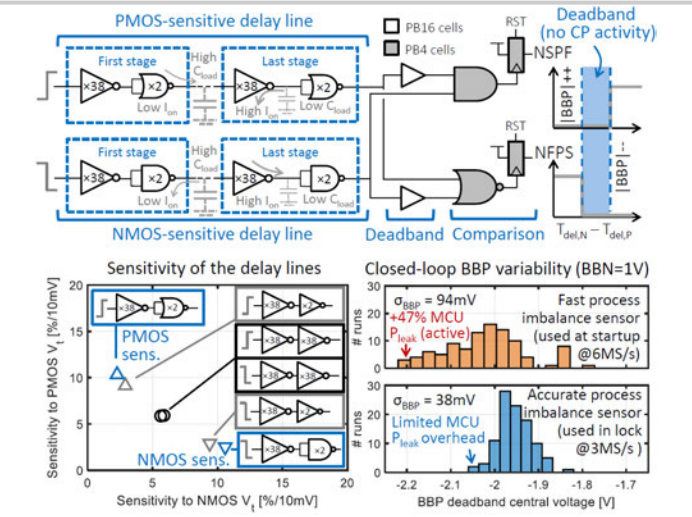


Figure 19.6.4: Schematic of the process imbalance sensor, sensitivity to NMOS/PMOS independent V_t variations and die-to-die variability of the closed-loop BBP voltage (100-run MC SPICE simulation results at TT corner, 25°C).

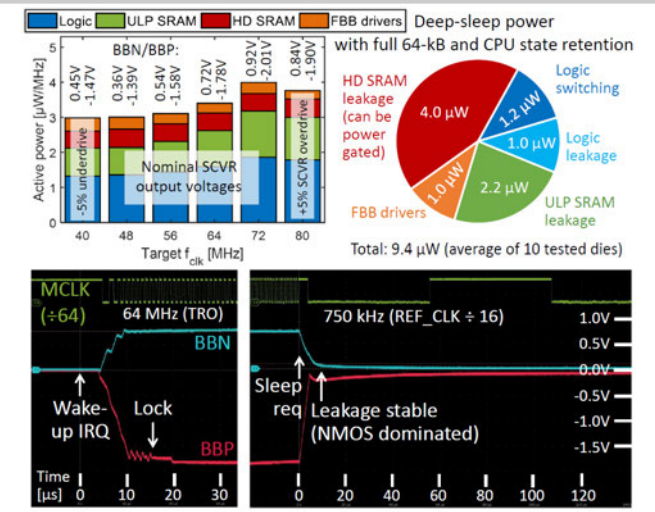


Figure 19.6.5: Power measurement results and power mode transition (adaptive FBB startup/stop). Logic power includes the adaptive FBB control, f_{max}-sensing TRO clock generator and process imbalance sensor.

	Bol, ISSCC, 2012	Myers, ISSCC, 2015	Paul, JSSC, 2017	Lallement, JSSC, 2018	Ambiq, Apollo2, 2018	Uytterhoeven, ESSCIRC, 2018	This work
CMOS technology	65nm LP/GP	65nm LP	14nm FinFET	28nm FDSOI	40nm ULP eFlash	28nm FDSOI	28nm FDSOI
CPU	oMSP430	CM0+	x86 IA	CM0+	CM4F	Zscale	CM0DS
Memory	18kB SRAM	24kB SRAM	72kB SRAM 16kB ROM	4kB SRAM 4kB ROM	256kB SRAM 1MB eFlash	64kB SRAM	64 kB SRAM
Closed-loop PVT compensation	✓	×	×	×	-	×	✓
Embedded power management (ePM)	✓	✓	×	×	✓	×	✓
CPU state retention in deep sleep mode	×	✓	✓ (S1) × (S0)	×	✓ (SS1) × (DS2)	×	✓
Max. frequency [MHz]	71	66	297	150	48	200	80
Deep-sleep retention power† [nW/kB]	95‡	20	-	88‡	914 (SS1) † 40 (DS2)†‡	-	147
Wake-up time	30 µs ‡	N/A	> 1 ms (S1) > 1 s (S0) ‡	-	1.9 µs (SS1) 16.5µs (DS2)‡	-	< 20 µs
Active power at MEP [µW/MHz]	6.1 @25 MHz	11.7 @0.7 MHz	27* @3.5 MHz	2.7 @16 MHz	44.3† @48 MHz	8.8 @22 MHz	3.0 @48 MHz
Peak efficiency [DMIPS/mW]	~ 66* @10 DMIPS	80 @0.65 DMIPS	~ 74* @7 DMIPS	× (not enough memory)	43† @92 DMIPS	114 @22 DMIPS	280 @40 DMIPS

* Computed with the CRO clock generator power. † Includes the ePM power as the numbers without ePM are not available. ‡ Without CPU state retention. * These are estimations assuming 0.4 DMIPS/MHz for oMSP430 and 2 DMIPS/MHz for x86.

Figure 19.6.6: MCU characteristics and measured performance compared to state-of-the-art ULP MCUs. For fair comparison, the power is reported without ePM (thus here excluding the SCVR inefficiency).

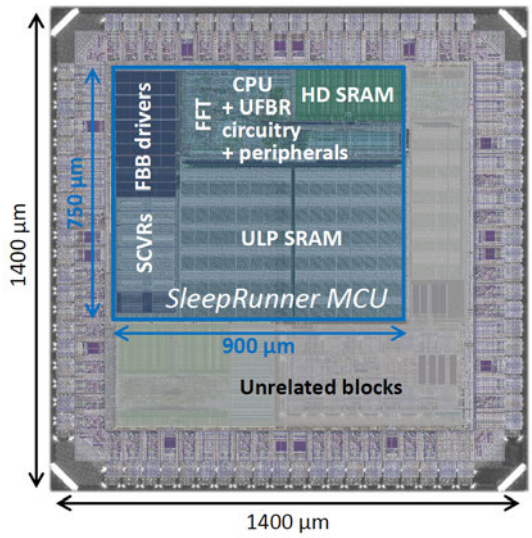


Figure 19.6.7: Microphotograph of the 2mm² SleepRunner SoC die in 28nm FDSOI with superimposed layout view. Active MCU area is below 0.7mm².

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

19.7 A Scalable Pipelined Time-Domain DTW Engine for Time-Series Classification Using Multibit Time Flip-Flops with 140Giga-Cell-Updates/s Throughput

Zhengyu Chen, Jie Gu

Northwestern University, Evanston, IL

Dynamic time warping (DTW), a variant of the dynamic programming algorithm, is widely used for time series classification [1]. Its strong capability for distance measurement for variable-speed temporal sequences makes DTW a popular method for time-series classification in broad applications, such as ECG diagnosis, motion detection, DNA sequencing, etc. [1]. Several efforts have proposed for accelerating the operation of DTW, including a recent demonstration of time-based design in DNA sequencing [2]. However, the demonstration was confined to single-bit operations, a fixed sequence length and low throughput due to non-pipelined operation and a large single-bit delay. To overcome such challenges, this work presents a general-purpose DTW engine for time-series classification using time-domain computing. Pipelined operation is enabled by a time flip-flop (TFF) leading to order-of-magnitude improvements in throughput and a scalable processing capability for time series. Compared with recent time-domain designs, which do not have time-domain memory elements, this work realizes a time-domain pipelined architecture [3].

Figure 19.7.1 shows the basic principle of DTW, which detects similarities among temporal signals having variable speed. As shown in Fig. 19.7.1, for two time series A and B, $D_{i,j}$ can be formulated as the summation of absolute difference $| A_i - B_j |$ and the minimum value of its three ancestor nodes $\min(D_{i-1,j}, D_{i,j-1}, D_{i-1,j-1})$ where A_i and B_j denotes the ith and j^{th} elements of A, B, and $D_{i,j}$ denotes the DTW value at node (i, j). A "warping path" is produced in order to align the two signals in time, as highlighted in Fig. 19.7.1. Time-domain design has shown significant advantages in performing the warping and comparison operation due to its efficiency in key operations, such as minimum (MIN) and absolute (ABS), which only require a few logic gates. The lower side of Fig. 19.7.1 shows the time-domain principle of DTW, where digital inputs are converted into quantized pulse widths and further processed in the time domain. The pulse $T(D_{i,j})$ can be obtained by the accumulation of the pulses of $| A_i - B_j |$ and $\min(D_{i-1,j}, D_{i,j-1}, D_{i-1,j-1})$, both of which are generated by the time-domain MIN module and ABS module. Both accumulation and pipelined operation are realized by the implemented TFF.

Figure 19.7.2 shows the ring-based multi-bit TFF. A 33-stage tristate inverter chain serves as the storage unit. During the reset phase, the rstb signal is used to reset voltages at the internal nodes of the TFF. During the write phase, input pulses are sent to the ring, which allows propagation of "0" through the ring with a duration of input pulses. Multiple input pulses can be repeatedly sent to the TFF and will be accumulated through propagation in the ring. During the readout phase, the stored pulse is sent from the output pin of the ring with pulse width equivalent to summation of the stored values. When the ring is filled, a carry signal rises and the ring will rotate back with remainder values stored inside. The "rotation" operation enables scalable operation in multi-bit groups. Simulations show the TFF can retain data for over 100ns, with less than 0.15b loss due to leakage. In addition, a minimum-pulse-generator circuit is used to create a removable offset to keep the pulse from being too narrow to be propagated. As opposed to conventional flip-flops, each TFF can process multibit signals. In this design, each TFF can store a 6b time-domain signal and two TFFs are used to construct a 10b time-domain value separated into MSB and LSB units, leading to a wide TFF module (WTFF), as shown in Fig. 19.7.4. In the WTFF, once the LSB TFF is full, a carry signal is sent to a pulse generator to generate an extra pulse to be stored in the MSB TFF extending the operation to 10b. Figure 19.7.2 also shows the MIN and ABS modules used in this work, consisting of only simple digital gates, e.g. NAND, rendering a 6× reduction compared with an equivalent digital implementation. A 4b digital-to-time converter (DTC) is implemented inside ABS to convert input digital values into time-domain pulses. The DTC consists of an inverter-based delay chain and multiplexers.

Figure 19.7.3 shows the pipelined DTW engine with 20×20 DTW unit cells and scalable operation to construct longer time series. The DTW matrix contains a group of DTW unit cells with a diagonal pipeline structure. The unit cell, depicted in Fig. 19.7.4, contains 2 WTFF modules, an ABS module, and a MIN module. The second WTFF module (marked in white) in the unit cell is used to copy the data from last pipeline stage, because the data stored in node $(i-1, j-1)$ is one pipeline stage earlier than the nodes $(i-1, j)$ and $(i, j-1)$. Due to the use of the TFF, in every clock cycle, the data pulses are propagated along the diagonal direction of the matrix. Fig. 19.7.3 also shows the data streaming flow for pipelined operation, where incoming digital data is fed from an on-chip register file (RF) and clock management unit. Each data item is piped through the DTW matrix as inputs to ABS modules both vertically and horizontally. The pipelined operation allows fixed dimensions of the DTW engine to be unfolded for longer data sequences, as shown in Fig. 19.7.3, ultimately limited by internal register storage capacity – 10b in this implementation. All output pulses from the bottom and right boundaries are decoded by shared time-to-digital converters (TDCs) every clock cycle and re-sent back for processing by subsequent sections. To speed up the operation for a simple data sequence, e.g. a DNA sequence, a non-pipelined mode is also possible by bypassing the TFF modules and allowing signal edges to directly propagate through the matrix. In addition, a 2b tunable delay cell is implemented in each unit cell to tune the output pulse width, compensating for process variations. A special DTW matrix calibration scheme is introduced to calibrate the DTW matrix. With calibration, the maximum DTW distance error drops from 5b to 1.5b.

Figure 19.7.4 shows the test chip implementation of the DTW engine in a 65nm 1V CMOS process. Two sets of TDCs, based on Vernier delay chains, are placed at the right and bottom sides to decode time-domain signals at the boundaries. A single-bit resolution of 40ps is used in the DTW design, while a resolution of 20ps is used in the TDC to reduce quantization errors at the boundary of operation. All the input and output data can be scanned in and out through a scan chain for verification.

Figure 19.7.5 shows measurement results. UCR time-series classification databases were used to test the architecture [4]. Five databases from four typical applications were selected. The DTW engine is configured in unfolded mode to adapt to the variable-input series length. The measured error rate for classification by the DTW engine is only 1.5% higher than ideal DTW operation (floating point results in software) mainly due to quantization (0.5%) and process variation (1%). 100 sets of DNA sequence data from the human genome database (GDB) were also tested for comparison between ideal DTW operation and measurement results [2]. The measured distance closely tracks the ideal results, having an error within 2.6%. The measured waveform (Fig. 19.7.6) confirms the expected output pulse at a frequency of 110MHz in pipelined mode and 3.1ns processing time in DNA-sequencing mode. The linearity of the implemented TFF is measured through an on-chip TDC showing the proper storage of a time signal within 0.5b under the quantization limitation of the TDC. The chip was also verified at different supply voltages in pipelined mode down to 0.7V, with a 2.3% increase in error rate compared with ideal DTW operation on the UCR database. Fig. 19.7.6 shows the comparison with prior work. Throughput of 140 giga-cell-updates-per-second (GCUPS) for DNA sequencing is achieved with 9× improvement over previous work [2]. A significantly higher throughput per area (GCUPS/mm^2) is observed compared with prior CPU, GPU and ASIC implementations. Compared with non-pipelined operation, the pipelined design shows 7× improvement in throughput for general DTW applications. In addition, the use of time flip-flops enables a pipelined architecture for time-domain design [3]. The die micrograph is shown in Fig. 19.7.7.

References:

[1] Hui Ding, et al., "Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures," *Proceedings of VLDB*, pp. 1287-1300, 2008.

[2] Advait Madhavan, et al., "A 4-mm2 180-nm-CMOS 15-Giga-Cell-Updates-per-Second DNA Sequence Alignment Engine Based on Asynchronous Race Conditions," *CICC*, 2017.

[3] A. Amravati, et al., "A 55nm Time-Domain Mixed-Signal Neuromorphic Accelerator with Stochastic Synapses and Embedded Reinforcement Learning for Autonomous Micro-Robots," *ISSCC*, pp.124-126, 2018.

[4] UCR Archive, http://www.cs.ucr.edu/~eamonn/time_series_data

[5] M. Farrar, "Striped Smith–Waterman Speeds Database Searches Six Times Over Other SIMD Implementations," *Bioinformatics,* vol. 23, no. 2, pp. 156-161, 2007.

[6] Y. Liu, A. Wirawan, and B. Schmidt, "Cudasw++ 3.0: Accelerating Smith-Waterman Protein Database Search by Coupling CPU and GPU SIMD Instructions," *BMC Bioinformatics*, vol. 14, no. 1, 2013.

[7] N. Neves, et al., "Multicore SIMD ASIP for Next-Generation Sequencing and Alignment Biochip Platforms," *IEEE Trans. VLSI*, vol 23, no. 7, pp. 1287-1300, 2015.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

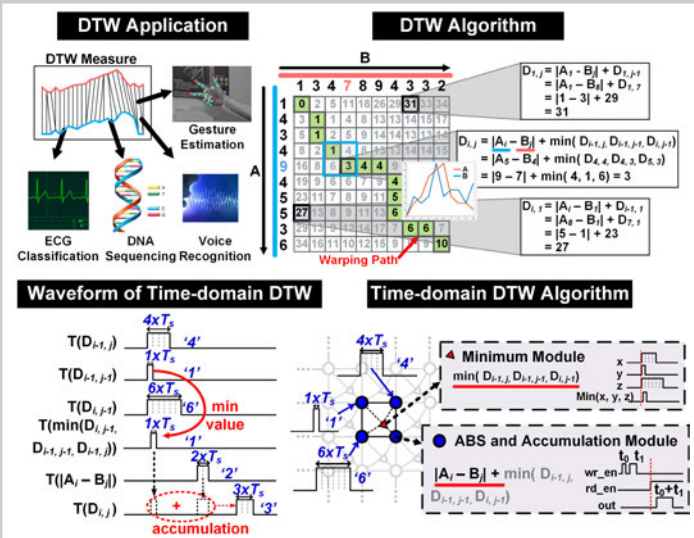


Figure 19.7.1: Application of DTW; concept of DTW algorithm; and, time-domain DTW implementation.

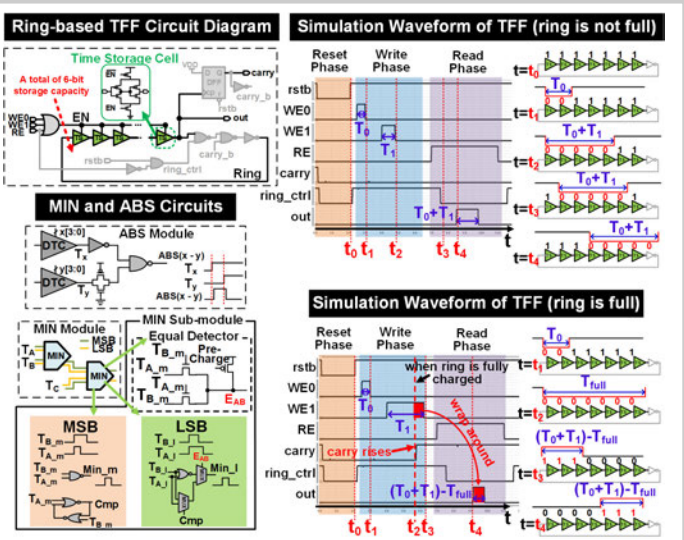


Figure 19.7.2: Circuit diagram and proposed ring-based time flip-flop, and other time-domain circuits.

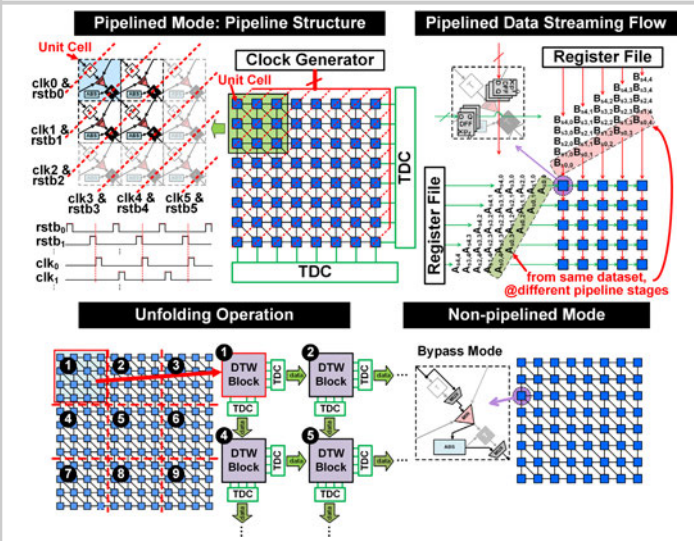


Figure 19.7.3: Pipelined mode of DTW engine; pipelined data streaming flow, unfolding for long sequences; and, non-pipelined mode of DTW engine.

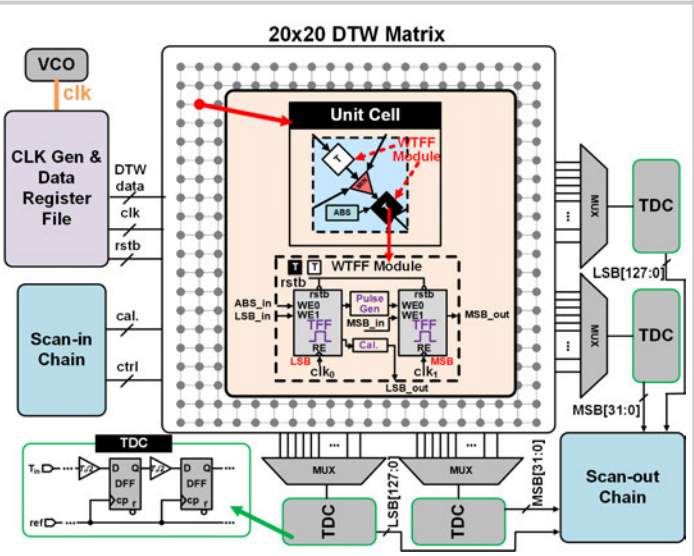


Figure 19.7.4: Overall test chip block diagram.

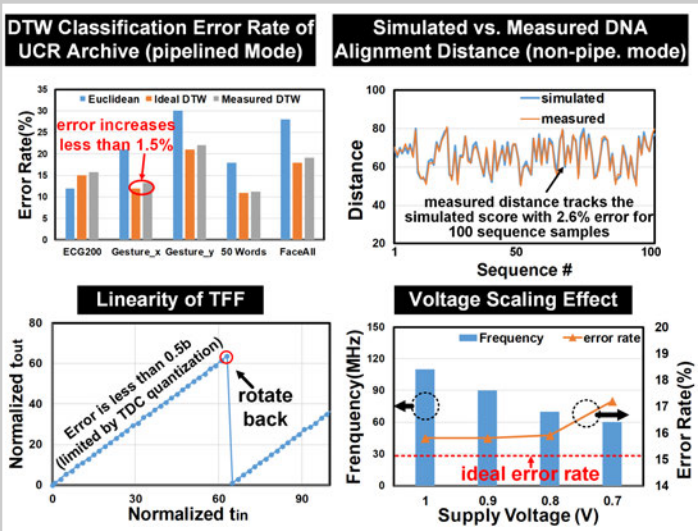
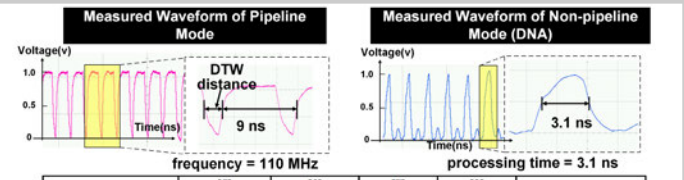


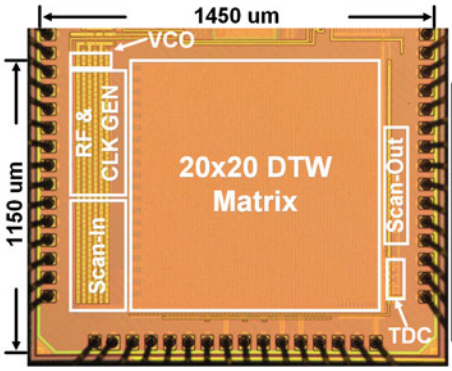
Figure 19.7.5: Measurement results of DTW classification databases, DNA sequencing; linearity of TFF; and, voltage scaling impact on the DTW engine.



	[5] Bioinformatics 2007	[6] Bioinformatics 2013	[7] TVLSI 2015	[2] CICC 2017	This work
Architecture	CPU	GPU	ASIC/ CPU	Time-domain ASIC	Time-domain ASIC
Process (nm)	65	28	90	180	65
Area (mm²)	143	300	6.4	4	1.67
Number of bit	floating point	floating point	32	1	4 (input) 10 (internal)
Power (mW)	9.5×10^4	2×10^5	2732	70	136 (pipeline) 35 (non-pipe.)
Clock period (GHz)	2	1	0.6	0.01	0.11 (pipeline)
Throughput for DNA sequencing (GCUPS)	3	119	9	15	140 *
Throughput per Area (GCUPS/mm²)	0.02	0.4	1.4	3.75	84
Throughput for general DTW App. (GCUPS)	-	-	-	-	71 (pipeline) 10 (non-pipe)
Error rate	-	-	-	2.9%	1.5~2.6%

* In DNA application, single bit non-pipeline mode is utilized for fair comparison with prior work.

Figure 19.7.6: Measurement waveform and performance comparison with prior work.



Technology			65nm
Die Area			1.67 mm^2
Frequency			110 MHz
Nom Vdd			1V
Power	DTW Core	Pipelined	111 mW
		Non-pipe.	20 mW
	RF& CLK GEN	Pipelined	25 mW
		Non-pipe.	15 mW

Figure 19.7.7: Die micrograph and specifications.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 20 Overview: *Noise-Shaped and VCO-Based ADCs*

DATA CONVERTERS SUBCOMMITTEE

Session Chair: ***Yun-Shiang Shu***
Mediatek, Hsinchu, Taiwan

Associate Chair: ***Bob Verbruggen***
Xilinx, Saggart, Ireland

Subcommittee Chair: Michael Flynn, University of Michigan, Ann Arbor, MI

Noise-shaping and VCO-based ADCs have been an area of continued interest in the data converter community. This session features 7 designs using these techniques, with bandwidths ranging from 25MHz to 2.5GHz and SNDRs from 45.2dB to 76.6dB, highlighting the wide applicability of these techniques. This session also marks the first 7nm ADC published at ISSCC.

8:30 AM

20.1 A 5GS/s 7.2 ENOB Time-Interleaved VCO-Based ADC Achieving 30.5fJ/conv-step
M. Baert, KU Leuven, Heverlee, Belgium
In Paper 20.1, KU Leuven describes a 28nm time-interleaved VCO-based ADC. It achieves 45.2dB SNDR with a Nyquist input at 5GS/s, while consuming 22.7mW.

9:00 AM

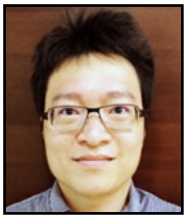
20.2 A 40MHz-BW 320MS/s Passive Noise-Shaping SAR ADC with Passive Signal-Residue Summation in 14nm FinFET
Y-Z. Lin, MediaTek, Hsinchu, Taiwan
In Paper 20.2, MediaTek demonstrates a technique for passive signal-residue summation in a noise-shaping SAR ADC in 14nm FinFET technology. It achieves 66.6dB SNDR in a 40MHz bandwidth consuming 1.25mW of power.

9:30 AM

20.3 A 50MHz-Bandwidth 70.4dB-SNDR Calibration-Free Time-Interleaved 4th-Order Noise-Shaping SAR ADC
L. Jie, University of Michigan, Ann Arbor, MI
In Paper 20.3, the University of Michigan shows how time-interleaving can be combined with noise-shaping for the first time-interleaved noise-shaping SAR ADC. Implemented in 40nm CMOS, the design achieves 70.4dB SNDR in a 50MHz bandwidth while sampling at 400MS/s and consuming 13mW.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

10:15 AM



20.4 An 8×-OSR 25MHz-BW 79.4dB/74dB DR/SNDR CT ΔΣ Modulator Using 7b Linearized Segmented DACs with Digital Noise-Coupling-Compensation Filter in 7nm FinFET CMOS

T-Y. Lo, MediaTek, Hsinchu, Taiwan

In Paper 20.4, MediaTek shows the first 7nm ADC at ISSCC, using a linearized DAC and digital noise coupling compensation. The design achieves 74dB SNDR in a 25MHz bandwidth while consuming only 3.8mW.

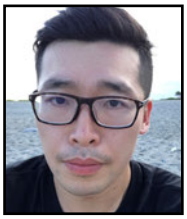
10:45 AM

20.5 A 76.6dB-SNDR 50MHz-BW 29.2mW Noise-Coupling-Assisted CT Sturdy MASH ΔΣ Modulator with 1.5b/4b Quantizers in 28nm CMOS

L. Qi, University of Macau, Macau, China

In Paper 20.5, the University of Macau demonstrates a 4th-order sturdy-MASH DSM using a noise-coupling technique with a zeroth-order second stage. The design achieves 76.6dB SNDR in a 50MHz bandwidth consuming 29.2mW of power, without requiring any DAC calibration or error shaping.

11:15 AM

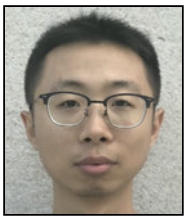


20.6 An 80MHz-BW 31.9fJ/conv-step Filtering ΔΣ ADC with a Built-In DAC-Segmentation/ELD-Compensation 6b 960MS/s SAR-Quantizer in 28nm LP for 802.11ax Applications

C-Y. Wang, MediaTek, Hsinchu, Taiwan

In Paper 20.6, MediaTek shows a dual-mode delta-sigma including variable gain and filtering for an 802.11ax analog baseband. The 28nm design achieves 64.9dB SNDR in 80MHz bandwidth consuming 7.3mW of power, or 76.8dB SNDR in 10MHz bandwidth consuming 4.15mW of power.

11:45 AM



20.7 A 72.6dB-SNDR 100MHz-BW 16.36mW CTDSM with Preliminary Sampling and Quantization Scheme in Backend Subranging QTZ

W. Wang, University of Macau, Macau, China

In Paper 20.7, the University of Macau shows a wideband CTDSM with a preliminary sampling and quantization technique to increase the time available for quantization. The design, implemented in 28nm CMOS achieves 72.6dB SNDR in 100MHz bandwidth consuming 16.3mW of power.

20.1 A 5GS/s 7.2 ENOB Time-Interleaved VCO-Based ADC Achieving 30.5fJ/conv-step

Maarten Baert, Wim Dehaene

KU Leuven, Heverlee, Belgium

Technology scaling has been very beneficial for digital circuits both in terms of speed and power. Traditional analog techniques however are challenged by the ever-decreasing supply voltages. Highly digital VCO-based ADCs are able to benefit directly from improved digital performance [1]; however, the resolution and sampling rate of state-of-the-art VCO-based designs are insufficient for most applications. This paper presents a faster and more efficient VCO-based ADC architecture based on an improved high-speed, low-power ring oscillator and an asynchronous counting strategy. The architecture is 8× time-interleaved and combined with on-chip calibration. The design is implemented in 28nm CMOS and achieves 45.2dB SNDR (7.2 ENOB) near Nyquist at 5GS/s while consuming only 22.7mW, resulting in a Walden FOM of 30.5fJ/conv-step. The core area is only 0.023mm². These results demonstrate that VCO-based ADCs are a viable choice for next-generation Ethernet and high-speed wireless communication.

Figure 20.1.1 shows the ADC architecture. The 5GHz clock is divided into 8 low-speed, phase-shifted sampling clocks. A timing diagram of the sampling process is shown in Fig. 20.1.4 (top). Each channel samples the input onto a 300fF capacitor using a bootstrapped switch, while the VCO-based ADC core is connected to Vcm. The pseudo-differential implementation rejects common-mode interference and even-order distortion. During the hold phase, the sampled voltage is presented to the ADC core, which consists of a voltage-controlled ring oscillator (VCRO) connected to a counter. The input voltage is used as the control voltage of the VCRO. A bank of sampling flip-flops samples the state of the VCRO and counter. The sampled counter value corresponds to the number of cycles that have elapsed, while the sampled VCRO state provides extra fractional (LSB) information, which improves the resolution. Once decoded and combined, the resulting value is proportional to the integral of the VCRO frequency. During each tracking phase, the state of the core is sampled and compared to the previous state. The difference is proportional to the VCRO frequency, which in turn is determined by the input voltage. The non-linearity, offset and gain error of the VCRO tuning curve is calibrated using a 32-segment lookup table followed by a linear interpolator.

Figure 20.1.2 shows the VCRO architecture and corresponding tuning curve. The VCRO has 8 stages and is based on current-starved inverters. It is implemented differentially and makes use of feedforward cross-coupling to suppress common-mode gain without the use of differential pairs. The feedforward path results in a higher oscillation frequency, which improves the resolution of the ADC. A single current starving tail transistor is used for the entire oscillator instead of separate transistors for each inverter. The sensitive current starving transistor is moved out of the signal path, so it can be upsized to improve matching and reduce noise. The tail transistor also isolates the ring oscillator from the supply voltage, since the oscillation frequency is determined solely by the current flowing through the tail transistor. As a result, this VCRO design is significantly less sensitive to supply noise. The resulting VCRO has a usable tuning range from 2 to 12.3GHz with good linearity and an average power consumption of only 0.5mW.

Previous VCO-based ADC designs used a low-speed ring oscillator with a large number of stages, either because they did not include a counter and needed a large number of output phases to obtain the required resolution, or because the oscillation frequency had to be reduced to be compatible with a slow, synchronous counter. In both cases, the large number of stages would have resulted in high mismatch and a poor INL, so the transistors of the oscillator had to be upsized to reduce the mismatch, which increased the power consumption. Additionally, the large number of stages required an equally large number of sampling flip-flops, again resulting in increased power consumption.

The design presented in this paper uses a different approach: a fast asynchronous counter with a maximum clock frequency of 15GHz is used. This counter, shown in Fig. 20.1.3, is an 8b ripple counter formed by four cascaded divide-by-4 blocks. Since only the first divider operates at the full clock speed, this circuit is very energy-efficient. Sampling the counter value is challenging though: since the counter value changes asynchronously with respect to the sampling clock, there is a risk that the sampled value will contain a mixture of old, new and metastable bits. The problem is exacerbated by the propagation delay of the dividers: when operating at the maximum clock frequency of 15GHz, the propagation delay of the entire counter is much longer than one clock cycle.

For synchronous counters, such a problem can be solved with double sampling [1]: the counter value is delayed by a half clock cycle, and both the original and delayed values are sampled. At any point in time, at least one of these two values will be uncorrupted by transitions. The correct counter value can then be selected based on the clock value, which is also sampled.

In this work, the same principle is extended to asynchronous counters by treating the reconstructed value of each divider as the clock for the next divider. Figure 20.1.3 shows the schematic of this asynchronous double sampling decoder, and Fig. 20.1.4 (bottom) shows the corresponding timing diagram. For each divider, muxes are used to select only the sampled signals which are guaranteed to be stable. For the first divider, this selection is made based on the state of the counter clock (VCRO output). For the remaining dividers however, the reconstructed value of the previous divider is used for this selection, in order to avoid errors due to the propagation delay of the dividers. This strategy avoids metastability errors even when the total propagation delay is longer than one VCRO cycle. With four extra XOR gates (shown only in Figure 20.1.3), the state of the dividers can be converted to a binary value, which makes the asynchronous counter behave just like a synchronous counter operating at 15GHz.

The calibration parameters for each channel are determined by using a 5MHz sine wave as a test signal. These parameters are then loaded into the chip and used for all other measurements. Thanks to the efficient lookup table and interpolation-based approach, the power required for on-chip digital calibration is only 2.7mW. Sampling time mismatch is corrected using a tunable delay in the clock path (as in [3]).

Measurement results are shown in Fig. 20.1.5. At 5GS/s, the SNDR peaks at 46.7dB and drops to 45.2dB at the Nyquist frequency. The SFDR peaks at 60.3dB and drops to 57.1dB at the Nyquist frequency. The SNR is 51.0dB. The total power consumption (including clock management and calibration) is 22.7mW when using a 1.0V analog and a 0.85V digital supply. As the clock frequency is reduced, the effective resolution of the VCO-based core increases since more time is available for integration, and the SNDR gradually increases up to 49.8dB at 2GS/s. Thermal noise becomes more dominant in this range. At 1GS/s the bootstrap capacitor experiences significant charge leakage, and the resulting distortion from the sampling switch limits the ADC performance.

Figure 20.1.6 compares the performance of this work to several state-of-the-art ADC designs. Thanks to the energy-efficient high-speed VCRO and counter, the presented architecture achieves a resolution and speed similar to traditional designs, yet with a significant power and area reduction. A die photo is shown in Fig. 20.1.7. These results demonstrate that VCO-based ADCs are competitive with traditional designs for high-speed, medium-resolution applications.

Acknowledgements:
Maarten Baert is an SB PhD fellow at Research Foundation – Flanders (FWO).

References:
[1] J. Daniels, et al., "All-digital differential VCO-based A/D conversion," *IEEE ISCAS*, pp. 1085-1088, May 2010.
[2] B. Murmann, "ADC Performance Survey 1997-2018," Online. Available: http://web.stanford.edu/~murmann/adcsurvey.html.
[3] J. P. Keane, et al., "An 8GS/s time-interleaved SAR ADC with unresolved decision detection achieving –58dBFS noise and 4GHz bandwidth in 28nm CMOS," *ISSCC*, pp. 284-285, Feb. 2017.
[4] M. Brandolini, et al., "A 5GS/S 150mW 10b SHA-less pipelined/SAR hybrid ADC in 28nm CMOS," *ISSCC*, pp. 468-469, Feb. 2015.
[5] J. Fang, et al., "A 5GS/s 10b 76mW time-interleaved SAR ADC in 28nm CMOS," *IEEE CICC*, pp. 1-4, Sept. 2015.
[6] J. Wu, et al., "A 5.4GS/s 12b 500mW pipeline ADC in 28nm CMOS," *IEEE Symp. VLSI Circuits*, pp. C92-C93, June 2013.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

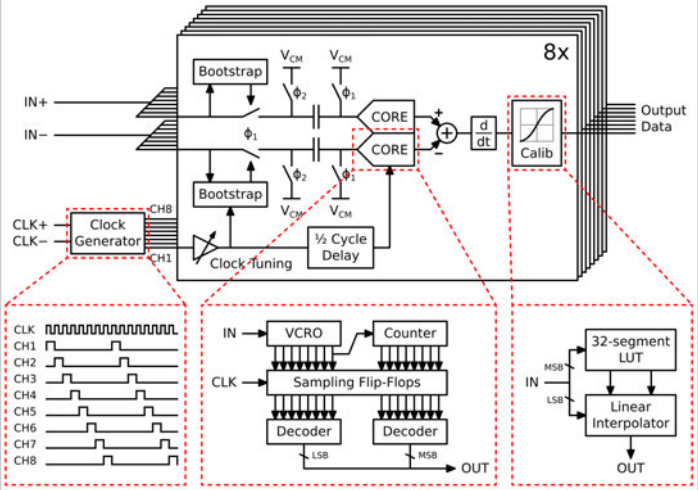


Figure 20.1.1: High-level ADC architecture with detailed view of clock generation, VCO-based ADC core, and on-chip digital calibration.

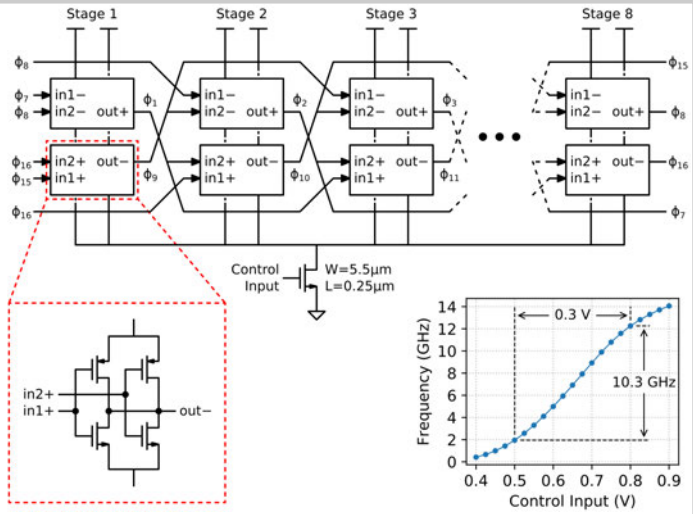


Figure 20.1.2: High-speed voltage-controlled ring oscillator (VCRO) with feedforward cross-coupling and shared tail transistor, and the corresponding tuning curve.

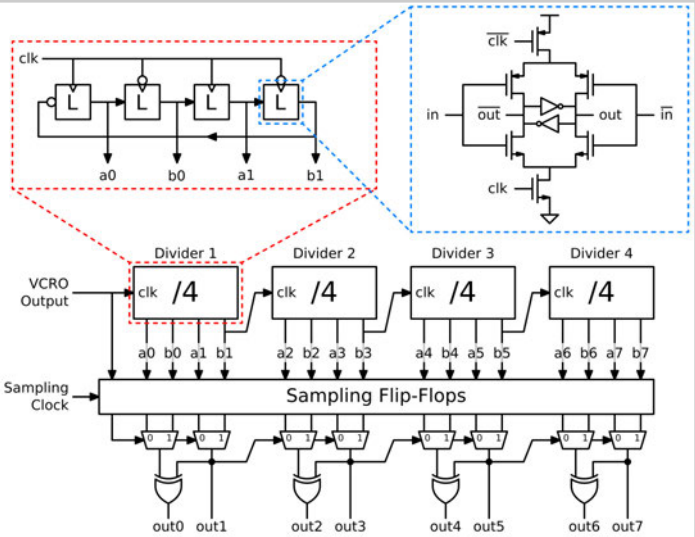


Figure 20.1.3: Asynchronous counter with decoder for asynchronous double sampling.

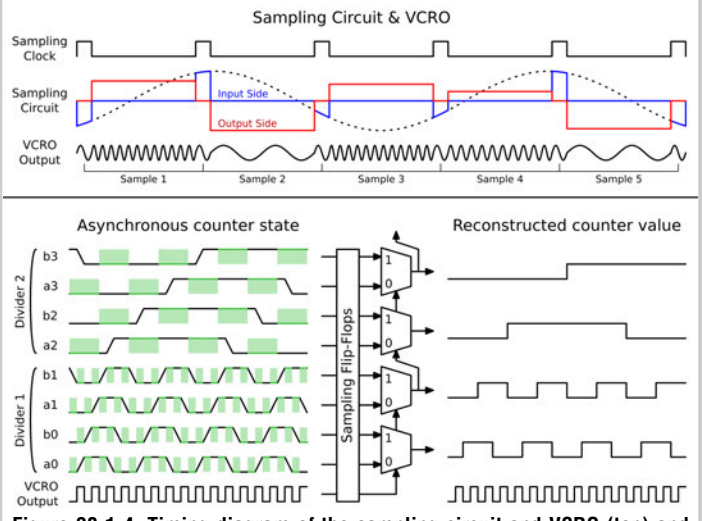


Figure 20.1.4: Timing diagram of the sampling circuit and VCRO (top) and asynchronous counter and decoder (bottom). The decoder selects only signals that are guaranteed to be stable (shaded areas).

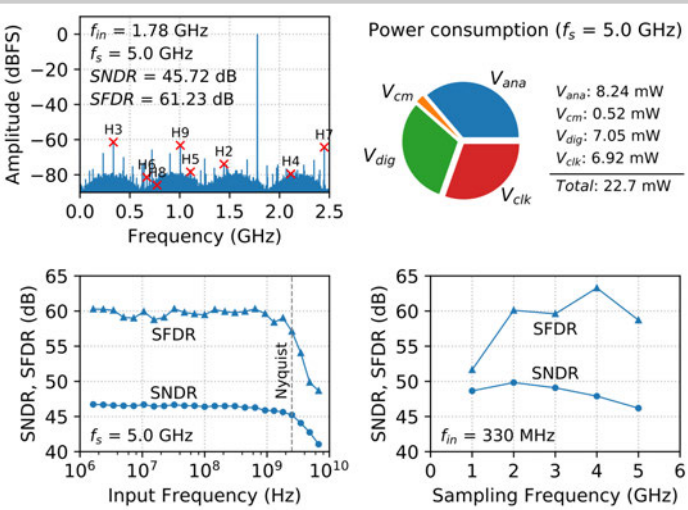
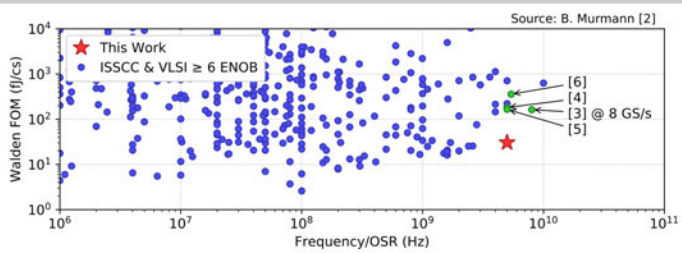


Figure 20.1.5: Measurement results: spectrum for single-tone input, power consumption breakdown and SNDR/SFDR as a function of input and sampling frequency (average of 3 samples).



	This Work	Keane [3]	Brandolini [4]	Fang [5]	Wu [6]
Published	ISSCC 2019	ISSCC 2017	ISSCC 2015	CICC 2015	VLSI 2013
Architecture	TI VCO	TI SAR	TI Pipe/SAR	TI SAR	TI Pipe
Technology	28nm	28nm FDSOI	28nm	28nm	28nm
Supply (V)	1.0/0.85	1.8/1.0/0.8	1.0	1.0	1.8/0.9
Sample Rate (GS/s)	5.0	5.0	5.0	5.0	5.4
Interleaving	8x	16x	2x/8x	12x	2x
Power (mW)	**22.7**	150	150	76	500
SNDR (dB)	45.2	51.7	46.1	42	50
SFDR (dB)	57.1	59.2	58.5	54.4	65
Walden FOM (fJ/cs)	**30.5**	95.5	181.9	165	358.4
Schreier FOM (dB)	**155.6**	153.9	148.3	147.2	147.3
Core Area (mm²)	**0.023**	0.184	0.45	0.57	0.4

Figure 20.1.6: Comparison with state-of-the-art ADCs achieving at least 6 ENOB [2-6]. There are no other VCO-based ADCs that approach the presented results.

20

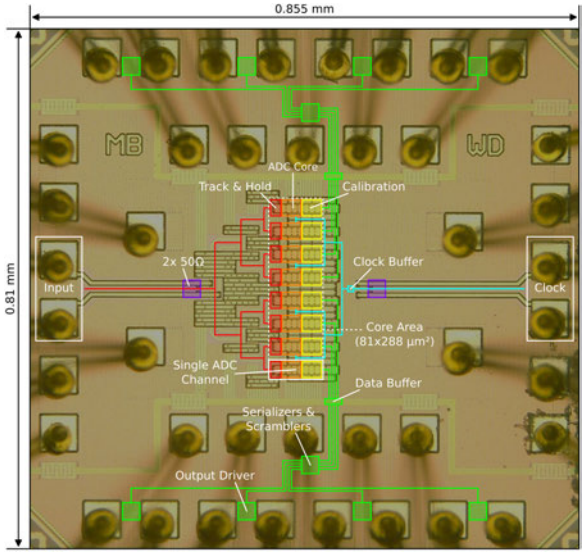


Figure 20.1.7: Die photo with important functional blocks highlighted.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

20.2 A 40MHz-BW 320MS/s Passive Noise-Shaping SAR ADC with Passive Signal-Residue Summation in 14nm FinFET

Ying-Zu Lin*, Chin-Yu Lin*, Shan-Chih Tsou, Chih-Hou Tsai, Chao-Hsin Lu

MediaTek, Hsinchu, Taiwan
*Equally-Credited Authors (ECAs)

SAR ADCs are popular in mobile WiFi applications due to their low power and small area. SNR of 60-70dB is necessary to meet the noise budget for the downlink chain in the 802.11 ac/ax standards. Comparator noise and quantization noise are typically the dominant noise sources limiting the SNR. Recently, noise-shaping SAR (NS-SAR) ADCs have become popular to increase SNR by reducing these two noise sources through noise shaping.

Active NS-SAR, first introduced in [1], uses an opamp to amplify and integrate residue. With residue amplification, the active NS-SAR achieves a noise transfer function (NTF) equivalent to high order noise shaping by performing charge sharing multiple times [1]. NS-SAR ADCs (passive or active) usually require a comparator with an additional input differential pair for the residue, which contributes extra thermal and kickback noise. An error-feedback (EF) NS-SAR uses a 1-input-pair comparator but its capacitive DAC (C_{DAC}) for input sampling must perform charge sharing with a residue capacitor for subtraction. Thus, an EF NS-SAR requires active residue amplification to maintain the signal amplitude and create a sharp NTF with small residue capacitance [2]. In a passive NS-SAR with no residue amplification, the residue amplitude degrades each time charge sharing occurs. Thus, passive noise shaping requires a gain ratio between the signal and residue input pairs of the comparator to compensate for the charge-sharing gain loss. Although the passive solution has weaker noise shaping, it is simpler in implementation and lower power [3-6]. In addition, the NS-SAR has a limited sampling rate because of extra timing slots required for charge sharing. The NS-SAR in [7] uses a ping-pong residue switching method to enhance the sampling rate, but the comparator still has 2 input pairs. This paper introduces a passive noise shaping technique using a comparator with only 1 input pair.

A conventional passive NS-SAR uses a 2-input-pair comparator, as depicted in Fig. 20.2.1. One pair for input, IP and IN, and the other for residue, RP and RN. For 2nd-order noise shaping, a 3-input-pair comparator is necessary [6]. C_{RES}, with the same capacitance as the integrating capacitor, C_{INT}, samples the residue after A/D conversion. To create a residue-to-input gain of 2, the width of the residue differential pair is sized 2× larger than that of the input pair. With the capacitance and gain ratios, the NTF will be $1-0.5*z^{-1}$ for a NS-SAR in [4]. In contrast, the NS-SAR in this work does not have to perform charge sharing between the C_{DAC} and C_{RES} during residue sampling. As a result, there will be a negative pole in the NTF, $(1-0.5*z^{-1})/(1+0.5*z^{-1})$. Because the negative pole suppresses low frequency noise, the noise shaping efficiency will be better than ideal 1st-order noise shaping for an OSR of 4. Ideally, $(1-0.5*z^{-1})/(1+0.5*z^{-1})$ provides 7.8dB suppression for quantization and comparator noise while the 1st-order noise shaping has 7dB noise reduction.

A 2-input-pair comparator induces more thermal noise than a 1-input-pair comparator. Moreover, the kick-back noise of a comparator is proportional to the total input transistor size, so the residue pair generates extra kick-back noise. In this 14nm design, a 1-input-pair comparator has 470μV (1σ) noise while a 2-input-pair one has 800μV. Considering the other noise sources including kT/C (270μV), reference (150μV), and quantization noise (370μV), 470μV and 800μV occupy 49% and 73% of the total noise power, respectively. In a conventional passive NS-SAR, the input and residue pairs should have the same common-mode voltage (V_{CM}) to maintain a stable gain ratio. A large difference in the V_{CM} between the differential pairs results in a large difference in transconductance. Consequently, a conventional NS-SAR cannot use a monotonic switching scheme, since the V_{CM} of the residue pair will be much lower than the initial V_{CM} of the input pair.

Conventional NS-SAR ADCs thus face a tradeoff between noise reduction due to noise shaping and noise increase due to the use of a multiple-input-pair comparator. The ADC in this paper avoids this tradeoff by using a passive signal-residue summation technique to achieve the desired NTF without the need for a multiple-input-pair comparator. As depicted in Fig. 20.2.2, the residue is sampled by two back-to-back capacitors C_{RES} connected across the top plates of the DAC arrays. This sampling method captures the signal component of the residue without the common-mode component. In addition, the sampled residue is 2× larger than that in conventional residue sampling. In other words, no active residue amplification by the comparator is necessary. The residue is put in series with the C_{DAC} holding the sampled input signal, thereby achieving summation. Since the passive signal-residue summation only needs a single-input-pair comparator, the comparator noise contribution is greatly reduced.

Figure 20.2.3 depicts the simplified block diagram of this NS-SAR ADC, which has 13 conversion cycles: the 5th and 12th are for redundancy and the 13th for implementing the noise quantizer technique [7]. The capacitors holding the residue ($C_{RP1,2}$ and $C_{RN1,2}$) operate in a ping-pong manner. C_{RP1} and C_{RN1} connect across the top plates of the DAC arrays, serving as attenuation capacitors when Φ_1 is high. When conversion finishes, C_{RP1} and C_{RN1} obtain the final residue with the same quantity but opposite polarity. Simultaneously, C_{RP2} and C_{RN2}, holding the final residue of the previous sample, perform charge sharing with integrating capacitors $C_{INTP,N}$, which are always connected between the comparator input nodes (SP and SN) and the top plates of the C_{DAC}. When Φ_2 is high, the roles of $C_{RP1,N1}$ and $C_{RP2,N2}$ are swapped. In [3-6], extra time slots are required to process residue switching and charge sharing. In this work, the non-overlap between Φ_1 and Φ_2 is less than 50ps. Hence, the noise shaping function does not affect the sampling (CLKS) and comparison (CLKC) timing. Moreover, the hardware overhead to implement the noise shaping function is only 6 capacitors and several switches.

Figure 20.2.7 displays the micrograph and layout of the NS-SAR ADC. The active area of the ADC core is $48 \times 45\mu m^2$. The C_{DAC} unit capacitance is 1.2fF, for a total DAC capacitance of 300fF, and C_{RES} is 40fF. With a supply voltage of 0.9V, the ADC consumes 1.25mW at 320MS/s. Figure 20.2.4 shows the FFT result with –0.1dBFS input amplitude. Figure 20.2.5 plots SNR and SNDR versus input amplitude for a 10MHz input. The peak SNDR is 66.6dB and peak SNR is 67dB. This passive noise shaping improves SNR by 4.4dB compared to a normal SAR ADC using only a noise quantizer technique. Using a larger capacitive DAC and residue sampling capacitor will further enhance SNR since kT/C noise induced by input and residue sampling cannot be shaped by a NTF. The resultant Walden FOM is 8.9 fJ/conv-step and the Schreier FOM is 171.7dB for a 40MHz bandwidth at 320MS/s.

Figure 20.2.6 summarizes the performance and comparison to state-of-the-art passive NS-SAR ADCs [8]. The residue and signal are summed passively without a multiple-input-pair comparator. Hence, the comparator thermal and kickback noise is greatly reduced. Only small hardware overhead is necessary to convert a conventional SAR ADC into the proposed one. The passive noise shaping with passive signal-residue summation, compatible with all types of C_{DAC} switching schemes (e.g., monotonic switching), can be seen as an SNR enhancement technique without having to increase DAC resolution or suppress comparator noise. The passive noise shaping technique presented here only slightly modifies the SAR ADC timing and does not decrease the operating speed. Finally, this ADC achieves the highest bandwidth compared with the other single-channel NS-SAR ADCs shown in Fig. 20.2.6.

Acknowledgment:
The authors would like to thank Stacy Ho and Chan-Hsiang Weng, Mediatek, for the technical discussions. We also appreciate the layout support by Chun-Ju Chuang, Mediatek.

References:
[1] J. Fredenburg, M. Flynn, "A 90MS/s 11MHz bandwidth 62dB SNDR noise-shaping SAR ADC," *ISSCC*, pp. 468-470, Feb. 2012.
[2] S. Li, et al., "A 13-ENOB 2nd-order noise-shaping SAR ADC realizing optimized NTF zeros using an error-feedback structure," *ISSCC*, pp. 234-235, Feb. 2018.
[3] Z. Chen, et al., "A 9.35-ENOB, 14.8 fJ/conv.-step fully-passive noise-shaping SAR ADC," *IEEE Symp. VLSI Circuits*, pp. 64-65, June 2015.
[4] W. Guo, N. Sun., "A 12b-ENOB 61μW noise-shaping SAR ADC with a passive integrator," *ESSCIRC*, pp. 405-408, Sept. 2016.
[5] Z. Chen, et al., "A 2nd order fully-passive noise-shaping SAR ADC with embedded passive gain," *IEEE ASSCC*, pp. 309-312, Nov. 2016.
[6] W. Guo, N. Sun, 'A 13b-ENOB 173dB-FoM 2nd-order NS SAR ADC with passive integrators,' *IEEE Symp. VLSI Circuits*, pp. 236-237, June 2017.
[7] Y.-Z. Lin, et al., 'A 2.4-mW 25-MHz BW 300-MS/s passive noise shaping SAR ADC with noise quantizer technique in 14-nm CMOS,' *IEEE Symp. VLSI Circuits*, pp. 234-235, June 2017.
[8] B. Murmann, "ADC Performance Survey 1997-2018," [Online]. Available: http://web.stanford.edu/~murmann/adcsurvey.html.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

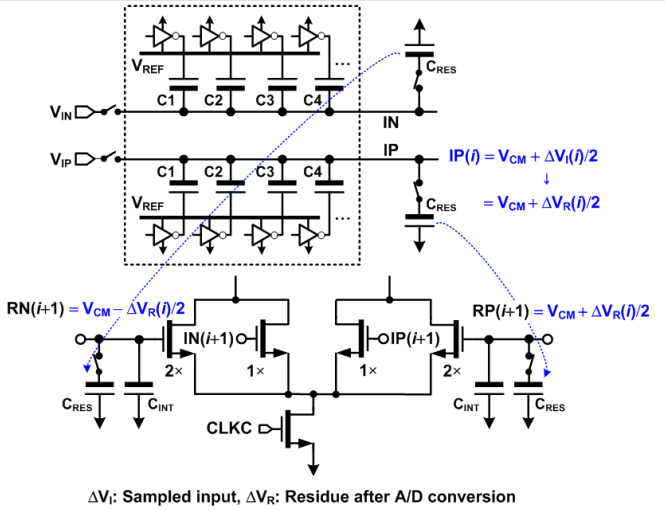


Figure 20.2.1: A conventional passive NS-SAR ADC using a 2-input-pair comparator. For simplicity, charge sharing between C_{INT} and C_{RES} is not considered here.

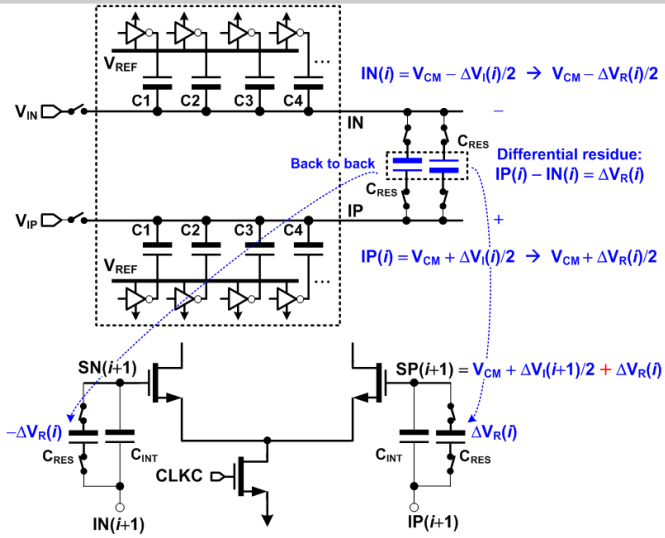


Figure 20.2.2: A passive NS-SAR ADC with passive signal-residue summation. For simplicity, charge sharing between C_{INT} and C_{RES} is not considered here.

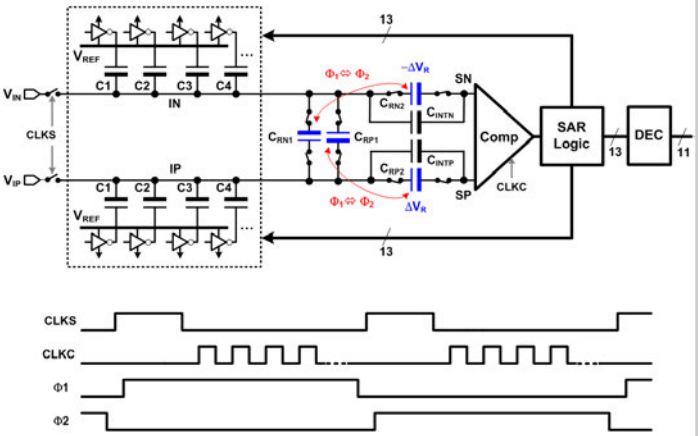


Figure 20.2.3: Simplified block diagram and timing diagram of the NS-SAR ADC using passive residue-signal summation.

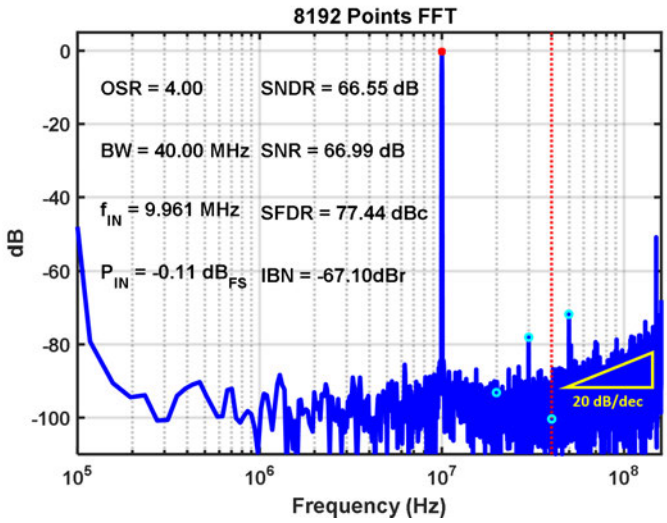


Figure 20.2.4: Measured FFT spectrum at 320MS/s.

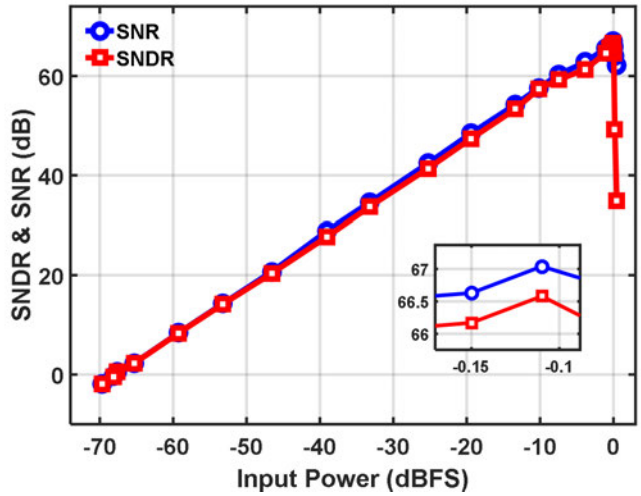


Figure 20.2.5: SNR and SNDR vs. input amplitude at 320MS/s.

	VLSI 15	ESSCIRC 16	ASSCC 16	VLSI 17	VLSI 17	This Work
Technique	Passive NS	Passive NS	Passive NS	Passive NS	Passive NS	Passive NS
Process (nm)	65	130	65	40	14	14
Supply (V)	0.8	1.2	1	1.1	1	0.9
Resolution (bit)	10	10	8	9	10	10
f_s (MS/s)	50	2	64	8.4	400	320
OSR	4	8	4	16	6	4
BW (MHz)	6.25	0.125	8	0.26	33	40
Power (mW)	0.12	0.06	0.25	0.14	2.8	1.25
SNDR (dB)	58	74	65	80	66	66.6
FOM_S (dB)	165	167	170	173	167	171.7
FOM_W (fJ/conv.-step)	14.8	60	10.9	33	25	8.9
Area (mm^2)	0.0123	0.13	0.013	0.04	0.0043	0.0021

Figure 20.2.6: Comparison to state-of-the-art passive NS-SAR ADCs.

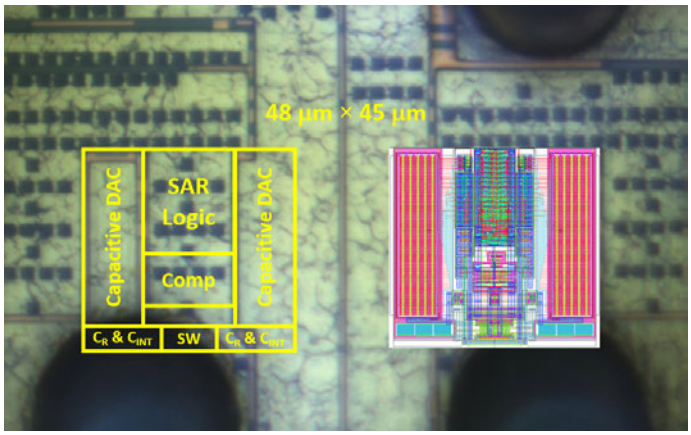


Figure 20.2.7: Chip micrograph and layout.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

20.3 A 50MHz-Bandwidth 70.4dB-SNDR Calibration-Free Time-Interleaved 4th-Order Noise-Shaping SAR ADC

Lu Jie, Boyi Zheng, Michael P. Flynn

University of Michigan, Ann Arbor, MI

Noise-Shaping SAR (NS-SAR) is an emerging ADC architecture that offers both high resolution and high energy efficiency. State-of-the-art NS-SAR ADCs eliminate the need for op-amps, which relaxes design complexity and technology scaling issues. However, existing NS-SAR ADCs, with high FoM, are limited in bandwidth [1,2] (typically in the MHz range). This makes NS-SAR ADCs unsuitable for applications that need bandwidths in the tens of MHz range, such as wireless communications. Traditionally, high-bandwidth, high-resolution applications utilize pipeline or continuous-time sigma-delta (CT-SD) ADCs, but these architectures are much more power hungry than the NS-SAR. Thus, to increase the bandwidth of NS-SAR ADCs and extend their low-power advantages, this work presents a new time-interleaved noise-shaping SAR (TINS-SAR) architecture that enables higher bandwidth. Although time-interleaving of ADCs is difficult for high resolution, interleaving impairments can be avoided when combining interleaving with noise-shaping. Our prototype 40nm CMOS TINS-SAR ADC has a measured SNDR of 70.4dB for a 50MHz bandwidth without calibration. It consumes only 13mW and occupies 0.061mm^2, making it a potential substitute for CT-SD ADCs.

Interleaving of NS-SARs is inherently challenging for the following reasons: 1) inter-channel feedback greatly increases system complexity; 2) residues are not available at the beginning of each conversion, making existing NS-SAR design methods not applicable; and 3) noise from channel mismatch can be large enough to cause the system to become unstable. Our proposed TINS-SAR architecture solves these problems, with an error-feedback-based, multi-phase, noise-shaping method to implement interleaving, shown in Fig. 20.3.1. Major interleaving impairments due to channel mismatch [3] can be relegated to out-of-band noise as long as we choose an OSR greater than or equal to channel number N, or keep the bandwidth less than Fs/(2N).

In our 4× interleaved design, the SAR core of each channel operates in four phases (P1-P4), each of which quantizes a quarter of the bits (Fig. 20.3.1). Instead of feeding the CDAC residue at the end of the channel conversion, (i.e. Q4), back into the same channel for the next conversion, we feed this residue to all four channels at different phases. This feedback is scaled with different coefficients, A,B,C,D in Fig. 20.3.1. Utilizing the delay, related to interleaving, a 4th-order error-feedback structure is formed and any form of noise transfer function (NTF) with four zeros can be realized. At any point during operation, only one channel generates a residue and provides feedback. Therefore only one shared analog bus is needed to pass the residue to the other channels, keeping the routing simple.

Unlike previous work on non-interleaved error-feedback-based NS-SARs [1], a charge-sharing implementation cannot be applied to a TINS-SAR since charge-sharing changes the stored-charge on the CDAC, disturbing the next phase conversion. Instead, in this work, we realize residue summation with the multi-input, summing pre-amplifier, shown in Fig. 20.3.2. The four capacitors (C1-C4) at the pre-amplifier input sample the pre-amplifier outputs (i.e. residue) of each of the four channels. In order to sample the residue of the channel itself, an extra auxiliary capacitor (C5) is added to first sample the output of the pre-amplifier. Afterwards, this sampled value is passed to capacitor C4 by simple charge sharing. The attenuation from charge-sharing (K_C) is considered part of the feedback coefficient. The pre-amplifier sums the voltages on the CDAC and the four capacitors to form the input of the comparator. Equivalently, the residues from all four channels are "added" to the CDAC. By adjusting the gain ($G_{A\text{-}D}$) of each pre-amplifier input, any form of 4th-order NTF can be realized.

The pre-amplifier approach has two inherent advantages: 1) its gain suppresses the noise and offset of the comparator; 2) it provides effective isolation between the comparator and feedback capacitors, protecting them from any kickback noise. Although the noise from the pre-amplifier sampled on the feedback capacitors is not shaped, this noise is set by the capacitance value and can be made negligible through careful design.

The active pre-amplifier only requires a low gain and operates in open loop. Therefore, it is much more power efficient and more amenable to process scaling compared to an op-amp in a pipeline or CT-SD ADC. In our prototype, we use a single-stage, differential amplifier, which consumes only 1.5mW per channel. A typical concern with this type of open-loop amplifier is PVT variation. However, since the inputs of the pre-amplifier are built with layout-matched differential pairs and drive the same load (Fig. 20.3.2), the gain ratio between the different inputs is precise and independent to PVT variation. As the transfer function strongly depends on gain ratios ($G_{A\text{-}D}$), but only weakly depends on the common gain variation, K, the performance of this method is PVT stable and no calibration is needed.

Overloading is a potential problem in a TINS-SAR since the large out-of-band noise spurs, induced by channel mismatch, can easily overload the quantizer. We prevent this issue both at the architectural and system levels. At the architectural level, we use a 10b SAR core with 6 redundant bits (Fig. 20.3.3). In addition to protecting against overloading, the redundant bits also provide some extra tolerance for CDAC settling errors. At the system level, we set the zeros of the NTF to 0.5, as shown in Fig. 20.3.3, for the following three reasons: 1) unlike a conventional 4th-order NTF with zeros at 1, this modified NTF provides a flatter in-band noise floor which is better for low-OSR applications; 2) the proposed NTF has good tolerance to coefficient variation, and thus no calibration or tuning is needed; and 3) most importantly, the feedback values for the proposed NTF are much smaller than for a conventional NTF, further preventing the quantizer from overloading.

Since the pre-amplifier has relatively large input pairs, the parasitic capacitance (C_{par} in Fig. 20.3.3) from the pre-amplifier input can cause noticeable distortion. To solve this, we add two extra switches ($S_{connect}$ and S_{short}) to isolate the CDAC and pre-amplifier. During the sampling phase, $S_{connect}$ is open and S_{short} is closed, shorting the input of the pre-amplifier to Vcm, resulting in input-independent charge storage. Capacitor C_n is added to limit the thermal noise from $S_{connect}$ and speed up settling. Bottom-plate sampling reduces the non-linearity caused by charge injection. To eliminate the effect of CDAC mismatch (~1%), dynamic element matching (DEM) shuffles the four MSBs in the CDAC. 23.4dB and 7.6dB improvements in SFDR and SNDR are observed, respectively, when all linearity enhancements are enabled.

Our prototype TINS-SAR ADC (Fig. 20.3.7) is fabricated in 40nm CMOS and has an active area of 0.061mm^2. Figure 20.3.4 shows the measured PSD at 400MS/s, with a peak SNDR and SFDR of 70.4dB and 88.0dB, respectively. Figure 20.3.4 also compares performance when noise-shaping and the linearity enhancements are enabled and disabled. With a 1V power supply, the ADC consumes 13mW at 400MS/s, where 1.9mW, 1.2mW, 5.9mW, and 4.0mW are dissipated by the reference, analog, pre-amplifier and digital circuitry, respectively, resulting in a Schreier FoM of 166.3dB. The measured performance vs input amplitude and vs input frequency are presented in Fig. 20.3.4. The measured dynamic range (DR) is 71.7dB. Figure 20.3.5 shows the measured performance of 10 devices. The average SNDR and SFDR exceed 69dB and 84dB without calibration. SNDR varies less than 0.3dB for a ±10% variation in pre-amplifier supply and 1.4dB for a 0-70°C variation in temperature. Figure 20.3.6 provides a performance summary and comparison with state-of-the-art, high-bandwidth NS ADCs. In conclusion, the new TINS-SAR architecture extends the bandwidth of the NS-SAR technique but keeps the advantages of high resolution and energy efficiency. Our design achieves a higher bandwidth among NS-SAR ADCs and approaches the performance of state-of-the-art, CT-SD ADCs.

Acknowledgements:
The authors would like to thank Intel and SRC for their support.

References:
[1] S. Li, et al., "A 13-ENOB 2nd-order noise-shaping SAR ADC realizing optimized NTF zeros using an error-feedback structure," *ISSCC*, pp. 234-235, Feb. 2018.
[2] C.-C. Liu, et al., "A 0.46mW 5MHz-BW 79.7dB-SNDR Noise-Shaping SAR ADC with Dynamic-Amplifier-Based FIR-IIR Filter," *ISSCC*, pp. 466-467, Feb 2017.
[3] N. Kurosawa, et al., "Explicit formula for channel mismatch effects in time-interleaved ADC systems," *IEEE Instrumentation and Measurement Technology Conference*, pp. 763-768, vol.2, 2000.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

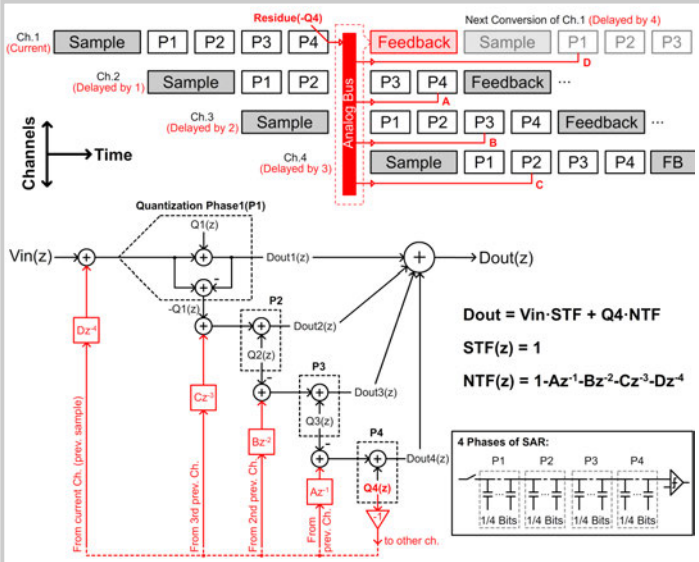


Figure 20.3.1: Interleaving sequence diagram and signal model diagram.

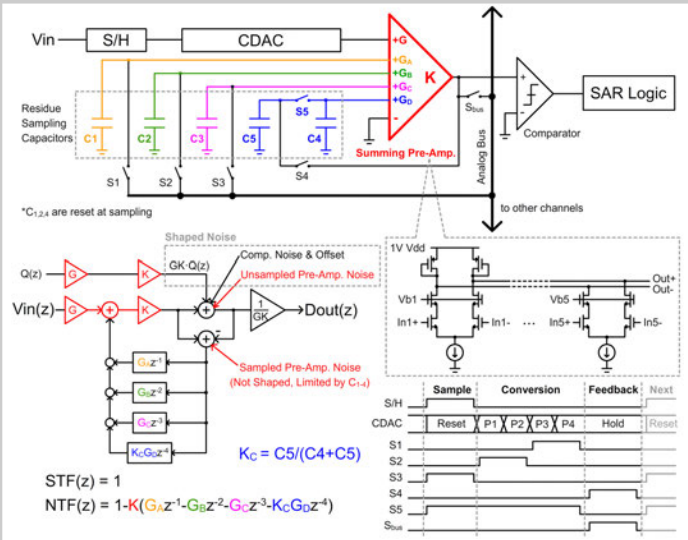


Figure 20.3.2: The proposed multi-input preamplifier realizing residue summation, the related signal model, and the feedback timing.

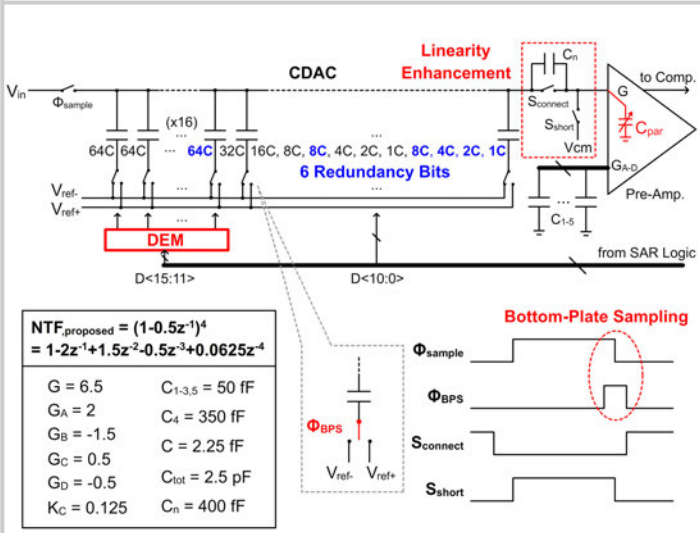


Figure 20.3.3: Details of the CDAC (single-ended) and the proposed linearity enhancements.

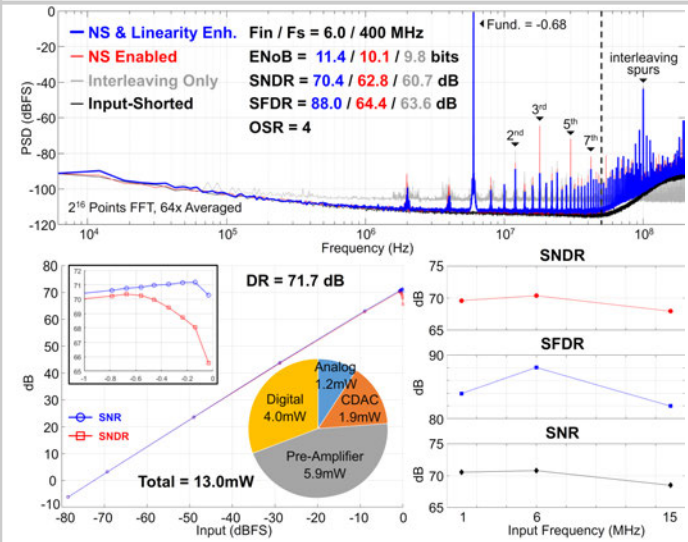


Figure 20.3.4: Measured output PSD (top), SNR/SNDR versus input power, power breakdown (left), and performance versus input frequency (right).

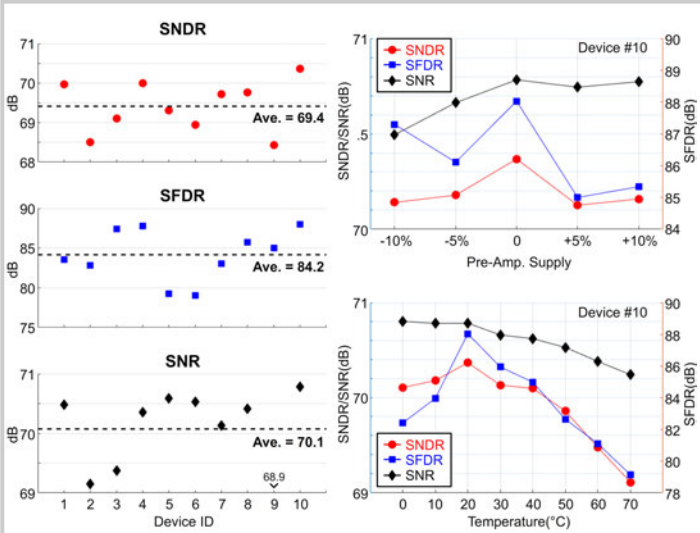


Figure 20.3.5: Measured performance of 10 devices (left), measured performance under supply voltage, and temperature variation (right).

	This work	[1]	VLSI 17 Lin	ISSCC 12 Fredenburg	ISSCC 18 He	VLSI 16 Jain
Architecture	**TINS-SAR**	NS-SAR	NS-SAR	NS-SAR	CT-SD	CT-SD
Calibration Free	√	X	√	√	X	X
Technology (nm)	**40**	40	**14**	65	28	65
Area (mm²)	**0.061**	0.024	0.0043	0.0462	0.25	0.07
Supply Voltage (V)	**1**	1.1	1	1.2	1.16/1.5	1.4
Power (mW)	**13.0**	0.084	2.4	0.8	64.3	13.3
Sampling Rate (MS/s)	**400**	10	300	90	2000	6000
OSR	**4**	8	6	4	20	50
Bandwidth (MHz)	**50**	0.625	25	11	50	60
NTF Order	**4**	2	1	1	4	4
SNDR (dB)	**70.4**	79	69.1	62.1	79.8	67.6
SFDR (dB)	**88.0**	89	78	72.5	95.2	77.4
DR (dB)	**71.7**	80.5	72	-	82.8	76
FoM_S (dB)	**166.3**	178	169.3	163.3	168.7	164.1
FoM_W (fJ/conv-step)	**48.1**	9	20.6	35.8	80.5	56.5

Figure 20.3.6: Performance summary and comparison with state-of-the-art high-bandwidth NS ADCs and CT-SD ADCs (last two columns).

20

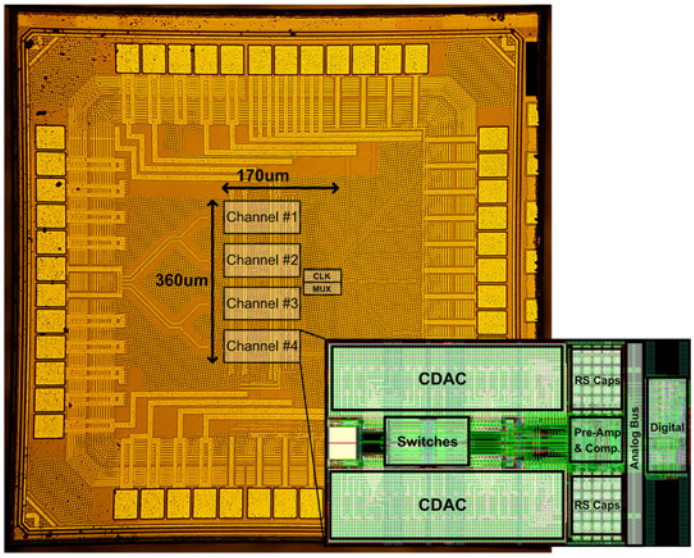


Figure 20.3.7: Die photo and zoomed-in layout of single channel.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

20.4 An 8×-OSR 25MHz-BW 79.4dB/74dB DR/SNDR CT ΔΣ Modulator Using 7b Linearized Segmented DACs with Digital Noise-Coupling-Compensation Filter in 7nm FinFET CMOS

Tien-Yu Lo, Chan-Hsiang Weng, Hung-Yi Hsieh, Ting-Yang Wang, Yun-Shiang Shu, Pao-Cheng Chiu

MediaTek, Hsinchu, Taiwan

A low-power and high-dynamic-range downlink ADC is the most critical building block in a cellular receiver design. A continuous-time delta-sigma modulator (CTDSM), which gets the benefit of inherent anti-alias filtering, is a common architecture choice for the ADC. However, low power dissipation dictates a low over-sampling ratio (OSR) since the bandwidth of the loop filter, DAC switching rate, and the clock frequency of the subsequent digital front-end are all proportional to the sampling frequency. Due to the desire for low OSR, a higher quantization level is required to maintain dynamic range, and the successive approximation register (SAR) quantizer, which gets the benefits of power and area efficiency compared to a flash quantizer, becomes more attractive in nanometer processes. However, the issue with a SAR quantizer in a high-speed CTDSM is the limited signal processing time available to achieve lower quantization noise and feedback DAC linearization. In this work, a 12 cycle, 400MHz passive noise-shaping SAR (NS-SAR) is embedded in a low-OSR CTDSM to fulfill the link budget requirement. Two techniques to satisfy the requirement of shorter signal processing time to DAC feedback path are introduced: segmented mismatch error shaping (S-MES) to achieve linear segmented DACs and digital noise coupling compensation filtering (DNCCF) to achieve low quantization noise.

Figure 20.4.1 shows the block diagram of the S-MES technique. The main purpose of this technique is to enhance the linearity of multi-level segmented DACs with a fast processing time. In previous work [1], the long signal processing time composed of the latency of the digital delta-sigma modulator, digital subtractor, and the DWA circuit, makes this method not suitable for a high-speed CTDSM. Compared to [1], S-MES can shape the gain error of multi-level segmented DACs in a short time period, while the mismatch error of the thermometer coded MSB DAC elements is shaped by data weighted averaging (DWA), and the binary coded LSB DAC is shaped by mismatch error shaping (MES) [2]. In S-MES, a second LSB DAC in the feedback path, $IDAC2_{LSB}$ (which is a copy of $IDAC1_{LSB}$ in Fig. 20.4.1) is designed to provide one cycle-delayed (z^{-1}·LSB) value, and a swapper is required to make sure the same LSB DAC is toggled by the delayed and current value between previous and present cycles. This means when one LSB DAC provides current LSB value in the feedback path, the other one can provide the z^{-1}·LSB value at the same time. Since the additional z^{-1}·LSB value is inserted into the modulator outer feedback path, it will be multiplied by the signal transfer function (STF_{DSM}) of the modulator and appears at the modulator output. Therefore, a duplicate z^{-1}·LSB value is added to the input of the quantizer (via $CDAC_{LSB}$ in Fig. 20.4.1), which means its value will be multiplied by the noise transfer function (NTF_{DSM}) at the modulator output. Since $STF_{DSM} + NTF_{DSM}$ is inherently equal to 1, the z^{-1}·LSB value appears at the modulator output and we can simply remove it by subtracting the same code at the modulator output to restore correct data ($Dout,_{SMES}$ in Fig. 20.4.1). Gain mismatch between the MSB and LSB DACs is another important issue. Fortunately, the z^{-1}·LSB value is added at the input of the quantizer, and it actually operates as a 1st-order digital delta-sigma modulator. As shown in Fig. 20.4.1, we can obtain $I1 = STF_{DSM}\cdot Vin + NTF_{DSM}\cdot Q_E - Q_L\cdot(1 - z^{-1})$ and $I2 = Q_L\cdot(1 - z^{-1})$, where Q_E is the quantization noise, and Q_L denotes the digital codes (LSB in Fig. 20.4.1) that are sent to the LSB DAC. By combining I1 and I2, Q_L will be cancelled if there is no gain mismatch between the MSB and LSB DACs. However, if a gain mismatch σ occurs, $\sigma\cdot Q_L\cdot(1 - z^{-1})$ appears in the feedback path, and thus the gain error will be 1st-order shaped.

Figure 20.4.2 shows the block diagram of the DNCCF. The main purpose of this technique is to reduce the quantization noise without inducing latency in the modulator feedback path. In this work, the digital noise coupling filter (DNCF) [3] is set to z^{-1}, which results in $Dout,_{DNCF}$ (DAC in Fig. 20.4.2) = $STF_{DSM}\cdot Vin + NTF_{DSM}\cdot[Q_R - (1-z^{-1})\cdot Q_T]$, where Q_T is the truncated digital code (TQN in Fig. 20.4.2), Q_R is the quantized noise, and Q_R is smaller than Q_T. Using DNCF, we can use fewer feedback bits in the CTDSM since the 1st-order shaped Q_T is smaller than Q_R by the high OSR. However, if the OSR is low, the 1st-order shaped Q_T is larger than Q_R and it still degrades the dynamic range within the signal bandwidth. As shown in Fig. 20.4.2, a DNCCF is implemented with $Hx = NTF_{DSM}$, $Hy = STF_{DSM}$, and Hz = -1 to cancel the shaped truncation error at the modulator output, resulting in $W = STF_{DSM}\cdot Vin + NTF_{DSM}\cdot Q_R + STF_{DSM}\cdot(1- z^{-1})\cdot Q_T$ and $Dout,_{DNCCF} = STF_{DSM}\cdot Vin + NTF_{DSM}\cdot Q_R$. In other words, $(1-z^{-1})\cdot Q_T$ is removed at the modulator output, while keeping the same feedback DAC bits. A 1st-order continuous-time loop filter is used in this work and thus the equivalent STF_{DSM} and NTF_{DSM} of the DNCCF are z^{-1} and $1 - z^{-1}$. A RC calibration circuit with 3% accuracy is implemented to minimize the coefficient mismatch between the continuous-time loop filter and the digital compensation filter, and the residual error is mitigated by 1st-order shaping, as shown in the equation.

Figure 20.4.3 shows the final block diagram of the proposed architecture, combining both S-MES and DNCCF techniques. A SAR quantizer which uses a 0.25Ts sampling period is utilized, where Ts is the sampling period of the ADC. The first 7 digital codes are sent to the segmented feedback DACs and the conversion time is limited to 0.5Ts ELD. The last 5 digital codes should be ready by the following 0.25Ts before the next sampling phase starts. A passive NS-SAR [4] with a NTF_{NSSAR} of $(1-0.5\cdot z^{-1})/(1+0.5\cdot z^{-1})$ is implemented, and the equivalent overall NTF of this modulator is $NTF_{DSM}\times NTF_{NSSAR}$. It results in more than 2nd-order shaping with low OSR. The 2 × DWA method, which doubles the rotation rate of the DWA state and switches the MSB DAC elements every 0.5·Ts, is utilized to get better suppression of DAC mismatch error. Figure 20.4.4 shows the detailed implementation of the proposed CTDSM. A feed-forward resistor is inserted in the 1st-order loop filter for ELD compensation, and 2-level thermometer/binary coded current DACs are implemented for S-MES. The SAR uses top-plate sampling with symmetric switching for the first 7 conversion cycles to maintain the input common mode voltage and speed of the comparator. The SAR then switches to a monotonic switching scheme for the following 5 conversion cycles, including 2 cycles for redundancy. The digital codes after the first 3 cycles are stored as the reference for the next sampling period to achieve S-MES and DNCCF operation. After the sampling period is finished, the reference sampling caps are reset and the information from previous cycles is subtracted in the charge domain.

The modulator is fabricated in a 7nm FinFET CMOS process. The loop filter is operated from 1V, the DACs from 1.3V, and the quantizer from 1V. The modulator consumes a total power of 3.8mW. Figure 20.4.5 shows the measured spectrum of the modulator with a 6.9MHz, -2dBFS sine wave. The THD improves from 63.1dB to 75.9dB by applying the S-MES technique, and the SNR improves from 71.3dB to 76.6dB by applying the DNCCF technique. When both S-MES and DNCCF are turned on, the modulator achieves 79.4dB DR and 77.4dB SNR, while peak SNDR is 74dB, with a THD of 76.6dB. Figure 20.4.7 shows the chip micrograph. The active area is 0.33mm × 0.17mm. Figure 20.4.6 summarizes the measured performance. Compared with state-of-the-art DSMs, this work demonstrates the combination of DNCCF and S-MES techniques to enable a CTDSM with a FoM_S of 177.6dB and FoM_W of 18.6fJ/conv-step.

Acknowledgements:
The authors thank Jui-Yuan Tsai, Stacy Ho, Hung-Chieh Tsai and Prof. Tsung-Hsien Lin for technical discussion.

References:
[1] R. Adams, et al., "A 113-dB SNR Oversampling DAC with Segmented Noise-Shaped Scrambling," *IEEE JSSC*, vol. 33, no. 12, pp. 1871–1878, Dec. 1998.
[2] Y.-S. Shu, et al., "An oversampling SAR ADC with DAC mismatch error shaping achieving 105 dB SFDR and 101 dB SNDR over 1 kHz BW in 55 nm CMOS," *ISSCC*, pp. 458–459, Feb. 2016.
[3] I.-H. Jang, et al., "A 4.2mW 10MHz BW 74.4dB SNDR Fourth-order CT DSM with Second-order Digital Noise Coupling Utilizing an 8b SAR ADC," *IEEE Symp. VLSI Circuits*, pp. 34-35, June 2017.
[4] Y.-Z. Lin, et al., "A 2.4-mW 25-MHz BW 300-MS/s Passive Noise Shaping SAR ADC with Noise Quantizer Technique in 14-nm CMOS," *IEEE Symp. VLSI Circuits*, pp. 234-235, June 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

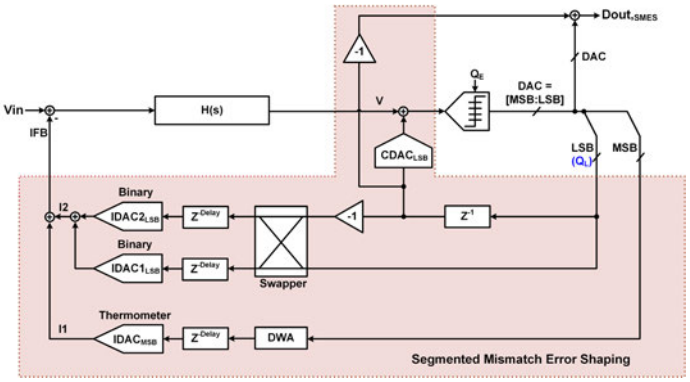


Figure 20.4.1: Block diagram of segmented mismatch error shaping topology (S-MES).

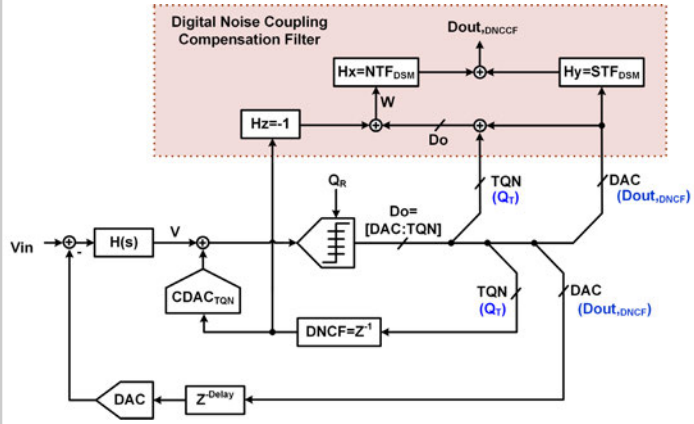


Figure 20.4.2: Block diagram of the digital noise-coupling compensation filter (DNCCF).

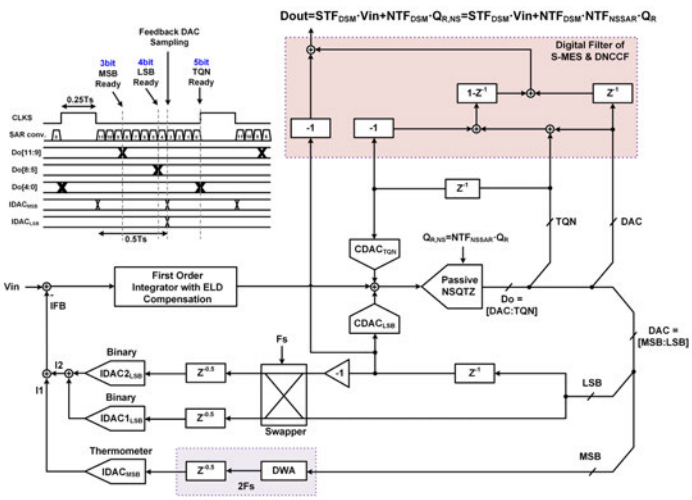


Figure 20.4.3: Architecture of the proposed implementation.

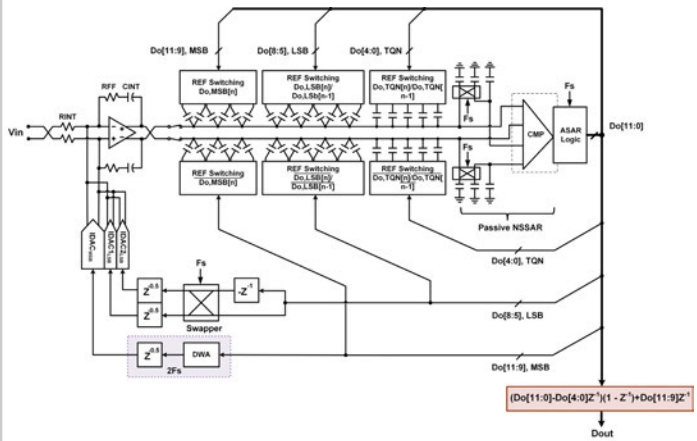


Figure 20.4.4: Detailed implementation of the proposed modulator.

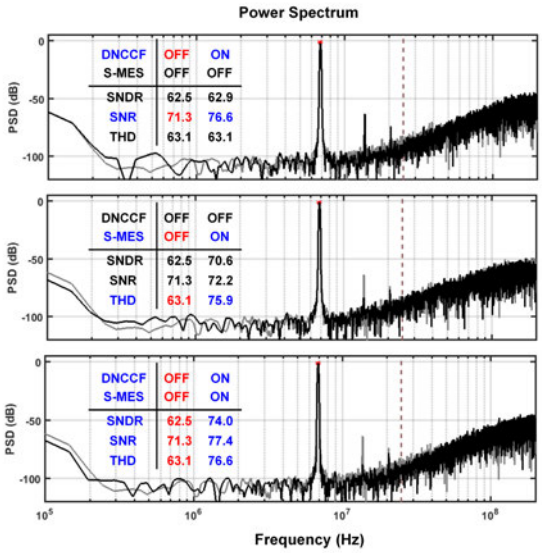


Figure 20.4.5: Measured output spectrum with a 6.9MHz input at -2dBFS.

	This work	ISSCC'18 He	ISSCC'17 Kim	ISSCC'16 Wu	ISSCC'15 Yoon	VLSI'15 Kao
Process [nm]	7	28	130	65	28	16
Fs [MHz]	400	2000	640	900	1800	832
BW [MHz]	25	50	15	45	50	29
OSR	8	20	21.3	10	18	14.3
DAC Linearity Calibration	Segmented mismatch error shaping (Background)	Off-Line calibration required	Sizing DAC Area	Off-Line calibration required	Off-Line calibration required	DWA (Background)
DR [dB]	79.4	82.8	82.9	82.5	85	75.1
Peak SNR [dB]	77.4	80	81.7	78.5	76.1	-
Peak SNDR [dB]	74	79.8	80.4	75.3	74.6	69.3
Power [mW]	3.8	64.3	11.4	24.7	78	12.4
Area [mm²]	0.056	0.25	0.17	0.16	0.34	0.23
FoMw [fJ/step]	18.6	80.5	44.1	57.7	177.7	89.7
FoMs [dB]	172.2	168.7	171.6	167.9	162.7	163
FoMsch [dB]	177.6	171.7	174.1	175.1	173.1	168.8

$FoMw = P/(2 \cdot BW \cdot 2^{(SNDR-1.76)/6.02})$ $FoMs = SNDR + 10 \cdot \log_{10}(BW/P)$ $FoM_{sch} = DR + 10 \cdot \log_{10}(BW/P)$

Figure 20.4.6: Performance summary and comparison to state-of-the-art ADCs.

20

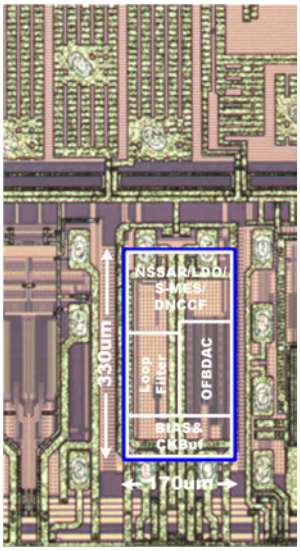


Figure 20.4.7: Chip micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

20.5 A 76.6dB-SNDR 50MHz-BW 29.2mW Noise-Coupling-Assisted CT Sturdy MASH ΔΣ Modulator with 1.5b/4b Quantizers in 28nm CMOS

Liang Qi[1], Ankesh Jain[2], Dongyang Jiang[1], Sai-Weng Sin[1],
Rui P. Martins[1,3], Maurits Ortmanns[2]

[1]University of Macau, Macau, China,
[2]University of Ulm, Ulm, Germany
[3]Instituto Superior Tecnico/University of Lisboa, Lisbon, Portugal

The demands for wider cellular bandwidth (BW) drive the development of continuous time (CT) ΔΣ modulators (DSMs). Oversampling ratio (OSR) is generally restricted due to high signal BW. To obtain an adequate resolution while maintaining good power efficiency, CT DSMs generally need to achieve an aggressive noise shaping and employ a multibit quantizer (QTZ) [1-3]. Though multibit operation requires a highly linear feedback (FB) DAC, dictating sophisticated linearization techniques [1-3]. Multi-stage noise-shaping (MASH) topologies can be employed to increase the order and they can apply multibit quantization only in the latter stages, where linearity requirements are highly relaxed. However, MASH DSMs suffer from quantization noise (QN) leakage due to the mismatch between analog and digital filters.

The sturdy MASH (SMASH) topology has relaxed matching requirements to eliminate the QN leakage. Reference [1] has first realized the CT SMASH using multibit QTZs in both stages. Inevitably, the multibit DAC FB from the QTZ of the 1st stage poses the same nonlinearity issue as in single-loop designs. On the other hand, the use of inherently linear single-bit 1st stage combined with multibit 2nd stage in SMASH is much more critical than in a MASH DSM. This occurs because single-bit QN is highly tonal and – in SMASH – the output of the 2nd stage is fed back to the sensitive input of the 1st stage. Moreover, the extraction and cancellation of the QN from the 1st stage is still challenging due to delays and phase shifts in the CT SMASH.

To explore a more robust CT SMASH topology and circumvent the use of linearization techniques for a multibit DAC, this work presents a noise-coupling-assisted CT SMASH DSM employing 1.5b/4b QTZs in the 1st/2nd stages. By effectively eliminating 1.5b QN of the 1st stage, this topology enjoys the benefits of multibit DAC FB. The noise coupling (NC) technique [4] not only increases the noise-shaping order by one, but also significantly mitigates the linearity requirement of the multibit DAC in the input front end. This SMASH renders the 4th-order 1.5b 1st stage effectively into an overall 4th-order multibit DSM without needing a highly linear DAC, thus achieving aggressive multibit QN shaping and large maximum stable amplitude (MSA) without employing area- or power-hungry DAC linearization techniques.

Figure 20.5.1 presents the block diagram of the SMASH DSM. It employs a 3-0 dual-stage architecture with 1.5b/4b QTZs in the 1st/2nd stages. The output of the 2nd stage is directly fed back into the 1st stage through DAC_2 without using a digital adder [1]. A 1st-order NC is applied in the 1st stage to increase the noise-shaping order, while a digital filter $(1-z^{-1})$ is correspondingly placed at the output branch of the 2nd stage. The digital outputs of two stages, V_1 and V_2, are combined off-chip to generate the final output V_{out}. Thereby, the overall output becomes

$$V_{out}=STF_1*V_{in}+NTF_1*(1-z^{-1})*(1-STF_2)*E_{q1} - NTF_1*(1-z^{-1})*E_{q2}$$

where NTF_1 is the noise transfer function (NTF) generated by the loop filter $H_{LF}(s)$. Instead of using the 1st-order feedforward (FF) topology as in [1], this work chooses zero-order for the 2nd stage to implement a more accurate and robust unity-gain signal transfer function (STF_2). Thereby, the 1.5b QN E_{q1} can be more precisely eliminated compared to [1]. To obtain SQNR>85dB, the gain error of the NC path should stay within ±5%, which can be reliably achieved by implementing it using a switched-capacitor (SC) circuit. Another beneficial feature of the NC branch in the 1st stage is that it works as dithering, significantly reducing idle tones and harmonic spurs for 1.5b QN E_{q1} [4]. Furthermore, before the output of the 2nd stage is fed back into the front end, it is shaped by the digital filter $(1-z^{-1})$, thus reducing the in-band tone power. Consequently, both features significantly mitigate the linearity requirement of the outermost 4b DAC_2.

However, with the presence of 4b DAC_2 mismatch and non-linearity, the in-band noise floor still gets increased due to the QN-folding generated by the out-of-band cross-modulation, thus degrading the SNR. To mitigate this, a 2-tap FIR filter $(1+z^{-1})/2$ is adopted to reduce the out-of-band QN from the 2nd stage, which is eventually combined with the $(1-z^{-1})$ filter due to NC to result in $(1-z^{-2})/2$, without requiring additional hardware. To implement $(1-z^{-2})/2$, a tri-level encoder is employed, followed by a tri-level DAC. Moreover, the FIR filter reduces the jitter sensitivity of the outermost DACs.

Figure 20.5.2 presents a simplified schematic of the CT SMASH DSM, using the NRZ tri-level scheme for all DACs. The loop filter $H_{LF}(s)$ employs a 3rd-order mixed FF/FB architecture with H_{inf}=2.3. This mixed mode separates the high-gain and high-speed requirements into different integrators, allowing better op-amp power efficiency. A local resonator improves the NTF, and the input FF path reduces the swing of the internal nodes. A simple FIR compensation filter Fc(z) – incorporated in the inner FB paths – restores the original NTF from the introduced FIR filter in the outermost FB. In [1], to correctly extract E_{q1}, an RC low-pass filter is used to generate a propagation delay for the continuous input of the 1st QTZ, thus matching the delay of the 1st QTZ. However, this is sensitive to the PVT variations. To circumvent this issue, E_{q1} is extracted using an SC array [2] integrated in the 1.5b QTZ_1. After correct extraction, E_{q1} passes an SC buffer, and injects back into the 3rd integrator to facilitate the 1st-order NC. Meanwhile, the output of the SC buffer directly drives the QTZ of the 2nd stage.

Figure 20.5.3 depicts the detailed 1st-order NC implementation and excess loop delay (ELD) compensation incorporated in the 1.5b QTZ_1. During the nth sampling phase, the bottom plates of the DAC capacitors C_{DAC1} and C_{DAC2} connect to the digital outputs V_1 and $V_2(1-z^{-1})$ from the previous cycle, while their top plates sample the output of the 3rd integrator. After that, by connecting the bottom plates to V_{CM}, the input node of the 1.5b QTZ_1 settles to $V_{Int3}-(V_1-V_2(1-z^{-1}))*z^{-1}*V_{Ref}$. Thereby, a negative FB path with a nominal unity gain is generated around the 1.5b QTZ_1, stabilizing the DSM with the 0.75*Ts ELD [5]. Since this ELD compensation (ELDC) method cannot help to reduce the swing of the last integrator, a scaling factor of 1/4 is applied in the 3rd integrator, reducing the swing as well as enhancing the feedback factor. To maintain the original NTF, the references of the 1.5b QTZ_1 are scaled down by a factor of 4, correspondingly. After the decision of the 1.5b QTZ_1, the digital codes V_1<1:0> are fed back to C_{DAC1} to produce the current residue E_{q1}. Setting the values of $C_{DAC1}=C_{res}$ and $C_{DAC2}=2*C_{res}$, the residue is correctly scaled as $E_{q1}/4$, and subsequently sampled by two capacitors C_{res} alternately [2]. Next, the residue charge is injected back to the 3rd integrator through a unity-gain SC buffer in a closed-loop operation. Thereby, a complete 1st-order NC path consists of the SC buffer and the summer formed by the 3rd integrator.

The prototype is fabricated in 28nm CMOS. Clocked at 1.2GHz, Fig. 20.5.4 shows the spectrum with a 6MHz, -1.6dBFS input signal. The measured SNDR/SFDR/DR are 76.6dB/87.9dB/80dB, respectively. The measured IMD3 is -84.6dBc/-82.5dBc with two -7.6dBFS inputs at 33.5MHz/36.5MHz. The MSA is around -1dBFS, which clearly shows the overall multibit operation of the SMASH DSM. With 1.2V/1.5V supply, the total power consumption is 29.2mW. Its breakdown is shown in Fig. 20.5.5. The achieved FOM_S is 168.9dB. Figure 20.5.6 compares the design to wideband CT DSMs. This work achieves state-of-the-art FOM with high linearity without employing any DAC linearization technique, thereby consuming an area of 0.085mm². Figure 20.5.7 shows the chip micrograph.

Acknowledgements:
This work was supported by Macao Science and Technology Development Fund (FDCT) under Grant 076/2017/A2 and SKL-AMSV-2017-2019(DP). The authors would like to sincerely thank Biao Wang and Jiaji Mao for technical support.

References:
[1] D.-Y. Yoon, et al., "A Continuous-Time Sturdy-MASH ΔΣ Modulator in 28nm CMOS," *IEEE JSSC*, vol. 50, no. 12, pp. 2880-2890, Dec. 2015.
[2] B. Wu, et al., "A 24.7mW 45MHz-BW 75.3dB-SNDR SAR Assisted CT ΔΣ Modulator with 2nd-Order Noise Coupling in 65nm CMOS," *ISSCC*, pp. 270-271, Feb. 2016.
[3] T. He, et al., "A 50MHz-BW Continuous-Time ΔΣ ADC with Dynamic Error Correction Achieving 79.8dB SNDR and 95.2dB SFDR," *ISSCC*, pp. 230-231, Feb. 2018.
[4] K. Lee, et al., "An 8.1mW, 82 dB Delta-Sigma ADC with 1.9 MHz BW and -98 dB THD," *IEEE JSSC*, vol. 44, no.8, pp. 2202-2211, Aug. 2009.
[5] Y. Shu et al., "A 28fJ/conv-step CT ΔΣ Modulator with 78dB DR and 18MHz BW in 28nm CMOS Using a Highly Digital Multibit Quantizer," *ISSCC*, pp. 268-269, Feb. 2013.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

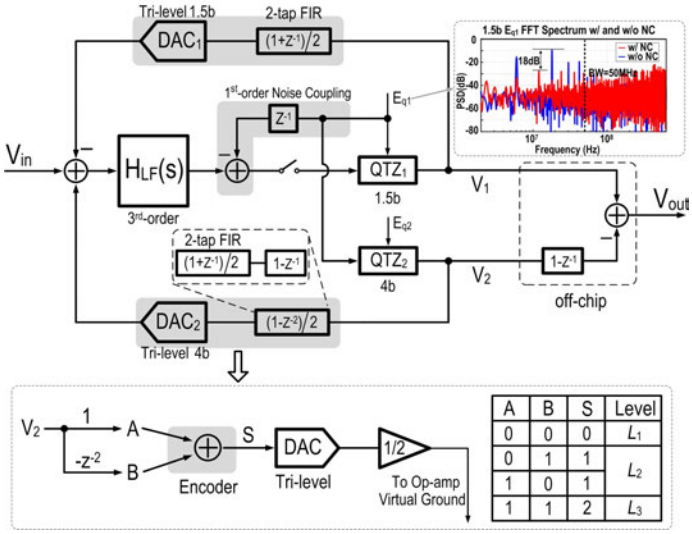


Figure 20.5.1: Block diagram of the CT SMASH DSM with illustration of the tri-level DACs.

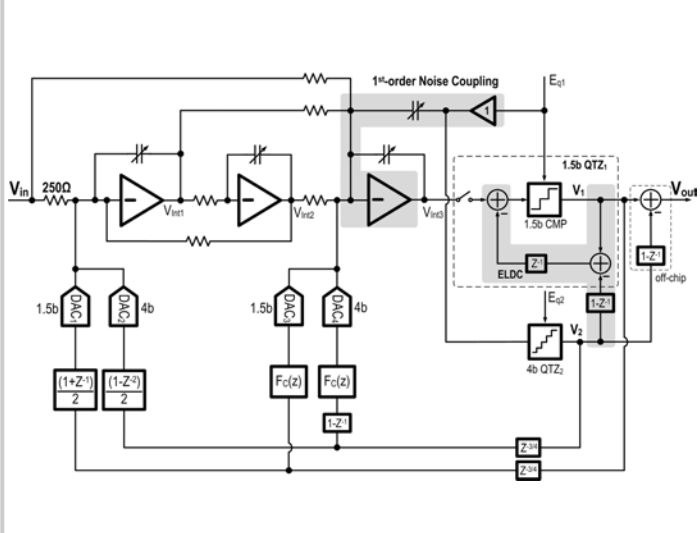


Figure 20.5.2: Simplified schematic of the CT SMASH DSM.

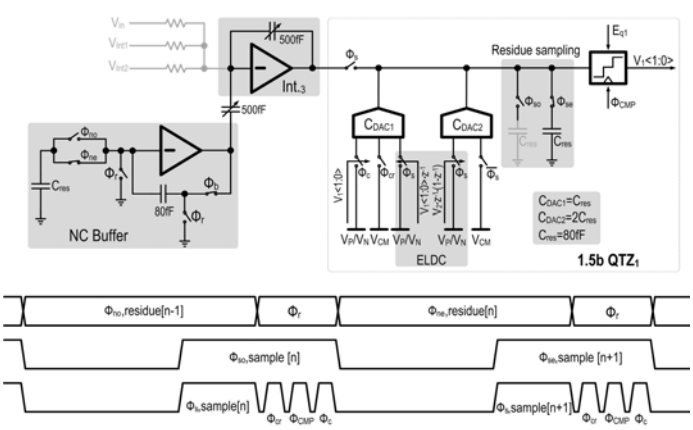


Figure 20.5.3: The 1st-order NC implementation and the ELDC incorporated in 1.5b QTZ_1 with the corresponding timing diagram.

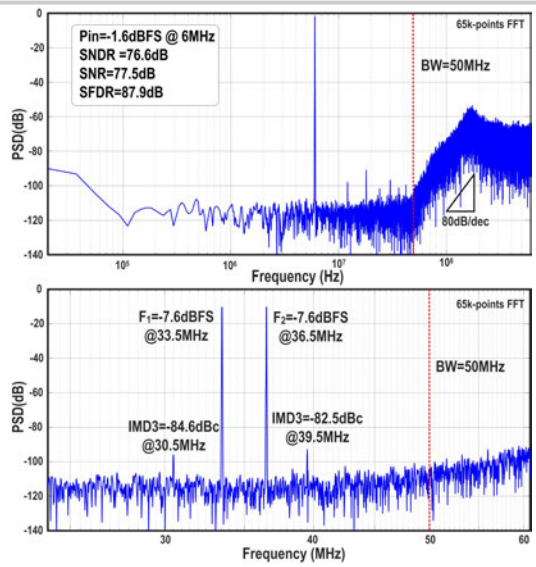


Figure 20.5.4: Measured spectrum of the CT SMASH DSM output and IMD3.

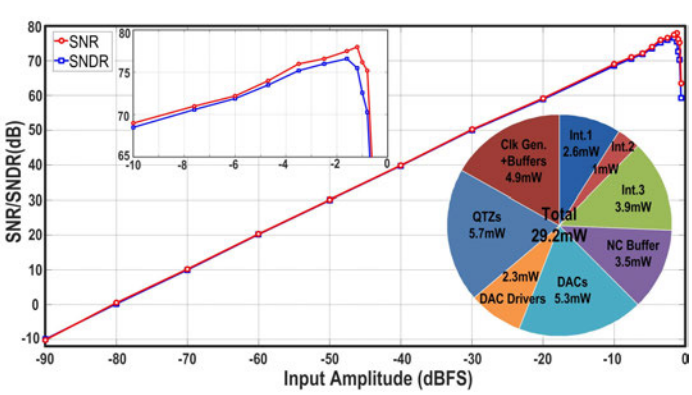


Figure 20.5.5: Measured SNR/SNDR vs. input amplitude and the power breakdown.

	This Work	ISSCC'18 He [3]	ISSCC'17 Huang	ISSCC'16 Wu [2]	ISSCC'15 Yoon [1]	VLSI'15 Loeda
Architecture	**SMASH 4th-order 1.5b/4b**	Single-loop 4th-order 4b	Single-loop 4th-order 7b	Single-loop 6th-order 4b	SMASH 4th-order 4b/4b	Single-loop 4th-order 1b
Process (nm)	**28**	28	16	65	28	40
Fs (GHz)	**1.2**	2	2.15	0.9	1.8	2.4
BW (MHz)	**50**	50	125	45	50	40
SNDR (dB)	**76.6**	79.8	71.9	75.3	74.6	66.9
SNR (dB)	**77.5**	80.0	72.6	78.5	76.1	N/A
DR (dB)	**80.0**	82.8	74.8	82.5	85	67.8
SFDR (dB)	**87.9**	95.2	N/A	83	89.3	N/A
THD (dBc)	**-83.9**	-94.1	-80	-78.1*	-79.9*	N/A
Supply (V)	**1.2/1.5**	1.16/1.5	1/1.35/1.5	1.2/1.8	1.2/1.5	N/A
Power (mW)	**29.2**	64.3	54	24.7	78	5.25
Area (mm²)	**0.085**	0.25	0.217	0.16	0.34	0.02
DAC Calibration	**Without**	**With on-chip**	**With on-chip**	**With off-chip**	**With on-chip**	**Without**
FOM_W (fJ/conv.-step)	**52.8**	80.5	67.2	57.7	177.7	36.3
FOM_S (dB)	**168.9**	168.7	165.5	167.9	162.7	165.7

$FOM_W=P/(2*BW*2^{(SNDR-1.76)/6.02})$ $FOM_S=SNDR+10*\log_{10}(BW/P)$

*THD calculated based on the difference between SNRs and SNDRs reported

Figure 20.5.6: Performance summary and comparison with the state-of-the-art wideband CT DSMs.

20

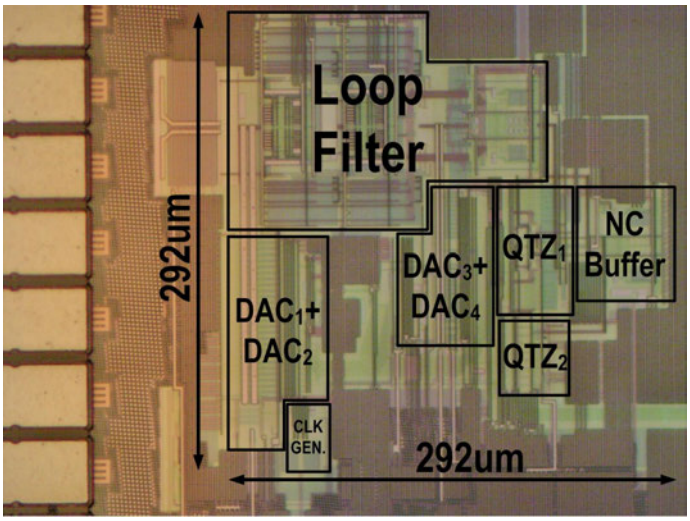


Figure 20.5.7: Chip micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

20.6 An 80MHz-BW 31.9fJ/conv-step Filtering ΔΣ ADC with a Built-In DAC-Segmentation/ELD-Compensation 6b 960MS/s SAR-Quantizer in 28nm LP for 802.11ax Applications

Chi-Yun Wang, Jen-Huan Tsai, Sheng-Yuan Su, Jen-Che Tsai, Jhy-Rong Chen, Chih-Hong Lou

MediaTek, Hsinchu, Taiwan

The demand for modern portable smart devices is growing at an unprecedented rate, which speeds up the development of the next-generation wireless-LAN (WLAN) standard. To improve spectrum efficiency and serve more users in crowded areas while increasing maximum throughput, 802.11ax supports 1024QAM modulation under MCS11 with OFDMA technology over 160MHz RF signal bandwidth (BW160). Taking channel fading margin into account, 1024QAM requires a receiver (RX) EVM better than -40dB when supporting multi-user MIMO. To satisfy the stringent RX EVM requirement, the intrinsic dynamic range (DR) of the RX analog baseband over such a wide in-band bandwidth (BW) of 80MHz must exceed 65dB to guarantee that the noise contribution is low enough, with sufficient headroom for the large peak-to-average-power-ratio of the OFDM-based 1024QAM signal. Moreover, the potential co-existence of the LTE-U TX application in the License Assisted Access band with 5G WLAN imposes extremely strict anti-aliasing requirements on the RX analog baseband. Hence, compared with discrete-time-based ADCs, which inevitably need the aid of intensive analog filtering, the continuous-time (CT) ΔΣ modulator (DSM) is a favorable RX ADC candidate due to its inherent anti-aliasing ability. The CTDSMs [1,2] achieve high DR over wide bandwidth by clocking at multi-gigahertz sampling rates to maintain a moderate oversampling ratio (OSR) with sufficient noise-shaping ability. The overhead is tens of milliwatts (mW) of power consumption per ADC, which may account for a significant portion to the total RX power and is not efficient enough for portable devices. This work presents a power-efficient 802.11ax RX analog baseband realized by a single filtering ΔΣ ADC [3] that uses a 6b 960MS/s SAR-quantizer combined with both low-latency DAC-segmentation (LLDS) functionality and excess-loop delay compensation (ELDC) [4]. While the prototype operates under an extremely low OSR of 6×, it achieves SNDR/DR of 64.9dB/68dB over an 80MHz BW with a total power consumption of 7.33mW in 28nm low-power (LP) CMOS. Meanwhile, an additionally embedded 12dB PGA gain range and a peaking-free signal transfer function (STF) make this work even more attractive for RX analog baseband applications.

Figure 20.6.1 shows the circuit implementation of the proposed filtering ΔΣ ADC. A 2nd-order CTDSM realized by a single-opamp resonator [5] is placed in the loop of an order-configurable analog filter for further suppressing the in-band noise, as well as smoothing out its inherent STF peaking. To compensate for the inferior loop-filter noise shaping ability under the extremely low OSR of 6×, the quantizer resolution is refined to 6 bits. While preserving a full-cycle (1T) feedback delay in the DSM loop, the successive approximation register (SAR) ADC is chosen as the quantizer due to its well-known energy efficiency with moderate resolution and conversion rate. DAC-segmentation function is combined with ELDC and realized inside the quantizer with no extra latency. After segmentation from 6b to 3 groups of 4b-1.5b-1.5b, the complexity of both the filter-DAC (DAC1) and the DSM-DAC (DAC2) is significantly reduced, and the number of their tri-level NRZ DAC units are reduced from 32 to 10, respectively.

The analog filter could be programmed as a 2nd-order filter (with a 1st-order active integrator cascading a 1st-order passive-RC) for intensive STF filtering. If a better in-band SNR is required, the passive-pole could be bypassed (C_2 off), and part of the integrator capacitance (C_{1B}) is sliced off to guarantee stronger noise suppression to the DSM stage ($1/2\pi R_1C_{1A}$ > 80MHz). Meanwhile, the higher loop gain of analog filter in the low-noise configuration is counteracted by introducing a negative feed-forward resistor (R_{FF}) connected between the integrator output and the center node of the resonator C-R-C T-network (V_X). Hence, with either analog filter configurations, the unconditionally stable nature of the 2nd-order DSM is maintained since the analog filter loop gain diminishes at high frequencies. Even if the input signal is amplified exceeding full scale, only clipping occurs instead of having the un-wanted overload behavior that needs a time-consuming reset. It is the key advantage of this architecture for acting as a self-contained analog baseband with built-in PGA functionality (by tuning the input resistor R_1). No extra noise-contributing PGA, nor swing-limiting preceding stage are needed.

Figure 20.6.2 shows the linear model of the proposed 6b SAR-quantizer. Due to the discrete-time (DT) characteristic of the SAR ADC, clock-based signal processing can be easily implemented inside it with less overhead. In order to divide the 6b DAC into segmented 4b-1.5b-1.5b groups with low latency, the 2 LSBs (B[1:0]) of the previous-cycle output are directly added back to the comparator input to be re-quantized. By doing this, the 4 output MSBs (B[5:2]) of the SAR-quantizer directly contain the information of both the real digitized output (V) and the 1st-order shaped truncation error (E_{T1} and E_{T0}) of the 2 missing LSBs (B[1:0]). Meanwhile, the shaped truncation error with opposite polarity, which is the information of the 2 LSB DAC groups, could be acquired without extra latency by merging the differentiation operation ($1-z^{-1}$) within the tri-level DAC latch. Unlike the conventional method, which needs an extra digital DSM with considerable loop delay, the segmentation function completes immediately after the LSB (B[0]) conversion is done. On the other hand, to compensate for the 1T ELD and stabilize the DSM loop with H_{inf}=3.5, a zero-order feedback path with coefficient k_0=2 is adopted. The ELDC is merged with LLDS and realized inside the SAR-quantizer using an FIR-like feedback manner. The building blocks of the 6b LLDS-ELDC hybrid SAR-quantizer and the related timing diagram are depicted in Fig. 20.6.3. The capacitor DAC (CDAC) is split into a 6b array mainly for conversion, and a 6b+2b auxiliary one which provides the insufficient coefficient weighting needed for LLDS and ELDC. Once the input is sampled, the top plates of all the CDAC cells in the two arrays are immediately reset (RST) to the reference voltage (V_{REF}) and complete the LLDS and ELDC feedback before the conversion phase. With a large ELDC coefficient k_0=2, the RST behavior also raises the common-mode voltage (V_{CM}) by about half of the SAR-quantizer full-scale (i.e., 250mV). The delicately lifted V_{CM} speeds up the operation of the NMOS-input dynamic comparator and facilitates 6b SAR conversion within 0.65T of the 960MHz clock cycle. With the aid of the LLDS, the binary-to-thermometer (B2T) transformation and data-weighted averaging (DWA) logic of the 4b MSB DAC operate in parallel with the conversion of the 2 LSBs and are no longer timing critical.

For Wi-Fi 802.11ax RX analog baseband applications, the proposed work is actually multi-mode configurable and supports modes from BW20 up to BW160 with a 2.1V full scale. Due to limited space, only the results of the two extreme modes, BW20 and BW160, are shown. Figure 20.6.4 shows the measured peak SNDR and the FFT spectra. The results of the two BW160 configurations are both demonstrated for comparison. When low-noise configuration is adopted, the noise is improved by about 2.3dB compared to the result derived from the filter-intensive mode, and the SNDR/SNR/DR measured over an 80MHz in-band BW are 64.9dB/65.7dB/68dB, respectively. Under BW20, the prototype is clocked at 480MS/s and achieves 76.8dB/79.6dB/82.5dB of SNDR/SNR/DR over a 10MHz in-band BW. The measured STFs over modes are shown in Fig. 20.6.5. The inherent STF peaking of the 2nd-order CTDSM is obviously suppressed by the analog filter, which prevents the potential DR loss at the blocker frequency. The BW160 low-noise mode tends to have less filtering ability than the filter-intensive mode does, which is the trade-off for better in-band performance. The bias circuit and DAC1/DAC2 consume current from a 2V supply while the rest circuits are powered by 1.15V. The total power consumption under BW160/BW20 is 7.33mW/4.15mW, respectively.

This chip is fabricated in 1P7M 28nm LP CMOS technology and occupies an active area of 0.06mm^2, as shown in the chip microphotograph in Fig 20.6.7. The table in Fig. 20.6.6 summarizes the performance of this work. When comparing to state-of-the-art CTDSMs with in-band BW ranges from 60MHz to 125MHz, the BW160 mode of this prototype demonstrates a comparable Schreier FoM of 165.3dB/168.4dB (in SNDR/DR), while achieving the best Walden FoM of 31.9fJ/conv-step with the lowest OSR of 6×. Although the most challenging target for a multi-mode design is to keep circuit performance over a wide BW programming range, the BW20 still delivers both competitive Schreier and Walden FoMs of 170.6dB/176.3dB (in SNDR/DR) and 36.7fJ/conv-step.

The proposed filtering ΔΣ ADC is more than an ADC. It acts in the role of a standalone RX analog baseband for 802.11ax applications, which is equivalent to an architecture of a blocker-rejection or anti-aliasing filter/PGA cascaded with a conventional ADC. Not limited to the WLAN, the prototype is also a promising solution to be utilized in the 5G mobile communication generation.

Acknowledgement:
The authors would like to thank George Chien, Hung-Chieh Tsai, and Chun-Cheng Liu, MediaTek, for paper review and technique discussion.

References:
[1] S. Ho, et al., "A 23mW, 73dB Dynamic Range, 80MHz BW Continuous-Time Delta-Sigma Modulator in 20nm CMOS," *IEEE Symp. VLSI Circuits*, June 2014.
[2] S.-J. Huang, et al, "A 125MHz-BW 71.9dB-SNDR VCO-Based CT ΔΣ ADC with Segmented Phase-Domain ELD Compensation in 16nm CMOS," *ISSCC*, pp. 470-472, Feb. 2017.
[3] M. Andersson, et al., "A Filtering ΔΣ ADC for LTE and Beyond," *IEEE JSSC*, vol. 49, no. 7, pp. 1535–1547, July 2014.
[4] G. Wei, et al., "A 13-ENOB, 5 MHz BW, 3.16 mW Multi-Bit Continuous-Time ΔΣ ADC in 28 nm CMOS with Excess-Loop-Delay Compensation Embedded in SAR Quantizer," *IEEE Symp. VLSI Circuits*, pp. C292-C293, June 2015.
[5] K. Matsukawa, et al., "A Fifth-Order Continuous-Time Delta-Sigma Modulator with Single-Opamp Resonator," *IEEE JSSC*, vol. 45, no. 4, pp. 697–706, Apr. 2010.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

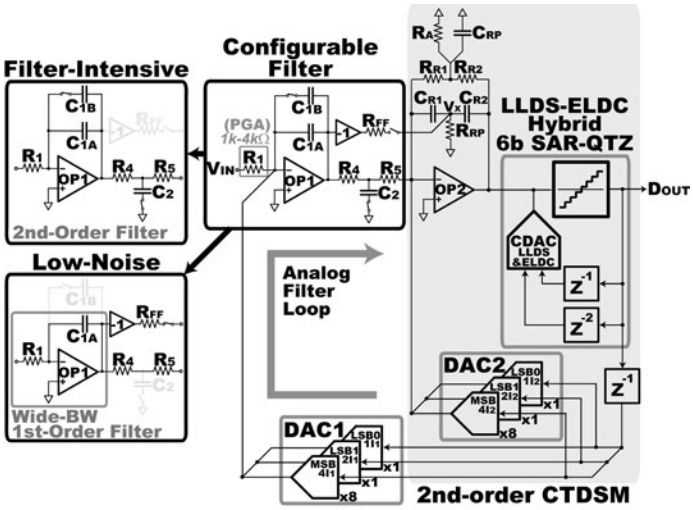


Figure 20.6.1: The circuit implementation of the 6b filtering ΔΣ ADC and its analog filter configurability.

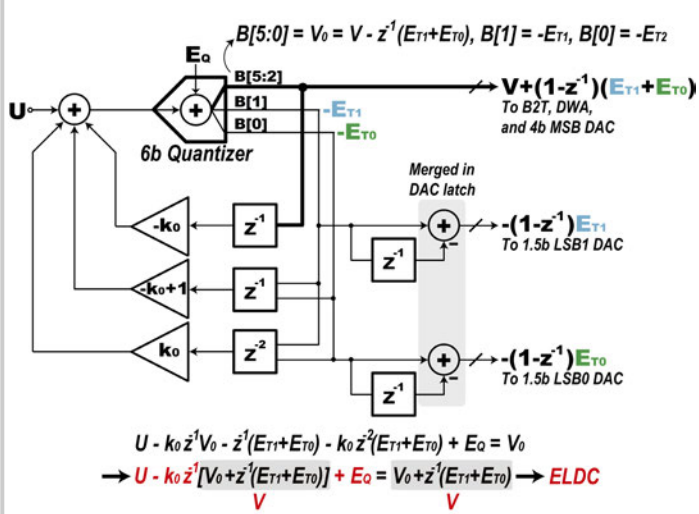


Figure 20.6.2: Linear model of the LLDS-ELDC hybrid 6b SAR-quantizer.

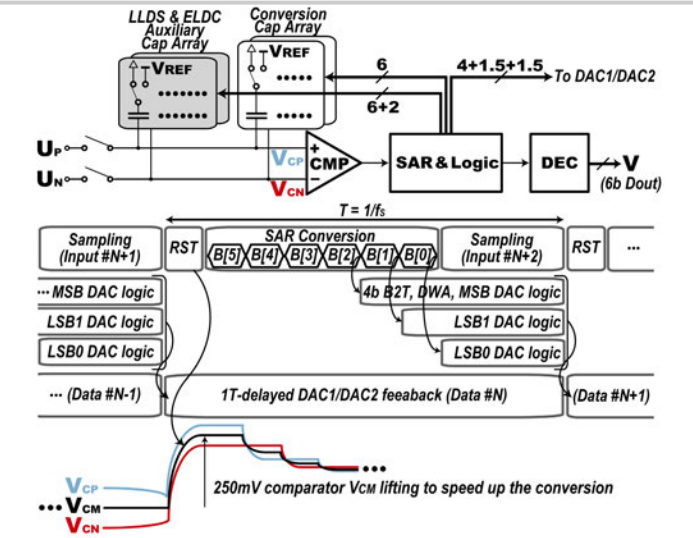


Figure 20.6.3: Schematic and related timing diagram of the LLDS-ELDC hybrid 6b SAR-quantizer.

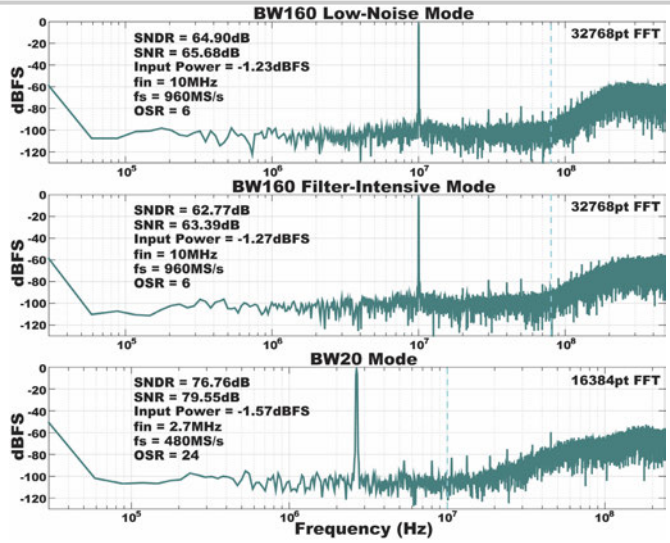


Figure 20.6.4: FFT spectra of BW160 low-noise mode, BW160 filter-intensive mode, and BW20 mode.

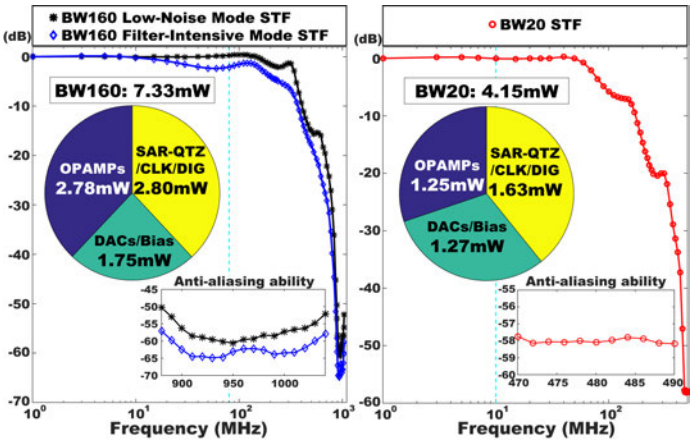


Figure 20.6.5: Measured STF plots (from 1MHz to fs+BW) and power consumption breakdown of both BW160 and BW20.

Group	BW = 60MHz ~ 125MHz					BW = 10MHz ~ 12.5MHz		
Work	This Work BW160	ISSCC17 Huang	VLSI16 Jain	VLSI15 Briseno-Vidrios	VLSI14 Ho	This Work BW20	VLSI18 Liu	VLSI17 Jang
Technology (nm)	**28**	16	65	40	20	**28**	40	28
Fs (GS/s)	**0.96**	2.15	6	3.2	2.184	**0.48**	0.5	0.32
OSR	**6**	8.6	50	21.33	13.7	**24**	20	16
BW (MHz)	**80**	125	60	75	80	**10**	12.5	10
SNDR (dB)	**64.9**	71.9	67.6	64.9	67.5	**76.8**	70.4	74.4
SNR (dB)	**65.7**	72.6	68.8	65.4	70	**79.6**	-	-
DR (dB)	**68**	74.8	76	67.7	73	**82.5**	73	80.8
Power (mW)	**7.33**	54	13.3	22.8	23	**4.15**	1.16	4.2
FoM_W (fJ/conv.)	**31.9**	67.2	56.5	105.8	74.2	**36.7**	17.1	49
$FoM_{S,SNDR}$ (dB)	**165.3**	165.5	164.1	160.1	162.9	**170.6**	170.7	168.2
$FoM_{S,DR}$ (dB)	**168.4**	168.4	172.5	162.9	168.4	**176.3**	173.3	174.6
Area (mm²)	**0.06**	0.217	0.07	0.09	0.1	**0.06**	0.029	0.1

$FoM_W = P/(2*BW*2^{(SNDR-1.76)/6.02})$; $FoM_{S,SNDR} = SNDR+10*\log_{10}(BW/P)$; $FoM_{S,DR} = DR+10*\log_{10}(BW/P)$

Figure 20.6.6: Performance summary and comparison with state-of-the-art CTDSMs of two BW groups, respectively.

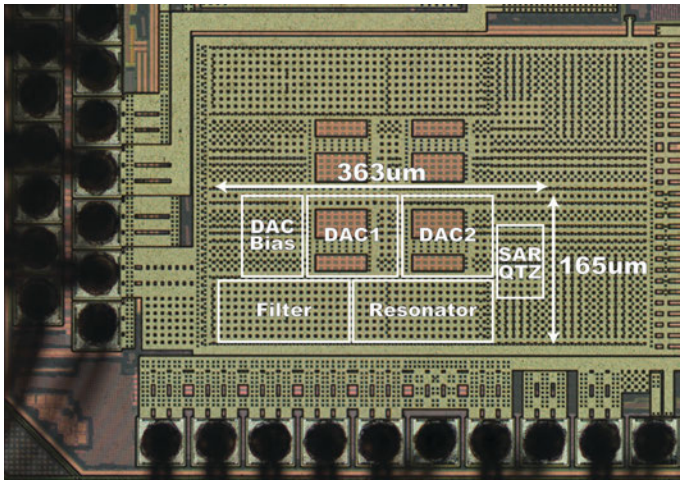


Figure 20.6.7: Chip microphotograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

20.7 A 72.6dB-SNDR 100MHz-BW 16.36mW CTDSM with Preliminary Sampling and Quantization Scheme in Backend Subranging QTZ

Wei Wang[1], Chi-Hang Chan[1], Yan Zhu[1], Rui P. Martins[1,2]

[1]University of Macau, Macau, China
[2]Instituto Superior Tecnico/University of Lisboa, Lisbon, Portugal

Driven by great demands for high data transfer rates in mobile communications, ADCs require wide bandwidths with low noise density and power consumption. The continuous-time $\Delta\Sigma$ modulator (CTDSM) is a suitable candidate for such applications as it has a simple slew rate requirement and inherently contains an anti-aliasing function. To satisfy the wide bandwidth (~100MHz for LTE-A) and simultaneously maintain the energy efficiency of the CTDSM, a multi-bit quantizer (QTZ) is often utilized that relaxes the OSR, jitter and loop filter stability. On the other hand, new design issues, such as DAC nonlinearity and high power consumption in the data feedback and QTZ, rise with large numbers of bits in the QTZ. While the DAC mismatch can be suppressed by calibrations [1,2] and the number of feedback DACs can be reduced to just one with feedforward [3] or QTZ-based excess loop delay (ELD) compensation [4], the QTZ design becomes the bottleneck in both the speed and resolution of the modulator. Here, we introduce a preliminary sampling and quantization (PSQ) technique that allows almost full utilization of a clock period for the quantization, thus leading to an extra 30% conversion time or an additional 12dB SQNR contribution when compared with the conventional approach. The modulator runs at 2GHz and attains 72.6dB SNDR and 76.3dB DR over a 100MHz bandwidth, while only consuming 16.36mW.

Conventionally, the ELD compensation limits the available conversion time (T_{av}) of the QTZ because it needs to cover the delays caused by the QTZ latency, DAC feedback and loop filter (LF) settling. Therefore, the available conversion time is highly confined under such circumstances, thus limiting the architectural choice of the QTZ. In this work, rather than squeezing all of the conversion within T_{av}, we manage to reuse the ELD period for quantization. Such an arrangement enables a 2-step operation in the backend QTZ that greatly relieves the timing and architectural constraints in conventional CTDSMs. Figure 20.7.1 illustrates the signal characteristics and timing diagram of the proposed CTDSM. We separate both the sampling and quantization into coarse and fine segments, where the coarse sampling is leading the fine sampling for a period (Δt_{CF}, details of its choice explained next). After the coarse QTZ determines the first two MSB decisions, they are employed by the fine QTZ and convert together with the fine sample. Then, the fine QTZ resolves the rest of the LSBs and finally feeds back both MSB and LSB decisions to the input through the DAC with ELD. ϕ_F and ϕ_C are the sampling instants of the fine and coarse QTZ, respectively, where the stability of the modulator (under a limited ELD compensation ability) determines ϕ_F. In conventional CTDSMs, the time before ϕ_F is squandered as the LF output is considered not yet ready. Here we utilize such an interval with a PSQ. Nevertheless, as the coarse and fine QTZs sample the output of the LF at different instants, they induce a sampling error ε_{SAM} related to Δt_{CF}, which trades off the available conversion time of the coarse QTZ.

Aside from Δt_{CF}, ε_{SAM} is also affected by the out-of-band gain, the order and response of the LF, and the input and the resolution of the QTZ. Figure 20.7.1 (top) shows the signal behavior of the LF output with an ideal and finite frequency response, respectively, with a 100MHz sinusoid input. Its response consists of two parts: bandwidth limited and input tracking. After the DAC feedback, the LF equivalently responds to a step input. While limited by the finite bandwidth of the opamps in the LF, the output deviates from its ideal value but eventually converges when the response becomes moderate during input tracking. Such a characteristic modulates ε_{SAM} diversely at different LF output conditions. At close to the zero crossing, ε_{SAM} decreases when comparing it with the ideal response, as the DAC feedback and the settling directions of the LF output are opposite. On the other hand, when the LF output is close to the peak value, ε_{SAM} rises because of the accumulated settling error. Consequently, with a limited opamp bandwidth of 7GHz, the emergence of the largest ε_{SAM} evolves from the ideal zero crossing case to its peak, while such error can still be corrected together with the first-stage residual offset by 1b overlapping between the coarse and fine QTZ. With 130ps Δt_{CF} in this design, the maximum ε_{SAM} is around 100mV, which is well covered by the 175mV redundancy range in the second stage.

Figure 20.7.2 exhibits the simplified schematic of the QTZ in a single-ended configuration while in the proposed design we utilize a differential structure. The coarse QTZ is a two-step architecture with the MSB resolved directly with the comparator C1 at ϕ_{ST1} after the coarse sampling ϕ_C. The decision from C1 then controls the bottom-plate switch of DAC2-4, which generates the reference voltages for C2-4 with different capacitance sizing ratios. C2-4 triggered together at ϕ_{ST2} resolve 2 more bits and the results from C1-4 are then fed to the segmented DAC of the fine SAR ADC directly without decoding. Finally, the fine SAR ADC resolves the remaining 4 bits after the fine sampling ϕ_F. In this design, the conversion of the coarse QTZ takes about 0.25 Ts (125ps), which benefits from the direct control and the small DACs (3fF). The offset of the comparators (C1-4) in the first stage is calibrated in foreground to a 4b level off-chip where the residual error is corrected by the redundancy. The whole conversion is running in an asynchronous manner where it waits until the validation of all the decisions from C2-4 to trigger the first comparison in the fine SAR ADC. The conversion time of the fine QTZ is close to 0.65 Ts (325ps), as we adopt a small DAC (9.6fF) and dynamic logic. As a result, the 7b subranging ADC can run at 2GHz while only consuming ~1.4mW of power.

Figure 20.7.3 presents the 4th-order CTDSM employing the PSQ scheme. It contains two simple RC integrators, one single amplifier biquad integrator, a 7b QTZ (with 1b redundancy) and a segmented NRZ DAC. We use a two-stage topology with feed-forward path and a Miller compensation amplifier [3] in the integrator, with the capacitors of the LF digitally programmable to cover the process variation on the transfer function. Similar to [1,3,4], we utilize the feedforward architecture to reduce the power from the opamps and realize the ELD compensation with 0.65 Ts ELD, such that only one main feedback DAC is necessary. Further, we design a 3-4b segmented NRZ DAC, which not only matches with the outputs of the QTZ but also helps to reduce the power of the DAC drivers while its mismatch is calibrated in the digital backend [1]. The supply voltage of the current steering DAC is 1.5V, which provides sufficient linearity and noise performance. Only one clock is necessary in this ADC while the others are generated asynchronously from the operation.

The CTDSM prototype fabricated in 28nm CMOS attains a 100MHz bandwidth with 2GHz sampling frequency. Figure 20.7.4 plots the spectrum of the modulator with -2dBFs input at 18MHz and an input common mode of 0.5V. The measured peak SNDR, SNR and SFDR are 72.6dB, 73.2dB and 83.6dB, respectively. The measured IMD3 is -72.4dB with two -15dBFS inputs at 83MHz and 87MHz. Figure 20.7.5 shows the SNR and SNDR versus the input amplitude and power breakdown. The measured dynamic range (DR) is 76.3dB. The modulator consumes 16.3mW from a 1.1V/1.5V supply, with 2mW drawn by the digital part, and 14.3mW by the analog part, where the QTZ only consumes ~1.4mW. Figure 20.7.6 summarizes the measured performance of this work and compares it with state-of-the-art CTDSMs having similar bandwidths. The proposed modulator achieves a Walden FOM_W of 23.4fJ/conv-step and a Schreier FOM_S of 170.5dB, revealing a CTDSM with excellent energy efficiency and a bandwidth ≥ 45MHz. The outstanding energy efficiency is derived mainly from the PSQ with 1-feedback DAC and simple clocking. Figure 20.7.7 exhibits the chip microphotograph with an active area of 0.019mm^2.

Acknowledgements:
This work was financially supported by the Macao Science & Technology Development Fund (FDCT) with Ref no: 077/2017/A2 and NSFC Grant 61604180.

References:
[1] B. Wu, et al., "A 24.7mW 45MHz-BW 75.3dB-SNDR SAR-assisted CT $\Delta\Sigma$ modulator with 2nd-order noise coupling in 65nm CMOS," *ISSCC*, pp. 270-271, Feb. 2016.
[2] Y. Dong, et al., "A 930mW 69dB-DR 465MHz-BW CT 1-2 MASH ADC in 28nm CMOS," *ISSCC*, pp. 278-279, Feb. 2016.
[3] C. Ho, et al., "A 4.5mW CT self-coupled $\Delta\Sigma$ modulator with 2.2MHz BW and 90.4dB SNDR using residual ELD compensation," *IEEE JSSC*, vol. 50, no. 12, pp. 2870-2879, Dec. 2015.
[4] S. Huang, et al., "A 125MHz-BW 71.9dB-SNDR VCO-based CT $\Delta\Sigma$ ADC with segmented phase-domain ELD compensation in 16nm CMOS," *ISSCC*, pp. 470-471, Feb. 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

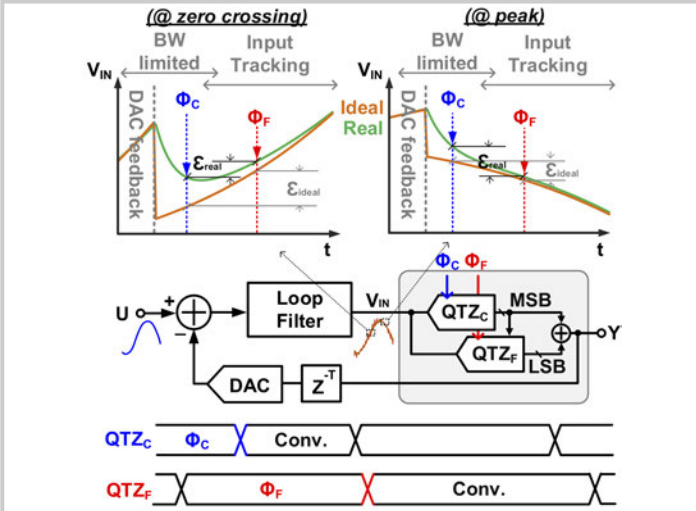


Figure 20.7.1: The conceptual block diagram and timing of the proposed CTDSM with preliminary sampling and quantization, including sampling error illustration (top).

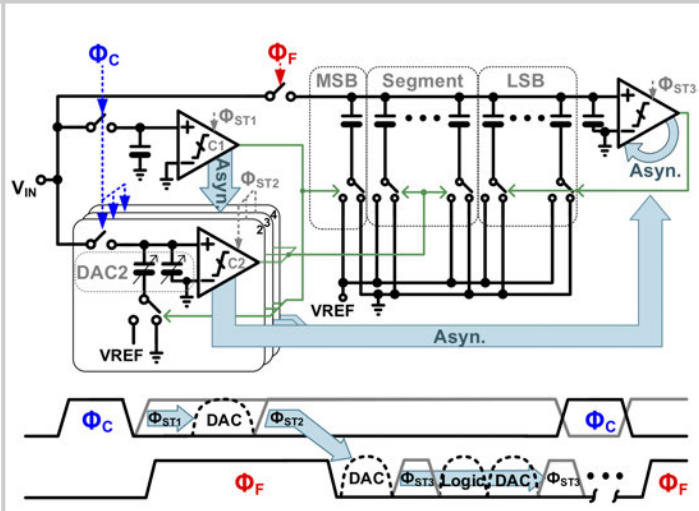


Figure 20.7.2: The simplified circuit schematic of the QTZ with asynchronous clocking path highlighted.

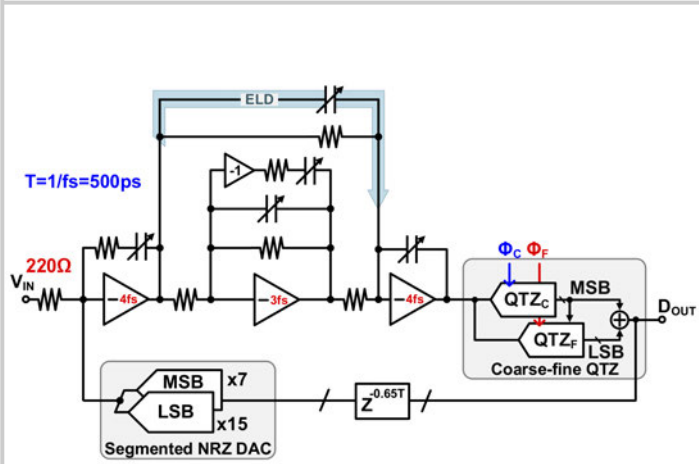


Figure 20.7.3: The block diagram of the modulator with the ELD path highlighted.

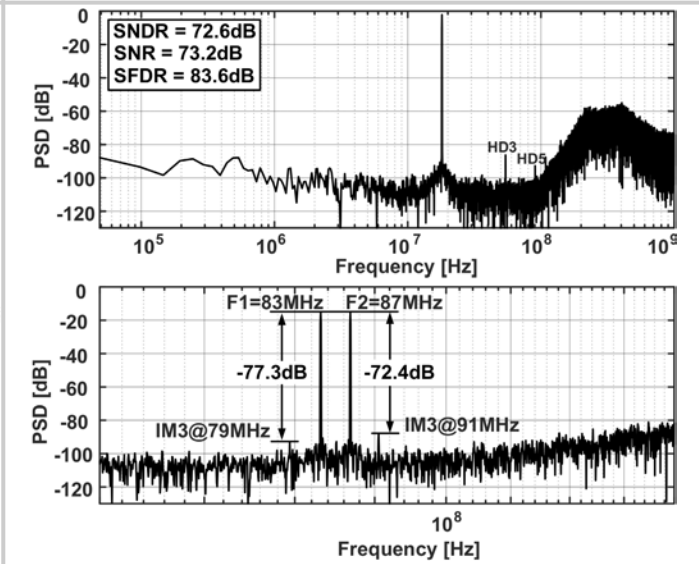


Figure 20.7.4: Measured 41k-point single-tone and two-tone output spectra.

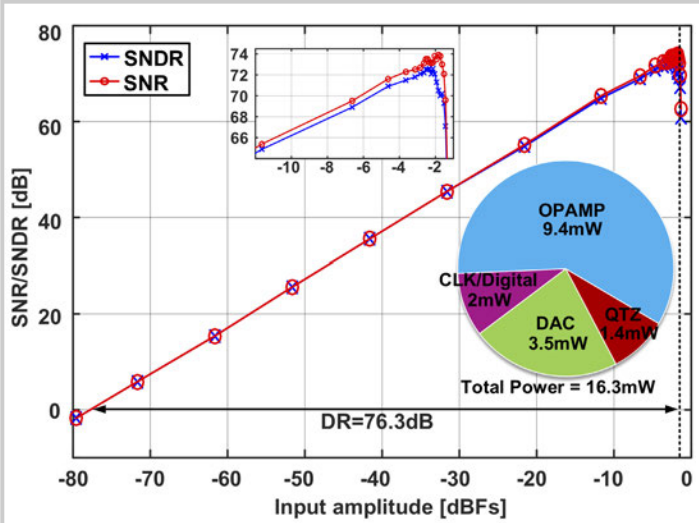
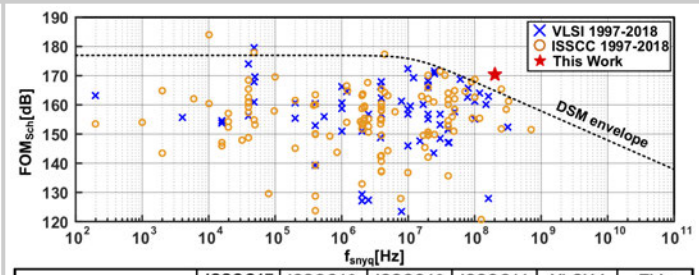


Figure 20.7.5: SNR/SNDR versus input amplitude (Fin = 18MHz) and power breakdown.



	ISSCC17 [4]	ISSCC16 [1]	ISSCC16 Wu	ISSCC11 Bolatkale	VLSI14 Ho	This work
Area (mm^2)	0.217	0.16	0.155	0.9	0.1	0.019
Technology (nm)	16	65	16	45	20	28
OSR	8.6	10	9	16	13.7	10
Fs (GHz)	2.15	0.9	2.88	4	2.184	2
Bandwidth (MHz)	125	45	160	125	80	100
Power (mW)	54	24.7	40	256	23	16.3
Peak SNDR (dB)	71.9	75.3	65.33	65	67.5	72.6
DR (dB)	74.8	82.5	72.1	70	73	76.3
FOMSch/SNDR (dB)	165.7	167.9	161.4	151.9	162.9	170.5
FOMSch/DR (dB)	168.4	175.1	168.1	156.9	168.4	174.2
FoMWa (fJ/conv.step)	67.2	57.7	82.8	704.7	74.2	23.4

Figure 20.7.6: Comparison with state-of-the-art ADCs and summary of measurement results.

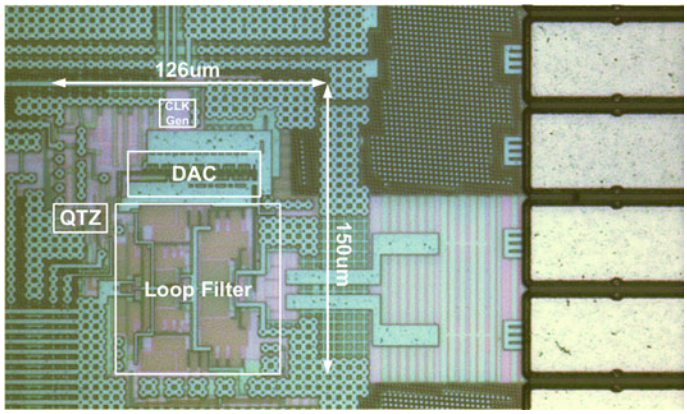


Figure 20.7.7: The chip micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 21 Overview: *4G/5G Transceivers*

WIRELESS SUBCOMMITTEE

Session Chair: ***Yiwu Tang***
Qualcomm Technologies, San Diego, CA

Associate Chair: ***Kyoo Hyun Lim***
FCI, Seongnam, Korea

Subcommittee Chair: Stefano Pellerano, Intel, Hillsboro, OR

Advancement from 4G to 5G transceivers in both sub-6GHz and mm-wave bands provides tremendous opportunities in high data throughput and low latency communications. This session includes advanced wireless technologies showing a 28GHz phased-array beamformer utilizing a neutralized bi-directional technique supporting dual-polarized MIMO, a 27-to-41GHz MIMO receiver with N-Input-N-Output using scalable cascadable autonomous array-based high-order spatial filters, a reconfigurable bidirectional 28/37/39GHz front-end supporting MIMO-TDD, Carrier Aggregation (CA) TDD and FDD/Full-Duplex with self-interference cancellation, an LTE-A transceiver with 4RX/2TX inter-band CA, 2-Carrier 4×4 MIMO and HPUE Capability, a 5G Sub-6GHz ZIF and mm-wave IF transceiver with MIMO and CA, a sub-6GHz 5G New Radio transceiver supporting EN-DC with 3.15Gb/s DL and 1.27Gb/s UL, and a receiver with mixed-signal circuit technique for cancellation of multiple modulated spurs in 4G/5G CA Transceivers.

8:30 AM

21.1 A 28GHz CMOS Phased-Array Beamformer Utilizing Neutralized Bi-Directional Technique Supporting Dual-Polarized MIMO for 5G NR

J. Pang, Tokyo Institute of Technology, Tokyo, Japan

In Paper 21.1, the Tokyo Institute of Technology and NEC Corporation present a 28GHz dual-polarized bi-directional beamformer achieving an SC data rate of 15Gb/s/pol. in 64QAM in a 1m OTA measurement. 2×2 DP-MIMO communication in 64QAM with 400MHz 5G NR signal is realized. A saturated EIRP of 45.6dBm can be supported by the 32H+32V array.

9:00 AM

21.2 A 27-to-41GHz MIMO Receiver with N-Input-N-Output Using Scalable Cascadable Autonomous Array-Based High-Order Spatial Filters for Instinctual Full-FoV Multi-Blocker/Signal Management

M-Y. Huang, Georgia Institute of Technology, Atlanta, GA

In Paper 21.2, the Georgia Institute of Technology reports a wideband 27-to-41GHz N-input-N-output MIMO RX with high spatial selectivity by employing several cascaded scalable array-based high-order Autonomous Spatial Filters (ASFs), which autonomously rejects multiple wideband modulated in-band blockers and receives desired signals with 3Gb/s 64QAM and 0.8Gb/s 256QAM over full FoV with μs response time.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

9:30 AM

21.3 A Reconfigurable Bidirectional 28/37/39GHz Front-End Supporting MIMO-TDD, Carrier Aggregation TDD and FDD/Full-Duplex with Self-Interference Cancellation in Digital and Fully Connected Hybrid Beamformers

S. Mondal, Carnegie Mellon University, Pittsburgh, PA

In Paper 21.3, Carnegie Mellon University presents a reconfigurable area-efficient bidirectional 28/37/39GHz front-end with self-interference cancellation in digital and power-efficient fully connected hybrid beamformers supporting MIMO and CA.

10:15 AM

21.4 An LTE-A Multimode Multiband RF Transceiver with 4RX/2TX Inter-Band Carrier Aggregation, 2-Carrier 4×4 MIMO with 256QAM and HPUE Capability in 28nm CMOS

C-C. Tang, MediaTek, Hsinchu, Taiwan

In Paper 21.4, MediaTek presents a 28nm CMOS LTE-A transceiver to support up to 4RX / 2TX inter-band CA and high-power user equipment (HPUE) as well as 2-carrier 4×4 MIMO with 256QAM to provide >2Gb/s downlink capability. The chip consumes 88mA battery current with 25mm^2 die area in a flip-chip chip-scale package.

10:45 AM

21.5 A 5G Sub-6GHz Zero-IF and mm-Wave IF Transceiver with MIMO and Carrier Aggregation

B. Jann, Intel, Hillsboro, OR

In Paper 21.5, Intel presents a 28nm bulk CMOS 4G/5G transceiver covering sub-6GHz bands in zero-IF while providing a 10.56GHz IF for mm-wave bands. It features an MPHY Gear 3 interface supporting MIMO and up to 800MHz bandwidth with carrier aggregation for a full 5G solution.

11:15 AM

21.6 A Sub-6GHz 5G New Radio RF Transceiver Supporting EN-DC with 3.15Gb/s DL and 1.27Gb/s UL in 14nm FinFET CMOS

J. Lee, Samsung Electronics, Hwaseong, Korea

In Paper 21.6, Samsung reports a 14nm FinFET CMOS single-chip RFIC that supports 2G, 3G, LTE, and SA/NSA sub-6GHz 5G NR with EN-DC in a die area of 38.4mm^2. The transceiver with 14 RX and 2 TX paths using on-chip DCXO and high-speed digital data interface with baseband SOC achieves a peak data throughput of 3.15Gb/s DL and 1.27Gb/s UL.

11:45 AM

21.7 A Mixed-Signal Circuit Technique for Cancellation of Multiple Modulated Spurs in 4G/5G Carrier-Aggregation Transceivers

S. Sadjina, Intel DMCE, Linz, Austria; Johannes Kepler University, Linz, Austria

In Paper 21.7, Johannes Kepler University and Intel present a 28nm CMOS receiver that can cancel multiple blocker-induced distortions without prior knowledge of their carrier frequencies by up to 33dB for blocker power levels of up to -10dBm at the LNA input.

21.1 A 28GHz CMOS Phased-Array Beamformer Utilizing Neutralized Bi-Directional Technique Supporting Dual-Polarized MIMO for 5G NR

Jian Pang[1], Zheng Li[1], Ryo Kubozoe[1], Xueting Luo[1], Rui Wu[1], Yun Wang[1], Dongwon You[1], Ashbir Aviat Fadila[1], Rattanan Saengchan[1], Takeshi Nakamura[1], Joshua Alvin[1], Daiki Matsumoto[1], Aravind Tharayil Narayanan[1], Bangan Liu[1], Junjun Qiu[1], Hanli Liu[1], Zheng Sun[1], Hongye Huang[1], Korkut Kaan Tokgoz[1], Keiichi Motoi[2], Naoki Oshima[2], Shinichi Hori[2], Kazuaki Kunihiro[2], Tomoya Kaneko[2], Atsushi Shirane[1], Kenichi Okada[1]

[1]Tokyo Institute of Technology, Tokyo, Japan; [2]NEC, Kawasaki, Japan

5G NR service will provide extremely-high-speed mobile data access using the millimeter-wave spectrum. To further boost the data-rate and spectrum efficiency, dual-polarized MIMO (DP-MIMO) will be introduced [1-4]. Separated beamformer arrays will be required for H and V polarizations. However, the increased free-space-path-loss for the 5G NR band n257 (26.5GHz to 29.5GHz) demands numerous elements to cover enough communication distance. Concerning the required considerable number of chips, an area-efficient design will be necessary for a DP-MIMO system targeting 5G NR FR2. This paper presents a 28GHz 4H+4V bi-directional beamformer chip supporting DP-MIMO in 65nm CMOS. The proposed neutralized bi-directional amplifier significantly reduces the required on-chip area. A bi-directional vector-summing phase shifter is also introduced. The measured RMS phase and gain errors are 0.4° and 0.2dB at 28GHz, respectively. The array module achieves a saturated EIRP of 45.6dBm/pol. at 0° scan with a 32H+32V array. In a 1m OTA measurement, a 4×4 sub-array module supports single-carrier data-rates of 15Gb/s and 6.4Gb/s per polarization in 64QAM and 256QAM, respectively. The measured 400MHz OFDMA TX-to-RX EVM for the 4×4 sub-array module at 0° scan is -34.4dB in 256QAM. 2×2 DP-MIMO communication with a 400MHz 5G NR channel bandwidth is also achieved with a 64-QAM EVM of 4.9% for the 4×4 sub-array module.

Figure 21.1.1 shows the system block diagram of the 4H+4V beamformer chip. In consideration of the large-sized phased-array and MIMO configuration in 5G NR, a bi-directional technique is utilized in this work for minimizing the on-chip area. The proposed array beamformer is capable of operating in TX and RX modes with the same bi-directional circuit chain. A single-element beamformer consists of a 2-stage PA-LNA, RF buffer, phase shifter, and isolation buffer. The 28GHz signal is distributed by a T-junction divider/combiner to each element in single-ended configuration. After the distribution, an isolation buffer is inserted for suppressing the influence of impedance variation caused by the phase tuning. Thus, independent phase tuning between each element can be realized. For minimizing the magnitude and phase errors between elements, a calibration circuit is integrated into the chip. The calibration paths are shared between H and V. The simulated cross-pol. isolation is higher than 39dB at 28GHz. The beamformer chip can be configured into H/V TX/RX calibration modes by the calibration select switches using quarter-wavelength TL. The redirected 28GHz signal is downconverted to a low frequency, and then the magnitude and phase detection is performed.

In this work, the bi-directional amplifiers are designed based on the proposed neutralized bi-directional core. Compared with the conventional switch-based bi-directional approach, the proposed bi-directional amplifier completely shares the inter-stage matching networks between the TX and the RX. Thus, the required on-chip area is further minimized. Figure 21.1.2 presents the circuit schematic of the proposed neutralized bi-directional core, which consists of two transistor pairs in cross-coupling connection. Transistors M1 and M2 are utilized for the TX mode, while M4 and M5 are utilized for the RX mode. The mode selection is realized by switching the tail bias (M3, M6). If the same transistor size is selected for M1-2 and M4-5, the C_{gd} of M1 and M2 will be neutralized by M4 and M5 in TX mode, while the C_{gd} of M4 and M5 will be neutralized by M1 and M2 in RX mode. The higher gain characteristics and the enhanced reverse isolation are realized by the proposed bi-directional amplifier, which contributes to low-noise and high-stability operation as well as an improved phase tuning isolation. Figure 21.1.2 also shows the circuit schematic of the bi-directional phase shifter. The active vector-summing architecture is adopted in this work due to its fine phase-tuning resolution with a compact size. The proposed phase shifter consists of two switchable polyphase filters (PPF) and I/Q bi-directional VGAs. The switchable PPF can be configured into normal PPF mode or adder mode by controlling the bypass switches in the capacitor paths. During the operation, one of the switchable PPFs will operate as a PPF while the other one will add up the I/Q signals. The I/Q VGAs are designed with two bi-directional cores with flipped input. For 5b phase shifting, the measured RMS phase and gain errors are 0.4° and 0.2dB at 28GHz.

Figure 21.1.3 demonstrates the circuit schematic of the proposed PA-LNA. Due to the different transistor size requirements for PA and LNA, an unbalanced bi-directional amplifier architecture is proposed. Extra capacitors C_{gd} are attached to the LNA transistor to compensate the additional C_{gd} from the PA transistor. As a result, neutralization can be realized in both operation modes. The PA and LNA share the antenna node in a switchless manner. A large capacitor in series with a small switching transistor is attached to the antenna node. By switching the transistor, the output matching network can be adapted to the required impedance for PA and LNA. The measured on-wafer single-path characteristics of the beamformer are also shown in Fig. 21.1.3. Within the frequency range of 26.5 to 29.5GHz, the beamformer achieves around 14dB gain in TX mode and 10dB gain in RX mode. The measured P_{sat} and P_{1dB} in TX mode are 15.1dBm and 11.3dBm at 28GHz. Regarding the RX, an NF of 4.2dB is achieved at 28GHz. The RX-mode output power and IM_3 are also measured and the RX SNDR is calculated with 100MHz signal bandwidth. The peak SNDR is 43.8dB at 28GHz. Regarding a large RX input, the SNDR can be further improved by decreasing the gain.

In this work, each sub-array module PCB has 4 chips on the front with a 4×4 dual-polarized array antenna on the back. Each antenna element has H and V ports for dual-polarized excitation, and each chip has 8 RF-signal ports and is connected to the 2×2 array of the 4×4. Multiple 4×4 sub-array module PCBs can be combined side by side to make a larger array such as 32×4. Figure 21.1.4 shows the measured V-pol. azimuth-plane beam pattern using two sub-array PCBs, which covers the beam angle from -50° to +50° with less than -10dBc sidelobe level. At 0° scan, the measured saturated EIRP for the 32H+32V array is 45.6dBm/pol.

Figure 21.1.4 also shows the measured on-wafer single-path transmitter-mode EVM and ACLR performances with 5G NR OFDMA signal. The EVM and ACLR are measured in QPSK, 16QAM, 64QAM, and 256QAM with 400MHz bandwidth against the output power. The proposed transmitter achieves 64-QAM (MCS19) and 256-QAM (MCS27) EVMs of -25.1dB and -32.5dB at the output power of 5.6dBm and 0.1dBm, respectively. The corresponding ACLRs are -28.9dBc in 64QAM and -33.6dBc in 256QAM at the above-mentioned output power.

Figure 21.1.5 shows the 1m OTA measurement summary at 0° scan for a 4×4 sub-array module. The SC-mode data-rates of 12Gb/s, 15Gb/s and 6.4Gb/s are achieved in 16QAM, 64QAM and 256QAM, respectively. Regarding the measurement with 5G NR OFDMA signals, the proposed module realizes 400MHz TX-to-RX EVMs of -34.6dB and -34.4dB in 64QAM (MCS19) and 256QAM (MCS27), respectively. The 2×2 DP-MIMO performance is also evaluated with the 5G NR uplink signals in QPSK, 16QAM, 64QAM and 256QAM. Simultaneously, two-stream input signals for the TX are generated from the Keysight AWG 8190A, while the RX output signals are analyzed by the Keysight digitizer M9703B along with the Keysight 89600 VSA software (2018 U_1.0 BETA5). 2×2 DP-MIMO communications with 400MHz are realized in QPSK, 16QAM, 64QAM with TX-to-RX EVMs of -26.7dB, -26.2dB and -26.2dB, respectively. Regarding the 256QAM, a TX-to-RX EVM of -31.1dB is achieved within 100MHz bandwidth.

Figure 21.1.6 compares this work with the state-of-the-art 28GHz dual-polarized phased-array transceivers [1-3]. The proposed beamformer chip reports a 2×2 DP-MIMO communication in 64QAM with a significantly reduced area.

Figure 21.1.7 shows the die micrograph. The proposed chip is fabricated in a standard 65nm CMOS process. The total chip area including pads is 12mm^2. The power consumption is 1.01W/pol. in TX mode and 0.46W/pol. in RX mode.

Acknowledgement:
This work is partially supported by the MIC/SCOPE #175003017, STAR and VDEC in collaboration with Cadence Design Systems, Inc., Synopsys, Inc., and Mentor Graphics, Inc.

References:
[1] J. Dunworth et al., "A 28GHz Bulk-CMOS Dual-Polarization Phased-Array Transceiver with 24 Channels for 5G User and Basestation Equipment," *ISSCC*, pp. 70-71, Feb. 2018.
[2] B. Sadhu et al., "A 28GHz 32-Element Phased-Array Transceiver IC with Concurrent Dual Polarized Beams and 1.4 Degree Beam-Steering Resolution for 5G Communication," *ISSCC*, pp.128-129, Feb. 2017.
[3] K. Kibaroglu et al., "A Dual-Polarized Dual-Beam 28 GHz Beamformer Chip Demonstrating a 24 Gb/s 64-QAM 2×2 MIMO Link," *IEEE RFIC*, pp.64-67, June 2018.
[4] H. Kim et al., "A 28-GHz CMOS Direct Conversion Transceiver With Packaged 2 × 4 Antenna Array for 5G Cellular System," *IEEE JSSC*, vol. 53, no. 5, pp.1245-1259, May 2018.
[5] J. Pang et al., "A 28GHz CMOS Phased-Array Transceiver Featuring Gain Invariance Based on LO Phase Shifting Architecture with 0.1-Degree Beam-Steering Resolution for 5G New Radio," *IEEE RFIC*, pp.56-59, June 2018.
[5] U. Kodak et al., "Bi-Directional Flip-Chip 28 GHz Phased-Array Core-Chip in 45nm CMOS SOI for High-Efficiency High-Linearity 5G Systems," *IEEE RFIC*, pp. 61-64, June 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

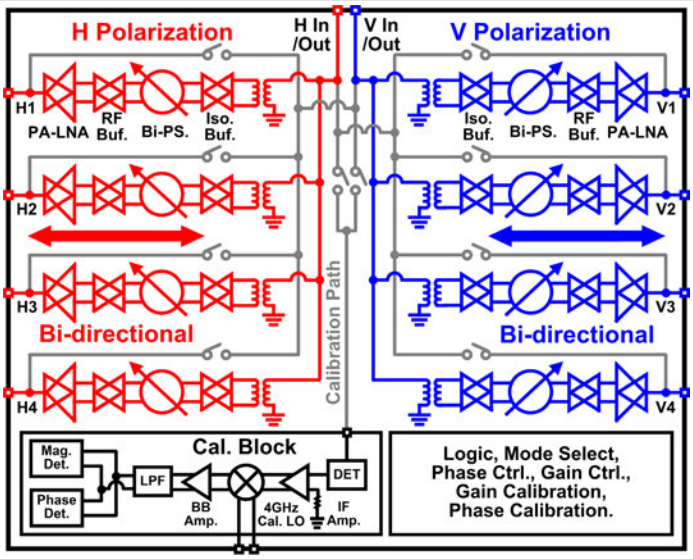


Figure 21.1.1: Block diagram of the proposed 4H+4V 28GHz beamformer chip.

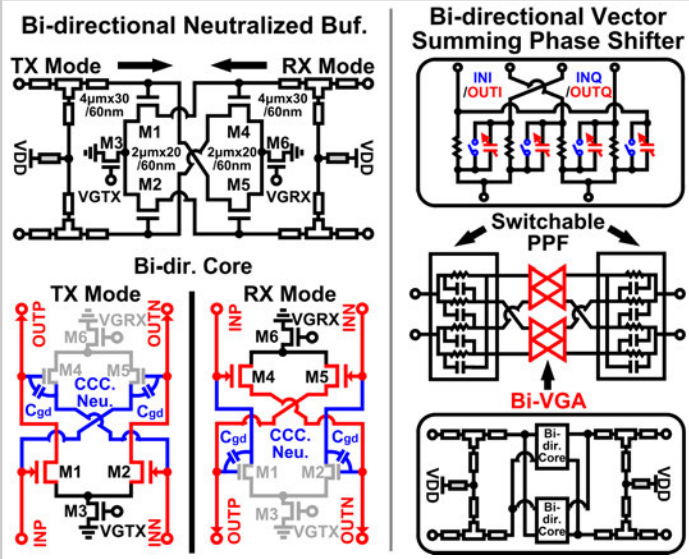


Figure 21.1.2: Proposed neutralized bi-directional core and phase shifter.

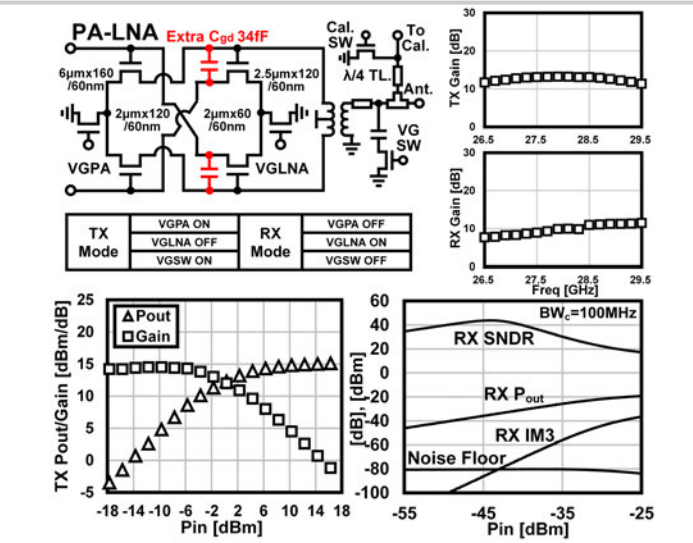


Figure 21.1.3: Proposed PA-LNA schematic and measured on-wafer V-pol. single-path TX/RX gain, TX Pout, RX SNDR.

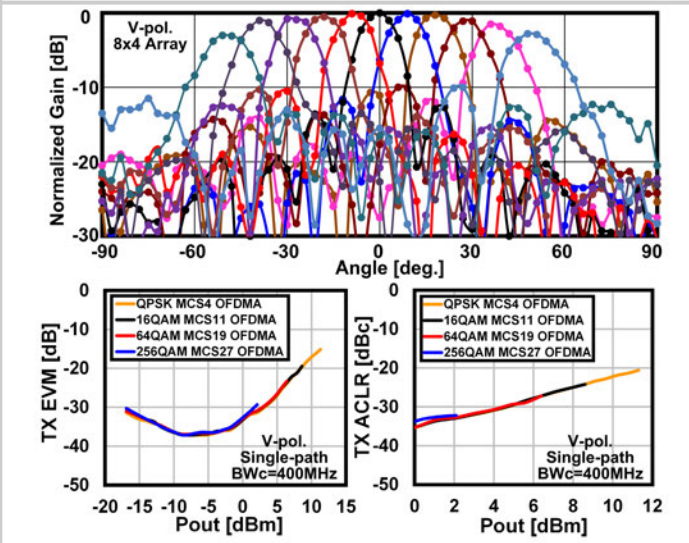
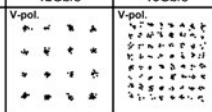
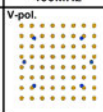
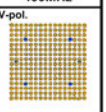
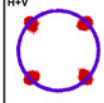
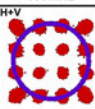
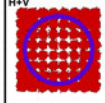
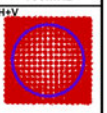


Figure 21.1.4: Measured V-pol. beam pattern for two 4×4 sub-array modules, single-path EVM, ACLR performances.

	Modulation	16QAM	64QAM	64QAM	256QAM
1-m OTA Measurement*	Mode	SC	SC	OFDMA	OFDMA
	Symbol rate/MCS**	3GSymbol/s	2.5GSymbol/s	5G NR MCS19	5G NR MCS27
	Data rate/BW_c	12Gb/s	15Gb/s	400MHz	400MHz
	Constellation**	V-pol. 	V-pol.	V-pol. 	V-pol. 
	TX EVM (RMS)***	-24.2dB	-24.2dB	-37.0dB (1.4%)	-36.8dB (1.4%)
	TX-to-RX EVM (RMS)****	-22.1dB	-22.6dB	-34.6dB (1.9%)	-34.4dB (1.9%)

	Modulation	QPSK	16QAM	64QAM	256QAM
2x2 DP-MIMO Measurement*	MCS**	MCS 4	MCS 10	MCS 19	MCS 27
	Uplink BW_c	400MHz	400MHz	400MHz	100MHz
	TX-to-RX Constellation for DP-MIMO *****	H+V 	H+V 	H+V 	H+V 
	TX-to-RX EVM (RMS) for DP-MIMO*****	-26.7dB (4.6%)	-26.2dB (4.9%)	-26.2dB (4.9%)	-31.1dB (2.8%)

*A 4x4 sub-array module is used.
**5G NR MCS index table 2 for PDSCH is used. (Table 5.1.3.1-2 in 3GPP TS 38.214 V15.2.0).
***Constellation, and TX EVM (referred to the RMS magnitude) are measured with an external downconverter.
****TX-to-RX EVM (referred to the RMS magnitude) is measured through TX and RX.
*****TX-to-RX constellation and EVM in DP-MIMO measurement are shown for H and V streams together.

Figure 21.1.5: OTA measurement summary for a 4×4 sub-array module.

		This work		[2] IBM ISSCC2017	[1] Qualcomm ISSCC2018	[3] UCSD RFIC 2018
Process		65nm CMOS		0.13μm SiGe	28nm CMOS	0.18μm SiGe
Carrier Freq.		28GHz (n257)		28GHz	28GHz (n257)	29GHz
P_{1dB}/path		11.3dBm		14.0dBm	12.0dBm	12dBm
P_{sat}/path		15.1dBm		16.4dBm	14.0dBm	N/A
EIRP@P_{sat}		32H+32V 45.6dBm/pol.		64H+64V 54dBm/pol.	8H+8V 35dBm/pol.	4H+4V 26.5dBm/pol. @P_{1dB}
RX NF		4.2dB		6.0dB	4.4~4.7dB	4.8dB
RMS Gain Error		0.2dB@28GHz		Gain var. <1.5dB	N/A	0.6dB
RMS Phase Error		0.4°@28GHz		<1°@28GHz	N/A	<4°
Integration/chip		4xH-Beamformer, 4xV-Beamformer		16xH-TRX, 16xV-TRX, IF, LO (w/o PLL)	24xTRX, IF, LO	4xH-Beamformer, 4xV-Beamformer
Calibration		Mag&Phase Cal.		N/A	N/A	N/A
P_{DC}/path		TX:252mW @11.3dBm/path RX:112mW		TX:319mW @16.4dBm/path RX:206mW	TX:119mW @11dBm/path RX:42mW	TX:220mW RX:150mW
Chip Area		12mm²		165.9mm²	27.8mm²	23.0mm²
SC Meas.	Constellation	4x4 array		N/A	N/A	2x2 array
		64QAM	256QAM			64QAM
	Max Data Rate	15Gb/s/pol.	6.4Gb/s/pol.	N/A	N/A	12Gb/s/pol.
5G NR Meas.	Modulation Supported	4x4 array		N/A	4x2 array	N/A
		QPSK, 16QAM 64QAM, 256QAM OFDMA			QPSK, 16QAM 64QAM OFDMA	
	2x2 DP-MIMO	4x4 array		N/A	N/A	N/A
		Uplink 64QAM MCS19 400MHz EVM=4.9%**				

* Estimated from the materials. ** Referred to the RMS magnitude of the constellation.

Figure 21.1.6: Performance comparison of 28GHz phased-array transceivers.

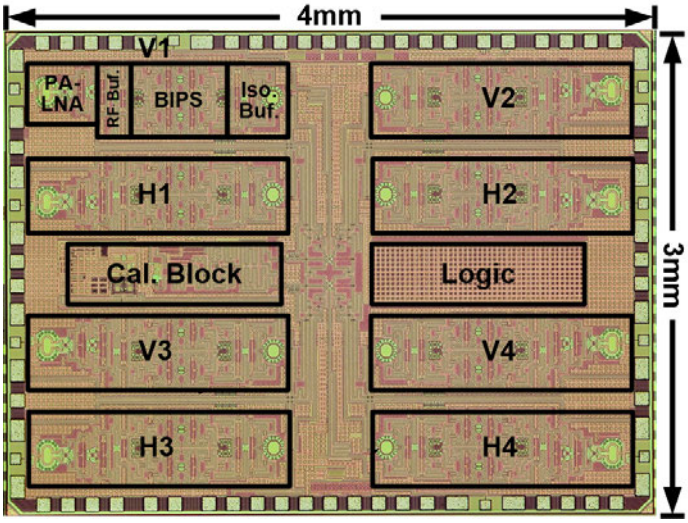


Figure 21.1.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

21.2 A 27-to-41GHz MIMO Receiver with N-Input-N-Output Using Scalable Cascadable Autonomous Array-Based High-Order Spatial Filters for Instinctual Full-FoV Multi-Blocker/Signal Management

Min-Yu Huang, Hua Wang

Georgia Institute of Technology, Atlanta, GA

Rapidly growing deployment of mm-wave links in commercial (e.g., 5G/automotive) and defense (e.g., fast-moving drones) applications often exposes the transceiver (TX/RX) front-ends in complex EM environments with multiple fast-changing yet unknown blockers (Angle-of-Arrival AoA/frequency/modulation). To preserve full field-of-view (FoV) and multi-beam/MIMO operations, digital arrays often skip FoV-limited front-end beamforming and rely on digital backends for spatial filtering [1]. However, the RXs and ADCs need high dynamic range (DR) to handle all the aperture information and avoid array saturation by strong signals/blockers that may hinder digital beamforming [2]. Therefore, to aid digital arrays and reduce RX/ADC DR, there is a critical need for agile spectral-spatial front-end filtering for instinctual blocker suppression and "power-equalizing" of desired signals.

In digital arrays, front-end spectral filtering, e.g., by mixers/filters, suppresses out-of-band blockers but not in-band spatial blockers. Most existing front-end spatial filters use analog beamformers (BFs) as spatial notches [3]. As a fundamental limitation, these approaches require the signal/blocker knowledge *a priori* for open-loop spatial notching or use digital backend for on-the-fly coefficient computation for the analog BFs, limiting them in dynamic use with unknown and fast-changing signals/blockers.

Conventional self-steering BFs automatically align the RX array towards the incident beam without backend computation. However, they either use Van-Atta-like simple reflectors, or employ coupled-oscillator/PLL autonomous BFs that are inherently narrowband and not scalable to large arrays due to loop stability [4]. Moreover, their intrinsic nonlinearity cannot support multi-beams or Gb/s complex modulations. Recently, [5] shows DLL-like autonomous BFs to suppress unknown strong signals/blockers and support Gb/s signals. However, their array size reduces by 2 after each beamforming stage and cannot support N-input-N-output MIMOs. Moreover, they only form 2-element spatial notch filters with limited spatial selectivity and ability to process multiple spatial signals/blockers.

We present a wideband 27-to-41GHz RX array for N-input-N-output MIMO systems (Fig. 21.2.1), which employs scalable cascadable array-based high-order Autonomous Spatial Filters (ASFs) as a "smart" spatial filter bank for instinctual multi-blocker/signal management to assist digital beamforming. Millimeter-wave wideband LNAs and passive mixers provide broadband front-end spectral filtering. The ASFs operate at IF, and each ASF employs an array-based phase-domain negative feedback for auto-beam-tracking and a feedforward path for spatial filtering. Identical ASF stages are cascaded to sequentially suppress multiple unknown spatial blockers or equalize multiple signals. An example 4-element broadband mm-wave RX array front-end includes a passive mixer spectral filter and three cascaded IF ASFs (Fig. 21.2.1).

Each IF ASF comprises phase shifters (PS), an array-based closed-loop autonomous BF, and feedforward signal combining/subtraction. The autonomous BF contains a power-aware phase detector (PD) [5] for feedback PS control-voltage generation. When multiple co-channel tones are received, the power-aware PD selects the strongest one, detects its adjacent-channel phase difference, and generates feedback PS control voltages, e.g., $-3\times V_{ctrl}$, $-1\times V_{ctrl}$, $+1\times V_{ctrl}$, and $+3\times V_{ctrl}$ for a uniform 1×4 array, so that PSs align the RX array to this tone (Fig. 21.2.2). Then, an auxiliary (Aux) path adds together N phase-aligned channels as a BF to accurately extract this strongest tone, which is then feedforward subtracted from the N-channel Main paths for spatial notching. The ASF equivalently rotates the array towards the strongest tone and spatially notches it by feedforward subtraction, so the array size and full-FoV are preserved. The ASF also acts only on RF power and needs no prior AoA/frequency/modulation knowledge or external BF controls. For the extracted tone in each ASF, although a one-step demodulation should verify whether it is signal or blocker, this consumes negligible time/resource compared to backend AoA computation in conventional arrays, ensuring rapid spatial filtering. Larger array sizes result in high-order ASF spatial filtering with sharpened selectivity (Fig. 21.2.2), while multiple cascaded ASFs can sequentially process more unknown co-channel signals/blockers. For instance, cascading 3 ASFs can create three independent spatial notches to suppress three similar-power blockers (case 1), suppress one strong blocker twice and reject one other medium-power blocker (case 2), or reject two blockers and beamform on one desired signal (case 3). The notch depth is also tunable by the Main/Aux-path gains if multiple co-channel signals require "power equalization".

A proof-of-concept mm-wave 4-input-4-output MIMO RX array with 3-stage ASFs in 45nm CMOS SOI was implemented (Fig. 21.2.7). All autonomous spatial notching or beamforming measurements were achieved only by on-chip ASFs with no external beamforming controls. The autonomous performance of one-stage ASF was first tested (Fig. 21.2.3), showing a 40-to-51dB spatial blocker notching or a flat normalized array factor >-0.2dB for desired signal beamforming over full-FoV. When two spatial co-channel tones are concurrently received, the power-aware PD can accurately discriminate the stronger one and achieve ≥28dB suppression, if the two tones have ≥1dB power difference. Response time of one ASF stage is approximately between 1.47 and 1.75μs over full-FoV, enabling fast beamforming for low-latency MIMOs.

Figure 21.2.3 shows the RX with cascaded 3-stage ASFs tested for different receiving cases. When three similar-power blockers are concurrently received (case 1), the three ASF stages will sequentially suppress them by three independent spatial notches with maximum 51dB cancellation. If one blocker is excessively strong (case 2), the first two ASFs will both track it and suppress it twice with a total 62dB deep spatial notch, and the 3rd ASF can reject one other blocker with a 46dB notch. When two blockers and one desired signal are simultaneously received (case 3), the two similar-power blockers are sequentially suppressed by two independent spatial notches with max 50dB rejection over full FoV. Compared to the fixed 2-element spatial filters in [5], the ASFs achieve 4-element spatial filters with sharpened selectivity that can be further improved for larger array sizes.

The RX achieves 27-to-41GHz wideband operation, 36dB conversion gain (CG) with 15dB controls, and 4.3-to-6.3dB NF with 3-stage ASFs on to support 5G multiband MIMOs (Fig. 21.2.4). When the 1st ASF stage is on, the measured in-band blocker P_{in1dB} achieves 15/17dB enhancement for 35/25dB CG. The measured in-band OIP3 of a desired signal achieves 18/24dB improvement for in-notch incidence at 35/25dB CG. Moreover, the RX array realizes "power equalization." When receiving three independent 64-QAM 100MSym/s signals (max/min=15dB) over 900MHz instantaneous bandwidth at different incidences, the RX array spatially equalizes and demodulates them with >26.2dB SNR, successfully relaxing the DR of downstream ADCs by 15dB.

The RX is also tested with concurrent multiple wideband modulated co-channel blockers (approximately between -35 and -40dBm) and one desired signal (-43dBm) with at least 50% frequency overlap (case 3) (Fig. 21.2.5). All the blockers/signals are independently modulated, and the blockers have various power levels (3 to 8dBc) to mimic practical MIMO scenarios. With the 3-stage ASFs, multiple wideband modulated blockers are autonomously rejected, and the desired signal is successfully demodulated for 3Gb/s 64QAM with -28.64dB EVM and 0.8Gb/s 256QAM with -32.7dB EVM. The desired signal is then swept for its incident angle. With a 4-element array, a clear spatial signal selection by the array-based ASFs is shown.

In summary (Fig. 21.2.6), this is the only N-Input-N-Output MIMO RX array in the state-of-the-art comparison table with array-based high-order scalable cascadable ASFs for autonomous suppression of multiple unknown blockers and desired signal beamforming supporting Gb/s complex modulations and μs low-latency.

Acknowledgements:
The authors would like to thank GlobalFoundries for chip fabrication.

References:
[1] A. F. Molisch et al., "Hybrid Beamforming for Massive MIMO: A Survey," *IEEE Commun. Mag.*, pp. 134-141, Sep. 2017.
[2] A. Agrawal et al., "A Concurrent Dual-Frequency/Angle-of-Incidence Spatio-Spectral Notch Filter Using Walsh Function Passive Sequence Mixers," *IEEE IMS*, pp. 1606-1609, June 2017.
[3] L. Zhang et al., "Arbitrary Analog/RF Spatial Filtering for Digital MIMO Receiver Arrays," *IEEE JSSC*, pp. 3392-3404, Dec. 2017.
[4] A. K. Gupta et al., "A Self-Steering Receiver Array Using Jointly Coupled Oscillators and Phased-Locked Loops," *IEEE TMTT*, pp. 631-644, Mar. 2014.
[5] M. Huang et al., "A 23-to-30GHz Hybrid Beam-Forming MIMO Receiver Array with Closed-Loop Multi-Stage Front-End Beam-Formers for Full-FoV Dynamic and Autonomous Unknown Signal Tracking and Blocker Rejection," *ISSCC*, pp. 68-70, Feb. 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

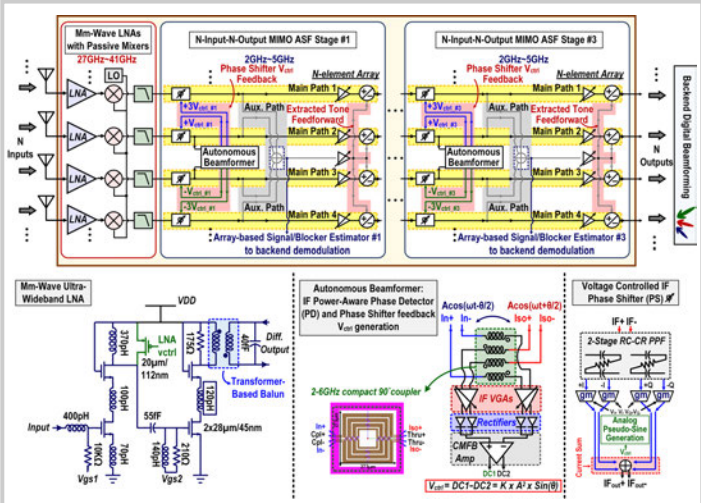


Figure 21.2.1: System architecture of the 27-to-41GHz N-Input-N-Output MIMO RX with scalable cascadable array-based high-order ASFs for instinctual full-FoV signal/blocker management.

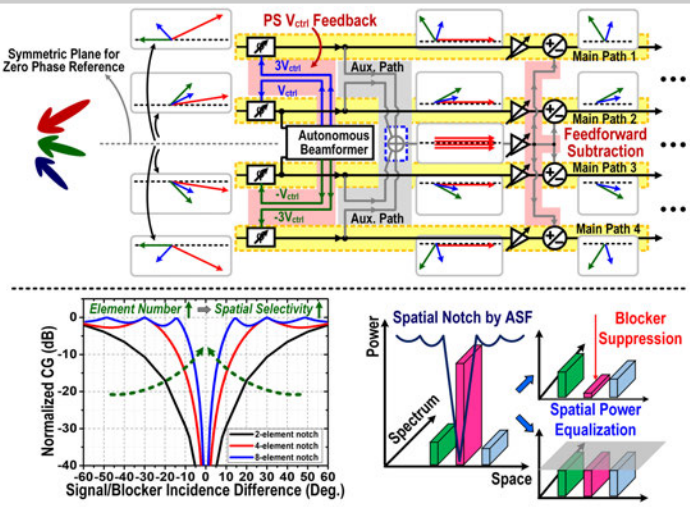


Figure 21.2.2: Conceptual diagram for the operation of the autonomous N-Input-N-Output array-based high-order ASF with a 4-element ASF example shown for spatial notching and desired signal power equalization.

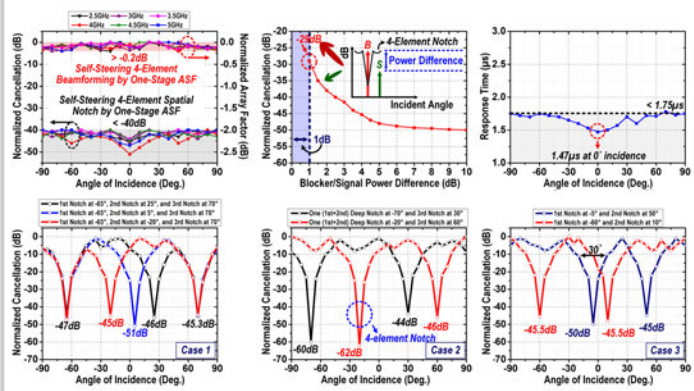


Figure 21.2.3: Measured single stage ASF performance over full FoV, including autonomous blocker rejection and desired signal beamforming, ASF autonomous discrimination of concurrently received multi-tones, and instinctual ASF response time, as well as 3-stage ASF performance in different receiving cases.

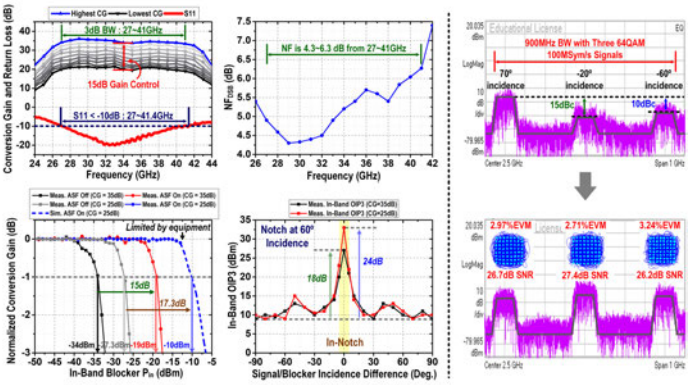


Figure 21.2.4: Measured wideband RX CG, S_{11}, NF, in-band blocker Pin_{1dB} with the 1st-stage ASF on/off, and in-band OIP3 versus signal/blocker incidence difference, as well as measured power equalization for three 64-QAM 100MSym/s desired signals.

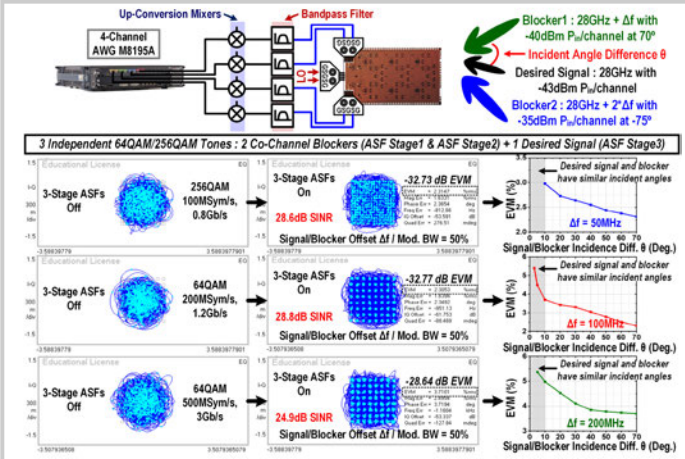


Figure 21.2.5: The RX with 3-stage array-based ASFs demonstrates autonomous rejection of multiple blockers and desired signal beamforming, when the blockers and desired signal have the same broadband modulation scheme and data-rate, showing successful demodulation of the desired signal after autonomous spatial suppression of multi-blockers by the 3-stage ASFs.

	Mm-Wave 5G Beamforming RX Array					Spatial Notch RX Array
	This Work	W. Roh Commun. Mag.'14	B. Sadhu JSSCC'17	J. Dunworth ISSCC'18	M. Huang ISSCC'18	L. Zhang JSSCC'17
Functionality	MIMO Array-based ASF	MIMO	MISO	MISO	MIMO SSA	MIMO ASF
Technology	45nm CMOS SOI	NR	130nm SiGe BiCMOS	28nm LPRF CMOS	130nm SiGe BiCMOS	65nm CMOS
Frequency (GHz)	27 – 41	28	28	25 – 30.5	22 – 30	0.1 – 3.1
# of Inputs / Outputs	4 / 4	32 / NR	32 / 1	12 / 1	8 / 2	4 / 4
Conversion Gain	36dB / element	NR	34dB / element	34dB / 4 element	33dB / element	43dB / element
$NF_{SSB,min}$ (dB)	4.3 – 6.3	NR	6[4]	4.4 – 4.8[4]	4.2 – 6.3	3.4 – 5.8
Beamforming	4-Element High-Order Array-based Closed Loop ASFs	Open Loop + Backend DSP	Open Loop + Backend DSP	Open Loop + Backend DSP	2-Element Closed Loop with SSA Front-End BFs	Open Loop + Backend DSP
Blocker Rejection	4-Element High-Order Array-based Closed Loop ASFs	No	No	No	2-Element Closed Loop with SSA Front-End BFs	Open Loop + Backend DSP
Spatial Order of Notch	4-element (Scalable)	No	No	No	2-element (Fixed)	4-element
Spatial Suppression 1- / 2- / 3 Notches (dB)	62 / 50 / 51	No	No	No	54 / 41 / No	56 / 48.6 / 41.8
>10dB Cancellation Spatial Notch Width (Deg.) Spatial Selectivity	27° – 32°	No	No	No	48° – 58°	~30°
Blocker Modulation Scheme	3Gb/s 64-QAM and 0.8Gb/s 256-QAM	No	No	No	3Gb/s 64-QAM and 0.8Gb/s 256-QAM	CW
Blocker/Signal Frequency Offset (Δf) / Modulation BW	< 50%	No	No	No	< 100%	>400%
# of Modulated Blockers Concurrently Received	1 or 2 or 3 (Scalable)	No	No	No	1 (Fixed)	2
Desired Signal Modulation Scheme	3Gb/s 64-QAM and 0.8Gb/s 256-QAM	2.1Gb/s 16QAM	NR	2.4Gb/s 64QAM	3Gb/s 64-QAM and 0.8Gb/s 256-QAM	2Mb/s QPSK
Signal EVM after Blocker Suppression	1.2Gb/s 64-QAM : 2.3%[1] 3Gb/s 64-QAM : 3.7%[1] 0.8Gb/s 256-QAM : 2.3%[1]	No	No	No	0.6Gb/s 64-QAM : 3.2%[2] 3Gb/s 64-QAM : 4.7%[2] 0.8Gb/s 256-QAM : 2.7%[2]	2Mb/s QPSK: 20.5%[3]
Response Time	1.47 μs	45 ms	NR	NR	< 1 μs	NR
Power Consumption / RX Element (mW)	70-85 : LNA/Mixer + 3 ASFs 55-65 : LNA/Mixer + 2 ASFs	NR	206	42.5	70	28.5 – 36.75
Area (mm^2)	23.4	NR	165.9[4]	29[4]	21.6	2.25

NR: Not reported.
The modulation test is based on
[1] 2 independent blockers + 1 desired signal and Δf / Modulation BW = 50%,
[2] 1 blocker + 1 desired signal and Δf / Modulation BW = 50%,
[3] 2 independent blockers + 1 desired signal and Δf / Modulation BW = 400%.
[4] It includes its T/R switch and transmitter design.

Figure 21.2.6: Comparison with the state-of-the-art RX array for mm-wave 5G beamforming or spatial notching.

21

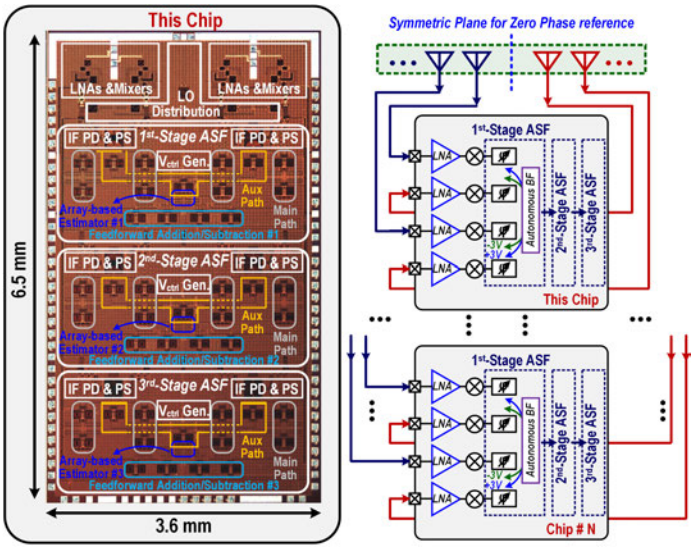


Figure 21.2.7: Die micrograph and its use in a large-scale MIMO system.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

21.3 A Reconfigurable Bidirectional 28/37/39GHz Front-End Supporting MIMO-TDD, Carrier Aggregation TDD and FDD/Full-Duplex with Self-Interference Cancellation in Digital and Fully Connected Hybrid Beamformers

Susnata Mondal, Rahul Singh, Jeyanandh Paramesh

Carnegie Mellon University, Pittsburgh, PA

This paper presents a 28/37/39GHz front-end applicable to beyond-5G wireless networks. It features three key contributions. First, a fully connected (FC) transmitter architecture is introduced for hybrid beamforming (HBF); it is shown that the power efficiency of FC-HBF is superior to the conventional partially connected (PC) HBF for a given modulation and antenna geometry. Second, a compact/low-cost circuit concept is introduced that supports bi-directional T/R operation concurrently at 28 and 37/39GHz, thereby facilitating multi-antenna carrier-aggregation (CA) or MIMO TDD with high antenna count. Third, a built-in mechanism for dual-band, *per antenna*, self-interference cancellation (SIC) is introduced, thanks to the FC-HBF architecture. The front-end is applicable to FDD or full-duplex (FD) multi-antenna systems; such SIC is not available in PC-HBF's. Also, the front-end is directly applicable to dual-band digital beamformers (DBF).

An FC-HBF receiver was introduced and compared to PC-HBF RX in [1], where the superior energy and spectral efficiencies of the FC-HBF RX for a given antenna geometry were noted. Figure 21.3.1 shows the proposed bi-directional FC T/R architecture and how it can be reconfigured between 3 modes and 10 sub-modes. Modes I and II are FC-TX and FC-RX TDD modes, respectively, and can be used for MIMO (in each band) or CA. Mode III applies to FDD/FD, where a hypothetical beamformer (see Fig. 21.3.6) comprising several of the proposed front-ends is partitioned into TX and RX parts. In the TX part, the complex-weights in one of the streams are left unused. In the RX part, one of the complex-weights is configured for reception, while the other complex-weight (which is available thanks to the FC architecture) is used to inject an independently weighted SIC signal for that element; thus, SIC can relax the dynamic range requirement of the following RX circuits. This SIC technique provides an elegant way to generate an independently weighted RF domain SIC signal from a single upconverted TX signal for each antenna in an FDD or FD beamformer.

The pros and cons of FC-TX versus PC-TX are discussed next with reference to Fig. 21.3.2. In a multi-antenna system, the EIRP scales quadratically with the number of antennas. Since each stream (N_S = #streams) accesses all N antennas in FC-TX and N/N_S antennas in PC-TX, FC-TX offers superior spatial combining. Consequently, each PA in FC-TX requires N_S times lower average output power in order to emit the same EIRP per stream as PC-TX. However, in an FC-TX, since independent signals from different streams are combined before the PA, the PAPR requirement of each PA is higher than the modulation-dependent PAPR of each stream. It can be shown that the effective PAPR that must be handled by each PA is N_S times higher in FC-TX than in PC-TX. Therefore, the peak power specification of each PA in PC-TX and FC-TX are equal. Based on these considerations, it is evident that for any PA that has lower DC consumption at back-off than at peak output power (i.e., non Class-A), the total power consumption of FC-TX is lower than PC-TX. Also, the margin of superiority of power efficiency of FC-TX depends on the PA's back-off efficiency, and improves with N_S. With ideal Class-B PA's, FC-TX has 0.64×/0.6× the power consumption of PC-TX for QPSK/1024-QAM for two streams while achieving identical spectral efficiency (Fig. 21.3.2). It is important to note that the PA's in a DBF have the same specifications as an FC-TX for the same number of streams.

The bi-directional, dual-band FC-HBF (Fig. 21.3.3) comprises several constituent circuit blocks designed for dual-band operation: a TX path PA (blue), an RX path LNA (red), a shared antenna interface network (AIN), a splitter-combiner structure, independent bi-directional complex-weight circuits for each MIMO stream and auxiliary on-chip test circuits. The AIN acts as a two-way transformer power-combiner in TX mode and as a matching network to a CG LNA in RX mode (Fig. 21.3.3 inset 1). Thin-gate MOS switches S_1-S_2 and thick-gate switches S_3-S_4 can be turned on or off to activate a high-power (HP) TX mode, a low-power (LP) TX mode or the RX mode (Fig. 21.3.3). Each PA slice comprises three driver stages followed by a Class-B output stage that drives the primary of the power-combiner. The Class-B stages employ a dual-band 2nd-harmonic shorting network (inset 2) to improve efficiency and P_{SAT}. In this network, the impedance Z_X is equivalent to an inductor having different values at the 2nd harmonic of the two frequency bands (56, 74GHz). Thus, the CM input impedance is equivalent to a series LC resonant network that concurrently acts as a short at the two frequencies. The signal path is designed so that additional T/R switches are not required over those used in the AIN. This includes the interface between the LNA-PA and the combiner-splitter, and also the bi-directional gain stages in the PGA and the complex-weights, designed as back-to-back transconductors, which are powered on or off to activate the TX or RX modes. The auxiliary circuits were included for testing purposes, but similar circuits (or from [2]) can be used to expand this front-end to a complete beamformer.

The 65nm CMOS prototype chip (Fig. 21.3.7) has core dimensions of 0.4mm×1.2mm. To achieve maximal compactness, passive structures were: (1) shared between TX and RX, (2) custom designed to minimize slice height. Designing for compactness is advantageous since it reduces interconnect lengths when such front-ends are combined into a larger beamformer. In the RX path, peak gain of 16/11dB and NF of 6.2/7.0dB were measured in the 28/37GHz band. The following metrics were measured in the HP TX mode: P_{SAT} = 15.8/16.8/16.7dBm, peak PAE = 20/21.6/22.2% and AM-PM distortion at P_{1dB} = -1.8°/-3.8°/2.7° at 28/37/39GHz. The same metrics for the LP mode are noted in Fig. 21.3.4.

Next, the front-end was characterized in one-stream PC-HBF, two-stream FC-HBF and multi-stream DBF configurations using 1-stream input, 2-stream input (both using on-chip complex-weight) and multi-stream input (using off-chip circuitry), respectively. In 28GHz TX mode, rms gain and phase error of the high-resolution complex-weight are 0.15/0.1dB and 1.45°/0.79° over all/constant-modulus weights (Fig. 21.3.5). The measured TX EVM at 28GHz is -26.3dB for 1.5Gb/s 64-QAM with average P_{OUT} of 7.1dBm. Average DC power consumption P_{DC} was measured in PC and FC TX configurations for the same peak P_{OUT} (Fig. 21.3.5, top right). For 2-stream operation, P_{DC} was measured for all QPSK symbols over all possible relative phases between the two streams. With Class-B unit PAs FC achieves 0.75×/0.64× lower P_{DC} than PC in HP/LP-HP mode (i.e, both unit PAs always on/both PAs turned on only for some symbol-phase combinations). For multi-stream DBF, Fig. 21.3.5 (bot-right) shows how the superiority of FC improves with N_S for both QPSK and 16QAM.

Figure 21.3.6 shows a conceptual FDD/FD beamforming node including TX (blue), RX (red) and *per-antenna* SIC (green). The RX signal passes through Z_{21} of the coupled-resonator. The SIC signal current is injected into the Z_{22} port of the coupled-resonator using a complex-weight cell (5b I/Q) followed by fine gain control PGA (5b, no sign) and fine phase control (<1°). Z_{21} and Z_{22} are designed to have similar frequency responses at the resonant frequencies (placed at the bands of interest) to enable dual-band SIC. To characterize SIC performance, the concept of Fig. 21.3.6 was simulated using external equipment to generate the RX signal, the SI and a reference version of the SI leakage. SIC greater than 36dB was measured over 360° SI signal phase variation and over 36dB range of SI power (Fig. 21.3.6).

Next, FDD operation was characterized using a desired 37GHz RX tone and a two-tone interferer near 28GHz. SIC weight settings were set once for maximum cancellation at band center. The two-tone spacing was swept to characterize SIC over BW. 28dB SIC was achieved for 0.4GHz RF BW (Fig. 21.3.6). FD mode operation was then demonstrated at 28GHz using a desired RX tone and a modulated SI; 26dB SIC was achieved over 500MHz RF BW of the interferer. Benchmarking against [1-4] is shown in Fig. 21.3.7. Despite having only a single element, our prototype showcases all the system concepts and configuration modes shown in Fig. 21.3.1, and can be directly incorporated in digital and hybrid beamformers, as demonstrated by our extensive measurements.

Acknowledgements:
This work was supported in part by the National Science Foundation under grants CCF-1314876 and ECCS-1343324.

References:
[1] S. Mondal et al., "A 25–30 GHz Fully-Connected Hybrid Beamforming Receiver for MIMO Communication," *IEEE JSSC,* vol. 53, no. 5, pp. 1275-1287, May 2018.
[2] S. Mondal et al., "A Reconfigurable 28/37GHz Hybrid-Beamforming MIMO Receiver with Inter-Band Carrier Aggregation and RF-Domain LMS Weight Adaptation," *ISSCC,* pp. 72-74, Feb. 2018.
[3] M. Vigilante et al.,"A Wideband Class-AB Power Amplifier With 29–57-GHz AM–PM Compensation in 0.9-V 28-nm Bulk CMOS," *IEEE JSSC,* vol. 53, no. 5, pp. 1288-1301, May 2018.
[4] T. Dinc et al., "A 60 GHz CMOS Full-Duplex Transceiver and Link with Polarization-Based Antenna and RF Cancellation," *IEEE JSSC,* vol. 51, no. 5, pp. 1125-1140, May 2016.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

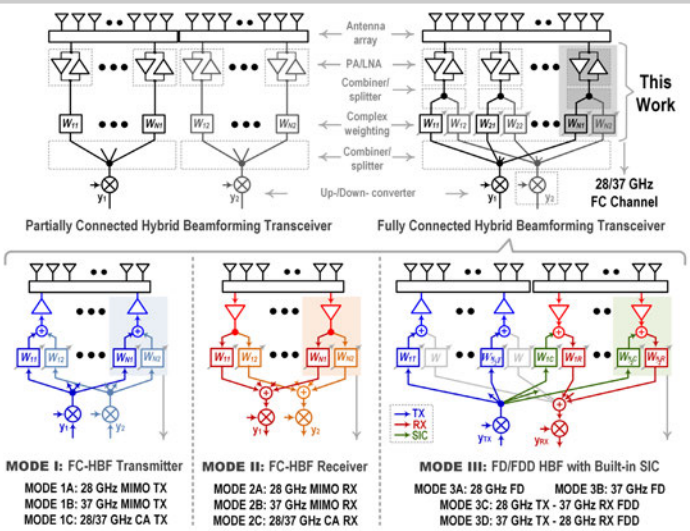


Figure 21.3.1: Partially/fully connected HBF TRX (top); FC-HBF supports fully connected transit/receive mode and FDD/full-duplex mode with SIC (bottom).

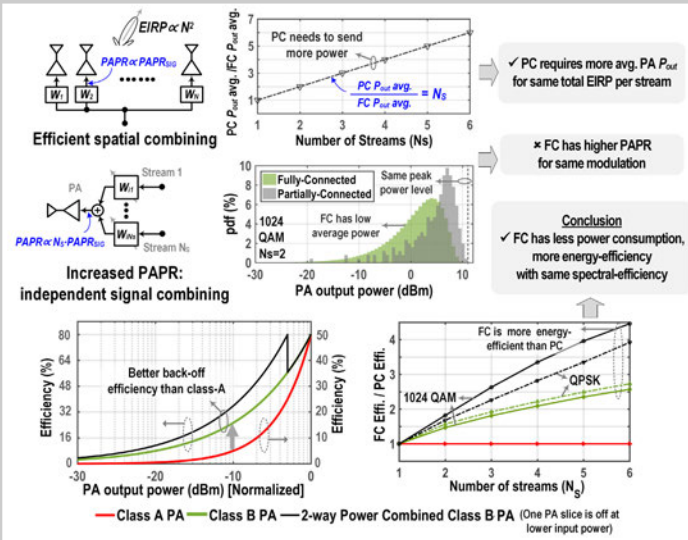


Figure 21.3.2: Partially connected versus fully connected HBF TX: Increased PAPR (top); Improvement for given PA characteristics (bottom).

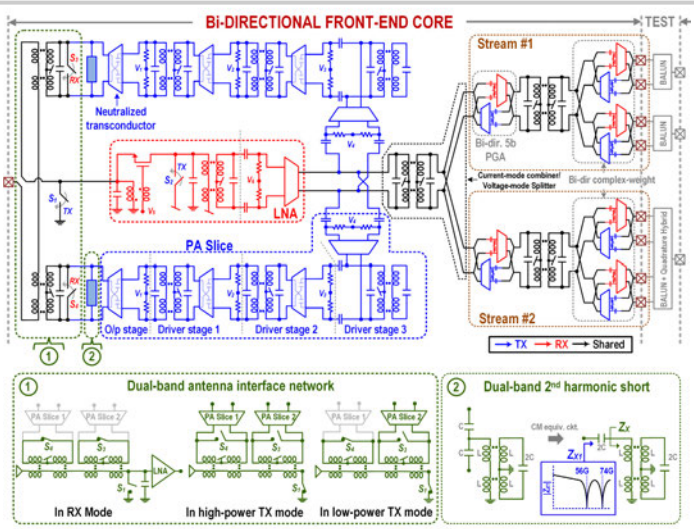


Figure 21.3.3: 28/37GHz bidirectional-TRX FC-HBF channel (top); Dual-band antenna interface switch and second-harmonic short design (bottom).

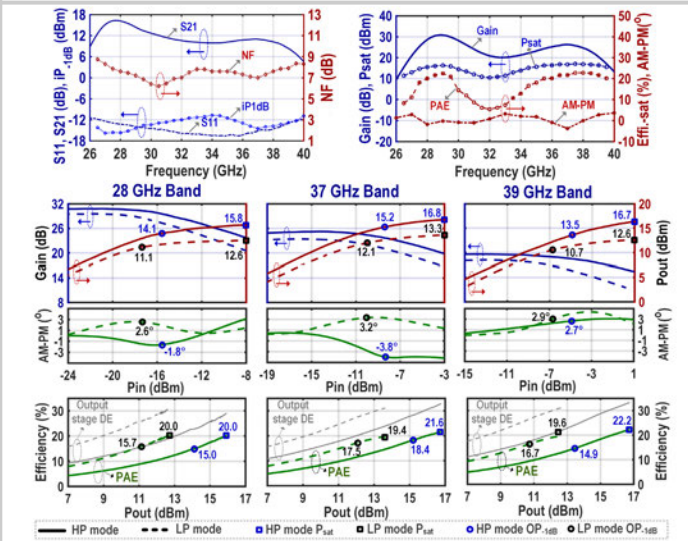


Figure 21.3.4: Receiver and transmitter characterization over frequency (top); 28/37/39GHz-band large-signal characterization of PA (bottom).

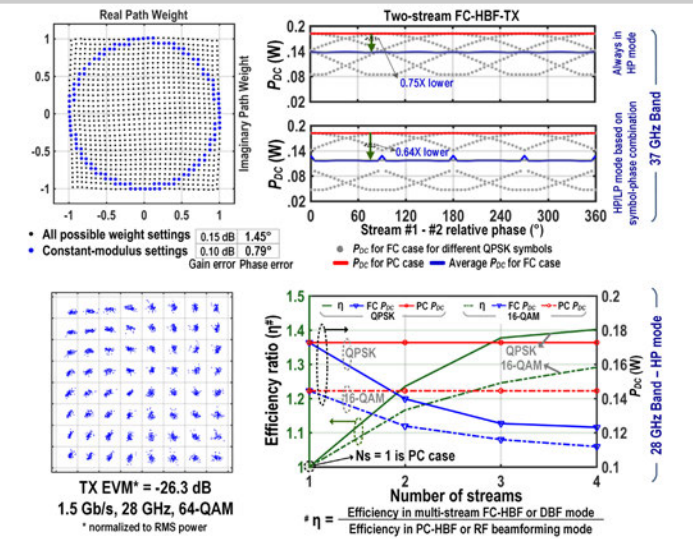


Figure 21.3.5: TX path complex-weight and EVM (left); Power consumption (P_{DC}) in PC-HBF, two-stream FC-HBF and multi-stream FC-HBF/DBF mode (right).

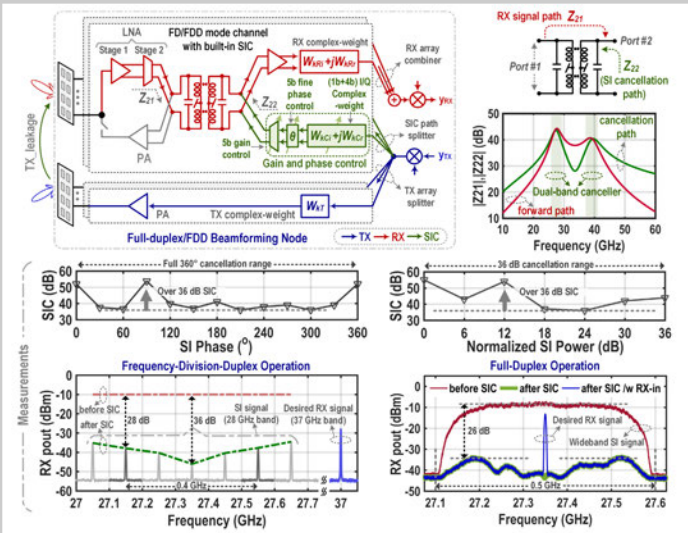
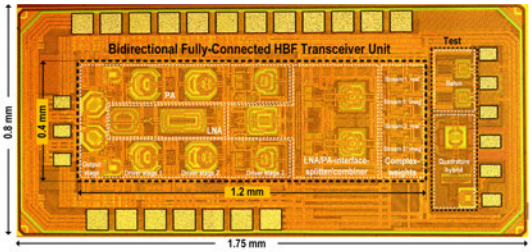


Figure 21.3.6: FDD/FD mode dual-band per-element SIC in FC-HBF TRX (top); SIC vs. phase and power of self-interferer, FDD/FD mode cancellation (bottom).



		This Work			Dunworth ISSCC'18	Mondal ISSCC'18		Sadhu ISSCC'17	Kim RFIC'17	Kibaroglu RFIC'17	Mondal RFIC'17	Vigilante RFIC'17			Hu ISSCC'17			Dinc JSSC'16
	Technology	65nm CMOS			28nm CMOS	65nm CMOS		130nm SiGe	28nm CMOS	180nm SiGe	65nm CMOS	28nm CMOS			130nm SiGe			45nm SOI
RX	Freq. (GHz)	28	37	39	25-30.5	28	37	28	28	28-32	25-30	30	40	50	28	37	39	57-66
RX	RX Gain (dB)	16.1	10.9	8.3	24	33[1]	26.5[1]	34	39	20	34[1]	-	-	-	-	-	-	40
RX	NF_{min} (dB)	6.2	7.0	7.9	4.4	5.7[1]	8.5[1]	-	6.7[1]	4.6	7.3[1]	-	-	-	-	-	-	4
RX	S11 (dB)	-13.2	-15.1	-12.8	<-10	<-10	<-15	-	-	<-10	<-8	-	-	-	-	-	-	<-7
RX	IP_{1dB} (dBm)	-15.7	-14.5	-12.8	-	-30	-23	-22.5	-	-22	-29	-	-	-	-	-	-	-32
TX	TX Gain (dB)	28.5	26.2	20.3	34-44	-	-	24-32	-	14	-	19[1]	20.8	19[1]	18.2	17.1	16.6	-
TX	P_{SAT} (dBm)	15.8	16.8	16.7	14	-	-	16	10.5	<13	-	16.6	15.9	15.1	16.8	17.1	17	11-15
TX	PAE@ P_{SAT} (%)	20.0%	21.6%	22.2%	-	-	-	-	-	-	-	24.2	18.4	14.9	20.3	22.6	21.4	15.3
TX	P_{1dB} (dBm)	14.1	15.2	13.5	12	-	-	13.5	9.5	<11.7	-	13.4	11.1	10.9	15.2	15.5	15.4	-
TX	PAE@ P_{1dB} (%)	13.0%	18.8%	14.9%	-	-	-	-	-	-	-	12.6	7.5	7.0	19.5	21.6	20.7	-
TX	AM-PM@ P_{1dB} (°)	-1.8[*]	-3.8[*]	2.7[*]	-	-	-	-	-	-	-	-0.7	-0.26	0.45	10	1.3	3.9	-
FD	SIC_{FDM} (dB)	36 dB			-	-		-	-	-	-	-			-			21[1]
FD	SIC BW (GHz)	0.5 GHz (26 dB SIC)			-	-		-	-	-	-	-			-			1
	Power (mW)*	37.6 (RX), 116.2 (TX-DC)			42[1](R),90[1](T)	52.5[1] (RX)		206[1](R),287[1](T)	50[1](R),85[1](T)	130(R),200(T)	27.5[1] (RX)	170 (PA-DC)			65 (PA-DC)			111(R),206(T)
	Area (mm^2)*	0.48			1.16[2]	0.46[2]		2.59[1]	0.91[1]	2.9	0.32[2]	0.16			1.735			4.42
	Functionalities	Multi-band bi-directional FC-HBF T/R front-end + FD/FDD mode SIC for multi-antenna systems			T/R RF Beamformer	Multi-band FC-HBF RX Beamformer		T/R RF Beamformer	T/R RF Beamformer	T/R RF Beamformer	FC-HBF RX Beamformer	Multi-band Class-AB PA			Multi-band Doherty PA			SIC in single-antenna system

* Single channel [1]Estimated [2]RF SIC Including [1]down-conversion [2]down-conversion and partial LO generation [3]down-convertion and full LO generation

Figure 21.3.7: Die micrograph (top); Comparison against state-of-the-art beamformers, PAs and single-antenna SIC (bottom).

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

21.4 An LTE-A Multimode Multiband RF Transceiver with 4RX/2TX Inter-Band Carrier Aggregation, 2-Carrier 4×4 MIMO with 256QAM and HPUE Capability in 28nm CMOS

Chih-Chun Tang[1], Yi-Bin Lee[1], Chih-Hao Eric Sun[1], Cheng-Chieh Lin[1], Jin-Siang Syu[1], Min-Hua Wu[1], YangChuan Chen[2], Tzu-Chan Chueh[1], Carl Bryant[3], Manel Collados[3], Mohammed Hassan[3], Joao Ramos[3], Yu-Lin Hsieh[1], Hsin-Hung Chen[1], Xiaochuan Guo[4], Hsinhua Chen[4], Changhua Cao[4], Daniel Li[4], Jon Strange[3], Caiyi Wang[4], Guang-Kaai Dehng[1]

[1]MediaTek, Hsinchu, Taiwan;
[2]MediaTek, San Diego, CA
[3]MediaTek, Kent, United Kingdom
[4]MediaTek, Austin, TX

With the increasing popularity of social media, artificial intelligence (AI) and mobile computing, higher data-rate wireless communication is required. In the existing limited 4G spectrum, 3GPP standards have defined several methods to increase the data-rate, e.g. carrier aggregation (CA), high-order modulation (HOM) and MIMO technique; however this comes at the cost of extra power consumption and potential sensitivity degradation due to the spurious components generated by CA. In addition, high-power user equipment (HPUE) extends the coverage range and uplink throughput, but introduces more stringent linearity requirements for transmitter design. In this work, we present a 28nm CMOS LTE-A transceiver, capable of supporting up to 4 inter-band downlink (or 2-carrier 4×4 downlink MIMO) and 2 inter-band uplink CA concurrently while adopting ADPLLs for RX and TX with 256-QAM capability. The techniques adopted to achieve these features are described in this paper.

The RX block diagram is shown in Fig. 21.4.1. Unlike the conventional single-carrier RX architecture, the LNAs are separated into 3 groups: LB (450 to 960MHz), MHB (1.4 to 2.7GHz) and UHB/LAA (3.4 to 6GHz) to maximize the isolation among CA cases with TX harmonic leakage directly landing on the RX channel, like B5 desensitization by 2nd-harmonic downconversion of B3 TX leakage mentioned in [1]. To mitigate CA-related spurs due to the digital-controlled oscillator (DCO)-to-LNA coupling in the highly integrated inter-band CA (ICA) receiver, multiple divider options are implemented in each receiver synthesizer (SRX) path to avoid intermodulation of TX leakage signal and DCO frequency as well as DCO frequency directly landing on the RX channel. In addition, a narrowband notch filter in the modem can help remove residual spurs caused by LO harmonics, and finally, a low-IF configuration can also be used to place spurs outside of the signal bandwidth to avoid RX desensitization [2]. All the LNA groups are connected to multiple mixers, not only to support most CA combinations, but also provide the additional flexibility on CA downconversion path selection to further improve isolation.

In a single band 4×4 MIMO configuration, 2 primary RX (PRX) LNAs and 2 diversity RX (DRX) LNAs are connected to the same mixer group driven by the same synthesizer/LO path to achieve high data-rates with reduced power consumption. Take UHB/LAA + MHB 2 ICA 4×4 MIMO as an example. 4 UHB/LAA LNAs can all be connected to the mixer group driven by SRX1 (or SRX4), and the mixer outputs are connected to PRX1 / DRX1 / PRX4 / DRX4 baseband chains. Meanwhile, 4 MHB LNAs can all be connected to the mixer group driven by SRX2 (or SRX3), and the mixer outputs are connected to PRX2 / DRX2 / PRX3 / DRX3 baseband chains. The bandwidth of each baseband chain is 60MHz. All 8 downconverted 2-ICA 4×4 MIMO baseband signals are passed to the modem chip. MHB + MHB 2-ICA 4×4 MIMO can also be supported in a similar fashion.

Figure 21.4.2 shows the TX block diagram. The I/Q baseband signals are filtered via the baseband filters (BBF) implemented as a 2nd-order Butterworth Multi-Feedback (MFB) topology driving a 25%-duty-cycle passive upconverter through a passive RC filter. The LC bandpass filter reduces the higher order mixer harmonic products and also implements impedance transformation such that the voltage swing across the mixer switches is reduced compared with a conventional passive modulator [3]. This improves the mixer linearity, which together with the reduction in harmonics entering the RF PGA, significantly reduces the TX counter-intermodulation (CIM) products. Additional linearization of the mixer is achieved by a bootstrap resistor [4] with programmable voltage offset. The RF PGA is implemented using a programmable dual bias scheme [3] that allows linearity (ACLR/CIM) to be flexibly exchanged against efficiency depending on required operating mode and band. RF PGA biasing always presents a tradeoff between bias DC current consumption, RF output-power variation, and linearity variation due to device matching limitations. To overcome this, a one-time calibration scheme replicates the bias condition of a larger unit element onto a smaller element (via a SAR loop) saving current in the biasing circuitry without reducing bias accuracy. TX gain control is implemented through a combination of segmenting the baseband into 4 slices (-10dB), -50dB implemented at RF with the remainder (-18dB) implemented in the Baseband Modem SOC. This together with the use of adaptive power-supply switching of the RF PGA achieves significant DC current-consumption savings as the TX RF power is reduced.

Figure 21.4.3 shows the all-digital phase-locked-loop (ADPLL) synthesizer architecture. A multi-modulus fractional divider (MMD) divides the DCO frequency down to the reference frequency. The MMD randomizes TDC input signal, and greatly relaxes the TDC gain and linearity requirements. One major MMD drawback is large quantization noise. However, the MMD quantization noise is actually a deterministic dithering signal and can be subtracted from the TDC output. The TDC has two 31-stage ring oscillators coupled together. Compared to a single ring, a coupled ring oscillator is faster and also provides differential inputs to the latches. The TDC uses a dynamic-element-matching (DEM) technique to alleviate TDC unit mismatch. The digital loop filter is a 5th-order IIR with adjustable bandwidth from 50kHz to 720kHz. Prior to the DCOs, the loop filter output is thermometer coded and split into 3 parts: 224 integer units, 5 fractional units and 40 dynamic-element-matching (DEM) units for GMSK frequency modulation. 2nd-order $\Delta\Sigma$ modulation (DCO DSM) is used to push the DCO tracking unit quantization noise to out-of-band. The DCO DSM frequency is programmable and chosen depending on the out-of-band phase-noise requirement. In addition, a hybrid digital-analog temperature control loop (TCL) keeps the DCO locked for more than 110°C ambient temperature change. Slow temperature drift is detected with a digital integrator at the loop filter output. The DCO frequency remains locked using an auxiliary analog varactor controlled by the digital integrator through a DAC. To achieve good integrated phase noise (IPN) in LAA bands, the DCO with coupled dual-coil inductor is designed to have 3dB phase-noise improvement compared to traditional single-coil design and the measured 12GHz DCO PN at 100kHz offset is -94dBc/Hz. The maximum in-band fractional spurs are below -57dBc across all channels including LAA band. The measurement results of IPN are 0.3° and 0.35° for B7 and LAA bands, respectively, which are sufficient for 256QAM. The GMSK frequency modulation can achieve -66.5dBc spectrum mask at 400kHz offset and 0.65° RMS phase error in the PCS band.

To lower power consumption, most transceiver circuits are connected to a 1.2V buck-converter output. The measured WCDMA B1 talking battery current consumption is 22mA. LTE B3 consumes 34mA and LTE 4-ICA RX / 2-ICA TX (B1+B3+B5+B7 downlink with B3+B7 uplink) consumes 98mA, with 0dBm TX output power. B3+B7 2-CA 4×4 MIMO consumes 88mA. This receiver achieves 2.4dB NF and 55dBm IIP2. As shown in Fig. 21.4.4, the TX CIM3 1RB performance for B13 (the most challenging LTE band for CIM3) and B41 HPUE is -92.3dBc/6.25kHz and -64.5dBc/1MHz respectively, which is more than 10dB better than the 3GPP requirement. Figure 21.4.5 shows the performance summary of this work. In addition, the sensitivity performance of B3 NCCA, shown in Fig. 21.4.6, is close to non-CA configuration. This paper demonstrates a highly integrated RF transceiver not only supporting the 2-carrier ICA 4×4 downlink MIMO operations with 256QAM but also with small die size and low power consumption. Figure 21.4.7 shows the die micrograph of this work. The die size is 24.9mm^2 in a 28nm CMOS process and it is housed in a flip-chip chip-scale package (FCCSP).

References:

[1] Tzung-Han Wu et al., " A 40nm 4-downlink and 2-uplink RF Transceiver Supporting LTE-Advance Carrier Aggregation," *IEEE RFIC*, pp. 316-319, June 2018.
[2] C.-S. Chiu et al., "A 40nm Low-power Transceiver for LTE-A Carrier Aggregation," *ISSCC*, pp.130-131, Feb. 2017.
[3] Y. Chen et al., "A Wideband Transmitter for LTE-A HPUE Using CIM3 Cancellation," *IEEE RFIC*, pp. 312-315, 2018.
[4] F. Tillman et al, "A Bootstrapping Technique to Improve the Linearity of CMOS Passive Mixers", *IEEE Symp. VLSI Circuits*, pp. 221-222, 2003.
[5] Qing Liu et al., "1.4-to-2.7GHz High-Efficiency RF Transmitter with an Automatic 3FLO-Suppression Tracking-Notch-Filter Mixer Supporting HPUE in 14nm FinFET CMOS," *ISSCC*, pp.172-174, Feb. 2018.
[6] M. Mikhemar et al., "A Rel-12 2G/3G/LTE-Advanced 3CC Cellular Receiver," *IEEE JSSC*, vol. 51, no.5, pp. 1066-1079, May 2016.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

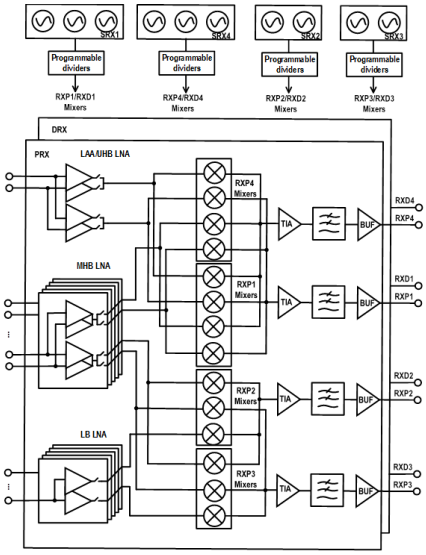


Figure 21.4.1: Receiver (RX) block diagram.

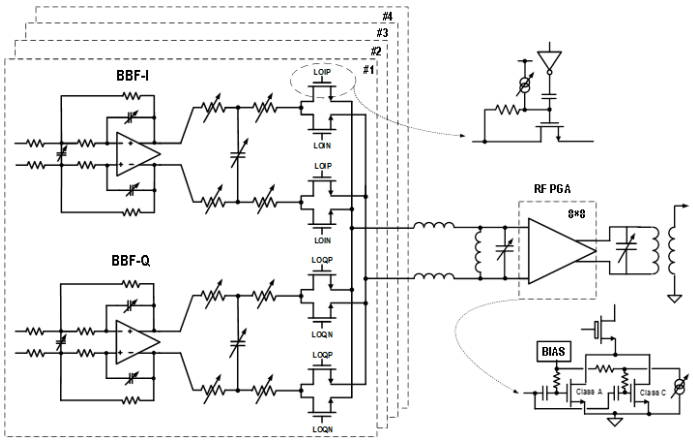


Figure 21.4.2: Transmitter (TX) block diagram.

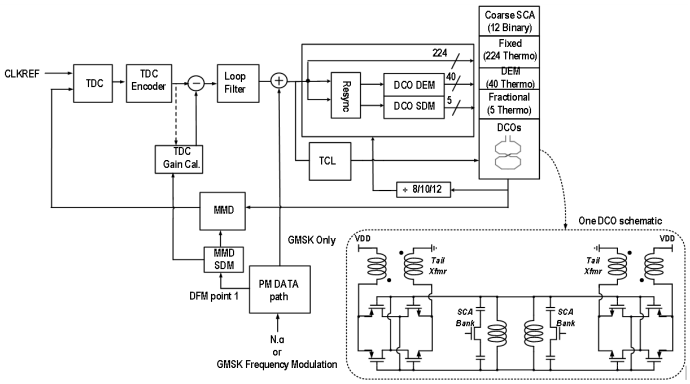


Figure 21.4.3: All-digital phase-locked loop (ADPLL) synthesizer architecture.

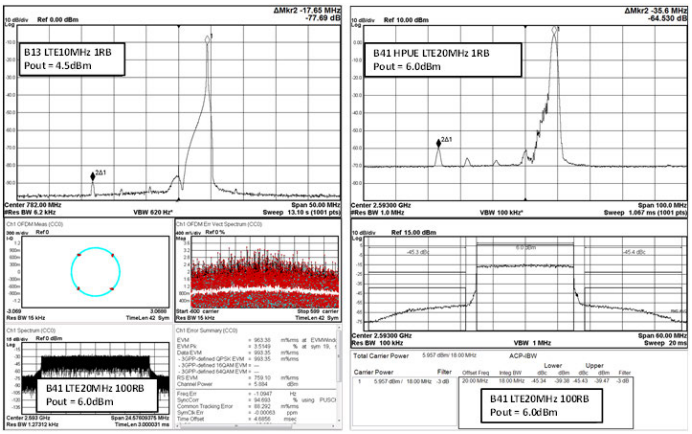


Figure 21.4.4: TX CIM3/EVM/EUTRA measurement results of B13 and B41 HPUE 1RB/100RB test cases.

Sub-sys	Parameter	[6] JSSC 2016	[5] ISSCC 2018	[1] RFIC 2018	This work	Unit
General	LTE-A feature	3 RX	1TX	4RX+2TX	4RX+2TX	
	RX MIMO capability	2x2	2x2	2x2	2CA 4x4	
	VDD	1.4	NA	1.2 / 1.8	1.2 / 1.8	V
	Technology	40	14	40	28	nm
	Die Area	11.3	1.04 (MHB only)	32 (16 + 16)	24.9	mm^2
	LTE B3 battery current	NA	NA	43	33.8	mA
RX	NF (< 2.7GHz)	2.5	NA	2.5	2.4	dB
	NF @ LAA band	NA	NA	3.5	3.5	dB
	ACS1 IIP3	NA	NA	-2	1	dBm
	OBB IIP3 (@45/91MHz offset)	1	NA	NA	0	dBm
	IIP2	55	NA	55	55	dBm
	LTE RX battery current	17.6	NA	NA	16.8 3)	mA
TX	Pout (B41, 1RB)	NA	5.1	5	6	dBm
	CIM3 (B41, 1RB)	NA	-54.5	NA	-64.5	dBc
	Pout (B41, FRB)	NA	6.3	5	7.5	dBm
	Pdc (B41, FRB, @6.3dBm)	NA	140.4 1)	NA	158.1 1)	mW
	EUTRA (B41, FRB)	NA	-41.7	-43	-45.3	dBc
	EVM (B41, FRB)	NA	NA	1.6	< 1.0	%
	Pout (B1, FRB)	NA	3.1	5	8	dBm
	Pdc (B1, FRB, @3.1dBm)	NA	113.2 1)	NA	108.6 2)	mW
	RX-noise (B1, FRB, 3.1dBm, @80MHz)	NA	-157.8	NA	-158.9	dBc/Hz

1) Total TX power consumption including LO generation circuits
2) Total TX power consumption including LO generation circuits and synthesizer (STX)
3) VBAT=3.8V, 90% efficiency to 1.2V and 83% efficiency to 1.8V

Figure 21.4.5: Performance comparison table.

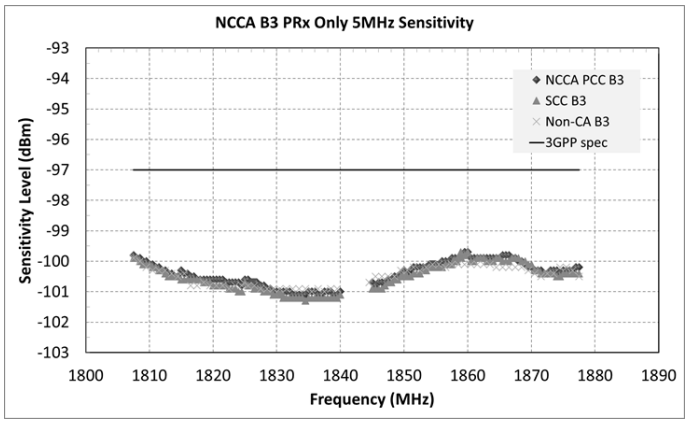


Figure 21.4.6: B3 NCCA and Non-CA sensitivity performance.

21

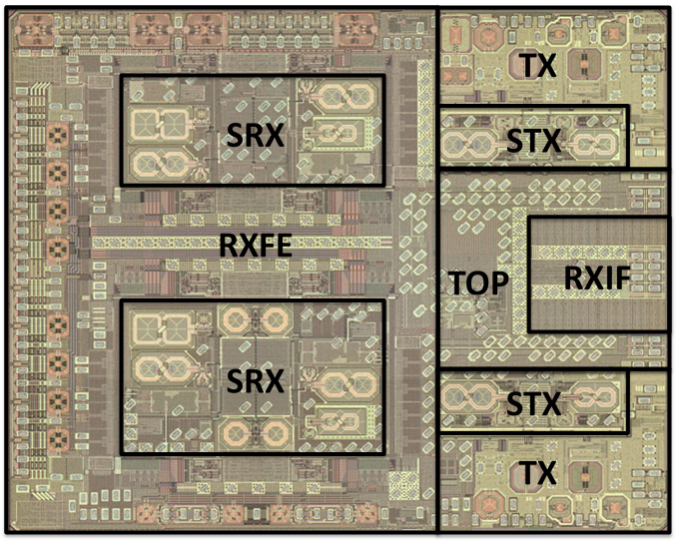


Figure 21.4.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

21.5 A 5G Sub-6GHz Zero-IF and mm-Wave IF Transceiver with MIMO and Carrier Aggregation

Benjamin Jann[1], Greg Chance[2], Ankur Guha Roy[1], Aishwarya Balakrishnan[1], Niranjan Karandikar[2], Thomas Brown[1], Xi Li[2], Brandon Davis[2], Jose Luis Ceballos[3], Nebil Tanzi[2], Kurt Hausmann[2], Hyun Yoon[2], Yen-ling Huang[2], Amit Freiman[4], Bruce Geren[2], Peter Pawliuk[2], Wayne Ballantyne[2]

[1]Intel, Hillsboro, OR; [2]Intel, Chandler, AZ
[3]Intel, Villach, Austria; [4]Intel, Santa Clara, CA

Now that 5G is transitioning from research to commercial products; fully functional systems are needed to detect issues with practical applications and real-life implementations. They are also needed to fine-tune the development of 3GPP standards as well as demonstrate interoperability and new-use cases in the field. 5G prototypes were demonstrated in the 2018 Winter Olympic Games in real mobile device form-factors connecting to a real network. To enable these prototypes the IC building blocks were developed including modem, transceiver and phased-array front-ends for the new sub-6GHz band, 28GHz band and 39GHz band. In this paper we will focus on the transceiver block.

While there have been publications [1,2] demonstrating 5G mm-wave front-ends, a key part of the system, the transceiver including baseband analog and mixed signal components has been neglected. The combination of wide bandwidth (800MHz RF) and high-order modulation such as 256-QAM OFDM presents a difficult challenge. In this paper we present a transceiver with Intermediate Frequency (IF) interface to the mm-wave phased-array front-end ICs supporting air interfaces for 27 to 29.5GHz, 37 to 40GHz as well as an RF front-end for the 3.3-to-4.2GHz band. The transceiver connects to the baseband modem via an industry standard MPHY Gear-3 interface for data & controls. In contrast to high-bandwidth analog I/Q interfaces with difficult matching requirements on the PCB and in the modem, a digital interface allows for easy system integration. The transceiver is manufactured in a 28nm bulk CMOS process with dimensions of 5261μm×5261μm and packaged in a wafer-level CSP package with 2 RDL layers. When coupled with the front-end the system supports a 2×8 dual-polarized 28GHz or 39GHz antenna array and two 4GHz antennas for dual-band MIMO with a peak data-rate of 6Gb/s. With dual-LO receive mixers and reconfigurable ADC allocation, non-contiguous intra-band carrier aggregation (CA) is supported with 400MHz RF bandwidth for one ADC and 800MHz aggregated bandwidth for two ADCs. 256QAM on the sub-6GHz downlink and 64QAM on the 28GHz/39GHz down-link and up-link are supported.

Figure 21.5.1 shows the architecture and the full system partitioning. To support all these features, the system includes two RF chains for 3.3 to 4.2GHz and two IF chains at 10.56GHz. Four DACs and ADCs interface the chains with the digital front-end. Four PLLs support the receivers for dual-band CA and two for the transmitters. In addition there is a clock PLL that supports data converters and other digital functions. The modem is interfaced with eight lanes of Gear-3 MPHY for data and one for control with 5.8368Gb/s per lane.

Figure 21.5.2 shows EVM contributions of various impairments. For sub-6GHz RX the main contributor is quantization noise from the ADC. However the RX IF chain is dominated by phase noise due to the higher-frequency carrier. For the sub-6GHz TX the driver amplifier non-linearity dominates with IQ imbalance close behind. The TX IF is dominated by driver and phase noise.

The transmitter, as shown in Fig. 21.5.3, takes digital data and up-samples in an interpolation filter to 2.46Gs/s. The high sample rate is needed to relax filter requirements. The 10b DAC and baseband filter consists of a current-steering differential pair loaded with a low-impedance current-mirror filter, necessary for the 800MHz RF bandwidth case where the zero-order hold is insufficient to reject the DAC replica for regulatory requirements. The three RC poles are placed at a high enough frequency so that group delay is flat in-band even over PVT. The filter bandwidth tracks with process with a digitally tunable resistor that matches a calibrated resistor. Because the DAC current consumption scales marginally with bandwidth, the entire channel is covered with a single DAC for all use cases. The current mirror has high bandwidth and interfaces to an active mixer as a linear current source. The quadrature LO is generated by dividing down the PLL frequency by two as near to the mixers as possible to maintain accurate quadrature relationship. The mixer upconverts the signal to either 3.3 to 4.2GHz in the case of the sub-6GHz band or to a 10.56GHz IF. In sub-6GHz a notch is placed to attenuate CIM3. After the mixer a driver stage amplifies the signal and drives the 50Ω interface to an external front-end in the case of sub-6GHz or to the phased-array IC in the case of the mm-wave bands.

The receive path for both sub-6GHz & mm-wave bands, shown in Fig. 21.5.4, consists of a low-noise-amplifier (LNA) mixer combination referred to as a micro-receiver (μRX) followed by a transimpedance amplifier (TIA), channel filter (ChF), an analog-to-digital converter (ADC) and Digital Front-End (DFE) that includes decimation. Two sets of LO signals are routed to each of the mixers to provide CA support. This reduces the requirements on the baseband filter and ADC as they need to support half the bandwidth and can be allocated only as needed in wideband use cases. The LNA is implemented as a single-stage inverter with no inductor matching and has a programmable resistive feedback to control gain and input impedance. The inverter topology was chosen for small footprint with no magnetic coupling. A large gain in the FE ICs relaxes the NF requirement on the LNA. The mixer is a single-balanced passive IQ mixer consisting of a switching quad, equipped with 10 attenuation steps of 2dB each. The output clock distribution networks for IF and sub-6GHz bands were implemented in current and voltage mode respectively. The buffers drive a quadrature divider with a 25% duty-cycle output. The quadrature divider outputs go through a quadrature buffer with a duty-cycle correction circuit, which drives the mixer. The μRX can be reconfigured to support CA by using both the LO signals simultaneously and routing the mixer outputs to two baseband blocks. In single-carrier mode either one of the LOs can be used. The lineup also includes a current-controlled DC-offset-correction DAC after the downconversion for IIP2 improvement. The TIA and ChF (3rd-order Butterworth) have 5 bandwidth options (25MHz, 50MHz, 100MHz, 150MHz, 200MHz) and 6 gain configurations (0dB, -3dB, -6dB, -9dB, -12dB, -15dB). The output of the channel filter is fed into a two-step Interpolative Flash ADC with a sampling rate of 1228MS/s. The ADC has foreground calibration, all capacitor internal DAC and reference ladder, redundancy (4.5b ADC for 4 LSBs), and bubble correction enabling it to achieve 7.7 ENOB with 200MHz bandwidth.

An on-chip clock generation and distribution network was used to derive all the clocks from a 38.4MHz crystal oscillator. To enable the carrier aggregation mode, two separate PLLs were used in the receiver for each band. For two transmit bands, receive operation and the clock PLL, 7 PLLs were used in the transceiver. Careful attention was given to supply regulation, floorplan and package to mitigate crosstalk with so many frequency sources on chip. Multiple packages were compared with isolated ground domains and one strong shared ground. It was found that one strong ground attenuated cross-talk and reduced ESD diodes, another source of cross-talk. High PSRR LDOs were utilized on sensitive supply nodes. RF routing was shielded on all sides. Return paths were drawn out to not cross sensitive nodes. Shown in Fig. 21.5.5, all PLLs use a fractional-N, dividerless digital-intensive architecture. A frequency band-select loop was used to bring the VCO frequency close to the required value. A stochastic flash TDC has been reported in [6] that achieves a sub-ps resolution required to meet the strict integrated phase-noise requirements of an RF DPLL. A similar architecture was used in the DPLLs reported in this work. The same TDC and digital loop filter were re-used in all the designs. VCO output frequency and the divider ratios were modified to ensure that the TDC operates within the same input frequency range. The digital loop-filter bandwidth was optimized independently for all the PLLs to reduce the output phase noise.

Figure 21.5.6 compares this work with the state of the art. Direct comparisons are difficult to find as fully integrated 5G transceivers have not been reported but it can be seen that this work has higher aggregated bandwidth than prior transceivers with integrated data converters such as LTE and supports better EVM performance and higher-order modulation than other wideband transceivers such as in 802.11ad/ay.

Acknowledgements:
The authors would like to thank the layout, digital and test teams for supporting this work.

References:
[1] J.D. Dunworth et al., "A 28GHz Bulk-CMOS Dual-Polarization Phased-Array Transceiver with 24 Channels for 5G User and Basestation Equipment", *ISSCC*, pp 70-72, Feb. 2018.
[2] H. Kim et al., "A 28GHz CMOS Direct Conversion Transceiver with Packaged Antenna Arrays for 5G Cellular System," *IEEE RFIC*, pp. 69-72, 2017.
[3] R. Wu et al., "64-QAM 60-GHz CMOS Transceivers for IEEE 802.11ad/ay," *IEEE JSSC*, vol. 52, no. 11, pp. 2871-2891, Nov. 2017.
[4] D. J. McLaurin et al., "A Highly Reconfigurable 65nm CMOS RF-to-Bits Transceiver for Full-Band Multicarrier TDD/FDD 2G/3G/4G/5G Macro Basestations," *ISSCC*, pp. 162-164, Feb. 2018.
[5] K. Lim et al., "A 65nm CMOS 2×2 MIMO multi-band LTE RF transceiver for small cell base stations," *ESSCIRC*, pp. 332-335, 2017.
[6] A. Farber, "Advanced Time to Digital Converter Architectures in 28nm CMOS Process", *M.Sc. thesis, Faculty of Electrical Engineering*, Technion, Haifa, Israel, 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

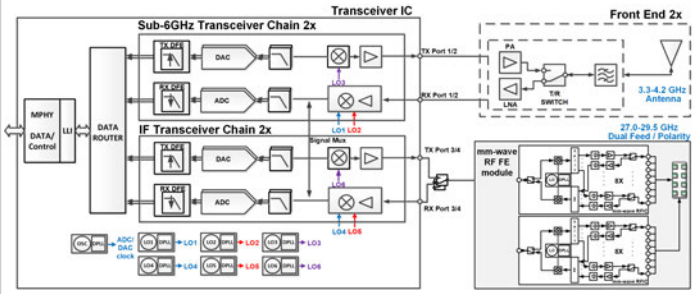


Figure 21.5.1: System architecture and block diagram.

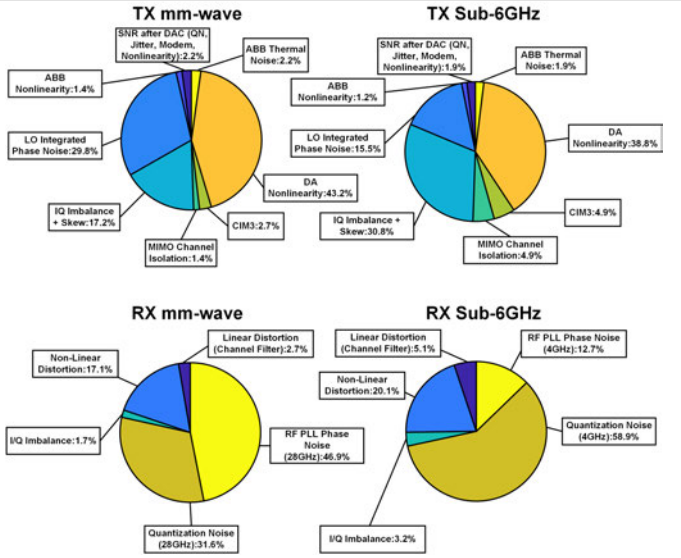


Figure 21.5.2: EVM budget for sub-6 and mm-wave TX and RX.

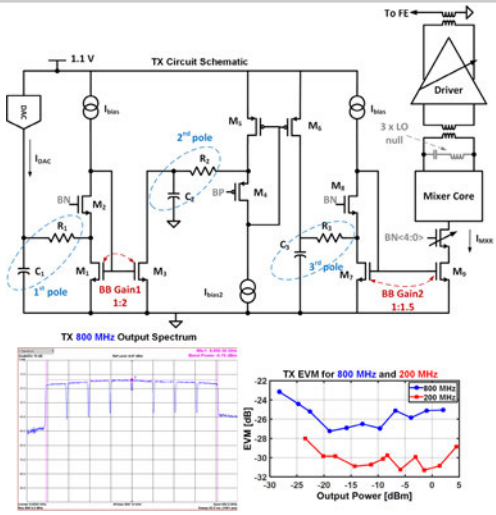
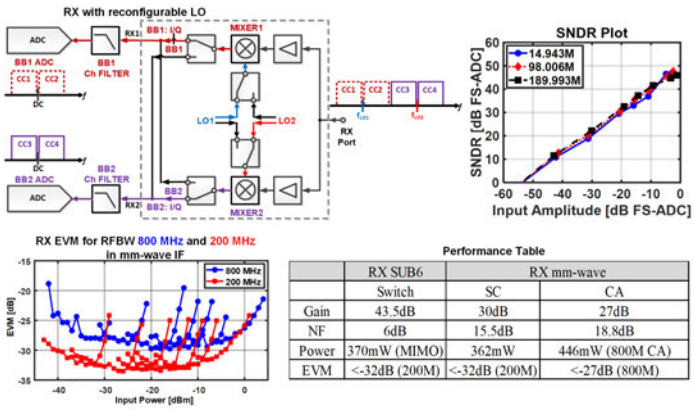


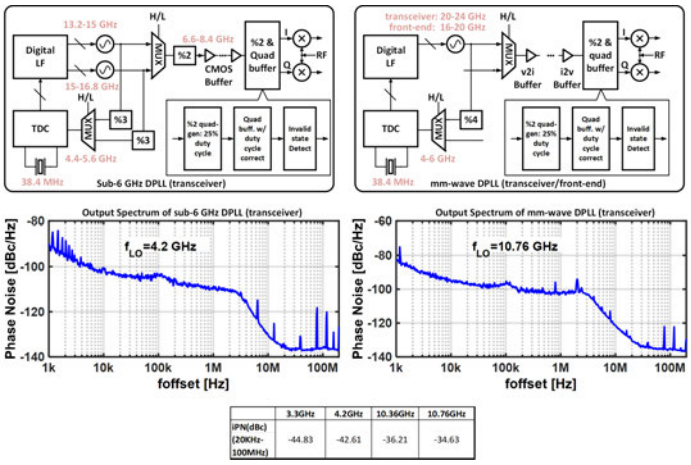
Figure 21.5.3: TX circuit diagram, 800MHz spectrum with 8-component carrier 5G OFDM 64-QAM waveform, EVM vs output power with 5G OFDM 64-QAM waveform in IF band.



Performance Table

	RX SUB6	RX mm-wave	
	Switch	SC	CA
Gain	43.5dB	30dB	27dB
NF	6dB	15.5dB	18.8dB
Power	370mW (MIMO)	362mW	446mW (800M CA)
EVM	<-32dB (200M)	<-32dB (200M)	<-27dB (800M)

Figure 21.5.4: RX reconfigurable LO, ADC SNDR,EVM vs power with 5G OFDM 64-QAM waveform, and performance table.



	3.3GHz	4.2GHz	10.36GHz	10.76GHz
iPN(dBc) (20KHz-100MHz)	-44.83	-42.61	-36.21	-34.63

Figure 21.5.5: Synthesizer block diagram, phase noise performance.

Parameter	This work	[2]	[3]	[4]	[5]
TX Modulation/EVM	64QAM OFDM/-31dB 200MHz 64QAM OFDM/-27dB 800MHz	64QAM/-33.1dB 10MHz	64QAM OFDM/-26.5dB	20MHz LTE @2.6GHz/-47dB	64QAM/-35.4dB 10MHz
RX Modulation/EVM	256 QAM/-34dB 200MHz 64 QAM/-30dB 800MHz	64QAM/-33.3dB 10MHz	64QAM OFDM/-21.3dB	N/A	64QAM/-31dB
Aggregated BW	800MHz	500MHz	4.3GHz	450MHz	10MHz
ADC BW/ENOB	200MHz/7.7	No ADC	No ADC	200MHz/71dB SNR	10MHz/10
Carrier Frequency	4.2GHz/10.56GHz	28GHz	61.56GHz	400MHz-6GHz	680MHz-6GHz
SoC Analog/RF Power Consumption TX/RX (highest performance mode)	596mW/446mW	680mW/400mW	544mW/432mW	4.1W	750mW/480mW
Noise Figure	6dB (4.2GHz)/15.5dB (10.56GHz)	6.7dB	5.7dB	12dB	<5.2dB
TX Output Power	5dBm (64QAM)	10.5dBm Psat	7dBm (64QAM)	7dBm Psat	-0.7dBm (64QAM)
Technology	28nm CMOS	28nm CMOS	65nm CMOS	65nm CMOS	65nm CMOS

Figure 21.5.6: Comparison table.

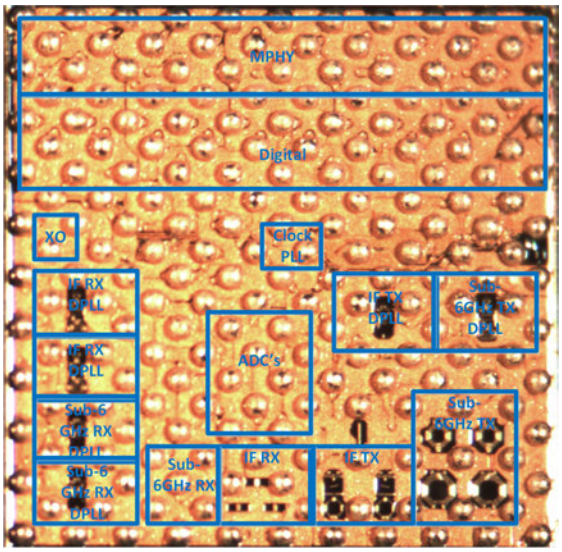


Figure 21.5.7: Die micrograph with 2-RDL-layer package.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

21.6 A Sub-6GHz 5G New Radio RF Transceiver Supporting EN-DC with 3.15Gb/s DL and 1.27Gb/s UL in 14nm FinFET CMOS

Jongwoo Lee, Sangwook Han, Joonhee Lee, Byoungjoong Kang, Jeongyeol Bae, Jaehyuk Jang, Seunghyun Oh, Suseob Ahn, Sanghoon Kang, Quang-Diep Bui, Kiyong Son, Hyungsun Lim, Daechul Jeong, Ronghua Ni, Yongrong Zuo, Ilyong Jong, Chih-Wei Yao, Seungchan Heo, Thomas Byunghak Cho, Inyup Kang

Samsung Electronics, Hwaseong, Korea

To carry on the explosion of mobile data traffic, cellular networks have evolved to enhance air capacity with emerging 5G New Radio (NR) technologies. Thanks to carrier aggregation (CA) and advanced-MIMO techniques, the NR devices can attain up to several Gb/s peak data-rate. The demand of high bandwidth has created a need for exploring high-frequency spectrum over 3GHz, while sustaining legacy LTE bands for LTE-NR dual connectivity (EN-DC). Since User Equipment (UE) requires small form-factor and low power consumption, a single-chip RF transceiver is essential to cover both NR and legacy protocols, simultaneously. This paper demonstrates an integrated CMOS RFIC that supports multimode and multiband applications including all the legacy 2G, 3G, 4G and stand-alone/non-stand-alone sub-6GHz 5G NR features.

The transceiver operates in all FDD/TDD 2G/3G/4G and NR bands from 450MHz to 5925MHz. It consists of primary and diversity receivers (RXs) with seven pipelines and six RX frequency synthesizers to support downlink (DL) 4-CA and three 4×4 MIMO sets as shown in Fig. 21.6.1. Low-noise amplifiers (LNAs) are designed with wideband characteristics, which are compatible with external LNA front-end modules. To support uplink (UL) two-CA and EN-DC, there are two independent transmitter (TX) chains with two dedicated TX frequency synthesizers. For closed-loop power control and TX calibrations, two feedback receivers (FBRXs) are used for respective TXs. In addition, an on-chip digitally-controlled crystal oscillator (DCXO) has been integrated to generate low-noise clocks for all internal frequency synthesizers.

The RFIC has two identical RXs (primary and diversity) to support antenna diversity and DL-MIMO. Each RX has 12 LNAs with dedicated RF input ports: two ports for low-band (LB) from 0.45 to 1GHz, seven ports for mid-high-band (MHB) from 1.4 to 2.7 GHz, and three ports for NR from 3.3 to 6GHz. For size efficiency, a frequency-band switchable transformer has been proposed [1]. The legacy 7-CA case requires only seven transformers by using the proposed pipeline structure in LNA output routing. Each input port has sufficiently wide bandwidth to cover up to intra-5CA in a lump. For SAW-less solutions, digitally assisted IIP_2 calibration is applied to compensate differential mismatch in the RX path. Each RX has seven sets of analog basebands (ABBs) for channel selectivity. Each ABB consists of a 3rd-order lowpass filter and a variable-gain amplifier followed by a SAR ADC. To support the 5G NR, channel bandwidth can be extended up to 100MHz, while DC offset cancellation and automatic bandwidth calibration are also supported. The oversampling SAR ADC architecture provides reconfigurable mode and effective blocker sensing. To save 70% of the ADC area, every set of four ADCs share a reference capacitor as shown in Fig. 21.6.2. To avoid synchronous kick-back noise that deteriorates the ADC linearity, a proprietary asynchronous MSB control is proposed. If adjacent ADC MSB events are asynchronous, a single small reference capacitor is sufficient to provide the required charge. The ADC has been measured over 79dB SNDR in 2G mode and over 66dB SNDR in NR mode at 250MS/s sampling rate.

Two SAW-less direct upconversion TXs are integrated to transmit LTE and 5G NR signals simultaneously (EN-DC). Each TX has dedicated LB, MHB, and NR paths with a baseband filter block. A flexible front-end configuration for multimode and multiband operation is implemented by allocating multiple RF ports on both the LB and MHB paths. The SAW-less operation is demonstrated by using a harmonic-rejection passive upconverter [2] and a harmonic notch filter at the mixer output [3]. To support the contiguous intra-band 2-CA LTE or NR with digital pre-distortion operation, the bandwidth of the baseband filter is widened up to 125MHz. In addition, +7dBm of the TX output power is achieved for the High-power UE in LTE B38, B40, B41, B42, B43, B47, B48, and NR n41, n77, n78, n79. In low-pass filters and driver amplifiers, a power-slicing technique and a multi-level supply-voltage switching technique are employed to reduce power consumption at the backed-off output power case.

Figure 21.6.3 shows the architecture of the RF frequency synthesizer which is comprised of an all-digital phase-locked loop (ADPLL) for the legacy bands and a sub-sampling PLL (SSPLL) for the NR bands over 3GHz and LAA bands. The ADPLL can avoid leakage-current issues in deep-submicron CMOS processes as well as reduce die size by removing a capacitive analog loop filter. It contains a phase-frequency detector (PFD), a multi-modulus feedback divider (MMDIV) driven by a MASH-111 $\Delta\Sigma$ modulator, a digital loop filter and a digitally controlled oscillator (DCO) [4]. To reduce DCO quantization noise, a divide-by-8 DCO clock is used for frequency dithering. For the NR bands over 3GHz, an SSPLL architecture is used with additional oscillators, which allow the narrow frequency range of the DCO and thus save power consumption without performance sacrifice [5].

The transceiver has a proprietary digital interface between the ADC output and the baseband modem to eliminate massive analog connections, which can be built on a highly confined phone board. It requires only 12 lines while the analog interface requires 56 lines to support the same data bandwidth. In the digital interface, each path requires 1.65Gb/s, 3.3Gb/s, 0.96Gb/s data rates for LTE, 5G NR, and 2G, respectively. The digital interface consists of an RX digital filter, a dedicated link, and a PHY block. The digital filter is used to reduce the data-rate of the ADC outputs by decimation. Then, the reduced digital I/Q signals are sent to the link for proper packing. The PHY delivers the packet data to the baseband modem.

A temperature-compensated crystal oscillator system has been proposed with a low-cost DCXO, a temperature-sensing unit (TSU), and a compensation-and-learning unit (CLU). The on-chip DCXO achieves sharper output signal rising/falling slopes under the same power budget with an off-chip TCXO due to reduced parasitic capacitance and resistance of pads and PCB routings. The TSU consists of a single-to-differential amplifier and a 15b SAR ADC that transfers ambient temperature information to the CLU. The proposed CLU employs an on-the-fly learning algorithm to adapt to the environmental variation. While the crystal oscillator has a frequency variation over $\pm$30ppm, the compensated frequency error is bounded within $\pm$2ppm from -30°C to 85°C.

As shown in Fig. 21.6.4, the measured RX sensitivity on the platform is -100.1/-100.2/-99.8dBm in LTE 10MHz B1/B5/B7, respectively, with 23dBm PA output power. The peak CINR is more than 39dB in LTE 10MHz B1/B5/B7. The EVM of LTE 20MHz with 64QAM in B1/B42/B46 shows 2.1% before I/Q calibration. The EVM of the NR 40MHz with 256QAM at n78 is measured at 1.8% before I/Q calibration. The out-of-band IIP_2 is at least 70dBm over temperature after the calibration. The RX consumes 18mA at 1V when operating the LTE one-CA.

The spectra of contiguous intra-band CA with the 40MHz LTE-CA in the B7 band and the sub-6GHz NR 100MHz in the n77 band are shown in Fig. 21.6.5. The measured ACLR is better than -41.2dBc, and the EVM does not exceed 1.94% at the highest output power. For the NR measurement, a cyclic-prefix orthogonal-frequency-division multiplexing (CP-OFDM) waveform having peak-to-average power ratio of 9.6dB is used. This TX consumes 145mA at 1.8V and 30mA at 1V, respectively, while generating output power of 5dBm at the 3.5GHz band for the High-power UE.

References:

[1] Y. Kim et al., "An RF Receiver for Multi-Band Inter- and Intra-Band Carrier Aggregation," *IEEE RFIC,* pp. 79-82, 2016.

[2] V. Bhagavatula et al., "A SAW-less Reconfigurable Multimode Transmitter with a Voltage-Mode Harmonic-Reject Mixer in 14nm FinFET CMOS ," *ISSCC*, pp. 220-221, Feb. 2017.

[3] Q. Liu et al., "A 1.4-to-2.7GHz High-Efficiency RF Transmitter with an Automatic $3F_{LO}$-Suppression Tracking-Notch-Filter Mixer Supporting HPUE in 14nm FinFET CMOS," *ISSCC*, pp. 172-174, Feb. 2018.

[4] C. Yao et al., "A 14nm Fractional-N Digital PLL with $0.14ps_{rms}$ Jitter and -78dBc Fractional Spur for Cellular RFICs," *ISSCC*, pp. 422-423, Feb. 2017.

[5] X. Gao et al., "A 2.2GHz 7.6mW Sub-Sampling PLL with -126dBc/Hz In-Band Phase Noise and $0.15ps_{rms}$ Jitter in 0.18μm CMOS," *ISSCC,* pp. 392-393, Feb. 2009.

[6] C. Chiu et al., "A 40nm Low-Power Transceiver for LTE-A Carrier Aggregation," *ISSCC*, pp. 130-131, Feb. 2017.

[7] T. Georgantas et al., "A $13mm^2$ 40nm Multiband GSM/EDGE/HSPA+/TDSCDMA/LTE Transceiver," *ISSCC*, pp. 160-161, Feb. 2015.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

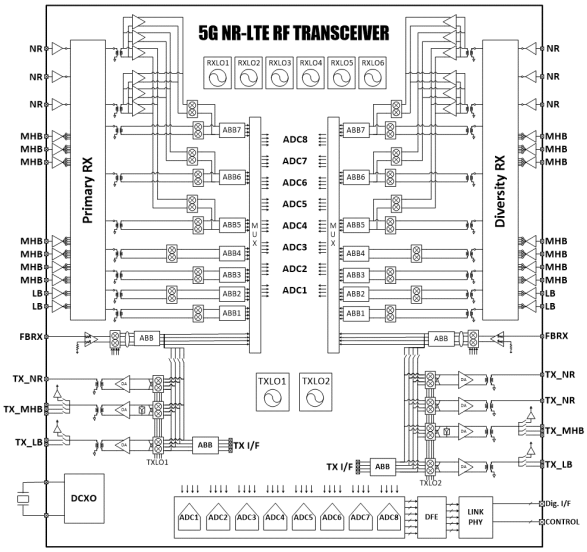


Figure 21.6.1: Block diagram of 5G NR-LTE RFIC.

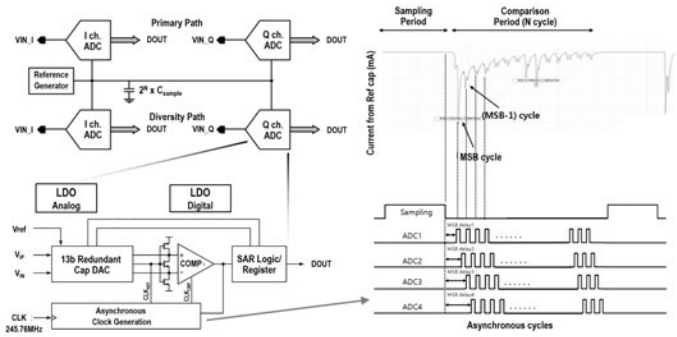


Figure 21.6.2: Asynchronous MSB event control for reference capacitor sharing of multi-ADC.

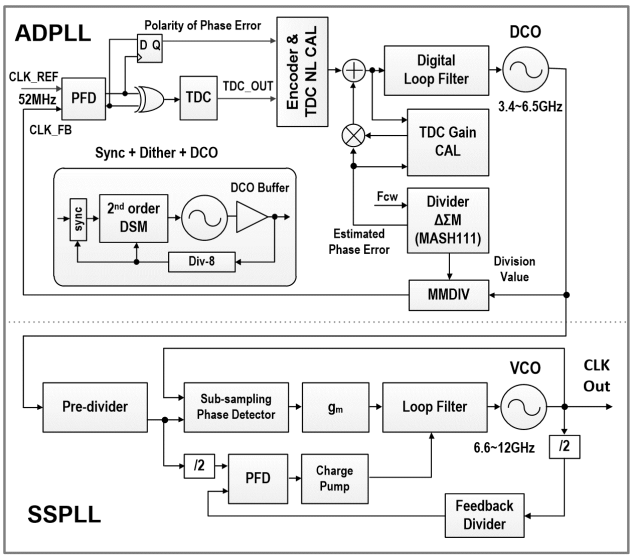


Figure 21.6.3: Proposed RF frequency synthesizer with ADPLL and SSPLL.

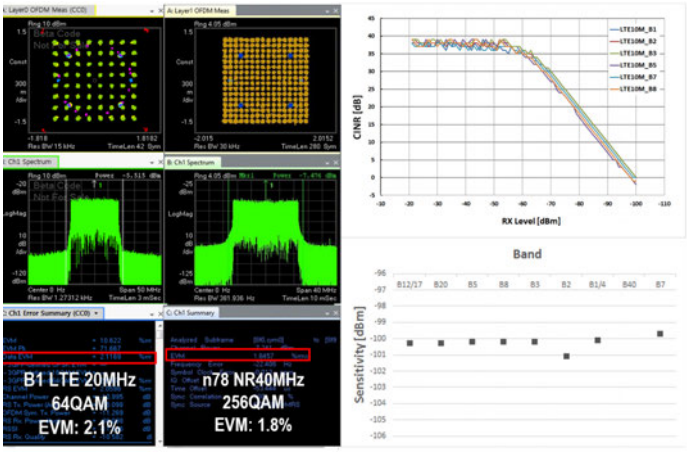


Figure 21.6.4: RX EVM for 64-QAM LTE 20MHz and 256-QAM NR 40MHz, CINR curve and sensitivity.

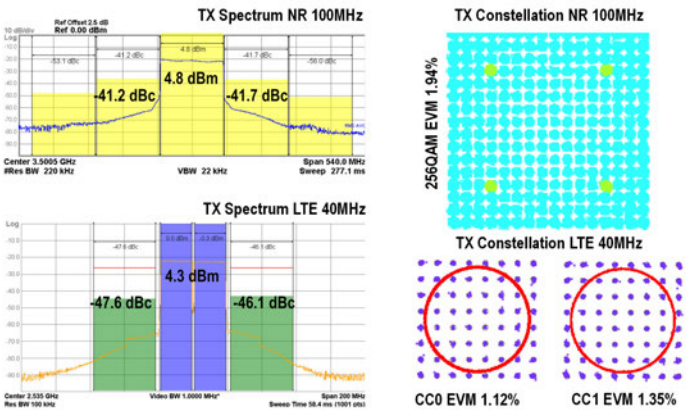


Figure 21.6.5: TX output spectrum for sub-6GHz NR 100MHz in n77 and LTE 40MHz in B7.

		This work	[6]	[7]	Units
Frequency band		**450-5925**	700-2700	700-2700	MHz
Supporting mode		**2G,3G,4G,LTE-A, SA/NSA Sub-6G NR**	2G, 3G, C2K, LTE-A	2G, 3G, LTE	
Technology		**14**	40	40	nm
Die size		**38.4**	13.9	13	mm^2
3G power		**117**	125	137	mW
LTE power/CC		**144**	212	241	mW
TX	UL path #	**2**	1	1	
	P_{out}	**>5 (n77)**	> 5 (B7)	4	dBm
	EVM	**< 1.94 (n77)**	< 2 (B7)	<1.8	%
	HPUE	**Yes**	No	No	
	Saw-less	**Yes**	-	Yes	
RX	DL path #	**14**	4	2	
	Sensitivity	**-101**		<-102	dBm
	EVM	**<2.1***		<1.5	%
	IIP_2	**>70**	55	>55	dBm

* Before I/Q Calibration

Figure 21.6.6: Comparison table.

21

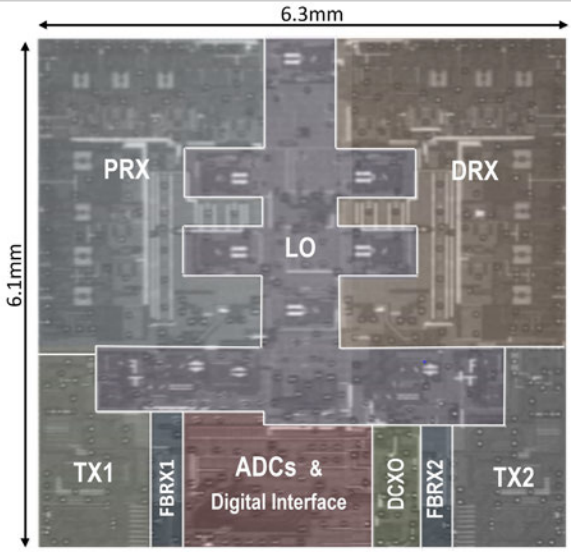


Figure 21.6.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

21.7 A Mixed-Signal Circuit Technique for Cancellation of Multiple Modulated Spurs in 4G/5G Carrier-Aggregation Transceivers

Silvester Sadjina[1,2], Krzysztof Dufrêne[1], Ram Sunil Kanumalli[1], Mario Huemer[2], Harald Pretl[1,2]

[1]Intel DMCE, Linz, Austria
[2]Johannes Kepler University, Linz, Austria

To enable high data-rates in mobile communications the LTE standard has implemented carrier aggregation (CA) to increase channel bandwidth. Several local oscillator (LO) signals are needed to drive a multitude of mixers to receive from a number of LTE bands in the different CA scenarios. Due to device nonlinearities in the LO signal distribution and cross-talk between LO paths spurs are generated inside the transceiver chip which may downconvert interfering signals located at the spurious frequencies into the receiver baseband (Fig. 21.7.1). These unwanted signals could either be the transceiver's own transmit signal when operating in frequency-division duplex (FDD), or external signals received by the antenna. An interferer downconverted by an RX LO-LO spur is referred to as a modulated spur and causes severe degradation of the signal-to-noise-and-distortion ratio (SNDR) of the desired signal.

Different modulated spur mitigation techniques have been proposed in the literature (Fig. 21.7.1). Digital-only interference cancellation (DIC) offers a flexible and configurable solution [1]. However, the cancellation performance is limited by the estimation process of the TX-RX leakage channel and is also limited to known interferers like the TX leakage signal. In contrast, mixed-signal cancellation (MSC) combines the simplicity of analog estimation and the reliability of digital cancellation [1]. An auxiliary (AUX) receiver is used to receive the interferer and to convert it to the digital domain to suppress the interference in the main receive chain by a digital cancellation system. Current MSC architectures are focused on the cancellation of one interfering signal. In this paper a mixed-signal architecture for 4G/5G CA transceivers is proposed targeting the cancellation of multiple, possibly unknown modulated spur interferences in LO-LO cross-talk scenarios.

The proposed architecture uses the serial-mixing concept [2] for the detection of modulated interferers in LO-LO cross-talk scenarios (Fig. 21.7.2 top). Two series-connected passive mixers downconvert the incoming RF signal using two LO signals, which are coherent with the LO signals in the main path. Due to the mixing, an interfering signal near to any of the spurious response bands is downconverted to the baseband (BB) at exactly the same low-IF carrier frequency as the troublesome modulated spur in the main path. This downconverted signal is filtered, digitized, and provided as a reference signal for the adaptive digital cancellation. In this architecture, cancellation of multiple modulated spurs is not possible because they may overlap in the baseband. To cancel multiple modulated spurs, the multiphase-mixing concept is proposed (Fig. 21.7.2 bottom). The phase of the modulated spur is dependent on the LO harmonic downconverting the interferer. To separate multiple, potentially overlapping, modulated spurs an additional phase-shift is introduced with this N-phase mixer structure. The LO phase dependency of the modulated spurs in the N-phase mixer structure can be used to extract this interference [3]. The extraction of a spur is similar to the calculation of a Fourier coefficient *R* in the N-phase mixer-path domain. For cancellation of M interferers, N=2×M+1 phases are needed.

The block diagram of the proposed architecture for a two-CA receiver is shown in Fig. 21.7.3. A 5-phase serial mixer structure was chosen for the separation of two modulated spurs, providing a reasonable trade-off between circuit complexity and number of interferers to be canceled. In parallel to the two main RX chains an AUX chain is implemented to extract the interfering signals. It re-creates the interferer downconversion behavior of the main path, and after separation, the interfering signals are provided as references to a digital cancellation block. Each of the five paths of the AUX RF part consists of an LNA followed by a fully differential serial-mixing architecture. The LOs used for both mixing stages are routed from the LOs of the main path provided by on-chip PLLs. Therefore, no additional PLLs are needed. To generate 5 phases for the second mixing stage a delay-locked-loop (DLL) is used. It consists of a five-stage differential inverter block whose delay is controlled by a phase detector regulating the power supply of the delay chain. The baseband (BB) part of the AUX path consists of two filter stages followed by a flash ADC. To reduce harmonic distortion due to quantization, dither is added to the signal before the second filter stage. The added dither is correlated between the paths and is therefore automatically separated from the interferer reference signals in the reconstruction unit. After the separation, the generated reference signals are fed to the digital cancellation system. A widely-linear LMS (WL-LMS) adaptive filter is used to estimate the modulated spur interference and its image [1].

The schematic of the prototype is shown in Fig. 21.7.4. In the main receive chain a Class-AB inverter-based LNA is implemented. The IQ single-balanced mixing stage uses a current-mode 25%-duty-cycle passive mixer. Succeeding the mixer, a first-order lowpass passive filter provides filtering of the BB signal, followed by a delta-sigma analog-to-digital converter ($\Delta\Sigma$-ADC) with a dynamic range of 89dB and sampling rate fs=1GHz. In the AUX path, a self-biased Class-AB inverter-based LNA is used. The serial-mixing part consists of a double-balanced mixer followed by a single-balanced mixer. 50% duty-cycle and 20%/80% duty-cycle LO signals are used for the first and second mixing stages, respectively. The phase shift is introduced in the second mixing stage. A 20%/80% duty-cycle is used to have maximum gain and non-overlapping switching times of the mixers. The differential implementation of the filter stages and the filtering in between are needed to suppress the desired signals located at f_{LO1} and f_{LO2}. After the mixing stage a passive filter followed by a first-order active filter performs the first filtering. Weighted resistors are used at the end of this filter stage to compensate the unequal duty cycle of the second mixing stage. After adding dither to the BB signal further filtering is done by a second-order Rauch filter. The filtered BB streams are then digitized by a flash ADC with an SNDR of 42dB and sampling rate fs=200MHz. The demonstrator has been implemented in 28nm CMOS as a part of a cellular transceiver. The analog detector part occupies an area of 0.41mm² (Fig. 21.7.7).

The RF performance of the AUX receive chain is shown in Fig. 21.7.5 (top). For the measurements, the receiver is configured for LTE band 2 and an RF bandwidth of 5MHz per carrier for a two-CA case with LO1 (=PCC) and LO2 (=SCC) set to 1950MHz and 1966MHz, respectively. The SNDR of the interferer reference signal of the AUX chain is dominated by the quantization noise, IMD3 of the chain and phase error of the DLL, and a peak SNDR of 33dB is achieved. Figure 21.7.5 (bottom) illustrates the cancellation performance. Two modulated spurs induced by interferers located at $2\times f_{LO1}-1\times f_{LO2}$=1.934GHz (external interferer) and $6\times f_{LO1}-5\times f_{LO2}$=1.870GHz (TX leakage) with -36dBm and -18dBm at the LNA input, respectively, can be canceled by 6dB below noise floor (Fig. 21.7.5 bottom left, step 1&2). In addition to both modulated spurs also the IMD2 component can be canceled (step 3). Single resource-block (RB) modulated interferers were chosen to better illustrate the cancellation performance. The effective noise figure, for different TX leakage power levels at the LNA input (Fig. 21.7.5 bottom right), can be restored to a value within 1dB of the reference value (CA disabled). For TX leakage power levels where the IMD2 component is dominating the NF can be restored below this reference value.

Acknowledgements:
The authors wish to acknowledge the contributions and support of Hubert Kusche for chip layout and their colleagues at Intel and the Christian Doppler Laboratory for Digitally Assisted RF Transceivers for Future Mobile Communications. The financial support by the Austrian Federal Ministry of Science, Research and Economy and the National Foundation for Research, Technology and Development is gratefully acknowledged.

References:
[1] A. Elmaghraby et al., "A Mixed-Signal Technique for TX-Induced Modulated Spur Cancellation in LTE-CA Receivers," *IEEE TCAS-I,* vol. 65, no. 9, pp. 3060-3073, Sept. 2018.
[2] S. Sadjina et al., "A Mixed-Signal Circuit Technique for Cancellation of Interferers Modulated by LO Phase-Noise in 4G/5G CA Transceivers," *IEEE TCAS-I*, vol. 65, no. 11, pp. 3745-3755, Nov. 2018.
[3] D. Murphy et al., "A Noise-Cancelling Receiver Resilient to Large Harmonic Blockers," *IEEE JSSC*, vol. 50, no. 6, pp. 1336-1350, June 2015.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

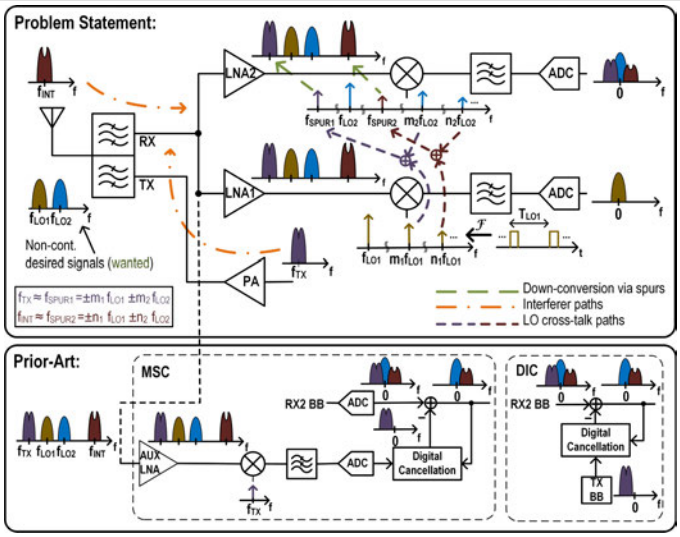


Figure 21.7.1: Mixed-signal and digital interference-cancellation systems for the mitigation of modulated spurs.

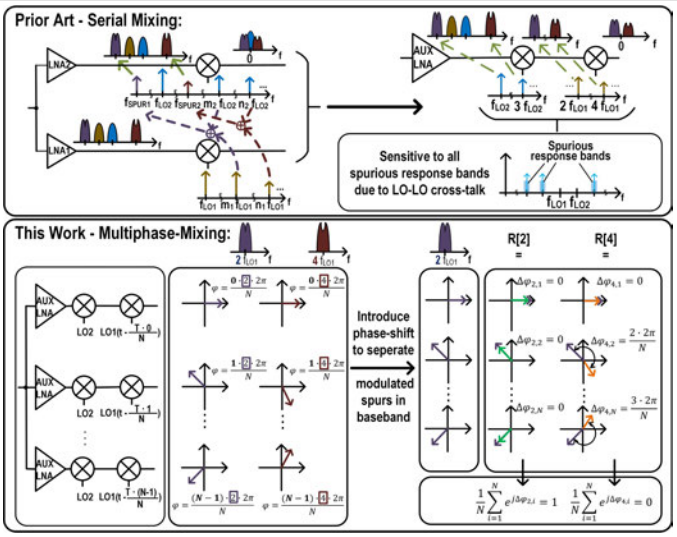


Figure 21.7.2: Proposed multiphase-mixing concept for the detection and separation of multiple modulated spurs.

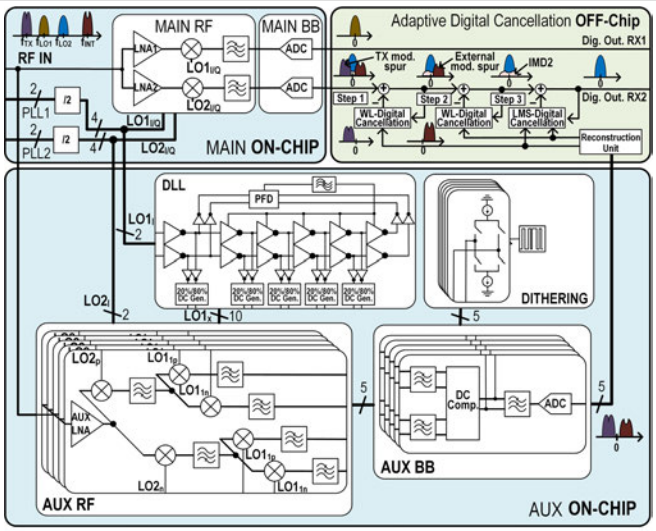


Figure 21.7.3: Block-level diagram of proposed mixed-signal cancellation system.

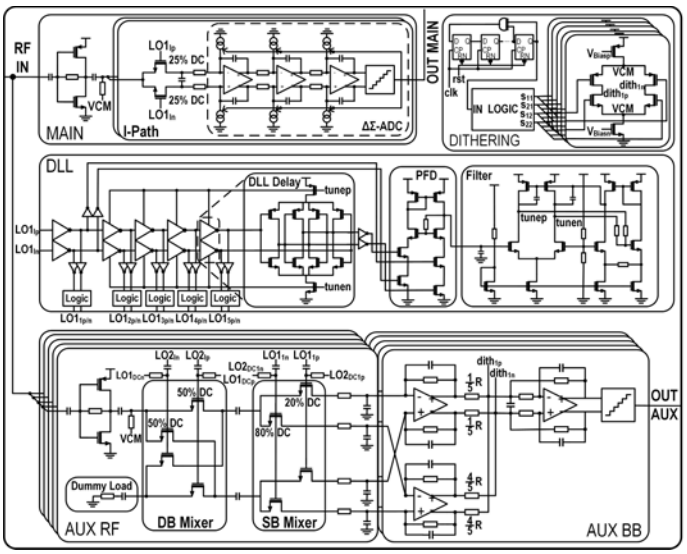


Figure 21.7.4: Circuit diagram of main and AUX receive chain.

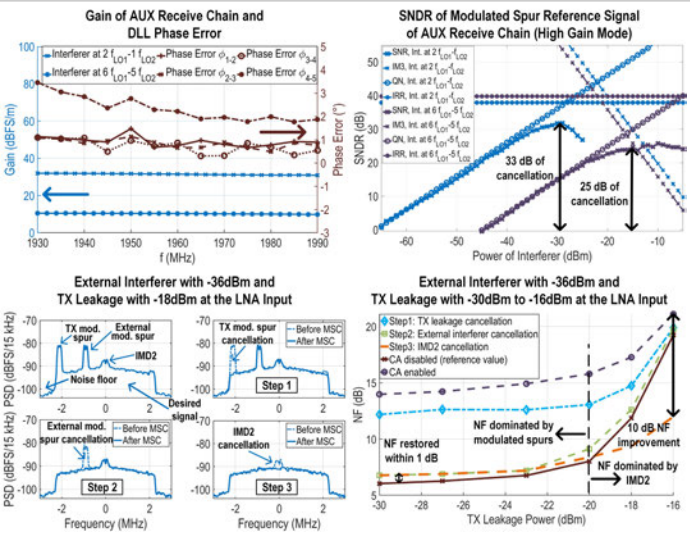


Figure 21.7.5: Measurement results of AUX receive chain (top) and illustration of the cancellation performance (bottom).

	This Work	[1]	[2]	[3]
Type	MSC	MSC	MSC	Analog Interference Cancellation
Point of Cancellation	BB	BB	BB	BB
Technology	28 nm LP	28 nm LP	28 nm LP	28 nm
RF Frequency (GHz)	LTE Band 2	1.85-1.91	0.7-3.8	0.1-3.0
NF at 2 GHz (dB)	5.9	3.4	3.5	1.7
OB-IIP3 (dBm)	9.6	9.4	-	10
OB-IIP2 (dBm)	64	61	-	49.5
Power Consumption (mW)	Main: 26 AUX: 20	76	52	38.8-70
Frontend Cancellation (dB)	-	-	-	-
Backend Cancellation (dB)	33	30	40	52-60
External Isolation (dB)	55	55	55	-
Convergence Time (us)	10	150	10	-
Targets Only Self-Interference	No	Yes	No	No
Digital Cancellation Algorithm	WL-LMS	WL-LMS	WL-LMS	-
Targets Only LO-LO Cross-Talk	Yes	No	Yes	No
Cancellation of Multiple Modulated Spurs	Yes	No	No	-
Cancellation of Image of Modulated Spur	Yes	Yes	Yes	-
Cancellation of IMD2	Yes	Yes	Yes	No

* This work re-creates interferer down-conversion behaviour of main path. [3], for example, can't cancel interferer at f_{LO1}-2 f_{LO2}.

Figure 21.7.6: Result summary and comparison with prior art.

21

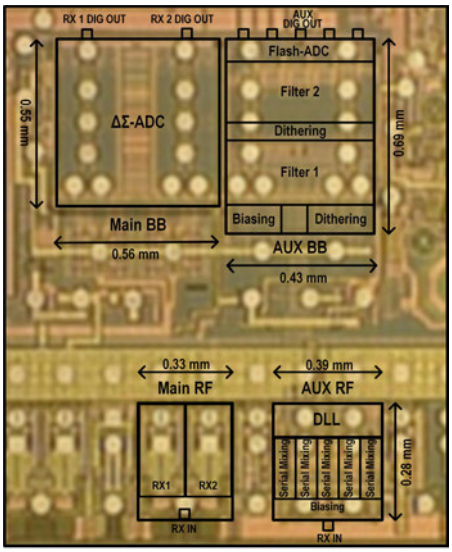


Figure 21.7.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 22 Overview: *Physiological Monitoring*

IMMD SUBCOMMITTEE

Session Chair: ***Esther Rodriguez-Villegas***
Imperial College, London, United Kingdom

Associate Chair: ***Rikky Muller***
University of California, Berkeley, Berkeley, CA

Subcommittee Chair: Chris Van Hoof, imec, Leuven, Belgium

This session involves wearable and implantable SoCs that record weak biopotential signals in the presence of real-life interference and under stringent power and size constraints. These new SoCs and corresponding techniques pave the way toward robust non-invasive devices that involve direct sensing, multi-sensor fusion and implantable closed-loop sensor/actuator systems. These advancements enable multimode physiological recordings from nearly every major organ system.

8:30 AM

22.1 A 769μW Battery-Powered Single-Chip SoC with BLE for Multi-Modal Vital Sign Health Patches

M. Konijnenburg, imec - Netherlands, Eindhoven, The Netherlands

In Paper 22.1, imec introduces an SoC for concurrent and synchronized ECG, bioimpedance and PPG recording, incorporating accelerators and an ARM Cortex-M4f microprocessor. The SoC also includes cryptography IP and an SRAM-based PUF solution. The 4320×4320μm^2 SoC consumes 769μW. The total average system power drawn from the battery, including LDO, LEDs and accelerometer is 2mW, allowing an operating time of more than 3 weeks on 2 ZnAir batteries.

9:00 AM

22.2 A Rugged Wearable Modular ExG Platform Employing a Distributed Scalable Multi-Channel FM-ADC Achieving 101dB Input Dynamic Range and Motion-Artifact Resilience

J. Warchall, University of California, San Diego, La Jolla, CA

In Paper 22.2, the University of California, San Diego presents an active electrode ExG platform achieving 101dB DR and motion artifact resilience, consuming 228μW per channel. The IC has been fabricated in a 65nm technology.

9:30 AM

22.3 A 0.5V 9.26μW 15.28mΩ/√Hz Bio-Impedance Sensor IC with 0.55° Overall Phase Error

K. Kim, KAIST, Daejeon, Korea

In Paper 22.3, KAIST presents a bio-impedance sensor IC for wearable/implantable blood status monitoring. The sensor IC demonstrates thoracic impedance variation (TIV) and impedance cardiography (ICG), consuming less than 10μW of power.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

9:45 AM

22.4 A 27.8μW Biopotential Amplifier Tolerant to 30V_{pp} Common-Mode Interference for Two-Electrode ECG Recording in 0.18μm CMOS

N. Koo, KAIST, Daejeon, Korea

In Paper 22.4, KAIST presents a biopotential amplifier that has a large tolerance to common-mode interference (CMI) for reliable two-electrode ECG recording. The IC demonstrates suppression of up to 30V of CMI, consuming 27.8μW.

10:15 AM

22.5 A Bio-Impedance Readout IC with Digital-Assisted Baseline Cancellation for 2-Electrode Measurement

H. Ha, imec - Netherlands, Eindhoven, The Netherlands

In Paper 22.5, imec presents a bio-impedance readout IC for low-noise 2-electrode measurement by implementing a circuit that simultaneously mitigates 1/f noise and cancels reference current noise. The IC is demonstrated in measurement on the human body.

10:45 AM

22.6 A 13-Channel 1.53-mW 11.28-mm^2 Electrical Impedance Tomography SoC Based on Frequency Division Multiplexing with 10× Throughput Reduction

B. Liu, Shanghai Jiao Tong University, Shanghai, China; National University of Singapore, Singapore

In Paper 22.6, Shanghai Jiao Tong University presents an impedance tomography SoC based on frequency division multiplexing, supporting 13 voltage-sensing channels. Data throughput is reduced by 10×, while low power (118μW) and area (0.87mm^2) per channel are achieved.

11:15 AM

22.7 A Programmable Wireless EEG Monitoring SoC with Open/Closed-Loop Optogenetic and Electrical Stimulation for Epilepsy Control

S-Y. Lee, National Cheng Kung University, Tainan, Taiwan

In Paper 22.7, National Cheng Kung University presents a highly integrated SoC that combines a wireless interface, neural signal sensing, computation, and both electrical and optical stimulation in the brain. The system is demonstrated to detect and suppress seizures in rodents.

11:45 AM

22.8 Adaptively Clock-Boosted Auto-Ranging Responsive Neurostimulator for Emerging Neuromodulation Applications

M. R. Pazhouhandeh, University of Toronto, Toronto, Canada

In Paper 22.8, the University of Toronto presents a system for bidirectional neurostimulation from scalp electrodes utilizing temporally interfering electric fields. EEG recording circuits and a sleep-state classifier trigger responsive neurostimulation.

22.1 A 769μW Battery-Powered Single-Chip SoC with BLE for Multi-Modal Vital Sign Health Patches

Mario Konijnenburg[1], Roland van Wegberg[1], Shuang Song[2], Hyunsoo Ha[1], Wim Sijbers[2], Jiawei Xu[1], Stefano Stanzione[1], Chris van Liempd[1], Dwaipayan Biswas[2], Arjan Breeschoten[1], Peter Vis[1], Chris Van Hoof[1,2,3], Nick Van Helleputte[2]

[1]imec - Netherlands, Eindhoven, The Netherlands
[2]imec, Leuven, Belgium
[3]KU Leuven, Heverlee, Belgium

Continuous vital-sign monitoring is of paramount importance in remote heath monitoring or rehabilitation environments for chronic diseases. Medical-grade wireless and wearable bio-sensor systems that can be used at home offer a much more attractive solution than hospital-based monitoring systems. We report an all-in-one battery-powered SoC designed for low-cost single-use health patches (Fig. 22.1.1), allowing continuous monitoring in a home setting to improve patient comfort and reduce cost of care by, e.g., reducing hospital stays. In addition to medical-grade signal quality, low power consumption is key in such a health patch system, to enable a comfortable form factor with miniature battery size and prolong the operational lifetime to at least several weeks.

The SoC (Fig. 22.1.1) has built-in support for concurrent and synchronized ECG (2 channels), Bio-impedance (BIOZ) and PhotoPlethysmoGraph (PPG). Data processing is performed by dedicated digital filters and accelerators for DMA, FFT, matrix calculations and sample-rate conversion. In addition, an ARM Cortex-M4f microprocessor performs system control and further data processing. Secured boot and data transmission are important features in medical systems for privacy and safety reasons. The SoC includes cryptography IP that supports a wide range of protocols combined with an SRAM-based PUF solution that guarantees a chip-specific unique cryptographic key. The SoC has several interfaces including a BLE radio and an USB2 interface. Two on-chip LDOs generate from a single battery source (1.3 to 2.0V; e.g., a ZnAir battery): a configurable voltage for the digital core (0.6 to 1.2V) and a fixed voltage (1.2V) for SRAM, BLE, AFE, and the on-chip PLL. The very high integration level of the SoC allows a minimum of system component count (Fig. 22.1.1), which saves area, cost, power and weight.

Ultra-low-power, low-noise biomedical analog frontends are traditionally designed in older technology nodes due to their better noise, matching and leakage performance than (deep) scaled technologies. To allow co-integration with advanced low-power digital and RF blocks in a single-chip solution, this SoC is designed in a 55nm technology node. The ECG and BIOZ readouts are based on a power-efficient current-feedback IA [1]. For the PPG channel, the main design challenges are achieving a high dynamic range (DR) to keep signal integrity in presence of large DC component and low power consumption, which is dominated by the LEDs [1,2]. The LED power is usually as high as several mW, dominating the system power consumption.

A pulsed LED synchronized to the readout reduces the power consumption. The signal quality is determined by the peak current and frequency of the LED pulses. Hence short pulses and fast readouts should be used to reduce LED power consumption. The achievable pulse widths are usually limited by the PhotoDetector (PD) capacitance and amplifier input impedance. Existing single-ended readout solutions provide a DC bias for the PD that reduces its parasitic capacitance, however the DR is limited. Fully differential solutions provide a high DR but lack DC biasing capabilities for the PD [4]. Figure 22.1.2 shows the fully differential PPG readout frontend, which achieves high DR and a bias for the PD at the same time. This PPG readout includes a current-sensing stage and a transimpedance stage. In the first stage, the sub-amplifiers A1 and A2 keep the voltage on both nodes of the PD to the desired value, set by V_{BIASP} and V_{BIASN}. Up to 0.6V DC bias voltage can be provided, resulting in about 30% reduction in the parasitic capacitance. Moreover, the input impedance is reduced by both the local and main feedback loop in the current-sensing stage (Fig. 22.1.2), resulting in a <100Ω input impedance. Therefore, the amplifier enables a power-efficient 10μs level LED pulse (Fig. 22.1.3) compared to 100μs pulses [5].

To achieve a high effective DR without relying on >16b ADCs, a digitally assisted DC current-cancellation loop is used [4]. The DC component is cancelled with a 6b current DAC directly at the input of the readout, resulting in a 75μA maximum input current. The PPG readout channel consumes 121μW while achieving a 28pA rms input noise in 20Hz band and 111dB effective DR with a 13b SARADC (pulse frequency at 2kHz, 16× oversampling). Figure 22.1.3 shows a PPG waveform recorded using a finger sensor with infrared light and an SpO_2 waveform derived from red and IR PPG during breath holding. In an HR detection only scenario the total LED power is reduced to 205μW because a 0.67% duty cycle is used which is a 52% reduction compared to the state of the art [5] (Fig. 22.1.6).

A 13b SARADC (Fig. 22.1.4) is reused in ECG, BIOZ and PPG channels due to its flexibility in exchanging power for resolution via oversampling. The BIOZ and PPG channel with a low BW of interest require a larger DR than ECG, and hence a higher oversampling ratio is employed. However, oversampling only reduces the noise, providing no improvement in linearity. Therefore, the voltage DAC employs randomization dynamic element matching in the 6 MSBs by a Pseudo-Random Number Generator (PRNG) controlled thermometer-coded elements, which turns the harmonic distortion into a noisy spectrum in the Nyquist band. Apart from quantization, sampling noise and nonlinearity of the DAC, the comparator thermal noise also contributes to the total noise spectrum. The low-power dynamic pre-amplifier and latch-based comparator in [3] suffers from time-varying and limited gain (8 to 12×). The CM voltage of the decision point (Node 1,2 in Fig. 22.1.4) is difficult to set, because for the latch it should be low to guarantee high gain, but for the pre-amplifier it should be high enough to keep the input devices in saturation to reduce IRN. Figure 22.1.4 shows the high-gain comparator. By inserting cascode transistors MN4/MN5 and gain boosting transistors MN6/MN7, the gain of the dynamic preamp is increased to 52× and is less time-dependent. A high gain reduces the noise contribution of capacitors at node 1,2 and the latch. As a result, the IRN of the comparator is reduced (2 to 3×) compared to [3] for the same power consumption. The SARADC consumes 0.9 μW for an ENOB of 12.8b in a BW of 0.5 to 150Hz at a sample frequency of 32 kS/s (OSR=64).

Figure 22.1.5 shows the measurement setup executing a realistic typical (health patch) application. It covers concurrent and synchronous data collection from ECG, BIOZ, PPG (2 streams, red and infrared LEDs, duty-cycled 10%) and accelerometer. Additionally, on-chip calculation of HR from ECG, respiration rate from BIOZ, and SpO_2 from the PPG streams is performed. The accelerometer data is processed to track activity for further data processing, e.g., motion artefact reduction. During data processing, the FFT, matrix processor, and sample rate converter accelerators are used. The relevant processed data is sent via a BLE link (duty-cycled 5%) to a viewer application on a computer. The SoC consumes 769μW (from the regulated LDO outputs). The total average system power drawn from the battery, including LDO, LEDs and accelerometer, is 2mW allowing an operating time of more than 3 weeks on 2 ZnAir batteries in series. Finally, the impact of the system on channel noise is demonstrated (Fig. 22.1.5). In general, a slight increase of noise from 10.3 to 11.7mΩ_{rms} is observed. Figure 22.1.6 benchmarks the SoC, the PPG channel and the SARADC. Summarizing, an all-in-one (highest level of integration) low-power SoC is proposed for low-cost miniaturized wearable devices (e.g., Health Patch). The improved comparator in the SARADC for all readouts, and specifically the differential PPG readout, assures medical standard signal quality for ECG, BIOZ and PPG with lowest power consumption among reported solutions (Fig. 22.1.6). Figure 22.1.7 shows the die, sized 4320×4320μm^2.

References:
[1] M. Konijnenburg, et al., "A Multi(bio)sensor Acquisition System with Integrated Processor, Power Management, 8 × 8 LED Drivers, and Simultaneously Synchronized ECG, BIO-Z, GSR, and Two PPG Readouts," *IEEE JSSC*, vol. 51, no. 11, pp. 2584-2595, Nov. 2016.
[2] P. Schönle, et al., "A Multi-Sensor and Parallel Processing SoC for Miniaturized Medical Instrumentation," *IEEE JSSC*, vol. 53, no. 7, pp. 2076-2087, July 2018.
[3] M. van Elzakker, et al., "A 10-bit Charge-Redistribution ADC Consuming 1.9μW at 1 MS/s," *IEEE JSSC*, vol. 45, no. 5, pp. 1007-1015, May 2010.
[4] J. Xu, et al., "A 665μW Silicon Photomultiplier-Based NIRS/EEG/EIT Monitoring ASIC for Wearable Functional Brain Imaging," *ISSCC Dig. Tech. Papers*, pp. 294-295, Feb. 2018.
[5] A. Sharma, et al., "A Sub-60-μA Multimodal Smart Biosensing SoC With >80-dB SNR, 35-μA Photoplethysmography Signal Chain," *IEEE JSSC*, vol. 52, no. 4, pp. 1021-1033, Apr. 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

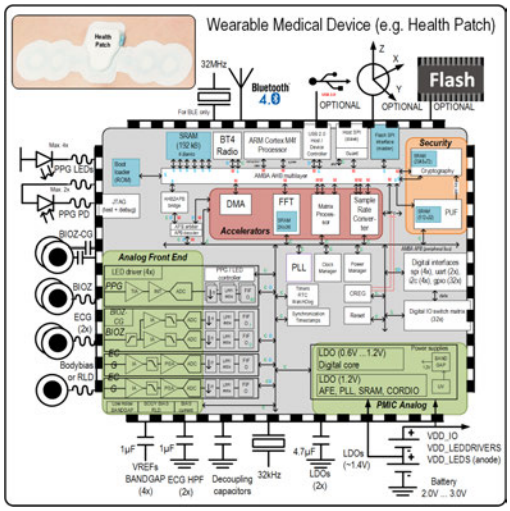


Figure 22.1.1: A fully battery-powered bio-medical multi-sensor acquisition ASIC for wearable medical devices. A relatively low amount of external components makes miniaturization easy to be performed.

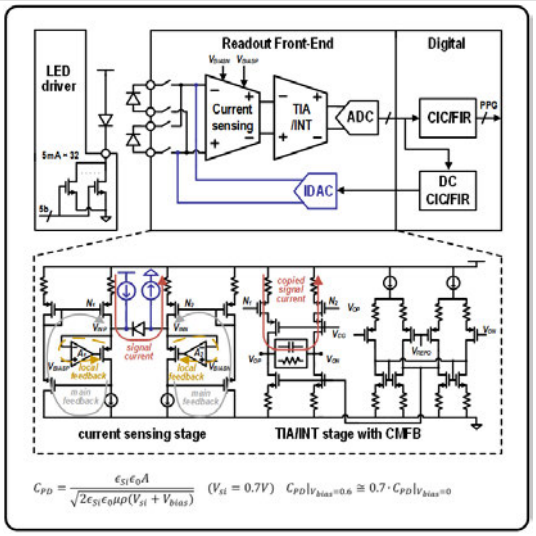


Figure 22.1.2: A two-stage fully differential PPG readout frontend together with a DC servo loop, providing a reverse DC bias voltage for the PD.

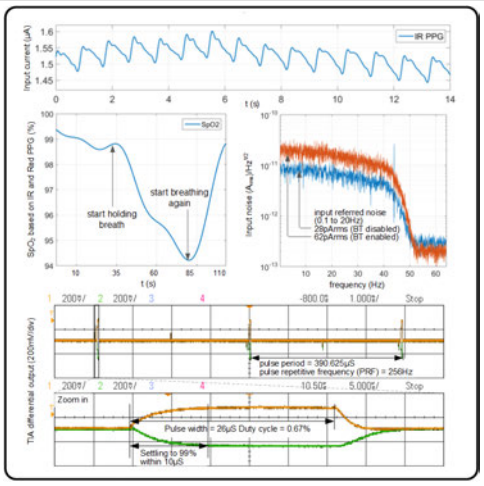


Figure 22.1.3: PPG measurement results, an IR PPG waveform, an extracted SpO2 waveform, the input referred noise spectrum when enable/disable the radio and the differential output of the TIA, showing a 10µs level LED pulse width can be used.

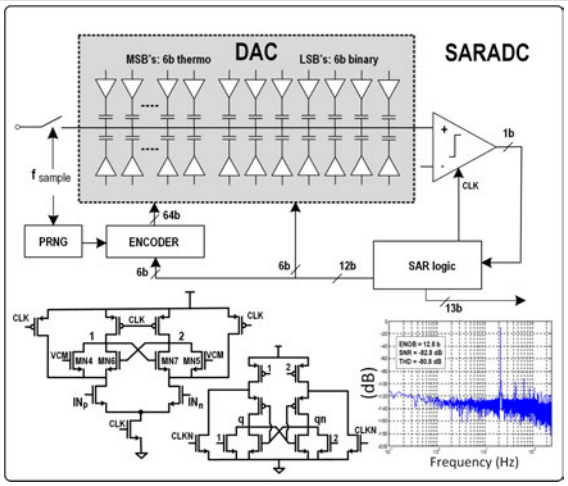


Figure 22.1.4: A 13b SARADC used in ECG, BIOZ and PPG channel. By adding cascode transistors MN4 and MN5 and gain boosting transistor MN6 and MN7 the preamp gain is increased to 52× and becomes less time-dependent. An ENOB of 12.8b and an SNR of 82.8dB are achieved.

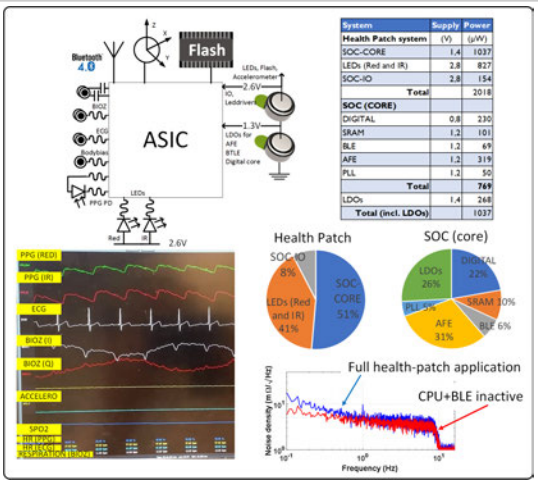


Figure 22.1.5: An health-patch setup executing a typical application (ECG, BIOZ, and PPG) consumes on average 2mW, powered by two ZnAir coin-cell batteries. The SOC consumes 769µW. The noise density of analog readouts is slightly (~10%) affected when radio and digital processing are active.

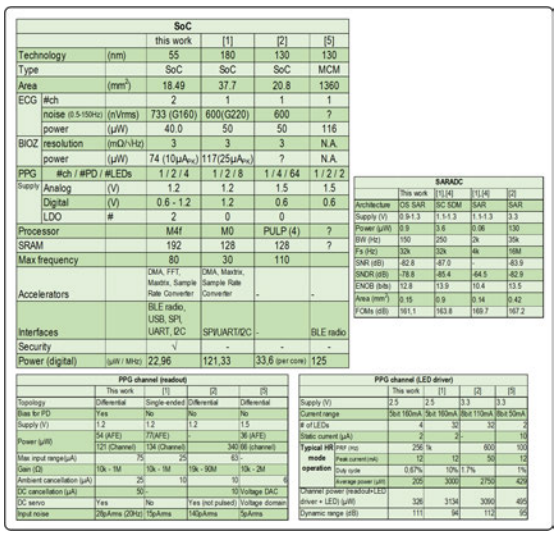


Figure 22.1.6: Benchmark tables for SoC, PPG channel (readout and LED driver) and SARADC.

22

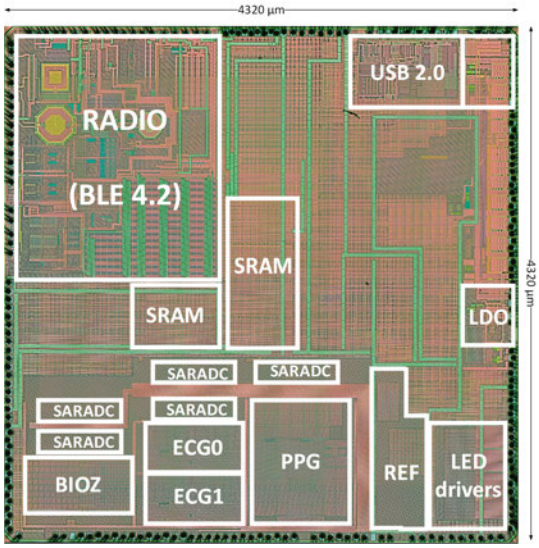


Figure 22.1.7: Die photo. The size of the die is 4320×4320µm².

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

22.2 A Rugged Wearable Modular ExG Platform Employing a Distributed Scalable Multi-Channel FM-ADC Achieving 101dB Input Dynamic Range and Motion-Artifact Resilience

Julian Warchall[1], Paul Theilmann[2], Yuxuan Ouyang[2], Harinath Garudadri[1], Patrick P. Mercier[1]

[1]University of California, San Diego, La Jolla, CA
[2]MaXentric Technologies, La Jolla, CA

Wearable ExG biopotential acquisition systems can potentially capture a wealth of clinically useful diagnostic information during activities of daily life. In practice, however, motions from common activities introduce large artifacts that can easily saturate traditional analog front-ends (AFEs) designed to sense biopotentials on the micro- to milli-volt scale. In addition, the wires that connect each electrode to an array of high-impedance AFEs can easily pick up interference and large, potentially saturating artifacts. For these reasons, many-channel monolithic biopotential sensing systems are often fragile and difficult to use in ambulatory environments. While active electrodes can be used to combat interference picked up by high-impedance wires, they require power-hungry drivers to deliver high-fidelity signals across relevant anatomy to an array of ADCs. Placing an ADC on each active electrode followed by a digital bus driver can eliminate analog driver power, resulting in a per-channel power consumption of 104μW in [1]. However, the 12b SAR ADC in [1] could not tolerate significant motion artifacts, and further increasing the ADC resolution to accommodate a larger dynamic range (DR) would require a quadratic increase in per-channel ADC power consumption.

This paper presents a wearable *N*-channel FM-ExG platform, shown in Fig. 22.2.1, that achieves high-DR motion-artifact-resilient yet low-power operation in a modular and rugged form factor by: 1) coupling each ExG electrode to an active IC comprising a sinusoidal VCO, whose output is a bandwidth-expanded frequency-modulation (FM) signal, with FM coding gain cubically proportional to the bandwidth expansion factor; 2) aggregating multiple FM signals with frequency division multiplexing (FDM) over a single wire via energy-efficient tuned-amplifiers, where each FM-ExG signal rests at a distinct center frequency set by each VCO; 3) exploiting VCO-enabled FM coding gain to digitize all *N* channels simultaneously with a single 12b SAR ADC at a wearable gateway while achieving a per-channel SNR of 101dB; 4) operating the SAR ADC at the aggregate FM spectrum bandwidth (1.28MHz) rather than Nyquist (~30MHz) by exploiting the low-harmonic nature of the sinusoidal VCO and bandpass sampling, reducing ADC power from ~5.7mW to 190μW; and 5) transmitting the digitized FM spectrum via an energy-efficient UWB TX based on [2] for off-body processing and storage. While the distributed FM-ADC-like [3] nature of the system is similar to [4], this paper represents the first IC-based embodiment that fully realizes advantages in energy efficiency and system motion artifact reduction.

As depicted in Fig. 22.2.1, *N* active electrode ICs are daisy-chained via a 4-wire bus consisting of $V_{DD,}$ GND, an on-body reference potential, and the FDM wire. Each active electrode IC is attached to a skin-contact electrode via a coupling capacitor, and up-converts the ~250Hz BW biopotential voltage to a 50kHz BW FM-modulated signal. In practice, FM-modulated signals are sensitive to phase, and thus system SNR is limited either by the lower bound between FM coding gain and phase noise. To maintain >100dB SNR for the given FM coding gain, simulations in Fig. 22.2.2 (top right) reveal a phase noise requirement for the VCO of <-120dBc/Hz at a 10kHz offset. This requirement is achieved at low power by utilizing an external 0603 inductor with *Q*=50 in a folded cross-coupled 14MHz oscillator (Fig 2 bottom left). To reduce power by a factor of 2.7×, the oscillator utilizes the minimum size transistors necessary to maintain oscillation after larger start-up transistors are switched out. Since the VCO cannot directly drive the FDM wire, a bandpass amplifier tuned by another 0603 inductor and varactor followed by a source-follower for additional isolation is employed (Fig. 22.2.2 bottom right), for a net savings of 80× in power over a conventional wideband driver.

Figure 22.2.3 (top left) shows the spectrum of a single FM channel driven by a full-scale 10mV 10Hz sinusoidal test message, while Fig. 22.2.3 (top right) shows the signal after demodulation, where worst-case VCO V-to-F nonlinearities are shown. The demodulated SNR and SNDR for varying amplitudes are shown in Fig. 22.2.3 (bottom left), demonstrating a peak SNR of 101dB, a peak SNDR of 84dB, and a useable DR of 101dB. This design does not employ a gain amplifier prior to the VCO, and instead directly up-converts the electrode's baseband signal to the FM domain. This feature provides resilience to out-of-range artifacts or disturbances if sufficient guard band in the FM domain is available. To demonstrate this, a 100 mV DC step (ostensibly 10× the full-scale range of the VCO's input) is applied concurrent with a 10mV test sinusoid. The resulting demodulated output, shown in Fig. 22.2.3 (bottom right) with and without the input coupling capacitor, tracks the 10mV test sinusoid well throughout the disturbance.

To demonstrate multi-channel access, six active electrodes were daisy-chained and driven with six different 1mV test sinusoids at 10, 50, 90, 130, 170, and 210Hz; the spectrum of the FDM wire here is shown in Fig. 22.2.4 (top). After being received by the SAR ADC sampling at 1.28MS/s (thanks to bandpass sampling) and after demodulation, the six message signals are individually distinguishable without any cross-talk as shown in Fig. 22.2.4 (bottom). An arbitrary number of active electrodes can be daisy-chained to realize massively parallel ExG acquisition by increasing the ADC sampling frequency to accommodate the added FM bandwidth.

The short 4-wire bus used to chain active electrode ICs together and connect to the gateway IC affords a rugged structure unlike other ExG systems previously published other than [1]. To demonstrate the advantages offered, the FM-ExG system and an OpenBCI system were connected side-by-side to measure ECG on an ambulatory subject. Both systems recorded a differential waveform with six electrodes placed on the chest. As is clearly seen in transient and spectrum recordings from the V2 chest electrode in Fig. 22.2.5 (top), the FM-EXG system tolerated subject motion much better than the OpenBCI system – 2.9× better according to normalized cross-correlations between still and motion data. The FM-ExG system was also validated in a SSVEP EEG-based experiment; Figure 22.2.5 (bottom) shows the results acquired after the UWB output correctly indicating a 10Hz component that corresponds to the 10Hz screen flashing in the SSVEP experiment.

The active electrode ICs and gateway IC are both fabricated in a 65nm process. The active electrode ICs expend 212μW/channel, while the 12b SAR ADC in the gateway IC consumes 190μW at 1.28MS/s, both at 1.2V. The UWB TX, whose FCC-complying output spectrum is shown in Fig. 22.2.6 (top right), consumes 11pJ/pulse, and 8 pulses are transmitted per bit for a total energy cost of 88pJ/b. For 101dB SNR operation, each FM channel occupies 50kHz, and with 50kHz of guard band between channels, 12 total channels are supported in the developed FM-ExG system (with more channels capable simply by increasing the SAR sampling rate). The digitization of multiple channels via a single low-power SAR ADC cements the power advantage of the proposed system, where each channel achieves 101dB SNR, yet with an aggregate power consumption of only 228μW/channel without the UWB TX, and 340μW/channel with. Despite the increased bit rate per channel due to FM-based oversampling and the corresponding extra energy required from the UWB TX to transmit at a higher data rate, the power efficiency and DR of the system reflected in the Schreier FoM exceeds that of prior-art active electrode systems that include the power of electrode-to-gateway drivers, and also exceeds conventional AFE arrays that do not (Fig. 22.2.6, bottom). Note that in all cases the 88pJ/b energy from the UWB TX is assumed in order to normalize TX performance across references. A photograph of the FM-ExG active electrode board is shown in Fig. 22.2.6 (top left), while die photos are shown in Fig. 22.2.7.

Acknowledgements:
This work was supported in part by the U.S. Army Research Laboratory under Contract W911QX-16-C-0003, and in part by the National Science Foundation Graduate Research Fellowship under Grant DGE-1144086. The authors acknowledge W. David Hairston for help and advice.

References:
[1] J. Xu, et al., "24.7 A 60nV/√Hz 15-Channel Digital Active Electrode System for Portable Biopotential Signal Acquisition," *ISSCC Dig. Tech. Papers*, Feb. 2014.
[2] A. T. Phan, et al., "Energy-Efficient Low-Complexity CMOS Pulse Generator for Multiband UWB Impulse Radio," *IEEE TCAS-I*, vol. 55, no. 11, pp. 3552-3563, 2008.
[3] J. Warchall, et al., "A 678-μW Frequency-Modulation-Based ADC with 104-dB Dynamic Range in 44-kHz Bandwidth," *IEEE TCAS-II*,vol. 65, no. 10, pp. 1370-1374, 2018.
[4] J. Warchall, et al., "A Multi Channel EEG System Featuring Single-Wire Data Aggregation via FM-FDM Techniques," *IEEE ISCAS*, pp. 526-529, 2016.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

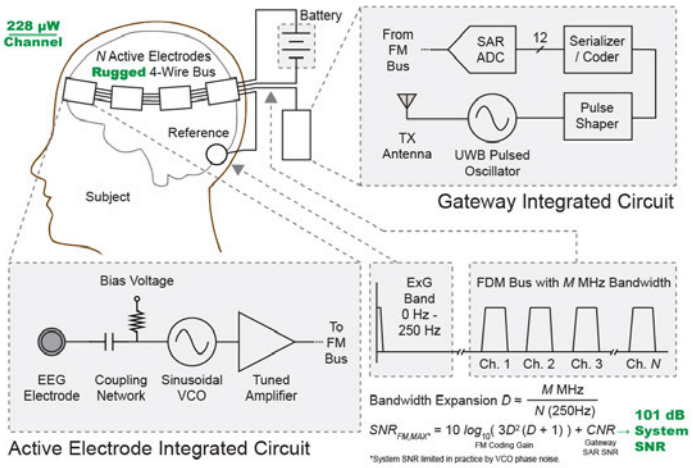


Figure 22.2.1: Top-level diagram of the FM-ExG system used for EEG with FM plan depicting N-channel frequency domain multiplexing.

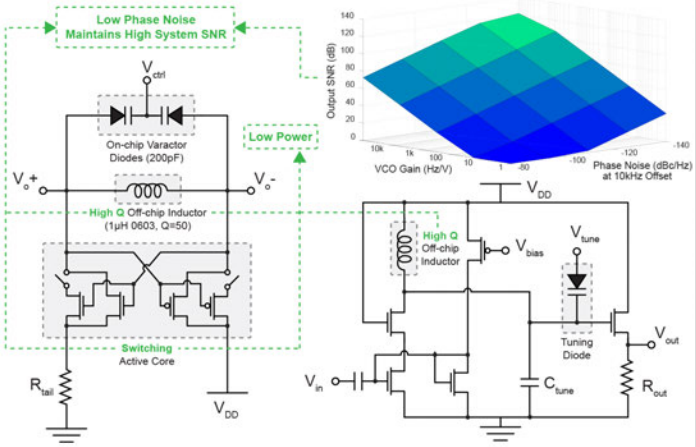


Figure 22.2.2: Circuit diagram of FM voltage-controlled oscillator (left); simulated effect of phase noise on an FM-FDM system's SNR (top right); circuit diagram of FM bus driving tuned bandpass amplifier (bottom right).

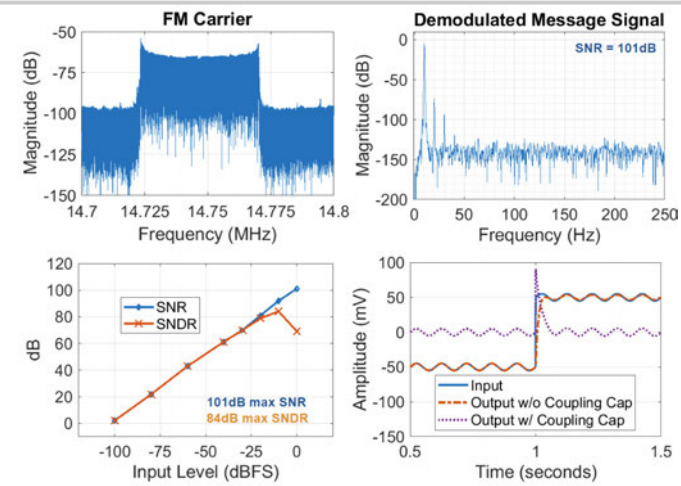


Figure 22.2.3: FFT of single channel driven with full-scale (10mV) 10Hz sine message (top left); same channel after demodulation (top right); system SNR and SNDR vs. input signal level (bottom left); demodulated transient step input interference for single channel configuration (bottom right).

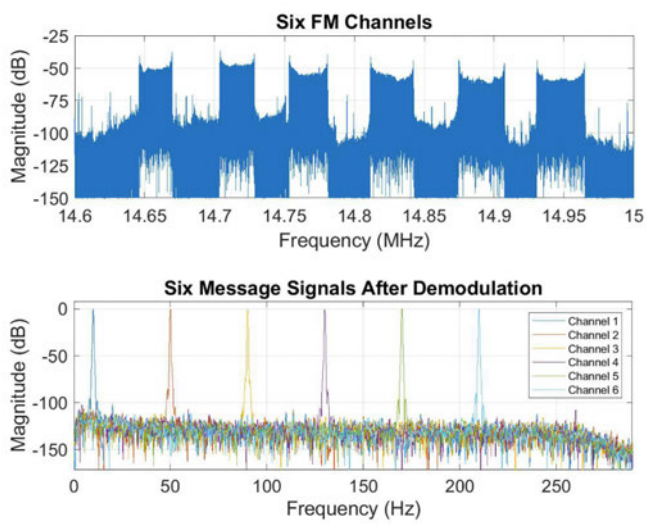


Figure 22.2.4: FFT of FDM bus for six channel test with various sine inputs (top); the six channels after demodulation, demonstrating spur-free DR (bottom).

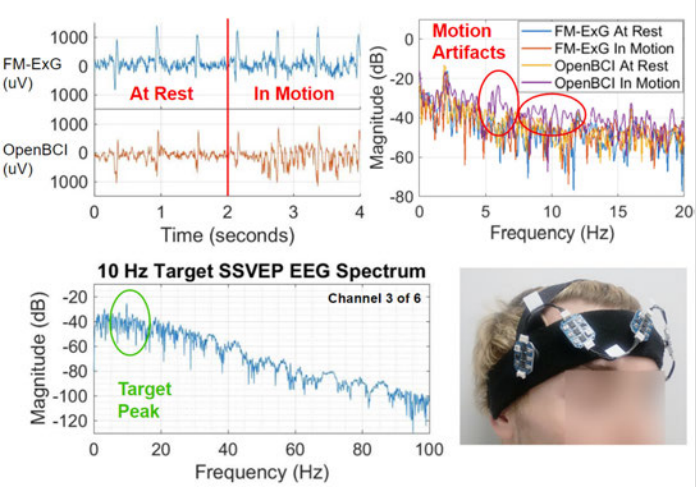
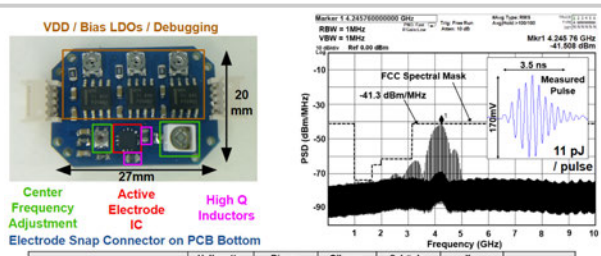


Figure 22.2.5: Demonstration of FM-ExG resilience to motion artifacts in ECG (top); FM-ExG system arranged for EEG application (bottom right); EEG spectrum from 10Hz SSVEP experiment using FM-ExG (bottom left).

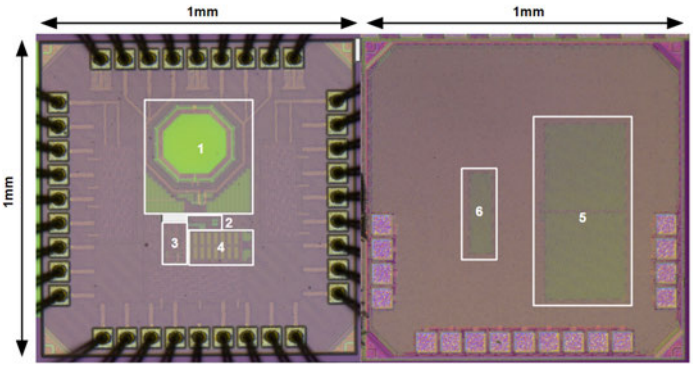


Design	Helleputte, JSSC 2015	Rieger, TCAS-I 2018	O'Leary, ISSCC 2018	Schönle, JSSC 2018	Xu, ISSCC 2014	This Work
Active Electrode?	No				Yes	
Process [nm]	180	180	130	130	180	65
IC Active Area [mm²]	49	0.06	7.6	20.8	15.8	1
Electrode Bus Drive Type	-	-	-	-	Digital I2C	FDM via Tuned Amplifier
Channels / Chip	5	3	32	9	1	1
IRN [uVrms] (BW)	0.61 (150Hz)	0.67 (100Hz)	1.6 (500Hz)	1.21 (150Hz)	0.65 (100Hz)	0.99 (250Hz) 0.63 (100Hz)
Acquisition Bandwidth [Hz]	250	7000	500	4000	100	250
Usable DR at Input [dB]	~83 (13.5-bit)	~50 (8-bit)	70	~83 (13.5-bit)	~74 (12-bit)	101
Output Bit Rate per Channel [Hz]	32000	56000	1000000	112000	2400	1280000
Acquisition Power per Channel [µW]	56	96.7	1.26	285	104.4	228
Calculated Wireless TX Power* per Channel [µW]	2.82	4.93	88	9.86	0.21	112.6
Total Power per Channel [µW]	58.82	101.63	89.26	294.86	104.61	340.6
FOM = Usable DR [dB] + 10 log10(BW [Hz] / Total Power per Channel [W])	149.3	128.4	137.5	154.3	133.8	159.7

*Calculated using 88 pJ/bit at digital output bit rate

Figure 22.2.6: Detail of active electrode application board (top left); Measured UWB pulse PSD showing FCC mask compliance (top right); Table of comparison (bottom).

22



1 - UWB TX
2- Pulse Shaper
3 - Serializer/Coder
4 - SAR ADC
5 - Sinusoidal VCO
6 - Bandpass Tuned Amplifier

Figure 22.2.7: Die photo of gateway IC including SAR ADC and UWB TX (left); Die photo of active electrode IC including sinusoidal VCO and tuned amplifier (right).

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

22.3 A 0.5V 9.26μW 15.28mΩ/√Hz Bio-Impedance Sensor IC with 0.55° Overall Phase Error

Kwantae Kim, Ji-Hoon Kim, Surin Gweon, Jiwon Lee, Minseo Kim, Yongsu Lee, Soyeon Kim, Hoi-Jun Yoo

KAIST, Daejeon, Korea

Continuous blood-status monitoring by thoracic impedance or impedance phase measurement [1] is critical in reducing chronic heart failure (CHF) because the symptoms of deterioration are not recognized by patients, thus preventing timely treatment. Therefore, a wearable/implantable bio-impedance (Bio-Z) sensor is required for ambulatory patients to detect early signs of deterioration for appropriate intervention in proper time. Continuous monitoring requires ultra-low-power consumption (<10μW) and it requires the high resolution (<0.1Ω) to detect small impedance variations [2]. Prior Bio-Z sensor ICs [2-4] do not satisfy these requirements simultaneously because the resolution depends inversely on the power consumption. Furthermore, previous Bio-Z sensors have a large phase error that is unable to accurately (<1°) and continuously measure phase due to the square-wave modulation scheme of the current generator [3], resulting in intolerable measurement errors [2], or the supply and temperature variations of measurement accuracy due to the limited bandwidth (BW) of readout [2]. Although the pre-demodulation [3] technique can relax BW requirement, conversion loss of sine to square wave demodulation causes SNR degradation by 2/π times, leading to 2.47× increased power consumption.

In this paper, a sub-10μW, 15.28mΩ/√Hz Bio-Z sensor IC is presented for long-term wearable/implantable blood-status monitoring with 0.55° phase error for 20kHz modulation frequency. The instrumentation amplifier (IA) operates at 0.5V supply and exhibits 45nV/√Hz noise performance, enhancing resolution. Further, a lead-compensated TC stage increases its BW up to 408kHz, while minimizing the total phase error even under supply and temperature variations. The LDO-based V-I converter also operates at 0.5V supply, enabling ultra-low-power operation.

Figure 22.3.1 shows the overall architecture of Bio-Z sensor IC, which is composed of 5 components: (1) an instrumentation amplifier (IA) for low-noise and low-phase-shift Bio-Z readout, (2) a LDO-based V-I converter for low-power current injection, (3) a look-up table (LUT), a voltage DAC, and a g_m-C low pass filter for low harmonic sinusoidal synthesis, (4) an I/Q clock generator to create the quadrature demodulation clock that is phase-aligned to the injection current, and (5) a body bias driver, a voltage, and a current references.

Figure 22.3.2 shows the circuit schematics of the current-feedback IA. To achieve low-supply IA operation, the solution in [4] consumes significant power (15.96μW) with replica-based adaptive biasing of the transconductance (TC) stage. In this work, a common-mode feedback (CMFB) loop is used to bias the drain voltage of the input transistor. The common-mode signal is extracted by pseudo-resistors and fed back to g_{m2} amplifier to control the amount of current of g_{m3}, biasing V_{GS} of g_{mff} to V_{REFTC}. Furthermore, all branches of the TC stage are designed to operate with only 3-transistor stacking for a 0.5V supply. This is achieved by midpoint voltage (TI_{Drain}) sensing of the input resistor (R_{IN}) rather than a cascode-based sensing method [4], to regulate the current mirror of the transimpedance (TI) stage. To achieve low noise, the TC stage should have low R_{IN} [5] which, however, results in a difference in loop dynamics between CM and DM. Since the R_{IN} forms a high impedance for the CM loop and a low impedance (2kΩ) for the DM loop, it is difficult to compensate the CM loop rather than the DM loop. In this work, a lead-compensated TC stage is proposed to fully utilize DM loop gain-bandwidth (GBW), compared with simple Miller compensation in previous work [5], which sacrifices DM loop GBW to achieve a stable CM loop response.

Figure 22.3.3 shows the lead-compensated TC stage, which is composed of 2 parts: (1) nested Miller compensation with feedforward stage and nulling resistor (NMCFNR) [6] for stabilization of 3-stage CMFB loop (g_{m1}-g_{m2}-g_{m3} with g_{mff}); and (2) lead compensation (LC) to stabilize for both of CM and DM loop. Since the LC uses a capacitive phase leading path instead of reducing GBW [5], it leads to not only a maximized GBW of DM loop but also sufficient phase margin of CM loop. The simulated DM loop response shows an increase in GBW of 5.85×, and its loop gain is enhanced by 17dB for 20kHz modulation frequency with the same CM loop phase margin, compared to previous design [5].

Figure 22.3.4 shows the schematic diagram of the LDO-based V-I converter. Two differential flipped voltage follower (FVF) LDOs with tri-input OTA [7] reject the even harmonics and enhance the accuracy of LDO output. The output stage is separated from FVF feedback loop to avoid parasitic effects of PCB and pads using the M_{1-3} current mirror. For 0.5V supply operation, the output load current is fixed to 8.4μA_{PP} relaxing V_{GS} variation of M_2 and a low-V_{TH} MOS design of M_{0-1} reduces V_{GS} to maximize the headroom. When the input signals are buffered to the $V_{2P,N}$, a voltage difference is generated across resistor R_{VI}. Since both of the DC and the AC voltages are converted to current by R_{VI}, residual DC currents ($I_{Residue}$) cause additional power dissipation. In this work, an auxiliary LDO is used to reduce $I_{Residue}$ by driving a predefined DC voltage (V_{AUX}) to R_{VI} instead of V_{DD} rail. The V_{AUX} voltage is set to nearly maximum swing level of $V_{2P,N}$ ($V_{2,MAX}$) by an on-chip voltage reference. As a result, $I_{Residue}$ is reduced from 12.6μA to 0.16μA. In addition, a bang-bang digital PLL-based I/Q clock generator is implemented which is phase-aligned to $V_{2P,N}$, providing a quadrature demodulation clock to IA.

Figure 22.3.5 shows the measurements of IA gain curve, overall impedance resolution, and output spectrum of the V-I converter. The IA exhibits 408kHz BW consuming only 3.95μW, which has a 16.93× improvement in power efficiency compared to [5]. The overall impedance resolution floor is measured to be 15.28mΩ/√Hz at 1Hz with 100Ω test resistor. The output spectrum of the V-I converter shows 0.66% THD performance and 44.64dBc SFDR performance, which comes from residual harmonics of pseudo-sine synthesis.

Figure 22.3.6 shows the measured phase error performance from 3 chips. All chips are tested over 10% supply and 0-to-70°C temperature variation with 20kHz modulation frequency. After 1-point trimming, it shows that the maximum and minimum value of total phase error are 0.59°, 0.52° respectively, and an average value of 0.55° is achieved. The Bio-Z sensor IC is also validated by in-vivo test, successfully measuring the thoracic impedance variance (TIV) and impedance cardiography (ICG), exhibiting the important points such as systolic peak, diastolic peak, and dicrotic notch of TIV and B, C, X-points of ICG. As a result, heart rates and left-ventricular ejection time are extracted.

Figure 22.3.7 shows the chip micrograph and performance comparison table. The Bio-Z sensor IC is implemented in 65nm CMOS and occupies 4.83mm² with 9.26μW power consumption and 15.28mΩ/√Hz resolution. It achieves the lowest power-resolution figure of merit (FoM) in the table while exhibiting the lowest phase error for continuous blood-status monitoring applications.

References:

[1] L. C. Martínez, et al., "Bioelectrical Impedance and Strength Measurements in Patients with Heart Failure: Comparison with Functional Class," *Nutrition*, vol. 23, no. 5, pp. 412-418, May 2007.

[2] S. Kim, et al., "A 20μW Intra-Cardiac Signal-Processing IC with 82dB Bio-Impedance Measurement Dynamic Range and Analog Feature Extraction for Ventricular Fibrillation Detection," *ISSCC Dig. Tech. Papers*, pp. 302-303, Feb. 2013.

[3] J. Xu, et al., "A 0.6V 3.8μW ECG/Bio-Impedance Monitoring IC for Disposable Health Patch in 40nm CMOS," *CICC*, 4 pages, Apr. 2018.

[4] N.V. Helleputte, et al., "A Multi-Parameter Signal-Acquisition SoC for Connected Personal Health Applications," *ISSCC Dig. Tech. Papers*, pp. 314-315, Feb. 2014.

[5] M. Kim, et al., "A 1.4mΩ-Sensitivity 94dB-Dynamic-Range Electrical Impedance Tomography SoC and 48-Channel Hub SoC for 3D Lung Ventilation Monitoring System," *ISSCC Dig. Tech. Papers*, pp. 354-355, Feb. 2017.

[6] K.N. Leung and P.K.T. Mok, "Nested Miller Compensation in Low-Power CMOS Design," *IEEE TCAS-II*, vol. 48, no. 4, pp. 388-394, Apr. 2001.

[7] Y. Lu, et al., "A 0.65ns-Response-Time 3.01ps FOM Fully-Integrated Low-Dropout Regulator with Full-Spectrum Power-Supply-Rejection for Wideband Communication Systems," *ISSCC Dig. Tech. Papers*, pp. 306-307, Feb, 2014.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

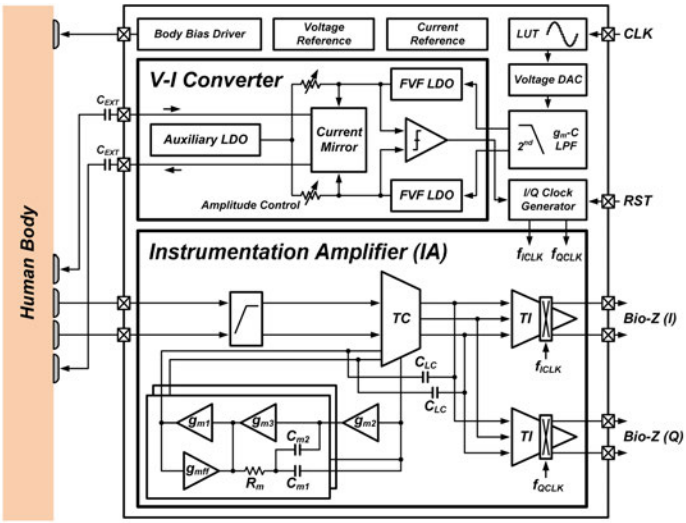


Figure 22.3.1: Overall architecture of bio-impedance (Bio-Z) sensor IC.

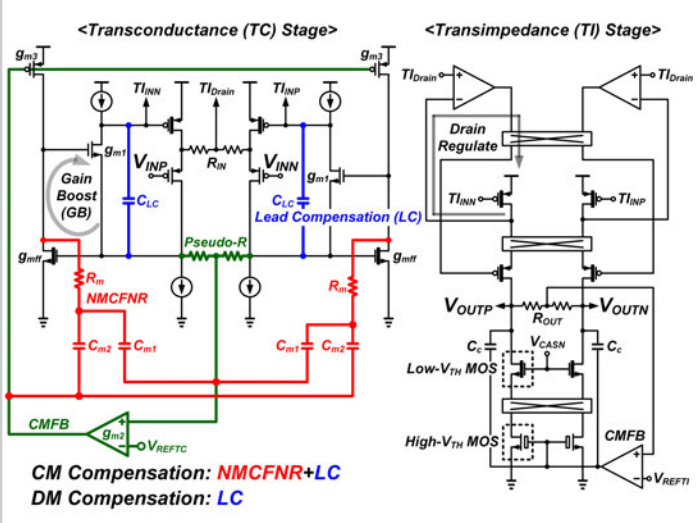


Figure 22.3.2: 0.5V-supply current-feedback instrumentation amplifier (IA).

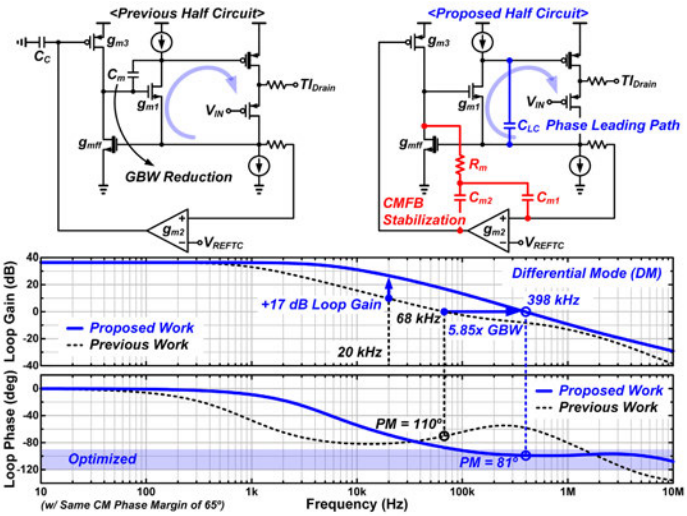


Figure 22.3.3: Lead-compensated TC stage with NMCFNR.

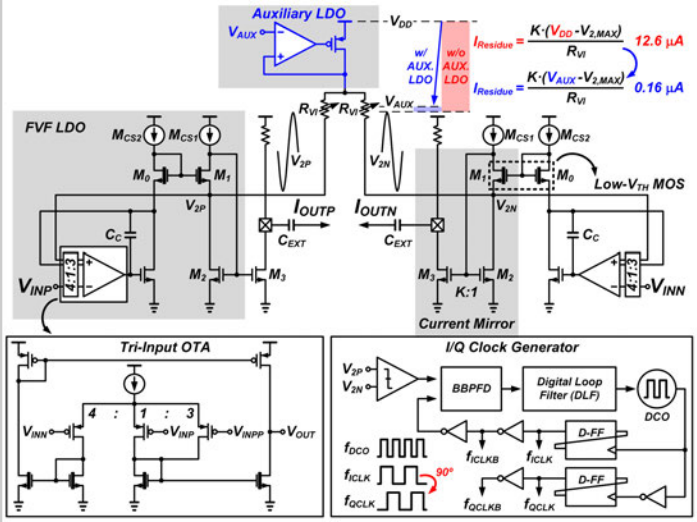


Figure 22.3.4: 0.5V-supply LDO-based V-I converter and I/Q clock generator.

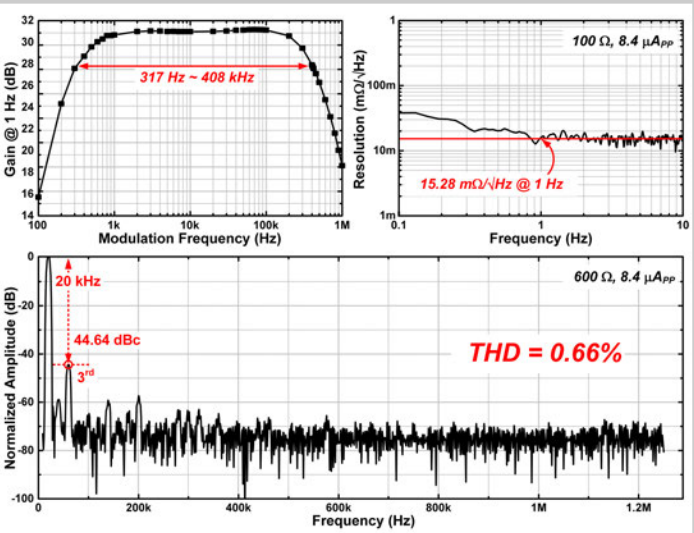


Figure 22.3.5: Measured IA gain curve (upper left), overall impedance resolution (upper right), and output spectrum of V-I converter (lower).

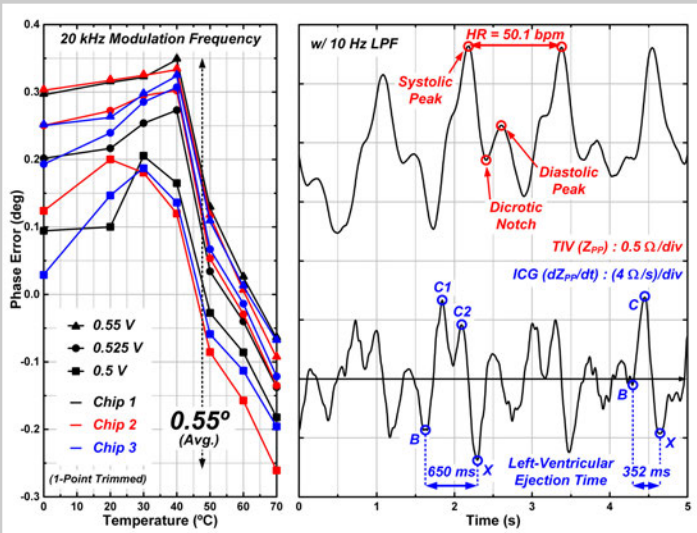
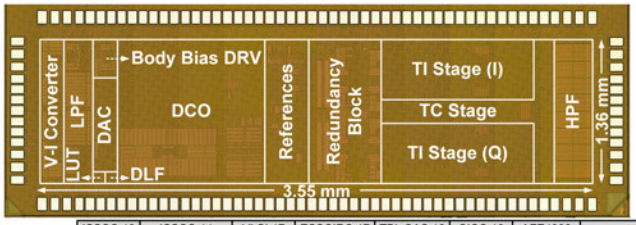


Figure 22.3.6: Measured supply and temperature sensitivity of phase error (left) and in-vivo TIV and ICG results (right).



	ISSCC 13	ISSCC 14	VLSI 17	ESSCIRC 17	TBioCAS 18	CICC 18	AFE4300	*This Work*
	S. Kim	N. V. Helleputte	H. Ha	K. Kim	M. Zamani	J. Xu	TI	
Process (nm)	180	180	180	180	180	40	-	***65***
Supply (V)	1.8	1.2	1.2/1.8	1.8	1	0.6	3	***0.5***
Phase Measure	Yes	Yes	Yes	No	Yes	No	Yes	***Yes***
Modulation	Pseudo-Sine	Pseudo-Sine	Square	Pseudo-Sine	Semi-Ramp	Square	Pseudo-Sine	***Pseudo-Sine***
Frequency (kHz)	20	20	1-1024	20	2	32	8/16/32/64	***20/50***
Injection Current (μA_{PP})	40	36	10	26	1	20	1060	***8.4***
Power (μW)*	56.2	58	81.7	24	1.55	15.78	2910	***9.26***
Resolution ($m\Omega/\sqrt{Hz}$)	8.8	8.6	3.84	12.49	222	35	70	***15.28***
Phase Error (°)	-	-	-	-	-	-	<0.69	***<0.55***
FoM†	495	499	314	300	344	553	203700	***141***

*Including Current Generator & Current Injection Power
†FoM = (Total Power)×(Resolution) [$\mu W \times m\Omega/\sqrt{Hz}$]

Figure 22.3.7: Chip micrograph and performance comparison table.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

22.4 A 27.8μW Biopotential Amplifier Tolerant to $30V_{pp}$ Common-Mode Interference for Two-Electrode ECG Recording in 0.18μm CMOS

Nahmil Koo, SeongHwan Cho

KAIST, Daejeon, Korea

Two-electrode ECG devices have gained popularity in the recent past to enable comfortable and long-term monitoring of cardiovascular health. As a ground or bias electrode is not used in a two-electrode ECG device, common-mode interference (CMI) caused by powerline coupling to the human body can be as large as a few tens of volts. Such a large CMI ruins the ECG recording, and thus the analog front-end of the ECG device must be immune to large CMI.

The input common-mode range of a front-end amplifier is limited as the input transistors must operate in the saturation region. Although a CM cancellation path can increase the CM range, it is useful only when the CMI is smaller than the supply voltage [1,3]. For large CMI that exceeds the supply rail, the ESD diode is turned on, and the input is saturated to supply or ground. Unfortunately, current state-of-the-art ECG devices available commercially or in research papers do not have a solution to this problem. In this paper, we present a biopotential amplifier for a wearable ECG device that can handle CMI up to 30V at 60Hz under a 1.2V supply. We propose a charge-pump-based common-mode feedback loop, which suppresses CMI so that a reliable two-electrode ECG recording is possible for dry or wet electrodes.

A basic diagram of the biopotential amplifier under CMI is shown in Fig. 22.4.1, where powerline coupling to the human body and dry electrodes are modeled by C_C and $R_{EL}||C_{EL}$, respectively. The key component of the proposed work is a common-mode charge pump (CMCP) which absorbs unwanted displacement current from the power line (I_{PL}). It adds a fixed voltage V_F that is negative, positive or zero to the input so that the input CM voltage (V_{INCM}) stays within the desired range. The CMCP controller determines the polarity of V_F according to V_{INCM}.

The operation principle of the CMCP is shown in Fig. 22.4.1 for the case when V_{IN} is larger than the upper threshold. A single-ended schematic is drawn for simplicity. During ϕ_1, a unity-gain buffer pre-charges the flying capacitor (C_P) to $V_{IN} - V_{DD}$. During ϕ_2, the bottom node of C_P is switched from V_{DD} to ground, thus creating a negative V_F that is added to V_{IN}. The magnitude of V_F is set by the voltage division of V_{DD} between $Z_{EL}+Z_C$ and Z_P. When V_{INCM} is smaller than the desired threshold, the bottom plate of C_P is switched from the ground to the supply, and thus a positive V_F is added to V_{INCM}. When V_{INCM} is within the desired range, the bottom plate of C_P is always tied to the ground, and thus V_{INCM} is not changed. To keep the V_{INCM} within the desired range, $|V_F|$ should be large enough or subtracted frequently enough from V_{INCM}. Since $|V_F|$ cannot be controlled, the clock frequency of the CMCP (f_{CLK}) is set high enough to ensure that V_{INCM} stays within the desired range. The differential mode (DM) is not affected by the CMCP since the CMCP pre-charges the top plate of C_P with the same voltage as the input node.

The schematic and timing diagram of the amplifier are shown in Fig. 22.4.2. The signal is amplified by using two stages of amplifiers, which are implemented using a capacitively coupled instrumentation amplifier (CCIA) with a PMOS-input folded-cascode OTA. The CMCP controller consists of two parts: a common-mode level detector and a digital control logic. The CM level detector sends flags (UP and DN) to the control logic to decide whether V_{INCM} should be increased, decreased, or stay still. The buffers and resistors produce V_{INCM} from differential inputs (V_{IN+} and V_{IN-}), and the comparators detect whether V_{INCM} exceeds the pre-defined threshold. Based on the output of the comparators, the control logic generates ϕ_P to control the voltage at the bottom of C_P.

There are a couple of non-idealities in the circuit that may affect its operation. First, 1/f noise of the unity-gain buffer and noise from switches (e.g., charge injection) have a severe effect on the input-referred noise due to the accumulating nature of CMCP. To reduce the noise from the unity gain buffer, $CHOP_1$ is used for chopping. To deal with the noise from the switches, the clock frequency of the CMCP is increased so that it moves out of the ECG band. Second, the mismatch between the CMCP flying capacitors generates an undesired CM-to-DM conversion. This is addressed by inserting $CHOP_2$, which operates as a 1b DEM to average the C_P mismatch as shown in the timing diagram.

The IC is implemented in a 0.18μm standard CMOS process. To verify the amplifier, function generators and components are set up as shown in Fig. 22.4.3 to mimic the human body ECG model shown in Fig. 22.4.1. The component values used are typical parameters verified in [2,6]. Note that capacitors and resistors with 5% mismatch are deliberately used to model the mismatch in the electrodes. The clock frequencies for the CMCP, $CHOP_1$, and $CHOP_2$ are 100kHz, 2.5kHz, and 50kHz, respectively. The transient waveform of the input CM is shown when a $10V_{pp}$ at 60Hz signal is applied as the CMI. It can be seen that the CMCP loop is activated when the V_{INCM} exceeds the threshold and thus keeps it below the threshold. Tolerance to large CMI is measured by observing the gain and CMRR as we increase the CMI at 60Hz when the differential input is 1mV at 10Hz. It can be seen that the proposed technique allows the gain and CMRR to be maintained until the CMI reaches 30V. By contrast, the amplifier without CMCP is not usable after the CMI exceeds 1.2V which is when the input transistors are no longer in the saturation region.

The transfer function of the amplifier is shown in Fig. 22.4.4 for two cases, with and without a $30V_{pp}$ CMI at 60Hz. The difference between the two cases is less than 2%, which implies that the amplifier operates robustly under large CMI. To see the impact of CMCP on noise, input-referred noise is measured and shown in Fig. 22.4.4 for two cases, with and without CMCP. The integrated noise in the ECG signal band within 100Hz is $5.04\mu V_{rms}$ and $5.05\ \mu V_{rms}$ when CMCP is turned off and on, respectively. The tones that result from the accumulation of noise from switches can be filtered out as they are above the ECG signal band.

Figure 22.4.5 shows the result of ECG measurement on a human subject. The ECG is measured using only two dry-electrodes at the chest. A powerline is placed about 50cm near the leg. The amplitude of the CMI measured at the electrode is about 15Vpp at 60Hz. With the proposed technique, the CM swing at the input node is suppressed and the amplifier output shows P-QRS-T complex after a low-pass filter with 45Hz cut-off frequency. Without the proposed technique, the amplifier is saturated by the large CMI, and the output does not show any QRS complex even after the low-pass filter.

The performance of this work is summarized and compared with the previous state-of-art biopotential amplifiers in Fig. 22.4.6. It can be seen that all previous amplifiers have CMI tolerance that is below the supply voltage. On the other hand, the tolerance of the proposed biopotential amplifier is more than 20 times the supply. The chip micrograph is shown in Fig. 22.4.7.

Acknowledgements:
This work was supported in part by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2018R1A2B6005773) and in part by the Nano Material Technology Development Program through the NRF, Ministry of Science, ICT and Future Planning (NRF-2016M3A7B4910637).

References:
[1] H. Chandrakumar, et al., “An 80-mVpp Linear-Input Range, 1.6GΩ Input Impedance, Low-Power Chopper Amplifier for Closed-Loop Neural Recording That Is Tolerant to 650-mV_{pp} Common-Mode Interference," *IEEE JSSC*, vol. 52, no. 11, pp. 2811-2818, Nov. 2017.
[2] M. Chen, et al., “A 400GΩ Input-Impedance, $220mV_{pp}$ Linear-Input-Range, $2.8V_{pp}$ CM-Interference-Tolerant Active Electrode for Non-Contact Capacitively Coupled ECG Acquisition,” *Symp. VLSI Circuits*, pp. 129-130, 2018.
[3] J. Xu, et al., “A 0.6V 3.8μW ECG/Bio-Impedance Monitoring IC for Disposable Health Patch in 40nm CMOS,” *CICC*, 4 pages, 2018.
[4] Y. P. Chen, et al., “An Injectable 64nW ECG Mixed-Signal SoC in 65nm for Arrhythmia Monitoring,” *IEEE JSSC*, vol. 50, no. 1, pp. 375-390, Jan. 2015.
[5] J. Lee, et al., “An Ultra-High Input Impedance Analog Front End Using Self-Calibrated Positive Feedback,” *IEEE JSSC,* vol.53, no. 8, pp. 2252-2262, Aug. 2018.
[6] Y. M. Chi, *et al.* “Dry-Contact and Noncontact Biopotential Electrodes: Methodological Review,” *IEEE Reviews in Biomedical Engineering*, vol. 3, no. 1, pp. 106-119, Oct. 2010.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

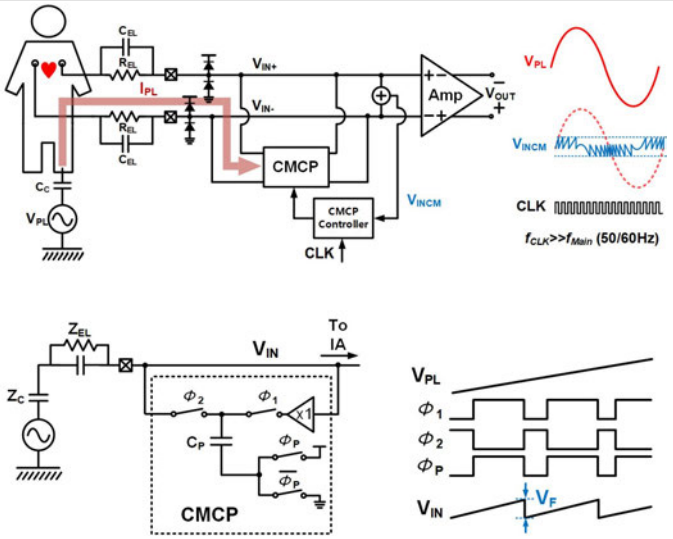


Figure 22.4.1: Concept of the proposed biopotential amplifier and simplified operation of the common-mode charge pump (bottom).

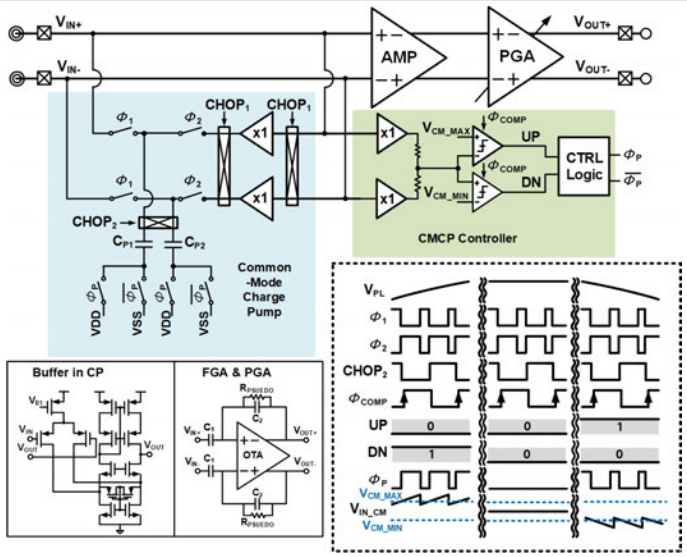


Figure 22.4.2: Overall schematic and timing diagram.

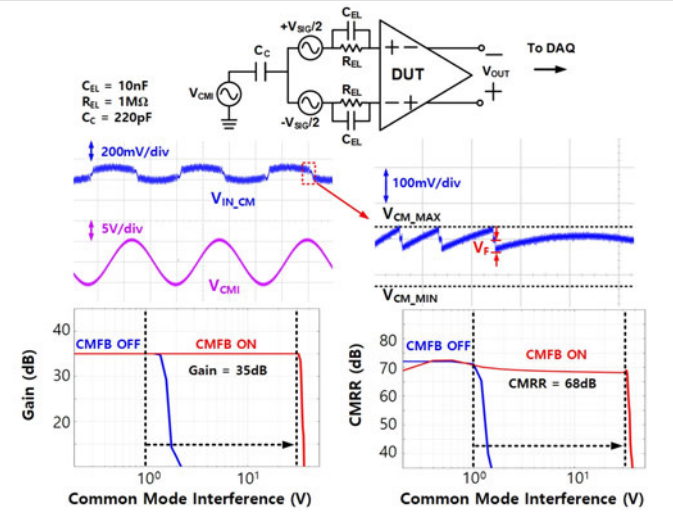


Figure 22.4.3: Measurement setup of the device (top), a transient waveform of the CMCP when 10Vpp@60Hz interference is applied (middle) and measured gain and CMRR under large common-mode interference (bottom).

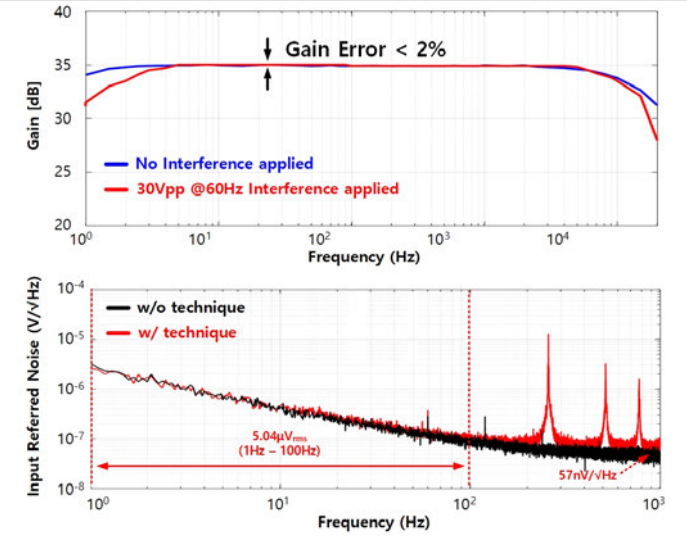


Figure 22.4.4: Measurement transfer function (top) and noise spectral density (bottom).

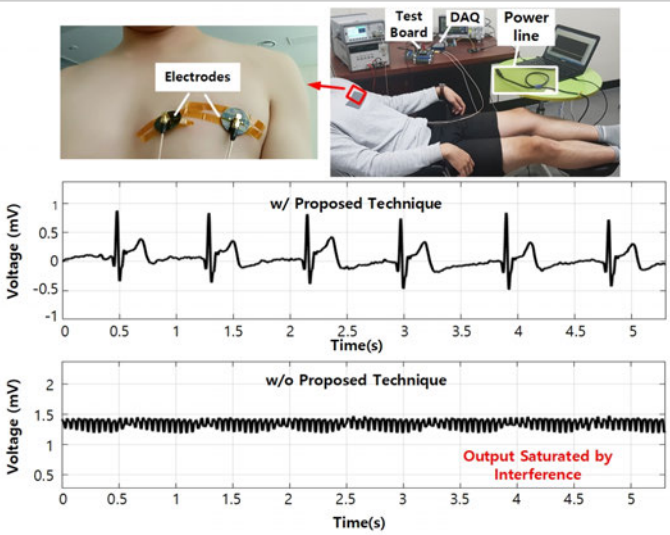


Figure 22.4.5: Setup for ECG measurement (top) and measured ECG waveform with/without the proposed technique (bottom).

	This work	JSSC17 [1]	VLSI18 [2]	CICC18 [3]	JSSC15 [4]	JSSC17 [5]
Technology	180nm	40nm	180nm	40nm	65nm	180nm
Application	ECG	Neural Recording	ECG	ECG	ECG	ECG
Supply (V)	1.2	1.2	5	0.6	0.6	0.8
Power Consumption (μW)	27.8	2.8	155	3.8[3]	0.064	0.289
Mid-band Gain (dB)	35	25.7	20	30	51-96	34
Bandwidth (kHz)	15	5	20	0.4	0.25	0.4
Tolerance to CM Interference (V)	**30**	0.65	2.8	0.4	N/A	N/A
CMRR (dB)	68[1]	85[1]	>70[1]	87[2]	55[2]	66[2]
Input referred noise (μV_{rms})	5.05 (1~100Hz)	1.8	3.7	1.8	6.52	8.26
Ground Electrode	X	O	O	O	X	N/A

1) CMRR with biopotential measuring environment
2) CMRR with IC only
3) ECG + BioZ

Figure 22.4.6: Comparison table.

22

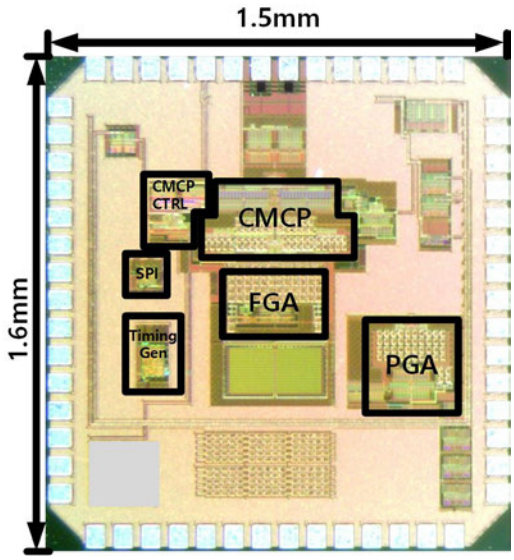


Figure 22.4.7: Die photograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

22.5 A Bio-Impedance Readout IC with Digital-Assisted Baseline Cancellation for 2-Electrode Measurement

Hyunsoo Ha[1], Wim Sijbers[2], Roland van Wegberg[1], Jiawei Xu[1], Mario Konijnenburg[1], Peter Vis[1], Arjan Breeschoten[1], Shuang Song[2], Chris Van Hoof[2,3], Nick Van Helleputte[2]

[1]imec - Netherlands, Eindhoven, The Netherlands
[2]imec, Heverlee, Belgium
[3]KU Leuven, Heverlee, Belgium

Recent advances in wearables allow for monitoring the health status of users without disturbing their daily lives (Fig. 22.5.1). Bio-Impedance (BIOZ) is an interesting sensing modality for impedance cardiography, respiration measurement, and body composition analysis [1-4]. However, existing solutions rely on a 4-electrode measurement, where two leads are used for the current generator (CG), and the others for the readout front-end (RFE). This eliminates the effect of the electrode-tissue impedance (ETI) and measures the actual BIOZ, which is required in body composition analysis where the absolute value is important. However, this degrades user comfort and adds to system complexity and cost. This work proposes a BIOZ readout IC supporting 2-electrode BIOZ measurement of body signals, where relative changes over time yields the desired information (heartrate, respiration, etc.), with a good accuracy as a 4-electrode setup (Fig. 22.5.1). The ASIC can even measure the heartrate at the wrist via 2 electrodes, which consumes significant less power than the Photoplethysmography.

In the 2-electrode setup, the increased baseline impedance (>500Ω) due to the ETI ($2\times Z_{ELEC}$) imposes new challenges (Fig. 22.5.1). First, the RFE should achieve a wide input range with a high Signal-to-Noise Ratio (SNR) to detect small variations of BIOZ (0.1 to 10Ω) over the large baseline. A Current-Balancing Instrumentation Amplifier (CBIA) is widely used for competitive performance (power efficiency, low-noise, etc.) [2-4]. However, CBIAs exhibit input-amplitude-dependent noise behavior, resulting in unacceptable high noise and limited SNR in the presence of a large input amplitude [5], which is the case in a 2-electrode setup. Second, current noise from the CG appears as voltage noise at the input of the RFE and is further amplified by the Z_{ELEC}. Hence it is even more important to minimize the current noise of the CG. While Dynamic Element Matching (DEM) can mitigate the flicker noise component of the CG noise ($i_{N.CG}$) significantly [2], the noise ($i_{N.REF}$) contribution of the reference current (I_{REF}) remains unaddressed. Consuming more power and adopting chopper stabilization are conventional ways to mitigate thermal and 1/f noise in $i_{N.REF}$; however, they increase power and design complexity. In this paper, we propose digital-assisted baseline impedance cancellation to mitigate the RFE noise dependency on the (large) input amplitude (impedance). Additionally, an I_{REF} correlated noise-cancelling architecture is introduced to reduce the CG's noise contribution without utilizing low-noise current references.

Figure 22.5.2 shows the block diagram of the IC, which contains a CBIA including transconductance (TC), transimpedance (TI) stages, and baseline-cancellation current DAC (IDAC) and current-generating circuitry including the CG and IDAC which share I_{REF}. The RFE is designed to obtain in- (I) and quadrature-phase (Q) BIOZ simultaneously, and a pre-demodulation is adopted to reduce CBIA power [2,4]. In the CBIA, the signal-dependent current flows to the current mirrors between the TC-TI stages (T1 and T2 in Fig. 22.5.2) and hence 1/f noise is also input-signal dependent. This nonlinear effect results in noticeable 1/f noise in the presence of a large input despite chopping [5]. We propose a baseline cancellation loop comprising a digital controller and a fully differential IDAC (Fig. 22.5.3) which effectively acts as a DC servo. The large baseline current (I_{BASE}) is subtracted from the signal current (I_{TC}) directly in the TC such that the TC-TI current mirrors only see a residual current (I_{TC}-I_{BASE}) effectively reducing the nonlinear effect and the associated 1/f noise in the presence of a large input (Fig. 22.5.3). The input and the cancellation current are modulated by an identical clock (f_{chop}) for synchronization. The digital controller monitors the output of the I-channel and determines the I_{BASE} (0 to 11.75μA in 250nA steps) that satisfies the required threshold $|I_{TC}-I_{BASE}|$<150nA to ensure negligible noise modulation (Figs. 22.5.2 and 22.5.3). Since the baseline is derived in the digital domain, the complete full-scale impedance can be reconstructed, which cannot be done with an analog DC servo loop [3]. This is useful for checking the contact status in the 2- and 4-electrode scenarios with a large body impedance (>100Ω).

Figure 22.5.3 compares the measured noise performance with and without baseline cancellation. All noise numbers are measured by applying a current to a known resistor and taking the digital output to include all noise sources (CG, RFE, IDAC and ADC). A noise density at 1Hz and an integrated noise (0.1 to 4Hz) have been reduced from 310 to 98nV/√Hz and 619 to 198nV_{RMS}, respectively (with the largest input for the conventional CBIA (25mV_{PK})). Furthermore, since the large baseline is subtracted in the first TC stage, the CBIA can be configured to the largest gain (160V/V) relaxing the ADC noise requirement over the whole extended input range (<125mV_{PK}). In the conventional case [2], the IA gain must be reduced to accommodate a large input resulting in noise fluctuation when gain changes (Fig. 22.5.3).

The noise contribution of the IDAC also matters. The total BIOZ noise is the sum of all uncorrelated noise powers of the building blocks. In this work, DEM is also adopted in the IDAC, hence most of the uncorrelated noises (1/f noise) of the CG and the IDAC are mitigated. It is noteworthy that since both the CG and the IDAC use the same reference current, the remaining noise (in I_{BASE}, I_{CG} and I_{TC}) are largely correlated (amplified noise of $i_{N.REF}$ as in Fig. 22.5.4). Therefore, since the circuit cancels most of the I_{TC} by subtracting I_{BASE}, the correlated reference noise is also cancelled in TI ($i_{N.TI}$) due to the identical amplification ratio of signal and noise: I_{REF} to I_{TC} and I_{BASE} are equal to $i_{N.REF}$ to $i_{N.TC}$ and $i_{N.BASE}$ (Fig. 22.5.4). The maximum $|I_{TC}-I_{BASE}|$ of 150nA is small enough to make the residual $i_{N.REF}$ negligible for BGR noise <10μV/√Hz, which is achievable in low power (<5μW) without chopping. The correlated noise cancellation reduces the noise dramatically in comparison to the noise with uncorrelated input signal (Fig. 22.5.4) (e.g., V_{IN}=125mV_{PK}, noise density at 1Hz and integrated noise (0.1 to 4Hz) is reduced from 1789 to 436nV/√Hz and 3441 to 869nV_{RMS}, respectively).

The IC is fabricated in 55nm CMOS and occupies 0.738mm^2 (Fig. 22.5.7). The IC consumes 18.9 to 34.9μW from the RFE (2-electrode case of BIOZ-I only, V_{IN}=0.1 to 125mV_{PK}) and 31.4 to 154.7μW from the CG (I_{CG}=5 to 100μA_{PK}) from a single 1.2V power supply. For the 4-electrode case where the I- and Q-channel are operating, the RFE consumes 39.1μW (V_{IN}=10mV_{PK}). The noise performance has been characterized over whole input range (V_{IN}=0.1 to 125mV_{PK} with Z_{LOAD}=20Ω to 24kΩ) in Fig. 22.5.5. The work achieves a wide input range of 24kΩ and a fine resolution of 0.8$m\Omega_{RMS}$ and 13.9$m\Omega_{RMS}$ for Z_{LOAD} of 20Ω and 2kΩ, respectively, resulting in a dynamic range of 141dB and a maximum SNR of 94dB. There are remaining noise dependencies to the input signal due to 1/f noise of the cascode transistors in CG and IDAC that are not mitigated by DEM and a thermal noise of CG and IDAC, which become non-negligible with large impedance. Nevertheless, the body signals are successfully measured with 2-electrode (gel) on chest and 2-electrode (dry) on wrist for respiration and heartbeat (an additional electrode was used for body bias), respectively. It demonstrates that the IC has the required sensitivity to detect very small variations of impedance (<100mΩ) over a large baseline, which has not previously been demonstrated at IC level (Fig. 22.5.5). Figure 22.5.6 benchmarks the proposed IC with the state of the art. The proposed work achieves the best power efficiency and the widest dynamic range among reported solutions by addressing 2 critical challenges with a single solution.

References:
[1] P. Schönle, et al., "A Multi-Sensor and Parallel Processing SoC for Miniaturized Medical Instrumentation," *IEEE JSSC*, vol. 53, no. 7, pp. 2076-2087, July 2018
[2] H. Ha, et al., "A Bio-Impedance Readout IC with Frequency Sweeping from 1k-to-1MHz for Electrical Impedance Tomography," *Symp. VLSI Circuits,* pp. C174-C175, June 2017.
[3] M. Kim, et al., "A 1.4mΩ-Sensitivity 94dB-Dynamic-Range Electrical Impedance Tomography SoC and 48-Channel Hub SoC for 3D Lung Ventilation Monitoring System," *ISSCC Dig. Tech. Papers*, pp. 354-355, Feb. 2017.
[4] H. Ko, et al., "Ultra Low Power Bioimpedance IC with Intermediate Frequency Shifting Chopper," *IEEE TCAS-II*, vol. 63, no. 3, pp. 259-263, Mar. 2016.
[5] H. Ha, et al., "Measurement and Analysis of Input-Signal Dependent Flicker Noise Modulation in Chopper Stabilized Instrumentation Amplifier," *IEEE Solid-State Circuits Letters*, vol. 1, no. 4, pp. 90-93, Apr. 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

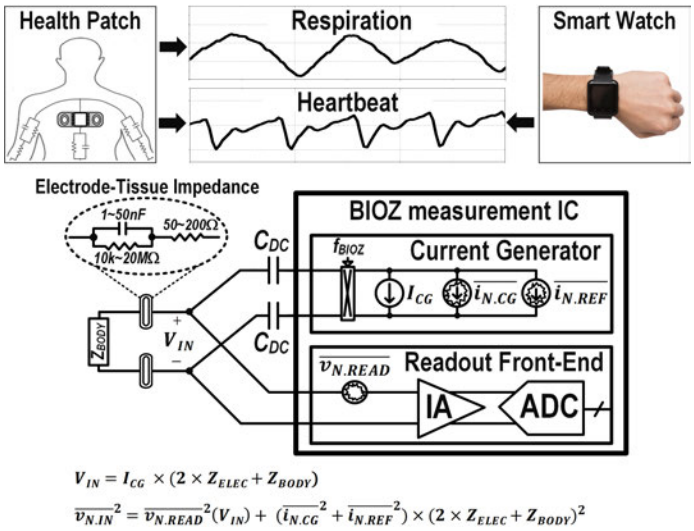


Figure 22.5.1: BIOZ measurement method via 2-electrodes and resulting input signal and input-referred noise from RFE, CG and reference.

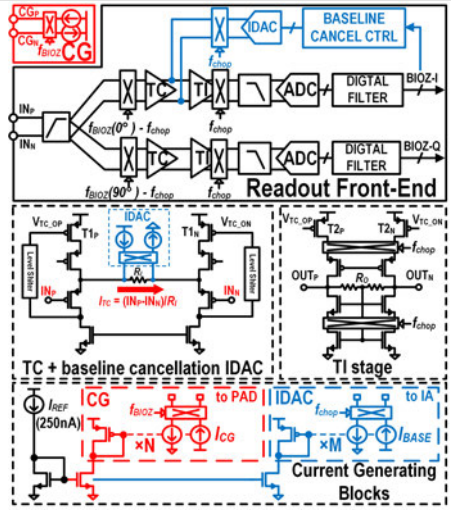


Figure 22.5.2: Top block diagram and circuit diagram of CBIA compromises TC and TI, as well as current-generating blocks (CG and IDAC) sharing a reference current.

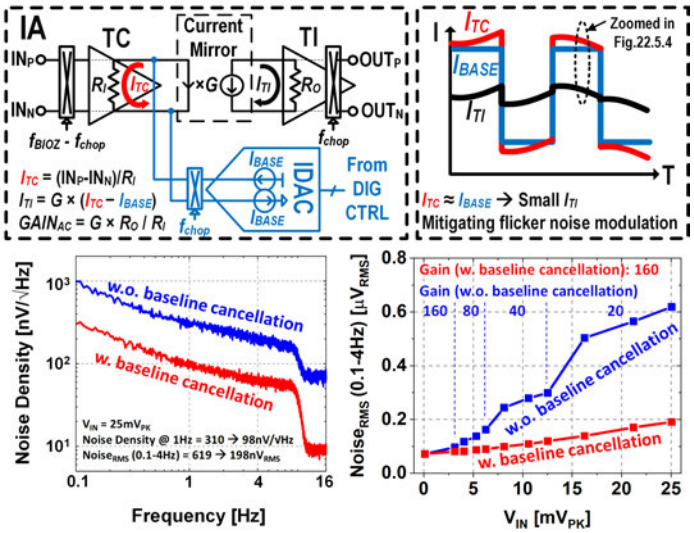


Figure 22.5.3: Operating principle of CBIA with baseline cancellation and resulting noise improvement.

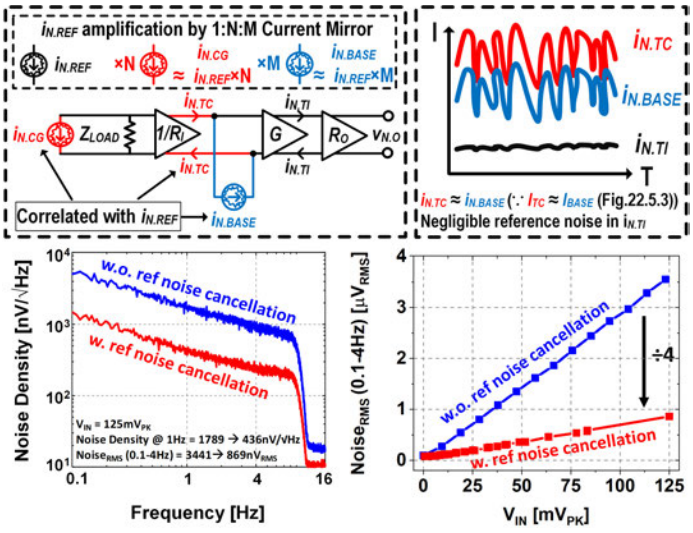
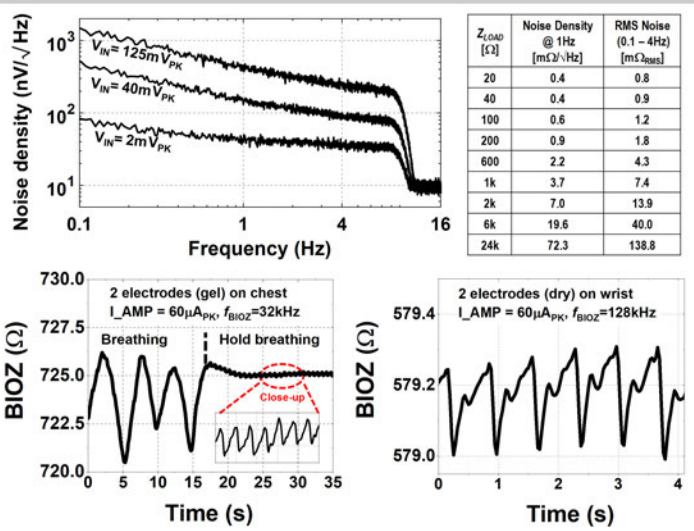


Figure 22.5.4: Correlated reference noise cancellation and resulting noise improvement.



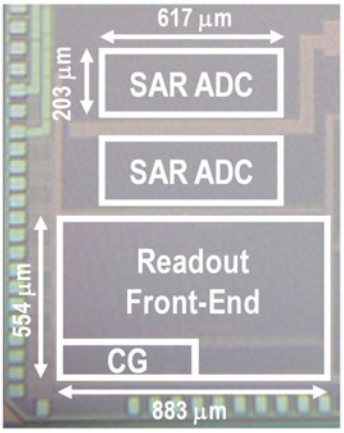
Z_{LOAD} [kΩ]	Noise Density @ 1Hz [mΩ/√Hz]	RMS Noise (0.1 – 4Hz) [mΩ$_{RMS}$]
20	0.4	0.8
40	0.4	0.9
100	0.6	1.2
200	0.9	1.8
600	2.2	4.3
1k	3.7	7.4
2k	7.0	13.9
6k	19.6	40.0
24k	72.3	138.8

Figure 22.5.5: Noise and body measurement results (respiration and heartbeat) from chest and wrist of human through 2 gel and dry electrodes, respectively.

		This work	JSSC 18 [1]	VLSI 17 [2]	ISSCC 17 [3]	TCAS-II 16 [4]	MAX30001	ADS1292R
Technology (nm)		55nm	130nm	180nm	65nm	130nm	N/A	N/A
Area (mm²)		0.74	0.96	0.88	3.2	13.8	N/A	N/A
Supply (V)		1.2	1.5/3.5	1.2/1.8	1.2	1.2	1.8	3
Stimulation signal		Square (I)	Pseudo-sine (I)	Square (I)	Pseudo-sine (I)	Square (I)	Square (I)	Square (V)
Input range (Ω)		24k	10k	1k	500[1]	623.5	5.625k[1]	10k
Current mag. (μA$_{pk}$)		5 – 100	5 – 50	5 – 100	50 – 500	0 – 169	8 – 96	0 – 100
Current freq. (kHz)		1 – 1024	4 – 100	1 – 1024	10 – 256	0.25 – 1240	0.125 – 131	32 – 64
Noise		0.4 mΩ/√Hz (R=40Ω, 100μA$_{PK}$) 7.0 mΩ/√Hz (R=2kΩ, 60μA$_{PK}$)	0.7 mΩ/√Hz (R<50Ω) 9.5 mΩ/√Hz (R=2kΩ)	1.1 mΩ/√Hz (R=20Ω) 26.0 mΩ/√Hz (R=1kΩ)	0.2 mΩ/√Hz[1,2] (R=N/A)	0.9 mΩ/√Hz[2] (R=N/A)	40mΩ$_{pp}$ (R=680Ω)	40mΩ$_{pp}$ (R=2kΩ)
Dynamic Range[3] (dB)		141 (BW=4Hz)	128 (BW=4Hz)	104 (BW=4Hz)	113 (BW=4Hz)	102 (BW=4Hz)	110 (BW=4Hz)	115 (BW=2Hz)
Power (μW)	Readout	18.9 – 34.9[4] / 39.1[5]	89[6]	18.7	6960	52	320	335
	CG	31.4 – 154.7	56 – 95	54 – 229		0 – 304[7]		

[1] Calculated from reported value in the paper
[2] not including CG noise
[3] $20\times\log(R_{MAX}/(2\sqrt{2}\times R_{MIN_RMS}))$, $R_{MIN_RMS}=R_{N.DENSITY}\times\sqrt{BW}$ or $R_{PP}/6.6$
[4] I-channel only (2 electrodes scenario), V_{IN} = 0.1 - 125mV$_{PK}$
[5] I- and Q-channel (4 electrodes scenario), V_{IN} = 10mV$_{PK}$
[6] without ADC
[7] Calculated as reported in the paper: 1.8×Injecting current

Figure 22.5.6: Comparison table with state-of-the-art BIOZ IC from academy and industry.



Building Block	Area [mm^2]
CG	0.054
Readout Front-End	0.435
SAR ADCs	0.250
Total	0.738

Figure 22.5.7: Microphotograph and area breakdown.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

22.6 A 13-Channel 1.53-mW 11.28-mm² Electrical Impedance Tomography SoC Based on Frequency Division Multiplexing with 10× Throughput Reduction

Boxiao Liu[1,2], Guoxing Wang[1], Yongfu Li[1], Hui Li[1], Yue Gao[1], Lei Zeng[2], Yixin Ma[1], Yong Lian[1], Chun Huat Heng[2]

[1]Shanghai Jiao Tong University, Shanghai, China
[2]National University of Singapore, Singapore, Singapore

Electrical-impedance tomography (EIT) has been reported as the only viable wearable real-time method for lung imaging [1]. Previous EIT chips generally employ Time-Division Multiplexing (TDM) or as active electrode to facilitate multi-channel read-out [2-4], and support a large number of electrodes by sharing a maximum of 6 parallel readout channels per chip. In this paper, we present an EIT-SoC with the following features: 1) early demodulation to relax the bandwidth requirement of the analog front-end, and minimize the impact of motion artifact and DC offset generated at the skin-electrode interface; 2) Frequency-Division Multiplexing (FDM) to combine 13 pairs of I/Q signals into 2 data streams for quantization by 2 $\Delta\Sigma$ Modulators (DSMs); 3) throughput reduction at a compression ratio of 9.75:1; 4) an inverted 'V-shape' gain configuration to minimize voltage variations across all channels; and 5) batch spectrum processing for data from all channels with no digital filtering.

Figure 22.6.1 shows the 16-electrode EIT imaging system. When current is injected into one electrode pair, voltage is measured from the rest electrodes simultaneously through 13 dedicated voltage-sensing channels. This action is repeated for all 16 electrodes through successive rotations to create one frame of image. Figure 22.6.2 shows the multi-channel system architecture. It includes: 1) a fully Differential Sinusoidal Current Stimulator (DSCS); 2) an Electrode Switching Control Circuit (ESCC) unit to manage the rotation of electrode pairs for current injection and voltage sensing; 3) 13 channels of read-out front-ends with each block containing an I/Q demodulator, two instrumentation amplifiers (IAs), and two programmable gain amplifiers (PGAs) to compensate amplitude variation among channels; and 4) a Data Compression and Sampling (DCS) unit that implements FDM with the help of 2 adders to assemble 13 I/Q signals into two data streams for quantizing by two 16b DSMs.

The I/Q demodulator converts injected frequency to DC. As its outputs pass through a chain of IA and PGA for signal conditioning, the low-pass characteristic of the IA and PGA suppresses the unwanted harmonics caused by demodulation. Thus, no low-pass filter (LPF) is required for the demodulator. The early demodulation has three advantages: 1) it decouples the frequency of injection current from the bandwidth requirement of IA and PGA, leading to significant power savings, especially for supporting high injection frequency; 2) it eliminates the impact of electrode DC offset and motion artifact without the need for a high-pass filter; and 3) as the up-conversion is performed after the early I/Q demodulation, the I/Q part simply corresponds to the amplitude of a particular frequency bin after FFT, and hence, no digital filter is needed for processing the DSM outputs.

It is common in EIT imaging systems that the sensed voltage signals display a 'V-shape' amplitude variation pattern, i.e., the electrode pairs closer to the injection electrode pair have higher amplitude, and those further away have lower amplitude, as illustrated in Fig. 22.6.3. Since an identical configuration of voltage-sensing channels is maintained during the rotation, we configure 13 PGAs with a fixed inverted 'V-shape' gain setting to compensate the voltage variation. This approach relaxes the dynamic range requirement, eliminates the need of adaptive gain control, and simplifies data processing. With the inverted 'V-shape' gain setting, the amplitudes of each channel are equalized as shown in the simulated spectrum in Fig. 22.6.3. Choppers are used to implement FDM by up-converting 26 I/Q baseband signals into a 5.1-to-8.7kHz band with 300Hz channel spacing. The up-converted I/Q signals are assembled into two data streams by two adders, which are digitized by two 16b DSMs. Compared to a TDM scheme, FDM brings two advantages: 1) it implements multiplexing without the need of large buffers for driving the analog MUX, which saves power, especially for a large number of channels; and 2) it enables data compression on-chip. For a conventional 13-channel EIT system with two 12b ADCs for both I/Q channels, it would produce a total throughput of 312 bits (13×2×12 bits) per reading. With the FDM scheme, the total throughput is reduced to only 32b (2×16b) per reading, realizing a compression ratio of 9.75:1.

Figure 22.6.4 shows the DSCS, which generates a fully differential sinusoidal output through an 8b capacitive DAC with the digital sinusoidal code (Dig_IN) as its input. It produces four frequencies, i.e., 15.625, 31.25, 62.5, and 125kHz. Two cascaded 1st-order active LPFs are used to suppress the harmonics. An operational-amplifier based V-to-I converter (VIC) converts the sinusoidal voltage from filter output to a sinusoidal current. Native transistors M_7-M_8 are employed to attain large AC swing while maintaining decent linearity. Ideally, only M_3-M_6 are needed to form the current mirror for the VIC. The addition of M_1-M_2 can offset the required minimum current and boost the amount of AC current that can be injected to the body. This increases $|I_{AC}|/|I_{DC}|$, thus improves output power efficiency by a factor of 4. As each DSM takes in 13-channel signals, which makes it susceptible to saturation, the output current is designed in the range of 20 to 200μA with 16 amplitude steps. The amplitude is controlled by 2 coarse bits of R_V and 2 fine bits from Dig_IN. The measured spectrum for the 4 different frequencies with load impedance of 1kΩ and current amplitude of 100μA is shown in Fig. 22.6.4 with the unwanted harmonics all below -58dBc.

Figure 22.6.5 shows the read-out channel implementation. DC coupled IA is selected for I/Q demodulated input. The IA exhibits 2kHz bandwidth and 26dB gain determined by the resistor ratio. The PGA gain is configurable from 0 to 18dB, which is sufficient for compensating the 'V-shape' pattern of V_1 -V_{13}. The measured spectrum from the DCS in Fig. 22.6.5 illustrates the combined multi-channel signals in the 5.1 to 8.7kHz band. For the DSM, it has a measured peak SNDR of 90.4dB and SFDR of 98dB. Foreground resistor calibration is performed to eliminate the gain error across all channels. By applying a Blackman-window corrected FFT algorithm, we can retrieve the amplitudes of all channels with less than 60dB spectrum estimation error at one go.

Figure 22.6.6 shows the measurement result of a chest model, a lung image, and the comparison table. The chest model is made of a 20cm-diameter water tank with two test objects modeling the shape of lung. EIDORS software is used to reconstruct the EIT image with 64×64 pixels. Heat map binning based pixel count is used to determine the test object region, which is marked with red circles. The measured area ratio error is 5.3% for the chest model image. The lung image is measured above the nipple level on a male subject with 200μA 125kHz injection current. The reconstructed image is projected to a body shape. The chip generates 5fps EIT images regardless of injection frequencies with a sensitivity of 9.36mΩ.

The SoC is fabricated in 130nm CMOS and occupies an area of 3.18×3.54mm². Figure 22.6.7 shows the chip micrograph and performance summary. The SoC supports simultaneous voltage sensing of 13 channels. Compared with previous works in the comparison table, the proposed SoC achieves the lowest power per channel of 118μW and the smallest area per channel of 0.87mm², showing 10× and 58% improvements, respectively. The resulting data throughput is also reduced by 10×, thanks to the inherent data compression.

Acknowledgements:
The work has been supported by the National Key Research and Development Program of China grant '2016YFE0116900' and Singapore A*STAR BMRC grant '1610500027'.

References:
[1] S. Hong, et al., "A 10.4mW Electrical Impedance Tomography SoC for Portable Real-Time Lung Ventilation Monitoring System," *IEEE JSSC*, vol. 50, no. 11, pp. 2501-2512, Nov. 2015.
[2] Y. Lee, et al., "A 4.84mW 30 fps Dual Frequency Division Multiplexing Electrical Impedance Tomography SoC for Lung Ventilation Monitoring System," *IEEE Symp. VLSI Circuits*, pp. C204-C205, June 2015.
[3] M. Kim, et al., "A 1.4mΩ-Sensitivity 94dB-Dynamic-Range Electrical Impedance Tomography SoC and 48-Channel Hub SoC for 3D Lung Ventilation Monitoring System," *ISSCC Dig. Tech. Papers*, pp. 354-355, Feb. 2017.
[4] S. Hong, et al., "A 4.9mΩ-Sensitivity Mobile Electrical Impedance Tomography IC for Early Breast-Cancer Detection System," *ISSCC Dig. Tech. Papers*, pp. 316-317, Feb. 2014.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

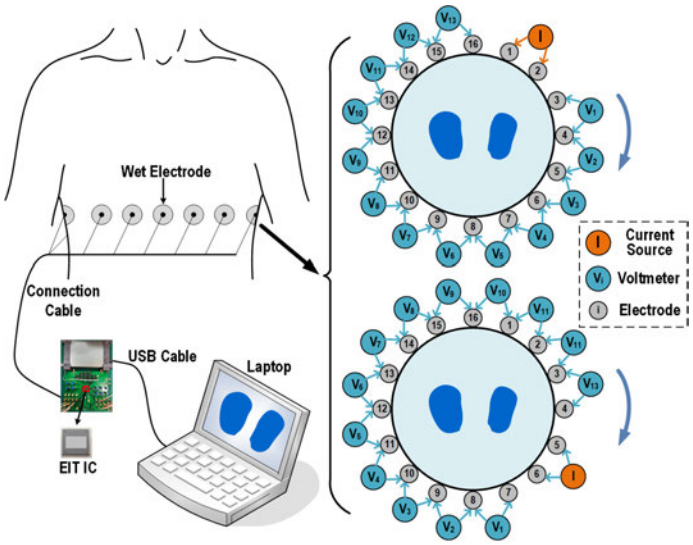


Figure 22.6.1: Proposed EIT imaging system.

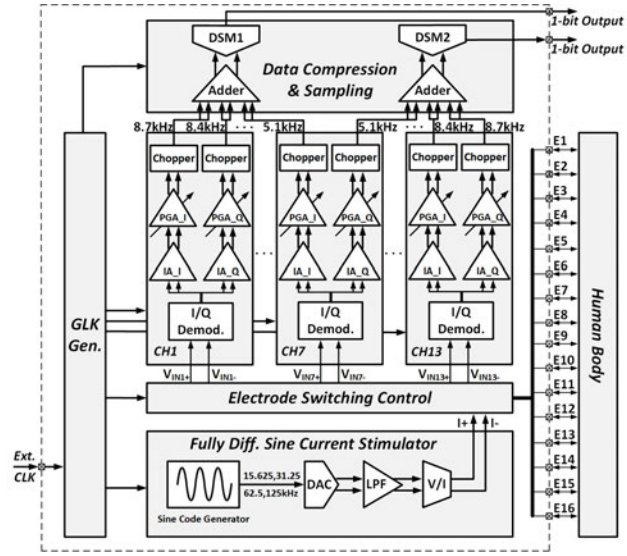


Figure 22.6.2: Multi-channel system architecture.

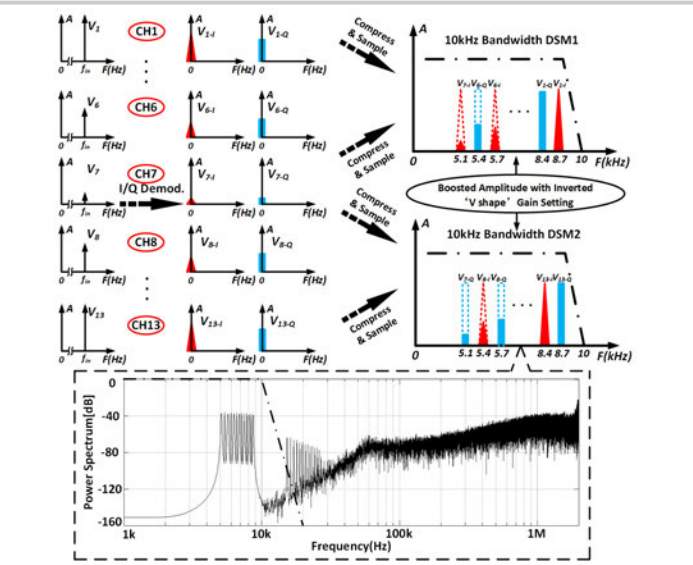


Figure 22.6.3: Inverted 'V-shape' gain configuration.

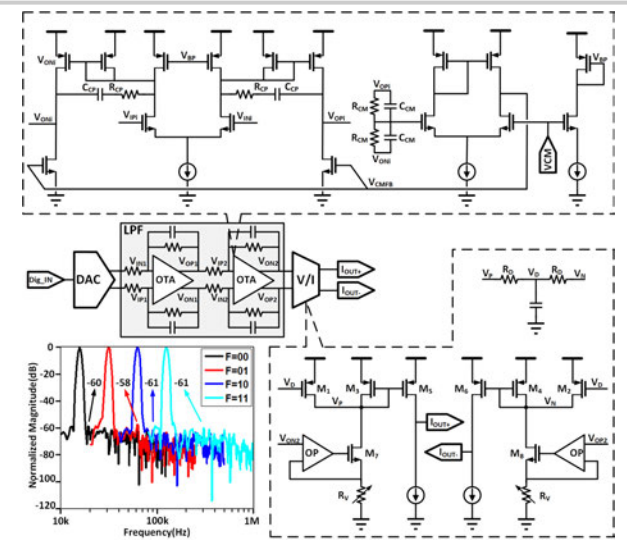


Figure 22.6.4: Differential sinusoidal current stimulator with measured spectrum.

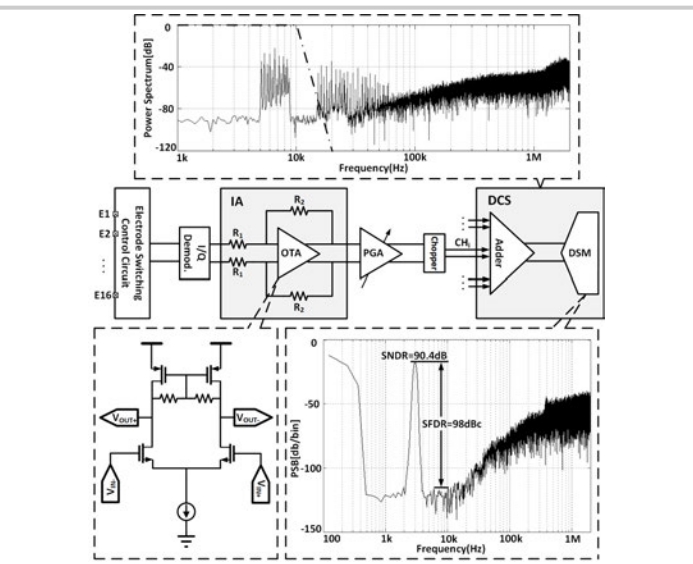
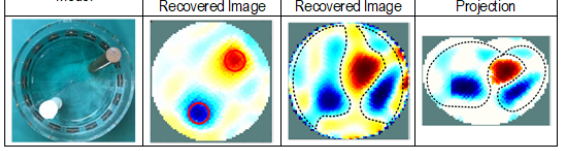


Figure 22.6.5: Read-out channel implementation.

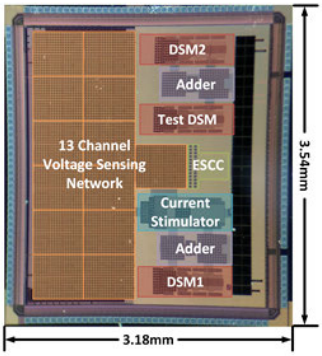
Nylon Bar (2cm) & Copper Bar (2cm)		Lung Measurement Result	
Model	EIDORS Recovered Image	EIDORS Recovered Image	Body Shape Projection



	JSSC'15[1]	VLSI'15[2]	ISSCC'17[3]	This Work
Technology	180nm	180nm	65nm	130nm
Channels/Chip	6	4(Dual Freq.)	1(AE Syst.)	13
Power/Channel	1.34mW	1.21mW	-	0.118mW
Area/Channel	4.17mm^2	2.38 mm^2	3.2mm^2	0.87 mm^2
Injection Amplitude	0.1-1mA (4 step)	0.125-0.5mA (4 step)	0.1-1mA (4 step)	0.02-0.2mA (16 step)
Injection Frequency	10-200kHz (4 step)	10-360kHz	10-256kHz (20 step)	15.625-125kHz (4 step)
Frame Rate	5-20fps (Vary with Freq.)	~30fps (Vary with Freq.)	~10fps (Vary with Freq.)	5fps (Fixed)
On-chip Data Compression	N	N	N	Yes 9.75:1
'V-shape' Variation	N	N	N	Utilized by Dedicated CH

Figure 22.6.6: Measurement result and comparison table.

22



Technology	130nm	
Die Size	3.18mm×3.54mm	
Supply	1V	
Power Consumption	1.53mW	
Current Stimulator	Frequency	15.625-125kHz
	Amplitude	20-200μA
	Harmonics Distortion	<-58dBc
Read Out Front End	# of Channels	13
	Gain	26-44dB
	Input Referred Noise	5μV (DC~2kHz, Simulated)
Read-out ADC	DSM (90.4dB)	
Operating Frequency	24MHz	

Figure 22.6.7: Chip micrograph and performance summary.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

22.7 A Programmable Wireless EEG Monitoring SoC with Open/Closed-Loop Optogenetic and Electrical Stimulation for Epilepsy Control

Shuenn-Yuh Lee, Chieh Tsou, Peng-Wei Huang, Po-Hao Cheng, Chi-Chung Liao, Zhan-Xien Liao, Hao-Yun Lee, Chou-Ching Lin, Chia-Hsiang Hsieh

National Cheng Kung University, Tainan, Taiwan

The number of studies on closed-loop detection and electrical stimulation systems [1-2] for efficient control of neurological disorders is increasing, because recent clinical studies have shown their efficiency and usefulness in symptom suppression. Electrical stimulation can produce enough stimulation to affect a large range of nerves. However, all nerves near the stimulus are excited and hurt, and over time, currents start to exceed acceptable limits. Therefore, optogenetic stimulation [3-4] has become compelling in recent years due to several advantages: (1) no artificial noise on the EEG; (2) ability to stimulate specific nerves; and (3) no injurious effects on nerves. In this study, a wireless programmable stimulating system-on-chip (WPSSoC) is reported that provides wireless open/closed-loop optogenetic and electrical stimulation to improve treatment for epilepsy suppression. The system is demonstrated on programmable stimulation parameters wirelessly controlled by a software Graphical User Interface (GUI) on a computer. Moreover, an animal experiment conducted on optogenetic tissue was successful, thereby demonstrating that the nerve injury on optogenetic stimulation is lower than that of electrical stimulation.

Figure 22.7.1 shows the block diagram of the whole system, which is composed of a headstage, a transgenic mouse and a remote control GUI with wireless adapter. The headstage includes a battery, an off-chip laser driver, an antenna, and a WPSSoC. From the recording electrode, the EEG is sensed by high-pass delta-sigma modulator (HPDSM) with decimator and transmitted by antenna with UART packaging and RF on-off keying (OOK) modulation under a transmission distance of 13cm. The software can monitor the real-time signal and wirelessly program the stimulation parameters, including optogenetic and electrical stimulating waveform (PWM or mono/bi-phasic) and open/closed-loop mode. When epilepsy is detected, the system controller enables the current-steering DAC and off-chip laser driver to stimulate the nerve and control the symptom.

As shown in Fig. 22.7.2, in which the comparison and electrode positions of optogenetic and electrical stimulations are presented, the position of the implanted electrodes and the *in vivo* experiment result of apoptosis and injection position are illustrated. Channelrhodopsin-2 (ChR2) is a typical optogenetic actuator. When ChR2 is expressed in neurons, blue light induces inward currents, which depolarize the membrane for the opening of voltage-gated sodium and calcium channels. Then, ChR2 permits Na+ and Ca2+ to enter the interneuron, which releases gamma amino butyric acid and suppresses the action potentials of the principal neuron. Thus, the activation of interneurons with depolarizing actuators is valid for seizure control. Thy1-ChR2-YFP transgenic mice at an approximate age of postnatal day 90 were used *in vivo* experiments. The adeno-associated virus (AAV) carrying ChR2 was injected in the CA1 of hippocampus of mice (right red spot). AAV vectors with transgene induced the therapeutic gene delivery for neurological disorders. ChR2-transgenic mice were treated with pilocarpine and spilt into mock, optogenetic, and electrical stimulation groups. In comparison with mock groups, the neuron apoptosis in immunohistochemical assay was attenuated in ChR2-transgenic mice with blue light stimulation, and was more significantly detected in ChR2-transgenic mice with electric stimulation. Thus, ChR2 with light stimulation exerted potent neuroprotection.

Figure 22.7.3 demonstrates block diagrams, including the digital signal processor, the measured impedance of the electrical stimulation path with the synthesized equivalent model, the bio-signal sensing system, and the stimulation circuit. The impedance is measured using a 20mV sine wave from 20Hz to 5kHz in the animal test to ensure the specification of the current-steering DAC, and this was hypothesized as an equivalent model with four causes. The EEG data was sensed by the bio-signal sensing system, which was composed of a chopper-based programmable amplifier (PGA) and a HPDSM. Then, the output signal was demodulated and filtered by the decimation filter with a bandwidth of 100Hz. In epilepsy detection, the features used for linear least squares (LLS) classifier were composed of signal amplitude, normalized frequency bins, which were calculated by single-bin discrete Fourier transform circuit, averaged-32-points approximate entropy (ApEn). The detection result is considered by using the output of the classifier to operate the stimulator. Based on the results of the simulation of the algorithm with raw data, the precision is about 86%. An 88b command is used to define the stimulation waveform and mode. The whole stimulation controller is operated at 96kHz to manipulate the on-chip current DAC and off-chip laser driver by the input command. The stimulation frequency can be adjusted between 24kHz and 5.8Hz.

As shown in Fig. 22.7.4, an OOK transceiver for 2.45GHz is proposed. The transmitter is composed of a bias-stimulation circuit (BSC), a current-reused self-mixing voltage controlled oscillator (CRSMVCO), and a quadruple-transconductance power amplifier (QTPA). The BSC can pre-emphasize the output waveform controlled by delay cell and control logic to increase the data rate of VCO. The frequency doubler (FD) and the cross-couple mixer (CCM) are realized in CRSMVCO for self-mixing, and use the same current path to save power consumption. The quadruple output current can be achieved by the current-reuse structure with two common drain/source amplifiers to enhance the output power under low power consumption. The receiver includes a single-to-differential envelope detector (SDED), a level shifter, a cascaded baseband amplifier (BA), and a hysteresis comparator. The SDED directly down-converts the single-ended RF input signal to differential baseband output without the external balun. It uses a square-law device detection to produce the symmetrical differential output and the gain-boosting technique to enhance the conversion gain. The level shifter acts as high-pass filter implemented with two capacitors for DC decoupling and a pseudo-resistor to redefine output bias voltage due to the low-frequency flicker noise. The baseband amplifier is implemented in open-loop to enlarge the baseband low/high level digital signal. The receiver consumes 12.7μW with a sensitivity of -35dBm at data rate of 1Mb/s, and the transmitter consumes 810μW with output power of -14.3dBm at a data rate of 16Mb/s.

Figure 22.7.5 demonstrates the animal study conducted by installing the headstage on the implanted electrodes, and reveals the EEG signals with or without epilepsy seizure, as well as the comparison results after electrical and optogenetic stimulation. The WPSSoC region on the headstage is designed to be $2.5\times2.1cm^2$ including the pin header, bias, and antenna. The other region includes laser driver module and Bluetooth interface for convenient usage communicated with smartphone during the animal study. The headstage receives an input RF signal with the configuration of stimulus parameters and LLS coefficients from a wireless controller at approximately 13cm distance. The output data rate to the on-chip UART and algorithm-controlled circuits is 375Hz. When epilepsy is automatically detected in the closed-loop mode or in the open-loop mode controlled by the GUI, the stimulus parameters are configured in the corresponding registers. The effective range of electrical stimulation in this animal test is approximately 150 to 300μA at 100Hz monophasic/biphasic waveform, and the optogenetic stimulation is approximately 15 to 50mW at 60Hz monophasic waveform. The signal saturation caused by side effects can appear on the electrical stimulation. However, these artifacts are very small during optogenetic stimulation. Figure 22.7.6 shows the measurement summary of WPSSoC, including a bio-signal sensing system, an RF transceiver, a digital signal processor, a stimulation controller with DAC, and a decimation filter with UART interface. Figure 22.7.7 presents a die microphotograph of the WPSSoC; also the performance of WPSSoC is compared with previous state-of-the-art works, both for electrical and optical stimulation.

Acknowledgements:
The authors would like to thank the Chip Implementation Center (CIC) and the Ministry of Science and Technology (MOST), Taiwan, R.O.C. for their support of this work (under grant nos. MOST 106-2314-B-006-001, MOST 107-2218-E-006-034, MOST 107-2622-8-006-009 -TE2) and the help of CBIC Lab members, Ming-Chun Liang, and Ching-Fu Tsou.

References:
[1] H. Kassiri, et al., "All-Wireless 64-Channel $0.013mm^2$/ch Closed-Loop Neurostimulator with Rail-to-Rail DC Offset Removal," *ISSCC Dig. Tech. Papers.*, Feb. 2017, pp. 452–453.
[2] W. M. Chen, et al., "A Fully Integrated 8-Channel Closed-Loop NeuralProsthetic CMOS SoC for Real-Time Epileptic Seizure Control," *IEEE JSSC*, vol. 49, no. 1, pp. 232–247, Jan. 2014.
[3] Y. Jia, et al., "A mm-Sized Free-Floating Wirelessly Powered Implantable Optical Stimulating System-on-a-Chip," *ISSCC Dig. Tech. Papers.*, pp. 468-469. Feb. 2018.
[4] G.G. Turcotte, et al., "A Wireless Headstage for Combined Optogenetics and Multichannel Electrophysiological Recording," *IEEE TBioCAS*, vol. 11, no. 1, pp. 1-14, Feb. 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

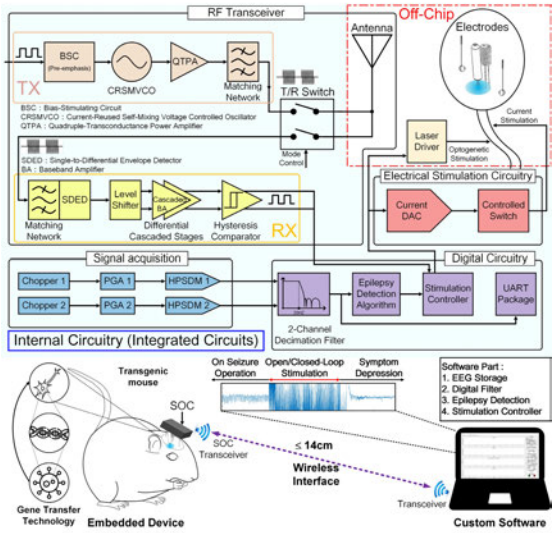


Figure 22.7.1: Block diagram of the open/closed-Loop optogenetic and electrical stimulation system with external device.

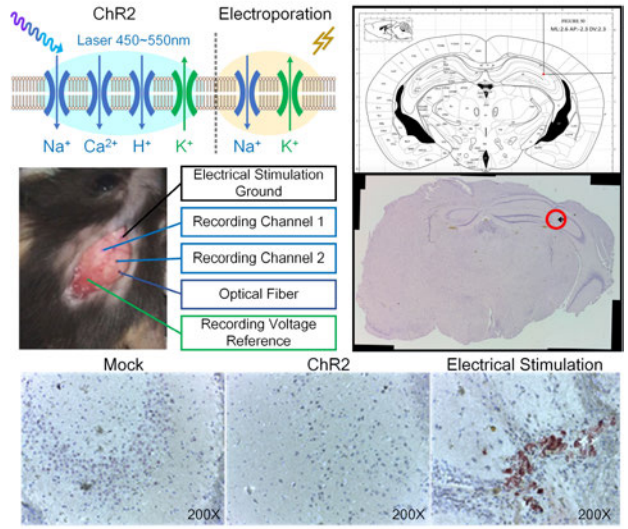


Figure 22.7.2: Opsin transgenic mice experiment with comparison between the two stimulations.

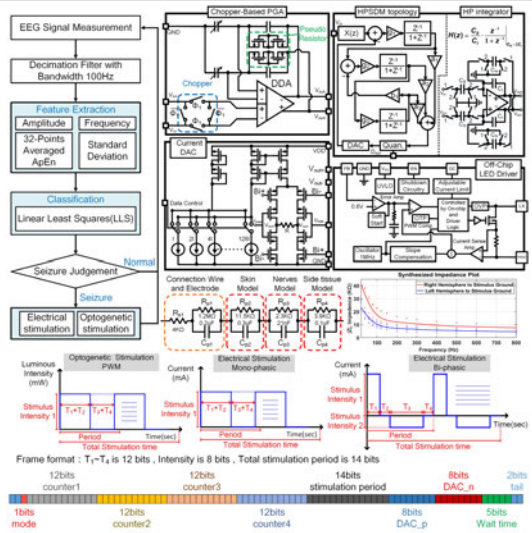


Figure 22.7.3: Bio-signal sensing system, algorithm, and programmable stimulation circuits.

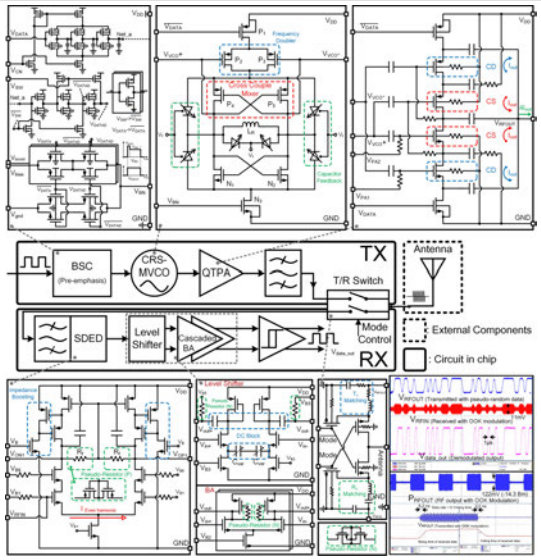
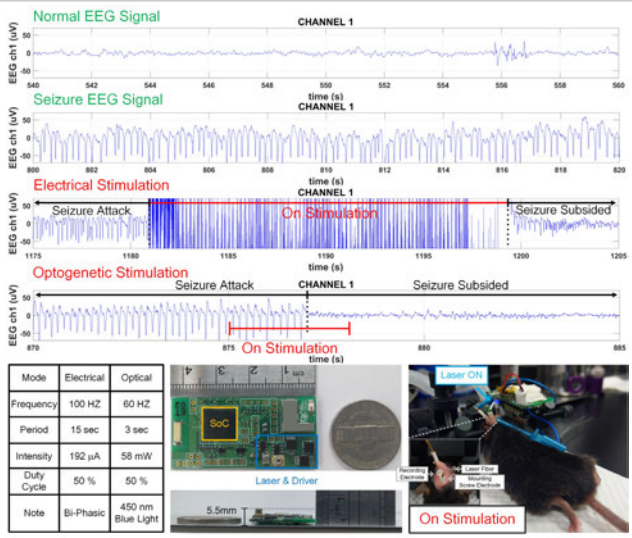


Figure 22.7.4: RF transceiver circuit structure.



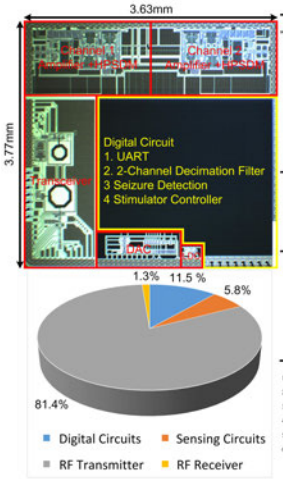
Mode	Electrical	Optical
Frequency	100 HZ	60 HZ
Period	15 sec	3 sec
Intensity	192 μA	58 mW
Duty Cycle	50 %	50 %
Note	Bi-Phasic	450 nm Blue Light

Figure 22.7.5: Seizure inhibition behavior in *vivo* experimental result with opsin transgenic mice induced by seizure EEG signal.

WPSSoC Summary			
General		***Digital Circuit Block @ 1.8V***	
Technology	TSMC 0.18μm 1P6M	Digital signal processor (DSP)	Two channel Seizure Detection → Feature Extraction (3 x SBDFT + ApEn) → LLS Classifier → Seizure Judgement Operation Frequency = 375 Hz
Chip area	3.63 x 3.77 mm^2		
Bio-signal Sensing System @ 1.2V			
Chopper-based amplifier	Chopper Frequency = 24 KHz Input refer noise = 1.3 μV_{RMS} Gain = 27 dB Power = 15.5 μW		
High-pass sigma-delta modulator	Bandwidth = 187.5 Hz Sampling Frequency = 48 KHz IMD_3 = 68 dB Power = 13.5 μW	Stimulation controller (With current-steering DAC)	Stimulation Frequency = 5.8Hz ~24KHz Stimulus Duration = 41 μs ~ 5s Operation Frequency = 96 KHz (Stimulus Current = 0 ~ 510 μA @3.3V, 6KΩ)
Total power consumption	29 μW		
RF Transceiver @ 1.2V, OOK, 2.45GHz			
Transmitter	Output Power = -14.3 dBm Data Rate = 16 Mbps Power Consumption = 0.81 mW Energy per bit = 50.6 pJ/bit	UART interface	Baud rate = 9600 / 115200 Data Output = 4 package/time Operation Frequency = 1.152 MHz
Receiver	Sensitivity (without LNA) = -35dBm @ 10^{-4} BER Data Rate = 1 Mbps Power Consumption = 12.7 μW Energy per bit = 12.7 pJ/bit	Decimation filter	Bandwidth = 187.5 Hz Word length = 16 bits IMD_3 = 68 dB
Total Power consumption	0.82 mW	Total power consumption	114 μW (without stimulation)

Figure 22.7.6: Summary of features and measured performance.

22



Publication	2014[2]	2017 [1]	2017 [4]	2018 [3]	This Work
Tech.(nm)	180	130	COTS[1]	350+COTS[2]	180+COTS[3]
Chip Area (mm^2)	13.47	5.98	-	1	13.67
Supply (V)	1.8	1.2/3.3	3.3	1.8	1.2/1.8/3.3
Power Diss. (mW)	2.8	1.07[4]	119[5]	1[5]	0.992[5]
# of Rec.Chan.	8	64	32	-	2
Signal Processing	Yes	Yes	Yes	-	Yes
Closed- loop Stim.	Yes	Yes	Yes	No	Yes
Wireless Comm.	Yes	Yes	Yes	Yes[6]	Yes
Range (cm)	10	10 / 200	-	< 3	10-14
Electrical Stim	Yes	Yes	No	No	Yes
Mode	Biphasic	Biphasic	-	-	Mono/ Biphasic
Stim. Current (mA)	0.03	0.01 - 1.35	-	-	0 - 0.51
Stim. Freq (Hz)	" 800	" 800	-	-	5.8~24K
Optogenetic Stim	No	No	Yes	Yes	Yes
# of Stim Chan.	-	-	32	16	1
Max $I_{optical}$	-	-	150 mA	10 mA	120 mA
Light Density (mW/mm^2)	-	-	70	10	0~50 (7 modes)
Device Weight	-	-	4.9 g	0.6 g	6 g
Optical Tether	-	-	Yes	No	No

[1] Whole system is implemented by commercial elements.
[2] The μLED is off-chip combined in the system.
[3] The laser and laser driver are off-chip combined in this system.
[4] Without the VCO-based UWB transmitter.
[5] Without the optogenetic stimulation driving power.
[6] Inductive link wireless interface.

Figure 22.7.7: WPSSoC die microphotograph and comparison table.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

22.8 Adaptively Clock-Boosted Auto-Ranging Responsive Neurostimulator for Emerging Neuromodulation Applications

M. Reza Pazhouhandeh[*1], Gerard O'Leary[*1], Iliya Weisspapir[2], David Groppe[2], Xuan-Thuan Nguyen[1], Karim Abdelhalim[1,3], Hamed Mazhab Jafari[1,4], Taufik A. Valiante[5], Peter Carlen[5], Naveen Verma[6], Roman Genov[1]

[1]University of Toronto, Toronto, Canada
[2]Krembil Neuroscience Center, Toronto, Canada
[3]now at Sitrus Technology, Irvine, CA
[4]now at Envirosen, Toronto, Canada
[5]Toronto Western Hospital, Toronto, Canada
[6]Princeton University, Princeton, NJ
*Equally-Credited Authors (ECAs)

An emerging neuromodulation approach to treat disorders such as intractable epilepsy and Alzheimer's disease is to use electrical stimulation triggered by pathological brain states, conventionally delivered via intracranially implanted electrodes as shown in Fig. 22.8.1 (top, left). A safer, non-invasive technology for closed-loop deep-brain neuromodulation (DBN) to treat neural disorders has been a holy grail of neuroscience for decades. Recent clinical studies of non-invasive transcranial stimulation are showing increasing promise. For instance, there is evidence that stimulation in certain stages of sleep can improve memory. This work demonstrates a platform technology for enabling such studies and protocols, where brain-state-dependent stimulation enables therapy in either conventional invasive or emerging non-invasive neuromodulation modalities.

Recently, a technique for non-invasive neuromodulation of the deep brain by temporally interfering (TI) electric fields has received wide attention [1]. The method enables the stimulation of a targeted deep-brain region that is electronically steerable using only scalp EEG electrodes. Figure 22.8.1 (top, right) depicts this principle, where two pairs of scalp electrodes deliver stimuli at two high frequencies with a small differential, f and $f+\Delta f$ (e.g., f=2kHz, Δf=10Hz), respectively. Neural activity is not directly modulated by these high-frequency signals independently; however, the two electric fields temporally interfere with a resulting modulation envelope of the low frequency Δf, which is most prominent in a focal region deep in the brain and diminishes elsewhere (Fig. 22.8.1 bottom, right). The neurons in that region are modulated by the stimulus envelope due to the nonlinear nature of their cell membranes. By changing the relative amplitude of the two currents, the influenced region is *remotely* steered within the brain. Significantly, the TI neuromodulation method lends itself well to standard low-voltage CMOS processes. At high frequencies (e.g., f=2 kHz), the scalp impedance is at least an order of magnitude lower than with conventional 5-to-200Hz stimuli (Fig. 22.8.1 bottom, left). This eliminates the need for a high-voltage CMOS process. This work demonstrates a 64-channel closed-loop neurostimulator that is validated in the emerging application of TI stimulation triggered by EEG sleep stage classification. The design consists of a rail-to-rail-input analog front end (AFE) with fast, adaptively clocked artifact recovery, and a feature-rich machine-learning neural signal classifier DSP. It is also demonstrated in a conventional clinical application: invasive intracranial-EEG seizure detection that triggers biphasic stimulation.

Closed-loop neuromodulation requires resolving small-amplitude neural signals in the presence of large stimulation artifacts. Figure 22.8.2 (top) depicts a conventional neural-recording architecture with a stand-alone high-gain low-noise amplifier (LNA), multiple feedback loops to ensure high input-impedance and reduced chopping ripple, as well as an analog-to-digital converter (ADC) [2]. Such amplifiers consume substantially more power and area than the subsequent ADC. When implemented separately from the ADC, such amplifiers are prone to output saturation by large stimulation transients. Figure 22.8.2 (middle) shows a state-of-the-art (SoA) architecture including an LNA and ADC combined in a single channel [3,4]. To ensure recovery from sharp artifacts, an event-driven artifact-detection and tracking mechanism is included in the feedback digital-to-analog converter (DAC) to avoid saturation. Such an architecture achieves low noise and power consumption, but its input transient tracking capability is limited by the speed of the ADC and of the artifact detection and tracking (ADT) circuit. This is not suitable for monitoring high-frequency TI stimulation (e.g., when f=2 kHz). Additionally, chopping at the capacitive input node limits the input impedance to tens of MΩ. This is too low for EEG recording as mismatches in the input impedance lead to common-mode noise contaminating the low-SNR differential signal. Figure 22.8.2 (bottom) depicts the presented adaptively clock-boosted AFE that integrates an amplifier within the ADC. During sharp input transients the clock is boosted (e.g., by a factor of 100×) so the AFE recovers from fast artifact transients such as those present during TI stimulation. DC-coupled input chopping yields high input impedance of hundreds of MΩ suitable for scalp EEG recording.

Figure 22.8.3 presents a block diagram of the presented AFE. The input chopper connects directly to the gate terminals of the thick-oxide input transistors yielding input impedance higher than 250MΩ at f_{CHOP}=10kHz, enabling high CMRR for both implantable and wearable applications. Instead of AC coupling, the AFE's highly oversampled loop delta-modulates the neural signal and removes differential electrode offset (DEO) rail to rail. To save power, the AFE employs a differential-difference dynamic comparator and a dynamic charge pump, that are nominally clocked at 100kHz, but no statically biased amplifiers. Both are open-loop circuits (i.e., not bandlimited integrators), allowing for a widely tunable oversampling ratio (OSR). When a sharp artifact is detected by the saturation detection block, the clock is boosted to f_{SAMPLE} > 10MHz enabling artifact recovery within microseconds. At the same time, radix-2 exponential auto-ranging of the loop feedback DAC is engaged (the DAC is also time shared with the biphasic stimulator). Once the AFE recovers from the sharp artifact transient, the sampling frequency and the DAC step size return to nominal values. To reduce chopping artifact without an additional loop [2], the chopper transitions take place just after the sampling phase, disabling ripple error propagation in the loop. The AFE output is reconstructed and filtered digitally to remove the quantization noise, the chopped preamplifier flicker noise, and in the case of TI stimulation, the residual stimulation tone.

Figure 22.8.4 depicts the SoC system architecture, including 64 neural ADCs that reuse 64 9b current-mode stimulating DACs. To detect brain states, the DSP extracts domain-specific signal biomarkers including physiological signal band energy (SE) and heartrate (HR) changes. These are computed by FIR/IIR filter banks and a QRS detector, respectively. A one-vs-all support vector machine accelerator with up to 200 support vectors and 200 input dimensions is trained off-line and classifies these features to distinguish among multiple brain states. The stimulator is controlled by an on-chip arbitrary stimulus waveform generator, which ensures charge-balanced waveform generation, such as high-frequency sine for stimulation. The SoC also includes on-chip wireless commands/data RX/TX interface, high-frequency multi-phase/frequency ring oscillators and regulation/distribution circuits for wirelessly received power.

Figure 22.8.5 illustrates the brain state classification performance of the SoC. Figure 22.8.5 (left) demonstrates the transcranial application with the use of multi-state classification for sleep-stage responsive neuromodulation. The DSP classifier distinguishes wakeful states from light sleep (stage 1 and 2), slow wave sleep (SWS; stages 3 and 4), and rapid eye movement (REM) sleep. TI allows for stimulation of deep-brain regions associated with memory formation in response to SWS detection. Figure 22.8.5 (right) demonstrates the SoC's use in intracranial applications, where conventional direct bi-phasic pulse stimulation is applied to the seizure onset zone to prevent an anticipated seizure. Figure 22.8.6 compares the design with the state of the art. Figure 22.8.7 depicts in vivo animal results and saline TI concept test results and the die micrograph.

Acknowledgements:
The authors thank CMC Microsystems for IC fabrication; Canadian Institutes of Health Research for funding; Maged El-Ansary for critical discussions; Dr. Hossein Kassiri and Dr. Nima Soltani for contributions to the wireless powering circuits design. The first two authors contributed equally (M. Reza Pazhouhandeh, AFE, and Gerard O'Leary, DSP).

References:
[1] N. Grossman, et al., "Noninvasive Deep Brain Stimulation via Temporally Interfering Electric Fields," *Cell* 169, 1029–1041.e16, June 2017.
[2] J. Xu, et al., "A 160 μW 8-Channel Active Electrode System for EEG Monitoring A 160 μW 8-Channel Active Electrode System for EEG Monitoring A 160 μW 8-Channel Active Electrode System for EEG Monitoring," *IEEE TBioCAS*, vol. 5, no. 6, pp. 555-567, Dec. 2011.
[3] B.C. Johnson, et al., "An Implantable 700μW 64-Channel Neuromodulation IC for Simultaneous Recording and Stimulation with Rapid Artifact Recovery," *IEEE Symp. VLSI Circuits*, pp. C48-C49, 2017.
[4] C. Kim, et al., "A 92dB Dynamic Range Sub-μVrms-noise 0.8μW/ch Neural-Recording ADC Array with Predictive Digital Autoranging," *ISSCC Dig. Tech. Papers*, pp. 470-471, Feb. 2018.
[5] G. O'Leary and M.R. Pazhouhandeh, et al., "A Recursive-Memory Brain-State Classifier with 32-Channel Track-and-Zoom $\Delta^2\Sigma$ ADCs and Charge-Balanced Programmable Waveform Neurostimulators," *ISSCC Dig. Tech. Papers*, pp. 296-297, Feb. 2018.
[6] N. Van Helleputte, et al., "18.3 A Multi-Parameter Signal-Acquisition SoC for Connected Personal Health Applications," *ISSCC Dig. Tech. Papers*, pp. 314-315, Feb. 2014.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

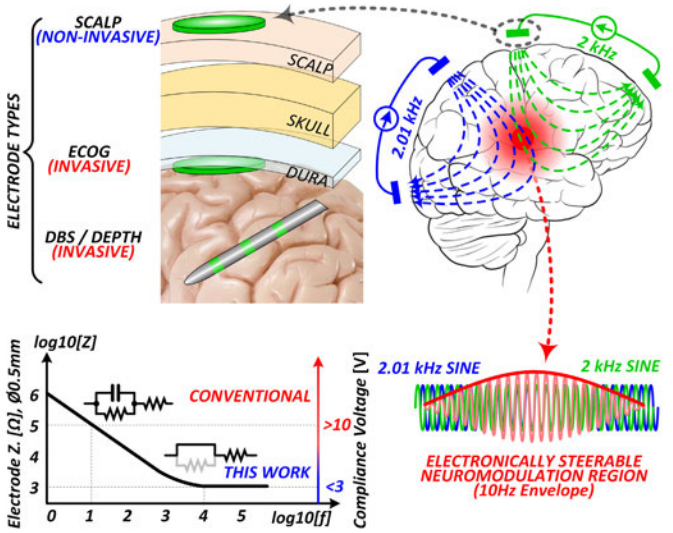


Figure 22.8.1: High-frequency non-invasive deep-brain neuromodulation concept [1] (top right), and lower required compliance voltage (bottom left).

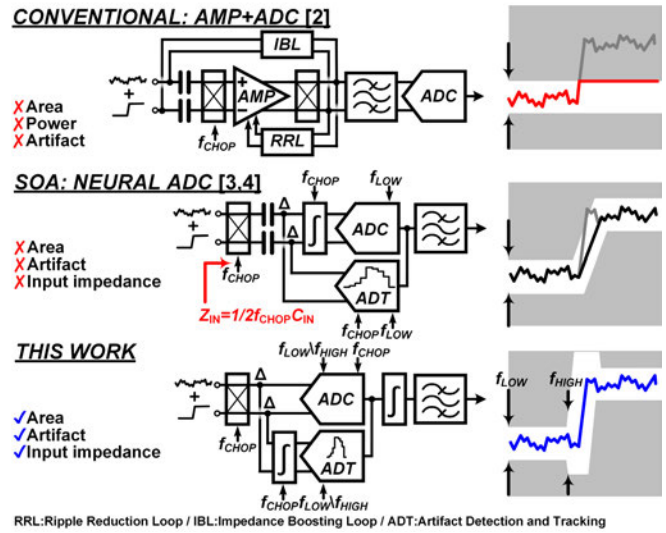


Figure 22.8.2: Conventional, state-of-the-art (SoA), and presented neural recording AFE circuit architectures.

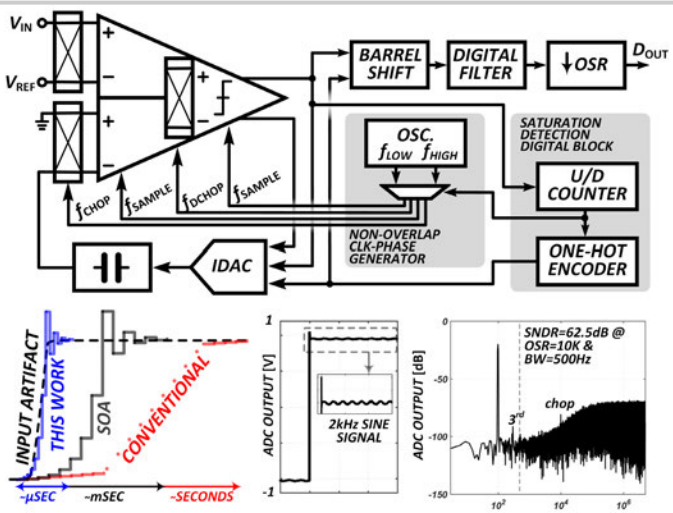


Figure 22.8.3: Clock-boosted AFE circuit (top); sharp-transient artifact recovery (bottom left); measured transient response (bottom middle); and noise spectral density (bottom right).

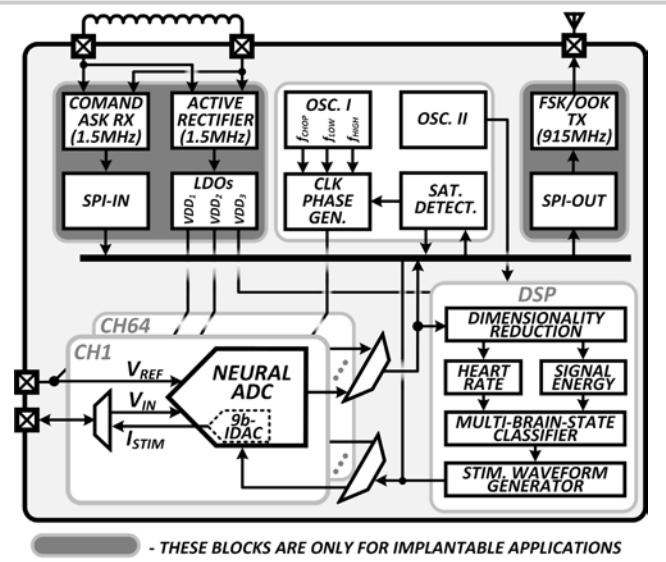


Figure 22.8.4: Simplified block diagram of the SoC.

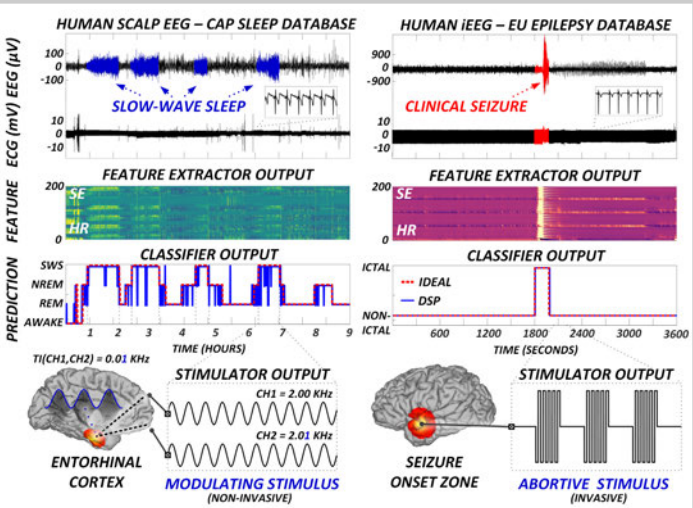


Figure 22.8.5. One-vs-all SVM uses SE and HR features to classify sleep stages in CAP Sleep Database (left), and seizure onset in EU Epilepsy Database (right). A responsive TI or DBS stimulus could then be applied (bottom).

SPECS.	VLSI17 [3]	ISSCC18 [4]	ISSCC18 [5]	ISSCC14 [6]	THIS WORK
Tech.(nm)/Supply(V)	180HV/1	65/0.8	130/1.2	180/1.2	130/1.2&3.3
Area/Ch (mm²)	-	0.024	0.023[1]	-	0.018[1]
Power/Ch (µW)	8	0.8[2]	1.26[2]	31	0.79[2]
# of Channels	64	16	32	3	64
IRN (µVrms)	1.6	0.99	1.6	0.6	2.1
Noise BW (Hz)	1-500	1-500	0.1-500	0.5-150	1-500
NEF/PEF (–ADC)	NA	NA	NA	9.6/110.5	NA
NEF/PEF (+ADC)	7.8/60.8	1.81/2.6	2.86/9.8	NA	2.98/10.6
Artifact Insensitive	YES	YES	YES	YES	YES
Tracking BW (Hz)	1M*	64k*	1M	10	10M
Input Imped. (MΩ)	30*	26	100	500	250
DEO Range (V_{pp})	±0.2	±0.13	±0.6	±0.4	±1.5
Bandwidth (Hz)	1-500	0.1-5k	0.1-1k	0.5-150	0.1-5k
DSP	NO	NO	YES	YES	YES
CPU	-	-	openMSP430	Cortex-M0	openMSP430
Accelerators	-	-	SE, PLV, CFC	MAU	MAC, FIR, IIR, QRS, CORDIC
Memory (kB)	-	-	96	128	96
Classifier	-	-	EDM-SVM	-	EDM-SVM
Sensitivity (%)	-	-	97.7	-	97.7
Stimulation/Wireless	YES/NO	NO/NO	YES/NO	NO/NO	YES/YES

[1] With Stimulator [2] Modulator Power -: Not Reported NA: Not Applicable
*: Estimated DEO: Differential Electrode Offset

Figure 22.8.6: Comparison with the state-of-the-art (SoA).

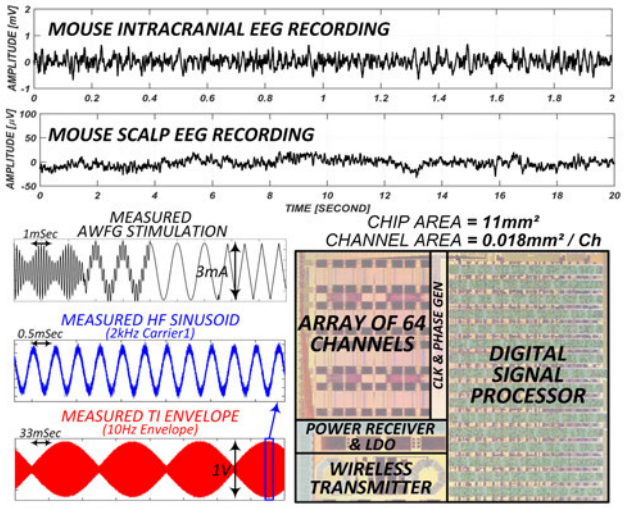


Figure 22.8.7: Examples of experimentally measured: in vivo signals in mice (top); and arbitrary-waveform and temporal interference (TI) envelope stimuli in saline (bottom left). The die micrograph (bottom right).

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 23 Overview: *DRAM*

MEMORY SUBCOMMITTEE

Session Chair: ***Wolfgang Spirkl***
Micron Semiconductor Deutschland,
Munich, Germany

Associate Chair: ***Dong Uk Lee***
SK hynix, Gyeonggi-do, Korea

Subcommittee Chair: Jonathan Chang, TSMC, Hsinchu, Taiwan

First devices for the new DRAM standards LPDDR5 and DDR5 are implementing improvements in bandwidth and power efficiency. The new devices will have a huge impact on a very wide range of applications, from IoT applications and smartphones to server and workstation applications. A new proposal for managed LRDIMM promises to reduce cost and power, and to provide capacities up to 512GB. For next generation DRAM interfaces a PAM-3 transceiver with 27Gb/s/pin on the base of 3 bits per 2 symbols is presented.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

8:30 AM

23.1 A 7.5Gb/s/pin LPDDR5 SDRAM with WCK Clocking and Non-Target ODT for High Speed and with DVFS, Internal Data Copy, and Deep-Sleep Mode for Low Power

K-S. Ha, Samsung Electronics, Hwaseong, Korea

In Paper 23.1, Samsung presents a 1st-generation 10nm class process LPDDR5 with schemes that increase the maximum bandwidth such as WCK clocking and non-target ODT (NT-ODT). It also reduces power consumption using techniques such as dynamic voltage frequency scaling (DVFS), and a deep-sleep mode (DSM).

9:00 AM

23.2 A 1.1V 1ynm 6.4Gb/s/pin 16Gb DDR5 SDRAM with a Phase-Rotator-Based DLL, High-Speed SerDes and RX/TX Equalization Scheme

D. Kim, SK hynix, Icheon, Korea

In Paper 23.2, SK hynix presents a DDR5 SDRAM to overcome bandwidth, power and capacity limitations of DDR4. This paper presents a 16Gb 6.4Gb/s/pin DDR5 SDRAM with a phase-rotator-based DLL, write-level training, and RX/TX with the enhanced DFE/FFE. The energy efficiency is improved by more than 30% with 1.1V/1.8V VDD and VPP.

9:30 AM

23.3 A 3-bit/2UI 27Gb/s PAM-3 Single-Ended Transceiver Using One-Tap DFE for Next-Generation Memory Interface

H. Park, Korea University, Seoul, Korea

In Paper 23.3, Korea University presents a 3bit/2UI 1.03pJ/bit PAM-3 single-ended TRX. An 27Gb/s PAM-3 symbol is generated with an output driver voltage of 0.6V and a 1/3-rate forwarded clock frequency of 9GHz. The RX adopts a 1-tap tri-level DFE, which has the same complexity as for NRZ signaling to equalize the PAM-3 signal. Fabricated in a 28nm CMOS technology, the proposed PAM-3 TRX can be utilized for the next generation memory interface with low power.

9:45 AM

23.4 A 512GB 1.1V Managed DRAM Solution with 16GB ODP and Media Controller

S. Lee, SK hynix, Icheon, Korea

In Paper 23.4, SK hynix presents a Managed DRAM Solution (MDS) DIMM with 512GB capacity. It consists of Octal-Die Package (ODP) structured memory (media) and a media controller with ECC. This paper deals with architecture and ODP structure of the media for cost minimization, and pre-CMD scheme for power reduction. Dies per wafer increase of 26% compared to conventional DRAM with same process and same capacity was achieved. The DIMM power consumption is 12W.

23.1 A 7.5Gb/s/pin LPDDR5 SDRAM with WCK Clocking and Non-Target ODT for High Speed and with DVFS, Internal Data Copy, and Deep-Sleep Mode for Low Power

Kyung-Soo Ha, Chang-Kyo Lee, Dongkeon Lee, Daesik Moon, Jin-Hun Jang, Hyong-Ryol Hwang, Hyungjoon Chi, Junghwan Park, Seungjun Shin, Dukha Park, Sang-Yun Kim, Sukhyun Lim, Kiwon Park, YeonKyu Choi, Young-Hwa Kim, Younghoon Son, Hyunyoon Cho, Byongwook Na, Hyo-Joo Ahn, Seungseob Lee, Seouk-Kyu Choi, Youn-Sik Park, Seok-Hun Hyun, Soobong Chang, Hyuck-Joon Kwon, Jung-Hwan Choi, Tae-Young Oh, Young-Soo Sohn, Kwang-Il Park, Seong-Jin Jang

Samsung Electronics, Hwaseong, Korea

High-speed and low-power techniques for the latest mobile DRAMs, such as LPDDR4/4X [1-3], have been developed to enable high-resolution displays, multiple cameras and 4G communication in mobile devices. However, DRAM with higher bandwidth and lower power consumption than LPDDR4X is indispensable to support 5G communication, on-device artificial intelligence and advanced driver assistance systems. In this paper, we present a 1st generation 10nm-class process LPDDR5. It includes novel schemes that increase the maximum bandwidth, such as WCK clocking and non-target ODT (NT-ODT). Power consumption is also reduced by low power schemes, such as dynamic voltage-frequency scaling (DVFS), an internal data-copy function called write-X and a deep-sleep mode (DSM).

Figure 23.1.1 shows a block diagram of the LPDDR5 that uses WCK clocking. WCK signals are used for the write and read operations in LPDDR5 like GDDR5/6 [4,5]. As the data rate increases and the voltage decreases for LPDDR4X, the power-noise induced jitter becomes the dominant factor that limits high-speed operation due to the long clock-to-DQS delay (tDQSCK). In order to minimize tDQSCK in LPDDR5, WCK signals are sent to each byte. To reduce power consumption, WCK signals are not transmitted in the idle state. Thereby, WCK-to-CK (WCK2CK) synchronization must be executed for domain crossing between CK and WCK signals before DQ operations, such as read and write operations. There are two types of WCK clock trees in LPDDR5. CML clocking that includes a duty cycle adjustor (DCA) for SI improvement is for high-speed mode and CMOS clocking is for low-power mode.

Figure 23.1.2 shows the timing and circuits used for WCK2CK synchronization. When WCK signals are not synchronized with CK signals, a WCK2CK synchronization command (CAS(WS_RD/WR/FS)) must be issued before DQ operation commands are issued. The WCK waveforms consist of enable timing (tWCKENL), static period (tWCKPRE_static), half-rate and full-rate toggles (tWCKPRE_toggle). Each value is changed depending on the speed grade. tWCKENL is needed to minimize static current through the WCK channels in WCK ODT enabled mode. tWCKPRE_static and tWCKPRE_toggle are the minimum times required to enable a WCK divider in a static input condition and to complete the WCK2CK synchronization. Two half-rate WCK clock pulses are required for WCK2CK synchronization. The first WCK pulse is for mitigating ISI affect due to channel bandwidth limitation and the other half-rate WCK clock pulse is used for the WCK2CK synchronization. The internal synchronization signal that is generated by the CAS(WS) command is sampled by the divided WCK signals (WCK/2_90, 270). Accordingly, the correct WCK/2 signals are selected for domain crossing operations between CK and WCK domains. After read or write operations are completed, WCK buffers in LPDDR5 are automatically turned off at the WCK2CK SYNC off time to reduce power consumption. New WCK2CK synchronization command should be issued prior to the next DQ operation, because the previous WCK2CK synchronization information was lost when the WCK buffers are turned off.

LPDDRx DRAMs using a two-rank configuration are implemented for high density. However, since one DQ package pad is shared by a DQ pad of each die, as shown in Fig. 21.1.3, the reflection noise due to the forked DQ connection causes signal distortion beyond 5.5Gb/s. Therefore, LPDDR5 supports NT-ODT mode that mitigates the reflection noise to improve SI during high-frequency operation. The values of the NT-ODT and the target ODT (T-ODT) can be controlled by the memory controller depending on the channel characteristics. In NT-ODT mode, the ODT of each rank DRAM is always enabled. Thus, the WCK drivers of the memory controller should drive a 'Hi-Z' or '0' signal to prevent static current through WCK channels. As shown in Fig. 23.1.3, the SI of the write operation in NT-ODT mode improves by 32.5% at 6.4Gb/s due to the reduced signal reflection on the proto test system.

There are low-power schemes such as DVFS [6] and DSM as well as lower *V*DD operation: 1.1V (LPDDR4/4X) → 1.05V (LPDDR5). DRAM is divided into 3 parts from a structural point of view: (1) cell & core, (2) data-path and (3) I/O. DVFS can be only used at low frequencies (≤1.6Gb/s for data-path and ≤3.2Gb/s for I/O). A memory controller or a higher system supplies *V*DD2H (1.05V) and *V*DD2L (0.9V) to the DRAM and the internal power switches are used in DRAMs for the DVFS as shown in Fig. 23.1.4. *V*DDQ for I/O can be controlled by an external voltage regulator which provides 0.5V for high-speed and 0.3V for low-speed (≤3.2Gb/s) operation. When DVFS is enabled, IDD4R and IDD4W decrease by 8% and 9% at 1.6Gb/s operation, however, lower VDD2 (VDD2H → VDD2L) causes a 4ns delay increase of data-path repeaters.

Figure 23.1.4 shows a block diagram for the write-X operation. For an all-zero write operation, the write-X command can be issued without transmitting WCK signals and DQ s. The power consumption of a write-X operation is 58% smaller than a normal write operation at 4.266Gb/s. Although WCK signals are transmitted to maintain WCK2CK synchronization state, the power consumption of the write-X operation decreases by turning off the WCK clock tree after the WCK divider.

In DRAM idle state, since the self-refresh power is critical, DSM is proposed. In DSM, the internal voltage generators (V_{INT_1}) that are not used for refresh operations are turned off to reduce leakage current of the internal voltage and static current of the internal voltage generators. The burst operation mode is also used to reduce self-refresh power. It modulates the self-refresh interval from regular (tREFI=3.9μs) to minimum (tRFCab=280ns) during certain period. In burst operation mode, the internal voltage generators (V_{INT_2}) which are used for refresh operations are turned on, but these generators are turned off in pause mode. Thus, the power consumption in self-refresh power-down mode is improved by 25%. The DSM entry command can be issued at idle or self-refresh state, as shown by the state diagram in Fig. 23.1.5 and CS toggling is required to exit DSM. In addition, the transient time for exiting DSM (tXSR_DSM>200μs) should be guaranteed to allow for the internal voltage generators to recover from their off-state. Upon exiting DSM, the refresh interval becomes its regular interval (tREFI) in normal self-refresh power-down mode and self-refresh mode.

Figure 23.1.6 shows a tCK shmoo and an eye shmoo of read and write operations from the ATE test. A maximum operation speed of 7.5Gb/s (*V*DD2H=1.05V) is achieved. The read and the write DQ eyes are 0.56UI and 0.6UI (UI=133ps). At a 4.266Gb/s data rate, the read and the write power consumption of LPDDR5 are improved by 21% and 33% compared to LPDDR4X. DVFS (*V*DD2L) for data-path and DVFS (*V*DDQ) for I/O can be utilized up to 1.6Gb/s and 3.2Gb/s. At 1.6Gbp/s in DVFS-enabled mode, IDD4R and IDD4W decrease by 8% and 9%. The power consumption of self-refresh power-down with DSM decreases by 25% compared to conventional self-refresh power-down. Figure 23.1.7 shows a micrograph of the fabricated DRAM which occupies 45.9mm²/channel (8Gb).

References:
[1] T.-Y. Oh, et al., "A 3.2Gb/s/pin 8Gb 1.0V LPDDR4 SDRAM with integrated ECC engine for sub-1V DRAM core operation," *ISSCC*, pp. 430-431, 2014.
[2] C.-K. Lee, et al., "A 5Gb/s/pin 8Gb LPDDR4X SDRAM with Power-Isolated LVSTL and Split-Die Architecture with 2-Die ZQ Calibration Scheme," *ISSCC*, pp. 390-391, 2017.
[3] K.-C. Chun, et al., "A 16Gb LPDDR4X SDRAM with and NBTI-Tolerant Circuit Solution, an SWD PMOS GIDL Reduction Technique, an Adaptive Gear-Down Scheme and a Metastable-Free DQS Aligner in a 10nm Class DRAM Process," *ISSCC*, pp. 206-207, 2018.
[4] M. Brox, et al., "An 8Gb 12Gb/s/pin GDDR5X DRAM for Cost-Effective High-Performance Applications," *ISSCC*, pp. 388-389, Feb. 2017.
[5] Y.-J. Kim, et al., "A 16Gb 18Gb/s/pin GDDR6 DRAM with Per-Bit Trainable Single-Ended DFE and PLL-Less Clocking," *ISSCC*, pp. 204-205, 2018.
[6] T. D. Burd, et al., "A dynamic voltage scaled microprocessor system," *ISSCC*, pp. 294-295, 2000.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

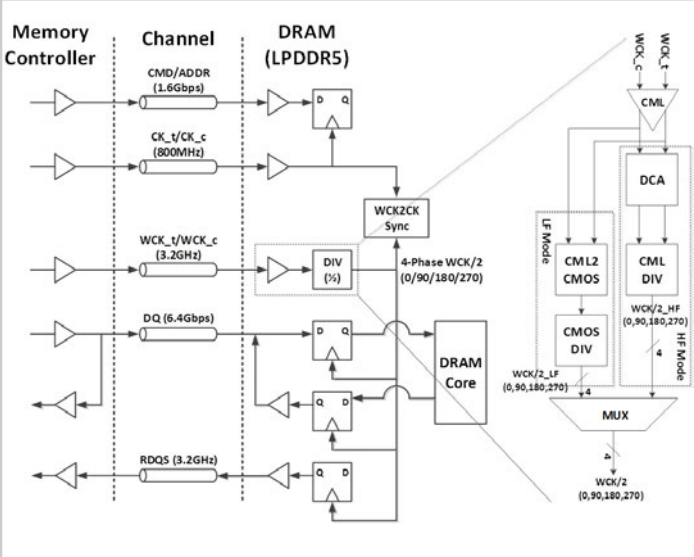


Figure 23.1.1: Block diagram including WCK clocking of LPDDR5.

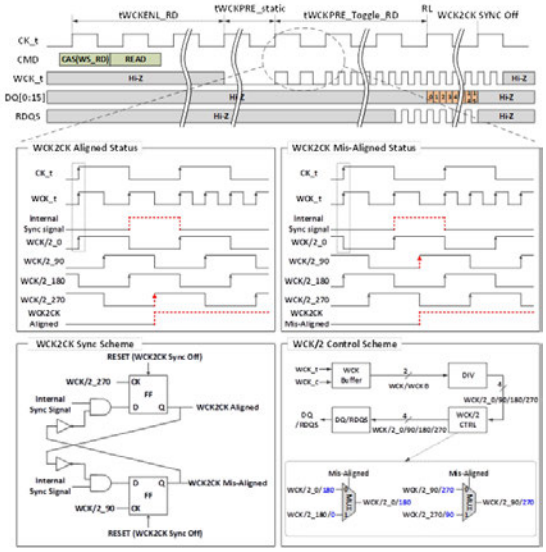


Figure 23.1.2: Timing of WCK clocking and WCK2CK synchronization.

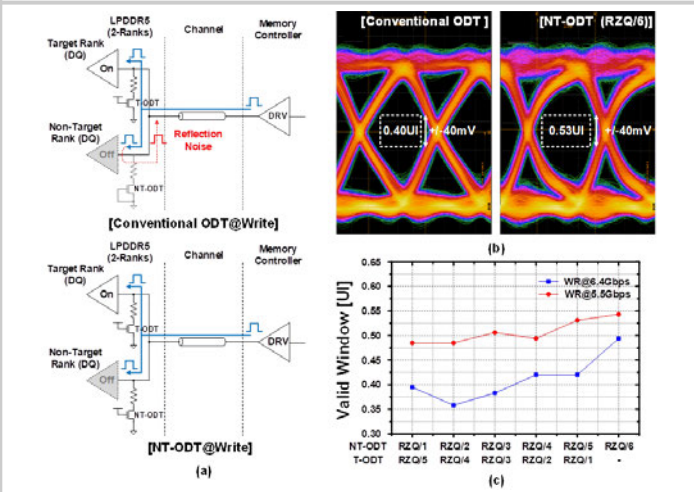


Figure 23.1.3: (a) Block diagram of conventional ODT and NT-ODT, (b) Measured eye diagram of 6.4Gbps data on proto test system (c) Write valid window depending on the combinations of NT-ODT value and T-ODT value on proto test system.

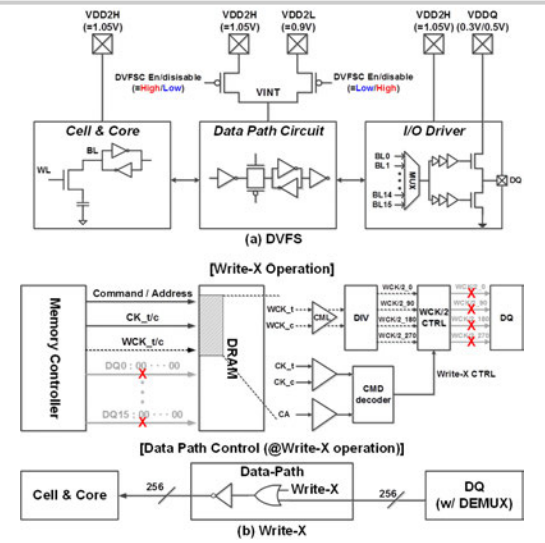


Figure 23.1.4: (a) Dynamic voltage frequency scaling (DVFS), (b) Internal data copy function (Write-X).

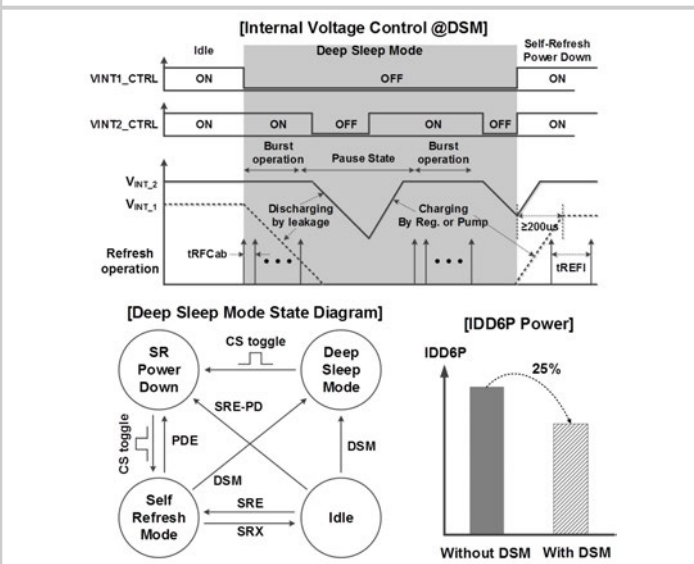


Figure 23.1.5: DSM voltage and timing control and state diagram.

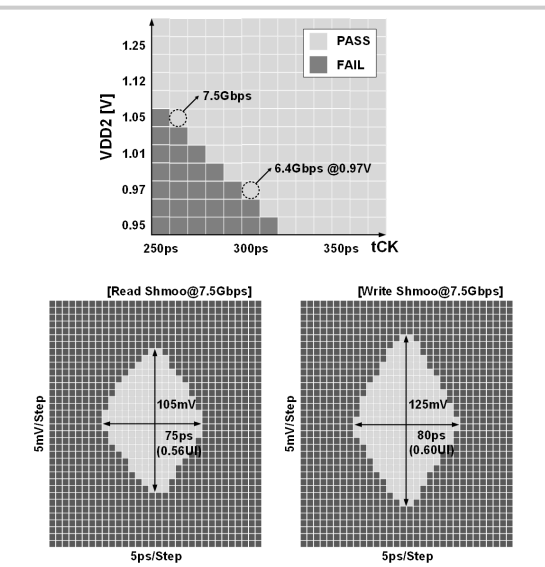


Figure 23.1.6: Measured tCK shmoo and read/write IO shmoo.

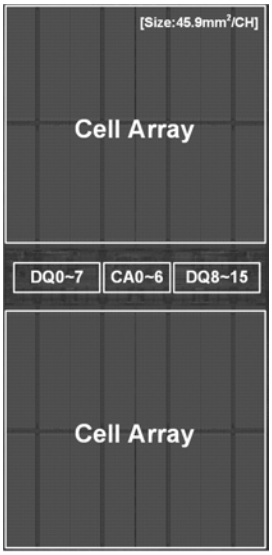


Figure 23.1.7: Chip micrograph of 8Gb/channel LPDDR5.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

23.2 A 1.1V 1ynm 6.4Gb/s/pin 16Gb DDR5 SDRAM with a Phase-Rotator-Based DLL, High-Speed SerDes and RX/TX Equalization Scheme

Dongkyun Kim, Minsu Park, Sungchun Jang, Jun-Yong Song, Hankyu Chi, Geunho Choi, Sunmyung Choi, Jaeil Kim, Changhyun Kim, Kyungwhan Kim, Kibong Koo, Seonghwi Song, Yongmi Kim, Dong Uk Lee, Jaein Lee, Daesuk Kim, Kihun Kwon, Minsik Han, Byeongchan Choi, Hongjung Kim, Sanghyun Ku, Yeonuk Kim, Jongsam Kim, Sanghui Kim, Youngsuk Seo, Seungwook Oh, Dain Im, Haksong Kim, Jonghyuck Choi, Jinil Chung, Changhyun Lee, Yongsung Lee, Joo-Hwan Cho, Junhyun Chun, Jonghoon Oh

SK hynix, Icheon, Korea

Required system performance for the computing and server, cost and power forces DRAM to improve its bandwidth, capacity and power. DDR5 SDRAM has been proposed as the next memory solution, with various new functions and circuit techniques to overcome the limitation of DDR4. Speed has been increased to 4.4 – 6.4Gb/s from DDR4's 3.2Gb/s. Energy efficiency has improved by over 30% using a 1.1V V_{DD} and a 1.8V V_{PP}. To achieve the performance and power consumption targets DDR5 adopted a small number of command pins with 2-cycle decoding, new write training methods, an on-die ECC, an un-matched DQ/DQS scheme, and an equalize scheme for the interface. This paper presents a 16Gb 6.4Gb/s/pin DDR5 SDRAM with a high-speed circuit techniques using a 1.1V DRAM process.

Figure 23.2.1 shows the functional block diagram of the proposed DDR5 SDRAM. It has 32 banks, an 8-bank group array with on die ECC for a 16Gb density. 14-address pins are implemented for command address decoding with a $2t_{CK}$ command instead of DDR4's 24 pins. This reduces the power consumption of the command address decoding operation. For a 2× bandwidth and maintaining high-channel efficiency, the burst length and the number of banks, and number of bank groups are changed. The burst length is extended to 16. The number of banks is 32, which is twice of DDR4. The number of bank groups is extended from 4 to 8. To improve the signal integrity, DFE and FFE are used for the receiver and transmitter. To test the BER of the transceiver a PRBS generator and bit error counter is used. To achieve an enhanced data vaild window, a DLL using a phase rotator (PR) and an injection locked oscillator (ILO) is proposed.

A conventional DRAM DLL controls the clock delay by changing the number of delay elements. As the clock speed increases, a large number of delay elements increases the clock jitter and clock duty cycle distortion. To alleviate this problem, the proposed DLL adopts a PR and an ILO as shown in Fig. 23.2.2 [1]. The PR interpolates between the input quadrature clocks and injects the output quadrature clocks into the ILO. To reduce switching glitches, the PR controls the amount of current to each inverter, instead of switching input clock phases. The ILO consists of a current-controlled two-stage differential ring oscillator, with the input quadrature clocks injected into its four oscillation nodes. The ILO improves PR linearity and reduces the quadrature-clock-phase error like an active poly-phase filter [2]. The ILO's free-running frequency is adjusted by a 6b-current DAC during the initial training period.

For a system using DQS[4], it is important to gate the DQS by the burst length after the DRAM receives a write command [3]. In order to gate the DQS, the internal write command which is synchronized by internal clock and DQS should be matched, but the write command-decoding delay exceeds several clock cycles at high speeds. In a DDR4, a delay is inserted internally to DQS to match with the internal command, as shown in Fig. 23.2.3. However, for DDR5 it is not possible to add a delay to DQS path because the additional delay causes the jitter that interferes the high speed operation. To solve the clock-domain-crossing problem, caused by this delay difference between CLK and DQS, we proposed a write-level training method that uses an unmatched CLK-to-DQS path. Figure 23.2.3 shows the block diagram of a DDR5 write with a DQS control scheme. Figure 23.2.3 also shows a flow chart of new write-level training sequence. While a conventional write leveling[4] synchronizes CLK and DQS at the external pin of the DRAM. Write level training of DDR5 synchronizes the CLK and DQS at the DQS gating point inside the DRAM. 1st step is external training like DDR4 for setting reference of multi rank DRAM. 2nd, 3rd step are internal training. 2nd step is coarse training, pulling the internal write enable signal (IWES) by t_{CK} through the mode register write (MRW) setting. 3rd step is fine training, sweeping the DQS signal to find the rising edge of IWES. After the training, IWES and DQS are matched without any additional delay. Therefore, it saves the power consumption of DQ/DQS delay and improves S/H characteristics.

For high-density servers a main memory should support multiple modules and have multi-drop topology. A multi-drop topology induces ISI noise from reflections and it is a bottleneck to increase interface speed. In the DDR5 SDRAM, a 4-tap DFE is defined to achieve data rate above 3.2Gb/s using a multi-drop channel topology. The DFE is an effective way to remove reflection noise. To minimize DFE overhead, the DFE adaptation is performed at the memory controller by the MRW. Figure 23.2.4 shows a block diagram of the DFE, which uses a 4-way interleaving architecture by using quad-rate 4-phase internal DQS signals. Because the 4-tap feedback data of each slicer are different, each slicer has separate DFE summer. The DFE summer adopts a CML-type circuit to support various data rates required for DDR5 SDRAM. A current integrating summer is widely used to reduce DFE feedback delay and to increase SNR. However, it requires an additional calibration sequence to normalize the integrating gain for various data rates.

FFE is widely used in interface circuit to improve operating speeds. In a DRAM process, however, full-rate flip-flops are difficult to implement, which makes conventional FFE schemes inadequate. The proposed FFE circuit optimized for DRAM process is shown in Fig. 23.2.4. A quarter-rate FFE is followed by 4:1 serialization. Since impedance should be matched with a fixed value when consecutive identical data are transmitted, the target impedance is divided into multiple parallel segments, each of which is controlled by MUX to change the tap coefficients of FFE coarsely. For finer control, we add one more segment at the cost of impedance mismatch.

DDR5 uses a high-speed serializer to multiplex 4-data inputs. The serializer comprises four data MUX blocks that use a cross-coupled latch, a tristate driver, and a clock signal controller (CLKCTL), as shown in Fig. 23.2.5. The clock signal controller is able to change the clock duty cycle via specific codes by enabling the additional transistor in the controller, according to the level of the clock signal and codes used: like "Duty H" and "Duty L" shown in top-right side of Fig. 23.2.5. The rising and falling slopes of the clock signal are different from each other, thereby finely adjusting the duty of the clock signal and finely adjusting the DQ output skew. This code value can be loaded externally and a read training can be used to find a code value that can correct the DQ skew.

The test methodology for high frequency random, deterministic jitter, duty cycle error and bounded uncorrected error is needed in single-ended signaling. Figure 23.2.5 shows the architecture of the on-chip BER circuit for a DDR5 system. The on-chip error-counter scheme for BER testing is applied to measure the bit error rate, and then to transfer the error rate to the tester. The on-chip BER circuit compares PRVG pattern from the tester with the pattern from DRAM and judges whether error occurs or not. If error occurs in DRAM, the BER system counts and tabulates them. Upon completion, the final counter value is loaded into the global IO to transfer data from DQs. The on-chip BER circuit can count and calculate up to 2^9 - 1 bit errors.

The 16Gb DDR5 SDRAM is fabricated in an 1*y*nm, 4-metal layer, DRAM process. Figure 23.2.6 shows the measurement results for the fabricated DDR5 SDRAM. The chip achieves an IO throughput of 6.4Gb/s/pin using a 1.04V supply voltage, as shown by the frequency-voltage shmoo. The Tx output window is measured at 1.1V and 4.4Gb/s. The horizontal and vertical eye opening are 170ps (0.75UI) and 450mV. The Rx eye-diagram shows the effect of DFE scheme at 1.1V and 4.4Gb/s. Time of data width (t_{DQ}) is improved by 20%; the voltage data height is improved by 60% by using DFE training. Figure 23.2.7 shows the chip microphotograph and a performance summary of the 16Gb DDR5 SDRAM.

References:
[1] F. O'Mahony, et al., "A programmable phase rotator based on time-modulated injection-locking," *VLSI*, pp. 45-46, 2010.
[2] K. Kim, et al., "A 5.4mW 0.0035mm2 0.48psrms-jitter 0.8-to-5GHz non-PLL/DLL all-digital phase generator rotator in 45nm SOI CMOS," *ISSCC*, pp. 98-99, 2009.
[3] S-M. Lee, et al., "A 0.6V 4.266Gb/s/pin LPDDR4X Interface with Auto-DQS Cleaning and Write-VWM Training for Memory Controller," *ISSCC*, pp. 398-400, 2017.
[4] DDR4 SDRAM Specification (JESD79-4), JEDEC Standard, JEDEC solid state technology association, Sept. 2012.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

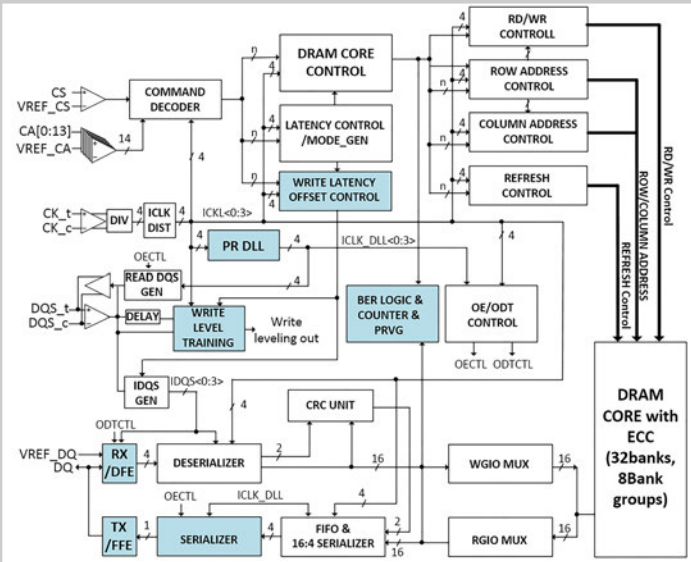


Figure 23.2.1: DDR5 Functional block diagram.

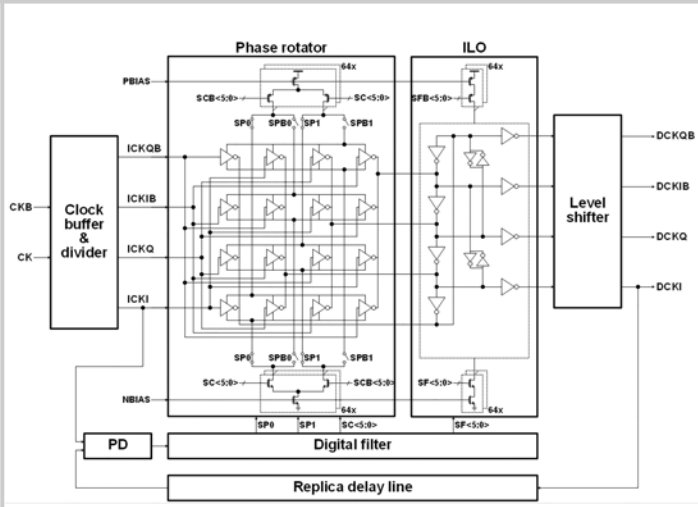


Figure 23.2.2: Implemented DLL based on a phase rotator and an injection-locked oscillator.

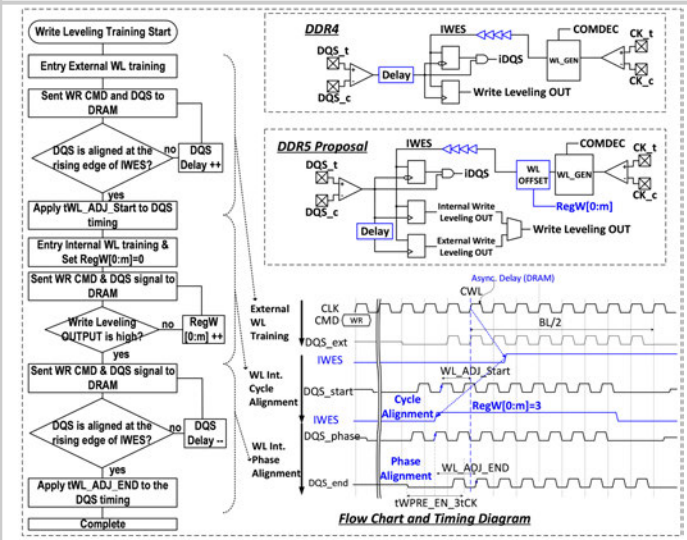


Figure 23.2.3: Write level training: Write leveling scheme on DDR4, Write level training scheme on DDR5, Write level training sequence and timing diagram.

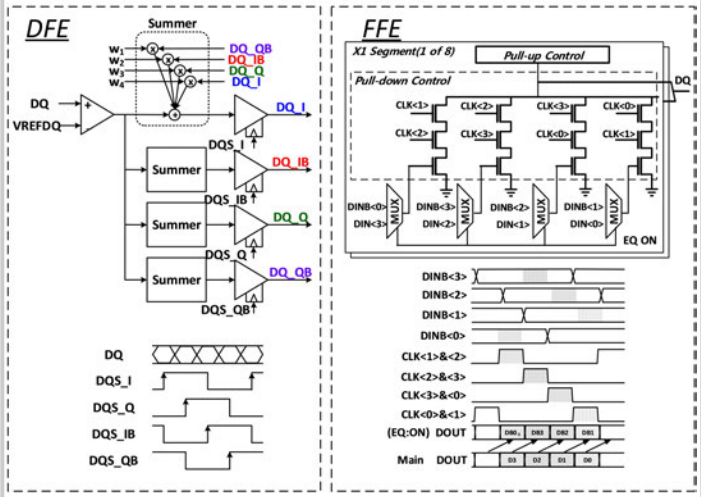


Figure 23.2.4: Equalizer scheme on DDR5: Block diagram and timing diagram of DFE(left), Block diagram and timing diagram of Quarter-rate FFE(right).

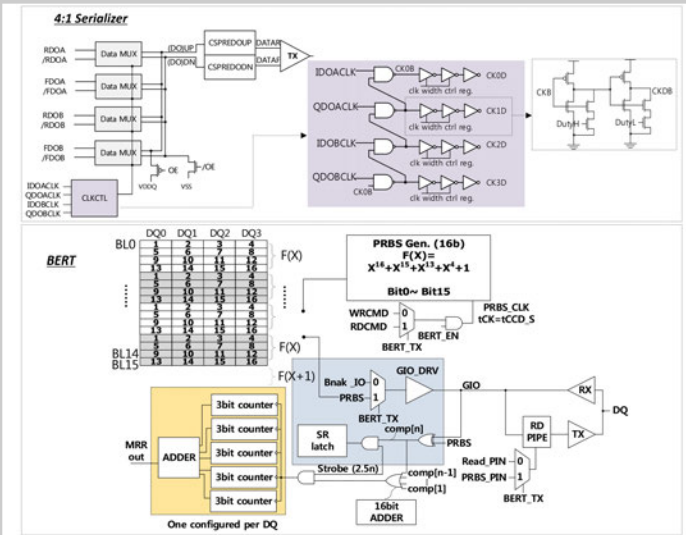


Figure 23.2.5: 4:1 Read serializer with clock control (top) and BER test scheme to evaluate for DDR5 interface (bottom).

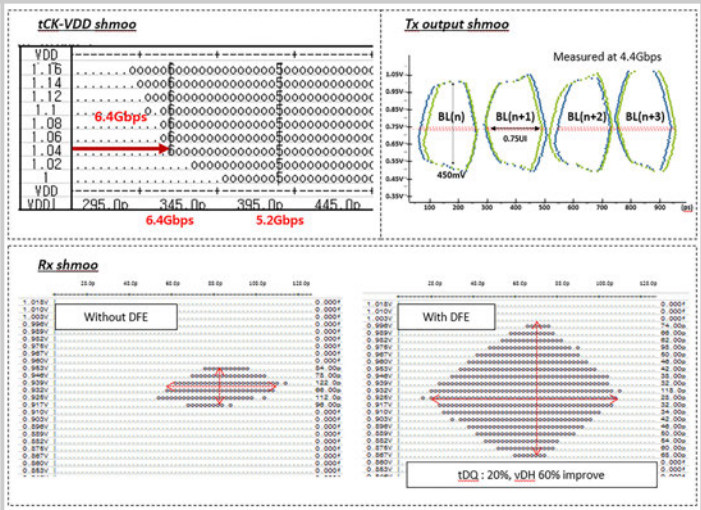
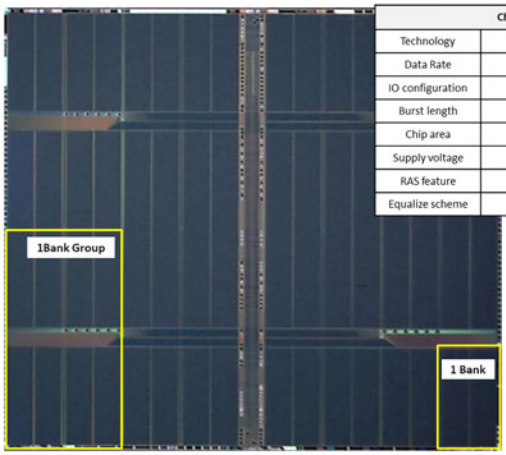


Figure 23.2.6: Measured tCK vs VDD shmoo (top left), Tx output shmoo(tDV) using PR-DLL (top right) and Rx shmoos according to adoption of DFE scheme at 1.1V and 4.4Gb/s (bottom).



Chip summary	
Technology	1y-nm 4-metal DRAM process
Data Rate	6.4 Gbps/pin
IO configuration	X4/x8/x16
Burst length	BC8, BL16 on-the-fly
Chip area	76.22mm^2
Supply voltage	VDD/VDDQ 1.1V, VPP 1.8V
RAS feature	In DRAM ECC
Equalize scheme	DFE/FFE

Figure 23.2.7: Chip micrograph and summary table.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

23.3 A 3-bit/2UI 27Gb/s PAM-3 Single-Ended Transceiver Using One-Tap DFE for Next-Generation Memory Interface

Hyunsu Park[1], Junyoung Song[2], Yeonho Lee[1], Jincheol Sim[1], Jonghyuck Choi[1], Chulwoo Kim[1]

[1]Korea University, Seoul, Korea
[2]Incheon National University, Incheon, Korea

Bandwidths of memory interfaces have been increased tremendously to enable high-data throughput while maintaining single-ended signaling and the supply voltage of I/O has been scaled down. Due to the increasing interface bandwidth the required area and power consumption has increased as well, resulting in higher I/O circuit design costs [3]. A high-loss channel causes ISI, which in turn limits the maximum data rate. Therefore, complex equalizers are needed for compensation, resulting in additional power dissipation and area overhead. As the data sampling rate increases, the deterministic and random noises degrade the data sampling margin and further limit the bandwidth. To lessen the negative impact of high channel loss and to reduce the forwarded clock frequency, multi-level signaling, such as PAM-4, can be used, as shown in Fig. 23.3.1 [2]. While the voltage sense margin for PAM-4 is theoretically $^1/_3$ of NRZ, in practice it is smaller due to simultaneous switching noise (SSN), crosstalk, and random noise in single-ended signaling. Eventually, the reduced voltage sense margin degrades the SNR, which causes a reduction in the BER. On the other hand, PAM-3's voltage sense margin is ½ of NRZ's. Duo-binary signaling is commonly used for PAM-3 signaling [1]. However, the pin efficiency and the forwarded clock frequency for duo-binary signaling is the same as for NRZ. In this paper, a 3b/2UI PAM-3 single-ended memory interface is proposed, with a pin efficiency of 150% and a reduced clock frequency, compared to NRZ signaling. To address PAM-3 equalizer inefficiencies a tri-level decision feedback equalizer (DFE) is implemented in the receiver (RX).

A high-level block diagram of the proposed PAM-3 single-ended transciever is shown in Fig. 23.3.1. In the TX, three 8:1 serializers convert 1.125Gb/s parallel data into 9Gb/s full-rate data (A, B, C) using an external 9GHz clock and an internal PRBS-7 pattern generator. The 3b/2UI half-rate encoder convert three data bits to the 18Gb/s up and down signals (UP/DN) to generate the 27Gb/s PAM-3 data via the N-over-N driver. The generated single-ended PAM-3 data stream is transmitted to the RX, along with the forwarded clock at $^1/_3$ of the data rate. In the RX, the single-ended-to-differential (S2D) amplifier with ground-termination for LPDDR interface converts the input data to the differential output. The 1-tap half-rate DFE compensates the post-cursor, and the three-original data are decoded by the 3b/2UI full-rate decoder. The quadrature-phase clocks, which are generated by the 4-phase generator with a duty-compensated half-rate clock, are used for the DFE and the decoder in the RX.

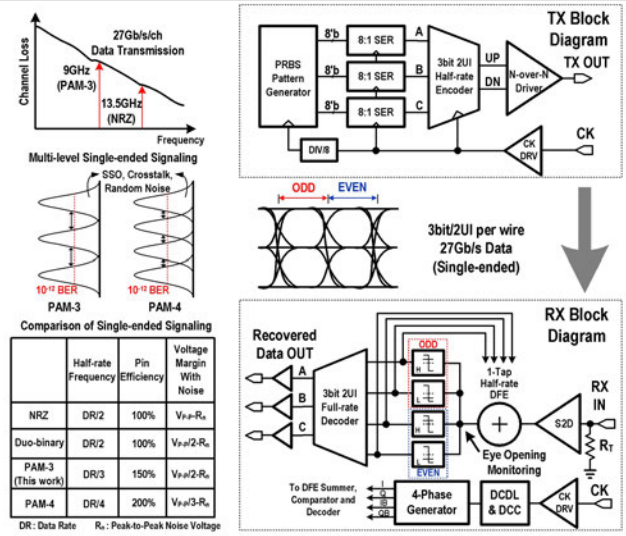
There are nine transition cases for the 3b/2UI PAM-3 signaling in the TX output, as shown in Fig. 23.3.2. Eight of which are mapped to express 3b for every 2UI. The remaining low-to-high transition are unassigned and used to detect the odd/even data in the RX, because there is no low-to-high transition between the odd and even data in normal operation. The generated UP/DN data are serialized by the 2:1 multiplexer and converted to singled-ended PAM-3 data by the N-over-N output driver. The output impedance of the binary-weighted driver is controlled for impedance matching with a ground-terminated 50Ω resistor in the RX. The supply voltage of the low-swing driver is 0.6V for the 308mV peak-to-peak voltage of the PAM-3 output.

In the proposed PAM-3 RX, the input stage consists of the ground terminated resistor and the S2D amplifier. A feedforward path is added to the first stage of the S2D amplifier to match the differential gain. The outputs of the S2D amplifier are fed to the 1-tap half-rate DFE for post-cursor cancellation. In the case that DFE has more than two input voltage levels, more than two current branches are used to express the different input levels. However, there is an unused case in the PAM-3 DFE which causes this inefficient in the PAM-3 DFE. In the proposed PAM-3 DFE, the unit tap of the summer is designed with only one current branch, like the DFE for NRZ signaling, while expressing three input voltage levels depending on the data as shown in Fig. 23.3.3 [3]. Comp_H and Comp_LB are applied to the tap as inputs, instead of the differential inputs for the three cases. When the gate voltages become zero, half of the tail current flows through each input MOS, and it expresses the middle level. This enables the DFE summer for the PAM-3 interface to be implemented with the same complexity as NRZ signaling. Due to the reduced number of current branches, the parasitic capacitance and the feedback delay is reduced. In the proposed DFE summer, Fig. 23.3.4, two current branches are used for half-rate operation, and the PMOS switches operate off the quadrature phase clocks (Q_{CK}, QB_{CK}). The strong-arm clocked comparators make the decision with I_{CK}, IB_{CK}, and the internal reference voltages. The outputs of the comparators are return-to-zero (RZ) signals and must be converted to non-return-to-zero (NRZ) signals for the proper operations of the DFE and the decoder. A CML-based SR-latch is commonly used for fast RZ-to-NRZ conversion; however, this consumes large power. In Fig. 23.3.4, the proposed RZ-to-NRZ converter is implemented in CMOS logic with a PMOS switch. When the clock is high, the output node of the RZ-to-NRZ converter is in high-impedance. This achieves both low-power and high-speed RZ-to-NRZ conversion without using a CML-based SR latch. The converted outputs are fed back to the DFE summer to cancel out the postcursor and to decode the data back to the three-original data (A, B, C) bits by the 3b/2UI full-rate decoder; which has the same logic depth as the encoder in the TX.

An external 9GHz clock, with an RMS jitter of 304fs, is applied to both the TX and the RX to generate and recover the data. The measured eye diagrams of the PAM-3 TX outputs are shown in Fig. 23.3.5. The 27Gb/s PAM-3 data are generated with a peak-to-peak swing of 308mV, and a sampling margin of 0.439UI for a single-channel without crosstalk. In the proposed RX, the delay of the clock, the threshold voltages of the comparators, and the coefficient of the DFE are manually controlled to achieve optimal BER. The measured eye diagrams and the bathtubs of the recovered PRBS-7 outputs are shown in Fig. 23.3.5. The bathtubs of the proposed PAM-3 interface are measured with a programmable error detector, by controlling the delay of the clocks in the RX. To prove the functionality of the proposed PAM-3 DFE, an internal eye-opening monitoring (EOM) circuit was added to the output node of the DFE summer. The eye was closed at the EOM circuit when the DFE is not used. On the other hand, with the use of the proposed PAM-3 DFE, the BER was measured at less than 10^{-12}, and an opened eye is obtained. A summary comparison with state-of-the-art memory interfaces is shown in Fig. 23.3.6. Compared to NRZ signaling, the proposed PAM-3 single-ended memory interface achieves a 150% pin efficiency, and the forwarded clock frequency is decreased by 33.3% for the same NRZ data rate. The power dissipation of the TX and the RX is 10.69mW and 17.03mW. For an 8b DQ pin case, the power efficiency will be further reduced since the forwarded clock can be shared. The power efficiency of the proposed transceiver is 1.03pJ/b at 27Gb/s due to the proposed energy-efficient PAM-3 DFE and the reduced forwarded clock frequency.

References:
[1] S. M. Lee, et al., "An 80 mV-Swing Single-Ended Duobinary Transceiver with a TIA RX Termination for the Point-to-Point DRAM Interface," *JSSC*, vol. 49, no. 11, pp. 2618-2630, Nov. 2014.
[2] W. Cho, et al., "A 38mW 40Gb/s 4-Lane Tri-Band PAM-4 / 16-QAM Transceiver in 28nm CMOS for High-Speed Memory Interface," *ISSCC*, pp. 184-185, 2016.
[3] J. Song, et al., "1V 10Gb/s/pin Single-Ended Transceiver with Controllable Active-Inductor-Based Driver and Adaptively Calibrated Cascade-DFE for Post-LPDDR4 Interfaces," *ISSCC*, pp. 320-321, 2015.
[4] A. Shokrollahi, et al., "A Pin-Efficient 20.83Gb/s/wire 0.94pJ/bit Forwarded Clock CNRZ-5-Coded SerDes up to 12mm for MCM Packages in 28nm CMOS," *ISSCC*, pp. 182-183, 2016.
[5] J. Wilson, et al., "A 1.17pJ/b 25Gb/s/pin Ground-Referenced Single-Ended Serial Link for Off- and On-Package Communication in 16nm CMOS Using a Process- and Temperature-Adaptive Voltage Regulator," *ISSCC*, pp. 276-277, 2018.

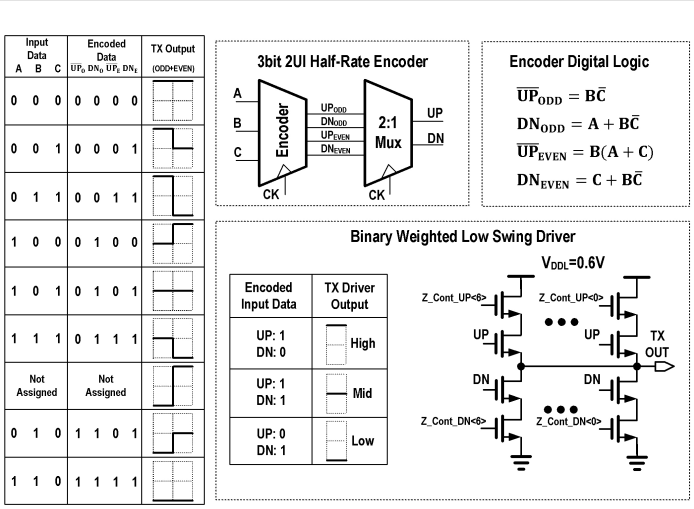
 978-1-5386-8531-0/19/$31.00 ©2019 IEEE



	Half-rate Frequency	Pin Efficiency	Voltage Margin With Noise
NRZ	DR/2	100%	V_{P-P}/R_n
Duo-binary	DR/2	100%	$V_{P-P}/2/R_n$
PAM-3 (This work)	DR/3	150%	$V_{P-P}/2/R_n$
PAM-4	DR/4	200%	$V_{P-P}/3/R_n$

DR : Data Rate R_n : Peak-to-Peak Noise Voltage

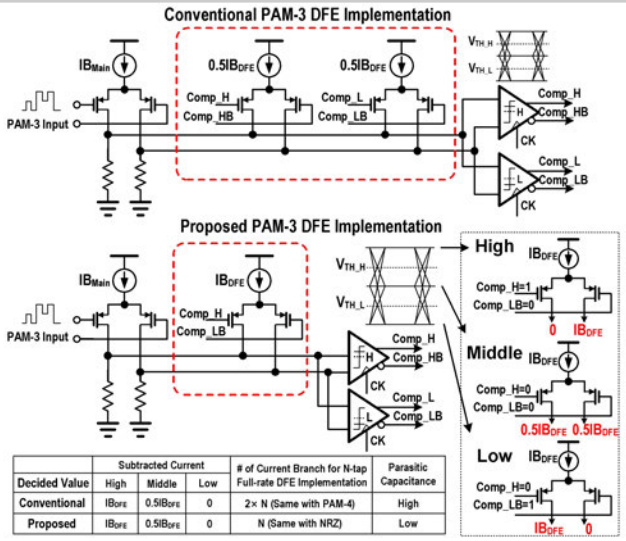
Figure 23.3.1: Top block diagram of the proposed PAM-3 single-ended transceiver.



Input Data A B C	Encoded Data $\overline{UP}_0$ DN_0 $\overline{UP}_E$ DN_E	TX Output (ODD+EVEN)
0 0 0	0 0 0 0	
0 0 1	0 0 0 1	
0 1 1	0 0 1 1	
1 0 0	0 1 0 0	
1 0 1	0 1 0 1	
1 1 1	0 1 1 1	
Not Assigned	Not Assigned	
0 1 0	1 1 0 1	
1 1 0	1 1 1 1	

Encoded Input Data	TX Driver Output
UP: 1 DN: 0	High
UP: 1 DN: 1	Mid
UP: 0 DN: 1	Low

Figure 23.3.2: Proposed 3bit/2UI encoding and low-swing driver for PAM-3 modulation.



Decided Value	Subtracted Current High	Middle	Low	# of Current Branch for N-tap Full-rate DFE Implementation	Parasitic Capacitance
Conventional	IB_{DFE}	$0.5IB_{DFE}$	0	2× N (Same with PAM-4)	High
Proposed	IB_{DFE}	$0.5IB_{DFE}$	0	N (Same with NRZ)	Low

Figure 23.3.3: Operation of the conventional and proposed PAM-3 DFE with 3 different cases.

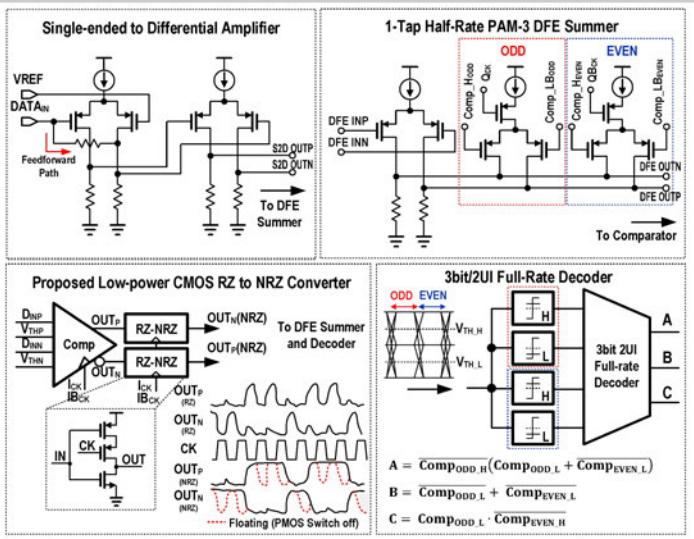


Figure 23.3.4: Implementations of the proposed PAM-3 receiver and 3bit/2UI decoding.

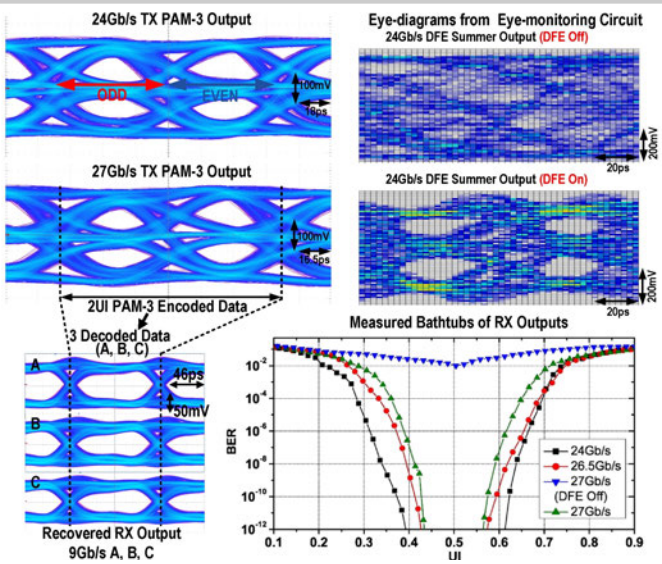
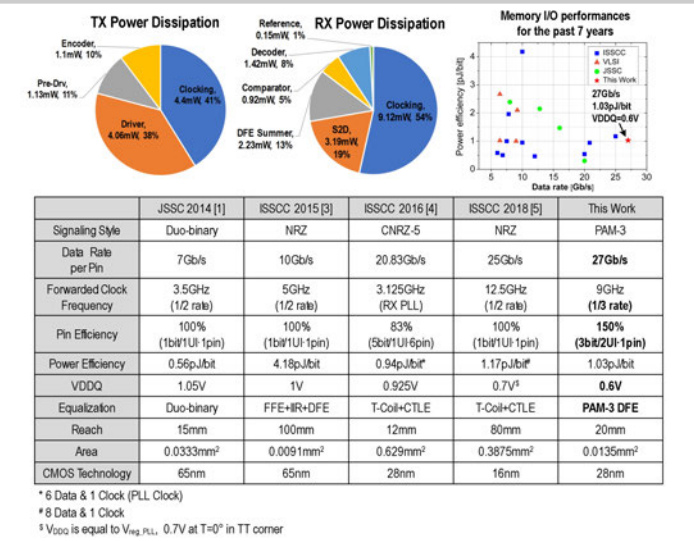


Figure 23.3.5: Measured eye-diagrams and bathtubs of the PAM-3 TRX.



	JSSC 2014 [1]	ISSCC 2015 [3]	ISSCC 2016 [4]	ISSCC 2018 [5]	This Work
Signaling Style	Duo-binary	NRZ	CNRZ-5	NRZ	PAM-3
Data Rate per Pin	7Gb/s	10Gb/s	20.83Gb/s	25Gb/s	**27Gb/s**
Forwarded Clock Frequency	3.5GHz (1/2 rate)	5GHz (1/2 rate)	3.125GHz (RX PLL)	12.5GHz (1/2 rate)	9GHz **(1/3 rate)**
Pin Efficiency	100% (1bit/1UI·1pin)	100% (1bit/1UI·1pin)	83% (5bit/1UI·6pin)	100% (1bit/1UI·1pin)	**150% (3bit/2UI·1pin)**
Power Efficiency	0.56pJ/bit	4.18pJ/bit	0.94pJ/bit*	1.17pJ/bit#	1.03pJ/bit
VDDQ	1.05V	1V	0.925V	0.7V§	**0.6V**
Equalization	Duo-binary	FFE+IIR+DFE	T-Coil+CTLE	T-Coil+CTLE	**PAM-3 DFE**
Reach	15mm	100mm	12mm	80mm	20mm
Area	0.0333mm²	0.0091mm²	0.629mm²	0.3875mm²	0.0135mm²
CMOS Technology	65nm	65nm	28nm	16nm	28nm

* 6 Data & 1 Clock (PLL Clock)
\# 8 Data & 1 Clock
§ V_{DDQ} is equal to V_{reg_PLL}, 0.7V at T=0° in TT corner

Figure 23.3.6: Power breakdowns, memory I/O performances, and comparison table.

23

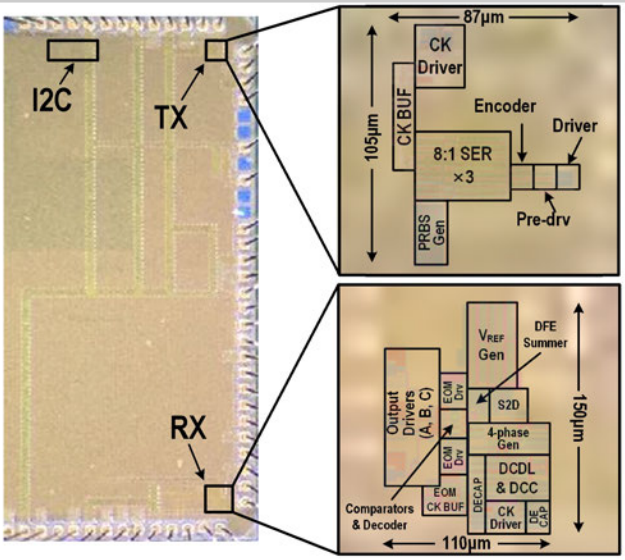


Figure 23.3.7: Microphotograph of the proposed PAM-3 TRX.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

23.4 A 512GB 1.1V Managed DRAM Solution with 16GB ODP and Media Controller

Seongju Lee, Byungdeuk Jeon, Kyeongpil Kang, Dongyoon Ka, Nayeon Kim, Yongseop Kim, Yunseok Hong, Mankeun Kang, Jinyong Min, Mingyu Lee, Chunseok Jeong, Kwandong Kim, Doobock Lee, Junghyun Shin, Yuntack Han, Youngbo Shim, Youngjoo Kim, Yongsun Kim, Hyunseok Kim, Jaewoong Yun, Byungsoo Kim, Seokhwan Han, Changwoo Lee, Junyong Song, Houk Song, Il Park, Yongju Kim, Junhyun Chun, Jonghoon Oh

SK hynix, Icheon, Korea

While the fast-growing big-data and cloud-computing markets are driving demand for server-oriented high-capacity memory, the high costs and high-power consumption due these high capacities can be major problems when building a server. In this situation, a high-capacity 512GB managed DRAM solution (MDS), bigger than any DIMM currently available, can be a good alternative. Despite such a large capacity, MDS has a price competitiveness due to 26% higher dies per wafer than a conventional dram, and uses 12W of power, which is similar to existing LRDIMM solutions that have a smaller density.

The MDS DIMM shown in Fig. 23.4.1 consists of the media, a media-controller, a data buffer (DB), and a voltage regulator (VR). Each media package is a 16GB octal-die package (ODP), with 8 slices of a 16Gb DRAM, each of ODP has 4 DQs. The DIMM consists of front and back physical ranks, with 20 ODPs per rank. 4 of the ODPs are used for ECC. The media-controller exchanges data with each other in the middle of the media and the DRAM controller. Communication with the DRAM controller uses the latency-relaxed DDR4 standard or a NVDIMM-P protocol operating at 3.2Gbps. For media communication, it operates at a speed of 2133Mbps via an asynchronous FIFO. Major functions include chip-kill, Reed-Solomon (RS) based error correction, link training, and module level test. Of these, chip-kill is a function that allows DIMM to work properly even if one of the packages in the DIMM fails. The Module level test consists of built-in self-test (BIST), built-in redundancy analysis (BIRA) and built-in self-repair (BISR). There are also DBs to speed up data transmission and a VR to provide a stable 1.1V V_{DD} supply to the media and some parts of the media-controller.

The media package has an ODP structure for high density and it is the first DRAM ODP in which all slices in the package share the same DQ. This structure uses wire-bonding thus it is possible to make ODP at a much lower cost than a conventional DDR4, which connects each slice using TSV. In addition, the structure uses only one wire-bonding per DQ on a single slice, unlike a conventional DDR4 that requires eight TSVs per DQ on a single slice [1]. Command/Address (C/A) signals are shared between 8 slices. However, in the case of DQ/DQS signal, cannot have this structure because signal integrity is reduced due to heavy load on the pad. As shown in the top of Fig. 23.4.2 only the DQ/DQSs of the two master slices (slice 0 and 4) are directly connected to the media-controller, using the main TX/RX. Slave slices 1, 2, 3 and slices 5, 6, 7 are connected to slice 0 and slice 4 via sub-TX/RX. Since only the pads of the two slices are directly connected to the media controller, the load on the pads is significantly reduced compared to when all eight pads are connected. This load reduced structure makes 2133Mbps data-transfer possible in ODP structures. The bottom of Fig. 23.4.2 shows the operation of the master and slave during write and read operations. When writing to the master, only the master operates; when writing to the slave, both the master and the slave operate. Master or slave write operations have almost the same latency because the data in both cases pass through the same number of circuits; similarly for a read operation.

Another key feature of MDS is cost minimization. Fig. 23.4.3 shows the media architecture, with a 16Gb density over 16 banks; this is a greatly increased matrix and bank size compared to a conventional DRAM. An increased matrix and bank size reduces the number of control blocks such as sense-amplifiers and XDECs and increases dies per wafer compared to the conventional DRAM. However, the increase in the matrix size may reduce the sensing margin and increase the defect rates. The ECC in MDS is used to resolve this issue. The top of Fig. 23.4.4 shows ECC operation. During read MDS DIMM outputs 80 DQs from 20 ODPs of 1 physical rank. From these, 64 DQs from 16 ODPs are used as data bits and 16 DQs from the remaining 4 ODPs are used as parity bits. Up to 8-symbol errors, from the set of 80 symbols, can be corrected during read. A DQ, including BL0-7, is a symbol and the symbol can be repaired regardless of how many bits have failed within that symbol [2]. With the help of ECC, we reduced the size of the media by 20% compared to existing memory with the same density and in the same technology node. This results in a 26% increase in dies per wafer [3]. In addition, ECC can contribute to cost reduction increasing media yield. Since ECC is located on the media-controller, ECC can be operated on DIMM test but not on wafer/package test. Thus, the media has an N-bit fail-waive scheme which ignores certain failure bits (symbol w.r.t. RS ECC) occurring in wafer/package test. The bottom of Fig. 23.4.4 shows the relevant block diagram. The expected data written at the latch are changed to the data suitable for test by the pattern generator. Then, the changed data is compared with the read data to detect failed bits. The fail bit counter counts the number of failed bits. If bit failures exceed the pre-set ignorable count, the failure is sent via the FLAG signal. The FLAG signal is sent to DQ pad to inform that failure has occurred and also sent to the fail address register to perform self-repair. The TBITSEL<0:1> signals determine the number of failed bits to ignore and can be selected between 0 and 3. This number is determined by considering ECC coverage of the DIMM. If fewer bits are failed than the bits set to be ignored, the flag signal does not generated and the test result is treated as pass. The yield of the wafer/package test is greatly increased because the fail bits that can be covered by the ECC is judged pass.

In high-capacity DIMMs, such as MDS, the standby current can be a big problem, so a pre-CMD operation has been devised to save power. The upper-left of Fig. 23.4.5 shows a timing diagram for the pre-CMD. First, CSB is input ahead of the other signals and activates only one physical rank. Since the signals following CSB go into only one rank the power consumption is reduced. This is an important scheme that reduces power consumption in conventional DDR4 [4]. In addition to this, in MDS, C[0:2] selects the operating slice and is input after CSB. At this time, only C[0:2] RXs of one rank selected by CSB are enabled. After which, only C/A RXs of one slice selected by C[0:2] is enabled, thereby reducing the C/A RXs current by one eighth. In addition, until an external command is entered, the bulk-bias scheme in each area is enabled, reducing the standby current. Depending on the command combination, the bulk-bias scheme is disabled and the external address is entered: t_{ACL} is the time required for bulk-bias stabilization. The upper-right of Fig. 23.4.5 shows the block diagram for this scheme. In the power CTRL block the BLUP_OFF signal is sent to the core area. This reduces the leakage current between BL and BLb by increasing the ground level of the sense amplifier's BL from V_{SS} to V_{BLP} in active standby mode (IDD3N). This block also sends a signal to the periphery to control the bulk-bias by area. This reduces the standby current by making the bulk-bias deeper than in normal operation. Another power-saving feature that uses the pre-CMD is smart DLL update. In standby mode, instead of always updating the DLL, it periodically updates to reduce the DLL current. Bottom of Fig. 23.4.5 shows the relevant circuit and its timing diagram. If the read/write command is input, the DLL changes to its always updating mode and t_{ACL} is required for stable operation. The use of these schemes reduces 5W power, which is possible because the MDS has its own controller that can freely control the timing of each signal.

The upper-left of Fig. 23.4.6 shows the ODP speed shmoo plot, which shows that the target speed 2133Mbps is reached at V_{DD} = 1.1V. Since the media-controller is currently under design, the input-output test solution (IOTS) that can replace DIMM test has been emulated using an FPGA. It performs media-controller functions including, link training and media test functions such as BIST/BIRA. The IOTS consists of the FPGA, two 128GB SO-DIMMs and a voltage regulator as shown in Fig. 23.4.7. One of the key IOTS validation items is to ensure that the number of failed bits generated in DIMMs, made with media using the N-bit fail waive scheme, is within ECC coverage. The IOTS results shown in Fig. 23.4.6 show that the maximum fail symbol is within the 8 symbols that can be covered by ECC. Fig. 23.4.7 shows the full chip micrograph, the wire-bonding structure in ODP, and the IOTS board.

References:
[1] R. Oh, et al., "Design Technologies for a 1.2V 2.4Gb/s/pin High Capacity DDR4 SDRAM with TSVs," *VLSI*, 2014
[2] C. L. Chen, "Symbol error correcting codes for memory applications," *in Proc. Annu. Symp. Fault Tolerant Computing*, pp. 200–207, 1996
[3] S. Shim, et al., "A 16Gb 1.2V 3.2Gb/s/pin DDR4 SDRAM with Improved Power Distribution and Repair Strategy," *ISSCC*, pp. 212-214, 2018.
[4] K. Koo, et al., "A 1.2 V 38 nm 2.4 Gb/s/pin 2 Gb DDR4 SDRAM with bank group and half-page architecture," *ISSCC*, pp. 40–41, 2012.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

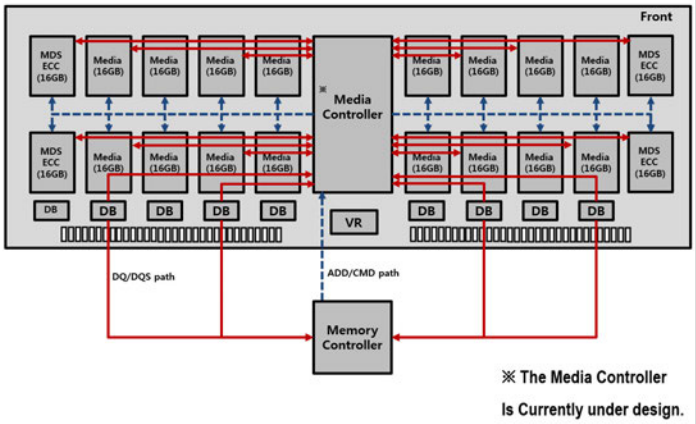


Figure 23.4.1: MDS DIMM Structure.

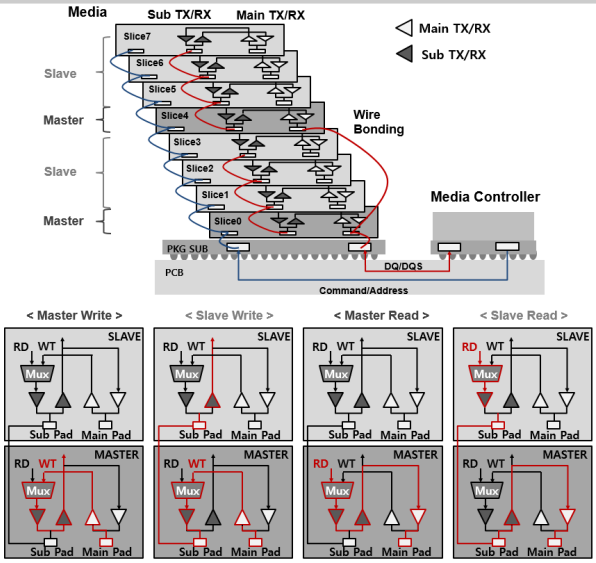
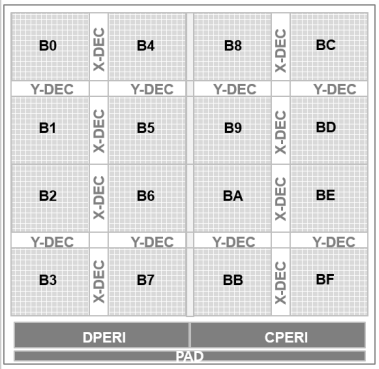


Figure 23.4.2: DQ Repeating Scheme at ODP.



	This work	[3]
Density	16Gb	16Gb
Process	18nm	18nm
Chip Size	64.93mm²	81.28mm²
Package Stack	Octal Die Package (Wire-Bonding)	Quad Die Package (TSV)
Data Rate	2133Mbps/pin	3200Mbps/pin
Power Supply (VDD/VDDQ/VPP)	1.1V/1.1V/2.5V	1.2V/1.2V/2.5V

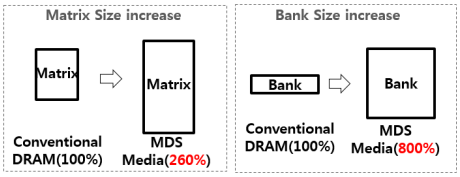


Figure 23.4.3: Media Architecture.

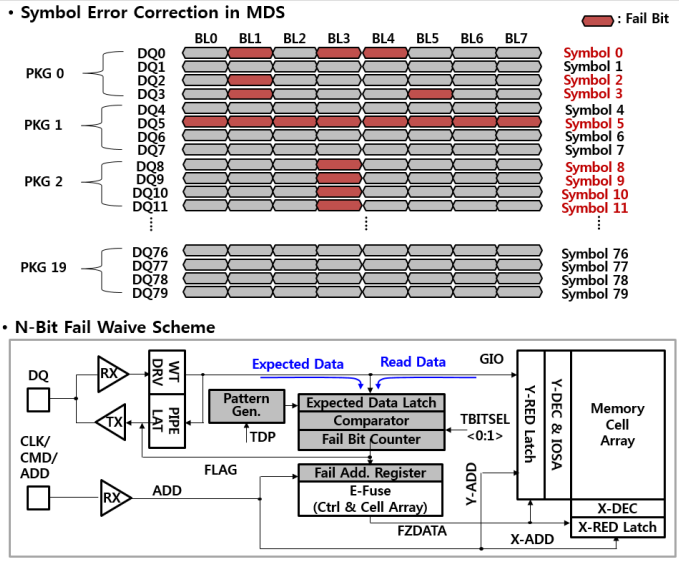


Figure 23.4.4: Symbol Error Correction in MDS and N-Bit fail Waive Scheme.

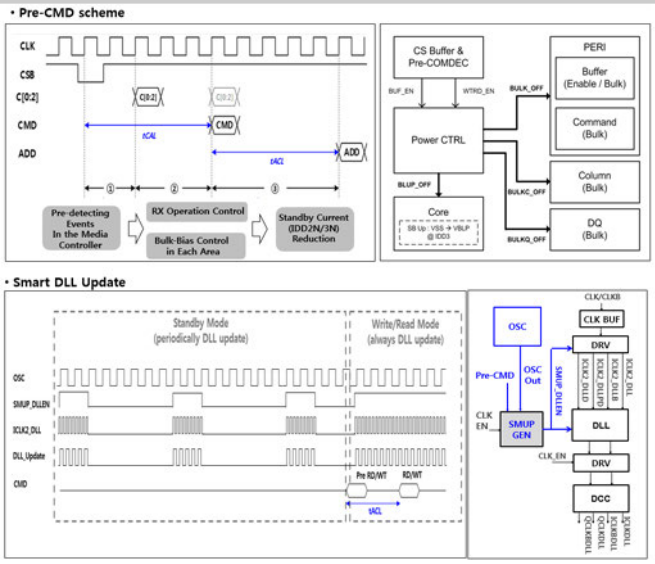
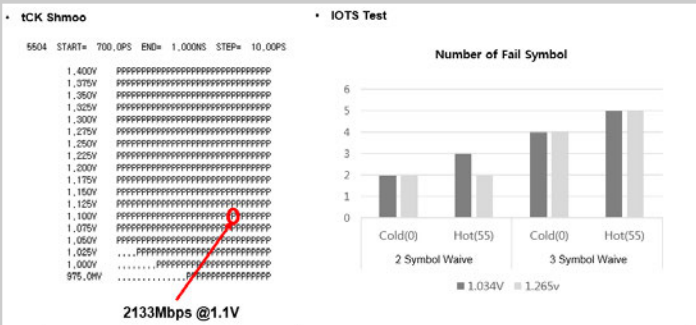


Figure 23.4.5: Power Reduction Scheme using Pre-CMD.



Density	512GB (DIMM)
I/O	X64 I/O (DIMM), X4 I/O (Package)
Size	101.7mm² (Package), 64.93mm² (Die)
Data Rate	2133Mbps/pin (Media), 3200Mbps/pin (Media-Controller)
CL/tRCD/tRP	30ns/30ns/25ns (Media)
Refresh rate	8K/512ms
MAX DIMM Power	12W (8.5W @Media, 3.5W @ Media-Controller and Data Buffer)

Figure 23.4.6: Measurement Results and MDS Summary Table.

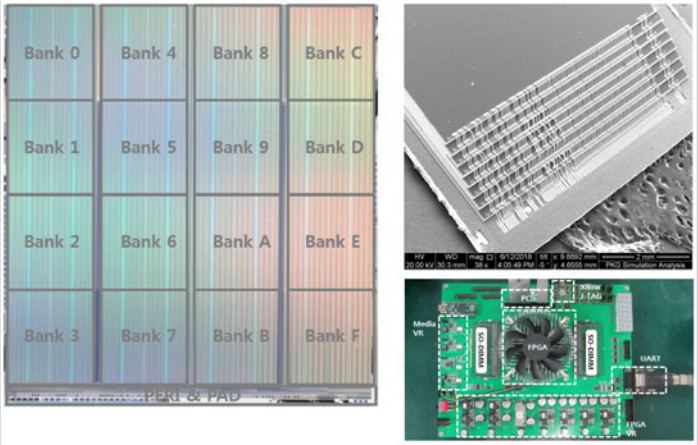


Figure 23.4.7: Micrograph of 16Gb Managed DRAM, the Wire-Bonding Structure in ODP and the IOTS Board.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 24 Overview: *SRAM and Computation-in-Memory*

MEMORY SUBCOMMITTEE

Session Chair: ***Shinichiro Shiratake***
Toshiba Memory, Yokohama, Japan

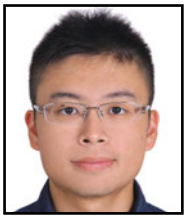
Associate Chair: ***Kyu-hyoun Kim***
IBM T.J. Watson, Yorktown Heights, NY

Subcommittee Chair: Jonathan Chang, TSMC, Hsinchu, Taiwan

Many state-of-the-art systems for machine learning are limited by memory in terms of the energy they require and the performance they can achieve. This session explores how this bottleneck can be overcome by emerging architectures that perform computation inside the memory array. National Tsing Hua University shows the multi-bit input/weight ReRAM as well as SRAM CIM macros. Southeast University shows the new computing architecture, which reduces the number accesses to the memory. On the other hand, SRAM continues to be the critical technology enabler for a wide range of applications from low-power to high-performance computing. This session showcases the leading-edge SRAM developments from the semiconductor industry. TSMC shows the 7nm dual-port SRAM architecture where read-disturb-write issues were mitigated. Samsung shows the dual-rail 7nm SRAM technology, which is tolerant to voltage and temperature variation.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

10:15 AM



24.1 A 1Mb Multibit ReRAM Computing-In-Memory Macro with 14.6ns Parallel MAC Computing Time for CNN-Based AI Edge Processors

C-X. Xue, National Tsing Hua University, Hsinchu, Taiwan

In Paper 24.1, National Tsing Hua University demonstrates the first ReRAM CIM macro capable of supporting CNN operations using multiple-bit input/weight and multiple-bit MAC output. This device achieved the fastest MAC operations among existing ReRAM-CIM macros (Tac = 14.6ns at 2b-input-3b-weight with 4b-MACout) and the best measured peak energy efficiency of 53.17 TOPS/W (binary mode).

10:45 AM

24.2 A 7nm 2.1GHz Dual-Port SRAM with WL-RC Optimization and Dummy-Read-Recovery Circuitry to Mitigate Read-Disturb-Write Issue

H. Fujiwara, TSMC, Hsinchu, Taiwan

In Paper 24.2, TSMC shows the 0.075um^2 dual-port SRAM bitcell in a 7nm technology node. It also mitigates the WL-R/C and cross-talk issue in the advanced technology node by process and layout co-optimization. They confirm that the dummy-read-recovery scheme can mitigate read-disturb-write issues in a dual-port SRAM bitcell. Silicon measurement results show Vmin <500mV at FS corner and 2.1GHz operation at SS and -40°C.

11:15 AM

24.3 A Voltage and Temperature Tracking SRAM Assist Supporting 740mV Dual-Rail Offset for Low-Power and High-Performance Applications in 7nm EUV FinFET Technology

I. Lee, Samsung Electronics, Hwaseong, Korea

In Paper 24.3, Samsung shows the dual-rail SRAM technology with voltage auto-tracking assist (VATA) and temperature auto-tracking assist (TATA) circuits. VATA and TATA have simpler architectures than others for avoiding additional power and area overhead. Specifically, VATA shows a new concept to cover the dual-rail offset voltage (740mV) to make dynamic and leakage power scaling possible.

11:30 AM

24.4 Sandwich-RAM: An Energy-Efficient In-Memory BWN Architecture with Pulse-Width Modulation

J. Yang, Southeast University, Nanjing, China

In Paper 24.4, Southeast University shows an energy-efficient in-memory binary-weight network (BWN) architecture with pulse-width modulation. It is the first in-memory BWN processor in which the feature is stored in memory. Its architecture is like a sandwich, memory-computing-memory. More precise analog computing based on pulse-width modulation is presented.

11:45 AM

24.5 A Twin-8T SRAM Computation-In-Memory Macro for Multiple-Bit CNN-Based Machine Learning

X. Si, National Tsing Hua University, Hsinchu, Taiwan
University of Electronic Science and Technology of China, Chengdu, China

In Paper 24.5, National Tsing Hua University demonstrates an SRAM-based CIM macro with twin 8T SRAM bitcells. A fabricated 55nm 3840b T8T SRAM-CIM macro achieved a better than 22× improvement in FoM (TOPS/W x GOPS/KB) and the fastest ever multiple-bit MAC operations (5ns for 2b-input, 5b-weight, 5b-MACout).

24.1 A 1Mb Multibit ReRAM Computing-In-Memory Macro with 14.6ns Parallel MAC Computing Time for CNN-Based AI Edge Processors

Cheng-Xin Xue, Wei-Hao Chen, Je-Syu Liu, Jia-Fang Li, Wei-Yu Lin, Wei-En Lin, Jing-Hong Wang, Wei-Chen Wei, Ting-Wei Chang, Tung-Cheng Chang, Tsung-Yuan Huang, Hui-Yao Kao, Shih-Ying Wei, Yen-Cheng Chiu, Chun-Ying Lee, Chung-Chuan Lo, Ya-Chin King, Chorng-Jung Lin, Ren-Shuo Liu, Chih-Cheng Hsieh, Kea-Tiong Tang, Meng-Fan Chang

National Tsing Hua University, Hsinchu, Taiwan

Embedded nonvolatile memory (NVM) and computing-in-memory (CIM) are significantly reducing the latency (t_{MAC}) and energy consumption (E_{MAC}) of multiply-and-accumulate (MAC) operations in artificial intelligence (AI) edge devices [1,2]. Previous ReRAM CIM macros demonstrated MAC operations for 1b-input, ternary-weighted, 3b-output CNNs [1] or 1b-input, 8b-weighted, 1b-output fully-connected networks with limited accuracy [2]. To support higher-accuracy convolution neural network heavy applications NVM-CIM should support multibit inputs/weights and multi-bit output (MAC-OUT) for CNN operations. One way to achieve multibit weights is to use a multi-level ReRAM cell to store the weight. However, as shown in Fig. 24.1.1, multibit ReRAM CIM faces several challenges. (1) a tradeoff between area and speed for multibit input/weight/MAC-OUT MAC operations; (2) sense amplifier's high input offset, large area, and high parasitic load on the read-path due to large BL currents (I_{BL}) from multibit MAC; (3) limited accuracy due to a small read/sensing margin (I_{SM}) across MAC-OUT or variation in cell resistance (particularly MLC cells). To overcome these challenges, this work proposes, (1) a serial-input non-weighted product (SINWP) structure to optimize the tradeoff between area, t_{MAC} and E_{MAC}; (2) a down-scaling weighted current translator (DSWCT) and positive-negative current-subtractor (PN-ISUB) for short delay, a small offset and a compact read-path area; and (3) a triple-margin small-offset current-mode sense amplifier (TMCSA) to tolerate a small I_{SM}. A fabricated 55nm 1Mb ReRAM-CIM macro is the first ReRAM CIM macro to support CNN operations using multibit input/weight MAC-OUT. This device achieves the shortest CIM-MAC-access time (t_{AC}) among existing ReRAM-CIMs (t_{MAC}=14.6ns with 2b-input, 3b-weight with 4b-MAC-OUT) and the best peak E_{MAC} of 53.17 TOPS/W (in binary mode).

Figure 24.1.2 presents the structure of the proposed ReRAM-CIM macro, using a 1T1R SLC ReRAM array, a current-aware BL clamper (CABLC), DSWCT, SINWP sampler-and-combiner (SINWP-SC), PN-ISUB, and TMCSA. This work places both positive and negative weights in the same array, but in different columns, unlike [1]. Each 3b-signed weight (W=1b-sign+2b-data) is stored in the same row of either the 2-column positive (PWG) or negative (NWG) group. In each PWG/NWG, the even BLs (BLM) represent the MSB and the odd BLs (BLL) represent the LSB. In other words, 4 SLC ReRAM cells are used to store 3b signed weights. To support configurable 1b/2b input (IN), this work translates 2b-input into two sequential single-bit (IN_0 and IN_1) WL pulses in one CIM clock cycle. For an n-by-n CNN kernel, n^2 weights are stored in n^2 consecutive rows. CABLC clamps the BL voltage (V_{BLC}) for current-mode sensing. A BL current (I_{BL}) that is equal to the sum of n^2 cell currents (I_{MC}) is generated via binary multiplication (IN×W) between a WL (IN) and a 1b-weight (W), as in [1]. The I_{BL} of BLM ($I_{BL\text{-}MSB}$) and BLL ($I_{BL\text{-}LSB}$) do not include their place-value processed in array (non-weighted I_{BL}). DSWCT and SINWP-SC then combine the $I_{BL\text{-}MSB}$ and $I_{BL\text{-}LSB}$ of a PWG/NWG to generate a weighted data-line (DL) current (I_{DL}) for the MAC value (MACV) using 2b-inputs and 2b-weights. PN-ISUB outputs the difference (I_{SUB}) between the I_{DL} of PWG ($I_{DL\text{-}P}$) and NWG ($I_{DL\text{-}N}$) as well as the sign bit ($DOUT_{SIGN}$=0 when $I_{DL\text{-}P} > I_{DL\text{-}N}$), where $I_{SUB}=|I_{DL\text{-}P} - I_{DL\text{-}N}|$. The current reduction provided by DSWCT and PN-ISUB allows for an I_{SUB} that is much smaller than the total current ($I_{DL\text{-}MSB} + I_{DL\text{-}LSB}$) of the accessed PWG and NWG. This allows TMCSA to use smaller transistors (achieving a lower parasitic load), which results in faster response times, smaller area and less power compared to schemes without DSWCT and PN-ISUB. Finally, each IO repeats the operation of TMCSA in 3 sequential phases to detect the same I_{SUB} with 3 different I_{REF} currents (I_{REF1} to I_{REF3}), and then outputs a 4b MACV with 3b data from TMCSA and a sign bit from PN-ISUB.

Figure 24.1.3 shows the operations of DSWCT, SINWP-SC, and PN-ISUB. DSWCT uses a current-mirror (CM) to translate $I_{DL\text{-}MSB}$ and $I_{DL\text{-}LSB}$ into lower weighted DL currents (I_{WDL}), $I_{WDL\text{-}MSB}= (1/p)\cdot 2\cdot I_{DL\text{-}MSB}$ and $I_{WDL\text{-}LSB}=(1/p)\cdot I_{DL\text{-}LSB}$. Choice of the reduced-amount ratio, p (p=4 for this work) is based on maintaining sufficient I_{SM}, while reducing power consumption and the transistor size (area and parasitic load) required for the remaining read-path. SINWP-SC operations occur in two phases. (1) For the 1st WL pulse (IN_0), INSW=1 and the gate-voltage ($V_{GCM\text{-}MSB}/V_{GCM\text{-}LSB}$) of CM_{MSB}/CM_{LSB} of DSWCT is sent to N3/N6 and stored in capacitor CM/CL to sample the $I_{WDL\text{-}MSB}/I_{WDL\text{-}LSB}$ resulting from IN_0×W. (2) For the 2nd WL pulse (IN_1), INSW=0 and $V_{GCM\text{-}MSB}/V_{GCM\text{-}LSB}$ is sent to N4/N7. The size of N4/N7 is 2× that of N3/N6. SINWP-SC then combines the drain currents of N3, N4, N6 and N7 to output a weighted I_{DL}, as follows:

$$I_{DL_LSB[0]}\cdot W_{DL\text{-}LSB[0]}/8+I_{DL_MSB[0]}\cdot W_{DL\text{-}MSB[1]}/4+I_{DL_LSB[1]}\cdot W_{DL\text{-}LSB[0]}/4+I_{DL_MSB[1]}\cdot W_{DL\text{-}MSB[1]}/2$$

In PN-ISUB, a comparator compares I_{DL_P} with I_{DL_N} to output the sign bit ($DOUT_{SIGN}$=0 when $I_{DL\text{-}P} > I_{DL\text{-}N}$), and $DOUT_{SIGN}$ enables PN-ISUB to connect the larger/lower current path of DL to the high/low-current input (I_{HC} / I_{LC}) terminal of a current-subtractor. If $DOUT_{SIGN}$ = 1, $I_{HC}= I_{DL\text{-}N}$, and $I_{LC}= I_{DL\text{-}P}$, then the current-subtractor generates the PN-ISUB output current I_{SUB} (=$I_{HC}- I_{LC}$, =$I_{DL\text{-}N}- I_{DL\text{-}P}$ if $DOUT_{SIGN}$ = 1). Compared to schemes that convert $I_{DL\text{-}P}$ / $I_{DL\text{-}N}$ to MACV separately before the digital combiner as in [1], PN-ISUB removes current leakage from HRS cells ($n^2\cdot I_{HRS}$) when $I_{DL\text{-}N} >> I_{DL\text{-}P}$, or enable larger current reduction (or small I_{SUB}) when $I_{DL\text{-}N} \sim= I_{DL\text{-}P}$ to improve sensing yield in the CSA.

Figure 24.1.4 shows the operation of TMCSA, which comprises two pairs of PMOS transistors (P1:P2 and P3:P4), four pairs of switches (SW1-SW4), two overdrive-coupling capacitors (C1, C2), four discharge NMOSs (DN1-DN4), an NMOS latch (N1-N3) and two current inputs (I_{IN} and I_{REF}). In stand-by mode, SW3 and SW4 are on, and SW1, SW2, and VDD_SA are off. DSD and CHD are high and DN1-DN4 are turned on to hold nodes DP1, DP2, LQ and LQB at 0V. In phase-1 (PH1, V_{TH} sampling), DSD, SW1 and SW2 are off and VDD_SA is on. The threshold voltages (V_{TH1} to V_{TH4}) of diode-connected P1 to P4 are stored on their gates (i.e. $V_{G1} = V_{DD} - V_{TH1}$, $V_{G3} = V_{DD} - V_{TH3}$). In phase-2 (PH2, V_{OV} sampling and coupling), SW4 is off, SW1 is on, and I_{IN} / I_{REF} flows through P1/P2. This results in $V_{G1} = V_{DD} - V_{TH1} - V_{OV\text{-}IN}$ and $V_{G2} = V_{DD} - V_{TH2} - V_{OV\text{-}REF}$, where $V_{OV\text{-}IN}$ / $V_{OV\text{-}REF}$ is the overdrive voltage of P1/P2 for sampling I_{IN} / I_{REF} as in [3]. In the meantime, V_{OV1} / V_{OV2} is coupled to V_{G4} / V_{G3} via C1/C2, so that $V_{G4} = V_{DD} - V_{TH4} - V_{OV1}$ and $V_{G3} = V_{DD} - V_{TH3} - V_{OV2}$. Ideally, $I_{P4} = 2\cdot I_{IN}$ and $I_{P3} = 2\cdot I_{REF}$ since transistors P3 and P4 are twice as big (2-finger style) as P1 and P2. In phase-3 (PH3, ΔI amplifying), SW2 is on and SW1 and SW3 are off. Thus, the current at node LQ is $I_{P3} - I_{IN} = 2I_{REF} - I_{IN}$. The current at node LQB is $I_{P4} - I_{REF} = 2I_{IN} - I_{REF}$. For a given period ($T_{PH3}$), the voltage difference ($V_{LQ\text{-}LQB}$) between node LQ and LQB is proportional to $[(2I_{REF} - I_{IN}) - (2I_{IN} - I_{REF})]\cdot T_{PH3} = 3\cdot(I_{IN} - I_{REF})\cdot T_{PH3}$, which is 3× the I_{SM} (=$I_{IN} - I_{REF}$) in conventional CSA. In phase-4 (PH4, latch), SAEN is high and N1-N3 are enabled to detect $V_{LQ\text{-}LQB}$ and generate a digital output at SAOUT.

Figure 24.1.5 shows the performance of the proposed schemes. For 3×3-CNN kernels DSWCT reduces the worst-case current by 3.6×. The combination of DSWCT and PN-ISUB schemes enables a 3.6–3.8× reduction in MACV-current. The SINWP scheme achieves an FoM (SM / (Energy × Area)) 1.4–6× better than those of sequential-input-parallel-weight (SIPW) and parallel-input-parallel-weight (PIPW) structures. TMCSA enabled 6× and 1.7× reductions in input offset, compared to conventional CSA and DR-CSA [1], using a 70μA input current.

Figure 24.1.6 presents the measurement results from a 1Mb ReRAM-CIM macro fabricated using 1T1R SLC ReRAM in a 55nm CMOS process. A demo system was built using our ReRAM-CIM with an FPGA host. For CNN operations using 3×3 kernels and 2b-input and 3b-weight, the captured waveforms confirm that t_{MAC} for a 4b MAC output (1b sign and 3b data) is 14.6ns, excluding the path-delay. Using 3×3-CNN kernels with 1b-input and 3b-weight, shmoo results confirm a t_{MAC} of 11.75ns per cycle at typical V_{DD}, and the system test achieve an 88.52% inference accuracy on the CIFAR-10 dataset. The measured peak energy-efficiency is 53.17TOPS/W in binary mode (1b-input, 3b-weight, 4b-MAC-OUT) and 21.9TOPS/W in multibit mode (2b-IN, 3b-weight, 4b-MAC-OUT) using CIM peripheral circuits and a reference generator. This work also achieves a 3.4× improvement in energy-efficiency and a 1.3× faster t_{MAC}, compared to [1]. Figure 24.1.7 presents the die micrograph and chip summary.

Acknowledgements:
The authors would like to thank MOST-Taiwan, CIC, NTHU-TSMC JDP for financial and manufacturing support.

References:
[1] W.-H. Chen, et al., "A 65nm 1Mb Nonvolatile Computing-in-Memory ReRAM Macro with sub-16ns Multiply-and-Accumulate for Binary DNN AI Edge Processors," *ISSCC*, pp. 494-495, 2018.
[2] R. Mochida, et al., "A 4M Synapses integrated Analog ReRAM based 66.5 TOPS/W Neural-Network Processor with Cell Current Controlled Writing and Flexible Network Architecture," *VLSI*, pp. 175-176, 2018.
[3] M.-F. Chang, et al., "An offset tolerant current-sampling-based sense amplifier for sub-100nA-cell-current nonvolatile memory," *ISSCC*, pp. 206-207, 2011.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

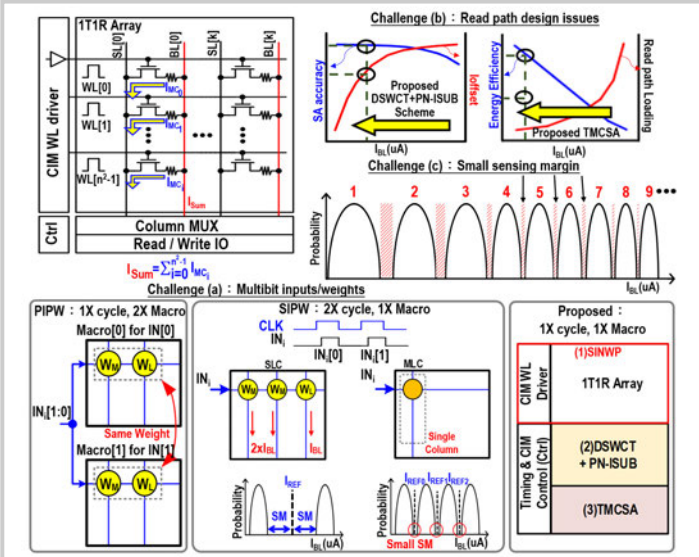
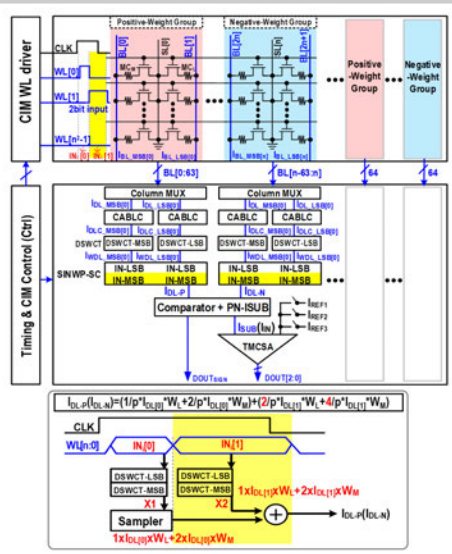


Figure 24.1.1: Multi-bit computation in nonvolatile memory.



Positive Weight Group

2 bit input /WL (IN) MSB	LSB	MC_W (2xI_{MC})	MC_L (1xI_{MC})	Weight Value (W)	Product (INxW)
0	0	LRS	LRS	+3	0
		LRS	HRS	+2	0
		HRS	LRS	+1	0
		HRS	HRS	0	0
0	1	LRS	LRS	+3	+3
		LRS	HRS	+2	+2
		HRS	LRS	+1	+1
		HRS	HRS	0	0
1	0	LRS	LRS	+3	+6
		LRS	HRS	+2	+4
		HRS	LRS	+1	+2
		HRS	HRS	0	0
1	1	LRS	LRS	+3	+9
		LRS	HRS	+2	+6
		HRS	LRS	+1	+3
		HRS	HRS	0	0

Negative Weight Group

2 bit input /WL (IN) MSB	LSB	MC_W (2xI_{MC})	MC_L (1xI_{MC})	Weight Value (W)	Product (INxW)
0	0	HRS	HRS	0	0
		HRS	LRS	-1	0
		LRS	HRS	-2	0
		LRS	LRS	-3	0
0	1	HRS	HRS	0	0
		HRS	LRS	-1	-1
		LRS	HRS	-2	-2
		LRS	LRS	-3	-3
1	0	HRS	HRS	0	0
		HRS	LRS	-1	-2
		LRS	HRS	-2	-4
		LRS	LRS	-3	-6
1	1	HRS	HRS	0	0
		HRS	LRS	-1	-3
		LRS	HRS	-2	-6
		LRS	LRS	-3	-9

Figure 24.1.2: Proposed ReRAM-CIM macro.

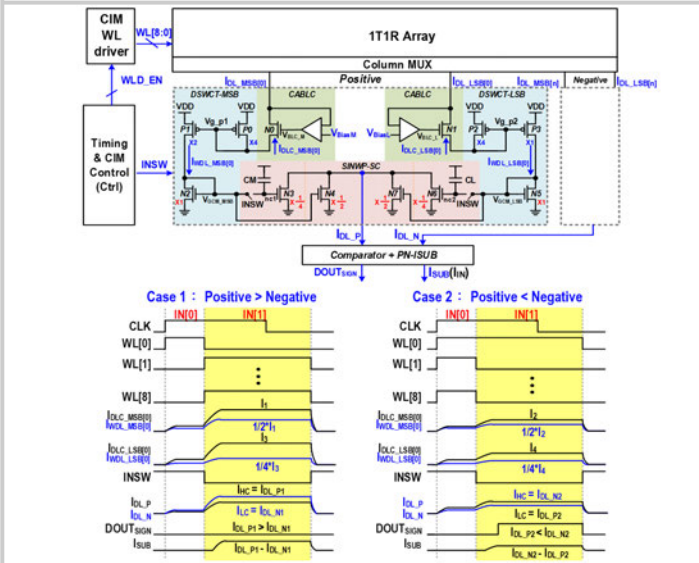


Figure 24.1.3: Sequential input parallel weight structure.

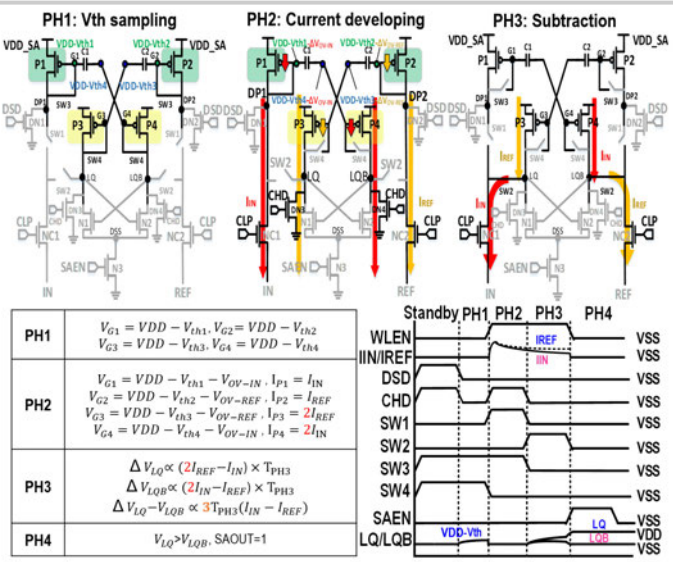


Figure 24.1.4: Structure and operation of TMCSA.

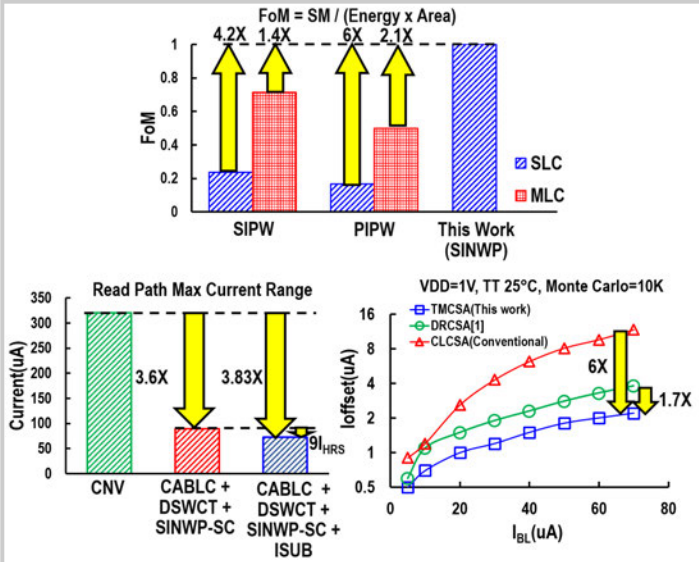


Figure 24.1.5: Performance comparison of this work vs prior-art.

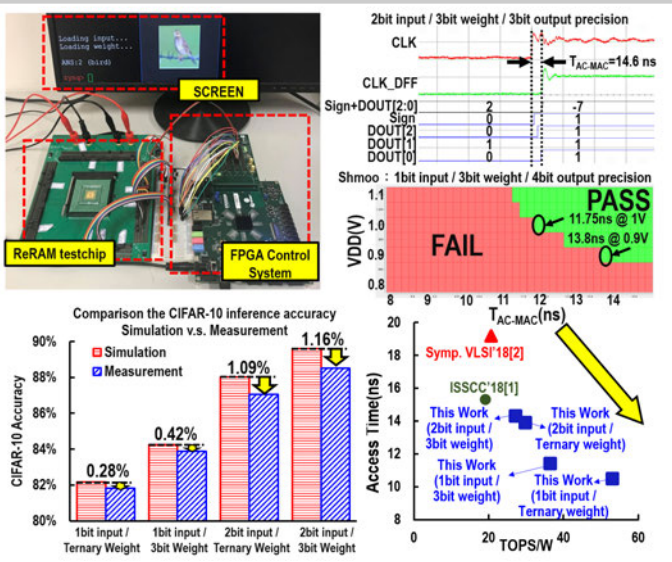
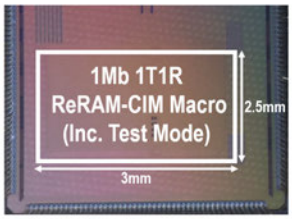
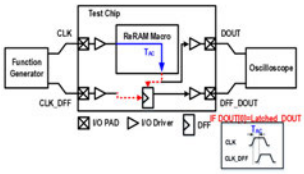


Figure 24.1.6: Measurement results.





Technology	55nm CMOS Logic Process
ReRAM	Logic-process ReRAM
Cell Size(1T1R)	0.2025um^2
ReRAM Mode	Memory/CIM
CIM Mode	CNN
Capacity	1Mb (8 Sub-array)
Sub-array	256rows x 512columns
Read Delay @ VDD =1V	Memory Mode: 3.16ns
	CNN Mode: 11.75ns (1bit input / 3bit weight) (3bit DOUT precision)
	CNN Mode: 14.6ns (2bit input / 3bit weight) (3bit DOUT precision)
Energy Efficiency @ VDD =1V	CNN Mode: 53.17 (TOPS/W) (1bit input / 3bit weight) (3bit DOUT precision)
	CNN Mode: 21.9 (TOPS/W) (2bit input / 3bit weight) (3bit DOUT precision)

Figure 24.1.7: Die micrography and summary Table.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

24.2 A 7nm 2.1GHz Dual-Port SRAM with WL-RC Optimization and Dummy-Read-Recovery Circuitry to Mitigate Read-Disturb-Write Issue

Hidehiro Fujiwara, Chih-Yu Lin, Hsien-Yu Pan, Cheng-Han Lin, Po-Yi Huang, Kao-Cheng Lin, Jhon-Jhy Liaw, Yen-Huei Chen, Hung-Jen Liao, Jonathan Chang

TSMC, Hsinchu, Taiwan

Continued transistor scaling increases random V_t variation, and wire routing resistance and capacitance; it degrades SRAM performance and results in SRAM design difficulties. Although dual-port (DP) SRAM is useful, because it can offer simultaneous read and write operations with two asynchronous clocks, circuit design is more complex than for a single-port (SP) SRAM when considering worst-case operation. In particular, WL resistance and capacitance (RC) degradation caused by process scaling and read-disturb-write (RDW) problem in a DP-SRAM bitcell are major challenges for DP-SRAM design in advanced technology nodes.

In this paper, we report a 16Kb DP SRAM in a 7nm process technology with a 0.075um^2 DP-SRAM bitcell. The DP bitcell improves wordline (WL)-RC and crosstalk between two WLs by process and layout co-optimization. The 16Kb DP-SRAM includes a sequential power wakeup scheme, which optimizes peak current (I_{PEAK}) and wakeup time. We also demonstrate that our dummy-read-recovery (DRR) scheme dramatically mitigates RDW issues. Measurement results show that the DRR scheme can achieve over 150mV improvement in V_{MIN}; V_{MIN} becomes lower than 500mV for 4Mb at 90% yield at –40°C. We also observed 1.3, 1.6 and 2.1GHz operation at 0.7, 0.8 and 1.0V, respectively.

Figure 24.2.1 shows the scaling trends of SP- and DP-SRAM bitcells [1-3]. A 0.027um^2 bitcell in a 7nm process and a 0.031um^2 bitcell in a 10nm process for SP high-density (HD) SRAM bitcell have been already reported [1,2]. In this paper, we use a 0.075um^2 DP-SRAM bitcell for a 16Kb DP-SRAM macro. Although a DP-SRAM bitcell is roughly 2.5-3× larger than a SP-HD bitcell, our DP-SRAM bicell is on the same scaling trend as SP-HD SRAM bitcells. The top side graph in Fig. 24.2.2 shows the WL-R trend for SP-SRAM bitcells [1]. WL-RC degradation caused by process scaling is one of the major challenges for advanced SRAM design [4]. WL-RC optimization for SP-SRAM macro designs has also been reported [3]. However, WL-RC in DP-SRAM is more serious because we have to draw two WLs within a bitcell and the cell size of DP SRAM bitcell is larger than that of a SP-SRAM bitcell. Hence, we also improved WL-RC in DP SRAM bitcell by using multiple-layer WLs along with process and layout co-optimization. This WL-RC optimization does not result in any additional penalties. As shown by the waveforms in the bottom left side, WL-delay from the near-end to the far-end can be improved from 209ps to 108ps with WL optimization. Furthermore, we also mitigate a crosstalk between AWL and BWL in a DP-SRAM bitcell. As shown by the waveforms in the bottom right side, both the magnitude and the duration of crosstalk noise can be suppressed. This WL-RC optimization can also reduce the design effort for crosstalk mitigation for DP-SRAM.

Our test chip implements a 16Kb DP-SRAM macro. The macro configuration is 512×32b with 4-way column multiplexing. We used the negative bitline scheme for write assist [1]. The 16Kb DP-SRAM macro includes power management circuitry, which can reduce leakage power in sleep mode. In order to achieve both fast wake up (~one cycle) and acceptable I_{PEAK} (wakeup I_{PEAK} < mission-mode I_{PEAK}), we introduced a sequential wakeup circuitry in the SRAM macro as shown in Fig. 24.2.3. We fine-tuned the wakeup sequence using logic and wire delays in order to minimize I_{PEAK}. If there is no circuit optimization, huge inrush current happens during the initial stage of wakeup from sleep mode because of a large V_{DS} bias to header PMOS devices. In order to suppress the inrush current for the initial stage, only header PMOS device in the control block (CTRL) are turned on first. After that, the rest of the header PMOS devices in the IO and WL driver (WLDRV) blocks are sequentially enabled. Before entering mission mode, all header PMOS devices enter the on-state. Current waveforms show uniform current behavior. We also confirmed I_{PEAK} in wakeup is smaller than I_{PEAK} in mission mode. Chip measurement results show that wakeup time is smaller than cycle time.

In addition to the WL-RC problem, RDW is also a key challenge in DP-SRAM design. Figure 24.2.4 summarizes RDW issues for a conventional PMOS clamp scheme [5]. RDW is worse when two WLs are activated for the same row from the two ports with a positive clock skew in the aggressor port. There are three phases of the RDW access in the conventional circuitry. (1) The first phase is a pure write operation, which occurs only from the victim port: it behaves the same as a single port write operation. (2) In the second phase, the write operation is disturbed by the aggressor port because both BL and BLB in the aggressor port are clamped to "H", and disturb current flows into the "L" write side. (3) The third phase is a single-port Static-noise-margin (SNM) mode, because WL in the aggressor port still keeps "H" even after write from the victim port is finished. Storage data in a memory cell after write operation must be kept in this phase. However, if the "H" storage node does not have a high enough voltage level in the second phase, SNM failure will happen in the third phase.

Figure 24.2.5 shows the DRR scheme [6]. In order to support a bit-write-mask function, we improved the logic in the DRR circuitry. Unlike the conventional circuitry, the "H"-write side BL in the aggressor port is charged to "H" through a PMOS, while the other side BL is kept floating in the RDW condition. The first phase of RDW is the same as for the conventional circuitry. In the second phase, disturb current flows only in the early stage. This disturb current is decreased as "L"-write side BL in the aggressor goes to "L". One of the advantages of the DRR scheme is faster voltage recovery in "H" storage node because the disturb current becomes weaker and disappears during write operation. Another advantage of the DRR scheme exists in the third phase. Because "L"-write side BL in the aggressor port has been discharged, there is no single port SNM mode in the third phase even if WL in the aggressor port is still kept at "H".

Although a BL-equalizing scheme between the A and B ports has been proposed, it requires address comparators that may impact setup time for address inputs [7]. In comparison, the DRR scheme does not require an address comparator and the area overhead in IO circuitry is comparable with [7].

A micrograph of the test chip fabricated in a 7nm high-K metal-gate (HKMG) FinFET process technology is shown in Fig. 24.2.7. We implemented 1587 16Kb DP-SRAM macros on a test chip; the total bit-capacity is 24.8Mb. Each 16Kb DP-SRAM macro includes both a conventional clamp and the DRR scheme for side-by-side performance comparisons. Figure 24.2.6 shows the measurement results from the test chip. V_{MIN} measurement results, which include worst-case clock skew, confirm that the DRR scheme improves V_{MIN} by more than 150mV. V_{MIN} with a 90% yield of 4Mb at -40°C is lower than 500mV. Furthermore, the two-port access V_{MIN} with the DRR scheme shows almost the same level as for a single-port access SRAM. As shown by the frequency vs. voltage shmoo plot, for a SS-corner wafer at -40°C, we realize 1.3, 1.6 and 2.1GHz operation at 0.7, 0.8 and 1.0V, respectively.

Acknowledgements:
The authors would like to thank the physical design team for the chip implementation, the RD team for wafer manufacturing, and the test department for help with measurements.

References:
[1] J. Chang, et al., "A 7nm 256Mb SRAM in high-k metal-gate FinFET technology with write-assist circuitry for low-Vmin applications," *ISSCC*, pp. 206-207, 2017.
[2] Z. Guo, et al., "A 23.6Mb/mm^2 SRAM in 10nm FinFET technology with pulsed PMOS TVC and stepped-WL for low-voltage applications," *ISSCC*, pp. 196-198, 2018.
[3] T. Song, et al., "A 10nm FinFET 128Mb SRAM with assist adjustment system for power, performance, and area optimization," *ISSCC*, pp. 306-307, 2016.
[4] 2007 International Technology Roadmap for Semiconductors - Interconnection.
[5] Y. Ishii, et al, "A 28 nm Dual-Port SRAM Macro With Screening Circuitry Against Write-Read Disturb Failure Issues," *JSSC*, pp. 2535-2544, vol. 46, no. 11, Nov. 2011.
[6] Y-H. Chen, et al., "A 16nm dual-port SRAM with partial suppressed word-line, dummy read recovery and negative bit-line circuitries for low VMIN applications," *VLSI*, 2016.
[7] Y. Ishii, et al, "A 28-nm dual-port SRAM macro with active bitline equalizing circuitry against write disturb issue," *VLSI*, pp. 99-100, 2010.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

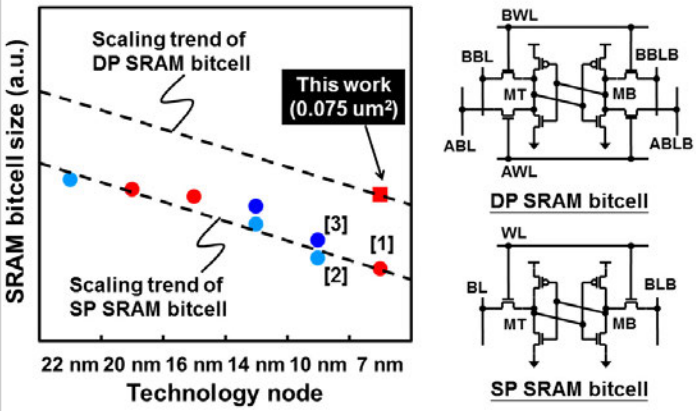


Figure 24.2.1: SRAM bitcell area scaling trends.

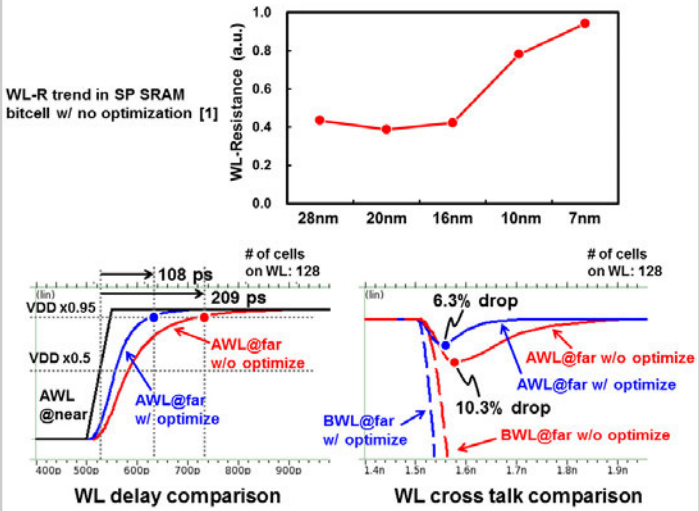


Figure 24.2.2: WL-RC improvement for a DP-SRAM bitcell.

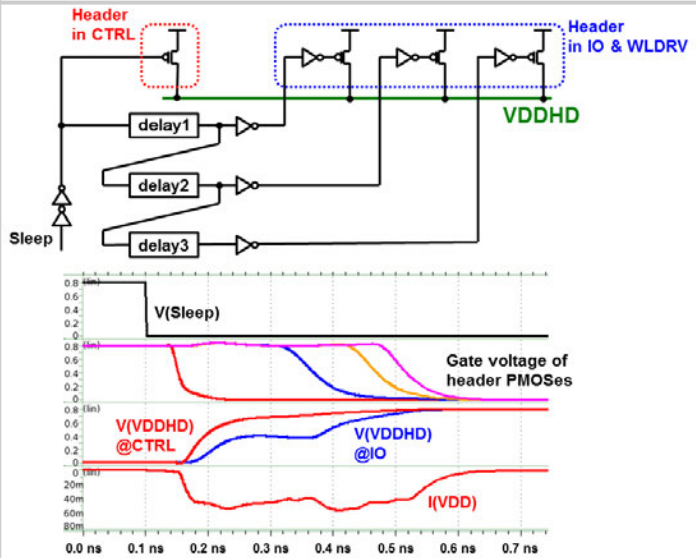


Figure 24.2.3: Sequential wakeup circuitry to reduce peak current.

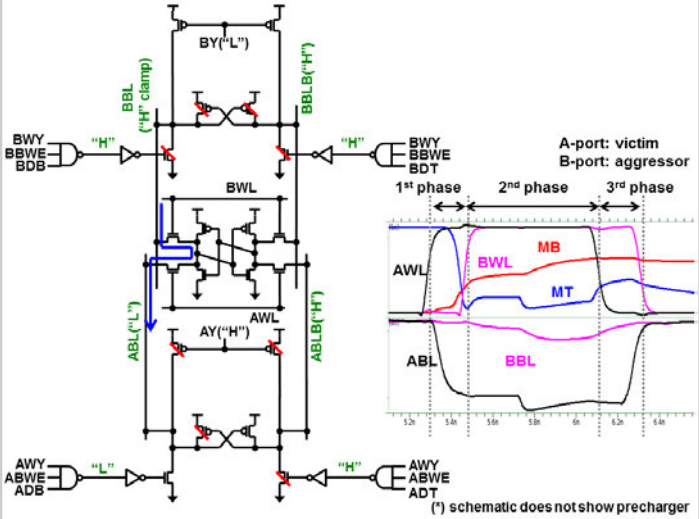


Figure 24.2.4: Conventional clamp scheme with PMOS in DP SRAM.

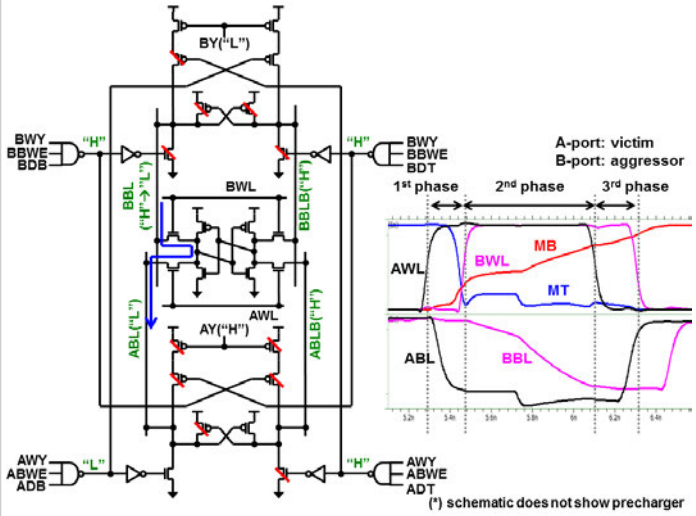


Figure 24.2.5: Dummy read recovery scheme to mitigate read-disturb-write.

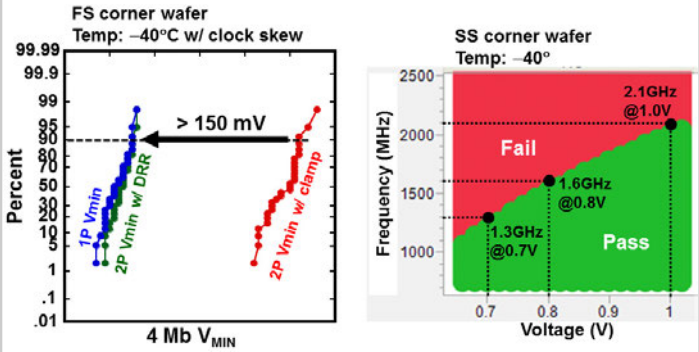
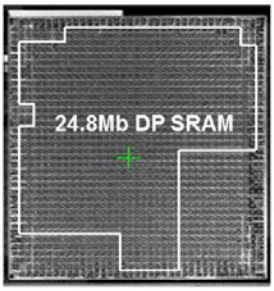


Figure 24.2.6: Chip measurement results.

24



Technology	7 nm HKMG FinFET CMOS
Chip size	4.91 mm x 4.99 mm
Bitcell	0.075 um^2 DP bitcell
Macro configuration	16 Kb (512 words x 32 IOs w/ MUX4)
Bit capacity	24.8 Mb (16 Kb x 1587 macros)

Figure 24.2.7: Test chip micrograph and table summarizing key metrics.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

24.3 A Voltage and Temperature Tracking SRAM Assist Supporting 740mV Dual-Rail Offset for Low-Power and High-Performance Applications in 7nm EUV FinFET Technology

Inhak Lee, Hanwool Jeong, Sangyeop Baeck, Siddharth Gupta, Changnam Park, Dongwook Seo, Jaesung Choi, Jaeyoung Kim, Hoon Kim, Jungmyung Kang, Sunyung Jang, Daeyoung Moon, Sangshin Han, Taehyung Kim, Jaehyun Lim, Younghwan Park, Hyejin Hwang, Jeonseung Kang, Jaeseung Choi, Taejoong Song

Samsung Electronics, Hwaseong, Korea

Embedded SRAM is a key component that typically determines the overall processor performance. A fully customized SRAM requires a long design cycle; hence, an SRAM compiler that can provide a wide range of user-specified SRAM designs, by utilizing predefined leaf cells, is widely used. An SRAM compiler must exhibit robustness across a wide range of voltage and temperature conditions. On top of optimizing power and performance for a variety of applications, a compiler also enables dynamic voltage and frequency scaling. However, the temperature and voltage sensitivity of an SRAM bitcell limits its voltage and temperature operating range, thus significantly degrading read and write stability. Several approaches to improve SRAM stability have been proposed [1-6]; however, the WL under-drive (WLUD) read-assist technique is the most widely used: owing to its simple architecture, and smaller power and area overhead [1-2]. However, this approach reduces the WL voltage at voltage and temperature conditions that do not require WLUD; thus, degrading read speed and the write margin (WRM).

Another approach uses a dual power-rail to enhance bitcell stability [3]. It applies a dedicated supply voltage (V_{DDC}) to the SRAM bitcell array and WL, while the peripheral circuits and the rest of the SoC are powered from a different supply voltage (V_{DDP}). Most dual-rail techniques use BL pre-charge circuits that are powered off V_{DDP}, since BL pre-charge energy is a large portion of the total dynamic power [4]. However, the core problem with this technique arises when V_{DDP} is reduced, thus increasing the dual-rail offset voltage ($V_{DDC} - V_{DDP}$). A large dual-rail offset voltage imbalances the high storage node of an SRAM bitcell during a read or a half-selected operation as shown in Fig. 24.3.1. To resolve this problem, a V_{DDP} tracking voltage regulator is used [5] that generates the optimal V_{DDC} based on the current V_{DDP}. However, this circuit is composed of a number of analog circuits: including a start-up circuit, a band-gap reference, an operation amplifier powered by a high supply voltage. These incur a substantial area and power overhead. Another issue is the limited temperature range, which is tackled by temperature tracking assist circuits such as the technique proposed in [6]. In which, a temperature sensor is utilized to generate the optimal WL level and prevent an excessive WLUD at low temperatures, as shown in Fig. 24.3.1. In doing so, the write margin and read speed can be enhanced. However, temperature sensing and voltage generation require numerous analog blocks powered off of a high-supply voltage, again drastically increasing area and power consumption.

To alleviate the aforementioned issues and to increase overall performance, this paper proposes a voltage auto-tracking assist circuit (VATA) and a temperature auto-tracking assist circuit (TATA). These circuits provide two main advantages. (1) VATA and TATA feature simple architectures, thereby limiting the additional power and area overhead required. (2) VATA and TATA under drive WL only when necessary; VATA is enabled only at a large dual-rail offset voltage condition and TATA is enabled only at high temperature conditions. Figure 24.3.2 shows a simplified circuit with VATA and TATA in a dual-rail SRAM operated from V_{DDP} and V_{DDC}.

VATA consists of a PMOS header switch P1, controlled by HDR_CTRL, that is always-on during read and write operations, and a pull-down PMOS P2, controlled by HVDDP, that retains an always-high state and is driven from V_{DDP}. Specifically, the size of P2 is comparable to P1 to sufficiently lower V_{DDWL}. In addition, the pull-down NMOS N1, controlled by VATA_EN, only turns on during a high WL condition. With this structure, P2 turns on when $V_{DDP} < V_{DDC}$ and turns off when $V_{DDP} \geq V_{DDC}$. Specifically, the larger the dual-rail offset voltage ($V_{DDC} - V_{DDP}$), smaller the V_{DDWL}; V_{DDWL} decreases with an increasing probability of the disturbance of half-selected and read-selected bitcells. This implies that V_{DDWL} self-adapts to track the dual-rail offset voltage condition. Figure 24.3.3 compares the access-disturb margin (ADM) [1] for SRAM with VATA, older architectures [5], and for SRAM without assist. A high V_{DDC} and low V_{DDWL} level maximizes the ADM, and unlike earlier architectures [5] VATA only reduces V_{DDWL} when V_{DDC} is unchanged. As a result, ADM levels are higher for VATA, compared to older architectures [5]. Moreover, VATA does not require an additional power regulator as in [5], thus minimizing area and power overhead.

Figure 24.3.2 illustrates the structure of TATA: the pull-down TATA_EN signal gates NMOS N2; hence N2 turns on only when WL is high. The diode connected N3, which is comparable in size to P3, makes the pull-down current stronger at higher temperatures. As a result, V_{DDWL} becomes lower at higher temperatures, where the bitcell disturbance degrades. Figure 24.3.3 compares the WLUD level of a conventional WLUD circuit [2] and TATA across -40 – 150°C. On the plot, *required WLUD* indicates the voltage level required to meet a target yield. Unlike a conventional WLUD circuit, which outputs the same WLUD level regardless of temperature, the TATA WLUD level is reduced at lower temperatures; becoming close to the required WLUD level, thus efficiently preventing performance degradation at lower temperatures.

Figure 24.3.4 illustrates the advantages of using VATA and TATA in terms of read speed and the WRM. When there is no dual-rail offset, a conventional WLUD assist lowers the WL, whereas there is no WL under drive with VATA. This is because the pull-down transistors in a conventional WLUD are large enough to generate WL level that satisfy the target ADM for case of large dual rail offset. As a result, VATA enhances read speed by 30%. WRM is compared based on the various offset values in the structure with and without VATA. When the dual-rail offset is small, WRM is significantly improved using VATA because of an optimized WLUD level. As a result, the write V_{MIN} is improved by 280mV at the zero dual-rail offset condition. TATA provides a similar benefit as well. At the worse-speed temperature corner TATA adaptively reduces WLUD, whereas the WL level is almost the same as for higher temperatures for a conventional WLUD assist circuit. Consequently, TATA improves the read speed by 15%, and V_{MIN} is improved by 100mV as well because the write V_{MIN} is determined at low temperatures.

VATA and TATA are evaluated in a fabricated test chip using the 7nm extreme-ultraviolet (EUV) FinFET technology. Figure 24.3.5 shows the measured V_{MIN} for a high-density (HD) SRAM macro with and without TATA. An additional V_{MIN} improvement of 60mV is achieved, compared to the SRAM without TATA. In Fig. 24.3.6, a shmoo plot compares a high-current (HC) SRAM macro with and without VATA. With VATA, the dual-rail offset can be up to 740mV; a 300mV improvement compared to the SRAM that does not use VATA. In addition, the minimum V_{DDP} can be lowered by 60mV at a given V_{DDC}. A die micrograph of the 7nm FinFET test chip with 1kb×80 6T-HD and HC dual-rail SRAM macros is shown in Fig. 24.3.7.

References:
[1] T. Song, et al., “A 7nm FinFET SRAM using EUV lithography with dual write-driver-assist circuitry for low-voltage applications,” *ISSCC*, pp. 198-200, 2018.
[2] T. Song, et al., “A 10 nm FinFET 128 Mb SRAM with assist adjustment system for power, performance, and area optimization,” *ISSCC* , pp. 208-209, 2016.
[3] M. Clinton, et al., "A 5GHz 7nm L1 cache memory compiler for high-speed computing and mobile applications," *ISSCC*, pp. 200-201, 2018.
[4] M. Clinton, et al., "A low-power and high-performance 10nm SRAM architecture for mobile applications," *ISSCC*, pp. 210-211, 2017.
[5] Y.H. Chen, et al., "A 0.6 V dual-rail compiler SRAM design on 45 nm CMOS technology with adaptive SRAM power for lower VDD_min VLSIs," *JSSC*, vol. 44, no. 4, pp. 1209-1215, Apr. 2009.
[6] H. Nho, et al., "A 32nm high-k metal gate SRAM with adaptive dynamic stability enhancement for low-voltage operation," *ISSCC*, pp. 346-347, 2010.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

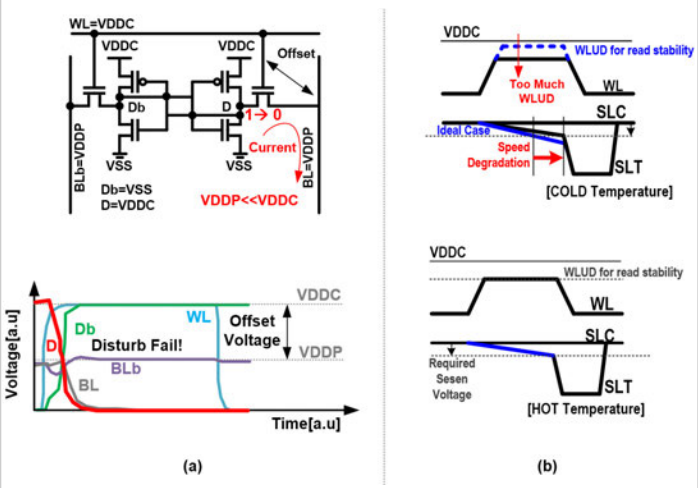


Figure 24.3.1: (a) Disturb margin issue at a high-voltage offset, (b) Speed penalty of conventional read assist at lower temperatures.

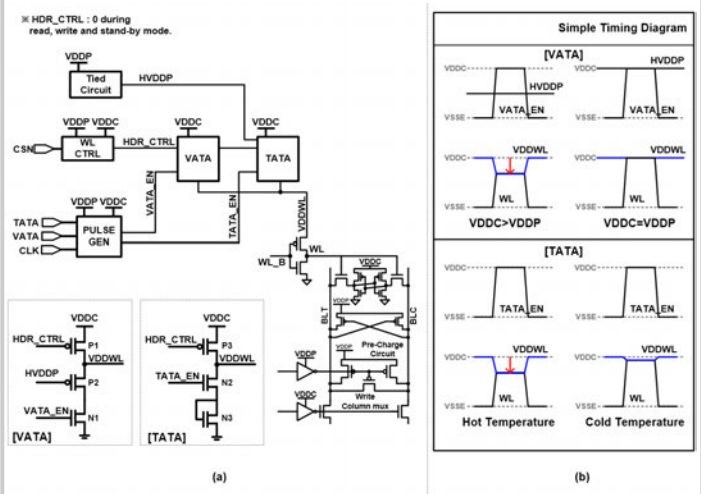


Figure 24.3.2: (a) TATA and VATA SRAM implementation, (b) Simple timing diagram with different offset voltage and temperature.

Offset Voltage↑ (VDDC-VDDP)	VDDC Level	VDDWL Level
Ideal	No Change	▼
Prior art [5]	▼	▼
This Work	No Change	▼

	Circuit Complexity	Area Overhead
Prior art [5]	High	Large
This Work	Low	Small

VDDWL Level	Hot Temperature	Cold Temperature
Ideal	▼	No Change
Prior art [2]	▼	▼
This Work	▼	No Change

	Circuit Complexity	Area Overhead
Prior art [6]	High	Large
This Work	Low	Small

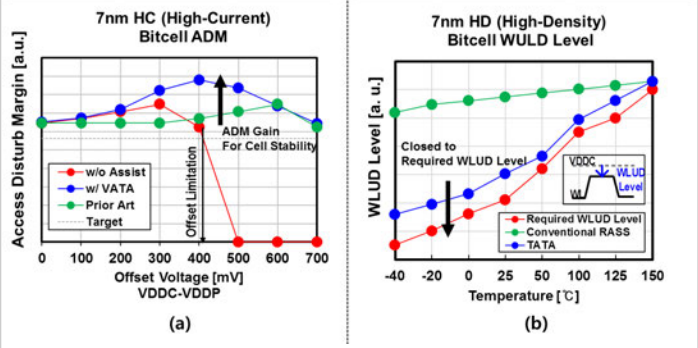


Figure 24.3.3: (a) Comparison of ADM values between VATA and prior art, (b) WLUD Level between TATA and prior art.

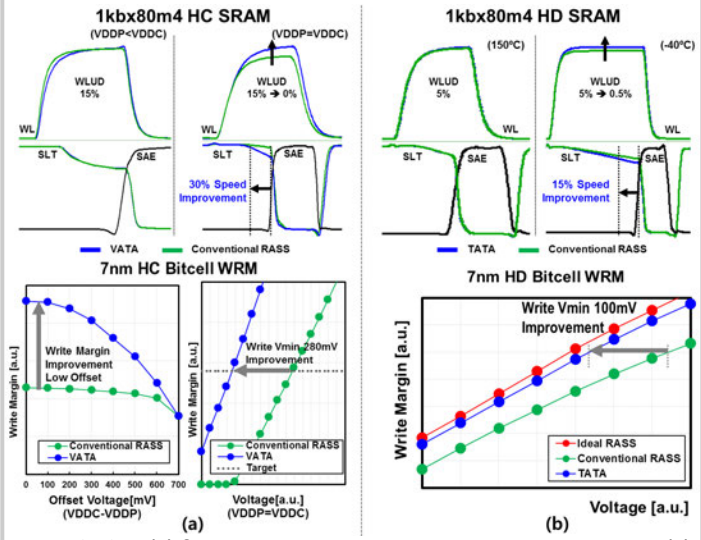


Figure 24.3.4: (a) Speed and write margin improvement by using VATA,(b) TATA.

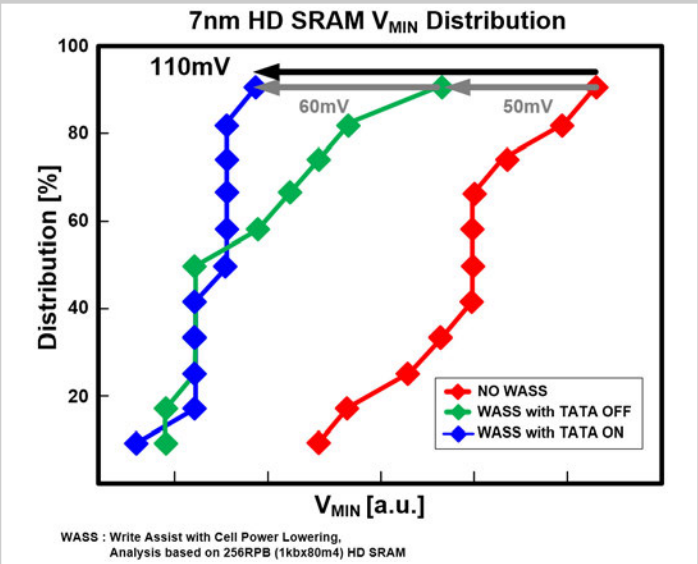


Figure 24.3.5: Silicon results of TATA HD SRAM V_{MIN}.

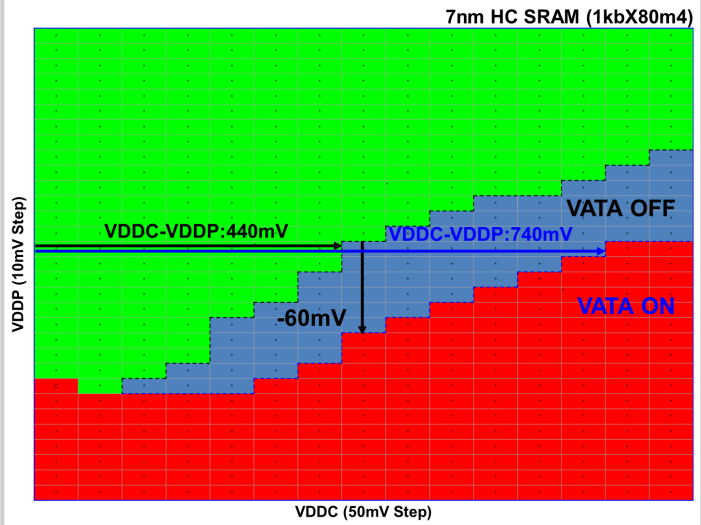


Figure 24.3.6: Silicon test data of HC SRAM with VATA - Voltage shmoo.

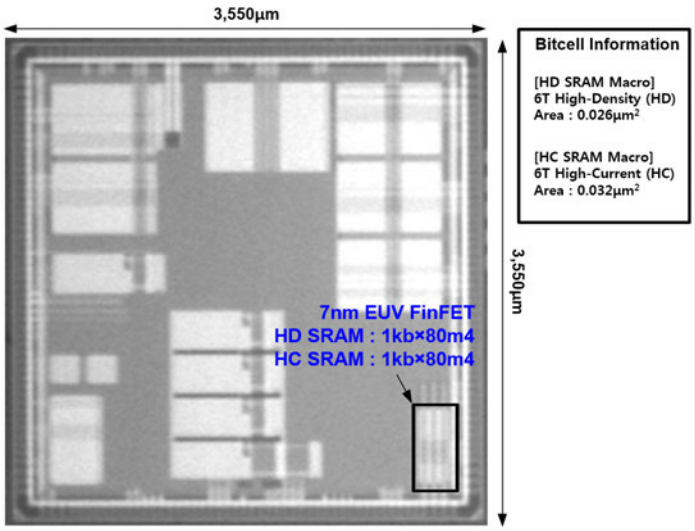


Figure 24.3.7: Die micrograph of 1kb×80 HD SRAM and HC SRAM.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

24.4 Sandwich-RAM: An Energy-Efficient In-Memory BWN Architecture with Pulse-Width Modulation

Jun Yang[1], Yuyao Kong[1], Zhen Wang[2], Yan Liu[1], Bo Wang[1], Shouyi Yin[3], Longxin Shi[1]

[1]Southeast University, Nanjing, China
[2]Boxing Electronics, Nanjing, China
[3]Tsinghua University, Beijing, China

Convolutional neural networks (CNN) achieve state-of-the-art results in the field of visual perception, drastically changing the traditional computer-vision framework. However, the movement of massive amounts of data prevents CNN's from being integrated into low-power IoT devices. The recently proposed binary-weight network (BWN) reduces the complexity of computation and amount of memory access. A conventional digital implementation, which is composed of separate feature/weight memories and a multiply-and-accumulate (MAC) unit, requires large amounts of data to be moved [3]. To reduce power the weight memory and the computations are integrated together, into an in-memory computation architecture [1,2,5]. However, feature data is still stored externally, so data movement has only been partially addressed, especially for BWN. This paper blends feature and partial-weight memory with a computing circuit together, like a sandwich, that achieves significantly less data access (Fig. 24.4.1). It also uses a reconfigurable analog-computation engine, based on pulse-width modulation, that is small and flexible enough to be inserted into the memory.

Figure 24.4.2 shows the overall architecture of the sandwich–shaped computing SRAM (Sandwich-RAM): consisting of a control block for configuration, a weight pre-processing unit to reduce weight complexity, and a post-processing unit including an accumulation memory. The system supports two modes: a conventional SRAM mode, and a computing SRAM mode for convolutional operations. Sandwich-RAM contains 28×112 processing blocks, a configurable pulse generator and a pulse-quantizer by replica cell (PQRC). Each block comprises of 6×8b 8T SRAM cells to store partial feature data, several shift registers for partial weight storage, and an analog-compute unit sandwiched between the memory arrays. A pulse computation unit is shared between 6 features, and one feature is selected for computation at a time. After one calculation, the weights can be shifted for the next calculation, repeating as such until all convolutions with the stored feature map are computed. The traditional in-memory computing architecture is inflexible as it only operates on a row-by-row or column-by-column mode. This results in low utilization when the filter kernel size varies. Sandwich-RAM can be reconfigured, based on the network, into three modes: a large-, medium- or small-window mode for large, medium or small kernels. It can perform convolution operations for different kernel sizes to increase Sandwich-RAM utilization. Depending on the configuration mode from the controller, an input 3D-feature map and weight kernel can be broken up into several 2D planes and dispersed into Sandwich-RAM. A fixed-width pulse is generated on the left side of Sandwich-RAM, modulated by each pulse-width modulation unit (PWMU) and quantized on the right. The PWMU performs multiplication and accumulation in the time domain, by selecting different control voltages, depending on the local feature, to modulate the pulse width: positive values make pulses wider, while negative values make them narrower. A delay-sensitive control-voltage generator (DSCVG) adjusts the control voltages in real time to minimize the impact of supply voltage fluctuations, and a PQRC is used to measure the pulse width, which makes the analog computations more accurate and stable.

Each of the four blocks in a row can be seen as a stacked 3D-convolution map, and are bypassed when these blocks do not need to be calculated (Fig. 24.4.3). The PWMU has three paths: an add path, a subtraction path and a bypass path. They are controlled by a 2b shift-register: a 1b-weight and a 1b-bypass flag. The add and subtraction paths make the input pulse width wider or narrower. The RWL selects computation feature data F[7:0] in the partial SRAM array, which are divided into two parts: F[7:4] modulates the MSB chain and F[3:0] modulates the LSB chain. Each 2b feature selects the control voltage for a delay unit via 2:4 decoder. The control voltage V_1 is supplied from V_{DD}, while voltages V_2, V_3 and V_4 are generated by the DSCVG to guarantee the delay is 2, 3 or 4× of the delay controlled by V_1. That is for F[1:0] = 0, the control voltage is V_{DD} and pulse width is modulated by 1 unit of change. Setting F[1:0] to 1, 2, or 3 modules the pulse width by a 2, 3 or 4× of the unit change. The PWMU for F[3:2] is 4 times larger than for F[1:0], which makes the pulse change linearly.

Figure 24.4.4 shows a PQRC, which consists of dynamic C^2MOS sample registers, compressed adder tree, accumulation circuits and a voltage controlled oscillator(VCO) for the sample CLK. The delay cell in the VCO is as same as that in PWMU. A replica delay cell, controlled by V_{DD}, is used to mimic the unit addition/subtraction in the PWMU. The clock frequency changes with the pulse modulation in the PWMU when global PT varies. The VCO can be configured to 4-quantization accuracies, depending on the CNN network type; the generated clock measures the pulse width of the accumulation result. The PWMU output varies by 63.0% under process and temperature variation, but the quantization fluctuation is only 16.7% after PQRC. In addition, pulse width modulation is very sensitive to the control voltage, especially under PVT variation. We propose using a DSCVG for control voltage generation. A set of PMOS and NMOS are controlled by the width of the ENABLE signals. Meanwhile, V_2, V_3 and V_4 are dynamically tuned by the number of conducting PMOSs, and determine the width of ENABLE1. Ultimately, the width of the ENABLE1 pulse will be equal to the width the ENABLE2 pulse, and the control voltage reaches a balanced value that guarantees the delay of voltage-controlled buffer in chain 1 is K× that of chain 2. The DSCVG does not need to provide supplement current for V_2, V_3 or V_4, so power consumption is negligible. Compared with a fixed and proportional control voltage scheme, the DSCVG linearity-tuned PWMU performs better.

Figure 24.4.5 presents two effective methods to increase energy efficiency. (1) Filtered weights computation, where the original kernel is decomposed into a base kernel and a filtered kernel. The base kernel consists of all -1 or 1 values, while the filtered kernel is the difference between the original one and the base kernel: filtered ("/") means that the result doesn't need to be calculated. The weight pre-processing unit counts the number of -1 and 1 values in the original weight kernel, and then selects a suitable base kernel to filter it in order to reduce its complexity by ~2.3×. In a conventional base/filtered kernel scheme, the computing logic for the filtered part cannot be gated completely. Using the pulse width modulation scheme allows for the filtered parts in the filtered kernel to be easily bypassed. Nearly 50% of computing energy is saved. (2) Feature-aware computation, where the pulse is transmitted through different paths based on the feature. When the feature data is relatively large (F[7:4] = 010X, 0110, etc.), the LSB chain will be bypassed, which achieves computation-energy saving at the expense of a little bit of accuracy loss. Conversely, when the feature data is relatively small (F[7:4] = 0000), the MSB chain is bypassed, which also saves energy, but without accuracy loss. Comparing to a conventional scheme, the BWN is feature-aware and more flexible. After using the stochastic gradient Descent (SGD) algorithm to train with the ImageNet dataset, a 30%-45% reduction in energy is seen for different configurations, accuracy is reduced by less than 1%.

Sandwich-RAM is fabricated in a 28nm CMOS technology; the die photograph is shown in Fig. 24.4.7. Figure 24.4.6 shows the computing differential nonlinearity (DNL), which is the accumulation result for one row under different voltage and temperature corners. The DNL is less than 0.5·LSB, because DSCVG dynamically adjusts the control voltage and PQRC tracks global PT variation. The energy efficiency is improved by 4.16× based on filtered-weight computation and feature-aware computation. A 119.7TOPS/W peak-energy efficiency is achieved at 0.6V. When running AlexNet convolutional layers, Sandwich-RAM achieves an energy efficiency of 46.6TOPS/W.

Acknowledgements:
This work was supported by the National Natural Science Foundation of China (NO. 61474022) and the National Science and Technology Project under Grant 2018ZX01031101-005.
We would like to thank the TSMC University Shuttle Program for chip fabrication. We also thank Taomei Zhou, Mengyang Sun and Lizheng Ren for their help.

References:
[1] A. Biswas, et al., "Conv-RAM: An Energy-Efficient SRAM with Embedded Convolution Computation for Low-Power CNN-Based Machine Learning Applications," *ISSCC*, pp. 488-489, 2018.
[2] D. Miyashita, et al., "A Neuromorphic Chip Optimized for Deep Learning and CMOS Technology with Time-Domain Analog and Digital Mixed-Signal Processing," *JSSC*, vol. 52, no. 10, pp. 2679-2689, Oct. 2017.
[3] R. Andri, et al., "YodaNN: An Architecture for Ultralow Power Binary-Weight CNN Acceleration," *IEEE TCAD*, vol. 37, no. 1, pp. 48-60, Jan. 2018.
[4] S. Yin, et al., "An ultra-high energy-efficient reconfigurable processor for deep neural networks with binary/ternary weights in 28nm CMOS," *VLSI*, pp. 37-38, 2018.
[5] W.-H. Chen, et al., "A 65nm 1MB Nonvolatile Computing-in-memory ReRAM Macro with Sub-16ns Multiply-and-Accumulate for Binary DNN AI Edge Processors," *ISSCC*, pp. 494-496, 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

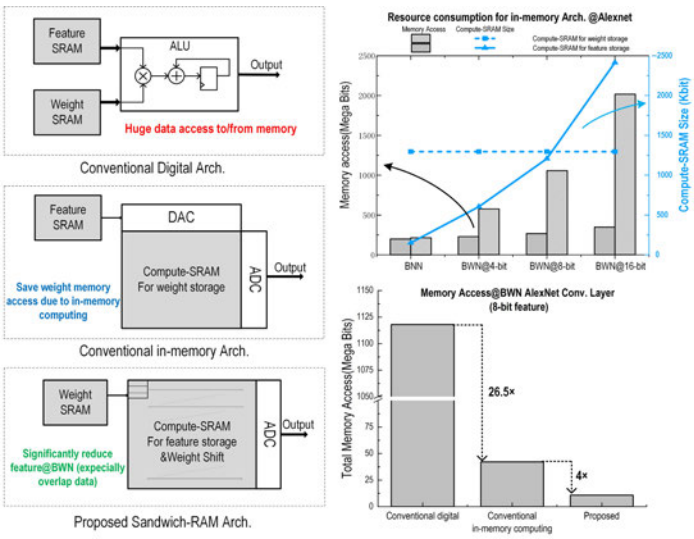
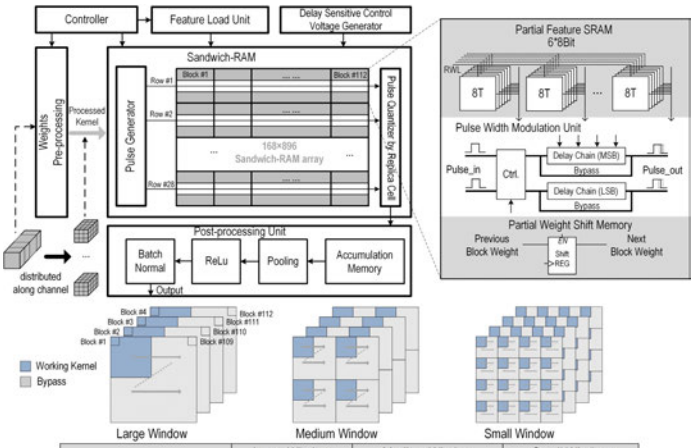


Figure 24.4.1: Proposed Sandwich-RAM architecture.



		Large Window	Medium Window	Small Window
Sandwich-RAM Blocks Map		(28×4)×(28×6)	(14×4×2)×(14×6×2)	(7×4×4)×(7×6×4)
Alexnet	Feature Map	28×28×24	14×14×96	7×7×384
	Weights	11×11×3	5×5×16	3×3×64
	Output	5×5×1×6cycles	10×10×1/6cycle	5×5×1/6cycle

Figure 24.4.2: Architecture, feature memory and CNN computation mode.

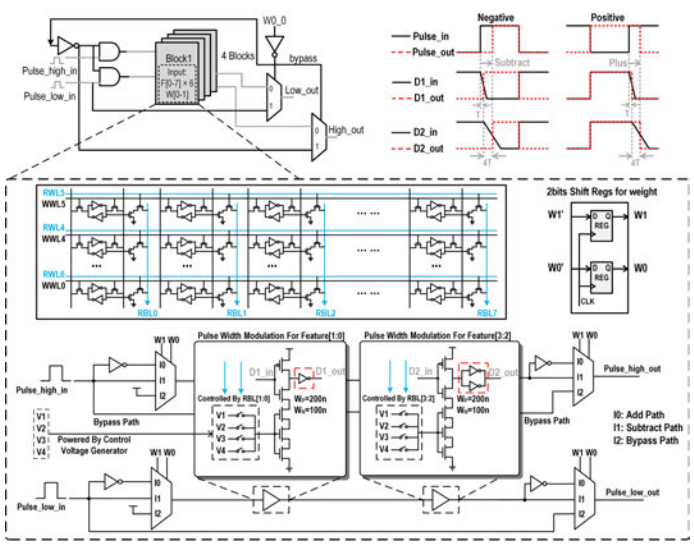


Figure 24.4.3: Processing block schematics including memory and PWMU.

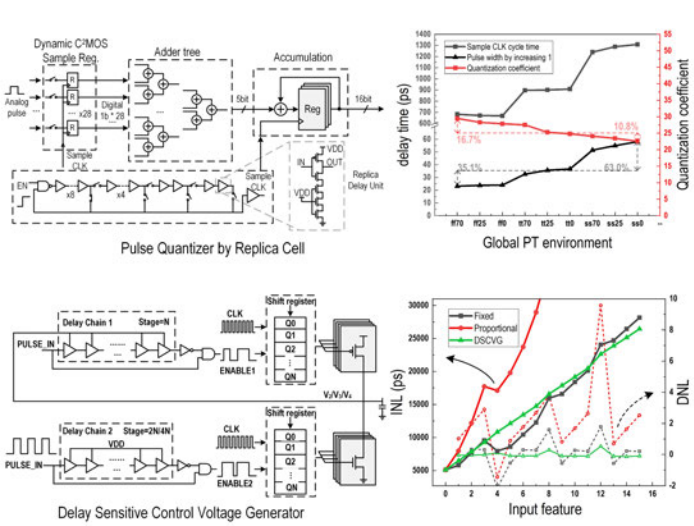


Figure 24.4.4: MAC accuracy improved by PQRC and DSCVG.

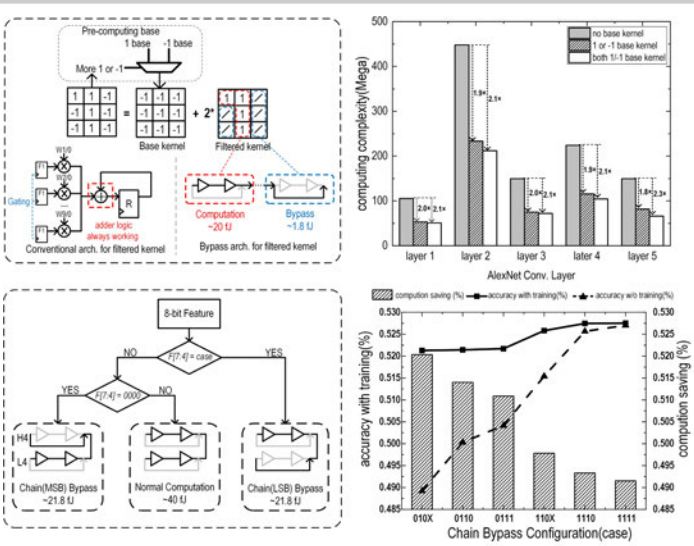
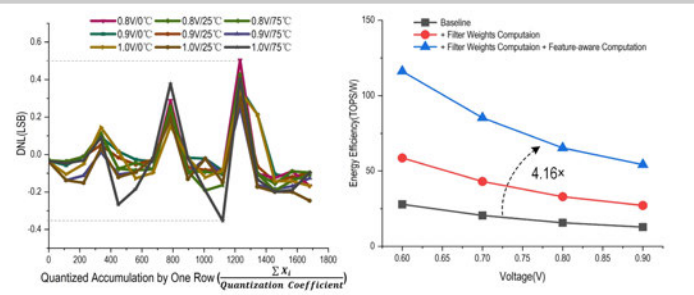


Figure 24.4.5: Complexity reduction by filtered-weight and feature-aware.

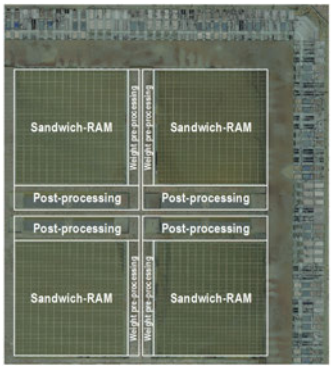


	Tech(nm)	Benchmark	Function	Computaion	Architecture	Energy Efficiency (TOPS/W)	
						Peak	Alexnet
Conv-RAM [1]	65	LeNet-5	BWN@6b	Analog(Voltage)	F ⊞	28.1	—
Miyashita [2]	65	LeNet-5	BNN	Analog(Time)	F ⊞	77	—
Yoda [3]	65	AlexNet Conv Layers	BWN@12b	Digital	F W +	—	2.2
YIN [4]	28	AlexNet Binary Weight	BWN@1~16b	Digital	F W +	95.8	19.9
This work	28	AlexNet Conv Layers	BWN@8b	Analog(Time)	⊞	119.7	46.6

* □means SRAM, +means MAC, ⊞means combine SRAM and MAC together.

Figure 24.4.6: Measurement and prior-art comparison table.

24



Technology	28nm
Supply voltage	0.6V ~ 0.9V
Frequncey	400MHz(max.)
Core size	0.424 × 0.51 mm^2
*Sandwich-RAM*4*	Feature:150Kb(28*28*4*6*8Bits) 8T Weight:6Kb(28*28*4*2)
Accumulation SRAM	2Kb
energy effiency (TOPS/W)	12.8~119.7

Figure 24.4.7: Die photograph and chip summary table.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

24.5 A Twin-8T SRAM Computation-In-Memory Macro for Multiple-Bit CNN-Based Machine Learning

Xin Si[1,2], Jia-Jing Chen[1], Yung-Ning Tu[1], Wei-Hsing Huang[1], Jing-Hong Wang[1], Yen-Cheng Chiu[1], Wei-Chen Wei[1], Ssu-Yen Wu[1], Xiaoyu Sun[3], Rui Liu[3], Shimeng Yu[4], Ren-Shuo Liu[1], Chih-Cheng Hsieh[1], Kea-Tiong Tang[1], Qiang Li[2], Meng-Fan Chang[1]

[1]National Tsing Hua University, Hsinchu, Taiwan
[2]University of Electronic Science and Technology of China, Chengdu, China
[3]Arizona State University, Tempe, AZ
[4]Georgia Institute of Technology, Atlanta, GA

Computation-in-memory (CIM) is a promising avenue to improve the energy efficiency of multiply-and-accumulate (MAC) operations in AI chips. Multi-bit CNNs are required for high-inference accuracy in many applications [1-5]. There are challenges and tradeoffs for SRAM-based CIM: (1) tradeoffs between signal margin, cell stability and area overhead; (2) the high-weighted bit process variation dominates the end-result error rate; (3) trade-off between input bandwidth, speed and area. Previous SRAM CIM macros were limited to binary MAC operations for fully connected networks [1], or they used CIM for multiplication [2] or weight-combination operations [3] with additional large-area near-memory computing (NMC) logic for summation or MAC operations.

This work presents a configurable SRAM CIM macro with 1-4b inputs, 1-5b weights, and a MAC-value (MACV) of up to 7b outputs. The CIM uses (1) a compact-rule compatible twin-8T (T8T) cell for weighted CIM MAC operations to reduce the area overhead and vulnerability to process variation; (2) an even-odd dual channel (EODC) array to double the input bandwidth; (3) a two's complement mapping and processing unit (C2PU) to enable MAC operations using positive and negative weights within a cell array, so as to reduce the area overhead and computation time. Our fabricated 55nm 64×60b T8T macro demonstrated the first configurable multiple-bit SRAM-CIM. This work also achieves the fastest CNN multi-bit MAC operations (3.2ns for 1b input, 2b weight and 3b MACV, and 5ns for a 2b input, 5b weight and a 5b MACV).

Figure 24.5.2 shows our configurable multi-bit T8T SRAM CIM macro, comprising of T8T-SRAM cells. Also shown is an even/odd dual-channel (EODC) array structure, a two's complementary processing unit (C2PU), an output combiner (OC), and a configurable reference column (CRC). The j-bit signed weights (W_i[j-1:0]) of each $n\times n$-CNN filter are stored in the T8T array using $2n^2$ rows and j columns, in two's complement form. Each T8T cell comprises of two read-decoupled 8T (RD8T) cells: a most-significant 8T (M8T) and least-significant 8T (L8T). The transistor width of N2/N1 in M8T is 2× that in L8T to provide a weighted cell current ($I_{MC\text{-}M8T} = 2\cdot I_{MC\text{-}L8T}$). The read decoupled transistor (RDT) structure enables, (1) full(large)-voltage swing on the RBL to increase the signal margin without encountering the read-disturb or write-disturb issues common for a 6T-SRAM CIM [1, 3]; (2) a small area overhead by using the foundry's compact-rule 8T-SRAM cells rather than the dedicated large-area logic-rule SRAM cells. Each input (IN) is applied to a read-WL (RWL) that supports binary- (2 levels: 0V and V_{WLL3}) or 2b- (4 levels: 0V, V_{WLL1}, V_{WLL2}, & V_{WLL3}) input precision. A dual-rail power distribution network allows for V_{WLL3} to exceed the standard supply voltage (V_{DD}). The WL voltage is used to generate a weighted cell current (I_{MC}): $^1/_3\cdot I_{MC}$, $^2/_3\cdot I_{MC}$, and I_{MC}. V_{WLL1}, V_{WLL2} and V_{WLL3} are determined by 1k Monte-Carlo simulations using different MAC patterns. For $n\times n$ MAC operations n^2 inputs (IN) are applied to n^2 RWLs in parallel, for each channel. Each T8T cell performs 2b-input 2b-weight multiplication with a discharge cell current ($I_{MC\text{-}T8T} = I_{MC\text{-}M8T} + I_{MC\text{-}L8T}$) based on the input-weight-product (IWP, which is IN[1:0]×W[1] + IN[1:0]×W[0]). Summing all I_{MC} on a RBL against a BL pull-up transistor results in a RBL voltage (V_{RBL}) that is the summation of n^2 2b-input 2b-weight products ($\Sigma IN_i[1:0]\times W_i[1:0]$). By combining several contiguous RBL lines, multi-bit MAC operations ($\Sigma IN_i[j\text{-}1:0]\times W_i[k\text{-}1:0]$) can be implemented via (C2PU) and OC.

The developed EODC scheme (Fig. 24.5.3) extends the bandwidth of CNN operations and reduces energy consumption. The even-row T8T cells, in a column, are connected to RBL-even (RBLE), while the odd-row cells are connected to RBL-odd (RBLO). This structure reduces in half the number of cells (and parasitic load) on an RBL in a typical cell array. By using the original RBL trace for each RD8T cell, the RBLO/RBLE in EODC does not consume additional area. The EODC scheme supports two operating modes: a single-channel (SC) and a two-channel-accumulation (2CA) mode. In SC mode, one of the even/odd channels (n^2 RWLE/RWLOs) is activated and its RBLs are connected to C2PUs. Thanks to the reduced BL-load in EODC, the settling time of V_{RBL} and power consumption on RBLs is lower than that without EODC. In 2CA mode, n^2-odd RWLs (RWLO) and n^2-even RWLs (RWLE) are activated simultaneously for parallel MAC operations on RBLEs and RBLOs. RBLE and RBLO on the same T8T column are dotted. The MACVs of these two channels are then accumulated using the same C2PU to double the bandwidth over that in SC mode. The even and odd channels both support configurable (1/2/4-bit) input precision. When the input precision is binary or 2b, T8T CIM requires only one cycle to compute MACV. When the input precision is 4b (IN[3:0]), each input is applied to an assigned channel using two cycles: IN[3:2] and IN[1:0]. The OC then combines the MACV results from the 1st cycle and 2nd to output the final multi-bit MACV.

Figure 24.5.4 presents the proposed two's-complement weight mapping and C2PU. In prior work [3], only multi-bit weights are implemented within an array, and all MAC operations are realized outside of the memory array (as NMC). In this work, the signed weights are stored in the T8T array in two's-complement form, which allows for the support of multiple RWL activations and to sum both positive and negative multiplied results together within the memory array. Owing to the proposed two's complement weight mapping scheme, the NMC area could be reduced by 158% compared to [3] for 2b-input 4b-weight MACs. For a given IN[1:0] codeword on an odd channel, the resulting RBLs are as follows: RBLS=$\Sigma(IN_i[1:0]\times W_i[4])$, RBLO[1]=$\Sigma(IN_i[1:0]\times W_i[3:2])$, RBLO[0]=$\Sigma(IN_i[1:0]\times W_i[1:0])$. When using C2PU and configurable reference columns (CRC), KS in Fig. 24.5.4 is switched from V_{DD} to MACS to lower the voltage of SUM (V_{SUM}) by $^1/_2$. In the meantime, K1 and K0 are switched from MAC[1:0] to V_{DD} to increase V_{SUM} by $^1/_8$, and $^1/_{32}$. KS, K1, and K0 have the similar structure to a 2:1 multiplexer, except with differently weighted capacitor loading. The total summation (TS) with a 5b-MACV is derived as follows: MACS×(-16) + MAC[1]×(+4) + MAC[0]×(+1) = $\Sigma(IN_i[1:0]\times W_i[4:0])$.

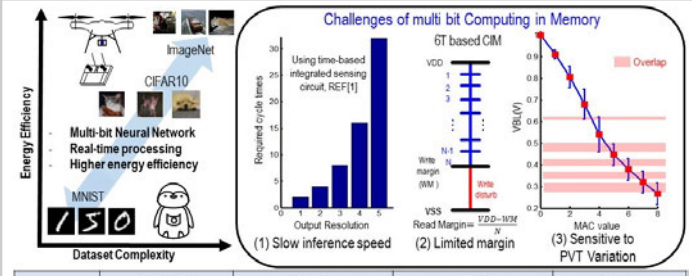
Figure 24.5.5 shows the simulated performance of the proposed schemes. A larger voltage swing (with no write-disturb) and a larger transistor width at the read-port allows the T8T cell to achieve a 1.4-1.6× increase in sensing margin, compared to 6T cells with the same MACV precision. Thanks to its compact area and double throughput, the EODC (SC mode) scheme achieves a throughput/area FoM that is over 1.98× larger than that of fully-serial and fully-parallel input/weight structures. When the EODC and 2CA schemes are combined they achieve over 2.35× better FoM. When using two's-complement weight mapping, C2PU and OC achieve computation times that are 2.6× (3b MACV) and 4.85× (5b MACV) faster than [2]. This is also 2.04× (3b MACV) and 2.08× (5b MACV) faster than [3]. [1] and [3] do not use a signed MACV.

Figure 24.5.6 presents the measured results from a test chip fabricated using a 55nm 64×60b T8T-SRAM CIM macro. For CNN operations with 3×3 kernels the shmoo plot confirms a $t_{AC\text{-}MAC}$ of 5ns for a 2b-input 5b-weight 5b-MACV. Using 4b-inputs, 5b-weights and a 7b-MACV the array achieves a 90.42% accuracy in the inference of 10k CIFAR-10 images. This work achieves over a 22× improvement in the energy-efficiency × throughput/capacity FoM (TOPS/W × GOPS/KB) compared to prior SRAM-CIM work [1-3]. Figure 24.5.7 presents the die photo and a chip summary.

Acknowledgements:
The authors would like to thank MOST-Taiwan, CIC, NTHU-TSMC JDP, NTHU-MTK-JDP, M31 for financial and manufacturing support.

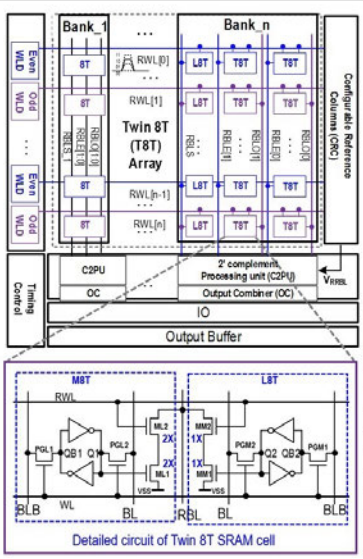
References:
[1] W.-S. Khwa, et al., "A 65nm 4Kb Algorithm-Dependent Computing-in-Memory SRAM Unit-Macro with 2.3ns and 55.8TOPS/W Fully Parallel Product-Sum Operation for Binary DNN Edge Processors," *ISSCC*, pp. 496-497, 2018.
[2] A. Biswas, et al., "Conv-RAM: An Energy-Efficient SRAM with Embedded Convolution Computation for Low-Power CNN-Based Machine Learning Applications," *ISSCC*, pp. 488-489, 2018.
[3] S. K. Gonugondia, et al. "A 42 pJ/Decision 3.12TOPS/W Robust In-Memory Machine Learning Classifier with On-Chip Training," *ISSCC*, pp. 490-491, 2018.
[4] K. Ueyoshi, et al. "QUEST: A 7.49 TOPS Multi-Purpose Log-Quantized DNN Inference Engine Stacked on 96 MB 3D SRAM Using Inductive-Coupling Technology in 40nm CMOS," *ISSCC*, pp. 216-217, 2018.
[5] J. Zhang, et al., "In-Memory Computation of a Machine-Learning Classifier in a Standard 6T SRAM Array", *JSSC*, pp. 915-924, 2017.

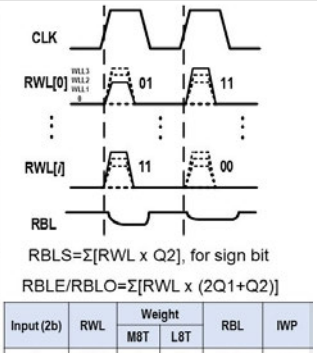
 978-1-5386-8531-0/19/$31.00 ©2019 IEEE



	This work	REF[2]	REF[3]	REF[1]
Structure	Twin 8T (T8T) Array (operation: x/+, Multi-bit)	Local 10T Array (operation: x, binary) / Local Analog Averaging (operation: +)	6T Array (operation: multi-bit weight combination) / Near memory computing (operation: x/+)	DSC 6T Array (operation: x/+, binary)
Model	Multiple-bit NN	Binary weight NN	Support Vector Machine	Fully connected layer
SRAM Cell	Twin 8T(T8T)	10T	Typical 6T	DSC6T
IMC with x/+ in array	Yes	No (binary multiplication)	No	Yes (binary)
Support CNN	Yes	Yes	No	No
Write disturb	No	No	Yes	Assist
Weight	2's complement	Binary (+1/-1)	1's complement	Binary (+1/-1)

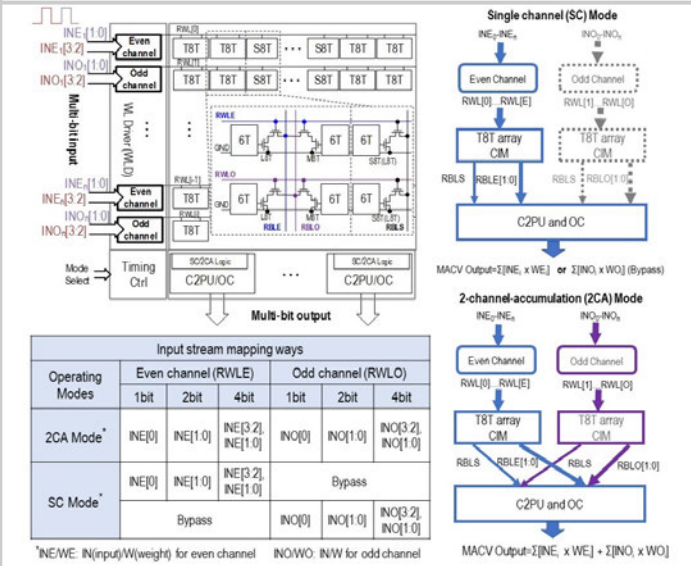
Figure 24.5.1: Overview of previous SRAM-CIM and this work.





Input (2b)	RWL	Weight M8T	Weight L8T	RBL	IWP
11	WLL3	1	1	3 x I_{MC}	9
10	WLL2	1	1	2 x I_{MC}	6
01	WLL1	1	1	1 x I_{MC}	3
11	WLL3	1	0	2 x I_{MC}	6
10	WLL2	1	0	4/3 x I_{MC}	4
01	WLL1	1	0	2/3 x I_{MC}	2
11	WLL3	0	1	I_{MC}	3
10	WLL2	0	1	2/3 x I_{MC}	2
01	WLL1	0	1	1/3 x I_{MC}	1
00	VSS	1/0	1/0	0	0

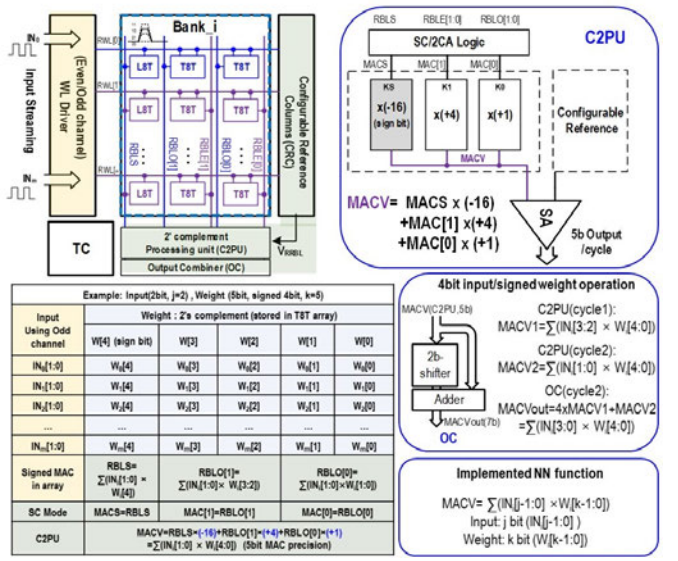
Figure 24.5.2: Macro level structure and twin 8T (T8T) SRAM cell.



Input stream mapping ways						
Operating Modes	Even channel (RWLE)			Odd channel (RWLO)		
	1bit	2bit	4bit	1bit	2bit	4bit
2CA Mode*	INE[0]	INE[1:0]	INE[3:2], INE[1:0]	INO[0]	INO[1:0]	INO[3:2], INO[1:0]
SC Mode*	INE[0]	INE[1:0]	INE[3:2], INE[1:0]	Bypass		
	Bypass			INO[0]	INO[1:0]	INO[3:2], INO[1:0]

*INE/WE: IN(input)/W(weight) for even channel INO/WO: IN/W for odd channel

Figure 24.5.3: Even odd dual channel (EODC) scheme for multi-bit input stream.



Example: Input(2bit, j=2) , Weight (5bit, signed 4bit, k=5)					
Input Using Odd channel	Weight : 2's complement (stored in T8T array)				
	W[4] (sign bit)	W[3]	W[2]	W[1]	W[0]
$IN_0[1:0]$	$W_0[4]$	$W_0[3]$	$W_0[2]$	$W_0[1]$	$W_0[0]$
$IN_1[1:0]$	$W_1[4]$	$W_1[3]$	$W_1[2]$	$W_1[1]$	$W_1[0]$
$IN_2[1:0]$	$W_2[4]$	$W_2[3]$	$W_2[2]$	$W_2[1]$	$W_2[0]$
...	...	...	...	...	...
$IN_m[1:0]$	$W_m[4]$	$W_m[3]$	$W_m[2]$	$W_m[1]$	$W_m[0]$
Signed MAC in array	RBLS= $\sum(IN_i[1:0] \times W_i[4])$	RBLO[1]= $\sum(IN_i[1:0] \times W_i[3:2])$		RBLO[0]= $\sum(IN_i[1:0] \times W_i[1:0])$	
SC Mode	MACS=RBLS	MAC[1]=RBLO[1]		MAC[0]=RBLO[0]	
C2PU	MACV=RBLS=(-16)+RBLO[1]×(+4)+RBLO[0]×(+1) =∑(IN[:1:0] × W[:4:0]) (5bit MAC precision)				

Figure 24.5.4: Two's complement weight mapping and signed MACV.

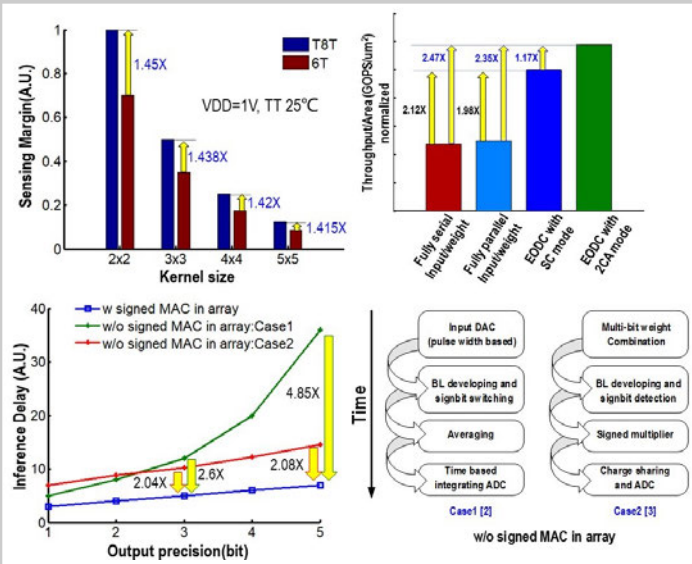


Figure 24.5.5: Simulated performance comparison.

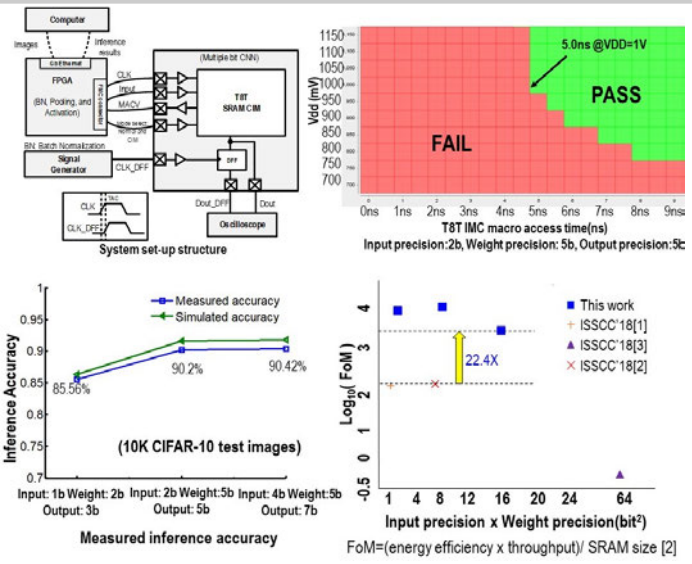
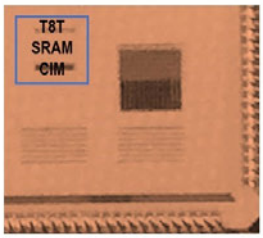
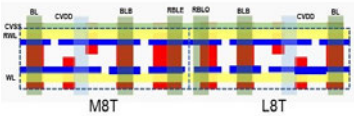


Figure 24.5.6: Measured results.

24





CHIP SUMMARY			
Technology	55nm CMOS		
Unit-macro Size	64X60b		
Bit cell size (In 65nm before shrinking to 55nm)	M8T(0.5um x 1.73um) L8T(0.5um x 1.59um)		
Input precision(bit)	1	2	4
Weight precision(bit)	2	5	5
Output precision(bit)	3	5	7
Inference time(ns)	3.2	5.0	10.2
Energy(pJ)	2.5	4.8	9.8
Energy efficiency (TOPS/W)	72.1	37.5	18.37
Measured accuracy MNIST	99.02%	99.18%	99.52%
Measured accuracy CIFAR10	85.56%	90.2%	90.42%

Figure 24.5.7: Die photo and summary table.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 25 Overview: *Circuits Enabling Security*

DIGITAL CIRCUITS SUBCOMMITTEE

Session Chair: ***Hirofumi Shinohara***
Waseda University, Kitakyushu, Japan

Associate Chair: ***Dennis Sylvester***
University of Michigan, Ann Arbor, MI

Subcommittee Chair: Edith Beigné, Facebook, Menlo Park, CA

This session focuses on two important aspects of securing hardware systems: 1) authentication, and 2) attack resistance. The first two papers offer new approaches to physically unclonable functions (PUFs), which provide unique keys for secure authentication of devices. Challenges in this space include achieving very low error rates, providing compact and energy-efficient key generation, and process scalability. These papers employ reconfigurability to ensure more stable key generation and provide gains in native bit error rates and/or energy. The third paper focuses on side-channel-attack resistance: this work leverages desynchronization of the regulator and AES clocks to reduce side-channel information leakage, making attacks significantly more difficult, and less likely to succeed.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

1:30 PM

25.1 A 562F^2 Physically Unclonable Function with a Zero-Overhead Stabilization Scheme

D. Li, Rice University, Houston, TX

In Paper 25.1, Rice University describes a compact PUF that exploits an analog-biased native NMOS header to provide supply voltage insensitivity. The PUF is also reconfigurable: one can chain together as many stages as needed to ensure stable full rail readout. The design operates across the widest reported temperature range (–55 to 125°C) with very low 0.00182% bit error rate after a new stabilization technique is applied. Total energy efficiency is 15.3fJ/b, and resiliency to aging is demonstrated.

2:00 PM

25.2 A Reconfigurable RRAM Physically Unclonable Function Utilizing Post-Process Randomness Source with <6×10^{-6} Native Bit Error Rate

Y. Pang, Tsinghua University, Beijing, China

In Paper 25.2, Tsinghua University, National Tsing Hua University, and Georgia Institute of Technology present a reconfigurable PUF based on a differential resistive RAM (RRAM) structure. The PUF achieves low native bit error rate (<6E-6) and is implemented in 0.13μm technology with each bit occupying 2.86μm^2.

2:30 PM

25.3 A 128b AES Engine with Higher Resistance to Power and Electromagnetic Side-Channel Attacks Enabled by a Security-Aware Integrated All-Digital Low-Dropout Regulator

A. Singh, Georgia Institute of Technology, Atlanta, GA

In Paper 25.3, Georgia Institute of Technology and Intel use a digital low-dropout regulator (LDO) to power 128b AES engines to provide side-channel attack resistance, specifically for power and electromagnetic probe type attacks. Security is further enhanced via injection of random switching noise, as well as pseudo-random clock-edge dithering. Mean-time-to-disclosure of AES keys is increased by 2182 to 3579× over a baseline design, with relatively low 10% performance overhead.

25.1 A 562F² Physically Unclonable Function with a Zero-Overhead Stabilization Scheme

Dai Li, Kaiyuan Yang

Rice University, Houston, TX

Internet of Things (IoT) devices bring a growing demand for secure, low-power, and low-cost secret key and ID storage solutions. Physically unclonable functions (PUFs) are one of the most promising alternatives to conventional non-volatile memory (NVM) solutions, which usually incur extra cost and are vulnerable to physical attackers. In recent years, significant research efforts have been devoted to improving the stability of PUF responses/keys across PVT variations, while maintaining the low-cost and low-power benefits. PUF design involves circuit topologies, stabilization methods, and error correction schemes. While redundancy-based stabilizing methods, such as extensive unstable bit masking and error correction codes (ECCs), are powerful to achieve 100% reliable keys [1,2,7], the on-chip area cost for redundancy, energy and latency during key extraction, and testing cost grow super linearly with the percentage of unstable bits in most cases (Fig. 25.1.1). Therefore, PUF designs with native reliability and lossless stabilizing methods are desirable. PUFs using metastability are fast and efficient [1,2], but suffer from transient noise. PUFs relying on shared digitizers [3,4] save area but require accurate comparators and suffer from noise during readout. PUF cells with static operation and local digitizing are shown to achieve the best stability [5-7], among which the amplifier chain structure shown in Fig. 25.1.1 demonstrates best native stability and low cost [6-7]. A NAND-based design [7] suffers from high power consumption and large area, while a 2T amplifier-based PUF [6] achieves a good trade-off among desired properties. However, the 2T amplifier in [6] does not scale well with technology, limited by its amplification gain and asymmetrical headroom (Fig. 25.1.1, top right). In this paper, a 562F² PUF cell, which can be reconfigured to a different PUF cell locally for lossless stabilization with zero area overhead (Fig. 25.1.1, bottom right), is implemented in 65nm CMOS. It features: 1) a 0.00182% native bit error after lossless stabilization and 0.44% across the military temperature range (-55 to 125°C); 2) a low-cost stabilization scheme combining a reconfigurable PUF cell with a body-bias-based V/T emulation method, which improves bit error rate (BER) by 149× and bit instability by 120×; 3) sub-threshold operation with only 0.062fJ/b core energy.

The PUF employs a similar structure as [6-7], which amplifies the differences in switching voltages between the first two stages to full rail after a few stages. Amplification gain is one of the dominant factors in output stability. With technology scaling, the stability of the state-of-the-art 2T PUF in [6] can be significantly reduced according to simulation. As shown in Fig. 25.1.1, inverters in the sub-threshold region consistently provide higher gain than the 2T amplifier [6] under the same current across technology nodes. In order to bias the inverter in the sub-threshold region and improve its resistance to voltage variations, a native NMOS is used for local voltage regulation introducing no power overhead. This is possible because of the low and stable current drawn by the inverters in the PUF. With proper sizing of the native transistor, the virtual V_{DD} can be relatively temperature insensitive, protecting the PUFs against temperature variations. In this simplified model, it is assumed that most current of a PUF cell goes through the first stage biased at roughly half of the VV_{DD}. Compared to using LDOs to provide a regulated supply, our design has higher efficiency and better flexibility because an arbitrary number of PUF cells can be grouped together, either as a large array or a single word embedded into other modules. In this work, a 32×128 array is implemented for prototyping. Native NMOS headers are shared by each column. The native transistor is biased by a reference voltage to keep the PUF cells in the desired operation region. Because the PUF cells can tolerate supply variations on the order of tens of mV, an ultra-low-power threshold voltage-based voltage reference [8] can be employed here. Measured voltage and temperature stability of the virtual V_{DD} are shown in Fig. 25.1.2 (bottom right). The native regulator shows less than 7mV/V line sensitivity across V_{DD} of 0.7-1.4V and VV_{DD} varies less than 10mV across -55-125°C. To achieve word-wide parallel access, two readout transistors, similar to an 8T SRAM, are added to the PUF cell. Complete read peripherals consist of sense amplifiers (SAs), SR-latches, WL drivers, and timing-generation circuits running at nominal 1.2V V_{DD} are implemented on-chip (Fig. 25.1.2, bottom left).

PUF cells are unstable when the switching voltages of the first two inverters have little mismatch, making the final PUF outputs dependent on thermal noise and/or V/T dependencies of transistors. To stabilize these bits, effective methods to locate and correct them are necessary. This work adds a reconfigure (RC) transistor connecting the input and output of the second stage for lossless stabilization with zero area overhead (Fig. 25.1.1 bottom right). Closing this switch forms a new 3-stage PUF cell almost independent from the original one. While the native BER of the reconfigured cell is larger than the original one, due to smaller mismatch of the combined "1st stage". The probability that both the original and the reconfigured cells are unstable is 0.16%, even across -55-125°C and 0.7-1.4V variations. PUFs within the same column share a vertical RC control line because they are never accessed at the same time. To apply reconfiguration, a reconfigure map (R-MAP) marking the locations of unstable bits must be determined. The ideal method is to sweep voltage and temperature over expected operating conditions during the enrollment phase and record all unstable bits. However, this approach adds significant testing costs. An alternate approach to finding the R-MAP is by emulating V/T variations using body bias (EVB). As shown in Fig. 25.1.3, changing VPW by ±0.4V can improve the BER by up to 70% when combined with reconfiguration. As expected, forward bias works better at 125°C and reverse bias is more effective at -55°C. However, the prediction accuracy of EVB is only 30%, implying large number of stable bits will be wasted if a conventional masking method is employed. The proposed reconfiguration method complements the body biasing method because bits falsely included in R-MAP and reconfigured still have a high probability to be stable. A flow chart for the enrollment of R-MAP is shown in Fig. 25.1.3.

The PUF test chip is fabricated in a 65nm LP process. Measured with 10 samples, the average native bit error rate (BER) at nominal conditions is 0.30%. After stabilization using temporal majority voting of 11 samples (TMV11) and reconfiguration using R-MAP acquired at nominal conditions, BER is reduced to 0.00182%. Unstable bits at nominal conditions are reduced from 2.95% to 0.024% using TMV11 and reconfiguration. V/T variation impacts are also shown in Figure. 25.1.4. The voltage regulator suppresses the errors induced by voltage variation to 0.4% (raw). The bottom-left plot in Fig. 25.1.4 compares three methods for getting R-MAP to reconfigure: 1) reconfigure only unstable bits at nominal conditions; 2) EVB with VPW from -0.4-0.4V; and, 3) sweeping temperature from -55-125°C. Measurements show that method 1 only improves BER by 17.4% , while EVB improves BER by 45.6% at -55°C and 66.8% at 125°C. Method 3 yields the lowest bit errors, but consumes excessive time and power, while method 2 achieves significant stability improvement with little overhead.

The intra and inter-die Hamming distance (HD) measured from 10 chips are shown in Fig. 25.1.5. 160× separation implying good identifiability is achieved without any stabilizing technique, while reconfiguration improves identifiability to 1020×. The inter-die HD is 0.4998, indicating excellent randomness and uniqueness of the PUF. Passing all NIST 800-22 and 800-90B randomness sub-tests requiring less than 40960 bits further validates the PUF's uniqueness. Two modes: low power (LP, VV_{DD} around 0.45V) and high performance (HP, VV_{DD} around 0.57V) are available by adjusting the native device bias voltage. In LP mode the PUF runs at 1.459Gb/s with 0.062fJ/b core energy efficiency (not including readout for fair comparison with prior works using different readout schemes). The total energy efficiency including all peripherals (WL driver, readout and timing generation circuits) is 15.3fJ/b, which is 6× lower than [6] (the best reported value among designs with similar array structures). In HP mode, throughput is 8.592 Gb/s and core energy is 0.076fJ/b.

Aging is another source of instability in addition to V/T variations. Thanks to the low regulated virtual V_{DD}, the aging effects within PUF cells can be minimized. An accelerated aging test is performed at 150°C and 1.4V supply. A total of 96-hours stress time causes aging effects (especially NBTI in this design) equivalent to multiple years of normal operation [4]. Bit error rates rise to 1.14% after accelerated aging when compared with golden data enrolled before aging, which is comparable to the best-reported aging resistance in [4] under similar stress conditions. Fig. 25.1.6 summarizes measurement results and compares with prior art. The die micrograph is shown in Fig. 25.1.7.

References:

[1] S. Satpathy, et al., "A 4-fJ/b Delay-Hardened Physically Unclonable Function Circuit With Selective Bit Destabilization in 14-nm Trigate CMOS," *IEEE JSSC*, vol. 52, no. 4, pp. 940-949, 2017.

[2] S. Mathew, et al., "A 0.19pJ/b PVT-Variation-Tolerant Hybrid Physically Unclonable Function Circuit for 100% Stable Secure Key Generation in 22nm CMOS," *ISSCC*, pp. 278-279, 2014.

[3] J. Lee, et al., "A 445F² Leakage-Based Physically Unclonable Function with Lossless Stabilization Through Remapping for IoT Security," *ISSCC*, pp. 132-134, 2018.

[4] J. Li and M. Seok, "Ultra-Compact and Robust Physically Unclonable Function Based on Voltage-Compensated Proportional-to-Absolute-Temperature Voltage Generators," *IEEE JSSC*, vol. 53, no. 3, pp. 2192-2202, 2016.

[5] S. Taneja, A. Alvarez, M. Alioto, "Fully Synthesizable PUF Featuring Hysteresis and Temperature Compensation for 3.2% Native BER and 1.02 fJ/b in 40nm," *IEEE JSSC*, vol. 53, no. 10, pp. 2828-2839, 2018.

[6] K. Yang, et al., "A 553F² 2-Transistor Amplifier-Based Physically Unclonable Function (PUF) with 1.67% Native Instability," *ISSCC*, pp.146-147, 2017.

[7] B. Karpinskyy, et al., "Physically Unclonable Function for Secure Key Generation with a Key Error Rate of 2E-38 in 45nm Smart-Card Chips," *ISSCC*, pp. 158-160, 2016.

[8] M. Seok, et al., "A Portable 2-Transistor Picowatt Temperature-Compensated Voltage Reference Operating at 0.5V," *IEEE JSSC*, vol. 47, no. 10, pp. 2534-2545, 2012.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

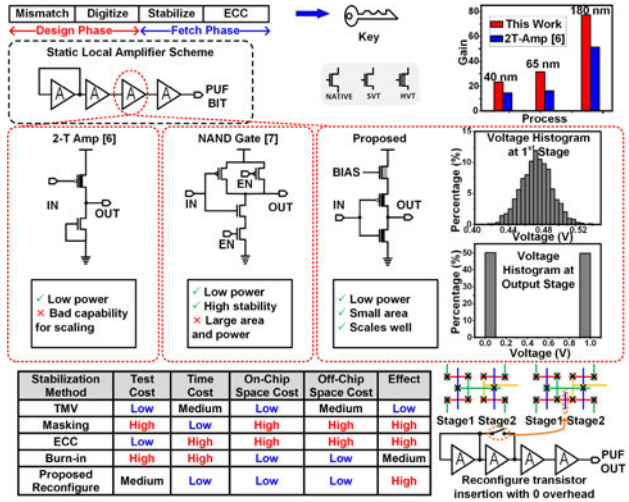


Figure 25.1.1: Schematic and concept of the proposed PUF cell, cell characteristics from Monte Carlo simulation. A reconfigurable approach by combining two inverters into one with no area overhead is shown.

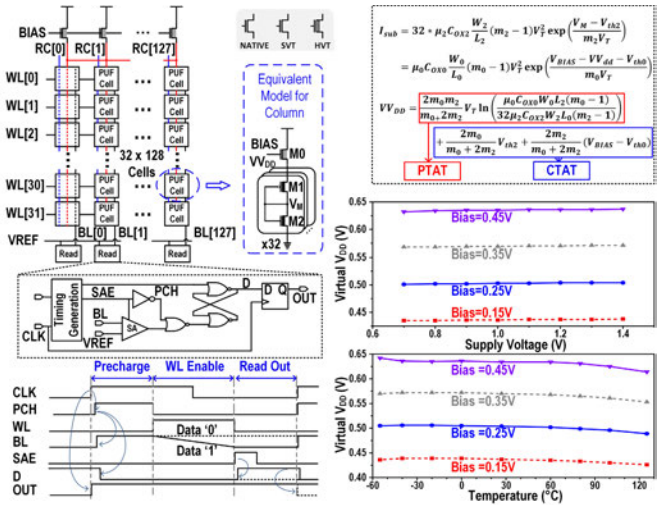


Figure 25.1.2: PUF array topology (with reconfigure control and VV_{DD} running vertically), along with readout circuits and timing diagram; equations for virtual V_{DD} generation and measured VV_{DD} versus V/T variations.

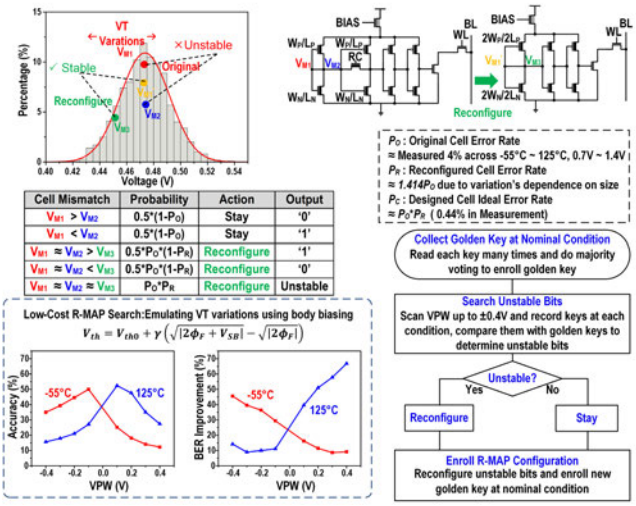


Figure 25.1.3: Concept and flowchart of reconfiguration; BER improvement and R-MAP accuracy when emulating V/T variations using body biasing (EVB) (accuracy: the ratio of truly unstable bits over all bits marked by EVB).

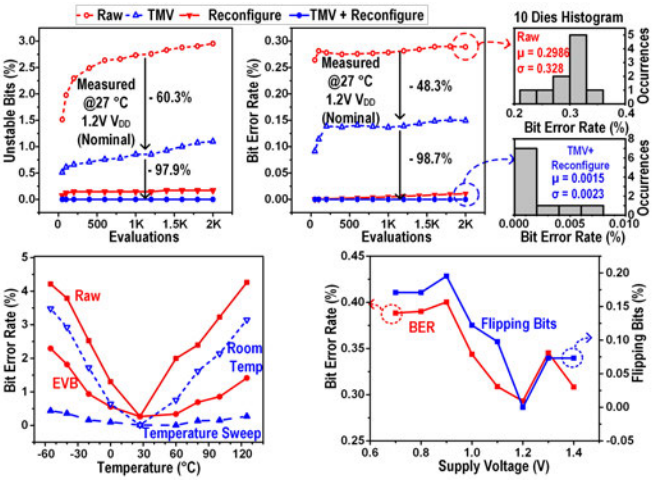


Figure 25.1.4: Measured BER and unstable bit percentage across 10 chips at nominal conditions, before and after various stabilizing techniques (top); BER across -55-125°C after applying different R-MAPs for reconfiguration (bottom left); and BER under V_{DD} variations (bottom right).

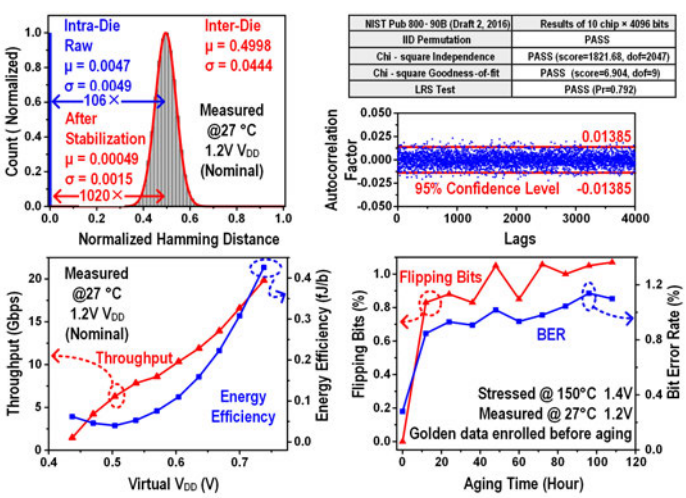


Figure 25.1.5: Measured intra-die and inter-die Hamming distances across 10 chips, NIST Pub 800-90B test results, bit stream autocorrelation, throughput and energy efficiency at different virtual V_{DD} and effects of accelerated aging.

		This work (27°C 1.2V)	JSSC' 18 [5]	ISSCC' 18 [3]	JSSC' 17 [1]	ISSCC' 17 [6]	JSSC'16 [4]	ISSCC' 16 [7]	ISSCC'14 [2]
Technology		65nm	40nm	180nm	14nm	180nm	65nm	45nm	22nm
PUF Cell Area/Bit (F^2)		562	3643	890	9387	782	726	2613	9628
Native Unstable Bits (Evaluations)		2.95% (2000)	2.55% (500)	5.62% (1000)	~27% (5000)	1.73% (2000)	6.54% (500)	-	30% (5000)
Native Bit Error Rates		0.30%	0.81%	0.69%	5.76%	0.18%		0.1%[b]	8.3%
Stabilization Method		TMV, EVB, Reconfigure	Hysteresis, T Compensation	Remapping, TMV	Burn-in, TMV, SBD[c]	Masking, TMV	TMV, Masking	Valid Map, SMV[d], ECC	Burn-in, TMV, Masking
Unstable Bits After Stabilization[a]		0.024%	-	0.08%	-	0.69%	2%	-	4.6%
Bit Error Rate after Stabilization[a]		0.00182%	-	0.019%	1.46%	0.08%	-	-	0.97%
Tested Conditions	Temp (°C)	-55~125	-40~125	0~80	25~110	-40~120	0~80	-25~85	25~50
	Supply (V)	0.7~1.4	0.8~1.0	1.2~1.8	0.55~0.75	0.8~1.8	0.6~1.2	-	0.7~0.9
Bit Errors per 10^6		0.12%[e]	0.15%	-	~0.17%	0.21%	0.44%	0.15%	
Bit Errors per 0.1V		0.057%[e]	0.36%	-	-	0.29%	0.13%	-	0.49%
Fractional Inter-PUF Hamming Distance		0.4998	0.4907	0.50001	0.486	0.499	0.5001	0.498	0.49
Entropy		0.9998	0.9972	-	0.99993	-	0.9998	0.999998	0.9997
Bit Rate (Mb/s)		8592 @HP[h] 1459 @LP[h]	24000	0.018	750	4832	10.2	1.92	2000
Core Energy[f] (fJ/bit)		0.076@HP 0.062@LP	1.02	-	4	13.5	-	-	13
Total Energy[g] (fJ/bit)		25.6@HP 15.3@LP	56.5	3600	-	91.1	548	-	-

a. At nominal condition, before using redundancy-based stabilization methods such as masking
b. Using glitch detection
c. Selective Bit Destabilization
d. Spatial Majority Voting
e. Results after EVB
f. Not including readout energy
g. Not considering energy overhead for stabilization
h. HP Virtual V_{DD} ~ 0.57V, LP Virtual V_{DD} ~ 0.44V

Figure 25.1.6: Summary of measured performance and comparison with prior art.

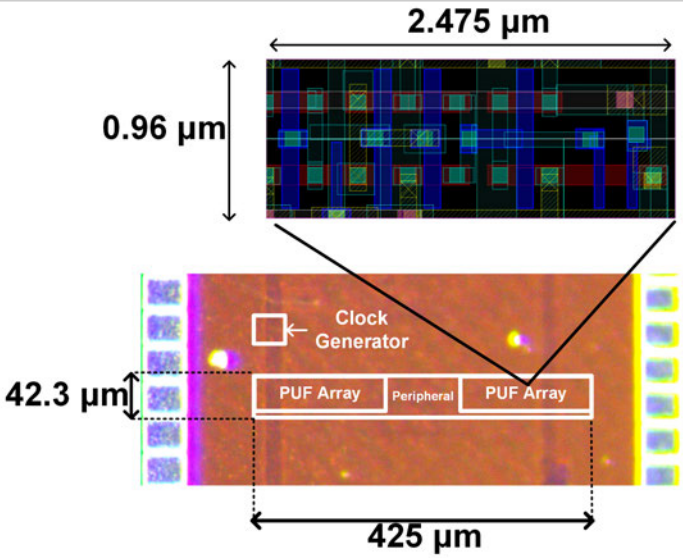


Figure 25.1.7: Die micrograph of 65nm PUF chip and PUF cell layout view.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

25.2 A Reconfigurable RRAM Physically Unclonable Function Utilizing Post-Process Randomness Source with <6×10^{-6} Native Bit Error Rate

Yachuan Pang[1], Bin Gao[1], Dong Wu[1], Shengyu Yi[1], Qi Liu[1], Wei-Hao Chen[2], Ting-Wei Chang[2], Wei-En Lin[2], Xiaoyu Sun[3], Shimeng Yu[3], He Qian[1], Meng-Fan Chang[2], Huaqiang Wu[1]

[1]Tsinghua University, Beijing, China
[2]National Tsing Hua University, Hsinchu, Taiwan
[3]Georgia Institute of Technology, Atlanta, GA

Physically unclonable functions (PUFs) are promising primitives for hardware security with wide applications in the Internet of Things (IoT), e.g., authentication and encryption key generation [1, 2]. Most silicon PUFs utilize process variability of semiconductor manufacturing [1, 3, 4]. These implementations are sensitive to variations in operating conditions (e.g., supply voltage and temperature variations) and undergo significant native bit-error-rates (N-BERs). Thus, additional stabilizing strategies, such as ECC, majority voting, and masking, are necessary. Furthermore, the PUF key after enrollment cannot be changed in prior implementations [1–5]. This could be unsafe if the PUFs are repeatedly used in insecure environments, as PUFs suffer from the challenges of ownership change and overuse (Fig. 25.2.1).

To achieve low N-BER, high reconfigurability, a compact area, and true randomness, we present a reconfigurable resistive random access memory (RRAM)-based PUF by using: 1) an intrinsic (or post-process randomness) physical randomness source instead of a process randomness source for the generation of a key; 2) a resistance-differential comparison between two RRAM arrays for the exact PUF key; and, 3) a split-resistance technique to achieve a low N-BER. The table in Fig. 25.2.1 compares the performance of different PUF implementations. In this study, a 130nm, 8Kb, RRAM PUF macro was implemented to realize a reconfigurable nonvolatile PUF with post-process randomness. This RRAM PUF achieves a low <6E^{-6} N-BER, with a bit-cell area of 2.86μm^2 and 47.29% reconfigure-Hamming Distance (reconfigure-HD) at runtime.

The PUF can obtain a true random encryption key by converting the random analog resistance distribution to a random digital distribution. Two 1-transistor-1-RRAM (1T1R) units are used as a PUF bit cell (PUF cell: RRAM1, differential cell: RRAM2). The RRAM has two resistance states when it is used for memory applications. These two states are called the high and low resistance states (HRS and LRS, respectively). A normal distribution of resistance is formed when programming an RRAM to one of the aforementioned states with a single programming pulse. The random distribution is attributed to the RRAM's device-to-device (D2D) variation, which is the entropy source for the RRAM PUF. Moreover, the distribution of different switching cycles differs considerably owing to the Write cycle-to-cycle (C2C) variation, which is the basis for the PUF reconfigurability (Fig. 25.2.2). To generate one stable PUF bit, we developed a split-resistance technique. A RESET pulse (as HRS has wider resistance distribution) was applied to RRAM1 and RRAM2, the resistances of which were then differentially compared through a small offset sense amplifier (SA). The cell with the smaller/larger resistance is SET/RESET to LRS (<100kΩ)/HRS (>1.5MΩ) with verification. As such, the sensing window is enlarged.

Figure 25.2.3 shows the architecture of the reconfigurable RRAM PUF macro. It includes two 8Kb 1T1R RRAM arrays as PUF and differential arrays. This macro also includes peripheral circuits, such as an X/Y decoder, a program control circuit, a driver, and a small offset SA. The PUF states are defined in this figure as follows: state 1, $R_{PUF\ cell} < R_{differential_cell}$; state 0, $R_{PUF\ cell} > R_{differential_cell}$. For the basic PUF bit, a 1×0.5μm^2 *n*-type transistor serves as the selector to provide enough drive capability for RRAM, and a 0.5×0.5μm^2 RRAM cell is integrated between M5 to M6. To extract the PUF key based on the resistance distribution of HRS, a small offset SA is required for small cell current (~500nA). The differential SA is shown in Fig. 25.2.3. A current-sampling-based SA (CSB-SA) technology [6] is used to distinguish the two resistances of RRAM1 and RRAM2. In addition, owing to the V_t-independent current sampling scheme using C1/C2, the SA can be implemented for differential current input. This approach uses the same device (P1/P2) to sample cell currents (Icell/Iref) and amplifies the current difference; this suppresses the input current offset due to device mismatch and small cell current.

Figure 25.2.4 shows the measured randomness test results among 20 PUF test chips (each PUF chip has 64 PUF keys, and each key is 128b) fabricated in the 130nm 1.8V/5V CMOS process. The tested spatial data mapping of 80 keys showed random and uniform (0/1 bias) distributions, and no apparent spatial correlations. For the Hamming weight (HW) of each key, the probability of 1 in a 128b key among 1000 PUF keys is very close to the ideal value (50%) with a tight normal distribution (μ/σ = 50.013%/2.866%). The embedded table shows the NIST randomness test results. The developed PUF passes all test indices with average P-value >> 0.01 (criteria value).

Figure 25.2.5 demonstrates the PUF's uniqueness and reproducibility among 20 test chips in normal conditions (25°C). For uniqueness, the data show an almost ideal value (50%) of inter-chip HD (μ/σ = 49.99%/4.35%) across 600 PUF keys. As for reproducibility, no error was detected in 10^{10} reproduction cycles, implying that the intra-chip HD is very close to the ideal value of 0 (BER < 1/160Kb ≈ 6.1E^{-6}). As mentioned above, almost all prior PUF implementations are vulnerable to environmental changes. Owing to the split-resistance technique, our PUF is robust under voltage and temperature changes (temperature: 25-150°C, voltage: 1.4-2.2V/2.5-5.5 V). No errors were found after long-term baking for 6000 minutes at 125°C. Furthermore, the spatial autocorrelation function (ACF) test shows no correlation among 30Kb PUF keys (confidence intervals: ±0.0075). This result implies PUF keys are independent, whether from the same or different chips.

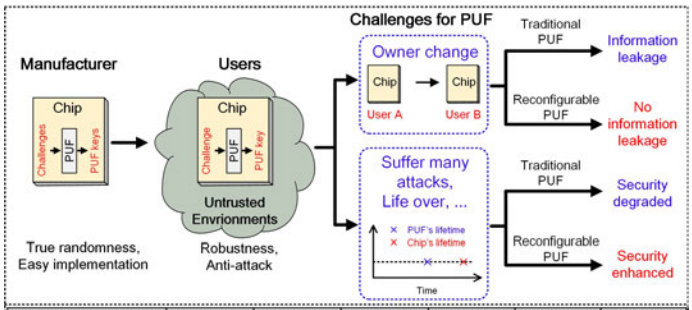
The most unique advantage of the developed RRAM-based PUF is its reconfigurability, indicating that the PUF's security can be enhanced. Fig. 25.2.6 shows the reconfiguration flow chart. The reconfiguration operation is easy to implement without a complex stabilizing method (e.g., masking, baking, and ECC). Regarding the uniqueness of the reconfigured PUF keys, 60 continuous reconfiguration cycles for the same PUF show values close to the ideal value of the reconfigure-HD (μ/σ = 47.29%/6.07%). In addition, the correlation matrix of 60 continuous reconfigured PUF keys demonstrates little correlation between the different reconfigured keys.

Figure 25.2.7 shows the top view of the PUF micrograph with a summary table and 2T2R PUF bit-cell layout. The micrograph also shows the details of the test chip, including an 8Kb PUF array, an 8Kb differential array, a program control circuit, an X/Y decoder, a driver, an SA, timing control, and other peripheral circuits. The area of the PUF array is smaller than peripheral circuits. This is because the PUF array only has 8Kb PUF bit cells, and the X/Y decoder and driver are designed to support a future 16M PUF bit array . In particular, the developed RRAM-based PUF shows the advantages of reconfigurability, easy implementation, high reliability, and good randomness.

Acknowledgements:
This work is supported in part by the NSFC (61674092), MOST of China (2016YFA0201803), Beijing Innovation Center for Future Chip (ICFC), Beijing Municipal Science and Technology Project (Z181100003218001), NSF-CNS-1615774 and SRC-2016-TS-2691, and Tsinghua and National Tsinghua joint project.

References:
[1] Y. Su, et al., "A 1.6pJ/bit 96% Stable Chip-ID Generating Circuit Using Process Variations,", *ISSCC*, pp. 406-407, Feb. 2007.
[2] N. Liu, et al., "OxID: On-chip One-Time Random ID Generation Using Oxide Breakdown," *IEEE Symp. VLSI Circuits*, pp. 231-232, 2010.
[3] S.K Mathew, et al., "A 0.19pJ/b PVT-Variation-Tolerant Hybrid Physically Unclonable Function Circuit for 100% Stable Secure Key Generation in 22nm CMOS," *ISSCC*, pp. 278–280, 2014.
[4] B. Karpinskyy, et al., "Physically Unclonable Function for Secure Key Generation with a Key Error Rate of 2E-38 in 45nm smart-card chips," *ISSCC*, pp. 158-159, 2016.
[5] M. Wu, et al., "A PUF Scheme Using Competing Oxide Rupture with Bit Error Rate Approaching Zero," *ISSCC*, pp. 130-131, 2018.
[6] M. Chang, et al., "An Offset-Tolerant Current-Sampling-Based Sense Amplifier for Sub-100nA-Cell-Current Nonvolatile Memory," *ISSCC*, pp. 206-207, 2011.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE



Entropy Source	SRAM	INV	SA	Photodetectors	Anti- Fuse	RRAM (This work)
Reliability	Low	Low	Low	Medium	High	High
Uniqueness	Low	Medium	Medium	Medium	High	High
Bit cell area	Large	Large	Large	Medium	Small	Small
Power consumption	Small	Small	Small	Large	Large	Medium
ECC, Masking, TMV ?	YES	YES	YES	YES	NO	NO
Re-configurability ?	NO	NO	NO	NO	NO	YES

Figure 25.2.1: Key features and challenges of PUFs in practical applications, showing the advantage of reconfigurable PUF and a brief comparison table of previous works and our implementation.

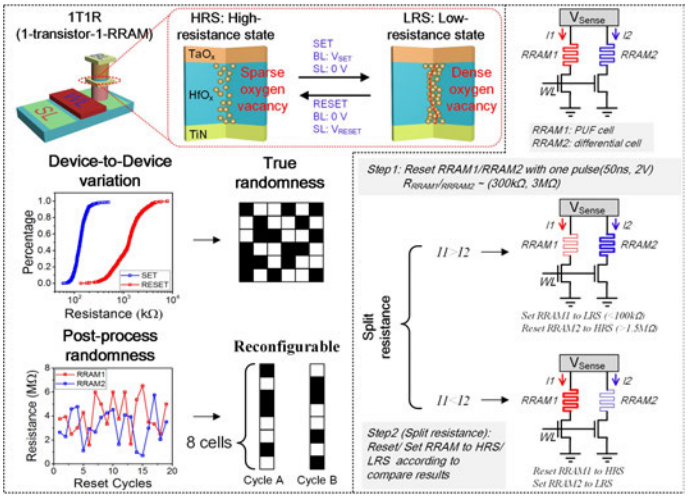


Figure 25.2.2: 1T1R RRAM structure with two typical states and two features of RRAM for PUF application; functional diagram of the split resistance technique to enhance the PUF's reliability.

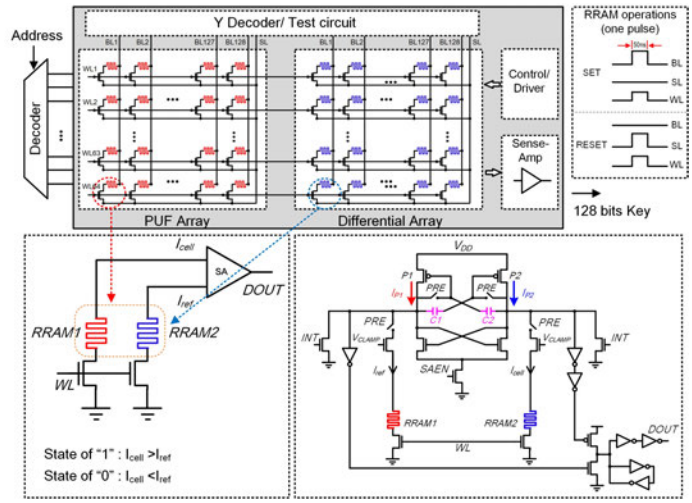
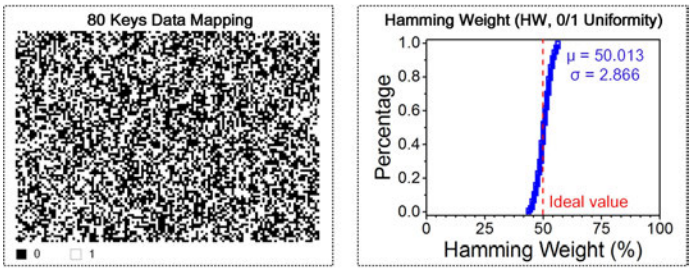


Figure 25.2.3: Architecture of RRAM PUF test chip with a PUF bitcell and a detailed schematic of differential sense amplifier. Each PUF key includes 128b.



NIST Test Results (Native data)

Test name	Stream length	No. of runs	Pass rate %	Avg. P-value	Pass?
Approximate entropy	1024	60	100	0.9973	YES
Block frequency	1024	60	100	0.9951	YES
FFT	1024	60	98.33	0.4983	YES
Frequency	1024	60	100	0.7338	YES
Longest runs	1024	60	98.33	0.5818	YES
Rank	1024	60	100	0.5059	YES
Runs	1024	60	100	0.5127	YES
Serial-1	1024	60	98.33	0.5102	YES
Serial-2	1024	60	100	0.5261	YES

Figure 25.2.4: Spatial data mapping of 80 keys from 20 PUF chips at typical conditions; bias of 1000 PUF keys from 20 test chips; summary table of NIST test results.

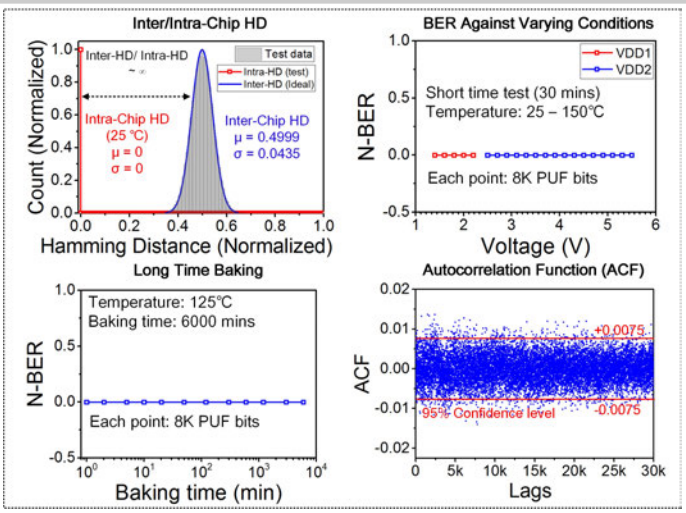


Figure 25.2.5: Measured inter- and intra- Hamming distance; N-BER vs. varying supplying voltage and temperature; N-BER under long-term baking; spatial autocorrelation function.

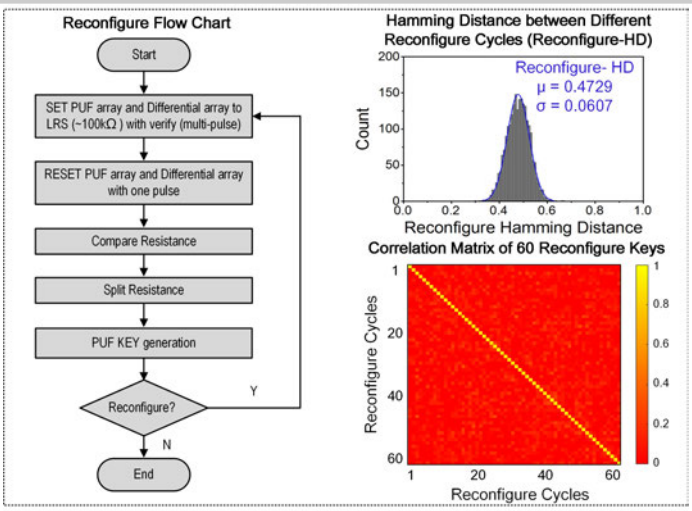
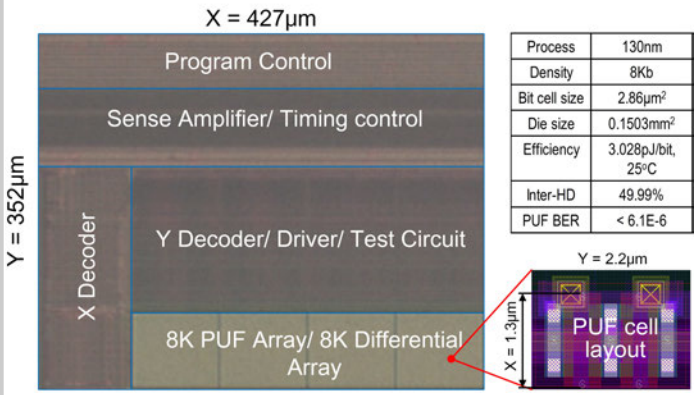


Figure 25.2.6: Flow chart of PUF key reconfiguration process; reconfigure-Hamming distance distribution between 60 reconfigured keys; correlation matrix of 60 reconfigured keys.



Process	130nm
Density	8Kb
Bit cell size	2.86μm^2
Die size	0.1503mm^2
Efficiency	3.028pJ/bit, 25°C
Inter-HD	49.99%
PUF BER	< 6.1E-6

Figure 25.2.7: Top view of chip micrograph with a measurement summary and two PUF-cell layout.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

25.3 A 128b AES Engine with Higher Resistance to Power and Electromagnetic Side-Channel Attacks Enabled by a Security-Aware Integrated All-Digital Low-Dropout Regulator

Arvind Singh[1], Monodeep Kar[2], Sanu Mathew[2], Anand Rajan[2], Vivek De[2], Saibal Mukhopadhyay[1]

[1]Georgia Institute of Technology, Atlanta, GA
[2]Intel, Hillsboro, OR

Side channel attacks (SCA) exploit data-dependent information leakage through power consumption and electromagnetic (EM) emissions from cryptographic engines to uncover secret keys. Integrated inductive voltage regulators (IVR) with a randomized control loop [1] or switching frequency [2], and random voltage dithering [3] have demonstrated improved power side-channel analysis (PSCA) resistance. Simulation studies have shown PSCA resistance via shunt linear regulators [4]. This paper demonstrates improved power and EM SCA resistance of standard (unprotected) 128b AES engines with parallel (P-AES, 128b) and serial (S-AES, 8b) datapaths via an on-die security-aware all-digital series low-dropout (DLDO) regulator, commonly used for fine-grain SoC power management. The security-aware DLDO improves SCA resistance using control-loop induced perturbations in a baseline DLDO, enhanced by a random switching noise injector (SNI) via power stage control and a randomized reference voltage (R-VREF) generator coupled with all-digital clock modulation (ADCM).

The fabricated 130nm CMOS test-chip (Fig. 25.3.7) contains P-AES and S-AES cores powered by a DLDO with 0.5-to-1.22V input and 1.9nF on-die metal-insulator-metal (MIM) load capacitor (Fig. 25.3.1). The DLDO is implemented with a 32-element PMOS array power stage. The feedback loop, sampled at 250MHz, includes a 4b analog-to-digital converter (ADC) and a digital type-III (two zeros and two poles) proportional-integral-derivative (PID) compensator. Measured supply current at the input supply ($V_{IN,DLDO}$) of the chip is a transformed version of the internal on-die supply current of the AES engine measured at local supply node (V_{AES}). The DLDO power stage acts as a low-pass filter with bandwidth dictated by the equivalent resistance of the power stage and output capacitor, thereby attenuating the high-frequency current signatures. The DLDO control loop induces frequency-dependent small signal perturbations dictated by the loop delay and zero/pole locations. PID coefficients (K_P, K_I, K_D) can be modified at runtime to reduce information leakage, while modulating the transient response of the DLDO to load/reference changes.

The SNI generates programmable pulses in each DLDO clock cycle to disable all devices in the PMOS array for a fraction of the clock cycle, thus disconnecting an information leakage path from V_{AES} to $V_{IN,DLDO}$ (Fig. 25.3.2). The SNI consists of nine pulse generators – each generating a pulse using four of the nine phases of a 9-stage current-starved differential-delay-cell-based voltage-controlled-oscillator (VCO). A maximum length 4b linear-feedback shift register (LFSR) generates a signal to select one of the nine pulses randomly. As the SNI is activated, the output decoupling capacitor supplies the current required for AES operation when PMOS's are all off. When the PMOS's are enabled again, the DLDO feedback loop responds to the voltage droop at V_{AES}, thus adding amplitude noise at V_{AES} (and $V_{IN,DLDO}$) even when the power stage PMOS's are ON. In addition, since the AES and DLDO clocks are not synchronized, current signatures at $V_{IN,DLDO}$ and V_{AES} are de-synchronized. The R-VREF circuit randomly (with 4b LFSR – LFSR3) selects a digital word for the reference voltage (V_{REF}) from pre-loaded registers and induces an output voltage transition in programmable regular intervals, thus randomizing the AES supply voltage (V_{AES}) (Fig. 25.3.2). Reference registers are pre-loaded with V_{REF} words such that the mean is equal to the target operating voltage, while the injected randomness is governed by the standard deviation. Unlike voltage dithering [3], where V_{AES} only changes between encryptions, the R-VREF runs at high frequency (DLDO clock/16) randomizing V_{AES} even during an encryption. An on-die ADCM circuit uses critical-path replicas to modulate edges of the AES clock and ensure correct operation under transient supply droops and variations induced by both SNI and R-VREF. ADCM utilizes the random noise in V_{AES} to inject additional timing and amplitude randomizations in the current signatures [3].

A test-control switch enables AES cores to operate in standalone (externally powered) or DLDO-powered modes (Fig. 25.3.1). Test vector leakage assessment (TVLA) and correlation power analysis (CPA) are performed for power and EM signatures in the time and frequency domains. Power signatures are measured as differential voltages across a 1Ω series resistor at V_{AES} (standalone AES) and $V_{IN,DLDO}$ (DLDO-AES) pins (Fig. 25.3.3). An external trigger is used to start AES encryptions. Measured patterns at $V_{IN,DLDO}$ show that the power signature, although distinguishable, is attenuated by ~5.6× compared to V_{AES}. When SNI and R-VREF are enabled, P-AES rounds are no longer distinguishable at $V_{IN,DLDO}$. EM signatures, though much weaker, show distinct peaks during the P-AES rounds. The power/EM signatures are filtered with narrow bandpass filters (10MHz) ranging from 5MHz to 305MHz to remove out-of-band noise. Filtered signatures are aligned with cross-correlation to remove the initial delay caused by use of an external trigger.

TVLA on P-AES shows high information leakage for the standalone AES (baseline) which reduces significantly for DLDO-AES (Fig. 25.3.4). When K_P=K_D=0 for the PID controller (integral compensator with K_I=1/32), TVLA leakage is maximum, while it is reduced for higher K_P and K_I values, which leads to underdamped response. Similarly, a faster R-VREF clock and larger SNI pulse width leads to more amplitude and timing (with ADCM) randomization, producing lower TVLA leakages. Although SNI increases noise at $V_{AES,}$ correct operation is ensured by ADCM with only 9.3% performance loss. The R-VREF generates V_{AES} levels with a mean close to the nominal value to ensure minimal (-1.1%) loss in the average throughput of the AES core.

CPA analysis on byte 9 (highest leaking byte) of the P-AES 128b key with 5 million measurement shows high correlation ratio (CR), defined as the ratio of absolute correlation for correct key guess and 2nd highest correlation for successful attack, across all bands for standalone AES. The CR value reduces for DLDO-AES with most of the bands still leaking (Fig. 25.3.5). With the R-VREF and SNI, CR reduces to 1.3 with only a few bands leaking. CPA in the frequency domain shows that byte 9 of the key in standalone P-AES can be uncovered with only 400 (minimum-traces-to-disclose, MTD) measurements, while 3.6 million measurements are required for DLDO-AES with the R-VREF and SNI. Note, when AES is powered externally through V_{OUT} (DLDO, R-VREF, SNI off) to emulate a design with an "on" power gate (test-control switch) with 1.9nF decap on the ungated input rail and 0.4nF on the virtual rail, the MTD increases to only 2200. All key bytes were extracted with up to 10 million measurements for frequency-domain CPA (Fig. 25.3.6). MTD for an 80% success rate (SR) (i.e. to reveal 13/16 key bytes) for baseline P-AES is 1900 (CPA) and 50K (CEMA), which improves to 8M (CPA) and 6.8M (CEMA) for DLDO-AES with the SNI and R-VREF. MTD considering both CPA and CEMA increases by 3579× for P-AES and by 2182× for S-AES. Including overheads due to the power stage, controller, and R-VREF/SNI, the DLDO-AES shows 36.9% higher area and 10.4% lower performance than the standalone AES, and 68% power-efficiency, while driving 40mA @ 0.84V.

A power injection attack (PIA) with a voltage glitch [2] is performed at V_{CTRL} (control supply that powers the DLDO loop, R-VREF, SNI, and LFSRs), and $V_{IN,DLDO}$. The PIA at V_{CTRL} introduced instability/noise at V_{AES} forcing random clock-skipping from ADCM, which de-synchronizes power/EM signatures [with 100mV glitch TVLA reduced to 5.8 (power) and 7.7 (EM)] but degrades encryption throughput. A PIA on $V_{IN,DLDO}$ can cause all power PMOS's to be always ON, rendering DLDO and R-VREF ineffective, but SNI remains effective; MTD of 8.4M (power) and 6.0M (EM) are observed for an 80mV glitch. The security of DLDO-AES can be further improved by using a TRNG to reduce predictability/repeatability of the LFSR and adding input sensing circuits at $V_{IN,DLDO}$ and V_{CTRL} to inhibit a PIA [2].

Acknowledgements:
This material is based on work supported in part by Intel Corporation and Semiconductor Research Corporation through TxACE (#2810.002 and #2712.002).

References:
[1] M. Kar, et al., "Reducing Power Side-Channel Information Leakage of AES Engines Using Fully Integrated Inductive Voltage Regulator," *IEEE JSSC*, vol. 53, no. 8, pp. 2399-2414, 2018.
[2] W. Yang, et al., "An Enhanced-Security Buck DC-DC Converter with True-Random-Number-Based Pseudo Hysteresis Controller for Internet-of-Everything (IoE) Devices," *ISSCC*, pp. 126-128, 2018.
[3] A. Singh, et al., "Improved Power Side Channel Attack Resistance of a 128-bit AES Engine with Random Fast Voltage Dithering," *ESSCIRC*, pp. 51-54, 2017.
[4] D. Das, et al., "ASNI: Attenuated Signature Noise Injection for Low-Overhead Power Side-Channel Attack Immunity," *IEEE TCAS-I*, vol. 65, no. 10, pp. 3300-3311, 2018.
[5] S. Lu, et al., "1.32GHz High-Throughput Charge-Recovery AES Core with Resistance to DPA Attacks," *IEEE Symp. on VLSI Circuits*, pp. C246-C247, 2015.
[6] C. Tokunaga, et al., "Secure AES Engine with a local Switched-Capacitor Current Equalizer," *ISSCC*, pp. 64-65, 2009.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

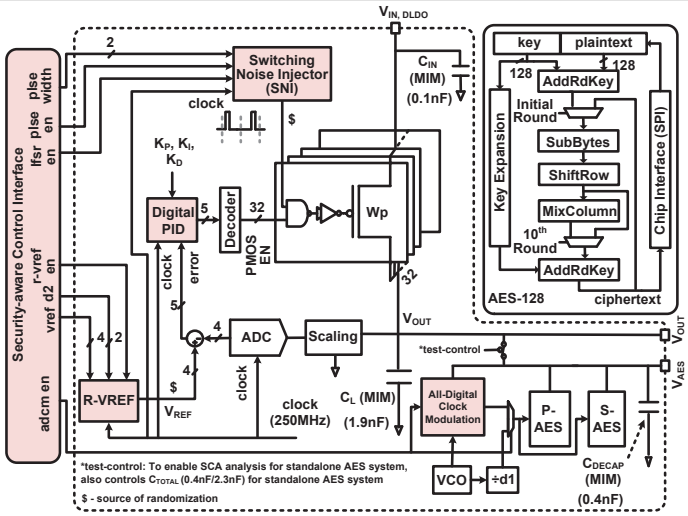


Figure 25.3.1: Architecture of security-aware digital LDO. Circuit blocks to enhance resistance to side channel attacks are shaded.

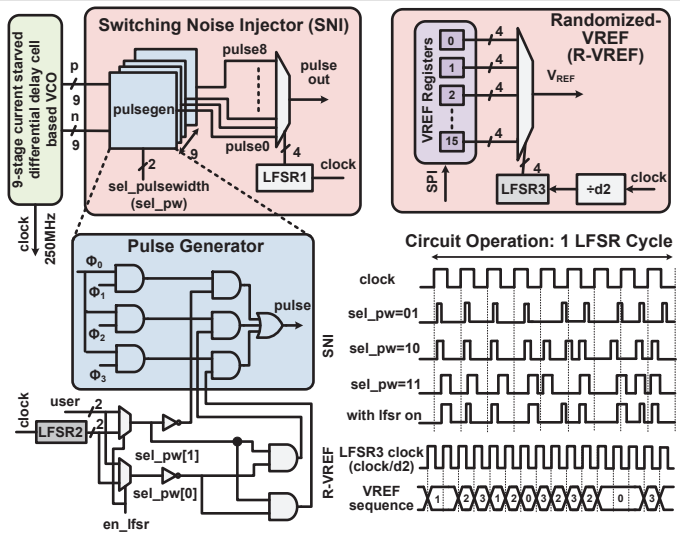


Figure 25.3.2: Circuit techniques and circuit operation for SNI and R-VREF circuits.

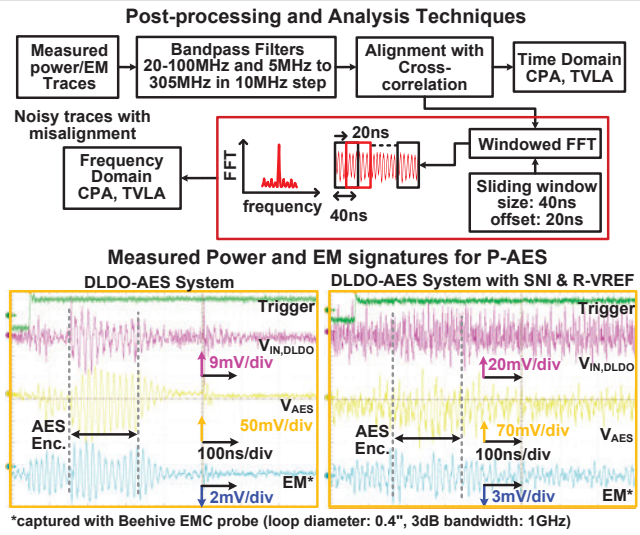
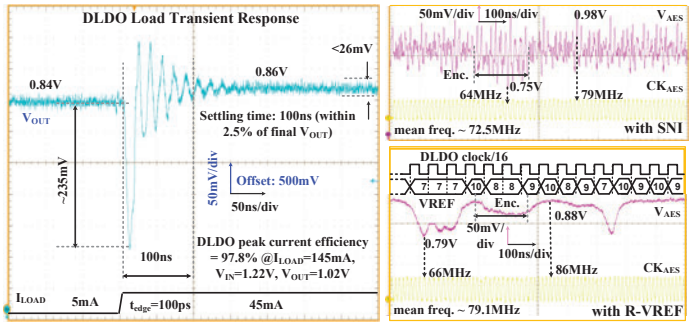


Figure 25.3.3: Filtering and alignment techniques for the measured signatures for SCA in time and frequency domains.



TVLA Analysis for Baseline and DLDO-AES system with SNI & R-VREF for P-AES

Circuit Technique	Baseline AES	DLDO-AES							
		PID Controller (K_P, K_I, K_D)			LFSR clock for R-VREF		Pulse width for SNI		
Configurations	-	22/32, 24/32, 9/32*	10/32, 7/32, 2/32	0, 1/32, 0	DIV16*	DIV64	$\frac{T_{clock}}{9}$	$\frac{2T_{clock}}{9}$	$\frac{3T_{clock}}{9}$ *
TVLA Peak	258	23.4	25.2	32.9	13.1	17.4	14	13.7	11.9

* configuration used for final SCA-resistant design

Figure 25.3.4: Load transient response of DLDO and measured SNI and R-VREF waveforms with ADCM and TVLA analysis with different configurations.

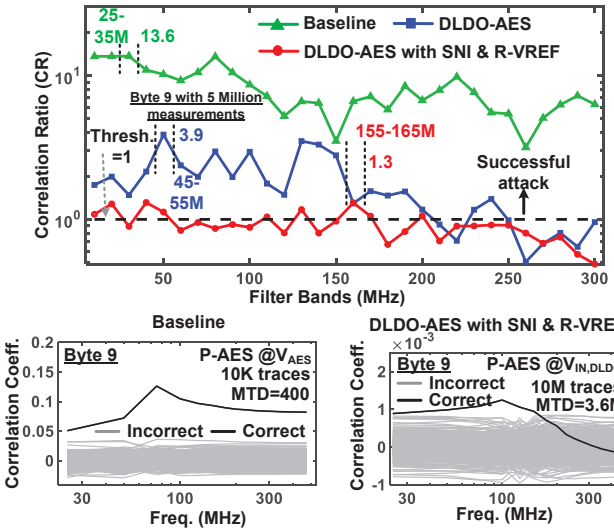
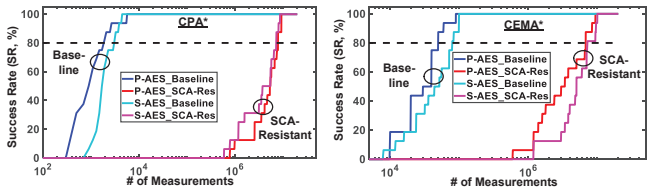


Figure 25.3.5: Correlation ratio plotted against filter bands and correlation plots in frequency domain showing successful key recovery.



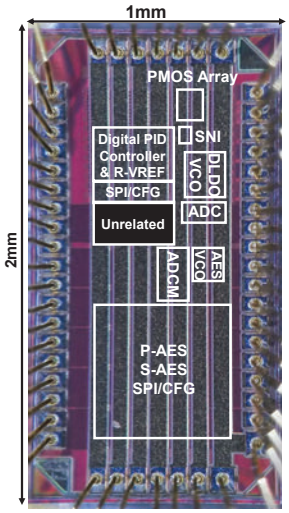
*powerModel: P-AES- hamming distance between 10th round input and ciphertext, S-AES- hamming weight for 1st SBOX output

Comparison with Existing Works

Metric		This Work	JSSC'18 [1]	ESSCIRC'17 [3]	VLSI'15 [5]	ISSCC'09 [6]	ISSCC'18 [2]
Countermeasure Technique		On-chip Digital Low Dropout Regulator with SNI & R-VREF	Integrated Buck Regulator	Random Fast Voltage Dithering	Charge Recovery Logic	Switch Capacitor Current Equalizer	Buck Regulator
Technology		130nm	130nm	130nm	65nm	130nm	55nm
AES power[a]		10.9mW @ 80MHz, 0.84V	10.5mW @40MHz	13.1mW @49.7MHz	138mW @1.32GHz	33mW @100MHz	No AES/other Encryption Engine present
Design Overheads[a]	Area	36.9%[b]	1%[c]	5.5%[c]	25%	33%	
	Power	32%[b]	5%[c]	5.9%[c]	30%	20%	
	Perf.	10.4%	3.33%	1.8%	0%	50%	
# of Measurements		10M	500K	500K	1M	10M	N/A[e] (No SCA)
SCA Analysis[d]	Time/Freq. Domain	Time, Freq.	Time, Freq.	Time	Time	Time	
	MTD for 80% SR	6.8M (3579×)[f]	>500K (100×)[g]	>500K (33×)[g]	940K (251×)[h]	>10M (2500×)[i]	
	Attack Mode	Power, EM	Power, EM	Power	Power	Power	

[a]With respect to parallel AES (P-AES). [b]Includes DLDO regulator area/power. [c]Doesn't include regulator area/power. [d]For P-AES. [e]Relative correlation based analysis. [f]Minimum of CPA MTD and CEMA MTD. [g]SCA performed for only 1 byte. [h]MTD to disclose all bytes. [i]None of the bytes could be disclosed.

Figure 25.3.6: Success rate for CPA/CEMA and comparison with existing works.



Test-chip	Process		130nm CMOS
	Package		LCC52
Digital LDO with SNI and R-VREF	Active Area		0.078mm^2
	V_{IN} [V]		0.5 – 1.22
	V_{OUT} [V]		0.35 – 1.17
	Current [mA]		5 – 145
	Load Cap [nF]		1.9 (0.44mm^2)
	Compensator		Type III
	ADC		4-bit delay line based
	Peak Current Eff. (DLDO)		97.8% @145mA, V_{IN}/V_{OUT}=1.22V/1.02V
ADCM	Clock [MHz]		600
	Op. range [V]		0.73 – 1.25
	Area		0.024mm^2
AES-128	Perf.	P-AES	80MHz/0.84V/10.9mW
		S-AES	80MHz/0.84V/2mW
	Arch.	P-AES	128b (datapath)
		S-AES	8b (datapath)
	Decap [nF]		0.4 (0.088mm^2)

Figure 25.3.7: Chip micrograph and design details.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 26 Overview:
Frequency Generation & Interference Mitigation
RF SUBCOMMITTEE

Session Chair: ***Ramesh Harjani***
University of Minnesota, Minneapolis, MN

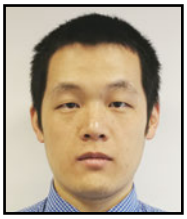
Associate Chair: ***Andrea Mazzanti***
Università degli Studi di Pavia, Pavia, Italy

Subcommittee Chair: Piet Wambacq, imec, Belgium

This session covers the latest advancements in frequency synthesis for radars, frequency sources, and RX techniques for interference mitigation. The first presentation in the session demonstrates an ultra-wide-band chirping FMCW modulator with a 16GHz synthesizer in 28nm CMOS. The next two presentations cover recent innovations in millimeter-wave CMOS frequency sources for 5G, achieving low phase noise and wide tuning range and exploiting multi-resonance LC tanks. The fourth paper demonstrates strong interference rejection by introducing envelope-based dynamic biasing in a 2.4GHz CMOS receiver. The last paper in the session proposes a low-noise oscillator for IoT operating down to 0.1V supply.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

3:15 PM



26.1 A Self-Calibrated 16GHz Subsampling-PLL-Based 30µs Fast Chirp FMCW Modulator with 1.5GHz Bandwidth and 100kHz rms Error

Q. Shi, imec, Heverlee, Belgium

In Paper 26.1, imec and Sony Semiconductors present a 16GHz FMCW radar modulator in 28nm CMOS. The modulator is based on a dividerless DTC subsampling PLL architecture and reaches a 30µs sawtooth chirping speed with a 1.5GHz chirping bandwidth. The rms frequency error is below 103kHz, which is less than 0.007% of the chirp bandwidth.

3:45 PM

26.2 A 0.08mm^2 25.5-to-29.9GHz Multi-Resonant-RLCM-Tank VCO Using a Single-Turn Multi-Tap Inductor and CM-Only Capacitors Achieving 191.6dBc/Hz FoM and 130kHz 1/f^3 PN Corner

Y. Chen, University of Macau, Macau, China

In Paper 26.2, the University of Macau and Instituto Superior Tecnico of the University of Lisboa propose a 25.5-to-29.9GHz VCO based on an RLCM tank with resonances at 1st, 2nd, and 3rd harmonics and tunable with only common-mode capacitors. The VCO, in 65nm CMOS, achieves up to 191.6dBc/Hz FoM@1MHz and demonstrates 1/f^3 phase noise corner between 130kHz and 230kHz.

4:15 PM

26.3 A 25-to-38GHz, 195dB FoM$_T$ LC QVCO in 65nm LP CMOS Using a 4-Port Dual-Mode Resonator for 5G Radios

A. Bhat, Indian Institute of Technology Madras, Chennai, India

In Paper 26.3, the Indian Institute of Technology, Madras, shows a wide-tuning-range millimeter-wave quadrature VCO exploiting a dual-mode resonator realized with 4-port coupled inductors. The VCO, in 65nm CMOS, is tunable from 25 to 38GHz with 17.5-to-21.6mW power dissipation from a 0.65V supply. The phase noise at 3MHz offset ranges from -121 to -112dBc/Hz, corresponding to a 181-to-187 dBc/Hz figure of merit.

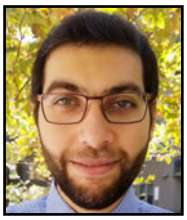
4:30 PM

26.4 A 2.4GHz 65nm CMOS Mixer-First Receiver Using 4-Stage Cascaded Inverter-Based Envelope-Biased LNAs Achieving 66dB In-Band Interference Tolerance and -83dBm Sensitivity

D. Ye, Fudan University, Shanghai, China

In Paper 26.4, Fudan University and the University of Washington present a 2.4GHz mixer-first receiver for IoT applications where the interference resilience is improved through an inverter-based LNA exploiting dynamic input biasing generated from the input envelope. With 2.05mA current consumption, the 65nm CMOS receiver achieves 66dB in-band interference tolerance and -83dBm sensitivity with 307kb/s OOK modulation.

5:00 PM



26.5 A 0.1-to-0.2V Transformer-Based Switched-Mode Folded DCO in 22nm FDSOI with Active Step-Down Impedance Achieving 197dBc/Hz Peak FoM and 40MHz/V Frequency Pushing

O. El-Aassar, University of California, San Diego, La Jolla, CA

In Paper 26.5, the University of California, San Diego, proposes an LC DCO based on a folded, transformer-based, circuit topology to improve phase noise performance at low supply voltages. Implemented in 22nm FDSOI, the circuit oscillates from 4.2 to 5GHz with a peak FoM of 197dBc/Hz, -142.1dBc/Hz PN at 10MHz, 40MHz/V pushing from a 0.15V supply. The oscillator can operate down to 0.1V V_{dd}.

26.1 A Self-Calibrated 16GHz Subsampling-PLL-Based 30µs Fast Chirp FMCW Modulator with 1.5GHz Bandwidth and 100kHz rms Error

Qixian Shi[1], Keigo Bunsen[2], Nereo Markulic[1], Jan Craninckx[1]

[1]imec, Heverlee, Belgium
[2]Sony Semiconductor Solutions, Atsugi, Japan

Frequency-modulated continuous-wave (FMCW) radars are critical for autonomous-driving applications. Sawtooth waveforms with short chirp time (<50µs) are desired to eliminate the speed-range ambiguity. The chirp signal needs to be linear to avoid spurious tones in the receiver baseband, which result in false detections. Furthermore, a wide chirping bandwidth (BW_{chirp} > 4GHz) improves the range resolution to 3.75cm. To meet these requirements, fractional-N PLLs with two-point-modulation (TPM) scheme are presented in [1-3] with a relative rms FM error of 0.05% when normalized to the chirping bandwidth. In conventional FMCW PLLs, the lowpass signal injection is implemented through division ratio modulation in the PLL feedback path. However, the divider power consumption and noise are not negligible. In [4], a subsampling GMSK modulator improves the EVM. By using a high-resolution DTC, the divider quantization noise contribution to the PM/FM error is eliminated. In this work, the frequency modulation bandwidth is extended to GHz range and a 16GHz subsampling PLL (SS-PLL) is used for fast-chirp generation. Thanks to the TPM technique and self-calibration, this PLL takes 30µs to chirp a sawtooth waveform with a 1.5GHz BW (~9.5% of the carrier frequency). The rms FM error is 100kHz, which is 0.007% of BW_{chirp}.

The diagram of the subsampling PLL is shown in Fig. 26.1.1. The FMCW control logic defines the chirping waveform and generates the frequency control word (FCW). In the lowpass path, the FCW is integrated to get the phase modulation signal. A 10b DTC delays the sampling edge based on the phase signal to ensure a near-zero-crossing detection of the VCO output during chirping. In parallel, the highpass FM signal is sent towards the VCO. With proper scaling and digital pre-distortion (DPD), the VDAC generates an FM chirping voltage ($V_{tune,FM}$) that periodically sweeps the biasing voltage of Var_2. If the nonlinearity and mismatch of the FM modulation path are well calibrated, the VCO frequency ramping is linear and the sub-sampler output is near the zero crossing. This leads to a zero-mean net current in the transconductor, and $V_{tune,PLL}$ remains constant over chirping cycles. The subsampling path is thus undisturbed by the direct FM path. This makes the chirp speed independent of the PLL bandwidth and allows fast-sawtooth modulation. For an SS-PLL modulator, the loop time resolution is the DTC LSB (~0.5ps). Compared with a delta-sigma-modulator-based fractional-N PLL, where the time resolution is one VCO period (~62.5ps), a DTC-based modulator injects much less quantization error, which improves the output FM error. Also shown in Fig. 26.1.1 is an FLL that ensures initial frequency acquisition and foreground calibration. It is disabled during chirping.

The VCO and FDAC schematics are shown in Fig. 26.1.2(a). The VCO topology is a current-biased Class-B NMOS type. A tail resonator suppresses the flicker-noise upconversion from the cross-coupled pair. In contrast to FDACs based on switched-capacitor arrays [2,5], an analog varactor driven by a voltage DAC adds less parasitic capacitance, which allows a compact layout and thus wider chirping range. The IDAC and varactor nonlinearity are calibrated by FDAC code pre-distortion. As the VDAC parasitics are not critical, large transistors are used to improve matching. The VDAC is composed of a tunable load resistor and a 13b current steering IDAC (segmented in 8b binary for the LSB and 5b thermometer-encoded for the MSB). With a nominal FDAC LSB of 195kHz/b, the maximal BW_{chirp} of 1.5GHz is reached.

Due to the absence of a divider chain, an SS-PLL usually has a narrower acquisition range compared to a conventional delta-sigma fractional-N PLL. Without proper pre-distortion at the FDAC path, the gain mismatch and varactor nonlinearity induce sampling outside of the acquisition range thus making the PLL lose lock. A foreground self-calibration must be executed before chirping starts. An external DC voltmeter is used in this prototype for simplicity. For the nominal K_{VCO2} gain calibration, the VDAC code is set to its minimal and maximal values, respectively. The corresponding oscillation frequency is determined by locking the PLL while varying the division-chain ratio to force $V_{tune,PLL}$ to settle at a mid-supply value (V_0). This frequency-tuning range (TR) is the maximal chirping bandwidth, and the nominal K_{VCO2} is calculated as the TR divided by the FDAC code length. For the FDAC nonlinearity calibration, the PLL is locked at the targeted intermediate frequencies by varying the fractional division ratio accordingly. The raw FDAC code is calculated based on the calibrated K_{VCO2}, and an offset value is added to settle $V_{tune,PLL}$ close to V_0. This calibration obtains a 64-element piece-wise-linear approximation of the code-to-frequency characteristic, and the offset values are stored in a look-up-table (LuT). During FM chirping, the chip further employs a fine online calibration of the FDAC nonlinearity. Namely, the nonlinear digital-to-frequency conversion induces instantaneous phase error in the following sampling event. The phase error is digitized by a current-sign extractor, see Fig. 26.1.2(b). Within the sample pulser opening window, the source follower pre-charges C_1 and C_2 to the transient $V_{tune,PLL}$. When the pulser window is closed, the switch to C_1 opens and the same current (i_{loop}) is then steered to further (dis)charge C_2 while C_1 keeps its voltage as a reference node. After a few nanoseconds, a voltage difference (ΔV) is set up for the comparator to determine the sign of the transient current. Derivative of the error stream is tracked in correlation with the FM modulation codes to continuously track the nonlinearity. The loop updates the LuT coefficients such that the phase error remains a zero-mean stream during the chirping period. Because of the steep FM slope, the SS-PLL could lose lock when starting with a blank LuT causing failure in the online tracking. Thus, a proper foreground calibration is required.

The FMCW SS-PLL is fabricated in a 28nm CMOS process. Figure 26.1.7 shows the die micrograph. The core area is below 1mm^2 excluding the test circuits. During chirping operation, the SS-PLL modulator consumes 44mW, and the details of sub-blocks are concluded in Fig. 26.1.3(a). The PLL discrete-tuning frequency ranges from 14.7 to 17.2GHz. Figure 26.1.3(b) shows the VCO phase noise at 16GHz by locking the PLL with a 100kHz BW_{loop}. At a 1MHz offset, the phase noise is -106.3dBc/Hz, which is equivalent to -92.6dBc/Hz when normalized to 79GHz. Figure 26.1.3(c) shows the divided-by-16 spectrum in the SS-PLL mode with a 1.5GHz BW_{chirp}. The chirp time is 30µs with a return time of 2µs. The idle time at both turn-around points (TAP) is 2µs. To access the linearity performance, the divide-by-16 output is analyzed with an oscilloscope. Figure 26.1.4 shows the divided waveform and FM error. Excluding the 1% chirp time at the TAP, the rms and peak FM errors in subsampling mode is 6.45×16=103kHz (~0.007% of BW_{chirp}) and 26×16=416kHz (~0.027% of BW_{chirp}), respectively. In comparison, when the PLL chirps under the conventional delta-sigma fractional-N PLL with the subsampling path off, the rms FM error is 564kHz. Figure 26.1.5 shows the $V_{tune,PLL}$ and $V_{tune,FM}$ transient waveforms during chirping with the 1.5GHz BW_{chirp}. With the foreground offline-calibrated LuT, the $V_{tune,PLL}$ experiences a 100mV ripple. After the fine online self-calibration, $V_{tune,PLL}$ stays constant, which indicates that the subsampling loop is undisturbed from the FM modulation as intended. The chip performance and the comparison with other prior-art FMCW PLLs are concluded in Fig. 26.1.6. Thanks to the DTC-based subsampling technique, this FMCW modulator achieves the widest BW_{chirp} in 30µs with a 10× improvement in relative rms FM error compared to prior-art in Fig. 26.1.6.

Acknowledgments:
The authors thank Panasonic for their support and IClink for the digital synthesis.

References:
[1] B. Ginsburg et al., "A Multimode 76-to-81GHz Automotive Radar Transceiver with Autonomous Monitoring," *ISSCC,* pp. 158-160, Feb. 2018.
[2] D. Cherniak et al., "A 23GHz Low-Phase-Noise Digital Bang-Bang PLL for Fast Triangular and Saw-Tooth Chirp Modulation," *ISSCC,* pp. 248-250, Feb. 2018.
[3] J. Vovnoboy et al., "A Dual-Loop Synthesizer with Fast Frequency Modulation Ability for 77/79 GHz FMCW Automotive Radar Applications," *IEEE JSSC,* vol. 53, no. 5, pp. 1328-1337, May 2018.
[4] N. Markulic et al., "A DTC-Based Subsampling PLL Capable of Self-Calibrated Fractional Synthesis and Two-Point Modulation," *IEEE JSSC,* vol. 51, no. 12, pp. 3078-3092, Dec. 2016.
[5] W. Wu et al., "A 56.4-to-63.4 GHz Multi-Rate All-Digital Fractional-N PLL for FMCW Radar Applications in 65 nm CMOS," *IEEE JSSC,* vol. 49, no. 5, pp. 1081-1096, May 2014.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

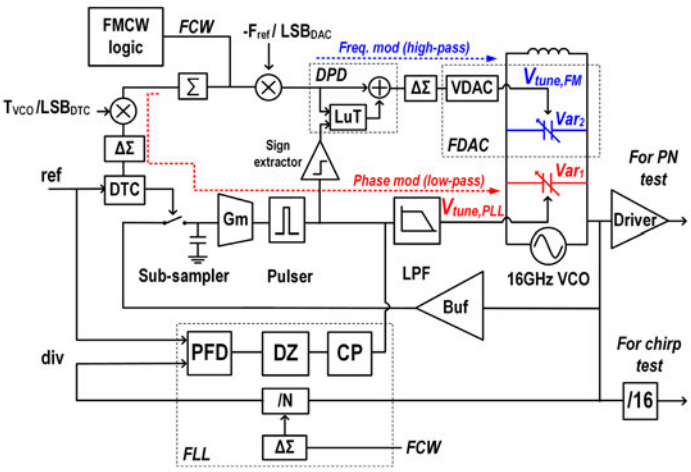


Figure 26.1.1: The block diagram of a subsampling PLL with a TPM scheme for a fast chirp.

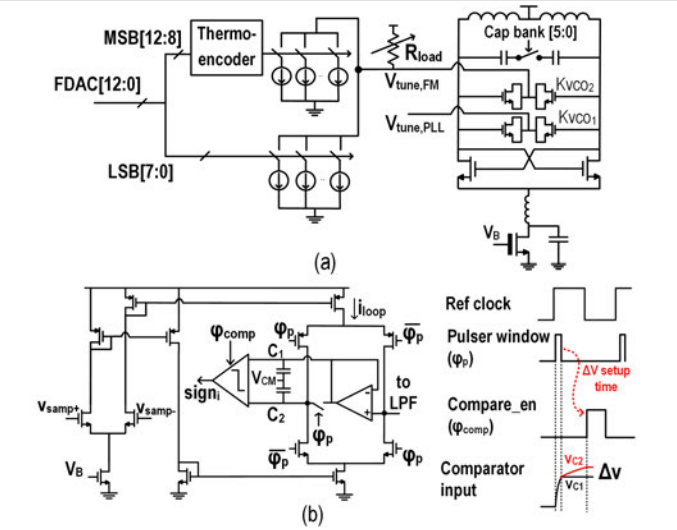


Figure 26.1.2: (a) The 16GHz VCO and the FDAC schematics. (b) Sign extractor embedded in the transconductor and signal timing waveforms.

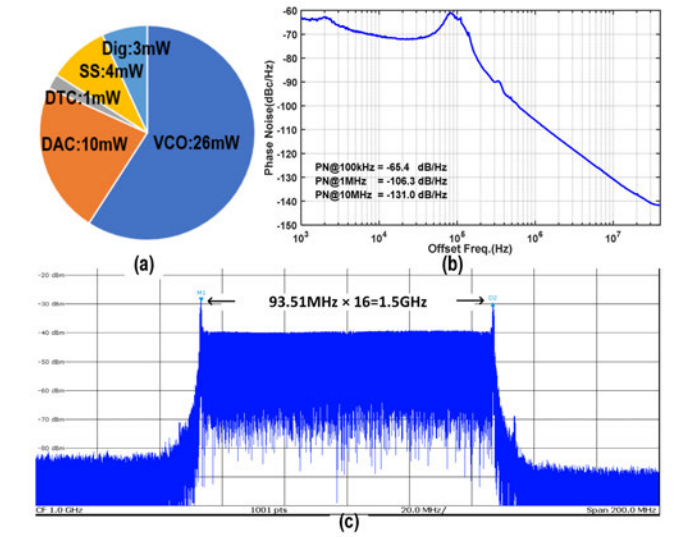


Figure 26.1.3: (a) Power consumption details. (b) 16GHz VCO phase noise. (c) Chirping power spectrum at the divided output under subsampling mode.

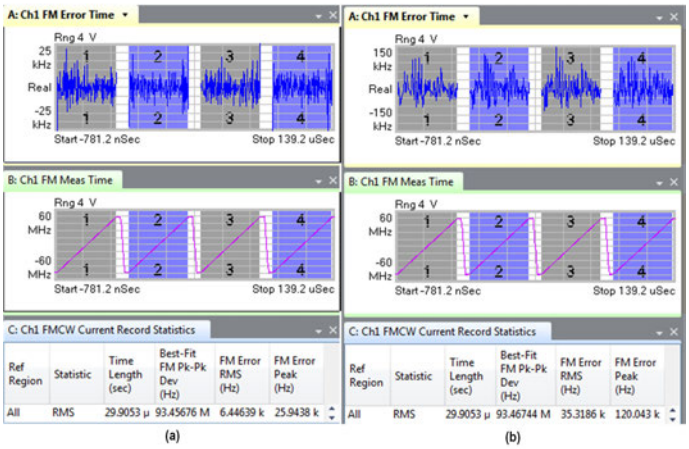


Figure 26.1.4: Chirping waveform and FM error at the div/16 output with a 1.5GHz BW and 30μs Tchirp: (a) with SS-PLL and (b) with DSM frac-N PLL.

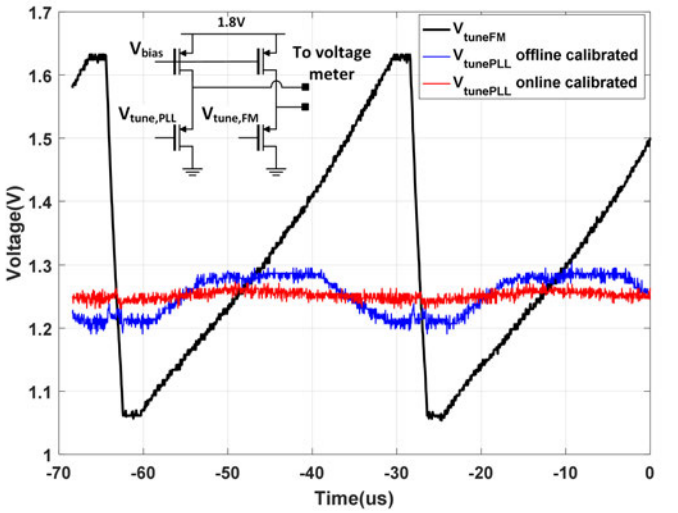


Figure 26.1.5: Waveforms of FM and PLL V_{tune} during chirping with the 1.5GHz BW_{chirp}, measured with a DC level shifter under a 1.8V supply.

	This work	[1]	[2]	[3]	[5]
Architecture	TPM/SS-PLL	TPM/fracN-PLL	TPM/BB-PLL	Frac-N PLL	TPM/AD-PLL
Fast chirp (<50 us)	Y	Y	Y	Y	N
Ref freq. [MHz]	80	1000[1]	52	125	40
Center freq. [GHz]	14.7~17.2	19~20.25	20.4~24.6	18.8~20.8	56.4~63.4
BW_{chirp}/T_{ramp} [GHz/μs] : rms FM error [kHz]	1.5/30: 103 1.5/100:46	1/40: 600[2]	0.2/1.2: 124	1.25/50: 800[3]	1/210: 384
BW_{chirp}/Center freq. [%]	9.375	5	0.9	6.25	1.67
rms FM error/BW_{chirp} [%]	0.003~0.007	0.06[2]	0.05/0.21	0.064	0.038
Phase Noise[4] [dBc/Hz]	-92.6	-91~-94	-92	-97	-88
Waveform	Sawtooth	Sawtooth	Sawtooth	Sawtooth	Triangle
Power (mW)	44	>120	20	240	48
Technology	28nm CMOS	40nm CMOS	65nm CMOS	130nm SiGe	65nm CMOS

1: A clean-up PLL is required to generate 1GHz reference clock for FMCW synthesizer.
2: Only maximal error reported.
3: With 5% chirping time excluded at turning points.
4: Normalized to 79GHz carrier at 1MHz offset.

Figure 26.1.6: FMCW SS-PLL performance and comparison with prior-art designs.

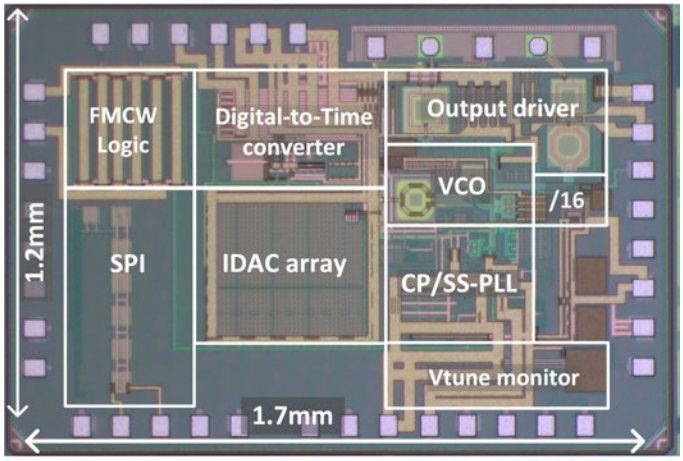


Figure 26.1.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

26.2 A 0.08mm² 25.5-to-29.9GHz Multi-Resonant-RLCM-Tank VCO Using a Single-Turn Multi-Tap Inductor and CM-Only Capacitors Achieving 191.6dBc/Hz FoM and 130kHz 1/f³ PN Corner

Hao Guo[1], Yong Chen[1], Pui-In Mak[1], Rui P. Martins[1,2]

[1]University of Macau, Macau, China
[2]Instituto Superior Tecnico/University of Lisboa, Lisbon, Portugal

VCO designs have evolved from single-resonant LC-tank VCOs to the recent multi-resonant RLCM (Resistor-Inductor-Capacitor-Mutual-inductance) tank VCOs that allow reshaping of VCO phase-noise (PN) impulse sensitivity function (ISF). Two recent RF VCOs [1,2] exploited the 2nd-harmonic resonance to impede the flicker-noise upconversion, achieving an $FoM_{@1MHz}$ of up to 195.4dBc/Hz (Fig. 26.2.1). Specifically, the RLCM tank in [1] tailors a multi-turn inductor with a positive mutual coupling factor (k>0) to generate an implicit common-mode (CM) resonance at 2× the oscillation frequency (F_{OSC}). Yet, the CM resonance has a quality factor (Q) that is ~50% of its differential-mode (DM) counterpart, due to partial magnetic-flux cancellation within the multi-turn inductor. In [2], the transformer-based RLCM tank offers two intrinsically high-Q resonances at F_{OSC} and $2F_{OSC}$ that can effectively suppress PN in the 1/f²-to-1/f³ PN regions. Still, when migrating to mm-waves both [1] and [2] face their respective challenges. A large fixed DM capacitor (C_{DM}) for fulfilling the DM-to-CM capacitance ratio of [1] largely limits the tuning range (TR) at mm-wave. For [2], its high-ratio multi-turn transformer cannot maintain its Q at mm-waves.

This paper proposes an mm-wave VCO using a multi-resonant RLCM tank (Fig. 26.2.1) that enables compact and high-FoM implementation. Specifically, a single-turn multi-tap inductor with k<0 and CM-only tunable capacitors, without the constraint of the DM-to-CM capacitance ratio, enable high-Q high-impedance resonances at F_{OSC}, $2F_{OSC}$, and $3F_{OSC}$.

The 1st-, 2nd-, and 3rd-harmonic voltages can be properly combined to nullify the DC value of the flicker-noise effective ISF ($\Gamma_{1/f,eff,dc}$). As shown in Fig. 26.2.2 (left), A_V and θ_V denote the gain and phase shift, respectively, of the 1st-harmonic voltage traveling from the drain (V_D) to the gate (V_G) of the MOS transconductor (G_M). For a 1-port RLCM tank, [A_V=1, θ_V=0] can ideally yield $\Gamma_{1/f,eff,dc}$=0. For a 2-port RLCM tank [3-5], we have [A_V>1, θ_V<0] that implies more amplitude and phase controllability. Specifically, we can obtain the normalized ISF at V_D: $\Gamma(t)=V'_D(t)/\max(|V'_D(t)|)$ and the noise modulation function: $m(t)=G_M(t)/\max[G_M(t)]$. In one oscillation cycle, G_M in the cutoff, linear, and saturation regions can be defined as 0, βV_D, and $\beta(V_G-V_{th})$, respectively. $\Gamma(t)$ is only a function of the amplitude ratios and phase differences of the 2nd-to-1st harmonics [A_1, θ_1] and 3rd-to-1st harmonics [A_2, θ_2]. Such variables allow plotting $\Gamma_{1/f,eff,dc}$ against θ_V as shown in Fig. 26.2.2 (right). For the Class-F_2 VCO [3] with [A_1=0.3, θ_1=0], or Class-F_3 VCO [4] with [A_2=0.33, θ_2=180°], $\Gamma_{1/f,eff,dc}$ is zeroed only at θ_V=0, being a constraint of the 2-port RLCM tank. Although [5] uses both $2F_{OSC}$ and $3F_{OSC}$ resonances, they are of low impedance and low Q, while the optimization only concerns A_1 and A_2. Here, we reveal that proper combinations of $A_{1,2}$ and $\theta_{1,2}$ can yield $\Gamma_{1/f,eff,dc}$=0, regardless of θ_V. To realize [A_V=2.95, θ_V=-10°] for our VCO, we choose [A_1=0.5, θ_1=90°] and [A_2=0.3, θ_2=90°] to nullify $\Gamma_{1/f,eff,dc}$, and obtain $2\Gamma^2_{rms}$=0.45 that is better than those in [1,3,4,6].

Figure 26.2.3 (left) depicts our VCO and its sizing parameters, excluding the varactors and switched-capacitor (SC) cells. An EM simulation at 28GHz aids extracting the equivalent inductive model (e.g., L_1 (r_1), L_2 (r_2), and k). With k<0, the CM-to-DM inductance ratio can be >1 (i.e. $L_{CM}/L_{DM}=(1-k)/(1+k)$), allowing the CM resonance to sit around $2F_{OSC}$, between the 2 DM resonances at F_{OSC} and $3F_{OSC}$. The optimal DM-to-CM capacitance ratio can be found by sweeping the factor X (Fig. 26.2.3, right) similar to [1]. Interestingly, our VCO with only CM tunable capacitors (i.e. X=0) helps correcting the frequency ratio between the 2nd and 3rd resonances from the ideal 0.67 to 0.72; the latter allows retrieving the impedance and phase information at $2F_{OSC}$. Thus, the Leeson's noise factor (F) reaches its theoretical minimum of 1.34 for a MOS channel-noise coefficient $\gamma\approx1$, and the VCO power efficiency is 0.48. At this optimum point, the flicker noise contribution of $M_{1,2}$ is also nulled, ideally yielding $FoM_{@10kHz}\approx FoM_{@10MHz}$.

Our single-turn inductor (Fig. 26.2.4, left) features 2 inner taps to merge L_1 into L_2, partially reducing the magnetic-flux cancellation between them as seen from EM simulations. At 28GHz, the EM-simulated Q_1 is 28.5 for L_1=126pH, and Q_2 is 27 for L_2=250pH; all realized in a compact area (320×250μm²). The varactor (C_V) and tunable C_{CM} (3 bits) at $V_{GP,N}$ target a TR of 16%, and their simulated Q is >25. Tunable C_2 (3 bits, LSB: 9.5fF) at $V_{DP,N}$ is for the resonance alignment: the 1st resonance is set by C_{CM} first, then the 2nd and 3rd can be tuned to align with the 1st via C_2. $\Gamma_{1/f,eff}(t)$ can be estimated by simulating the impedances ($\Gamma_{H,i}$) and phases ($\phi_{\Gamma}^{(i)}$) at F_{OSC}, $2F_{OSC}$, and $3F_{OSC}$. Multiplying them with the higher harmonics of m(t), $\Gamma_{1/f,eff,dc}$ can be nullified within one oscillation cycle in the symmetric $\Gamma_{1/f,eff}(t)$ as exhibited in Fig. 26.2.4 (right). The impedance variations of each resonance over the TR lead to imperfect zeroing of $\Gamma_{1/f,eff,dc}$, accounting for the FoM and 1/f³ PN corner variations.

Our VCO prototyped in 65nm CMOS includes an on-chip divider-by-2 to ease the measurement similar to [3,6]. To ensure the reliability with thin-oxide transistors, we selected V_{DD}=0.48V. Figure 26.2.5 illustrates the measured PN and FoM profiles. It reaches an $FoM_{@1MHz}$ of up to 191.6dBc/Hz and $FoM_{@10MHz}$ of up to 190.3dBc/Hz over the de-embedded TR from F_{min} of 25.5GHz to F_{max} of 29.9GHz. With a 16% TR, the corresponding $FoM_{T@1MHz}$ peaks at 195.7dBc/Hz. Limited by the increased noise floor at the 10MHz offset due to the small-size divider-by-2 and the test buffer, the $FoM_{@10MHz}$ is inferior to $FoM_{@1MHz}$. The 1/f³ PN corner raises from 130kHz at F_{min} to 230kHz at F_{max} due to AM-PM conversion dominated by the parasitics of the varactors and the SC cells. The varactors cover a tuning range of 491MHz at the low frequency and 766MHz at the high frequency, for the control voltage from 0 to 1.2V. The frequency pushing is +114MHz/V at F_{min} dominated by the varactors, but -535MHz/V at F_{max} is originated mainly in the SC cells. Multi-chip measurements (5 samples) confirm that there is no CM-oscillation risk, that the average $FoM_{@1MHz}$ is 190.65dBc/Hz (±0.95dB), and that the average 1/f³ PN corner is 135kHz (±15kHz). With our tunable capacitance range and resolution, the PN and FoM after manual resonance alignment is 5dB (8dB) better at a 1MHz (100kHz) offset. When operating in a PLL, C_{CM} can be tuned for frequency locking first, then C_2 to optimize the PN performance.

Benchmarking with the recent mm-wave VCOs in CMOS [6,7] and BiCMOS [8], this work succeeds in improving both FoM and FoM_T over a wide range of frequency offsets as Fig. 26.2.6 tabulates. The 1/f³ PN corner is at least 2.4× smaller than in [6], and is comparable with [7] that, however, entails an extra mm-wave buffer to enlarge the 3rd-harmonic output, penalizing its FoM. For the die area, our VCO is 1.9× smaller than [7], and 12.5× smaller than [8] that relies on quad-core coupling. Measuring a TR of 16% from 25.5 to 29.9GHz, our VCO achieves a $FoM_{@1MHz}$ up to 191.6dBc/Hz that is at least 2dB better than the prior-art mm-wave VCOs in both CMOS and BiCMOS in Fig. 26.2.6.

Figure 26.2.7 (left) depicts the VCO die micrograph. The active and capacitive elements are laid out inside the inductor footprint for area savings. Figure 26.2.7 (right) plots the achieved 1/f³ PN corner and the FoM relative to other mm-wave VCOs.

Acknowledgements:
The work is funded by the University of Macau - MYRG2017-00167-AMSV, and Macau FDCT - SKL Fund.

References:
[1] D. Murphy et al., "Implicit Common-Mode Resonance in LC Oscillators," *IEEE JSSC,* vol. 52, no. 3, pp. 812-821, Mar. 2017.
[2] C.-C. Lim et al., "An Inverse-Class-F CMOS VCO with Intrinsic-High-Q 1st- and 2nd-Harmonic Resonances for 1/f²-to-1/f³ Phase-Noise Suppression Achieving 196.2dBc/Hz FoM," *ISSCC,* pp. 374-376, Feb. 2018.
[3] M. Babaie and R. Staszewski, "An Ultra-Low Phase Noise Class-F_2 CMOS Oscillator with 191 dBc/Hz FoM and Long-Term Reliability," *IEEE JSSC,* vol. 50, no. 3, pp. 679-692, Mar. 2015.
[4] M. Babaie and R. Staszewski, "A Class-F CMOS Oscillator," *IEEE JSSC,* vol. 48, no. 12, pp. 3120-3133, Dec. 2013.
[5] M. Shahmohammadi et al., "A 1/f Noise Upconversion Reduction Technique for Voltage-Biased RF CMOS Oscillators," *IEEE JSSC,* vol. 51, no. 11, pp. 2610-2624, Nov. 2016.
[6] H. Guo et al., "A 0.083-mm² 25.2-to-29.5 GHz Multi-LC-Tank Class-F_{234} VCO with a 189.6-dBc/Hz FoM," *IEEE SSCL,* vol. 1, no. 4, pp. 86-89, Apr. 2018.
[7] Y. Hu et al., "A Low-Flicker-Noise 30-GHz Class-F_{23} Oscillator in 28-nm CMOS Using Implicit Resonance and Explicit Common-Mode Return Path," *IEEE JSSC,* vol. 53, no. 7, pp. 1977-1987, July 2018.
[8] F. Padovan et al., "A Quad-Core 15GHz BiCMOS VCO with -124dBc/Hz Phase Noise at 1MHz Offset, -189dBc/Hz FoM, and Robust to Multimode Concurrent Oscillations," *ISSCC,* pp. 376-377, Feb. 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

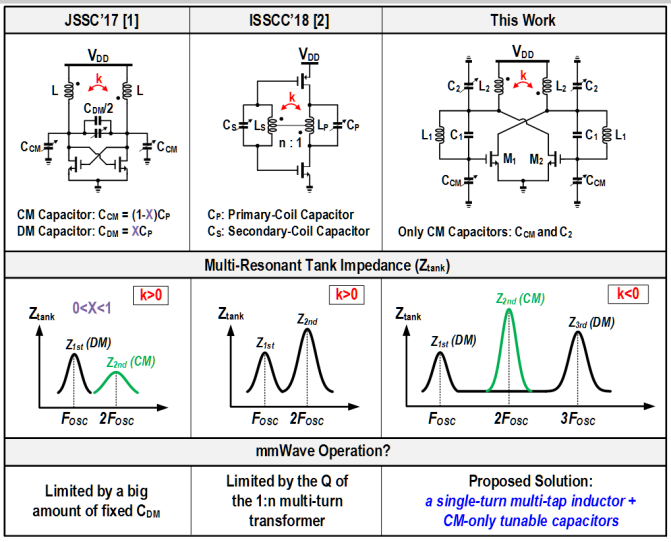


Figure 26.2.1: Simplified schematics and key features of the VCOs in [1,2] and this work. All apply RLCM tanks to generate the multi-resonant tank impedance.

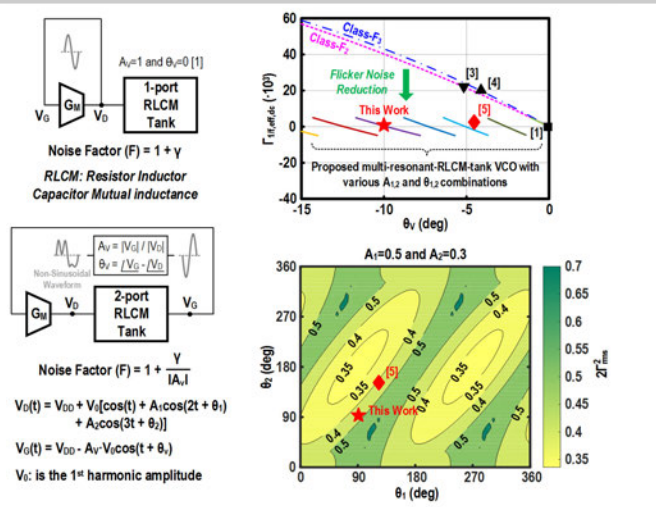


Figure 26.2.2: *Left:* 1-port and 2-port RLCM tanks. *Right*: flicker-noise ($\Gamma_{1/f,eff,dc}$) reduction versus θ_V, and the selected $A_{1,2}$ and $\theta_{1,2}$ combination in this work for $2\Gamma^2_{rms}$=0.45.

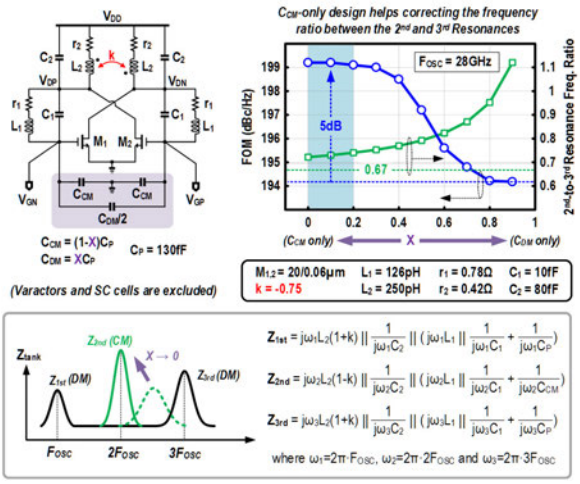


Figure 26.2.3: *Left:* detailed schematic of the proposed VCO. *Right*: changing the C_{CM}-to-C_{DM} ratio via the factor X. An improved FoM is achieved by using only C_{CM} that boosts Z_{2nd}, while correcting the 2nd-to-3rd resonance frequency ratio from the ideal 0.67 to 0.72.

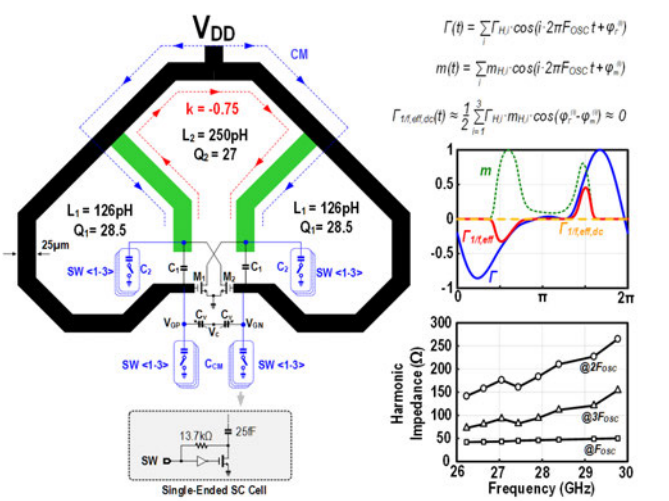


Figure 26.2.4: *Left:* Proposed compact single-turn multi-tap inductor with high Q_1 and Q_2. *Right:* Simulated Γ, m, and $\Gamma_{1/f,eff}$ in one cycle and harmonic impedances over TR.

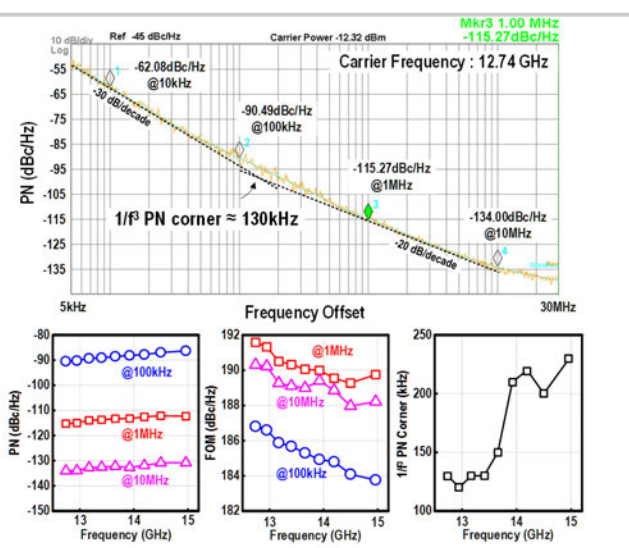


Figure 26.2.5: *Upper:* Measured PN at 12.74 GHz. *Lower*: Performance metrics versus the TR.

Parameters		This Work		JSSC'2018 [7]		SSCL'18 [6]	ISSCC'18 [8]
Key Techniques		Multi-Resonant RLCM tank: Single-Turn Multi-Tap Inductor + CM-only Tunable Capacitors		Implicit Class-F_{23} + 3rd-Harmonic Extraction		Class-F_{234} + DM Tunable Capacitors	Class-C + Quad-Core Coupling
Supply Voltage (V)		0.48		1		0.55	3
Tuning Range (TR) (F_{min} to F_{max}GHz)		16% (25.48 to 29.92)		14% (27.3 to 31.2)		15.7% (25.2 to 29.5)	16% (11.8 to 15.6)
Output Frequency (GHz)		25.48 (12.74 *)	29.92 (14.96 *)	27.3	31.2	28.466 (14.233 *)	15
PN (dBc/Hz)	@ 1MHz Δf	-115.27	-112.31	-106	-104	-114.7	-124
	@ 10MHz Δf	-134.00	-130.77	-126	-125	-131.9	-142.6
Power Consumption P_{DC} (mW)		3.8	4	22 &	23 &	6.6	72
FOM (dBc/Hz)	@ 1MHz Δf	**191.6**	**189.8**	181	180	189.59	189
	@ 10MHz Δf	**190.3**	**188.2**	181 &	181 &	186.79	187.6
FOM_T (dBc/Hz)	@ 1MHz Δf	**195.7**	**193.9**	184	183	193.51	193
	@ 10MHz Δf	**194.4**	**192.3**	184 &	184 &	190.71	191.6
1/f³ PN Corner (kHz)		**130**	**230**	120	210	550	<50
Die Area (mm²)		**0.08**		0.15		0.083	1
Technology		65nm CMOS		28nm CMOS		65nm CMOS	130nm BiCMOS

FOM = $|PN| + 20\log_{10}(f_0/\Delta f) - 10\log_{10}(P_{DC}/1mW)$ $FOM_T = FOM + 20\log_{10}(TR/10)$

* Measured after on-chip divider-by-2 & Included the power of the first-stage O/P buffer for harmonic extraction [7]

Figure 26.2.6: Chip summary and benchmark with [6,7] in CMOS and [8] in BiCMOS.

26

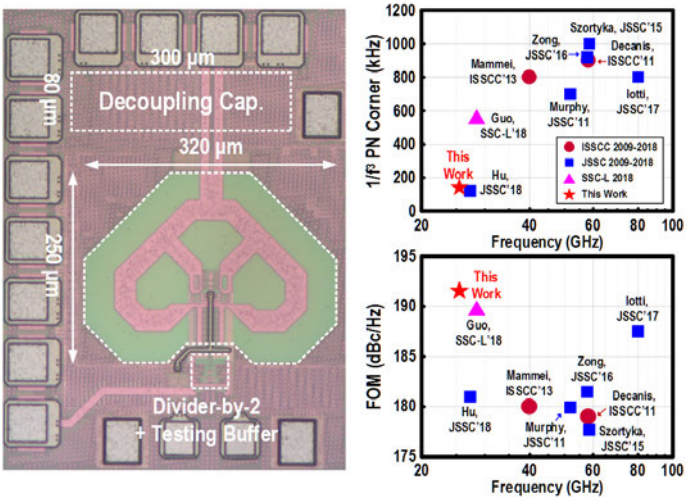


Figure 26.2.7: ***Left***: **Die micrograph of the fabricated mm-wave VCO in 65nm CMOS.** ***Right***: **Plots of the FoM and 1/f³ PN corner with respect to prior-art mm-wave VCOs.**

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

26.3 A 25-to-38GHz, 195dB FoM_T LC QVCO in 65nm LP CMOS Using a 4-Port Dual-Mode Resonator for 5G Radios

Abhishek Bhat, Nagendra Krishnapura

Indian Institute of Technology Madras, Chennai, India

Because of their promise of wide bandwidths and high data-rates, fifth-generation (5G) wireless systems are expected to be deployed worldwide over the next decade. Several bands in the 25-to-40GHz range are allocated for 5G radios. A single widely tunable LC VCO covering the required bands is difficult to design, but is more compact than a set of multiple narrowly tuned VCOs, and is, therefore, preferred. In practice, it is difficult to simultaneously achieve low phase noise and wide tuning range by varying only the capacitor in an LC VCO because (a) switchable capacitor banks have a poor quality factor, Q, at millimeter-wave frequencies and degrade the phase noise, and (b) using wide switches to increase Q increases parasitics and reduces the tuning range.

The VCO in [1] exploits lower parasitics of BJTs in a SiGe BiCMOS process to attain wide-tuning range with good efficiency. Widely tunable LC VCOs in CMOS processes typically also change the effective tank inductance. One such technique, proposed in [2], has a lower phase noise and is more robust compared to techniques such as those that short out a portion of the tank inductor or use multiple cross-coupled negative resistors across different portions of the tank inductor. The technique in [2] uses two identical inductors L_0 with a mutual inductance M_0 between them (Fig. 26.3.1a). The circuit operates in two modes: one in which voltage across ports P1-M1 and P2-M2 are out of phase and another in which they are in phase. The effective tank inductance is L_0+M_0 and L_0-M_0 in the two modes. Loss-compensating negative resistors and the tank capacitors are connected across each port P1-M1 and P2-M2. The desired mode is forced and the undesired mode is suppressed by connecting the transconductors of appropriate polarities between the two ports. Since the two ports are adjacent to each other, these connections can be made without adding significant parasitic inductance.

When designed for a reasonably high Q and self-resonance frequency, the concentric structure with a two-turn inner coil [2] (Fig. 26.3.1a) has an inductance L_0 in excess of 300pH. Therefore, it is not directly usable at millimeter-wave frequencies, which require a small inductance value of 100 to 200pH. Figure 26.3.1b shows the layout of the proposed 4-port single-turn coupled inductor. The central portions, L, of the two mirrored inductors are parallel and closely spaced to maximize the mutual inductance. The total self-inductance of each coil is $2L_s+L$. The dual-mode oscillator using this is shown in Fig. 26.3.1c. As in [2], loss-compensating negative resistors, $-1/G_m$, and the tank capacitors are connected across the ports P1-M1 and P2-M2. Additional ports P3-P4 and M3-M4, each of which has terminals close to each other, are defined at the ends of the central segments. For suppressing odd-mode oscillations, positive resistors R_{xl} are connected across each of these ports. For suppressing even-mode oscillations, differential negative resistors, $-R_{xh}$, which have a positive common-mode resistance, are connected across each of these ports. Since they appear only across a part of the tank inductor, R_{xl} and R_{xh} have to be lower than if they were connected to the ends of the coupled coils.

In the low-frequency mode LB = 1, HB = 0, and the differential negative resistance is off. If R_{xl} = 0, P3 and M3 are shorted to P4 and M4 respectively. This forces $V_{p3} = V_{p4}$ and $V_{m3} = V_{m4}$, and consequently, $V_{p1} = V_{p2}$ and $V_{m1} = V_{m2}$, i.e. the two resonators operate in phase (even mode). With in-phase currents, the total inductance across C is $2L_s+L+M$. The oscillation frequency is $\omega_l = [(2L_s+L+M)C]^{-½}$. Since no current flows through R_{xl} in the even mode, it does not contribute to the resonator phase noise.

In the high-frequency mode: LB = 0, HB = 1. The positive resistance is off and differential negative resistance is on. As shown in Fig. 26.3.1c, the latter is designed to have a negative resistance $-R_{xh}$ between its terminals for a differential signal, which augments loss compensation for the odd mode, and positive resistance of $R_{xh}/2$ between each terminal and ground for common-mode signals, which increases loss in the even mode. If R_{xh} is sufficiently low, this forces odd-mode operation with $V_{p3} = -V_{p4}$ and $V_{m3} = -V_{m4}$, and consequently $V_{p1} = -V_{p2}$ and $V_{m1} = -V_{m2}$. With anti-phase currents, the total inductance is $2L_s+L-M$. The oscillation frequency is $\omega_h = [(2L_s+L-M)C]^{-½}$. Apart from thermal noise of $-R_{xh}$, there is no additional loss added by R_{xh} in the odd mode, hence there is no phase noise penalty.

Figure 26.3.2 shows the conditions on G_m, R_{xl}, and R_{xh} for sustaining oscillations in the desired mode and suppressing the undesired mode. In the prototype, $L_s \approx 40$pH, $L \approx 50$pH, and $M \approx 30$pH, which results in the effective inductance values of 160pH and 100pH in the low- and high-frequency modes, respectively. Figure 26.3.2 shows the effective inductance and Q obtained from EM simulations. The worst-case Q of the inductor is 18. This is good enough since the quality factor of the tank is dominated by the capacitor bank and the varactor ($Q \approx 10$) from 20 to 45GHz. The impedance seen into P1-M1 of the coupled resonator in Fig. 26.3.1c excluding G_m was determined from EM simulations. The impedance peak can be tuned over a wide frequency range and the undesired mode can be effectively suppressed as shown in Fig. 26.3.2.

Figure 26.3.3 shows the VCO cell schematic. Cross-coupled pairs (G_m), 3b capacitor banks, and varactor banks are connected between P1-M1 and P2-M2. Switches controlled by LB realize R_{xl}. Cross-coupled pairs controlled by HB realize $-R_{xh}$. Since both G_m and $-R_{xh}$ compensate for resonator loss in the high-frequency mode, whereas only G_m does so in the low-frequency mode, the cross-coupled pairs controlled by LB are activated only in the low-frequency mode to augment G_m. The cross-coupled pairs realizing G_m and $-R_{xh}$ and the NMOS switch R_{xl} are sized to satisfy the conditions shown in Fig. 26.3.2 for optimal phase noise. The average current consumed in $-R_{xh}$ is about 1.3 times smaller than that in G_m. Two such VCO cells are coupled (only at P1-M1) with a coupling factor ≈ 0.4 to realize a parallel coupled quadrature VCO (PQVCO). To reduce phase-noise degradation, RC phase shifters have been used in the coupling path to reduce the phase difference between the injected coupling signal and the oscillation waveform of the VCO cell. Quadrature outputs V_I and V_Q are taken at ports P2-M2 of the two coupled cells.

The proposed wide-tuning-range QVCO was fabricated in the TSMC 65nm LP CMOS process and encapsulated in a QFN48 package. Figure 26.3.7 shows the block diagram of the chip and the die micrograph. Two on-chip dividers [3] were used to divide the QVCO signals by a factor of eight for easier measurement. The divided outputs were fed to a Keysight E5052B signal-source analyzer for phase-noise measurements through on-chip CMOS buffers and off-chip baluns. Figure 26.3.4 shows the tuning curves of the QVCO covering a range from 25 to 38GHz. The average K_{vco} is 1.4GHz/V. The measured phase-noise spectra and the phase-noise values at a 3MHz offset frequency across the tuning range are shown in Fig. 26.3.5. The phase noise at a 3MHz offset frequency is –120.6 and –111.8dBc/Hz at 25 and 38GHz, respectively. The QVCO consumes 27 to 33.2mA from a 0.65V supply. The figure of merit (FoM) of the QVCO varies from 180.5 to 186.5dB across the tuning range. The tuning-range FoM (FoMT) varies from 192.8 to 198.8dB. Figure 26.3.6 summarizes the performance and compares it with prior-art wide-tuning-range millimeter-wave VCOs. Note that [5] exhibits a good FoMT, but with a 28nm CMOS process with inherently lower parasitic capacitances. By avoiding series or shunt switches in the inductor to switch between modes and by using closely-spaced single-turn inductors to maximize mutual coupling with minimum routing parasitics to active circuitry, the proposed QVCO achieves a wide tuning range with a low phase noise, which compares well with the prior art.

Acknowledgements:
The authors acknowledge the Ministry of Electronics and Information Technology, Government of India, for support and Integrand software for EMX simulator license.

References:

[1] Q. Wu et al., "A 10mW 37.8GHz Current-Redistribution BiCMOS VCO with an Average FoM_T of 193.5dBc/Hz," *ISSCC*, pp. 150-151, Feb. 2013.
[2] G. Li et al., "A Low-Phase-Noise Wide-Tuning-Range Oscillator Based on Resonant Mode Switching," *IEEE JSSC,* vol. 47, no. 6, pp. 1295-1308, June 2012.
[3] A. Ghilioni et al., "A 6.5mW Inductorless CMOS Frequency Divider-by-4 Operating up to 70GHz," *ISSCC*, pp. 282-284, Feb. 2011.
[4] T. Nakamura et al., "A Push-Push VCO With 13.9-GHz Wide Tuning Range Using Loop-Ground Transmission Line for Full-Band 60-GHz Transceiver," *IEEE JSSC*, vol. 47, no. 6, pp. 1267-1277, June 2012.
[5] Y. Shu et al., "A 20.7-31.8GHz Dual-Mode Voltage Waveform-Shaping Oscillator with 195.8dBc/Hz FoM_T in 28nm CMOS," *IEEE RFIC*, pp. 216-219, June 2018.
[6] J. Rimmelspacher et al., "36% Frequency-tuning-range dual-core 60 GHz push-push VCO in 45 nm RF-SOI CMOS technology," *IEEE IMS*, pp. 1356-1358, June 2017.
[7] U. Decanis et al., "A mm-Wave Quadrature VCO Based on Magnetically Coupled Resonators," *ISSCC*, pp. 280-282, Feb. 2011.
[8] T. Xi et al., "Low-Phase-Noise 54GHz Quadrature VCO and 76GHz/90GHz VCOs in 65nm CMOS Process," *IEEE RFIC*, pp. 257-260, June 2014.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

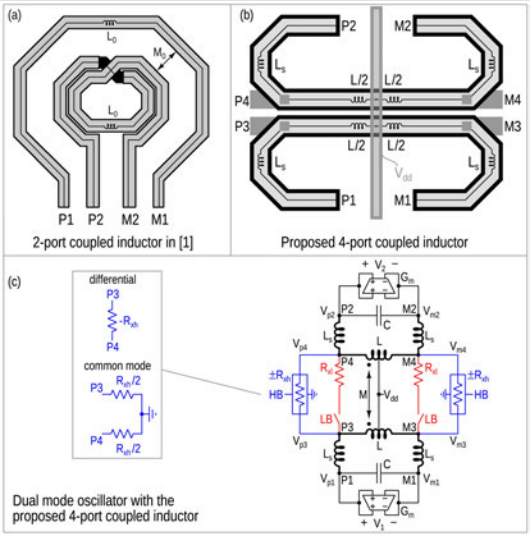


Figure 26.3.1: Coupled inductor in [2], proposed layout, and proposed dual-mode resonator with mode suppression.

Mode	Constraint	Condition	
Low frequency HB = 0, LB = 1 $\omega_l = \frac{1}{\sqrt{(2L_s+L+M)C}}$	Sustain even mode ($V_1 = V_2$)	$-G_m + G_{pe}(\omega_l) < 0$	$k_o = \frac{L-M}{2L_s+L-M}$
	Suppress odd mode ($V_1 = -V_2$)	$R_{xl} < \frac{k_o^2}{G_m - G_{po}(\omega_h)}$	
High frequency HB = 1, LB = 0 $\omega_h = \frac{1}{\sqrt{(2L_s+L-M)C}}$	Sustain odd mode ($V_1 = -V_2$)	$-G_m - \frac{k_o^2}{R_{xh}} + G_{po}(\omega_h) < 0$	$k_e = \frac{L+M}{2L_s+L+M}$
	Suppress even mode ($V_1 = V_2$)	$R_{xh} < \frac{k_e^2}{G_m - G_{pe}(\omega_l)}$	

G_{pe} and G_{po} are even and odd-mode half-circuit loss conductance across P1,M1 excluding G_m.

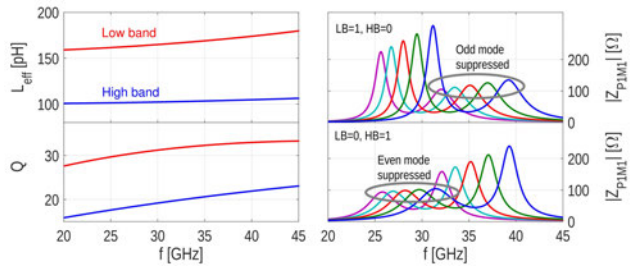


Figure 26.3.2: Oscillation conditions, EM-simulated inductor characteristics, and impedance looking into P1-M1.

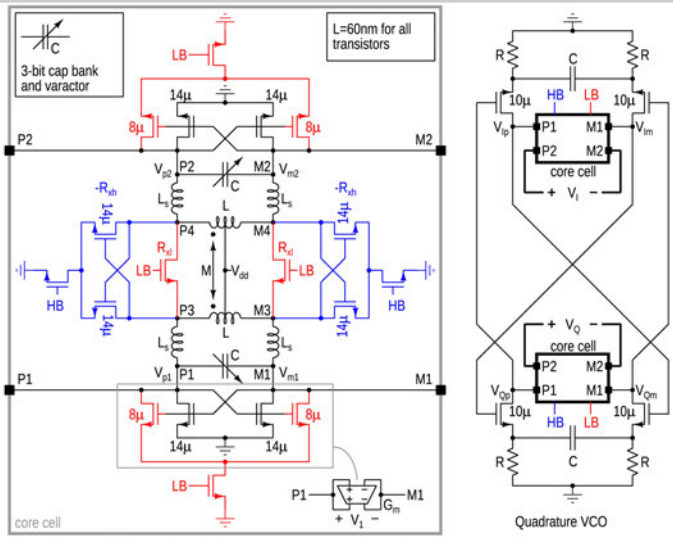


Figure 26.3.3: Dual-mode VCO cell and quadrature VCO using two coupled cells.

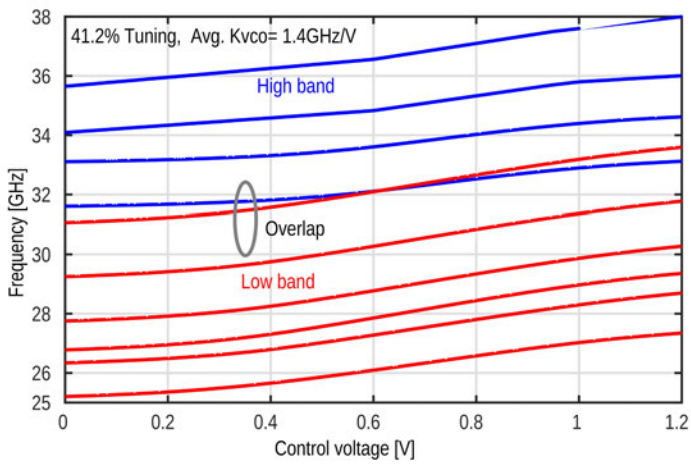


Figure 26.3.4: Measured continuous tuning range (25 to 38GHz). 3b binary-weighted capacitor bank and 2b varactor bank used for course and fine tuning, respectively.

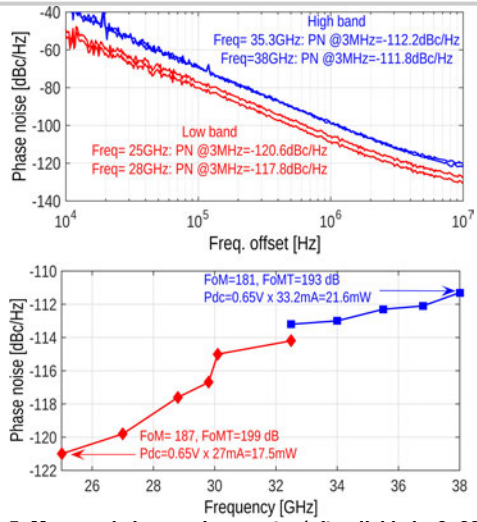


Figure 26.3.5: Measured phase-noise spectra (after divide by 8, 20log8 [dB] is added to the measurements to obtain the VCO phase noise), phase noise at a 3MHz offset.

Reference	This work	[1] ISSCC'13	[4] JSSC'12	[5] RFIC'18	[6] IMS'17	[7] ISSCC'11	[8] RFIC'14
Type	**QVCO**	VCO	VCO	VCO	VCO	QVCO	QVCO
Tech.	**65nm CMOS**	0.13μm SiGe	0.18μm SiGe	28nm CMOS	45nm SOI	65nm CMOS	65nm CMOS
Supply (V)	**0.65**	1.5	0.6	0.9	1	1	1
Power (mW)	**17.5 to 21.6**	10	132	5.5	65	22	24
f_0 (Hz)	**25G 38G**	32.9G	52.5G	26.2G	33G	58.2G	54G
Phase noise (dBc/Hz) @Δf = 3MHz*	**−121 −112**	−114	−118	−115	−102.6	−109	−109
Tuning range (%)	**41.2**	30	26.5	42.3	36	7.5	9.1
FoM (dB)	**187 181**	184.9	181.7	185.1	161.7	181.3	180.0
FoMT (dB)	**199 193**	194.4	190.1	197.6	172.8	178.4	179
Area (mm^2)	**0.08**	0.03	0.168	0.07	0.12	0.075	0.032

FoM=|PN (dBc/Hz)|+ $20\log_{10}(f_0/\Delta f)$- $10\log_{10}$(Power(mW)); FoMT = FoM + $20\log_{10}$(% Tuning range/10%);

*PN at 3MHz offset estimated from the phase-noise plots in the compared works.

Figure 26.3.6: Performance summary and comparison.

26

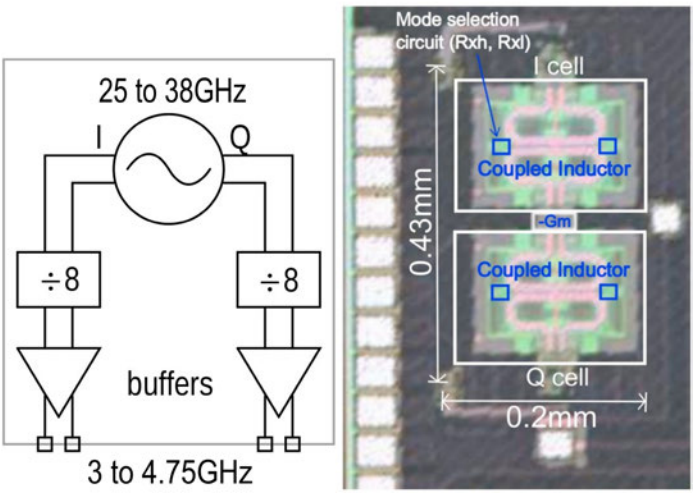


Figure 26.3.7: Die block diagram and micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

26.4 A 2.4GHz 65nm CMOS Mixer-First Receiver Using 4-Stage Cascaded Inverter-Based Envelope-Biased LNAs Achieving 66dB In-Band Interference Tolerance and -83dBm Sensitivity

Dawei Ye[1], Rongjin Xu[1], C.-J. Richard Shi[1,2]

[1]Fudan University, Shanghai, China; [2]University of Washington, Seattle, WA

Intensive interest in internet-of-things devices has crowded the ISM bands (e.g., 2.4GHz). A robust receiver for each wireless link in such a coexistence environment is required not only to have high rejection of out-of-band interference but also to tolerate strong in-band interference, which cannot be filtered out by the front-end BAW/SAW devices. Recently, nonlinear suppressors [1,2] have been demonstrated to amplify a weak desired signal while suppressing strong interferers. In contrast to conventional frequency-dependent techniques, which consume high power to implement high-Q filtering [3-6], the nonlinear suppressor is frequency independent and only needs the envelope information of the interferer to adjust the nonlinear transfer function for interference suppression, reducing power consumption. On the other hand, unlike traditional linear receivers, in particular mixer-first receivers, which require high power to lower LO phase noise to alleviate the impact of reciprocal mixing, nonlinear-suppressor-based receivers can suppress the phase noise inherently and hence have high tolerance to reciprocal mixing. Unfortunately, a major issue for nonlinear suppression is in its severely increased noise figure (NF), which degrades the system sensitivity and limits its potential applications to short ranges.

In this work, a nonlinear suppression technique without severe NF reduction is introduced. A circuit implementation of an inverter-based envelope-biased LNA (IBEB-LNA) is proposed. The IBEB-LNA dynamically adjusts its bias from input envelope tracking. This suppresses the strong interferer in the same way as a transconductance-amplifier (TA) based nonlinear suppressor [1,2]. But unlike TA-based suppressors, the IBEB-LNAs are mostly in the off state (no static current during the operation) thus alleviating the NF deterioration. Further, a cascaded IBEB-LNA architecture with negative feedback to compensate PVT variations in the envelope tracking is introduced, which further increases interference suppression and reduces the NF. With a 4-cascaded IBEB-LNA chain as a nonlinear suppressor, a mixer-first receiver has been implemented, which yields the measured carrier-to-interference ratio (CIR) of -66dB and the sensitivity of -83dBm, improving significantly over the previously reported CIR of -49dB and sensitivity of -61dBm [1], at 30× faster data-rate with only 10× more power.

Figures 26.4.1 shows the interference suppression principle of the IBEB-LNA compared to that of the IB-LNA. The IBEB-LNA uses an envelope-tracking path (ETP) to dynamically adjust the bias from the half V_{DD} to the input envelope. The ETP consists of an envelope detector (ED), a voltage doubler (VD), and a replica inverter (RI). The V_{IN} of the IBEB-LNA is firstly fed to the ED. Then its envelope V_P is extracted from V_{IN} and sent to the VD to compensate a 2× voltage reduction from the VD output to V_{INa}. After that, the input biasing voltage V_{REF} is changed to $V_{REF} + V_P$, realizing envelope-tracked biasing to enhance the CIR at V_{OUT}. To verify the principle, the signal conversion gain G_{sig} and interferer conversion gain G_{int} are simulated, which clearly show that compared to the IB-LNA, a ~20dB CIR improvement ($=G_{sig}/G_{int}$) is achieved by the IBEB-LNA when V_{INT}=0.2V. Yet, both of their NFs become worse as V_{INT} is increased and the corresponding current consumptions are also proportionally decreased. This is because the strong interferer turns off the IB/IBEB-LNA in most of their time period and hence reduces the average current. It implies that for a large V_{INT}, increasing the size of gain stage can compensate the NF degradation while keeping constant current consumption.

To further improve the CIR and NF, multiple stages of IBEB-LNAs are cascaded. Figure 26.4.2 shows the schematic of the first 2 stages of a 4-stage cascaded IBEB-LNA chain. The multiplier of each gain stage is tunable. For a stronger interferer, the multiplier is set to a larger number (e.g., M=16) to maintain the NF performance. Since half of V_{OUT1} (V_{OUT2}) is clipped and its phase is inversed with respect to V_{IN} and V_{OUT1}, a negative unit gain buffer is added after ED1 and ED2 to ensure the unclipped part of V_{OUT1} (V_{OUT2}) can be amplified. Moreover, to alleviate the influence of PVT variation on the envelopes, an envelope feedback loop (EFL) is introduced. The output envelope V_{P2} of V_{OUT1} is extracted by the ED2 and then is fedback to the error amplifier G_{E1}, which amplifies the envelope error between $-V_{P1}$ and V_{P2}. After passing through the VD2 and the RI2, the amplified error adds onto the LNA input to adjust V_{REF}. Although the EFL can minimize the spectrum regrowth of V_{OUT1} caused by the PVT, it also presents a trade-off between the envelope tracking speed and loop stability. The ETP and EFL are configurable and can be easily disabled to save the power.

Figure 26.4.3 shows the system diagram of a fully integrated 2.4GHz passive-mixer-first OOK receiver with the proposed IBEB-LNA chains. After an external matching network, the RF signal is first down-converted by a RF passive mixer driven by a 2.16GHz PLL to the IF1 band. Due to reciprocal mixing, the noise deposited on the desired signal deteriorates the RX NF when suffering interference from a strong blocker. However, since the IBEB-LNAs reject the blocker in the amplitude domain, the noise in the blocker phase can be accordingly suppressed [2], alleviating the impact of reciprocal mixing. The output signal of the IBEB-LNA chain is down-converted to the IF2 band. Then a 4-phase N-path filter is used to reject the residual power of the interferer. Finally, the signal is down-converted to the baseband (consisting of a 4th-order lowpass filter, a VGA, an ED, and a 1st-order lowpass filter) for further filtering.

The chip is fabricated in a 65nm CMOS process and the die micrograph is shown in Fig. 26.4.7. The power supply of the receiver is 1V except for VDs, which use 1.6V. The data-rate of the receiver is 307kbps. The measured total power consumption is 2.05mA, where 51% is used by PLL/buffers/XOs and only 29% is used by IBEB-LNAs. Without the blocker, the measured NF and conversion gain from the RF input to the 4th IBEB-LNA output is 11dB and 32dB, respectively. The S_{11} of the receiver is -10dB. The blocker NF of the IBEB-LNA chains is measured and shown in Fig. 26.4.4. The blocker is 10MHz away from the desired band. As can be seen, by enabling the ETP and EFL and adjusting the multiplier number, the blocker NF can be kept around 15dB. However, it is degraded by 11.5dB for a -12dBm blocker when disabling the ETP and EFL. The related power consumptions for the IBEB-LNA chains are also measured and shown. Obviously, the power consumption approximately remains constant due to the nearly zero DC current of the gain stages. The noise-and-power trade-off is optimized. Finally, the G_{sig} and G_{int} are measured and shown. With the ETP and EFL, the maximum CIR improvement is up to 36dB compared to that without the ETP and EFL. Figure 26.4.5 shows the measured RX specifications. The sensitivity is -83dBm for the BER of 0.1%. Moreover, the PLL phase noise at a 10MHz offset is -133dBc/Hz. In principle, this results in the 29dB blocker NF for a -12dBm blocker after RF mixing. However, with the IBEB-LNA chains, the blocker NF is suppressed to 15dB (shown in Fig. 26.4.4), improving the sensitivity. To test the CIR, a continuous wave is used as the interferer and the signal power is 6dB higher than the sensitivity. The relationship between the CIR and frequency offset (f_{int}-f_{sig}) is shown in the bottom-left of Fig. 26.4.5. The CIR at a 5MHz offset is improved by 40dB when enabling the ETP and EFL. On the other hand, a GFSK and an ASK (50% modulation depth) interferers with different symbol rates are also used for CIR measurement. The frequency offset is fixed to 5MHz. As shown in the bottom-right of Fig. 26.4.5, the symbol rate of the GFSK interferer has less impact on the CIR since it has a constant envelope, which can be easily tracked by the EFL. However, when using the ASK interferer for measurements, the CIR degrades as the interferer symbol rate increases. The comparison with prior art is shown in Fig. 26.4.6.

Acknowledgement:
This work is supported by Science and Technology Commission of Shanghai Municipality under Grant 16JC1400102. We thank Liangjian Lyv and Yuting Tu for their technical and layout support.

References:
[1] D. Ye et al., "An Ultra-Low-Power Receiver Using Transmitted-Reference and Shifted-Limiters for In-band Interference Resilience," *ISSCC,* pp. 438-439, Feb. 2016.
[2] E. Janssen, "Methodologies for Multi-Radio Coexistence: Self-Interference Suppression Techniques," Ph.D. Dissertation, Dept. of Electrical Engineering, Technische Universiteit Eindhoven, Jan. 2014.
[3] R. Ni et al., "A 2.4 GHz Hybrid Polyphase Filter Based BFSK Receiver with High Frequency Offset Tolerance for Wireless Sensor Networks," *IEEE JSSC,* vol. 48, no. 5, pp. 1250-1262, May 2013.
[4] J. Cheng et al., "A 1.3mW 0.6V WBAN-Compatible Sub-Sampling PSK Receiver in 65nm CMOS," *ISSCC,* pp. 168-169, Feb. 2014.
[5] H. Seok et al., "A 2.4GHz, -102dBm-Sensitivity, 25kb/s, 0.466mW Interference Resistant BFSK Multi-Channel Sliding-IF ULP Receiver," *IEEE Symp. VLSI Circuits,* pp. 70-71, June 2017.
[6] H. Liu et al., "An ADPLL-Centric Bluetooth Low-Energy Transceiver with 2.3mW Interference-Tolerant Hybrid-Loop Receiver and 2.9mW Single-Point Polar Transmitter in 65nm CMOS," *ISSCC,* pp. 444-445, Feb. 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

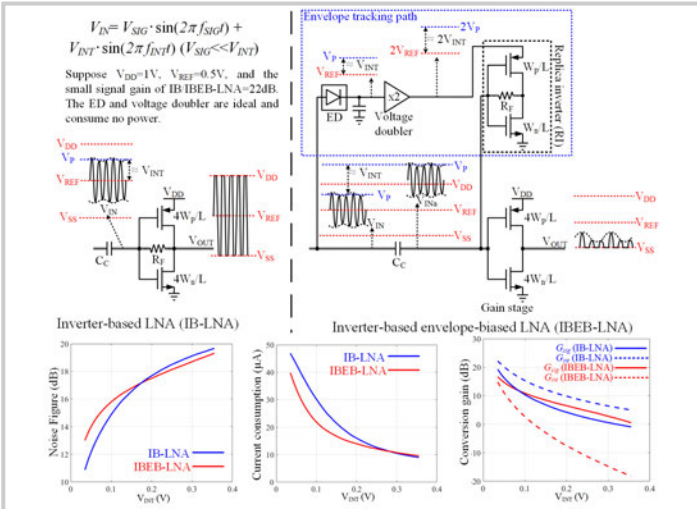


Figure 26.4.1: The comparison of in-band interference resilience between IB-LNA and the IBEB-LNA and their simulated NF, current consumption, and conversion gain with respect to V_{INT}

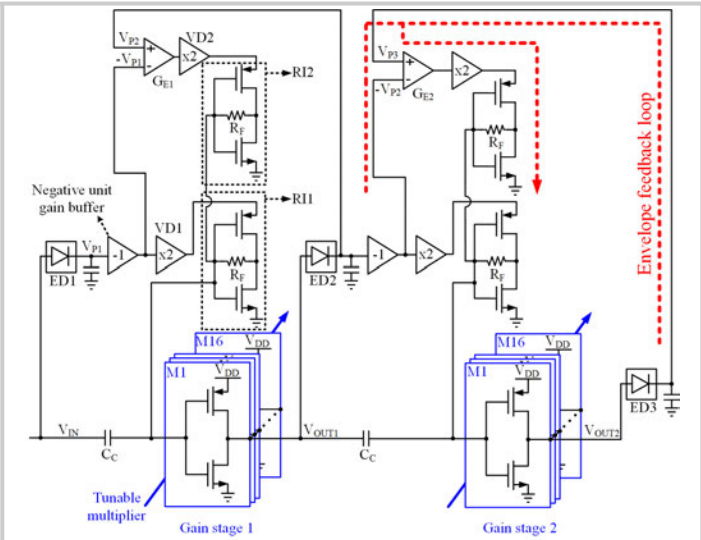


Figure 26.4.2: Simplified schematic of the first 2 stages of a 4-stage cascaded IBEB-LNA chain.

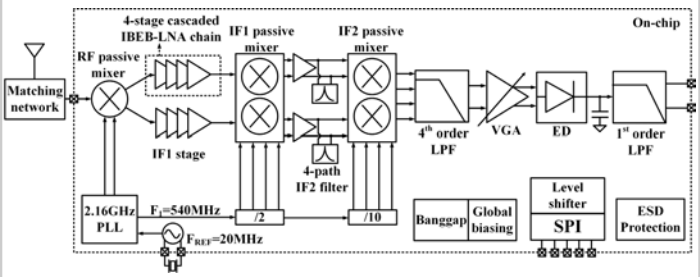


Figure 26.4.3: The system diagram of the proposed 2.4GHz passive-mixer-first receiver with IBEB-LNA chains.

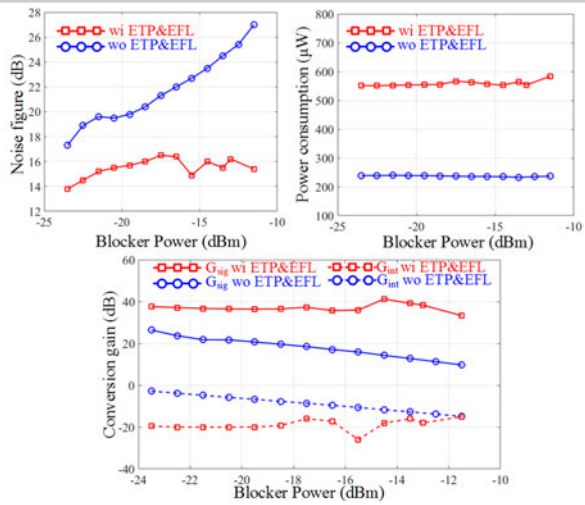


Figure 26.4.4: Measured blocker NF and corresponding power consumption of the IBEB-LNA chains and the signal conversion gain (G_{sig}) and interferer conversion gain (G_{int}) for the proposed cascaded IBEB-LNA chains.

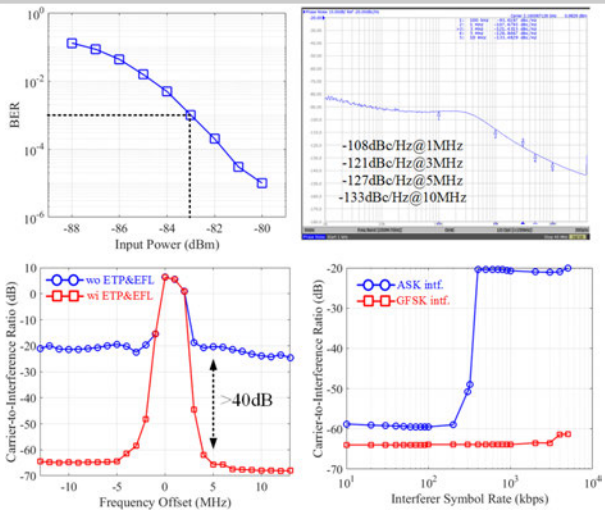


Figure 26.4.5: Measured system specifications including the BER with respect to input power, the phase noise of the PLL, the CIR with respect to frequency offset and the CIR for different modulated interferers.

	[1] ISSCC16	[3] JSSC13	[4] ISSCC14	[5] VLSI17	[6] ISSCC18	This work
Frequency (MHz)	915	2400	2400	2400	2400	2400
Techniques for in-band blocker rejection	Shifted-limiter +Transmitted reference	Hybrid PPF	Q-enhanced LNA	Cascaded N-path +Hybrid PPF	ADPLL reused as ADC with feedback DAC calibration	IBEB-LNAs
Modulation	OOK	BFSK	π/2-DBPSK	BFSK	GFSK	OOK
Power supply (V)	1.0	1.0	0.6	0.6	1.0	1.0/1.6(3)
Instant. Power (μW)	175	1950(1)	1301	466	2000(2)	2050
S11 (dB)	-13	<-15	N/A	N/A	N/A	-10
Noise figure (dB)(4)	N/A	6	6	N/A	6	11
Conversion gain (dB)(4)	N/A	38.8	N/A	N.A.	68.8	32
Sensitivity (dBm)	-61	-84	-96	-102	-94	-83
LO Phase noise (dBc/Hz) @3MHz @5MHz @10MHz	No LO(5)	External LO	N/A N/A N/A	(6) -114 -115 -115	N/A N/A N/A	-121 -127 -133
In-band CIR (dB) @+/- 3MHz @+/- 5MHz	-49/-49 -49/-49	-25/-26 -41/-41	N/A N/A	-44/-45 -52/-51	-37/-36 N/A	-45/-59 -66/-65
Data rate (kbps)	10	1000	486	10	1000	307

(1). excluding the power consumption of LO;

(2). excluding the power consumption of digital baseband;

(3). the 1.6V supply is only for the voltage doubler;

(4). [5] measure the NF and conversion gain of the whole RX chain;

(5). the RX uses ED for frequency down-conversion;

(6). the LC-VCO uses off-chip inductor.

Figure 26.4.6: Performance summary and comparison with the prior art.

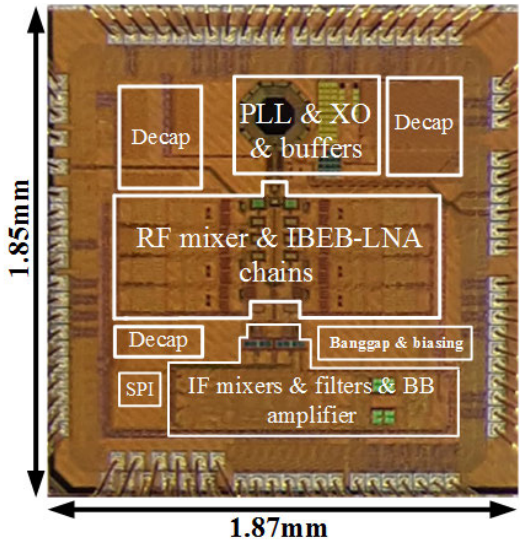


Figure 26.4.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

26.5 A 0.1-to-0.2V Transformer-Based Switched-Mode Folded DCO in 22nm FDSOI with Active Step-Down Impedance Achieving 197dBc/Hz Peak FoM and 40MHz/V Frequency Pushing

Omar El-Aassar, Gabriel M. Rebeiz

University of California, San Diego, La Jolla, CA

The continuous improvement in ultra-low-power/voltage (ULP/V) circuits paves the way for new near-battery-free IoE applications. In this work, we propose an LC oscillator that can significantly improve phase-noise (PN) performance and the figure of merit (FoM) at ULV supply levels. The structure is applied to a 4.15-to-4.97GHz switched-mode DCO that operates from a less-than-100mV supply and achieves a peak FoM of 197dBc/Hz at a 150mV supply.

A trifilar-coil oscillator [1] can operate from an ULV supply. The structure benefits from reduced drain impedance and passive voltage gain, provided by the source feedback and the Class-F transformer, to achieve low PN and reliable startup at 0.4V. However, at 0.2V, the PN is penalized from the limited loop gain and voltage swing, and an FoM of 188dBc/Hz is achieved. For the same headroom and using N-coupled oscillators, it is possible to decrease the PN by 10log(N) with the same loop gain and FoM at the expense of chip area. Another approach is to reduce the impedance seen by the transistor to achieve a higher power while also increasing the loop gain for reliable startup.

We first consider a modified version of the trifilar-coil design by modifying the Class-F transformer to a Class-F_{23} [2], which provides a common-mode (CM) resonance around the second harmonic ($2F_{LO}$) [3,4] at the drain nodes (Fig. 26.5.1-left). This is particularly useful in reducing flicker up-conversion while decreasing the impulse sensitivity function (ISF) and results in a lower overall PN. By adding a common-gate PMOS in a folded-cascode topology (Fig. 26.5.1-right), the main tank impedance (L_D side) can be effectively divided between both NMOS and PMOS devices under the same supply without sacrificing the loop gain. The resonator at the supply/folding node (L_F) is designed to have a high-quality-factor (Q) intrinsic resonance around $2F_{LO}$ for PN and frequency-pushing improvements. A weak positive coupling, k_{FD} (between L_F and L_D), is engineered along with the PMOS transconductance (Gm) to control the swing at the folding node and thus the impedance division between the NMOS and PMOS devices. Therefore, the single-side Gm-cell resembles a folded-stacked topology more than a folded cascode. The CM resonance at the PMOS drain is preserved by controlling the ratio of the differential-to-single-ended capacitance at this node. A source feedback resonator L_S is used for gain boosting and harmonic shaping. Instead of aligning the source resonator with $3F_{LO}$, a process similar to the source/load pull in power amplifiers is adopted to maximize the oscillator efficiency. Non-purely real, but adequately high, impedance values at the 4th and 5th harmonics were found to maximize the efficiency for the implemented design.

The maximum FoM theoretical value (FoM_{max}) is dictated by the Q of the passives for a given process. The FoM peaks when the CM resonance is adjusted around $2F_{LO}$ (Fig. 25.6.2-top-left). There remains, however, an FoM gap, for practical implementations, due to the limited CM-resonance Q and the oscillator switching power loss. These two factors are addressed in the switching mechanism of the proposed folded oscillator (Fig. 26.5.2-top-right). Under a large signal swing, both NMOS and PMOS on one side (M_{1N} and M_{1P}) are ON at different halves of the oscillator cycle. When the NMOS is in the triode region, an OFF state PMOS prevents the NMOS from de-Qing the main tank (L_D/L_G), while a high-Q resonance at the folding node at $2F_{LO}$ forces the NMOS current noise to circulate in its channel resistance ($1/g_{dsN}$). Similarly, when the PMOS is in triode, the intrinsic $2F_{LO}$ resonance at L_F, in addition to the CM resonance at L_D, minimizes the current noise injection in the tank. When both NMOS and PMOS are in saturation, the channel-length-modulation loading, which is critical for short-channel devices, is reduced due to the cascode structure while the transimpedance loop gain is boosted, thus allowing the gate voltage to swing 2× higher for a given supply at V_F compared to the Class-F_{23} with a source feedback. At a 0.2V supply, the single-ended peak-to-peak voltage swing at the gate node is 1.3V, providing rail-to-rail operation for the following on-chip buffer. The harmonic-rich drain-to-source (VDS) voltage for both NMOS and PMOS shapes the time domain ISF (Γ), (Fig. 26.5.2-bottom), while the small I/V overlap shifts the noise modulation function (NMF) from the ISF, thus yielding a negligible device effective ISF (Γ_{eff}) with a simulated switching efficiency higher than 85% at 5GHz (Fig. 3-right).

A prototype of the folded DCO was fabricated in 22nm FDSOI. Low-VT (250mV for NMOS and 300mV for PMOS) thin-oxide devices are used to provide a larger loop gain and better PN performance (particularly in the thermal region). Reliable start-up is ensured, at low supply levels, by optimizing the gate bias for the NMOS/PMOS pairs. The 4-winding transformer (Fig. 26.5.3-top-left) is designed in a planar fashion while stacking the upper two thick metals (copper and aluminum) for each winding. This is necessary to reduce the mandatory metal fill in the upper layers, which reduces Q. The simulated Q values for L_G, L_D, L_F, and L_S are 11, 11.2, 16, and 17, respectively, at 5GHz. A 3b differential switched-capacitor bank is placed in parallel with the lower swing inductor L_D for frequency tuning (Fig. 26.5.3-left). Therefore, thin-oxide low-VT NMOS switches can be used with their smaller R_{on}-C_{off}. Two digital supply levels are employed to bias the switches in ON and OFF states. This avoids accidental switch turn on from the large voltage swings while ensuring that the swing across the OFF switch remains within reliable limits.

Figure 26.5.4 presents the measured PN and FoM at the maximum (f_{max}=4.97GHz) and minimum (f_{min}=4.15GHz) frequencies for three different supplies (0.2, 0.15, 0.1V). At the upper band (f_{max}), the 10MHz-offset PN of -144.3, -142.1, and -136.5dBc/Hz is measured for 0.2, 0.15, and 0.1V, respectively. The achieved FoMs are higher than 193.7dBc/Hz for all supply levels in the thermal region and reach 197dBc/Hz at 0.15V. The flicker corner frequency, where the FoM is 3dB lower than its peak value, is between 180 and 700kHz at f_{max} (the simulated flicker corner for low-VT thin SOI devices is 5 to 10MHz). Also, at f_{min}, the flicker corner is greatly reduced to 50 to 70kHz for all three supplies. This validates the effectiveness of flicker reduction by the resonator L_F, but also shows the importance of tuning the tank at the folding node. The L_F tank is more tolerant to de-Qing from capacitive tuning since it is isolated from the main L_G/L_D tank. The FoM is reduced at f_{min} due-to the loading of the capacitor bank and the reduced gain, but remains better than 190dBc/Hz at both 0.2 and 0.15V. The PN and FoM results across the frequency range (Fig. 25.5.5-left) reach best values at the thermal region near f_{max}, when the capacitor bank is OFF, while the flicker is most reduced near f_{min}. The folded structure also decouples the supply from the main tank, thus allowing reduced frequency pushing. In addition, the dominant non-linear capacitors, C_{GDN} and C_{SGP} (Fig. 26.5.5-right), change in opposite ways with increasing supply, and provide complementary capacitance-change compensation at the supply node. The implemented topology allows the DCO to operate from 75mV (~2.25× less than any reported CMOS/FinFET oscillator in Fig. 26.5.6) to 300mV, and with a frequency pushing no larger than 40MHz/V across the whole frequency range (Fig. 25.5.5-bottom-right).

Compared with prior art, low-supply, LC oscillators employing common-mode resonance (Fig. 26.5.6), our folded DCO operates from the lowest supply and achieves the highest FoM at a 10MHz offset, and one of the highest FoMs at a 100kHz offset. The PN achieved at a 200mV supply is nearly 10dB better than the 16nm-FinFET design in [1] operating from the same supply, while it is comparable to the CMOS 65nm inverse Class-F design [5] operating from 3× the supply.

Acknowledgements:
The authors would like to thank Analog Devices for their financial support, Integrand for EMX electromagnetic simulator and GlobalFoundries for chip fabrication.

References:
[1] C.-C. Li et al., "A 0.2 V Trifilar-Coil DCO with DC-DC Converter in 16nm FinFET CMOS with 188dB FoM, 1.3kHz Resolution, and Frequency Pushing of 38MHz/V for Energy Harvesting Applications," *ISSCC*, pp. 332-333, Feb. 2017.
[2] M. Shahmohammadi et al.,"A 1/f Noise Upconversion Reduction Technique Applied to Class-D and Class-F Oscillators," *ISSCC*, pp. 444-445, Feb. 2015.
[3] E. Hegazi et al., "A Filtering Technique to Lower LC Oscillator Phase Noise," *IEEE JSSC*, vol. 36, no. 12, pp. 1921-1930, Dec. 2001.
[4] D. Murphy et al., "A VCO with Implicit Common-Mode Resonance," *ISSCC*, pp. 442-443, Feb. 2015.
[5] C.-C. Lim et al., "An Inverse-Class-F CMOS VCO with Intrinsic-High-Q 1st- and 2nd -Harmonic Resonances for $1/f^2$-to-$1/f^3$ Phase-Noise Suppression Achieving 196.2dBc/Hz FoM," *ISSCC*, pp.374-376, Feb. 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

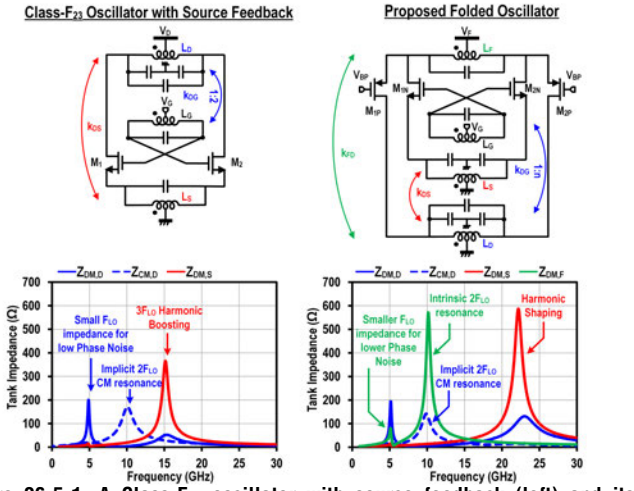


Figure 26.5.1: A Class-F_{23} oscillator with source feedback (left) and its composite tank impedance (bottom-left); Proposed folded oscillator using an intrinsic $2F_{LO}$ resonator at the folding node and its composite tank impedance (right and bottom-right).

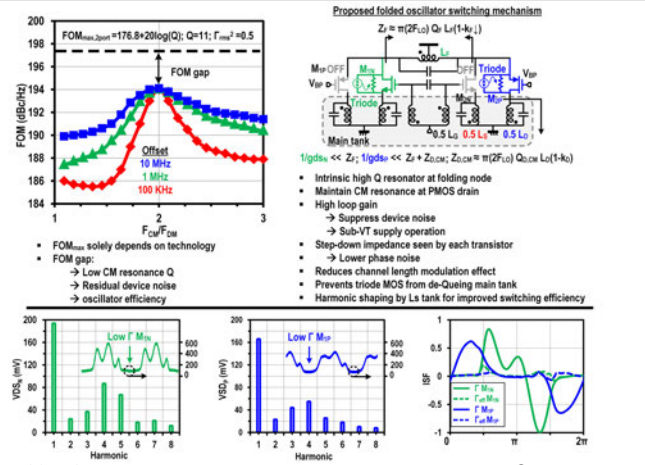


Figure 26.5.2: Dependence of the FoM on the accuracy of the CM resonance and existence of an FoM gap from sub-optimal operation (top-left). Folded-oscillator switching mechanism to bridge the FoM gap (top-right). Simulated drain-to-source voltage harmonics for both NMOS and PMOS and their respective ISF (bottom).

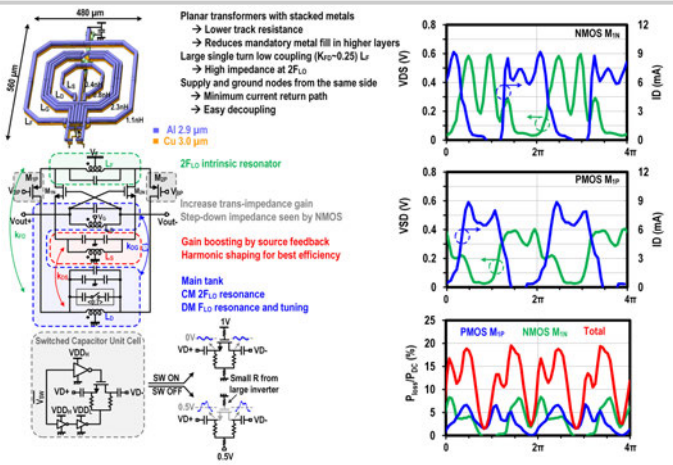


Figure 26.5.3: Folded DCO and its transformer layout (top-left); switched-capacitor unit cell and its biasing scheme in the ON and OFF states (bottom-left). Simulated drain-to-source voltage, current, and the normalized power loss with respect to the DC power for the switching FETs per cycle showing an efficiency higher than 85% (right).

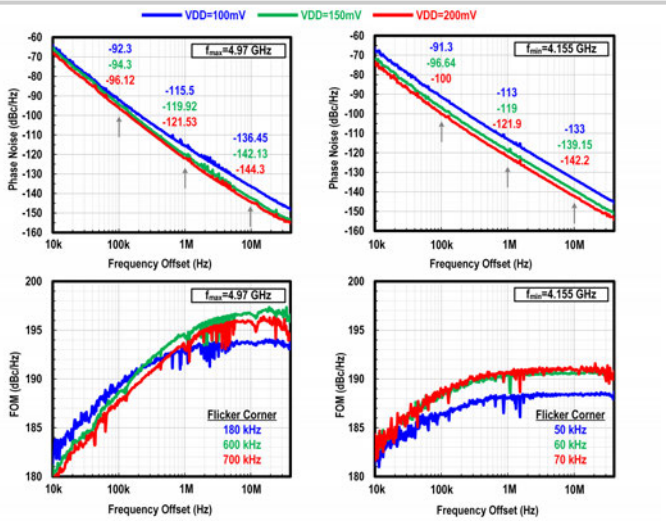


Figure 26.5.4: Measured phase noise and FoM for the 0.1V, 0.15V, and 0.2V supplies versus offset frequency for the high band (left) and the low band (right).

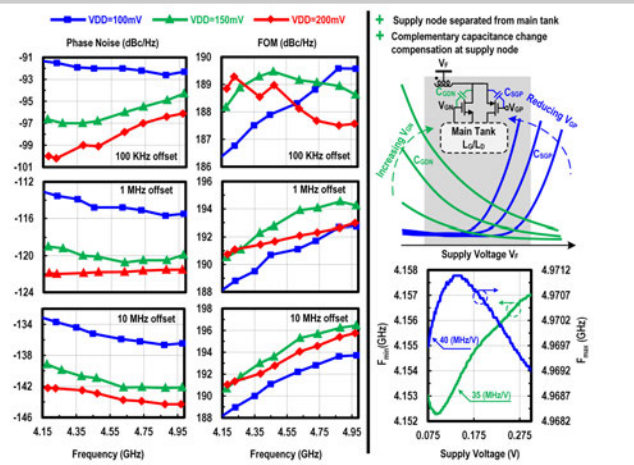


Figure 26.5.5: Measured phase noise and FoM versus oscillation frequency for 0.1V, 0.15V, and 0.2V supplies (left). Frequency pushing reduction mechanisms (top-right); Measured frequency pushing for the highest and lowest band demonstrating lower than 40MHz/V pushing for sub-threshold supply levels (bottom-right).

Design			[2] ISSCC'15	[4] ISSCC'15	[5] ISSCC'18	[1] ISSCC'17		This Work		
Topology			Class-F_{23}	NMOS implicit CM resonance	CMOS Inverse Class-F	Trifilar-Coil		Folded, 4-winding transformer		
Technology process			40nm	28nm	65nm	16nm FinFET		22nm FDSOI		
Frequency (GHz)			5.4-7	2.85-3.75	3.49-4.51	3.2-4		4.15-4.97		
Tuning Range (%)			25%	27.2%	25.5%	22%		18%		
VDD (V)			1	0.9	0.6	0.4	0.2	**0.2**	**0.15**	**0.1**
Power (mW) @ f_{min} / f_{max}			12/10	7.2/6.35	1.2/1.14	3.8	0.6	**2.28/ 1.77**	**1.22/ 0.91**	**0.54/ 0.46**
Phase noise (dBc/Hz)	@ 100 kHz	f_{min}	-105.3/ -102.05	-107/ -102	-102.4/ -98.5	-99*/ -95*	-88*/ -90*	-100/ -96.1	-96.7/ -94.3	-91.3/ -92.3
		f_{max}								
	@ 10 MHz	f_{min}	-146.7/ -144.5	-152†/ -148†	-145.6/ -143.7	-145/ -143	-133/ -135	**-142.2/ -144.3**	**-139.2/ -142.1**	**-133/ -136.5**
		f_{max}								
FOM (dBc/Hz)	@ 100 kHz	f_{min}	189.1/ 188.9	186.5/ 186.5	192.5/ 191	183.3#/ 181.3#	180.4#/ 184.2#	188.9/ 187.6	188.2/ 188.7	186.3/ 189.6
		f_{max}								
	@ 10 MHz	f_{min}	190.5/ 191.4	192.2^Δ	195.6/ 196.2	190	188	**191/ 195.8**	**190.7/ 196.5**	**188/ 193.7**
		f_{max}								
1/f³ Corner (kHz)			60-130	200	100-300	120	40	70-700	60-600	50-180
Frequency pushing (MHz/V)			12-23	N.R.	4.5-15	3.6-27.3	12.6-38.2	**< 40 (from 75mV to 300mV)**		
Core Area (mm²)			0.13	0.19	0.14	0.11		0.272		

N.R. = Not Reported
* FOM = -PN + 20log$_{10}$(f_0/Δf) - 10log(P_{DC}/1mW)

† Normalized from 5 MHz offset
^Δ Best-case across the reported range
* Estimated from plots
# Calculated based on the provided P_{DC}

Figure 26.5.6: Comparison of the folded DCO with prior-art oscillators employing common-mode resonance showing the best FoM, and lowest supply of operation.

26

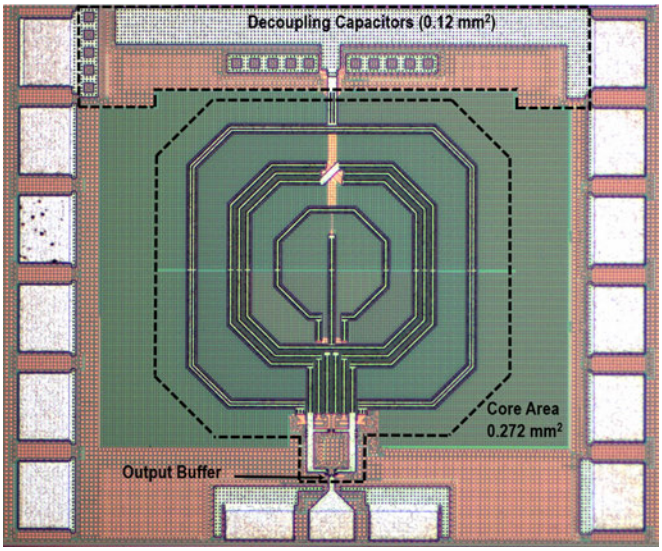


Figure 26.5.7: Micrograph of the fabricated folded-DCO die in 22nm FDSOI.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 27 Overview:
Energy Harvesting & DC/DC Control Techniques
POWER MANAGEMENT SUBCOMMITTEE

Session Chair: ***Yuan Gao***
Institute of Microelectronics, Singapore

Associate Chair: ***Chan-Hong Chern***
TSMC, Hsinchu, Taiwan

Subcommittee Chair: Yogesh Ramadass, Texas Instruments, Santa Clara, CA

The first four papers in this session focus on energy harvesting. New circuit techniques to enhance power-conversion efficiency, reduce circuit self-startup voltage and achieve adaptive circuit control for a wide range of inputs are presented. The next four papers present control techniques for envelope tracking modulator, DC-DC converter, and battery charger to improve efficiency, speed up settling and load transient response times, improve charging accuracy and charging time, etc.

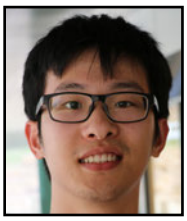
1:30 PM

27.1 An 84% Peak Efficiency Bipolar-Input Boost/Flyback Hybrid Converter with MPPT and On-Chip Cold Starter for Thermoelectric Energy Harvesting

P. Cao, Fudan University, Shanghai, China

In Paper 27.1, Fudan University presents a bipolar-input boost/flyback hybrid converter for thermoelectric energy harvesting. Implemented in 0.18μm CMOS, the circuit achieves maximum end-to-end power conversion efficiency of 84% and 79% when V_{TEG}=260mV and V_{TEG}=-300mV, respectively.

2:00 PM



27.2 An Adiabatic Sense and Set Rectifier for Improved Maximum-Power-Point Tracking in Piezoelectric Harvesting with 541% Energy Extraction Gain

Y. Peng, University of Michigan, Ann Arbor, MI

In Paper 27.2, the University of Michigan presents a Sense-and-Set rectifier that performs maximum-power-point-tracking for piezoelectric energy harvester and adapts to different excitation types and levels. The circuit is implemented in 0.18μm CMOS with 0.47mm^2 core area and achieved 5.41× and 4.59× power extraction improvement for periodic and shock excitations, compared to an ideal FBR.

2:30 PM

27.3 A Piezoelectric Energy-Harvesting Interface Using Split-Phase Flipping-Capacitor Rectifier and Capacitor Reuse Multiple-VCR SC DC-DC Achieving 9.3× Energy-Extraction Improvement

Z. Chen, University of Macau, Macau, China; Fudan University, Shanghai, China

In Paper 27.3, the University of Macau presents a Split-Phase Flipping-Capacitor Rectifier that uses 4 flying capacitors to achieve 21-phase operations and the capacitors are further reused to construct a multi-VCR buck/boost/unity SC DC-DC. Fabricated in 0.18μm CMOS, the chip demonstrates a 9.3× improvement on energy extraction compared with a conventional full-bridge rectifier interface.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

2:45 PM

27.4 Multi-Beam Shared-Inductor Reconfigurable Voltage/SECE-Mode Piezoelectric Energy Harvesting of Multi-Axial Human Motion

M. Meng, Pennsylvania State University, University Park, PA

In Paper 27.4, Pennsylvania State University presents a self-starting reconfigurable energy harvesting circuit that adaptively operates in voltage and SECE-modes for a custom inertial harvester with multiple flexible beams. Implemented in 0.35µm CMOS, the chip can improve harvested energy by 511% compared with a full-wave active rectifier.

3:15 PM

27.5 A 91%-Efficiency Envelope-Tracking Modulator Using Hysteresis-Controlled Three-Level Switching Regulator and Slew-Rate-Enhanced Linear Amplifier for LTE-80MHz Applications

P. Mahmoudidaryan, Arizona State University, Tempe, AZ

In Paper 27.5, Arizona State University presents an envelope tracking modulator using a hysteresis-control method for 3-level regulator. Fabricated in a 65nm CMOS process with an active area of 1.7mm×1.6mm, the chip has 3.2× higher switching frequency than previous PWM-controlled 3-level switching regulators in envelope tracking applications and achieves 91% efficiency for LTE-80MHz BW, ACLR of -32.5dBc at 24dBm P_{out} and 8% PAE improvement at 4dB back-off.

3:45 PM

27.6 Background Capacitor-Current-Sensor Calibration of DC-DC Buck Converter with DVS for Accurately Accelerating Load-Transient Response

Y.-W. Huang, National Cheng Kung University, Tainan, Taiwan

In Paper 27.6, National Cheng Kung University proposes background capacitor-current-sensor (CCS) calibration and ADC-based CCS techniques for a buck converter to precisely accelerate load-transient response. Fabricated in a 0.18µm CMOS process with an active area of 1.8mm^2, this chip has the smallest undershoot/overshoot voltage with 2.2× faster settling speed and achieves 6× faster calibration compared with other state of the art.

4:15 PM

27.7 A Synthesizable Digital AOT 4-Phase Buck Voltage Regulator for Digital Systems with 0.0054mm^2 Controller and 80ns Recovery Time

M. Choi, Seoul National University, Seoul, Korea; Samsung Electronics, Hwaseong, Korea

In Paper 27.7, Seoul National University and Samsung Electronics propose a digital adaptive on-time (DAOT) buck voltage regulator using ripple-injection schemes. Fabricated in a 28nm CMOS process with an active area of 0.39mm^2, this 4-phase voltage regulator achieves stable and fast output regulation under fixed switching frequency and phase synchronization of multiphase operations and exhibits 80mV droop with a recovery time of 80ns under a 1A/10ns load step up.

4:45 PM

27.8 Dynamic-Charging Current-Scaling Technique with Dual Accurate Current Control and Temperature Loops with Charging-Current Accuracy up to 99.6% for 1.6× Faster Lithium-Ion Battery Charging

W-T. Lin, National Chiao Tung University, Hsinchu, Taiwan

In Paper 27.8, National Chiao-Tung University proposes a buck-boost battery charger using dynamic charging-current scaling (DCCS) techniques to scale the charging current according to the temperature with dual accurate current control (ACC) and temperature-assisted loop (TAL). Fabricated in a 0.18µm BCD process with an active area of 3.3mm^2, the proposed ACC and TAL techniques along with Build-In Resistance-Detection and Adaptive Thermal-Loss-Calibration (ALTC) techniques achieve higher accuracy on charging current up to 99.6% and reducing charging time by 38.6%.

27.1 An 84% Peak Efficiency Bipolar-Input Boost/Flyback Hybrid Converter with MPPT and On-Chip Cold Starter for Thermoelectric Energy Harvesting

Peng Cao[1], Yao Qian[1], Pan Xue[1], Danzhu Lu[2], Jie He[2], Zhiliang Hong[1]

[1]Fudan University, Shanghai, China
[2]Analog Devices, Shanghai, China

Recent advances in energy-harvesting techniques have allowed wearable and IoT devices to operate without batteries. A thermoelectric generator (TEG) is one energy source that can provide a bipolar voltage that is proportional to the temperature difference between its two sides. Most of the prior TEG-harvesting systems could handle the positive input voltage [1-3]; however, the TEG's input voltage polarity may reverse due to environmental changes. Only a few prior publications provide a method to deal with the bipolar input voltage [4-6]. To harvest bipolar input energy, a switch-matrix-based boost converter is proposed in [4], but it cannot self-start and dynamically track the maximum power point. [5] and [6] provide methods for bipolar self-startup with an ultra-low input voltage, but both of them suffer from a low conversion efficiency. Hence, there is no complete system suitable for bipolar-input TEG energy harvesting. A solution must address three challenges: (1) the transfer of energy from a bipolar input voltage via a high-efficiency converter; (2) system self-start-up with a reasonably low bipolar input voltage; (3) a maximum power point tracking (MPPT) method for bipolar input voltage. This paper proposes a bipolar-input boost/flyback hybrid converter (BFHC) with an on-chip cold start-up for TEG energy harvesters.

Figure 27.1.1 illustrates the architecture of the TEG energy-harvesting system. The bipolar ring-oscillator-based startup circuits charge C_{DD} for self-startup; this block is disabled to save power once self-startup is complete. The BFHC extracts energy from the TEG source and transfers it to the output. The VCO generates the system clock (clk_{SYS}) according to the input voltage, and the duty cycle of S_1 in BFHC is adaptively controlled by the MPPT block. Due to MPPT's requirements the BFHC always operates in discontinuous conduction mode (DCM) mode, and a zero-current-sensing (ZCS) module is necessary to modulate the discharge time of the magnetic winding. The BFHC has 3 output voltages (V_{DD}, V_{OUT}, and V_{SUP}) for system-control circuits, the external load and for energy storage. V_{DD} is charged to 0.9V to enable control circuit operation after self-start, and V_{OUT} is regulated to an amplitude of 0.8V. If the energy harvested exceeds the load consumption, the surplus energy is then stored in C_{SUP}, which can directly share charge with V_{OUT} when the load requires.

The key challenge for a bipolar-input TEG harvesting system is how to operate and automatically switch the bipolar input voltages. To deal with this problem, a BFHC with a bipolar ring-oscillator-based self-starter is proposed as shown in Fig. 27.1.1. The system's bipolar operating mechanism is depicted in Fig. 27.1.2. The start-up stage consists of two types of ring oscillators: an N-type for $V_{IN}<0$ and a P-type for $V_{IN}>0$. Each of these is followed by a 32-stage Dickson charge pump. Thus, these two types of ring oscillators form a complementary group, which can operate over the whole bipolar input-voltage range. The output clock's amplitude approaches $|V_{IN}|$ to drive the following Dickson charge pump to charge V_{DD} to 0.66V. Three PMOS transistors are added at the end of the start-up circuit to prevent the current from flowing back into the charge pump.

After self-start, the control circuits are powered up and the system is able to detect the input-voltage polarity by comparing it with ground. The converter automatically select its operating mode based on the input polarity. When $V_{IN}<0$, the converter operates in flyback mode and current flows from the secondary side of the transformer to the output. When $V_{IN}>0$, the converter operates in boost mode and current flows directly from the inductor to the output. The design of the driver and bulk switch is also a key point for the BFHC. The high output voltage of S_1's driver is always V_{DD}, but if 0V is applied to the gate of S_1 when $V_{TEG}<0$ then S_1 will not fully shut off. Hence, there will be a large drain-to-source leakage current that reduces efficiency. In this design, V_{LX1} is used to turn off S_1 when $V_{TEG}<0$, while 0V is used to turn off S_1 when $V_{TEG}>0$. The body of S_1 is also automatically connected to V_{LX1} or 0V (the lower of the two) depending on the input polarity. Using this technique, the BFHC can quickly and smoothly switch between the two modes during a V_{IN} polarity reversal.

Figure 27.1.3 shows the operating principle of the VCO-based clock and MPPT for the bipolar input-voltage TEG energy harvester. The half-open-circuit-voltage MPPT method is used to control the duty cycle of the NMOS in BFHC. A sample-and-hold circuit periodically samples the half-voltage of V_{TEG} and stores it on C_{MPPT}. Then, the converter input voltage V_G is compared to V_{MPPT} to decide the counter's mode signal, and the counter value is updated according to the mode signal on every cycle. The duty cycle is determined by the final value of the counter so that V_G hold wave nearby the sampled voltage V_{MPPT} in order to track the maximum power point. To further improve the efficiency over the whole input-voltage range, the system clock must be dynamically adjusted based on the input voltage; the theoretical optimal frequency is given in the right-half of Fig. 27.1.3. For this design, a frequency that is proportional to $|V_G|$ is used to obtain an approximately linear model. An absolute-voltage-to-current generator (AVCG) converts the bipolar voltage V_G into a current ($|V_G|/R$). The AVCG is a reconfigurable operational circuit that operates as a unit-scale inversely-proportion amplifier when $V_G<0$, and as a voltage follower when $V_G>0$. Its mode of operation is automatically selected via CMP_{POLAR}. A current controlled ring oscillator (CCRO) follows to generate a clock, who's frequency is proportional to $|V_G|$.

The test chip is fabricated in 0.18μm CMOS. The bipolar cold-start-up waveforms at minimum V_{TEG} are shown in the top two waveforms of Fig. 27.1.4. The minimum self-start-up voltage are +129mV and -140mV. The bottom two waveforms in Fig. 27.1.4 shows that the hybrid converter can adaptively and quickly switch modes during an input polarity reversal, and also show that it operates normally during a load current step. As shown in Fig. 27.1.5, this energy harvester achieves a peak efficiency of 84% at V_{TEG}=260mV, and 79% V_{TEG}=-300mV. A comparison table to state-of-the-art systems is shown in Fig. 27.1.6. This work features high-efficiency and a fully-integrated cold start-up circuit among the bipolar-input TEG harvesters listed. Figure 27.1.7 shows the chip micrograph.

Acknowledgements:
The authors would like to thank Analog Devices for supporting this research, and to acknowledge ICC for MPW assistance.

References:
[1] J. Katic, et al., "A Dual-Output Thermoelectric Energy Harvesting Interface With 86.6% Peak Efficiency at 30μW and Total Control Power of 160nW," *JSSC*, vol. 51, no. 8, pp. 1928-1937, Aug. 2016.
[2] A. Shrivastava, et al., "A 10 mV-input boost converter with inductor peak current control and zero detection for thermoelectric and solar energy harvesting with 220 mV cold-start and -14.5 dBm, 915 MHz RF kick-start," *JSSC*, vol. 50, no. 8, pp. 1820-1832, Aug. 2015.
[3] Y. Lu, et al., "A 200nA Single-Inductor Dual-Input-Triple-Output (DITO) Converter with Two-Stage Charging and Process-Limit Cold-Start Voltage for Photovoltaic and Thermoelectric Energy Harvesting," *ISSCC*, pp. 368-369, 2016.
[4] M. Alhawari, et al., "An Efficient Polarity Detection Technique for Thermoelectric Harvester in L-Based Converters," *IEEE TCAS-I*, vol. 64, no. 3, pp. 705-716, Mar. 2017
[5] Analog Devices, "LTC3109 Auto-Polarity, Ultralow Voltage Step-Up Converter and Power Manager,"
http://www.analog.com/media/en/technical-documentation/data-sheets/3109fb.pdf, 2012.
[6] NV. Desai, et al., "A Bipolar ±40 mV Self-Starting Boost Converter with Transformer Reuse for Thermoelectric Energy Harvesting," *ISLPED*, pp. 221-226, 2014.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

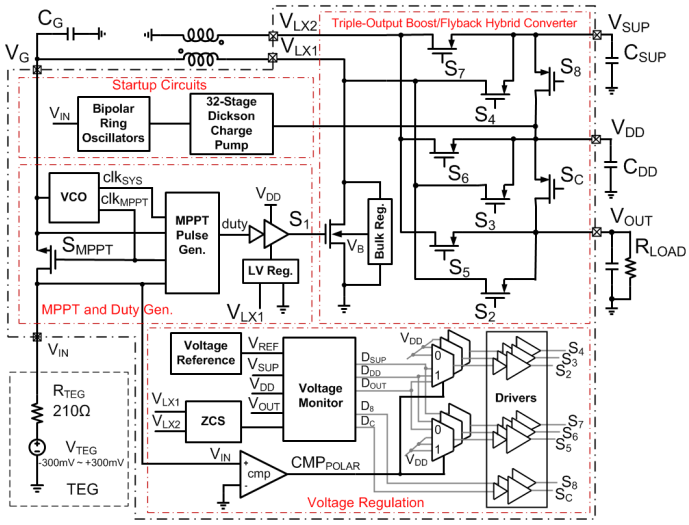


Figure 27.1.1: Architecture of the bipolar-input TEG energy harvester.

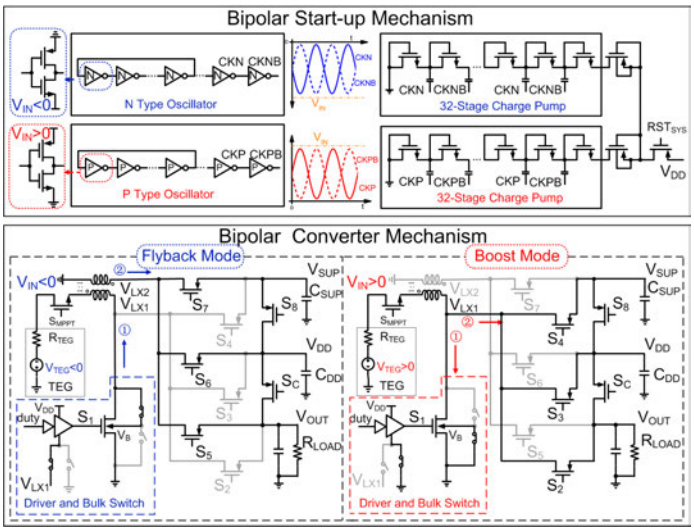


Figure 27.1.2: Bipolar operating mechanism.

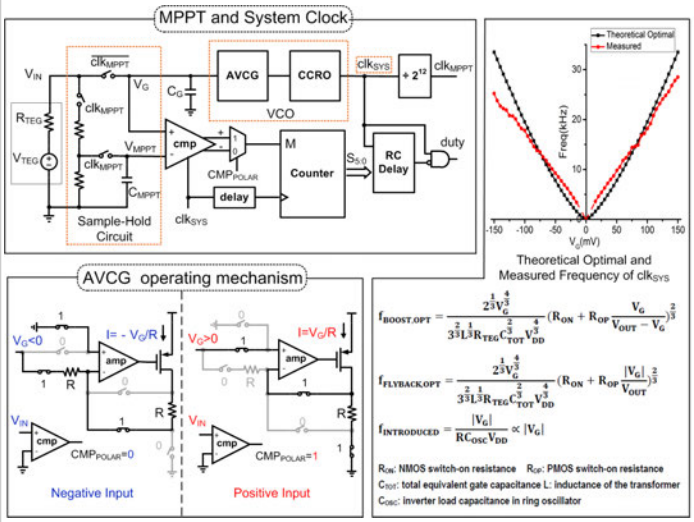


Figure 27.1.3: VCO based clock and MPPT method.

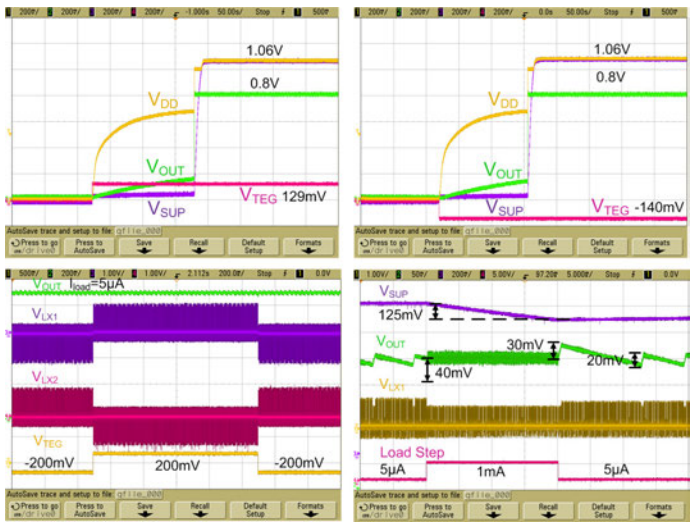


Figure 27.1.4: Measured bipolar start-up (top left and top right), input polarity reversal (bottom left), load step (bottom right) waveforms.

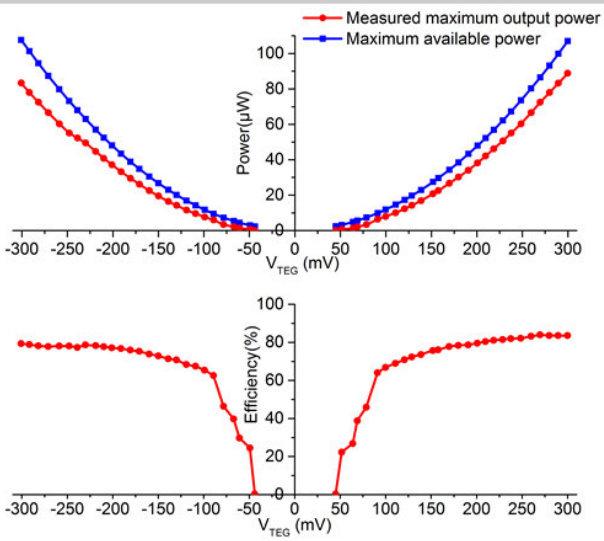


Figure 27.1.5: Measured maximum output power and conversion efficiency of the harvesting system.

	JSSC 2016 [1]	JSSC 2015 [2]	ISSCC 2016 [3]	TCAS-I 2017 [4]	Analog Devices [5]	ISLPED 2014 [6]	**This Work**
Process	0.18μm	0.13μm	0.35μm	65nm	N/A	N/A	**0.18μm**
R_{TEG}	210Ω	2Ω	N/A	7Ω	2Ω	1-5Ω	**210Ω**
MPPT	yes	yes	yes	yes	no	yes	**yes**
Adaptive switching frequency	no	yes	N/A	no	N/A	N/A	**yes**
Min Input Voltage V_{IN}	15mV	10mV	30mV	±20mV	±30mV	±30mV	**±25mV**
Auto-Polarity Method	Not Implemented	Not Implemented	Not Implemented	Switch Matrix	Two Parallel Flyback Converters	Flux-commutating Flyback Topology	**Boost/ Flyback Hybrid Converter**
Magnet Components	1 Inductor	1 Inductor	1 Inductor	1 Inductor	2 Transformers	1 Transformer	**1 Transformer**
Peak Efficiency (end to end)	86.6%	83% (boost only)	85% (boost only)	70%	55% (flyback only)	68% (flyback only)	**84%**
Start-up Mechanism	Not Implemented	On-chip Ring Oscillator	POR	Not Implemented	Off-chip Resonant Oscillator	Off-chip Meissner Oscillator	**On-chip Ring Oscillator**
Start-up Voltage	-	220mV	300mV	-	±30mV	±40mV	**+129mV, -140mV**

Figure 27.1.6: Performance comparisons with state-of-the-art works.

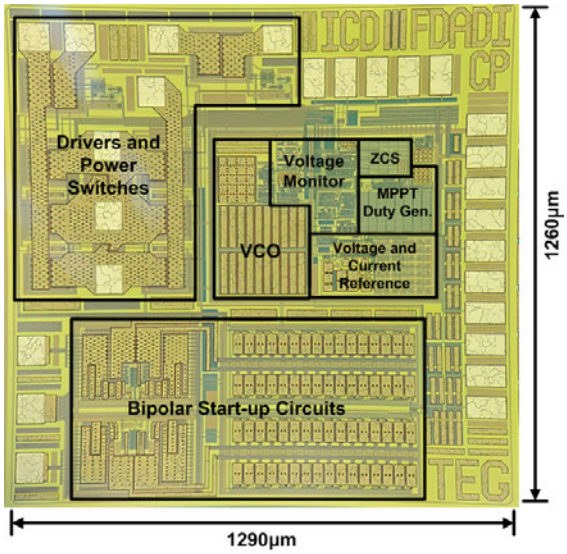


Figure 27.1.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

27.2 An Adiabatic Sense and Set Rectifier for Improved Maximum-Power-Point Tracking in Piezoelectric Harvesting with 541% Energy Extraction Gain

Yimai Peng, David Kyojin Choo, Sechang Oh, Inhee Lee, Taekwang Jang, Yejoong Kim, Jongyup Lim, David Blaauw, Dennis Sylvester

University of Michigan, Ann Arbor, MI

Piezoelectric energy harvesters (PEHs) convert mechanical energy from vibrations into electrical energy. They have become popular in energy-autonomous IoT systems. However, the total energy extracted by a PEH is highly sensitive to matching between the PEH impedance and the energy extraction circuit. Prior solutions include the use of a full-bridge rectifier (FBR) and a so-called synchronous electric-charge extraction (SECE) [1], and are suitable for non-periodic vibrations. However, their extraction efficiency is low since the large internal capacitance C_p (usually 10's of nF) of the PEH (Fig. 27.2.1) prevents the output voltage from reaching its maximum power point (MPP) under a typical sinusoidal and transient excitation ($V_{MPP} = ½·I_pR_p$). A recently proposed technique [2,3,4], called bias-flip, achieves a higher extraction efficiency by forcing a predetermined constant voltage at the PEH output, V_p, which is then flipped every half-period of the assumed sinusoidal excitation (Fig. 27.2.1, top left). To flip V_p, the energy in capacitor C_p is extracted using either a large external inductor [2,3] or capacitor arrays [4]. It is then restored with the opposite polarity (Fig. 27.2.1, top). However, V_{MPP} of the PEH varies with sinusoidal current I_p; hence, the two fixed values of V_p in the flip-bias technique either over- or underestimate V_{MPP} for much of the oscillation cycle (pattern filled regions in Fig. 27.2.1, top right). In addition, none of the prior approaches compensate for V_{MPP}-waveform amplitude changes, due to input intensity variations or decaying oscillations after an impulse, further degrading efficiency.

This paper proposes a sense-and-set (SaS) MPP-tracking interface circuit that extracts near-optimal power from the PEH. The SaS circuit dynamically measures V_{MPP} and sets the PEH output voltage (V_p) to match and follow the V_{MPP} waveform (step A, Fig. 27.2.1 bottom). After adjusting V_p to V_{MPP}, all switches are turned off and I_p naturally charges C_p (step B), achieving near-optimal extraction efficiency. The circuit harvests energy from C_p after a short, pre-determined time ($t_{A.1}$) to maintain V_p near V_{MPP} (step C). To reduce energy overhead, SaS performs all operations adiabatically using only a single external inductor and capacitor. Since SaS does not presume any particular PEH current waveform, it applies to both periodic and non-periodic use cases. SaS was implemented in 180nm CMOS and achieves a 5.41× power-extraction improvement (FOM) for periodic excitations and a 4.59× improvement for shock excitations compared to an ideal FBR.

The key challenge for efficient harvesting is repeatedly determining V_{MPP} that changes with the PEH current variation, with low overhead. SaS performs a series of steps (A.1–A.4, Fig. 27.22) that measure I_p and convert it to a voltage V_T on capacitor C_T, which becomes a reference for charging or discharging V_p to reach V_{MPP}. Initially in step A.1, S7 and S5 are closed and C_p and C_p' are charged to equal potentials. Then, S7 is opened for a brief time ($t_{A.1}$ ≈100μs) and I_p only charges C_p. Since I_p determines V_{MPP}, and observing that I_{Rp} can be approximated as constant during $t_{A.1}$, V_{MPP} is derived from the resulting voltage difference ΔV between C_p and C_p', as expressed by Eq. 1 shown in Fig. 27.2.2. Usually ΔV is very small (~1mV), since C_p is large and it is desirable to keep $t_{A.1}$ short, hence it is challenging to resolve using a conventional comparator or ADC. However, the stored energy difference between C_p and C_p' is relatively large (~pJ), thus we propose using a charge-based amplifier to transfer this energy difference into a small capacitor (C_T) using the inductor (L) in step A.3, thereby obtaining a much higher voltage V_T. In the absence of parasitics we obtain $V_{MPP} = ½·(V_T + V_1)$ (a variation of Eq. 3 in Fig. 27.2.2) by appropriately sizing $C_T = ½·C_p·(t_{A.1}/R_pC_p)^2$. After we have generated V_{MPP} from V_T and V_1 (as shown in Fig. 27.2.3) we adjust V_p in step A.4 to V_{MPP} adiabatically using the same inductor.

Figure 27.2.3 shows the circuit details for SaS. All switches are implemented using transmission gates and are controlled using a pre-configured sequence of pulses. A 5-stage ring oscillator and pulse generator produce the correct pulse length and sequence. Both of these employ leakage-based invertor cells [5] to ensure high energy-efficiency across a very wide frequency range (10-100kHz). Two clocked comparators further modulate (truncate) critical pulses that operate at a faster frequency (5MHz). Optimally sized drivers then buffer the pulses to generate the final switch-control signals. To generate $V_{MPP} = ½·(V_T + V_1)$ (Eq. 3 in Fig. 27.2.2), V_1 is sampled by capacitor C_{V1} (same size as C_T) and shorted to C_T. However, the off-chip inductor L introduces significant parasitics to ground, reducing the value of V_T. We compensate for this using the series-parallel structure (Fig. 27.2.3) and adjusting C_{V1} appropriately to generate $V_{MPP} = ½·n·(V_T + V_1)$, where n can be tuned between 1-5. Comparator CMP_1 then compares PEH output V_p with the generated V_{MPP} to set $V_p = V_{MPP}$ (Fig. 27.2.3, red path). Once V_p arrives at the correct value, a recycling path for the remaining energy in L is turned on until the second comparator CMP_2 detects the zero-crossing point of the inductor current necessary for adiabatic operation (Fig. 27.2.3, green path).

The value of C_p varies across the PEH type; hence, for each new PEH C_T must adjusted. We enable this by first choosing an appropriate C_T from an on-chip capacitor array (0.2–9.2pF with 0.2pF steps, omitted from Fig. 27.2.3 for clarity), followed by tuning the duration of step A.1 ($t_{A.1}$). The proposed harvester automatically performs this tuning by running through steps A.1–A.4 using a modified operation of the series-parallel capacitor structure such that V_p is set to V_{oc} ($2·V_{MPP}$) instead of V_{MPP}. Then, an additional sensing step A.1–A.3 is applied and the polarity of V_T is determined. If $t_{A.1}$ is tuned correctly and V_p is set at V_{oc}, then $\Delta V = 0$ at the end of step A.1; therefore, $V_T = 0$ after step A.3. By repeating this process while sweeping the value of $t_{A.1}$ and observing the change in sign of V_T, SaS automatically finds the point where $\Delta V = 0$ and $t_{A.1}$ is correctly tuned.

When I_p changes polarity so does V_{MPP}, which results in additional operating complexity for both a positive and negative V_{MPP}. Instead, SaS detects the I_p sign change by comparing V_{MPP} against ground with comparator CMP_1 (for simplicity switches that connect ground to the comparator'sinverse input are not shown in Fig. 27.2.3). SaS then performs two additional steps: (1) it extracts the remaining energy from C_p into L, and (2) it flips the two terminals of the PEH (flip switch in Fig. 27.2.3) and restores the extracted energy. It then proceeds as before, tracking a positive V_{MPP}. As a result, no negative voltages, thus simplifying harvester design.

The proposed SaS circuit is fabricated in 180nm CMOS and occupies 0.47mm². It is tested with a commercial PEH (Mide PPA-1022) on a shaker table (Unholtz-Dickie model 650) with a 10.6g tip mass. Figure 27.2.4 shows the measured MPPT-operation waveform for both periodic (85Hz) and shock excitations (1 pulse/s). During the start-up phase the maximum value of V_P is limited by V_{out}. As V_{out} builds the efficiency of the SaS circuit gradually increases and eventually tracks the MPP. After start-up the SaS circuit first auto-calibrates to adapt to the specific PEH parasitics (C_p and R_p) by tuning $t_{A.1}$. Following calibration, V_p dynamically adjusts to various PEH input characteristics across different excitation types and amplitudes, including weak amplitudes that are typically difficult to track (Fig. 27.2.4, weak periodic). Due to voltage limitations of our CMOS process, V_p is limited to 2V, even though the MPP voltage reaches higher than that for strong amplitudes (Fig. 27.2.4 strong periodic and shock input). An external voltage supply is used for accurate measurement of the SaS circuits' power overhead. Measured leakage and active power for the SaS circuit under a 2V-supply voltage are 7nW and 230nW.

Figure 27.2.5 (top) shows SaS measured output power across different output voltages (V_{OUT}) and different excitation levels. As long as $V_{OUT} < V_{MPP}$ the SaS circuit can perform true MPPT and extract the maximum power available. For stronger excitations ($V_{MPP} > V_{OUT}$) SaS operates off-MPPT due to the aforementioned CMOS voltage limit, but still harvests energy at a higher efficiency than a conventional FBR or a bias-flip circuit. Note that the circuit could be reconfigured in step A.4 to serve as a boost converter (requiring an additional switch configuration), or implemented with a high voltage process to extend the voltage range. Figure 27.2.5 (bottom left) shows the measured output power using different MPP tracking frequencies. Higher tracking frequencies allow SaS to adjust V_p more frequently and track V_{MPP} more accurately, improving harvesting efficiency, but this also leads to larger losses during energy transfer: due to both conduction and switching losses. The optimal frequency was found to be 2.5kHz for a 85Hz excitation.

Measurements show a 5.41 and 5.56× power extraction improvement (FoM) compared to an ideal FBR with strong and weak periodic excitation input. Shock excitation improvement is calculated to be 4.59×. Since the FoM depends on the PEH Q value (ωR_pC_p), we also report the improvement normalized to Q in Fig. 27.2.6. The proposed SaS circuit achieves nearly 2× energy extraction improvement, based on a Q-normalized FoM, compared to prior-art. It also offers the unique and practical ability to adapt to different excitation types and levels dynamically.

Acknowledgments:
The authors would like to thank Robert Gordenker and Professor Khalil Najafi at the University of Michigan, for their support and contribution to this project.

References:
[1] A. Quelen, et al., "A 30nA Quiescent 80nW-to-14mW Power-range Shock-optimized SECE-based Piezoelectric Harvesting Interface with 420% Harvested-energy Improvement," *ISSCC*, pp. 120-121, 2018.
[2] Y.K. Ramadass, et al., "An Efficient Piezoelectric Energy Harvesting Interface Circuit Using a Bias-Flip Rectifier and Shared Inductor," *JSSC*, vol. 45, no. 1, pp. 189-204, Jan. 2010.
[3] S. Du, et al., "A fully integrated split-electrode synchronized-switch-harvesting-on-capacitors (SE-SSHC) rectifier for piezoelectric energy harvesting with between 358% and 821% power-extraction enhancement," *ISSCC*, pp. 152-154, 2018.
[4] Z. Chen, et al., "A 1.7mm² Inductorless Fully Integrated Flipping-capacitor Rectifier (FCR) for Piezoelectric Energy Harvesting with 483% Power-extraction Enhancement," *ISSCC*, pp. 372-373, 2016.
[5] I. Lee, et al., "A Constant Energy-Per-Cycle Ring Oscillator Over a Wide Frequency Range for Wireless Sensor Nodes," *JSSC*, vol. 51, no. 3, pp. 697-711, Mar. 2016.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

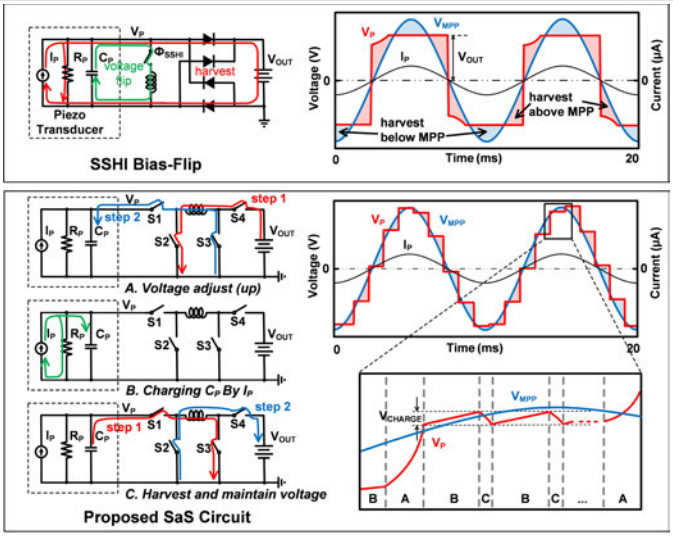


Figure 27.2.1: Circuit diagram and operation waveform for PEHs with SSHI (bias-flip) technique and proposed SaS circuit.

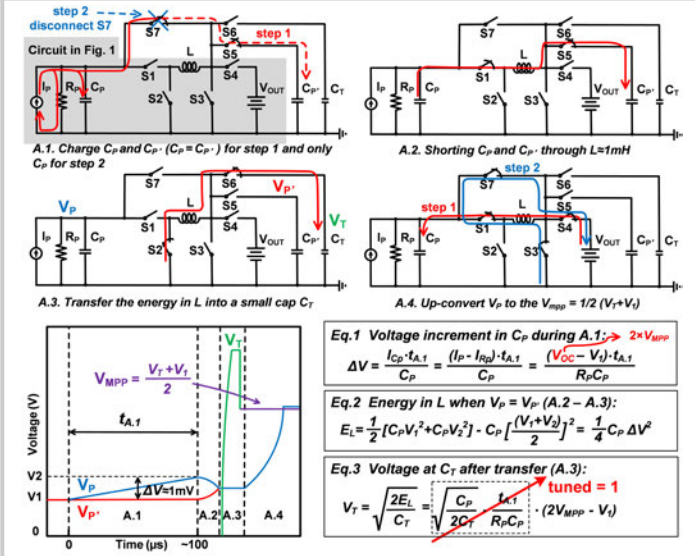


Figure 27.2.2: SaS operation sequences/waveforms for MPP-tracking.

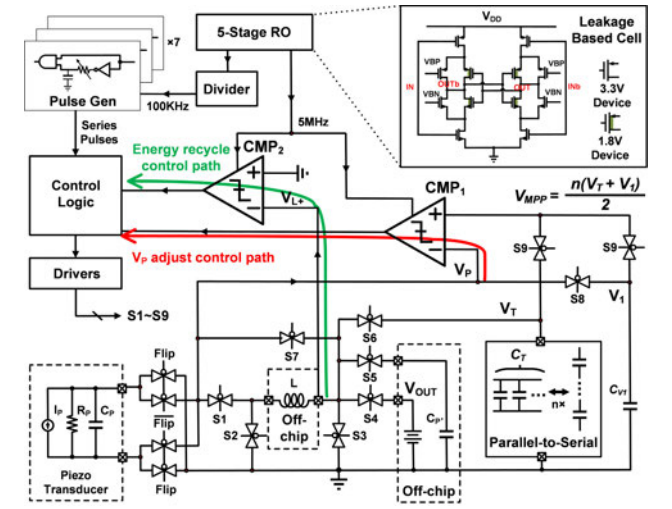


Figure 27.2.3: Proposed SaS circuit details: Inductor-sharing switches, leakage-based low power clock/timing circuit and clocked-comparator circuit for switch control.

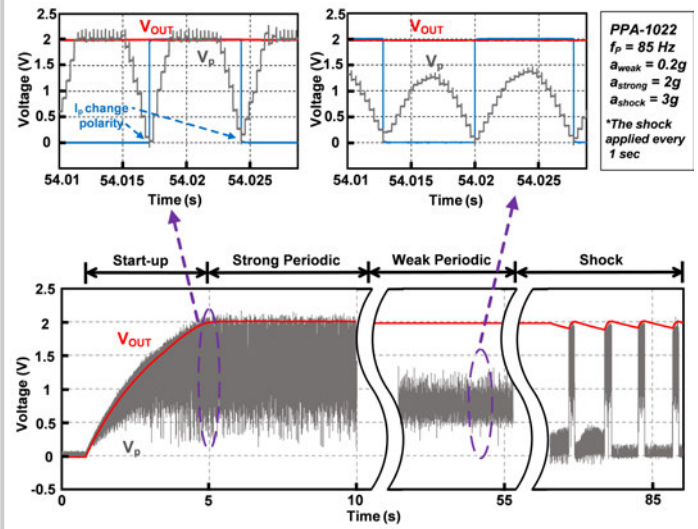


Figure 27.2.4: Measured waveform of MPPT process for strong periodic, weak periodic and shock excitation by the proposed SaS circuit.

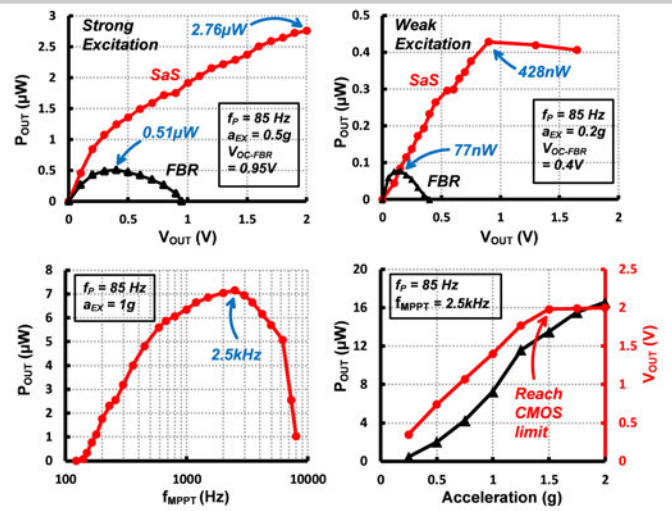


Figure 27.2.5: Measured SaS harvested power with different excitation types, acceleration levels and MPP tracking frequency, compared to a full-bridge rectifier.

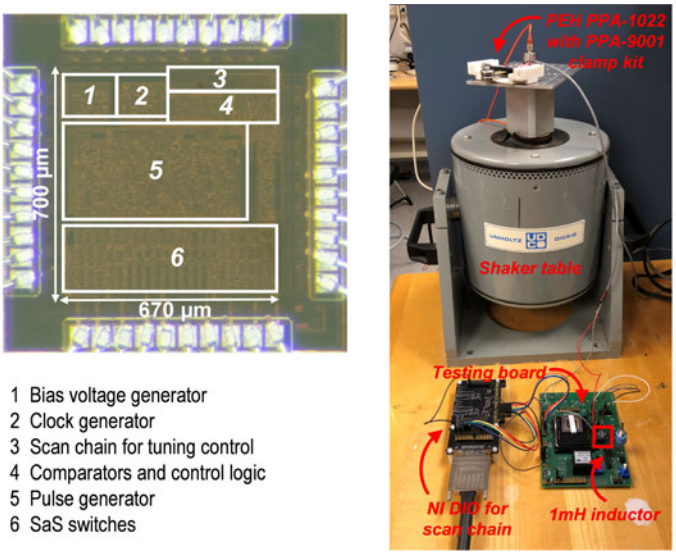
References	This Work		ISSCC'18[1]	JSSC'10[2]	ISSCC'16[3]	ISSCC'18[4]
Technology	0.18 μm		40 nm	0.35 μm	0.35 μm	0.18 μm
Energy Extract Technique	SaS		SECE	SSHI	SSHI	SSHC
Technique's Key Component	L = 1 mH $C_{P'}$ = 10 nF		L = 2.2 mH	L = 0.82 mH	L (value not reported)	C_{flip} = 8× C_p
PEH Type	MIDE PPA-1022		MIDE PPA-1011	MIDE V22B	MIDE V21B & V22B	CUSTOM MEMS
C_p	8 nF		43 nF	9 nF	9 – 26 nF	1.94 nF
Frequency	85 Hz		75.4 Hz	225 Hz	134.6 - 229.2 Hz	219 Hz
Excitation Type	Periodic & Shock		Periodic & Shock	Periodic	Periodic & Shock	Periodic
FoM [a] (Periodic)	(strong) **5.41×**	(weak) **5.56×**	3.14×	4×	4.40×	8.21×
FoM / Q_{PEH} [b] (Periodic)	**0.316**	**0.325**	0.132	0.168	0.119	Unknown
PEH V_{OC} for Reported FoM	**0.95 V**	**0.4 V**	2.85 V	2.4 V	0.95 V	2.5 V
FoM (Shock)	**4.59×**		4.20×	N/A	2.69×	N/A
Auto-maintain MPPT [c]	**Yes**		No	No	No	No

(a) $FoM = \frac{\max(P_{out})}{\max(P_{FBR})}$, which for periodic excitation can also be calculated by $\frac{\max(P_{out})}{C_p f_P V_{oc}^2}$

(b) Q_{PEH} is measured for this work and calculated for prior arts, according to the reported PEH type and frequency;

(c) SECE and Bias-Flip need to lower V_{OUT} for smaller excitations to maintain MPPT, while SaS does not require this.

Figure 27.2.6: Measured performance and comparison table to prior-art.



1 Bias voltage generator
2 Clock generator
3 Scan chain for tuning control
4 Comparators and control logic
5 Pulse generator
6 SaS switches

Figure 27.2.7: Die photograph and testing setup.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

27.3 A Piezoelectric Energy-Harvesting Interface Using Split-Phase Flipping-Capacitor Rectifier and Capacitor Reuse Multiple-VCR SC DC-DC Achieving 9.3× Energy-Extraction Improvement

Zhiyuan Chen[1,2], Yang Jiang[1], Man-Kay Law[1], Pui-In Mak[1], Xiaoyang Zeng[2], Rui P. Martins[1,3]

[1]University of Macau, Macau, China
[2]Fudan University, Shanghai, China
[3]Instituto Superior Tecnico/University of Lisboa, Lisbon, Portugal

Piezoelectric energy harvesters (PEH) exhibit promising features to scavenge the ambient vibration energy for ubiquitous miniaturized internet of things (IoT) devices. For the traditional PEH using a full-bridge rectifier (FBR), the inherent capacitance (C_P) limits the extractable AC-DC electrical power. By extending the damping duration, PEH interfaces employing non-linear techniques, such as the parallel-synchronized-switch harvesting-on-inductor (P-SSHI) [1], can increase the harvestable energy. However, they typically require bulky external high-Q inductors to enhance the extracted power. Recently, various inductor-less PEH interfaces exploiting only capacitors for flipping the PEH voltage during the zero-crossing of the PEH current (I_P) have been reported [2-4]. However, their achievable energy extraction improvement depends highly on the number of flipping phases. This work proposes a split-phase flipping-capacitor rectifier (SPFCR) implementation entailing only 4 capacitors, while achieving the highest number of phases (21) when compared to prior art [2-4]. To resolve the lack of input power (P_{in}) adaptability in [2-4], this work proposes a capacitor-reuse multiple voltage-conversion-ratio (VCR) switched-capacitor (SC) DC-DC converter to reduce the charge redistribution loss. Maximum power point tracking (MPPT) is also accomplished using the fractional FBR open circuit voltage ($V_{OC,FBR}$) for relaxed voltage tolerance, while raising the PEH extracted energy. This work demonstrates a measured 9.3× energy-extraction improvement when compared to a conventional FBR interface.

Figure 27.3.1 depicts our PEH system that consists of 4 flying capacitors ($C_{1\text{-}4}$), an AC-DC rectifier, a mode selector, a phase generator, two switch arrays with drivers, as well as a MPPT arbiter. Generally, a DC-DC converter is necessary to reduce the efficiency loss due to the mismatch between the system output V_{OUT} and the PEH MPP voltage (V_{MPP}). As the PEH voltage flipping time (t_{FLIP}) is typically less than 10% of the external excitation period to ensure a large conduction time, both SPFCR and SC DC-DC operations can be realized using the same capacitors ($C_{1\text{-}4}$). Theoretically, PEH achieves maximum power at $V_{OC}/2$, and the fractional V_{OC} MPPT approach is widely used for its simplicity. However, as V_{OC} is twice the nominal V_{MPP}, high voltage devices or device stacking is necessary; inevitably jeopardizing the system cost and/or robustness. This problem is even worse for PEH interfaces similar to P-SSHI, with the PEH voltage biased to be much higher than that in conventional FBR implementations. As shown in Fig. 27.3.1, there exists an empirical ratio between $V_{MPP,SPFCR}$ and $V_{OC,FBR}$ (depending on the PEH characteristics and interface parameters) which is almost constant over a wide P_{IN} range. We exploit this correlation for efficient MPPT without sustaining a high $V_{OC,SPFCR}$.

For capacitive PEH interfaces, different flying capacitors (C_{FLY}) are conventionally connected to bias the PEH with equal voltage steps [2-4], leading to inefficient voltage flipping. Figure 27.3.2 presents the capacitor reconfiguration examples from [2] and the proposed capacitor phase splitting technique using three C_{FLY} during the first-half of the positive transition cycle (PTC) of I_P. By systematically arranging C_{FLY}, we can generate extra augmented phases (i.e. connecting different number of flying capacitors in series) and extended phases (i.e. connect one single capacitor in each phase instead of parallel connections) during the sharing/recharging period. This work implements a 21-phase SPFCR using only 4 flying capacitors (out of the theoretical 31) taking into consideration the voltage flipping efficiency and the implementation complexity.

Figure 27.3.3 displays the MPPT arbiter and its operation, which harnesses the correlation between $V_{OC,FBR}$ and V_{MPP}. It comprises of a C_P reset circuit, a peak detector, a capacitive divider, and a 3-level ADC with processing. The exertion of V_{RST} triggers V_M when $I_P = 0$, enabling the MPPT arbiter while disconnecting the rectifier and capacitors from the PEH. The peak detector samples the peak-to-peak PEH V_{OC}, as in conventional FBR operation. The sampled voltage (V_S) is further processed by the capacitive divider, and subsequently compared with V_{REF} to determine the best VCR that matches V_{OUT} with $V_{MPP,SPFCR}$. One extra excitation cycle is used to reduce the residual V_S error during sampling. V_{OUT} is regulated using pulse frequency modulation (PFM) with an external f_S for testing flexibility. The pulse generator block employs differential pulse-delay cells (10 each) followed by a digital processing block to generate the required 21 sequential digital pulses, which can be also adjusted externally. The mode selection block generates V_{MODE} by utilizing the ϕ_{-10} and ϕ_{10} signals, with proper delays, to ensure robust mode switching. For the SC DC-DC converter, the implemented VCRs are {2, 1, 2/3, 1/3}, mainly determined by the flying capacitor voltages at the end of the 21-phase SPFCR ($V_{C1} = 0.27 \cdot V_{RECT}$, $V_{C2} = 0.32 \cdot V_{RECT}$, $V_{C3} = 0.19 \cdot V_{RECT}$ and $V_{C4} = 0.1 \cdot V_{RECT}$). This can reduce the charge redistribution loss due to the capacitor voltage mismatch during the SPFCR and SC DC-DC operations. Figure 27.3.3 also shows the capacitor utilization at different VCRs. At 2×, C_{EQ1} (i.e. $C_{1\text{-}4}$ connected in series) is charged up during ϕ_C, and stacked on top of V_{RECT} to generate V_{OUT} at ϕ_D. This can significantly reduce the charge redistribution loss when compared with parallel-connected $C_{1\text{-}4}$. Similarly, we use C_{EQ2} (C_2) and C_{EQ3} (C_3 in series with C_4) at VCR = $^1/_3$ and $^2/_3$, and disconnect all $C_{1\text{-}4}$ at VCR = 1.

The proposed PEH interface with a 21-phase SPFCR is fabricated in 0.18μm 1.8/3.3/6V CMOS, occupying an area of 0.2mm^2. We placed the MIDE PEH (PPA-1021) on a shaker at an excitation frequency (f_{EX}) of 200Hz. Each $C_{1\text{-}4}$ is 68nF, and the corresponding C_P, C_R and C_L are 22, 100 and 30nF. The 1.5V supply powers the phase generator for flexible delay adjustment. Figure 27.3.4 plots the SPFCR operation with the 21-phase voltage flipping steps at $V_{OUT} = 2V$ and an acceleration (a_G) of 0.12g; as well as the corresponding PEH voltage swing before (VCR = 2) and after (VCR = $^2/_3$) MPPT. The extracted ratio between $V_{MPP,SPFCR}$ and $2 \cdot V_{OC,FBR}$ from measurement demonstrates a pseudo-constant relationship within 2.2-2.35 with $P_{IN,FBR} > 0.4\mu W$ (a_G varying from 0.08-0.19g), which can lead to ~2% energy loss when compared to the exact MPP case. As $P_{IN,FBR}$ drops, the ratio reduces due to the excessive control loss, resulting in up to 25% MPPT energy loss when $P_{IN,FBR} = 0.1\mu W$ ($a_G = 0.05g$). As observed, the $V_{RECT,SPFCR}$ (3.2V) is successfully biased to ~2.3× of $2 \cdot V_{OC,FBR}$ (1.4V) after MPPT, which is consistent with the extracted $V_{MPP,SPFCR}$. Figure 27.3.5 shows the measured P_{OUT} under different V_{RECT}, achieving a MOPIR of up to 9.3× at $a_G = 0.12g$ and $V_{OUT} = 1V$. The MPP at $V_{OC,SPFCR}/2$ can clearly be observed, and $V_{OC,SPFCR}$ can exceed the device breakdown voltage at strong vibrations ($a_G = 0.16g$). With a wide P_{IN} range (measured using FBR), the proposed PEH system can improve the achievable MOPIR using different VCRs, with a slight drop at low $P_{IN,FBR}$ due to the increased intrinsic loss. The FCR can deliver up to 64μW at $f_{EX} = 200Hz$ and $a_G = 0.16g$. Figure 27.3.6 show a table summarizing key chip characteristics and performance benchmarks. Without a bulky external inductor, this work achieves a high MOPIR of 9.3× compared to [2-3]. Instead of 8 capacitors to achieve 17 phases as in [4], this work demonstrates a 21-phase SPFCR operation using half the capacitors. It also features multiple-VCR SC DC-DC and $V_{OC,FBR}$-based MPPT, improving the system P_{IN} adaptability without using extra passives nor tolerating excessively high V_{OC}. Figure 23.3.7 shows the die micrograph.

Acknowledgement:
The authors thank the Macau FDCT (FDCT069/2016/A2) and the University of Macau (MYRG2018-AMSV-00196) for financial support.

References:
[1] D. A. Sanchez, et al., "A 4-μW-to-1mW Parallel-SSHI Rectifier for Piezoelectric Energy Harvesting of Periodic and Shock Excitations with Inductor Sharing, Cold Start-up and up to 681% Power Extraction Improvement," *ISSCC*, pp. 336-337, 2016.
[2] Z. Chen, et al., "Fully-Integrated Inductor-less Flipping-Capacitor Rectifier (FCR) for Piezoelectric Energy Harvesting," *JSSC*, vol. 52, pp. 3168-3180, Dec. 2017.
[3] S. Du, et al., "A Fully Integrated Split-Electrode Synchronized-Switch-Harvesting-on-Capacitors (SE-SSHC) Rectifier for Piezoelectric Energy Harvesting with Between 358% and 821% Power-Extraction Enhancement," *ISSCC*, pp. 152-153, 2018.
[4] S. Du, et al., "An Inductorless Bias-Flip Rectifier for Piezoelectric Energy Harvesting," *JSSC*, vol. 52, pp. 2746–2757, Oct. 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

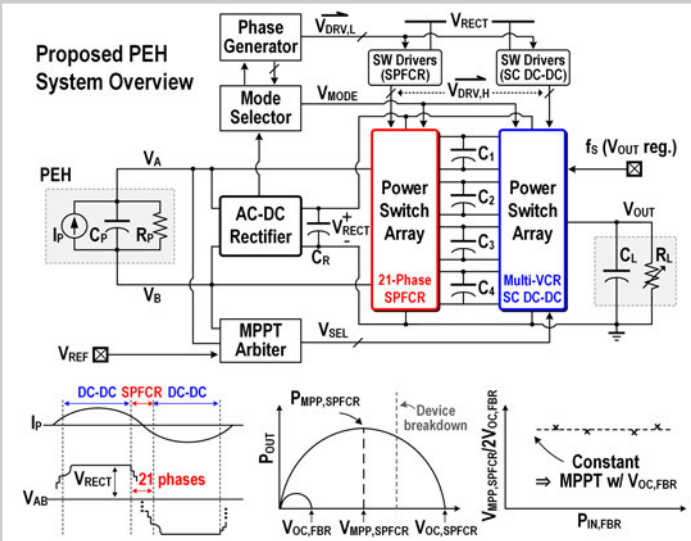


Figure 27.3.1: Proposed split-phase flipping-capacitor rectifier (SPFCR) for piezoelectric energy harvesting with capacitor-reuse SC DC-DC and MPPT.

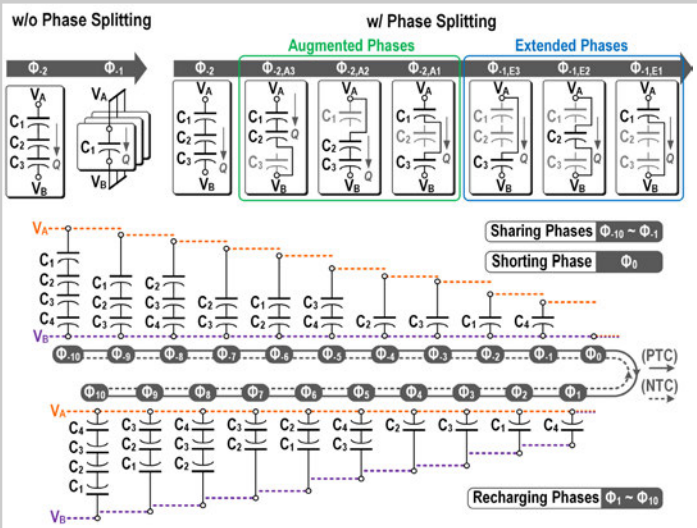


Figure 27.3.2: Proposed SPFCR with 3 C_{FLY} during half PTC (top); and 21-phase SPFCR for both the positive/negative transition cycles (PTC/NTC) (bottom).

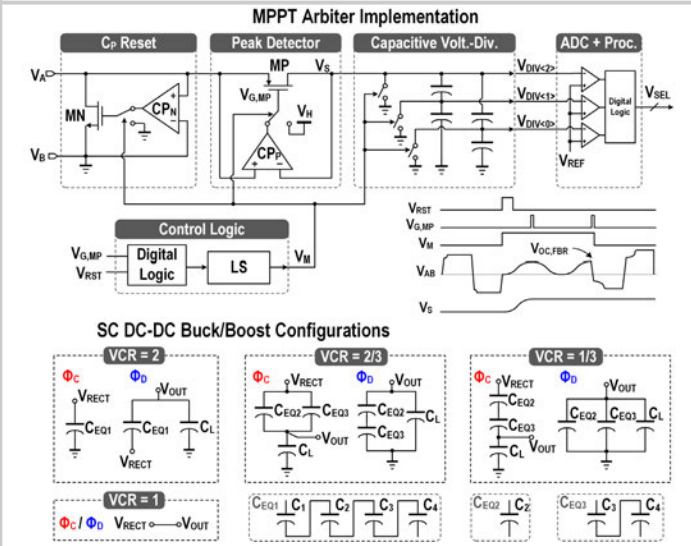


Figure 27.3.3: MPPT arbiter implementation (top) and SC DC-DC configurations (bottom).

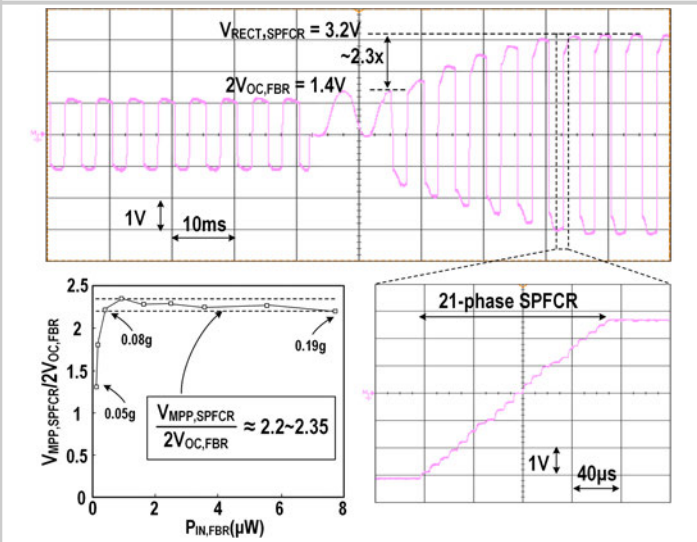


Figure 27.3.4: Measured MPPT operation (top); the 21-phase SPFCR and the extracted ratio between $V_{MPP,SPFCR}$ and $2 \cdot V_{OC,FBR}$ at VCR = 1 (bottom).

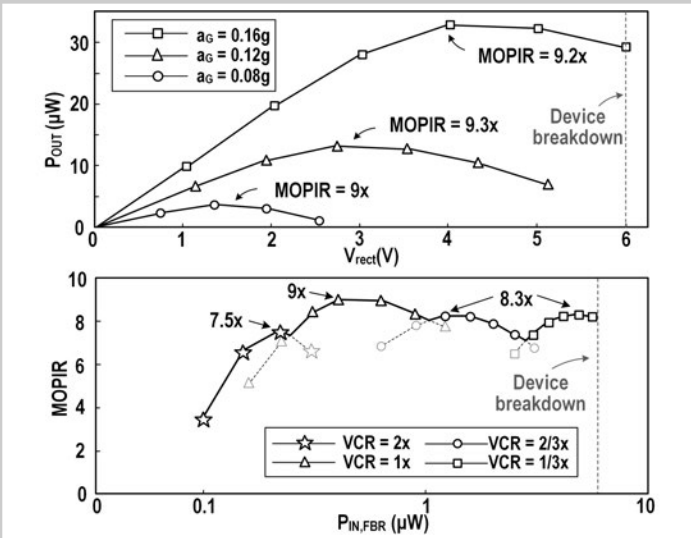


Figure 27.3.5: Measured P_{OUT} vs. V_{RECT} at different a_G (top); and MOPIR vs. $P_{IN,FBR}$ with VCR = {2, 1, $^2/_3$, $^1/_3$} at a_G = 0.05-0.19g and V_{OUT} = 2V (bottom).

	This work	JSSC'17 [4]	JSSC'17 [2]	ISSCC'18 [3]	ISSCC'16 [1]
Technology	**0.18 μm**	0.35 μm	0.18 μm	0.18 μm HV	0.35 μm
Energy Extraction Technique	**SPFCR**	SSHC	FCR	SE-SSHC	P-SSHI
Piezoelectric Harvester	**MIDE PPA1021**	MIDE V21BL	Piezo Systems Inc. (P5A4E @ 5mm^3)	Custom MEMS	MIDE V21B & V22B
Key Component	**4 Capacitors 21 phase**	8 capacitors 17 phase	4 Capacitors 7 phase	8 Capacitors 17 phase	Inductor
MOPIR	**9.3x**	9.7x	4.83x	8.21x	6.81x
P_{in} adaptation	**Capacitor-reuse Multi-VCR SC DC-DC**	no	no	no	Shared inductor
MPPT	**yes**	no	no	no	yes
Chip Size	**0.2 mm^2**	2.9 mm^2	1.7 mm^2	5.3 mm^2	0.72 mm^2
Output Power	**0.5 to 64 μW**	161.8 μW	50.2 μW	186 μW	160.7 μW
Operating Freq.	**200Hz**	92 Hz	110 kHz	219Hz	225 Hz

Figure 27.3.6: Performance summary and benchmark with prior art.

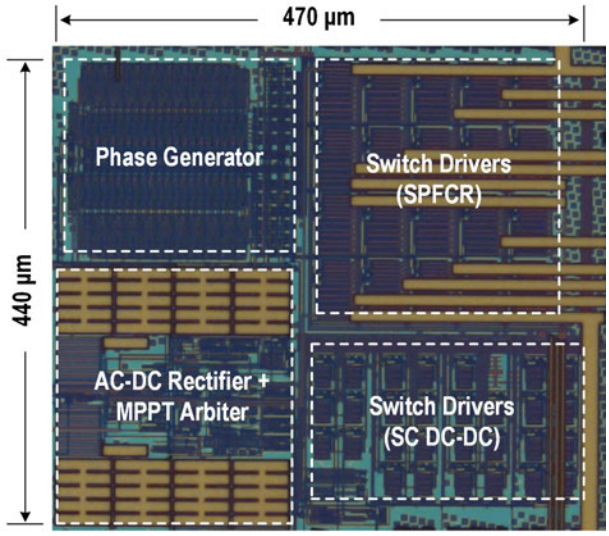


Figure 27.3.7: Die micrograph of the proposed PEH interface in 180nm 1.8/3.3/6V CMOS.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

27.4 Multi-Beam Shared-Inductor Reconfigurable Voltage/SECE-Mode Piezoelectric Energy Harvesting of Multi-Axial Human Motion

Miao Meng[1], Ahmed Ibrahim[1], Tiancheng Xue[2], Hong Goo Yeo[3], Dixiong Wang[1], Shad Roundy[2], Susan Trolier-McKinstry[1], Mehdi Kiani[1]

[1]Pennsylvania State University, University Park, PA
[2]University of Utah, Salt Lake City, UT
[3]Daegu- Gyeongbuk Institute of Science and Technology, Daegu, Korea

The past few years have witnessed a growing demand for self-powered wearables that can enable vigilant health monitoring, with 24/7 operation. Energy harvesting from human-body motion is attractive for wearables; however, conventional unidirectional single-cantilever-beam piezoelectric energy harvesters (PEHs) [1-4] suffer from several body-motion harvesting challenges: such as multi-axial motion, irregular frequencies, and unpredictable amplitudes with frequent low-power levels [5]. To address these challenges, an eccentric rotor-based inertial PEH has been developed, which utilizes multiple magnetically plucked flexible thin-film (60μm) PZT-nickel-PZT beams to significantly increase the harvested energy within a small volume [5]; compared to bulk-PZT beams that are more feasible in direct-force-driven PEHs. The wrist-worn multi-beam PEH, shown in Fig. 27.4.1, converts multi-axial body motion into AC voltages with different phases and decaying amplitudes (up to several volts) within the frequency range of 90-160Hz for each beam.

Current PEHs, utilizing synchronized-switch harvesting on inductor (SSHI) [1], energy investing [2], and synchronous electrical charge extraction (SECE) [3,4], suffer from multiple shortcomings: (1) they can only interface to one beam; (2) SSHI and SECE inherently suffer from harvesting shock responses and large inputs (large power loss in inductor), respectively; and (3) SECE with passive negative-voltage converters (NVCs) suffers from harvesting low PEH voltages, which often occur due to body motion. This paper presents a PEH chip (Fig. 27.4.1) with five unique capabilities: (1) simultaneous energy harvesting from up to 6 piezoelectric beams in a modular fashion using a single shared off-chip inductor; (2) seamless self-reconfiguration between operating as an efficient full-wave active rectifier (voltage mode, VM) or as an SECE to improve overall efficiency, extract the maximum energy, and to protect the chip against large inputs; thus, eliminating the need for off-chip components [3] or the use of a high-voltage process [4]; (3) extension of the input-voltage range (as low as 35mV) by utilizing active NVCs; (4) adaptive SECE operation (peak and zero-crossing time detection) to address the input frequency and beam variation; and (5) an optimal cold start (self-starting) with a tiny active-voltage doubler. Fig. 27.4.1 shows a block diagram of the proposed chip including 6 identical full-wave active rectifiers with active NVCs (VM: one per beam), a shared SECE circuit for 6 beams, a control circuit for optimal switching and reconfiguration, and a voltage doubler connected to an additional 7th beam (only for startup when active NVCs cannot operate). The chip externally requires one inductor (L_{EXT}), one storage capacitor (C_{STORE}) generating V_{STORE}, and one cold-start capacitor (C_{CHIP}), which with the beam's internal capacitance (C_P) forms a voltage doubler (eliminating one external capacitor) to supply the chip's internal circuitry via V_{CHIP}.

As shown in Fig. 27.4.1, the outputs from the beams ($V_{P,n}$) are first full-wave rectified by the active NVCs, generating $V_{REC,n} \approx |V_{P,n}|$, where n is the beam number, 1-6. As long as $V_{REC,n} < V_{STORE}$ beam-n's active rectifier is off ($P_{1,n}$:OFF), and $C_{P,n}$ is charged by the beam's internal current ($I_{P,n}$). Figure 27.4.2 shows key operational waveforms for the proposed reconfigurable VM-SECE scheme. When the peak $V_{REC,n} > V_{STORE}$, the chip seamlessly transitions between VM and SECE operation. Otherwise, the chip operates only based on SECE. For VM-SECE operation, when $V_{REC,n}$ slightly surpasses V_{STORE}, $P_{1,n}$ turns on, charging C_{STORE} with a high efficiency (95.6%), via $I_{P,n}$ (called VM), until $I_{P,n}$ reaches zero. At which time, $V_{REC,n}$ goes slightly below V_{STORE}, turning $P_{1,n}$ off and triggering SECE, which is time multiplexed for the 6 beams. Similar to a conventional SECE [4], the remaining stored energy in $C_{P,n}$ is extracted by first transferring it to L_{EXT} for a short time period by enabling $SW_{1,n}$ (transmission gate $TG_{1,n}$:ON) and SW_2 (N_1:ON) and disabling SW_3 (N_2:OFF) until $V_{REC,n}$ reduces to zero. Then L_{EXT} energy is transferred to C_{STORE} via a diode (P_2) by disabling $SW_{1,n}$, SW_2 ($TG_{1,n}$, N_1:OFF) and enabling SW_3 (N_2:ON). If the peak $V_{REC,n} < V_{STORE}$, the chip adaptively skips VM, and only operates in SECE; by following the same protocol except that SECE is triggered by $V_{REC,n}$ peak instead of $I_{P,n}$ zero-crossing time (rising edge of $V_{VM,n}$). Employing VM not only improves overall efficiency, but it also limits the beams' voltages to $\approx V_{STORE}$, i.e., efficient over-voltage protection. Figure 27.4.2 also shows key waveforms for asynchronous and synchronous inputs. For synchronous inputs (worst case), VM operations occur simultaneously via their own paths, while the shared SECE is time multiplexed between beams with minimal loss since SECE operation requires only 50μs relative to the inputs' lower frequency (90-160Hz).

Figure 27.4.3 shows schematic diagrams for the active NVC and control block: which includes a multi-beam sweep, peak and zero-crossing time detection (PKD&ZCD), and switching control. In the NVC a comparator switches pass transistors to full-wave rectify the input actively with a minimal voltage drop. Within a multi-beam sweep, the beams' voltages ($V_{P,n}$) are checked one-by-one every 60μs in a loop by 6 cascaded beam-control blocks. There are 4 possible conditions: (1) $C_{P,n}$ is charging, $V_{REC,n} < V_{STORE}$, (2) C_{STORE} is charging via VM, $V_{REC,n} > V_{STORE}$, (3) VM operation has ended ($V_{VM,n}$ is high) with some charge left on $C_{P,n}$, and (4) $V_{REC,n}$ reaches its peak, $V_{REC,n} < V_{STORE}$. Under conditions 1 and 2 (PKD_n=0), beam-n is skipped through flip-flop $FF_{2,n}$ followed by a 60μs delay to check the next beam. Under conditions 3 and 4 (PKD_n=1), SECE immediately starts and beam control waits until SECE operation ends, i.e., $SW_{1,n}$ goes low to clock $FF_{1,n}$, triggering the next beam after 60μs. In PKD&ZCD, a positive edge is generated at $V_{REC,n}$ peak and combined with $V_{VM,n}$ to clock a flip-flop for generating high PKD_n (indicating the start of SECE), which is reset to low at $V_{REC,n}$ zero-crossing time (ZCD_n=1). Finally, switching control generates required $SW_{1,1\text{-}6}$, SW_2 and SW_3 signals.

The proposed chip was fabricated in a 0.35μm 4M-2P standard-CMOS process, occupying 1.9mm^2 of active area. The chip was integrated with the inertial harvester (6 beams for charging C_{STORE}, and 1 beam for cold start), as shown in Fig. 27.4.1. In all measurements, L_{EXT} = 2.2mH, C_{STORE} = 47μF, C_{CHIP} = 10μF. Figure 27.4.4 shows the chip's measured transient waveforms, when the inertial harvester is gently shaken. During a cold start, C_{CHIP} was first charged via P_{C2} in Fig. 27.4.1 in a passive manner to ≈1.5V. At which point, the active voltage doubler starts to operate, charging C_{CHIP} efficiently to V_{CHIP} = 3.2V, with a $V_{P,7}$ peak-to-peak voltage as large as 4V. Similarly for V_{CHIP} < 1.5V, C_{STORE} is charged via the 6 $P_{1,n}$ transistors in a passive manner to ≈1.5V. At which point, the active rectifiers in VM start to operate, further charging C_{STORE} to ≈1.9V via VM, for V_{CHIP} < 3V (chip not fully functional yet). When V_{CHIP} > 3V, then the chip operates in the reconfigurable VM-SECE mode (fully-functional) to charge C_{STORE} more efficiently (at a faster rate in Fig. 27.4.4) to V_{STORE} > 1.9V in the presence of different voltage/frequency variations on each beam.

Figure 27.4.5 shows how the chip improves stored power (P_{STORE}) vs. V_{STORE} for weak periodic and shock vibrations to a single beam (open-circuit voltage, $V_{P,OC}$ = 1.6V). The chip (reconfigurable VM-SECE) harvested 2.43× and 3.65× more power than the maximum energy harvested using the 95.6%-efficient full-wave active rectifier (on-chip VM) for periodic and shock vibrations, respectively. Note, that due to the low-frequency operation, offset compensation is not required. This leads to a high shock figure-of-merit (FoM): 365% for a single beam. Utilizing all 6 beams, the FoM further increases to 511% because VM-SECE scheme can extract almost all of the energy from the individual beams, regardless of their voltage levels. Whereas, 6-parallel active rectifiers cannot. The table in Fig. 27.4.6 benchmarks our chip against the prior-art. This chip provides an adaptive reconfigurable VM-SECE scheme that maximizes efficiency and the extracted power for a wide range of inputs. It can interface 6 beams in a modular fashion, with inputs as low as 35mV. Allowing it to harvest energy from multi-axial body motion, which is often weak, using multiple flexible thin-film beams.

References:

[1] D. A. Sanchez, et al., "A 4μW-to-1mW parallel-SSHI rectifier for piezoelectric energy harvesting of periodic and shock excitations with inductor sharing, cold start-up and up to 681% power extraction improvement," *ISSCC*, pp. 366–367, 2016.
[2] D. Kwon and G. Rincon-Mora, "A single-inductor 0.35-μm CMOS energy investing piezoelectric harvester," *ISSCC*, pp. 78–79, 2013.
[3] P. Gasnier, et al., "An autonomous piezoelectric energy harvesting IC based on a synchronous multi-shot technique," *JSSC*, vol. 49, no. 7, pp. 1561–1570, July 2014.
[4] A. Quelen, et al., "A 30nA quiescent 80nW to 14mW power range shock-optimized SECE-based piezoelectric harvesting interface with 420% harvested energy improvement," *ISSCC*, pp. 150-152, 2018.
[5] T. Xue, et al., "Wearable inertial energy harvester with sputtered bimorph lead zirconate titanate (PZT) thin-film beams," *Smart Mat. Struc.*, vol. 27, no. 8, July 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

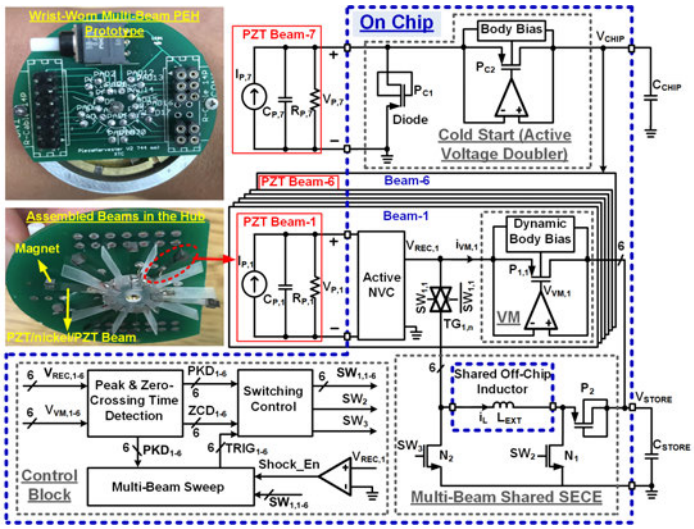


Figure 27.4.1: Wrist-worn multi-beam PEH prototype and block diagram of the proposed chip.

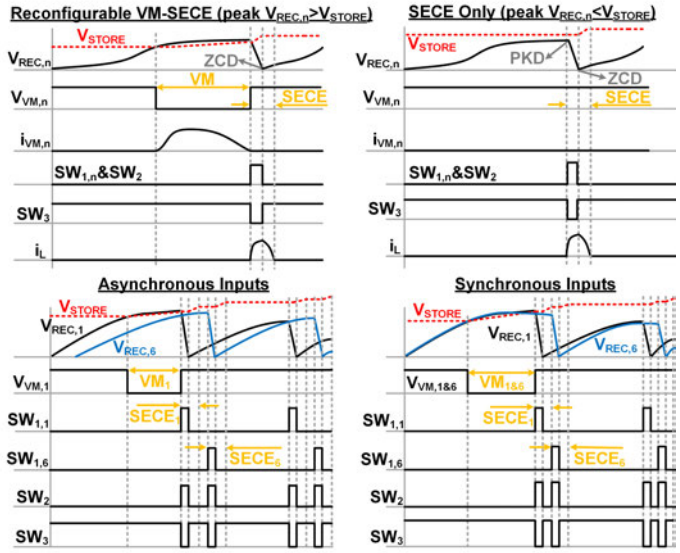


Figure 27.4.2: Key operational waveforms of the proposed chip.

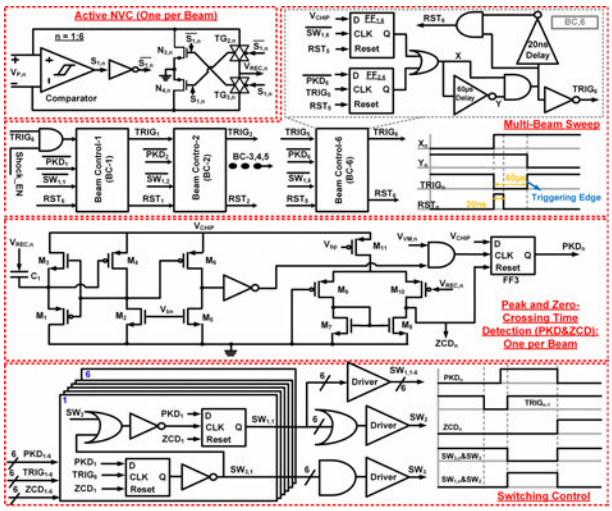


Figure 27.4.3: Schematic diagrams of the active NVC and control block: including a multi-beam sweep, peak and zero-crossing time detection, and switching control.

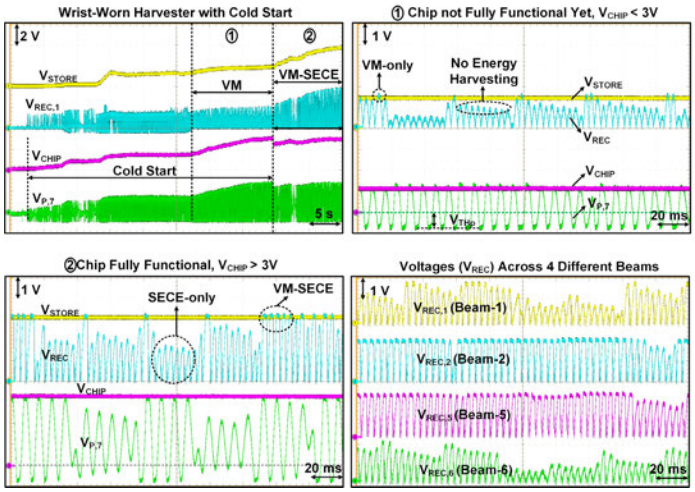


Figure 27.4.4: Measured transient waveforms of the proposed chip interfacing with a wrist-worn multi-beam PEH.

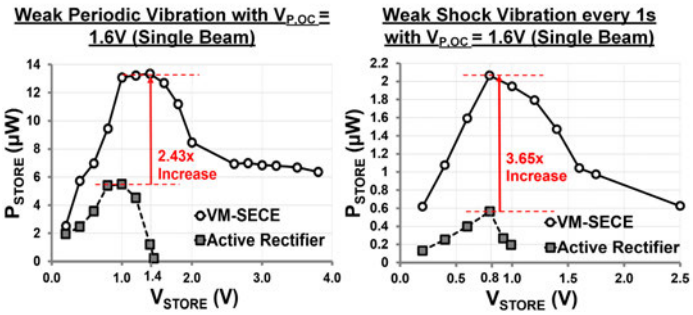


Figure 27.4.5: Comparison of measured harvested power (P_{STORE}) vs. V_{STORE} for a single beam between our VM-SECE chip and a full-wave active rectifier (on-chip VM-only) for weak periodic and shock excitations ($V_{P,OC}$ = 1.6V). For stronger vibrations with $V_{P,OC}$ = 2V, the harvested power increased by 1.6× on average.

Publication	[1]	[2]	[3]	[4]	This work
CMOS Tech (nm)	350	350	350	40 (10V)	350 (5V)
Chip Size (mm²)	0.72	2.34	3.6	0.55	1.9
Scheme Type	SSHI	Energy Investing	Multi-Shot SECE	SECE	Reconfigurable VM-SECE
# of Beams (Inputs)	1	1	1	1	6 (Modular)
Piezoelectric Energy Harvester	MIDE V21B & V22B	MIDE V22B	Murata	MIDE PPA1011	Custom PZT/nickel/ PZT (Wrist-Worn)
C_p (nF)	26	15	23	43	17-49
Beam Dimension, l×w×t (mm)	6.25 × 1.5 × 0.7	6.25 × 1.5 × 0.7	-	71 × 25.4 × 0.7	12 × 3 × 0.06 (each beam, flexible)
Excitation Type	Periodic & Shock	Periodic	Periodic	Periodic & Shock	Multi-Axial Body Motion
Operation Freq. (Hz)	225	143	100	75.4	90-160
FoM* (%) @ $V_{p,oc}$ (V)	269 @ 2.5	-	-	330 @ 3.4	1-Beam: 365 @ 1.6 6-Beam: 511 @ 1.6
Cold Startup	Yes	No	Yes	Yes	Yes
Maximum End-to-end Efficiency (%)	≈88**	69.2	61	94	VM: 95.6 VM-SECE: 84.6
Minimum Input Voltage (mV)	670	570	-	-	35
Quiescent Current (μA)	1	0.1	0.3	0.03	0.5 (6 beams)

*FoM = max(P_{out})/max($P_{out,FAR}$) for shock input (FAR: Full-wave active rectifier). **Calculated.

Figure 27.4.6: Comparison table of this work to prior-art.

1	Cold Start
2	Beta Multiplier (Current Reference)
3	Active Negative Voltage Converter (NVC)
4	Power-Path Switches
5	Voltage Mode (VM)
6	Peak and Zero-Crossing Time Detection (PKD&ZCD)
7	Switching Control
8	Multi-Beam Sweep
9	Shock Detection

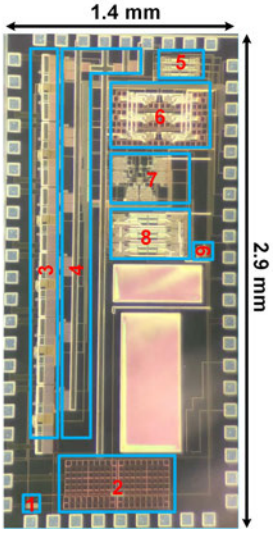


Figure 27.4.7: Proposed chip micrograph, occupying 1.9mm² of active area, and its key building blocks.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

27.5 A 91%-Efficiency Envelope-Tracking Modulator Using Hysteresis-Controlled Three-Level Switching Regulator and Slew-Rate-Enhanced Linear Amplifier for LTE-80MHz Applications

Parisa Mahmoudidaryan, Debashis Mandal, Bertan Bakkaloglu, Sayfe Kiaei

Arizona State University, Tempe, AZ

Envelope tracking (ET) is widely used to improve the efficiency of linear power amplifiers (PAs) in applications such as LTE, LTE-Advanced inter- or intra-band carrier aggregation (CA), and for high-speed uplink packet access (HSUPA) with a high peak-to-average power ratio (PAPR). The hybrid ET modulator (ETM), where an efficient switching regulator (SWR) operates in parallel with a fast class-AB linear amplifier (LA), is one method to address the bandwidth (BW), power-efficiency and output ripple requirements.

AC-coupled LAs [1-3], with an adjusted LA supply voltage [1-2], improve the system efficiency, but they require off-chip AC-coupling capacitors, and additional inductors for extra buck and boost converters. Multilevel SWR [3] reduces the inductor size to achieve a faster response, but the BW is still limited to 20MHz. In these hybrid ETMs, the LA compensates for the SWR's slow response by providing high-frequency current, resulting in low efficiency modulators [1-3, 5-6]. To improve the efficiency, [4] replaces the LA with a multilevel SWR, and the dual SWR in [5] decreases LA current; however, the BW is limited to 10MHz and 20MHz in [4] and [5], respectively. Thus, design of an efficient ETM for wideband application is still challenging [1-6].

This paper presents an 80MHz BW, 91%-efficient, hybrid ETM with a hysteresis-controlled 3-level SWR in parallel with a class-AB, slew-rate enhanced LA, as shown in Fig. 27.5.1. Employing the proposed 3-level hysteresis-controlled SWR, instead of a conventional PWM loop, resolves instability concerns which exist for the PWM loop and removes the required compensation circuit that carries with it stringent considerations. In addition, the proposed technique eliminates the flying capacitor (C_F) calibration loop by using concurrent C_F voltage sensing and regulation scheme; and results in a high-speed control loop due to the higher loop bandwidth of the hysteresis control, as high as the switching frequency. Thus, the wideband SWR reduces the LA burden by providing high-frequency current and the high-frequency current percentage provided by the LA reduces, and the ETM achieves over 91% efficiency. The LA uses slew-rate enhancement (SRE) that improves the slew-rate above 325V/µs and BW up to 275MHz, thus the ETM can operate up to LTE-80MHz.

The proposed hysteresis controller (HC) for the 3-level SWR uses concurrent regulation of the C_F voltage (V_{CF}) to $V_{DD}/2$. Thereby, eliminating the required additional calibration loop for V_{CF}. A high-speed current-mode voltage-sensing (CMVS) circuit is used to attain a fast loop for sensing V_{CF}, and for its regulation. In a 3-level SWR, the inductor current ripple reduces by 50% relative to the 2-level SWR. This improves the overall ET noise performance and reduces the switching losses by more than 50% [4] due to effectively halved switching rate of the low-parasitic thin-gate-oxide power transistors and drivers.

As shown in Fig. 27.5.2, the inputs of HC are V_{CF}, the input envelope voltage (V_{ENV}) and V_{SENS}. A sense resistor (R_{SENS} in Fig. 27.5.1) that is connected between V_{OUT} and the sensing stage, with a current that is scaled down to $^1/_N$th of the class-AB output stage, generates V_{SENS}. Thus, the R_{SENS} loss is removed from the main path for I_{LIN} [6]. The SWR operation is divided into two ranges based on the V_{ENV} level (Fig. 27.5.2). (1) For $0 < V_{ENV} \leq V_{DD}/2$, the output of comparator A1 selects the ctrl-a path. When the sinking I_{LIN}, and hence V_{SENS}, passes the hysteresis window $+V_{HYST}$ of comparator A2, the output of A2 is high which sets signals N1, P1 and N2, P2 to high, turning on M_{N1} and M_{N2}, and turning M_{P1} and M_{P2} off; thus, decreasing I_{SW} and I_{LIN}. When the sourcing I_{LIN} makes V_{SENS} cross $-V_{HYST}$, A2 output is low and V_C becomes the dominant control signal. For $V_{CF} < V_{DD}/2$, N1 and P1 is low, and N2 and P2 is high, turning on M_{N2} and M_{P1} to charge C_F. Conversely, for $V_{CF} > V_{DD}/2$, M_{N1} and M_{P2} are turned on to discharge C_F. Hence, to increase I_{SW}, the switching node voltage (V_{SW}) is set to $V_{DD}/2$. Thus, V_{SW} switches between 0V and $V_{DD}/2$. (2) For $V_{DD}/2 < V_{ENV} \leq V_{DD}$, the low output of A1 selects the ctrl-b path at the MUX outputs, and V_{SW} switches between $V_{DD}/2$ and V_{DD}, to minimize the inductor current ripple. The HC operates similarly to the former case and regulates V_{CF} to $V_{DD}/2$ per cycle, independent of parasitics and process-voltage-temperature (PVT) variations.

Given the voltage variance (ΔV) set by the thin-gate transistors' tolerance voltage, the objective is to ensure that V_{CF} is in the desired $V_{DD}/2 \pm \Delta V$ range. The CMVS circuit shown in Fig. 27.5.2, converts V_{CF} to I_{CF} using a diode-connected transistor and resistor (R). Similarly, the two reference currents $I_{REF} \pm \Delta I$ are proportional to $V_{DD}/2 \pm \Delta V$ and are generated from $V_{DD}/2$ and mirrored into a comparison stage with I_{CF}. In current mode, I_{CF} is compared to $I_{REF} \pm \Delta I$ and triggers proper set or reset signal for an RS latch to produce the control signal V_C for HC to regulate V_{CF}.

The LA uses SRE, maximum current selection (MCS), a rail-to-rail input stage and a push-pull class-AB output stage as shown in Fig. 27.5.3. It operates in unity-gain mode, which removes the extra pole in the feedback-path of a non-unity-gain configuration [1-3, 5-6], resulting in a faster loop response. The MCS circuit shown in Fig. 27.5.3 compares the NMOS and PMOS input-pair currents; and selects the larger of the two and transfers it to the output, independent of the input common mode range (CMR). This stabilizes the gain-BW of the LA and the class-AB bias current; and ensures push-pull operation over the CMR.

The two complementary transistors M_{PAB} and M_{NAB}, form a floating voltage source to generate the class-AB biasing for the output pair M_{PO} and M_{NO}, and the current sensing pair M_{PS} and M_{NS}. The SRE circuit mirrors the input stage branch currents and subtracts them to generate the auxiliary slewing currents (I_{SRD} or I_{SRC}). These speed up the discharging and charging of the V_X and V_Y nodes. This circuit also adjusts the floating voltage sources' V_{PAB} and V_{NAB} to maintain a minimum output-stage quiescent current (I_Q) during the SRE. When $V_{ENV} - V_{OUT}$ is positive, I_{SRD} increases to pull down V_X to supply current to V_{OUT} via M_{PO}; and increases V_{NAB} to maintain I_Q at M_{NO}. A similar operation is expected for the negative slewing of $V_{ENV} - V_{OUT}$ by increasing I_{SRC} and decreasing V_{PAB}. Using this method, the slew rate improves from 178V/µs to 325V/µs and from -169V/µs to -307V/µs. Using SRE method also improves the BW of the LA from 208 to 220MHz. I_Q adjustment further improves the BW to 275MHz. The DC current consumption of the SRE circuit is 200uA.

The proposed ETM is fabricated in 65nm CMOS, occupies 1.7×1.6mm^2 (including a 12nF on-chip C_F), and operates from a 2.4V supply. Figure 27.5.4 shows the measured waveforms for an LTE-80MHz uplink signal with a 7.84dB PAPR, using a 4.7Ω resistive load. The 3-level V_{SW} node switches between 0, 1.2 and 2.4V, with an effective frequency up to 80MHz. V_{CF} is regulated to 1.2V within ± 20mV of accuracy. When the SRE circuit is off, the modulator operates for LTE-40MHz with a peak efficiency of 93% and less than 2.3% input-output tracking error. When the SRE circuit is turned on, the ETM BW is over 80MHz and meets the LTE-80MHz standard, with a measured peak efficiency of 91%, and it tracks the signal envelope within less than 3.2% error. Measured ETM output noise at a 30MHz offset is -135dBm/Hz. Figure 27.5.5 shows the measured power-added-efficiency (PAE) versus output power (P_{OUT}) of an external RF-PA, with and without supply modulation, while an 8.1% PAE improvement is achieved at -4dB back-off using ETM. The output spectrum for LTE-80MHz with a 24dBm P_{OUT} is also shown with an ACLR of -32.5dBc. The performance comparison with prior-art designs is given in Fig. 27.5.6. The proposed ETM provides the highest tracking BW (80-MHz, similar to [1]); and has the highest peak efficiency of 93%. This method uses a 538nH off-chip inductor without an AC-coupling capacitor. The die micrograph is shown in Fig. 27.5.7.

Acknowledgment:
This project was funded and supported by the NSF's Connection One Center and Qualcomm Inc. We would like to thank Dr. Shahin Mehdizad Taleie for his technical support.

References:

[1] C-Y Ho, et al., "An 87.1% Efficiency RF-PA Envelope-Tracking Modulator for 80MHz LTE-Advanced Transmitter and 31dBm PA Output Power for HPUE in 0.153µm CMOS," *ISSCC*, pp. 432-433, 2018.
[2] J.-S. Paek, et al., "An RF-PA Supply Modulator Achieving 83% Efficiency and -136dBm/Hz Noise for LTE-40MHz and GSM 35dBm Applications," *ISSCC*, pp. 354-355, 2016.
[3] X. Liu, et al., "A 2.4V 23.9dBm 35.7%-PAE -32.1dBc-ACLR LTE-20MHz envelope-shaping-and-tracking system with a multiloop-controlled AC-coupling supply modulator and a mode-switching PA," *ISSCC*, pp. 38-39, 2017.
[4] P. Amò, et al., "Envelope modulator for multimode transmitters with AC-coupled multilevel regulators," *ISSCC*, pp. 296-297, 2014.
[5] M. Hassan, et al., "A CMOS dual-switching power-supply modulator with 8% efficiency improvement for 20MHz LTE Envelope Tracking RF power amplifiers," *ISSCC*, pp. 366-367, 2013.
[6] Y. Jing, et al., "A High Slew-Rate Adaptive Biasing Hybrid Envelope Tracking Supply Modulator for LTE Applications," *IEEE TMTT*, vol. 65, no. 9, pp. 3245-3256, Sept. 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

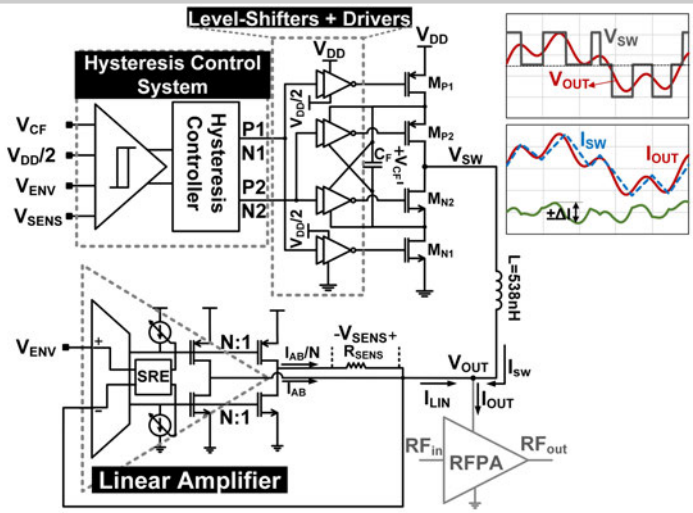


Figure 27.5.1: Block diagram of the proposed hybrid ET modulator with hysteresis-controlled 3-level switching regulator and a slew-rate enhanced linear amplifier.

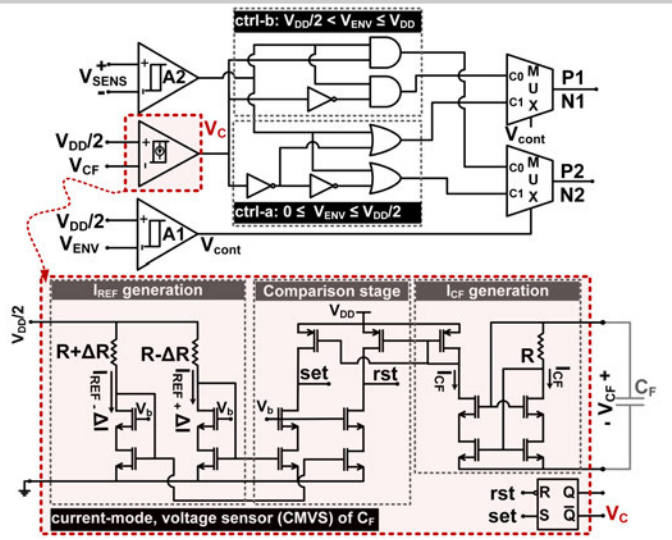


Figure 27.5.2: Proposed hysteresis control system at the top with a high-speed current-mode, voltage sensor for flying capacitor (V_{CF}) at the bottom.

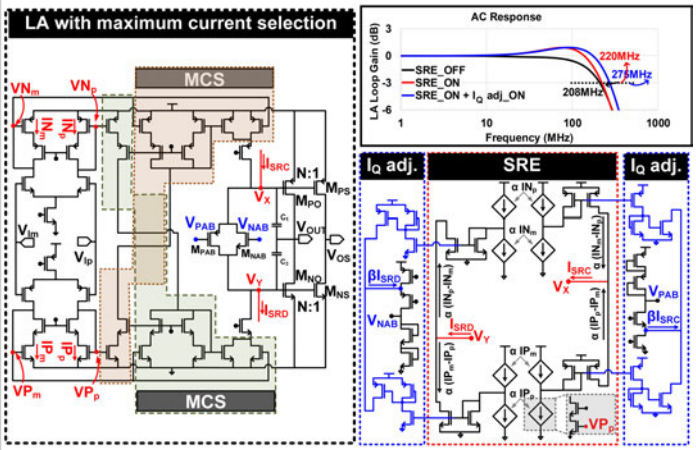


Figure 27.5.3: Simplified LA circuit with SRE and I_Q adjustment during SRE, and AC response of the LA showing -3dB BW improving from 208MHz to 220MHz using SRE and to 275MHz using both SRE and I_Q adjustment.

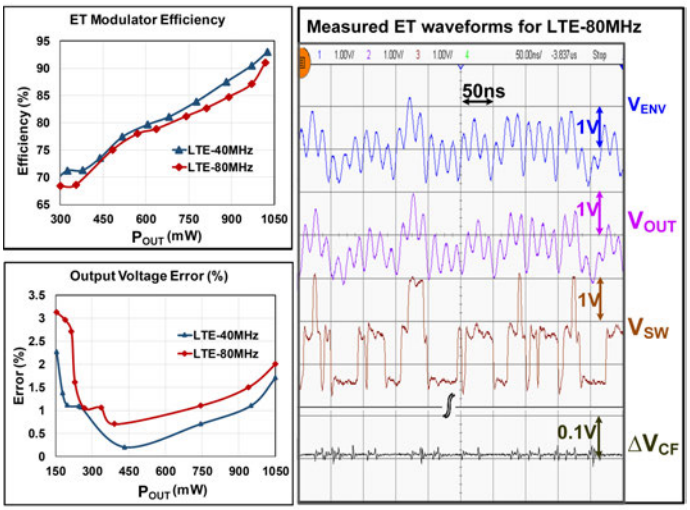


Figure 27.5.4: Measured ETM efficiency and output-input error vs. P_{OUT} for 40MHz/80MHz, and ETM waveforms for LTE-80MHz, 3-level operation with ±20mV V_{CF} variation.

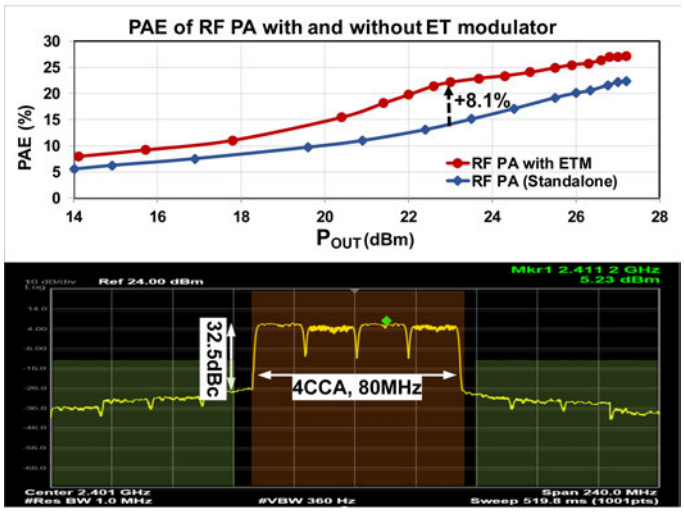


Figure 27.5.5: Measured PAE of PA with and without ET supply modulation, and output spectrum of supply modulated PA for intra-band CA LTE-80MHz signal at 24dBm P_{OUT}.

Parameters	This Work	ISSCC 2018 Media Tek [1]	ISSCC 2016 Samsung [2]	ISSCC 2017 HKUST [3]	ISSCC 2014 STM, Czech [4]	ISSCC 2013 UCSD [5]
CMOS Technology	**65nm**	153nm	130nm	65nm	130nm	180nm
Supply Voltage (V)	**2.4**	3.8	4	2.4	3.8	5.5
BW (MHz)	**80**	80	40	20	10	20
ETM Peak Efficiency (%)	**91 @ 80MHz 93 @ 40MHZ**	81.2 @ 80MHz 85.5 @ 40MHz	83	88.7	86.2	83
RF P_{OUT} (dBm)	**24 @ 80MHz 26 @ 40MHz**	26 @ 80MHz 26 @ 40MHz	26.5	23.9	26.3	28.3
ACLR (dBc)	**-32.5 @ 80MHz -34 @ 40MHz**	-38.1 @ 80MHz -41.5 @40 MHz	-41.2	-32	-40	-41.4
PAPR (dB)	**7.84 @ 80MHz 8.06 @ 40MHz**	8.3 @ 80MHz 7.54 @ 40MHz	NA	NA	6.7	6.7
Off-chip Ind / AC-Cap	**538nH / none**	1μH+0.68μH / 4.7μF	2ind / 1cap	206nH / 1μF	4.7μH+22nH / 0.47μF	1μH+0.3μH / none

Figure 27.5.6: Performance summary and comparison to prior-art.

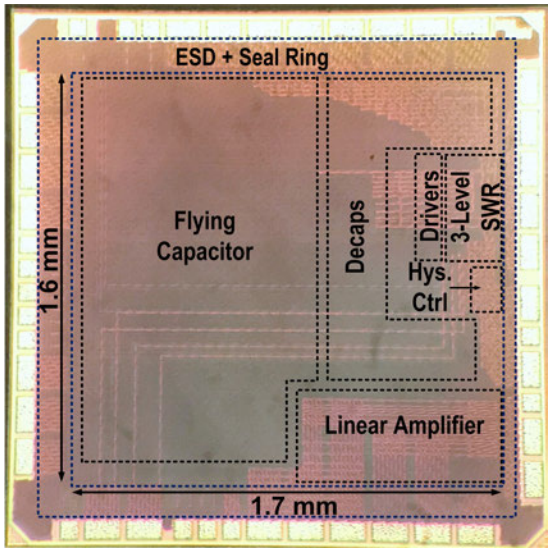


Figure 27.5.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

27.6 Background Capacitor-Current-Sensor Calibration of DC-DC Buck Converter with DVS for Accurately Accelerating Load-Transient Response

Tai-Haur Kuo, Yi-Wei Huang, Pai-Yi Wang

National Cheng Kung University, Tainan, Taiwan

Switching buck converters with dynamic voltage scaling (DVS) for high-efficiency high-performance computing applications need to reduce the output-voltage undershoot/overshoot (V_{US}/V_{OS}) and settling time t_S under a large and fast-changing load current (I_{load}). A multiphase topology with a fast load-transient response meets these requirements. The load-transient response can be accurately accelerated to reduce V_{US}/V_{OS} and t_S to near their ideal values by measuring the output-capacitor current I_{Co} to control the inductor's energizing and de-energizing times, since I_{Co} instantly reflects the load-current transients. An integrated capacitor-current sensor (CCS) [1] can be used to sense I_{Co} by emulating the output-capacitor impedance Z_{Co}: comprising capacitance C_O, the equivalent series resistance R_{ESR}, and inductance L_{ESL}. However, I_{Co} will be inaccurately sensed if Z_{Co} varies with different output voltages V_O, manufacturing variations, PCB parasitics, temperature, and aging. The state-of-the-art CCS calibration technique [1] for such Z_{Co} variations is suitable for foreground operation and DVS with pre-characterized V_O levels, since calibration starts immediately after being enabled and runs continuously until it ends. The CCS in [1] is calibrated with a low-power cost-effective comparator and successive approximation logic, with an acceptable calibration time T_{CAL} for foreground operation. To broaden the range of applications, this work proposes an ADC-based CCS and a background CCS calibration (BCC) controller. The proposed CCS uses a flash ADC with a dynamic reference to shorten T_{CAL}. The BCC controller automatically finds a quasi-steady state (QS), namely a short period of steady-state behavior when there is no load transient or DVS event, to trigger CCS calibration, and can interrupt CCS calibration when a load transient or a DVS event occurs. Since QSs generally exist, the BCC with a short T_{CAL} can increase the flexibility of scheduling both load transients and DVS events. Thus, it is suitable for DVS with numerous V_O levels that account for in situ parameter variations.

Figure 27.6.1 shows a 4-phase (ϕ_{1-4}) buck converter with the proposed ADC-based CCS and BCC controller. CCS calibration aims to adjust C_S, R_S, and L_S to precisely emulate $Z_{Cs} = k{\cdot}Z_{Co}$, where k is a scaling factor. Thus, the CCS current $I_{Cs} = I_{Co}/k$ is correctly sensed. An amplifier then emulates the impedance of L_S and converts I_{Cs} to a voltage (V_{ICs}). The ADC-based CCS uses the ripple information from the ϕ_1-inductor current (I_{L1}) and I_{Cs} (measured via Δv_{IL1} and Δv_{ICs}), to generate the calibration result for Z_{Cs}. The BCC controller can be event- and periodically-enabled to calibrate Z_{Cs} against short-term and long-term Z_{Co} variations. Once enabled, the QS detector finds a QS to trigger CCS calibration; then, the state controller generates ϕ_L, ϕ_C, ϕ_R, and ϕ_{S1-4} for the timing control of CCS calibration. When a load transient or a DVS event is detected by the transient detector, an interrupt handler pauses calibration, and can choose to continue or restart it at the next QS. Also, the CCS can read the previous calibration result from memory for a DVS event before a new calibration is restarted. The OCB uses V_{ICs} to precisely control V_{PWM1-4} to recover V_O in one switching period T_{SW} for large load transients, thus approaching the ideal load-transient response.

Figure 27.6.2 top-right shows the converter's modes of operation. In proportional-integral-derivative (PID) mode, V_{PWM1-4}, shown in Fig. 27.6.1, are controlled by the PID compensation, which is formed by a type-II compensator output V_C, the CCS output V_{ICs}, and the interleaved ramp V_{SAW1-4}. Upon load transient detection, the converter enters OCB mode, which enables ϕ_{1-4} and uses V_{ICs} to precisely control V_{PWM1-4}. The converter enters calibration (CAL) mode and remains in closed-loop PID control until a QS is detected. When the QS detector is enabled, four consecutive I_{L1} variations smaller than a programmable range are regarded as a QS. This drives TRG high to trigger CCS calibration. Then, V_{PWM1}'s duty ratio is set to V_O/V_I with only ϕ_1 enabled. Figure 27.6.2 top-left shows the concept of CCS calibration, which consists of 3 states: L_S, C_S, and R_S calibration. To calibrate L_S, the switching frequency f_{SW} changes to f_{SWL} to ensure that $Z_{Co} \approx s{\cdot}L_{ESL}$ and $Z_{Cs} \approx s{\cdot}L_S$. The goal is to tune L_S[3:0] until $\Delta i_{L1} = k{\cdot}\Delta i_{Cs}$ to ensure that $Z_{Cs}(f_{SWL}) = k{\cdot}Z_{Co}(f_{SWL})$ [1]. Figure 27.6.2 bottom-left shows the operating principle of ADC-based CCS calibration. For ease of circuit realization, Δi_{L1} and $k{\cdot}\Delta i_{Cs}$ are converted to Δv_{IL1} and Δv_{ICs}. The L_S calibration state has 4 phases (ϕ_{S1-4}). (1) f_{SW} is changed to f_{SWL}, and Δv_{IL1}'s amplitude is sampled as $V_{IL1,AP}$ near the end of ϕ_{S1}. (2) L_S[3:0] is set to 0 and the resulting Δv_{ICs}'s amplitude is sampled as V_{REFp}. (3) L_S[3:0] is set to 15 and Δv_{ICs} sampled as V_{REFn}. (4) The ADC converts $V_{IL1,AP}$ via a dynamic-reference ladder, which is derived from $V_{REFp} - V_{REFn}$, to find the calibration result code L_S[3:0]. C_S and R_S are calibrated using a similar 4-phase approach. Figure 27.6.2 center-column shows the timing diagrams for BCC. Upon a load transient, CAL is interrupted, OCB takes over, and then CAL continues from the unfinished phase, when the next QS is detected. When a DVS occurs, CAL is interrupted, PID takes over, and then CAL restarts from L_S calibration upon the next detected QS, since Z_{Co} varies with V_O. Furthermore, if the converter encounters a DVS event before a new calibration is restarted, Z_{Cs} can be determined by directly reading the previous calibration results from memory to ensure a fast-transient response. Figure 27.6.2 bottom-right illustrates OCB, which uses V_{ICs} to estimate T_1 and calculate T_{OPT} to switch V_{PWM1-4} at the optimal timing point where $Q_1 = Q_2$. T_{OPT} is equal to $T_1{\cdot}(V_O/V_I)^{1/2}$ and $T_1{\cdot}[1-(V_O/V_I)]^{1/2}$ for step-up and step-down load-transient responses, respectively [1].

Figure 27.6.3 shows the circuit implementations of ADC-based CCS, BCC controller, and OCB. For the ADC-based CCS, Z_{Cs} comprises C_{S0}, R_{S0}, L_{S0}, and their trimming arrays. The dynamic reference samples the respective Δv_{ICs} at ϕ_{S2}=1 and ϕ_{S3}=1 to generate V_{REFp} and V_{REFn} for the dynamic reference ladder of the 4b-flash ADC. Memory, which stores calibration results, can be realized via a standard cell library. For the BCC controller, the periodic- & event-enable block uses a timer and digital logic to enable QS detection with EN = 1. The QS detector uses V_{QSD} to indicate if V_{IL1} variations are less than the programmable window V_{WQ}, and the situation of $V_{QSD} = 0$ for four consecutive T_{SW} is regarded as a QS. ϕ_L, ϕ_C, and ϕ_R are generated by the state controller to calibrate L_S, C_S, and R_S, while ϕ_{S1-4} means the execution of steps 1–4 in CAL. The transient detector regards a load transient (DVS event) if V_{ICs} ($\beta{\cdot}V_{REF}$) step is greater than the programmable window V_{WI} (V_{WV}). The interrupt handler pauses CAL when ϕ_{LT} or $\phi_{DVS} = 1$. For $\phi_{LT} = 1 \rightarrow 0$, CAL continues since ϕ_L, ϕ_C, and ϕ_R all remain unchanged. If $\phi_{DVS} = 1 \rightarrow 0$, then the interrupt handler resets ϕ_L, ϕ_C, and ϕ_R to 0 to restart CAL. The circuit implementation of OCB is similar to that in [1].

The chip is fabricated in a 0.18μm CMOS process using 3.3V devices. Measurements are performed using a nominal $V_I = 3.3$V, $V_O = 1.8$V, $L_{1-4} = 100$nH, and $C_O = 840$nF. Figure 27.6.4 top shows that CAL is interrupted by DVS and then restarts calibrating L_S. With the programmable f_{SWL} = 15MHz and f_{SWC} = 2.5MHz, f_{SWR} and T_{CAL} are measured to be 3.6MHz and 18μs, respectively. Figure 27.6.4 bottom shows CAL interrupted by a load transient and then continuing to calibrate C_S. Figure 27.6.5 top shows that when CAL encounters a load transient, the load-transient responses are faster with BCC. Figure 27.6.5 bottom-left shows that when the converter encounters a DVS event followed by a load transient, the load-transient responses remain fast. Figure 27.6.5 bottom-right shows that when a step-up (step-down) $\Delta I_{load}/\Delta t$ = 2A/5ns occurs, the measured results are V_{US}=34mV (V_{OS}=28mV) and t_S=56ns (t_S=45ns), which are near the ideal values V_{US}=33.7mV (V_{OS}=27.1mV) and t_S=54.4ns (t_S=42.6ns) for a standard buck converter. Equations for the ideal values are from [1] and listed in Fig. 27.6.6.

Figure 27.6.6 shows that this work has the smallest/shortest V_{US}, V_{OS}, and t_S compared to prior state-of-the-art [1-3], and outperforms the hybrid buck converter in [3]. The deviation-from ideal (DFI) rate, defined as |real–ideal|/ideal·100%, shows how close the real value is to the ideal one. The DFI rates for V_{US}, V_{OS}, and t_S are close to 0% and up to 23% lower than those of [1, 2]. T_{CAL} is 6× faster than in [1]. The background calibration can accelerate the response to large load-transients at any time. Figure 27.6.7 shows the die micrograph.

Acknowledgements:
The authors would like to acknowledge the chip fabrication support provided by the National Chip Implementation Center (CIC), Taiwan.

References:
[1] Y.-W. Huang, et al., "A Four-Phase Buck Converter With Capacitor-Current-Sensor Calibration for Load-Transient-Response Optimization That Reduces Undershoot/Overshoot and Shortens Settling Time to Near Their Theoretical Limits," *IEEE JSSC*, vol. 53, no. 2, pp. 552-568, Feb. 2018.
[2] B. Lee, et al., "A 25MHz 4-Phase SAW Hysteretic DC-DC Converter with 1-Cycle APC Achieving 190ns t_{settle} to 4A Load Transient and Above 80% Efficiency in 96.7% of the Power Range," *ISSCC*, pp. 190-191, 2017.
[3] L. Chen and W.-H. Ki, "A 30MHz Hybrid Buck Converter with 36mV Droop and 125ns 1% Settling Time for a 1.25A/2ns Load Transient," *ISSCC*, pp. 188-189, 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

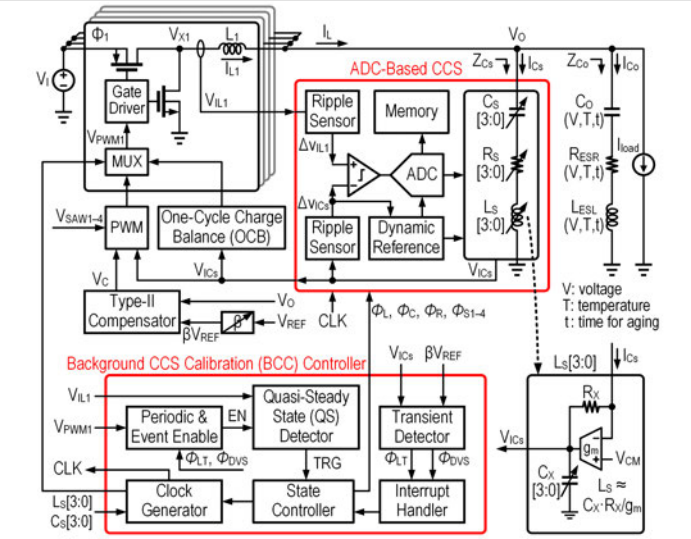


Figure 27.6.1: Block diagram of a buck converter with the proposed ADC-based CCS and BCC controller.

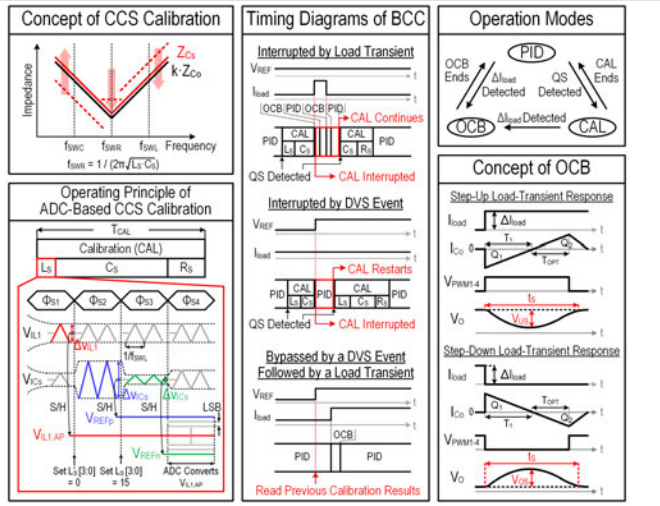


Figure 27.6.2: Concept of CCS calibration; operating principle of ADC-based CCS calibration; timing diagrams of BCC; operation modes of the proposed converter; and, concept of OCB.

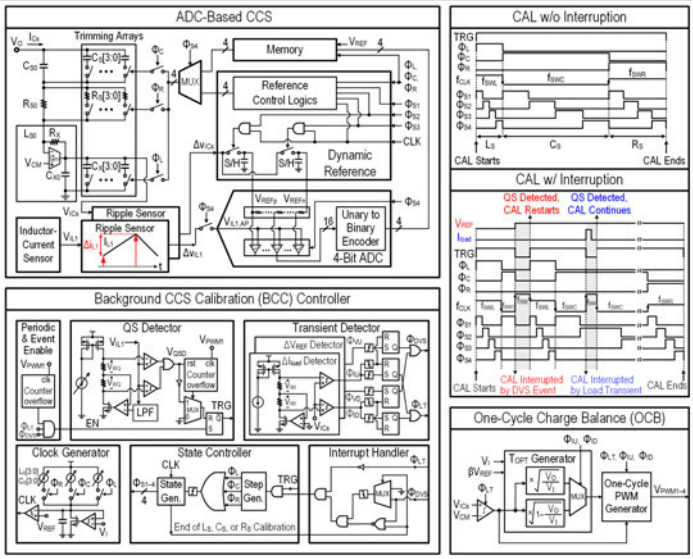


Figure 27.6.3: Circuit implementations and key timing diagrams.

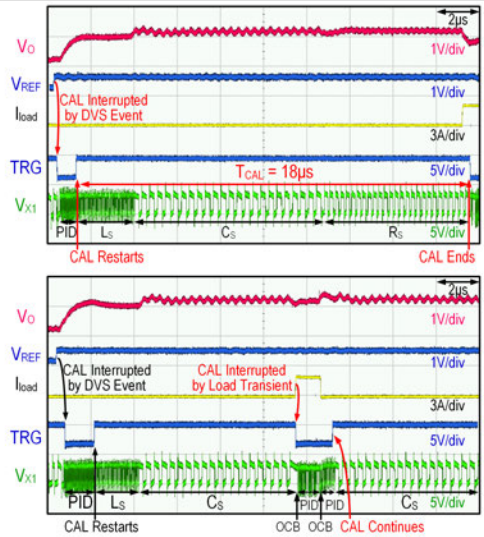


Figure 27.6.4: Measured results: CCS calibration interrupted by DVS events and load transients.

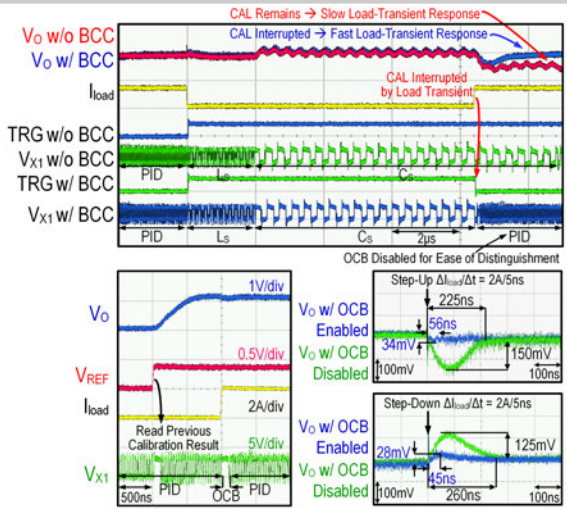


Figure 27.6.5: Measured results: CCS calibration interrupted by a load transient; operation with a DVS event followed by a load transient; and, load-transient responses.

		This Work	JSSC 2018 [1]	ISSCC 2017 [2]	ISSCC 2017 [3]
Topology		Buck	Buck	Buck	Hybrid Buck
# of Phases		4	4	4	1
Process (μm)		0.18	0.18	0.35	0.13
V_I Range (V)		3.3	3.3	3.3	3.3
V_O Range (V)		1.0–2.5	0.7–3.0	0.3–2.5	1.2–2.4
f_{SW} (MHz)		30	30	25	30
L_O (nH)		100	220	200	90
C_O (nF)		840	620	2470	790
Peak Efficiency		91%	88.1%	88.1%	90.7%
Chip Area (mm²)		**1.8	1.93	1.88	1.2
Load-Transient Response	V_I, V_O (V)	3.3, 1.8	3.3, 1.8	3.3, 1.6	3.3, 1.8
	ΔI_{load} / Δt (A/ns)	2 / 5	1.8 / 5	4 / 5	1.25 / 2
	Step-Up V_{US} (mV) [*DFI Rate]	34 [0.8%]	100 [12.9%]	103 [12.9%]	36 [****]
	Step-Up t_S (ns) [*DFI Rate]	56 [2.8%]	133 [5.7%]	190 [***]	125 [****]
	Step-Down V_{OS} (mV) [*DFI Rate]	28 [3.3%]	75 [3.3%]	123 [26.6%]	38 [****]
	Step-Down t_S (ns) [*DFI Rate]	45 [5.5%]	110 [5.6%]	237 [***]	125 [****]

* Deviation-from-ideal (DFI) rate = (|Ideal – Real| / Ideal)×100%; the ideal V_{US}, V_{OS}, and t_S are listed below.
** Excluding bonding pads.
*** Cannot be calculated since V_O did not settle within ±1% of its steady-state value before the load transient.
**** Cannot be calculated since the DFI rates are only available for standard buck converters.

Step-up load-transient response (L_{eff}: effective output inductance. In this work, $L_{eff}=(L_1+L_2+L_3+L_4)/4$.)

$$\text{Ideal } V_{US} = \frac{\Delta I_{load}}{2C_O}\left(\Delta I_{load}\cdot\frac{L_{eff}}{V_I - V_O} - \Delta t\right),\quad \text{Ideal } t_S = \Delta I_{load}\cdot\frac{L_{eff}}{V_I - V_O}\left[1+\sqrt{\frac{V_I}{V_O}\left(1-\frac{V_I - V_O}{L_{eff}}\cdot\frac{\Delta t}{\Delta I_{load}}\right)}\right] - \sqrt{2L_{eff}C_O}\cdot 1\%$$

Step-down load-transient response

$$\text{Ideal } V_{OS} = \frac{\Delta I_{load}}{2C_O}\left(\Delta I_{load}\cdot\frac{L_{eff}}{V_O} - \Delta t\right),\quad \text{Ideal } t_S = \Delta I_{load}\cdot\frac{L_{eff}}{V_O}\left[1+\sqrt{\frac{V_I}{V_I - V_O}\left(1-\frac{V_O}{L_{eff}}\cdot\frac{\Delta t}{\Delta I_{load}}\right)}\right] - \sqrt{2\frac{V_O}{V_I - V_O}L_{eff}C_O}\cdot 1\%$$

Figure 27.6.6: Measured performance comparison with other state of the arts.

27

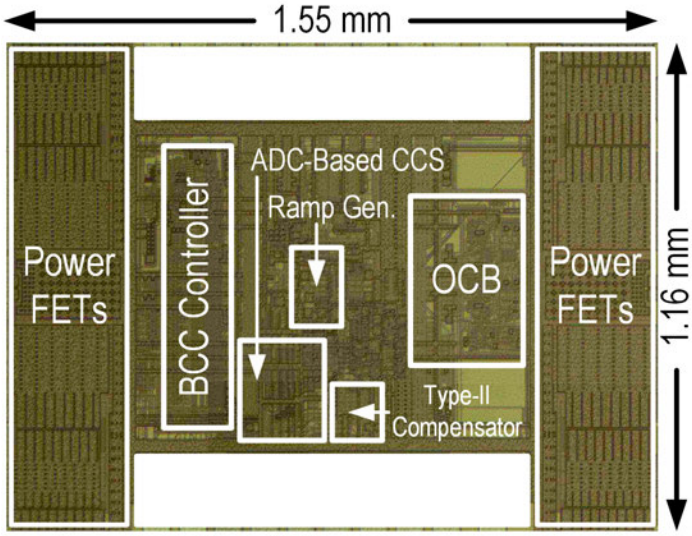


Figure 27.6.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

27.7 A Synthesizable Digital AOT 4-Phase Buck Voltage Regulator for Digital Systems with 0.0054mm² Controller and 80ns Recovery Time

Minho Choi[1,2], Chan-Ho Kye[1], Jonghyun Oh[1], Min-Seong Choo[1], Deog-Kyoon Jeong[1]

[1]Seoul National University, Seoul, Korea
[2]Samsung Electronics, Hwaseong, Korea

Achieving a fast-transient response is a major challenge when designing a switching regulator for a processor. Furthermore, high-frequency operation with small passive devices and robust control with a small area footprint in a CMOS-logic process are essential for integrated digital-system voltage regulators. A 4-phase hysteretic converter [1] has a superior transient response using a 3.3V input voltage. However, the analog control circuit under a low input voltage suffers from limited dynamic range. Although a time-based design [2] minimizes the use of analog circuits, it still requires an accurate analog-to-time converter and exhibits only a moderate transient response time. A buck voltage regulator (VR) with a digital controller has been proposed for modern digital systems [3-4], since it can take advantage of the advanced digital process. However, with a conventional digital proportional-integral-derivative (PID) controller it is difficult to offer high bandwidth due to the large power and chip area of the PID and the required multi-bit ADC. Thus, an additional circuit, the resistive transient assist (RTA), was proposed [4]. This paper presents an all-digital synthesizable VR using a digitally-adaptive on-time (DAOT) controller implemented in a CMOS logic process to achieve a fast recovery time.

Ripple-based controllers with ripple-injection circuits, such as the hysteretic or adaptive on-time (AOT) controllers, are well-known schemes to achieve a fast-transient response time to output voltage perturbation [5]. As illustrated in Fig. 27.7.1, the proposed ripple-injection scheme controls the comparator offset digitally, compared to a conventional analog ripple injection controller that uses bulky capacitors and resistors for sensing the inductor's current and that uses analog summation. The digital controller emulates the inductor current slope at a given clock frequency which is faster than the switching frequency of the regulator. Since the comparator of the AOT structure is only activated during the off time, the proposed scheme requires that the digital code only responds to the output voltage, disregarding the input voltage. A single-phase implementation of the DAOT controller using digital ripple injection is shown on the bottom of Fig. 27.7.1. The DAOT controller clocks the comparator using the system clock to immediately respond to load events occurring on each clock cycle of the digital system. The on-time generator, with a digitally-controlled delay line (DCDL), adjusts the on time from the minimum value provided by the constant on time (COT) generator according to the digital code corresponding to the reference voltage. The on time is adapted based on the error information from the ripple-injection process.

For stable and fast output regulation under a fixed switching frequency and for phase synchronization of the multiphase operation, the conventional ripple-based controller uses additional circuits: such as voltage or current sensors, passive filters, and a phase-locked loop [1,5]. In contrast, the DAOT controller digitally adjusts the waveform of the injected ripple to improve the regulation characteristics as illustrated in Fig. 27.7.2. As shown in the top of Fig. 27.7.2, the injection during t_1 is not required due to the fixed on-time operation regardless of the comparison result. The large injection during t_2 determines the worst-case sensing voltage, V_{MIN}, of the output perturbation, thereby the DAOT controller flattens the ripple injection waveform during t_1 and t_2. By flattening the injection waveform, increasing V_{MIN}, while ensuring output stability, the DAOT controller can improve the transient response time. The middle of Fig. 27.7.2 shows the on-time adaptation, which is based on the ripple-injection error for accurate output regulation and for a fixed switching frequency. Since the target value of the ripple injection when COMP_OUT goes to high is "0" in steady state, the error value (ERROR) of on time can be obtained easily using the injection code and COMP_OUT. ERROR is reflected in DLY_CODE via the digital filter, while the DAOT adjusts the on time accordingly via the DCDL. For multiphase operation, the bottom of Fig.27.7.2 shows the self-synchronization scheme using ripple injection. Because the multiphase DAOT controller uses a single comparator for matching between converters, ripple injection should reflect the current information in each phase sequentially. The injection code is increased by ΔV_{INJ} from the injection code at the end of the current phase comparison, so that the injection target always converges to the desired phase shift value, Δt_{PH}. Consequently, the phase of each sub-converter is automatically synchronized after V_{OUT} settles. In addition, since the ripple-injection error in each phase is generated sequentially, the phase is also synchronized by the on-time adaptation loop.

The overall 4-phase structure and the timing diagram of the transient response are depicted in Fig. 27.7.3. Each sub-converter shares a comparator and a DCDL to minimize the voltage and timing mismatches in each phase; thereby, maximizing the conversion efficiency. The digital filter for adaptation is implemented by a simple proportional-integral (PI) controller. Thanks to its structural similarity, the DAOT controller can support a multiphase burst mode for higher light-load efficiency, by simply adding a flip-flop based digital V_{SW} sensor [4] to determine the proper off time for the NMOS power transistor. The 4-phase DAOT controller responds immediately and sequentially to the fluctuations in V_{OUT} via the fast AOT loop control. Following which, the on time is appropriately adapted to the load condition and each phase is re-synchronized.

The DAOT controller with a gate driver is implemented in this work using standard digital cells and an automated design flow. Figure 27.7.4 shows the circuit implementations of the offset-controlled comparator and the on-time generator. The comparator is designed using 3-input NAND gates [6]. The offsets of the comparator is controlled by adjusting the parasitic capacitance of the OP and ON nodes implemented using 2-input NAND gates. To share the DCDL, implemented with delay cells, the short pulse generated on the falling edge of COT_OUT is forwarded to the DCDL and the output of the DCDL is latched to generate the falling edge of AOT_OUT.

The 4-phase VR is fabricated in a 28nm CMOS process. The measurement results are based on a 1.2GHz system clock (CLK_{sys}) frequency and the switching frequency of each phase is 75MHz with a 90° phase shift between each phase. As shown in Fig. 27.7.5, a 1A 10ns load step, at V_{OUT}=0.9V, only causes an 80mV droop with an 80ns recovery time. A -1A load step results in a 105mV overshoot with an 115ns recovery time. For the fast voltage recovery, the switching nodes show the immediate response with a minimum off time of 0.8ns for the droop and a maximum off time of 8.3ns for the overshoot. After the recovery of the output voltage, the switching nodes show that each phase has stabilized and is synchronized. The converter exhibits an 89% peak efficiency, and over 85% efficiency at 1V V_{OUT} over a 50mA to maximum load current range. Figure 27.7.6 compares the performance of the presented system with prior work. The DAOT VR shows a competitive transient response time with the smallest and a fully synthesizable digital controller. Figure 27.7.7 shows the chip micrograph.

References:
[1] B. Lee, et al., "A 25MHz 4-Phase SAW Hysteretic DC-DC Converter with 1-Cycle APC Achieving 190ns t_{settle} to 4A Load Transient and Above 80% Efficiency in 96.7% of the Power Range," *ISSCC*, pp. 190-191, 2017.
[2] J.-G. Kang, et al., "A 10MHz Time-Domain-Controlled Current-Mode Buck Converter with 8.5% to 93% Switching Duty Cycle," *ISSCC*, pp. 424-425, 2018.
[3] H. K. Krishnamurthy, et al., "A Digitally Controlled Fully Integrated Voltage Regulator with On-Die Solenoid Inductor with Planar Magnetic Core in 14nm Tri-Gate CMOS," *ISSCC*, pp. 336-337, 2017.
[4] M. Kar, et al., "An All-Digital Fully Integrated Inductive Buck Regulator with a 250-MHz Multi-Sampled Compensator and a Lightweight Auto-Tuner in 130-nm CMOS," *JSSC*, vol. 52, no. 7, pp. 1825-1835, July 2017.
[5] R. Redl, et al., "Ripple-Based Control of Switching Regulators—An Overview," *IEEE Trans. Power Electron.*, vol. 24, no. 12, pp. 2669-2680, Dec. 2009.
[6] S. Weaver, et al., "Digitally Synthesized Stochastic Flash ADC Using Only Standard Digital Cells," *IEEE Trans. Circuits Syst.I*, vol. 61, no. 1, pp. 84-91, Jan. 2014.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

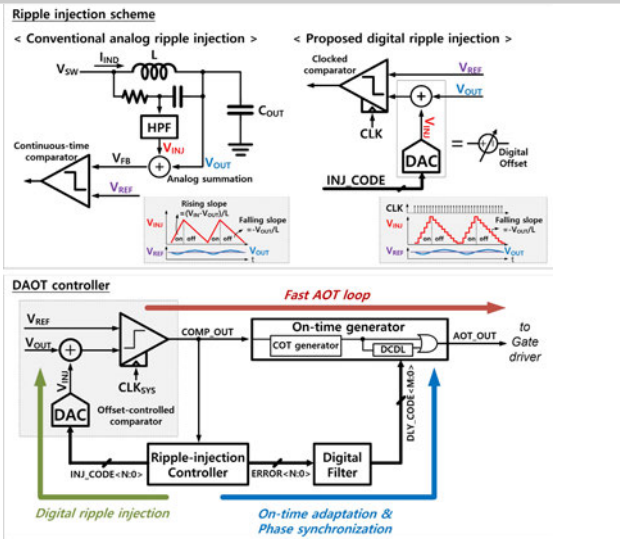


Figure 27.7.1: Comparison between analog and digital ripple-injection (top), and the single-phase block diagram of DAOT controller.

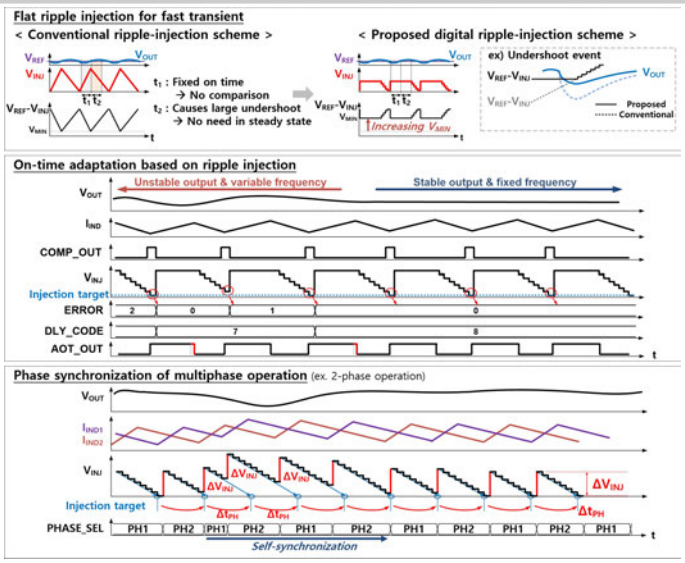


Figure 27.7.2: Various digital ripple-injection schemes with key waveforms.

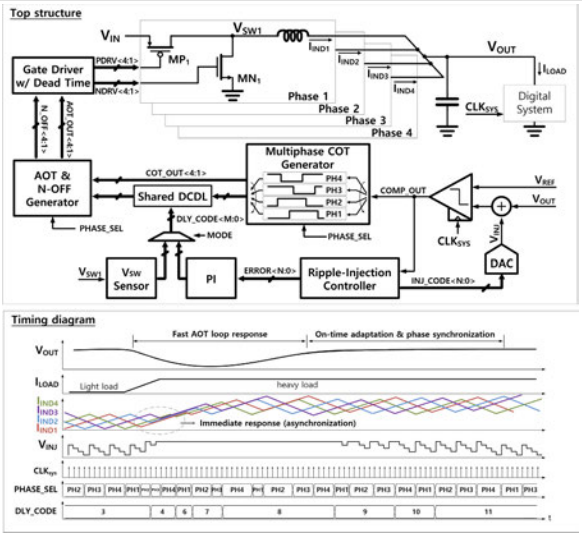


Figure 27.7.3: Structure of the 4-phase VR and the timing diagram of the transient response.

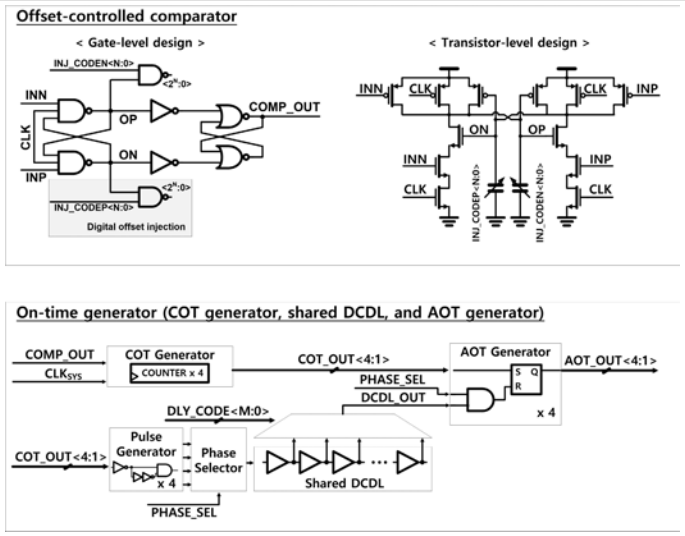


Figure 27.7.4: Circuit implementation of the offset-controlled comparator (top) and the on-time generator (bottom).

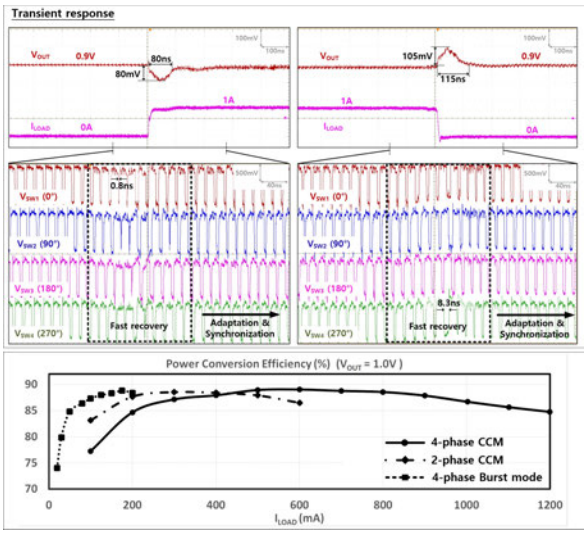


Figure 27.7.5: Measured load-transient waveforms (top) and efficiency plot (bottom).

		ISSCC'17 [1]	ISSCC'18 [2]	ISSCC'17 [3]	JSSC'17 [4]	This work
Technology		0.35um	65nm	14nm	0.13um	28nm
Control		Analog hysteretic	Time-domain	Digital Type-III	Digital PID + RTA	Digital AOT
Mode		CCM	CCM	CCM	CCM + DCM	CCM + Burst mode (DCM)
Phase		4	1	2	1	4
V_{IN} (V)		3.3	1.8	1.5	1.2	1.2
V_{OUT} (V)		0.3 – 2.5	0.15 – 1.69	1.15	0.45 – 1.05	0.6 – 1.0
Maximum I_{LOAD} (A)		6	0.6	0.36 *	0.07	1.2
Peak efficiency (%)		88.1	94.9	84	71	89.0, 87.0, 84.2 (@V_{OUT}=1.0V, 0.9V, 0.8V)
f_{SW} (MHz)		25	10	100	150/250	75
L (nH)		200	220	1.5	11.8	15
C_{OUT} (nF)		2500	4700	5	3.2	200
Area (mm^2)	Controller	0.450 *	0.036	0.011 *	0.139	0.0054
	Die	1.881	2.118	0.4 *	1.19	0.39
I_{LOAD} step		4A/5ns	0.48A/100ns	N.A.	0.075A/100ps	1A/10ns
V_{OUT} droop (mV)		103	100	N.A.	120 *	80
1% settling (ns)		190	3500	N.A.	>100 *	80
Synthesizable design with standard cells?		No	No	No	No	Yes

* estimated from the paper

Figure 27.7.6: Performance summary and comparison.

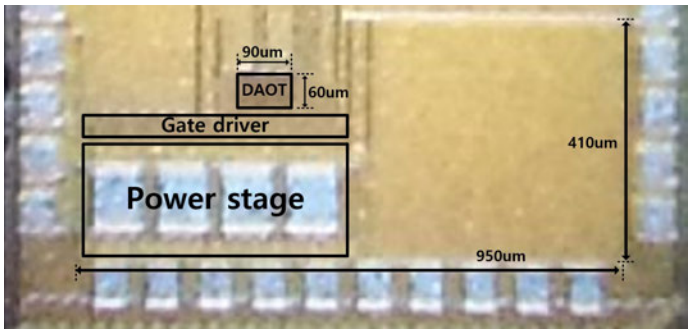


Figure 27.7.7: Chip photomicrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

27.8 Dynamic-Charging Current-Scaling Technique with Dual Accurate Current Control and Temperature Loops with Charging-Current Accuracy up to 99.6% for 1.6× Faster Lithium-Ion Battery Charging

Wei-Tin Lin[1], Zong-Yi Lin[1], Chia-Hao Liu[1], Chia-Ming Huang[1], Li-Cheng Chu[1], Ke-Horng Chen[1], Yin-Hsi Lin[2], Shian-Ru Lin[2], Tsung-Yen Tsai[2], Hann-Huei Tsai[1], Ying-Zong Juang[1]

[1]National Chiao Tung University, Hsinchu, Taiwan
[2]Realtek Semiconductor, Hsinchu, Taiwan

Reliability is one of the most important issues for lithium-ion (Li-Ion) battery fast charging. Due to the limited charge-acceptance rate of a Li-Ion battery, a fast charging technology needs to consider the main problems that can result in Li-Ion battery damage [1]. Pumping large amounts of energy into a Li-Ion battery, faster than the battery chemicals can react to the charge, can result in local polarization of the electrode, overheating, and unpredictable chemical reactions that can damage the Li-Ion battery. Conventional battery chargers use over-temperature and under-temperature protection [2], while some advanced battery chargers contain a temperature-controlled current to limit the charging current for safe charging [3]. In the event of overheating, a fast charging mechanism should stop until the temperature drops to a safe level. Fast charging on/off control is not the single unique factor that determines charging time [4]. Similar to dynamic voltage and frequency scaling (DVFS) technology in power management ICs (PMICs), this paper proposes a dynamic charge current scaling (DCCS) technique, which dynamically adjusts the charging current and the termination charging voltage in constant-current (CC) charge mode depending on the battery temperature (T_{BAT}) and the built-in resistance detection (BIRD) technique.

The proposed DCCS technique, which is assisted by a temperature assisted loop (TAL), is shown in Fig. 27.8.1. DCCS increases or decreases the battery charge current (I_{BAT}) depending on T_{BAT}, which is reflected by V_{TS}. The relationship between battery capacity (C_{BAT}) and T_{BAT} is illustrated in the middle-left of Fig. 27.8.1, which shows that charging the battery under a high T_{BAT} will greatly reduce C_{BAT}. Therefore, it is important to maintain a nominal temperature during charging to avoid reducing C_{BAT}. Based on the T_{BAT} detected by the temperature-to-digital converter (TDC), the TAL can increase the charging current for fast charging by changing the bias voltage V_{DD_GMV} of the GM_V using a digital low-dropout (DLDO) regulator. In addition, I_{BAT} varies under different T_{BAT}. Therefore, it is necessary to determine the termination voltage via BIRD, which is also varies with T_{BAT}, in the TAL by setting the reference voltage $V_{REF(IR)}$ at the input of GM_V. A real-time BIRD allows the proposed battery charger to monitor temperature changes and also monitor different BIRs for safety during fast charging. In an over-temperature condition, the TAL will allow for a rest time and will reduce the charging current by changing V_{DD_GMV} to ensure a reliable charging process and T_{BAT} to avoid damage to the battery (lower right in Fig. 27.8.1). Deviations in I_{BAT} can cause serious damage to the Li-Ion battery, especially during fast charging; therefore, it is necessary to improve the accuracy of charging current under the CC mode. Compared with the inaccurate peak current detection in conventional designs, the proposed accurate current control (ACC) technique obtains an accurate average charging current [4] to ensure that I_{BAT} is precise. The accuracy of I_{BAT} can be increased to approximately 99.6% via the ACC.

The structure of the TAL is shown at the top of Fig. 27.8.2: consisting of a TDC converter and DLDO. The TDC converter is divided into two parts: a temperature sensing core (TSC) and a frequency-to-digital converter (FDC). In TSC, the V_{TS} reflects the T_{BAT} trend and determines the difference between V_H and V_L, which are generated by a folded-cascode sensing core comprising of M_1-M_4. The GM-controlled oscillator (GCO) is shown in the left middle of Fig. 27.8.2. The frequency of f_{CLKT}, which is mainly determined by the GM of the transmission gate, increases when the difference between V_H and V_L decreases. The right middle part of Fig. 27.8.2 depicts a single stage GCO consisting of an inverter and a transmission gate. The delay time of each stage is $(C_{PAR}.(1+G_{MTX}.R_{INV}))/G_{MTX}$. $1/G_{MTX}$ can be expressed as $V_{DD}/(2.I_{TG})$, where I_{TG} is I_{CHARGE} or $I_{DISCHARGE}$, depending on the charging or discharging period. TDC's operating timing diagram is shown in the bottom of Fig. 27.8.2. Based on the current T_{BAT}, f_{CLKT} is converted into a digital code (TDC[7:0]) by the FDC. In FDC, CNT plays the role of counting the trigger times of f_{CLKR} during each cycle of f_{CLKT}, where f_{CLKR} is a temperature independent clock. The output of CNT (D[4:0]) represents the difference between f_{CLKT} and f_{CLKR}. At the same time, TDC[7:0] decides the driving ability of the power MOSFETs in DLDO, which is generated from the look-up table (LUT) and a decoder. The dropout voltage of DLDO raises up as D[4:0] decreases, which means V_{DD_GMV} will be lower to reduce the charging current during CC mode, to prevent the Li-Ion battery from being damaged by a high T_{BAT}.

The ACC controller, in Fig. 27.8.3, can be divided into a dual loop, a constant voltage (CV) loop and a CC loop. The CV loop regulates V_{BAT} and the CC loop regulates I_{BAT} with a high-precision charging current. Due to the nature of the transconductance amplifier, if the battery is not fully charged, V_{EA_GMV} will rise to V_{DD_GMV} during CC mode. The current sense voltage V_{CS}, with a weighting factor related to V_{IN} and V_{BAT}, will reflect I_{BAT} and track V_{DD_GMV} via transconductance GM_C. Therefore, I_{BAT} can be precisely controlled by adjusting the V_{DD_GMV} generated by TAL. Compare the V_{EA_GMC} with the dual ramp to drive the power switches and obtain the corresponding I_{BAT}. The frequency domain analysis of the ACC loop is shown in the lower-left of Fig. 27.8.3. To ensure an accurate I_{BAT} in CC mode, one low-frequency dominant pole is introduced to achieve a high gain, where the zero point needs to be placed at high frequencies to increase the phase margin; which is around 90 degrees. The bottom-right of Fig. 27.8.3 compares a conventional peak current mode (PCM) with the proposed ACC loop. In CC mode, the accuracy of PCM suffers from an offset between V_{EA_GMV} and V_{CS} due to the lack of GM_C. In other words, if PCM is employed, an inaccurate I_{BAT} occurs. In contrast, the proposed ACC loop will effectively eliminate the offset by GM_C.

To ensure that the battery is fully charged, BIRD with an adaptive thermal loss calibration (ATLC) technology generates the $V_{REF(IR)}$ to compensate for the loss of the BIR. The BIRD structure at the top-left of Fig. 27.8.4 has a dual path that reflects the output current and BIR in real time and generates $V_{REF(IR)}$ to compensate for V_{BAT0}. In BIRD, amplifier OP_1 produces a current that is proportional to the output current without the need to add an additional sense resistor to the power path. On the other hand, to handle different BIRs, the methods in [5] and [6] use two charging currents to calculate the slope of V_{BAT}, but these changes are too subtle to be detected. The proposed ALTC technique (bottom of Fig. 27.8.4) senses heat loss ($I^2.R_{BAT}$) and passes the differences in V_{TS} at different I_{BAT} values to I_{DIFF}. The operation of the ALTC technique is shown in the middle-of-the-right side of Fig. 27.8.4. Finally, the delay line analog-to-digital converter (ADC) determines the digital code RC[7:0] corresponding to the different BIRs.

The test chip was fabricated in a 180nm CMOS BCD process. The proposed battery charger with a DCCS technique is demonstrated in the left-top and middle of Fig. 27.8.5. The mode transition of the dual CC and CV loops in the ACC is measured in the upper-right of Fig. 27.8.5. The middle-right of Fig. 27.8.5 shows the benefit of the BIRD technique, where the charging period is reduced by 38.6%; equal to 1.6× speed-up. The accuracy of I_{BAT} is illustrated in the bottom-left of Fig. 27.8.5. The proposed ACC loop can effectively eliminate offset, and the accuracy can be raised up to 99.6%. The bottom-right of Fig. 27.8.5 depicts the charging time with and without the BIRD and ATLC techniques. The charging time can be reduced in the case of different BIRs. Figure 27.8.6 sumarizes the results of this work and compares it to prior art. The BIRD technique can effectively suppress the variation in termination voltage under a wide temperature range and the worst regulation is 0.45% at I_{BAT} = 2A. The peak efficiency is 96.8% for 7V-to-4.2V conversion. Compared to [3], the proposed battery charger, with DCCS technique, can adjust I_{BAT} corresponding to the current T_{BAT}. Whereas, R_{BAT} can be detected in real-time in comparison with [4] due to BIR detection. Figure 27.8.7 shows the chip micrograph.

References:

[1] Y.-M. Ju, et al., "A Hybrid Inductor-Based Flying-Capacitor-Assisted Step-Up / Step-Down DC-DC Converter with 96.56% Efficiency," *ISSCC*, pp. 184-186, 2017.

[2] P.-J. Liu, L. Chien, "A High-Efficiency Integrated Multimode Battery Charger with an Adaptive Supply Voltage Control Scheme," *IEEE Trans. on Power Electronics,* vol. 33, no. 8, pp. 6869-6876, Aug. 2018.

[3] M.-G. Jeong, C. Yoo, "Switching Battery Charger Integrated Circuit for Mobile Devices in a 130-nm BCDMOS Process," *IEEE Trans. on Power Electronics,* vol. 31, no. 11, pp. 7943-7952, Nov. 2016.

[4] Y.-H. Jung, et al., "A Fast and Highly-Accurate Battery Charger with Accurate Built-In Resistance Detection," *IEEE Trans. on Power Electronics*, vol. 33, no. 12, pp. 10051-10054, Dec. 2018.

[5] C.-H. Lin, et al., "A Li-Ion Battery Charger with Smooth Control Circuit and Built-In Resistance Compensator for Achieving Stable and Fast Charging," *IEEE TCAS-I, vol. 57, no. 2,* pp. 506-517, Feb. 2010.

[6] T.-C. Huang, et al., "Fast Charging and High Efficiency Switching-Based Charger with Continuous Built-In Resistance Detection and Automatic Energy Deliver Control for Portable Electronics," *JSSC*, vol. 49, no. 7, pp. 1580-1594, July 2014.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

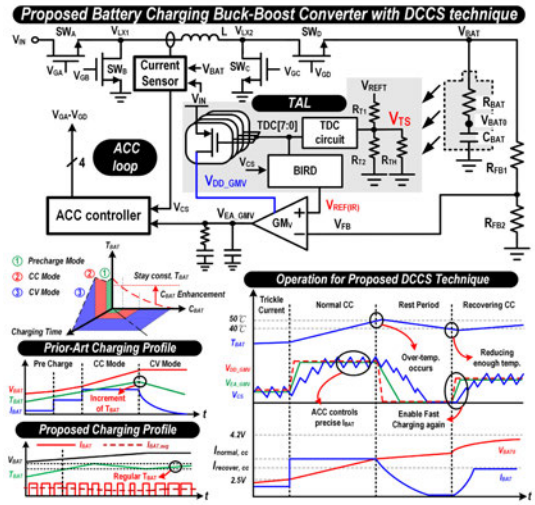


Figure 27.8.1: Structure of the proposed buck-boost battery charger with the TAL technique (top); Analysis of the relationship with temperature and charging time (bottom).

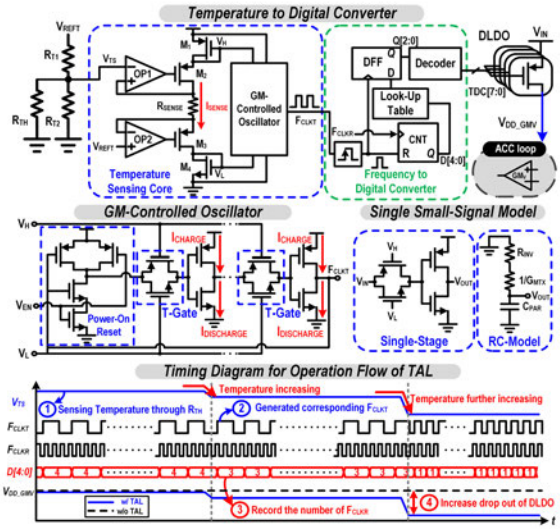


Figure 27.8.2: Circuit Implementation of proposed TAL with the TDC circuit (top); Timing diagram for the operation flow of TAL (bottom).

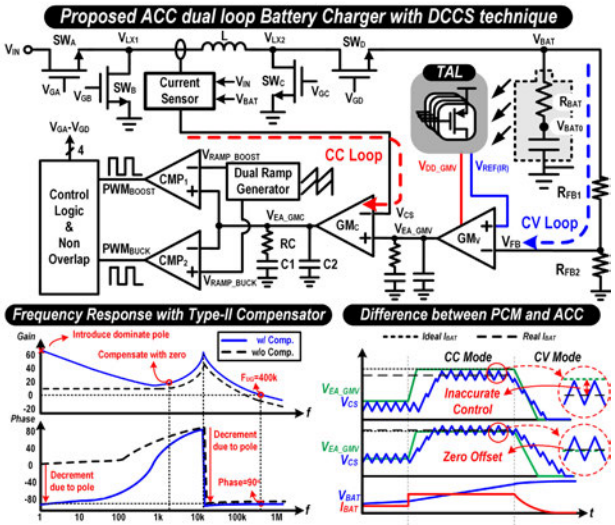


Figure 27.8.3: Structure of ACC controller with the TAL and its frequency response and comparison between PCM and ACC loop.

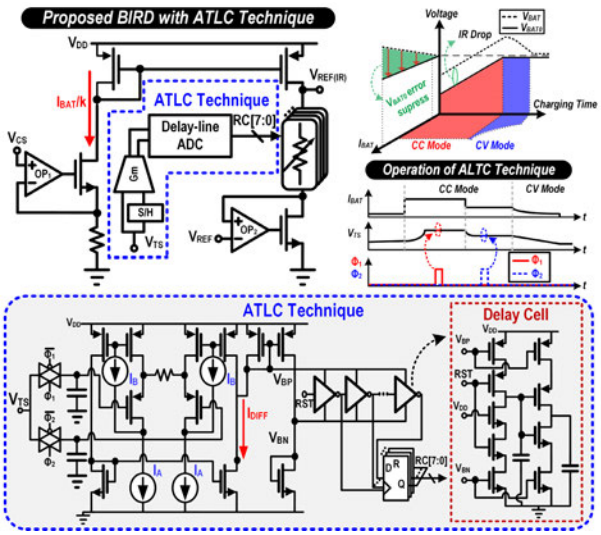


Figure 27.8.4: Circuit Implementation of the proposed BIRD with ALTC technique and its timing diagram.

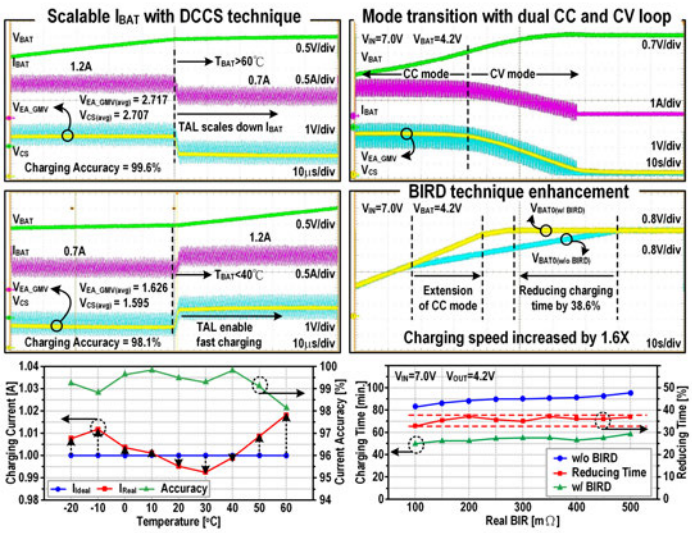
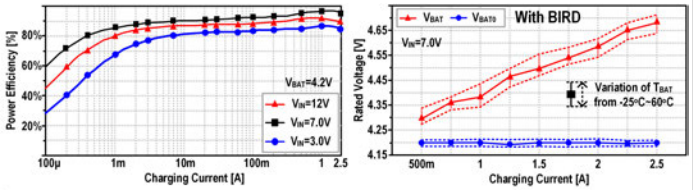


Figure 27.8.5: Measurement results of the battery charger using DCCS technique and BIRD.



	[1] ISSCC'17	[3] TPE'16	[4] TPE'18	This Work
Process	0.18μm BCD	0.13μm BCD	0.18μm	0.18μm BCD
Topology	Buck-Boost	Buck	Buck	**Buck-Boost**
Input voltage range [V]	2.7-4.2	6.0-16.0	4.5-5.5	**2.0-16.0**
Output voltage [V]	3.4	2.5-4.2	3.0-4.2	**2.5-4.2**
Temperature loop	No	Yes	No	**Yes**
Maximum Output Current [A]	2	1.5	1.974	**2.5A**
BIR detection	No	No	Continuous	**Real-time**
Reduce charging time	No	No	41.2%	**38.6%**
Max. Power Efficiency	96.56%	92%	87.4%	**96.8%**
Switching frequency	10MHz	1.5MHz	N/A	**2.4MHz**
Die size [mm²]	5.67mm2	12.25mm²	4mm²	**3.3mm²**

Figure 27.8.6: Performance of the battery charger using DCCS technique including variation of termination voltage, and efficiency under different output charging currents (top); Comparison table with state-of-the-arts (bottom).

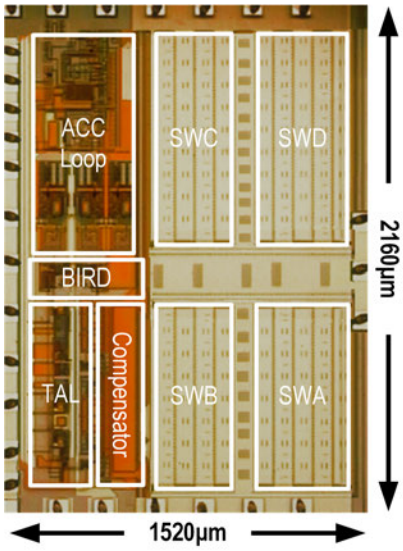


Figure 27.8.7: Chip micrograph with a silicon area of 3.3mm^2.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 28 Overview:
Techniques for Low-Power & High-Performance Wireless
WIRELESS SUBCOMMITTEE

Session Chair: ***Danielle Griffith***
Texas Instruments, Dallas, TX

Associate Chair: ***Alan Wong***
EnSilica, Abingdon, United Kingdom

Subcommittee Chair: Stefano Pellerano, Intel, Hillsboro, OR

Wireless technologies continue to penetrate and support a wide range of application areas. This session includes state-of-the-art improvements in wake-up radios enabling ultra-low-power and increased blocker rejection, a highly efficient Bluetooth Low Energy transmitter, and a wireless power transfer system supporting enhanced data rates. Furthermore techniques to boost wireless performance are presented, including non-magnetic circulators for improved full-duplex communications at mm-wave and RF frequencies, a wideband blocker-tolerant receiver with low LO leakage and a modified Cartesian feedback transmitter that addresses the noise-linearity trade-off of the conventional architecture.

1:30 PM

28.1 A 0.42nW 434MHz -79.1dBm Wake-Up Receiver with a Time-Domain Integrator
V. Mangal, Columbia University, New York, NY
In Paper 28.1, Columbia University describes a wake-up receiver in 65nm CMOS LP using a 40-stage MOS RF envelope detector and time-domain windowed integrator. The receiver has -78.3/-79.1dBm sensitivity at 151.25/434.4MHz with 110ms latency, while consuming 370/420pW from 0.4V.

2:00 PM

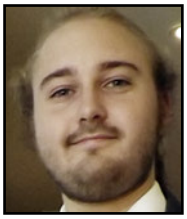
28.2 A 220μW -85dBm Sensitivity BLE-Compliant Wake-up Receiver Achieving -60dB SIR via Single-Die Multi-Channel FBAR-Based Filtering and a 4-Dimentional Wake-Up Signature
P-H. P. Wang, University of California, San Diego, La Jolla, CA
In Paper 28.2, the University of California, San Diego describes a BLE-compliant wake-up receiver in 65nm CMOS using single-die 3-channel FBAR-based filtering and a frequency-hopping time-counting majority-voting algorithm providing 4D signature detection. The receiver achieves -85dBm sensitivity and up to 60dB blocker rejection while consuming 220μW.

2:30 PM

28.3 A 606μW mm-Scale Bluetooth Low-Energy Transmitter Using Co-Designed 3.5×3.5mm^2 Loop Antenna and Transformer-Boost Power Oscillator
Y. Shi, University of Michigan, Ann Arbor, MI
In Paper 28.3, the University of Michigan describes a Bluetooth Low Energy transmitter in 65nm CMOS using a co-designed 3.5×3.5mm^2 loop planar antenna and transformer-boost power oscillator for improving output power and transmit efficiency. The transmitter achieves 23.6% overall efficiency with a measured phase noise of -118.5dBc/Hz at 1MHz offset, resulting in 2.1% FSK error and <10kHz frequency drift during BLE packet.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

2:45 PM



28.4 A High-Q Resonant Inductive Link Transmit Modulator/Driver for Enhanced Power and FSK/PSK Data Transfer Using Adaptive-Predictive Phase-Continuous Switching Fractional-Capacitance Tuning

H. Kennedy, University of Southampton, Southampton, United Kingdom

In Paper 28.4, the University of Southampton describes a 10kHz-to-2MHz combined antenna tuner/driver and modulator for high-Q inductively-coupled systems in 0.18μm CMOS/LDMOS using adaptive-predictive phase-continuous switching fractional-capacitance tuning. The technique enables FSK/PSK modulation beyond normal Q-factor/bandwidth restrictions, achieving 296kb/s QPSK at 2MHz with Q = 50.

3:15 PM

28.5 Non-Magnetic 60GHz SOI CMOS Circulator Based on Loss/Dispersion-Engineered Switched Bandpass Filters

A. Nagulu, Columbia University, New York, NY

In Paper 28.5, Columbia University describes a non-magnetic CMOS 60GHz circulator based on spatio-temporal conductivity modulation (STCM) across a loss/dispersion-engineered bandpass filter. The circulator achieves 3.6/3.1dB loss, isolation >40dB over 1.3GHz, 3.2dB NF, >+19.5dBm IP_{1dB} and spurs <-30dBc.

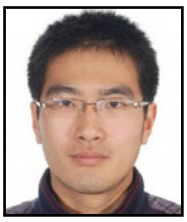
3:45 PM

28.6 Full-Duplex 2×2 MIMO Circulator-Receiver with High TX Power Handling Exploiting MIMO RF and Shared-Delay Baseband Self-Interference Cancellation

M. Baraani Dastjerdi, Columbia University, New York, NY

In Paper 28.6, Columbia University and Oregon State University describe a 2×2 MIMO full-duplex receiver with integrated N-path-filter-based non-magnetic circulators. The solution features area- and power-efficient passive RF and baseband active wideband MIMO cancellation with shared delay elements to address the $O(N^2)$ challenge while enhancing TX power handling by 8dB.

4:15 PM

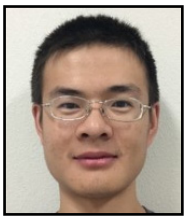


28.7 A Wideband Blocker-Tolerant Receiver with High-Q RF-Input Selectivity and <-80dBm LO Leakage

H. Wang, University of California, Irvine, CA

In Paper 28.7, the University of California Irvine describes a CMOS wideband blocker-tolerant receiver realizing high-Q RF-input selectivity for linearity improvement while maintaining a low LO leakage. It achieves a noise figure of 2.45dB, out-of-band IIP3/IIP2 are +14/+60dBm respectively and LO leakage is below -80dBm from 0.2 to 2GHz.

4:45 PM



28.8 A 21dBm-OP_{1dB} 20.3%-Efficiency -131.8dBm/Hz-Noise X-band Cartesian-Error-Feedback Transmitter with Fully Integrated Power Amplifier in 65nm CMOS

J. Li, University of California, Davis, CA

In Paper 28.8, the University of California Davis describes an X-band Cartesian error-feedback transmitter with fully integrated PA in 65nm bulk CMOS. The chip achieves 21dBm OP_{1dB} with 20.3% system efficiency, and -131.8dBm/Hz out-of-band noise.

28.1 A 0.42nW 434MHz -79.1dBm Wake-Up Receiver with a Time-Domain Integrator

Vivek Mangal, Peter R. Kinget

Columbia University, New York, NY

PicoNodes have applications in warehouse inventory, smart homes, or integrated patient monitoring and require ultra-low power consumption to eliminate battery replacement or enable batteryless systems. A small form factor is key for ubiquitous deployment. Recently, wake-up receivers [1,2] have traded off latency for improved sensitivity for non-latency-critical applications. However, the improved sensitivity requires a high-Q front-end inductor that limits operation to below 200MHz; this stands in the way of a small antenna form factor. Higher-frequency operation makes a smaller antenna possible, but entails lower-Q inductors and higher package parasitics. A receiver with a gate-biased self-mixer [3] has demonstrated a moderate sensitivity at 550MHz with a relatively low-Q matching network, however, the self-mixer introduces extra input capacitance.

Fig. 28.1.1 shows the proposed wake-up receiver architecture to receive an 11b wake-up code OOK modulated at 100b/s on the RF carrier. We introduce gate-biasing for a MOS envelope detector (ED) to reduce its input capacitance and enable high-frequency operation, as well as a time-domain, baseband windowed integrator and a first-order DLL bias loop, to drastically reduce power consumption. The RF input signal passes through the L-C matching network and hits a 40-stage MOS ED. The DC gate potential, V_{gate_bias}, generated using a current-biased replica ED, biases the ED transistors in the weak-inversion linear region. The matching network provides 23dB of passive gain, and combined with the ED, the RF-to-baseband conversion $v_{bb,edo}/P_{in}$ is 21mV/nW. The ED baseband output $v_{bb,edo}$ is amplified by 26dB with a g_m-reuse amplifier and then fed to a clocked voltage-controlled delay line ($VCDL_A$). When the OSC CLK is high, $VCDL_A$ and a reference $VCDL_B$ are put in a VCO configuration and encode the signal and reference $V_{delay,ref}$ in the time domain. They operate as windowed integrators with a sinc frequency response that rejects high-frequency noise before digitization. Its outputs pass through delay lines to set the comparator threshold and are then compared using a PFD and latched. The comparator output is fed to an 11b digital correlator to detect the wakeup signature. The outputs of the VCDLs are also compared using a PFD and fed back to the ED reference input $v_{ed,ref}$. This first order, low bandwidth, delay-locked loop rejects the DC signal due to continuous-wave interferers at the receiver input and any DC offsets introduced by the baseband circuits.

Optimizing the RF front-end design for wake-up receivers is critical to maximize sensitivity. The maximum Q-factor of the inductor available for the matching network and the input capacitance of the ED decide the achievable front-end passive gain and operating frequency. For a given inductor Q-factor, the achievable passive gain increases with the input resistance of the ED, $R_{in,ed}$ (Fig. 28.1.2). While a high $R_{in,ed}$ is desired for maximum passive gain, due to the passive nature of the ED, with increasing input resistance, the output resistance also increases, thus increasing the noise floor at the output of the ED. Optimizing the SNR at the output of the ED requires a careful selection of $R_{in,ed}$. With gate biasing, the ED $R_{in,ed}$ can be tuned. A 3.5dB improvement in sensitivity is obtained by reducing the $R_{in,ed}$ from 1MΩ (value for max. gain) to 50kΩ (value for max. SNR) for a 108nH inductor with a Q-factor of 200 at 433MHz. Additionally, the gate-biasing technique doesn't affect the conversion gain, makes the ED independent of processes variations, and provides a significant reduction of the input capacitance to only 44fF for a 40-stage ED. In our design, we selected a 200kΩ $R_{in,ed}$, which is only 1dB removed from the optimum while giving 2× advantage in power consumption.

A 40-stage gate-biased ED with an $R_{in,ed}$ of 200kΩ has an $R_{out,ed}$ of 320MΩ. For a current-reuse baseband amplifier (Fig. 28.1.1), the input-referred noise is $4kT\gamma/(2g_m)$. Assuming γ=1.2 and g_m/I_d=29, an I_d=370pA offers a simulated 1dB amplifier NF compared to the ED noise and provides a 26dB gain, while only consuming 150pW at 0.4V. The amplifier is current-biased using a PMOS current mirror with AC coupling while the NMOS input transistor is biased through the delay-locked loop. Figure 28.1.2 shows the noise PSDs demonstrating a 1dB amplifier NF in the signal bandwidth from 10Hz to 100Hz. The low-frequency flicker noise is rejected by the DLL.

A time-domain windowed integrator was proposed in [4] for power-efficient ADCs, but can be used here to serve as a matched detector for the rectangular bit shape; it filters the high-frequency baseband noise and improves SNR before sampling. Two voltage-controlled delay lines ($VCDL_A$ and $VCDL_B$) with clocked feedback realize a V-to-T signal conversion and time-domain integration (Fig. 28.1.3). The integration begins at the rising edge of the OSC CLK running at 200S/s for 2× oversampling, which is derived from the REF CLK provided by an on-chip oscillator operating at 1.6kHz.

In the signal path, current-controlled delay cells with a different delay (τ_n and τ_p) in each branch after VCDLs determine the comparison threshold, which allows to keep the false-alarm rate below 1/hr. A PFD compares the output phases and a clocked latch samples the output at the end of the integration period. The latch output then passes through a D-flipflop to the digital correlator. The sampling instances are not synchronized with the incoming data and, due to possible misalignment, with 2× oversampling every alternate bit can be unreliable. The on-chip 100pW 11b sliding-window digital correlator skips every alternate bit and thus correlates with the most reliable data and is flexible across different codes. Figure 28.1.4 shows the corresponding correlator coefficients, the correlator implementation and receiver response to a -79dBm wake-up signal at 434.4MHz.

In the DLL path (Figs. 28.1.1&3), a PFD-CP samples the phase offset of VCDLs and feeds the lowpass filtered DC offset back at the reference input of the ED. The 1pA CP with a 40pF load is enabled every alternate sample for 1/8× the clock period at the end of integration cycle to reduce the loop bandwidth to 5Hz. The integrator response from the CP ensures that the DC phase offset at the VCDL output is minimized, which is exactly what is desired for the comparator in the signal path. This allows for drastically reducing the size of VCDL integrators without affecting the baseband offset, and then for operating them at only 100pW.

The 65nm LP CMOS wake-up receiver prototypes (Fig. 28.1.7) have been measured at 151.25MHz and 434.4MHz, providing a 3dB bandwidth of 4MHz and 10MHz, using two ICs with the respective matching networks; the corresponding S_{11}'s are in Fig. 28.1.5. The power consumption from the 0.4V supply was measured using an 8.5-digit digital multimeter; at 434MHz the total consumption was 420pW (breakdown in Fig. 28.1.7) and at 151.25MHz it was 370pW. The missed detection ratio (MDR) was measured (Fig. 28.1.5) using the on-chip correlator with a code of "11100100110" at 100bps. For an MDR of 10^{-3}, a -78.3dBm/-79.1dBm sensitivity was obtained for 151.25MHz/434.4MHz. The signal-to-interference ratio (SIR) to maintain an MDR of 10^{-3} was measured (Fig. 28.1.5) with the desired signal at -77dBm. For a worst-case AM interferer 100% modulated with 1010-pattern at 100bps, the SIR was -3.7dB/+5.8dB at 3MHz offset for 151.25MHz/434.4MHz; the DLL does not offer rejection to AM interference and the SIR frequency profile follows the S_{11} profile. For a continuous-wave interferer at 3MHz offset, the SIR is as good as -23.3dB/-13.8dB for 151.25MHz/434.4MHz, thanks to the additional 20dB rejection by the DLL; the maximum rejection is limited by the dynamic range of the DLL CP. Figure 28.1.6 shows the comparison with state-of-the-art wake-up receivers. The receiver provides an 8dB better sensitivity at 434.4MHz and a 3dB better sensitivity at 151.25MHz, both at a 10× lower power consumption compared to the prior art.

Acknowledgments:
The authors thank partial support from Analog Devices and NSF EECS 1309721.

References:
[1] J. Moody et al., "A –76dBm 7.4nW Wakeup Radio with Automatic Offset Compensation," *ISSCC*, pp. 452-454, Feb. 2018.
[2] P. P. Wang et al., "A Near-Zero-Power Wake-Up Receiver Achieving –69-dBm Sensitivity," *IEEE JSSC*, vol. 53, no. 6, pp. 1640-1652, June 2018.
[3] V. Mangal and P. R. Kinget, "An Ultra-Low-Power Wake-Up Receiver with Voltage-Multiplying Self-Mixer and Interferer-Enhanced Sensitivity," *IEEE CICC*, pp. 1-4, 2017.
[4] S. E. Hsieh and C. C. Hsieh, "A 0.4V 13b 270kS/S SAR-ISDM ADC with an Opamp-less Time-Domain Integrator," *ISSCC*, pp. 240-242, Feb. 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

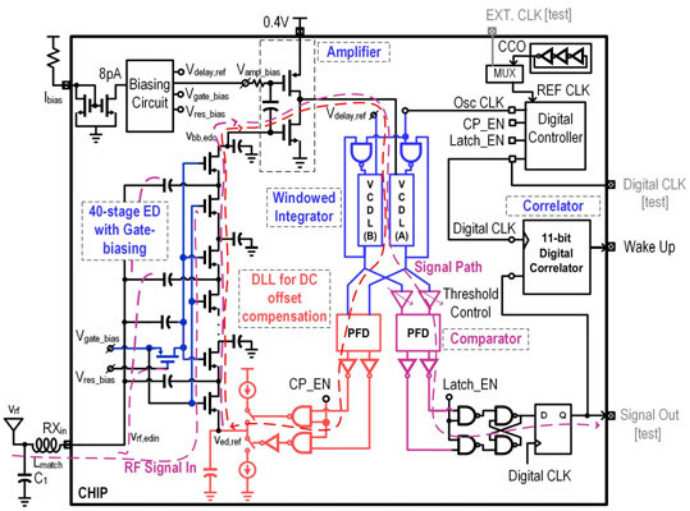


Figure 28.1.1: Fully integrated wake-up receiver using an ED with gate-biasing, a time-domain windowed integrator, an 11b correlator, and a first order DLL.

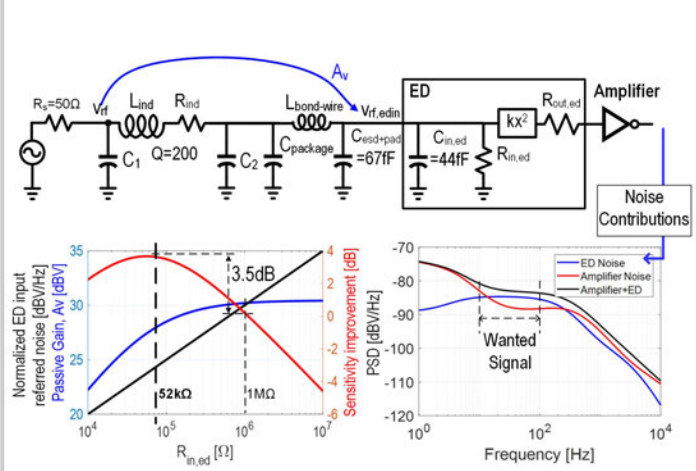


Figure 28.1.2: Optimization of the energy detector's input resistance $R_{in,ed}$ for optimal sensitivity at 434MHz.

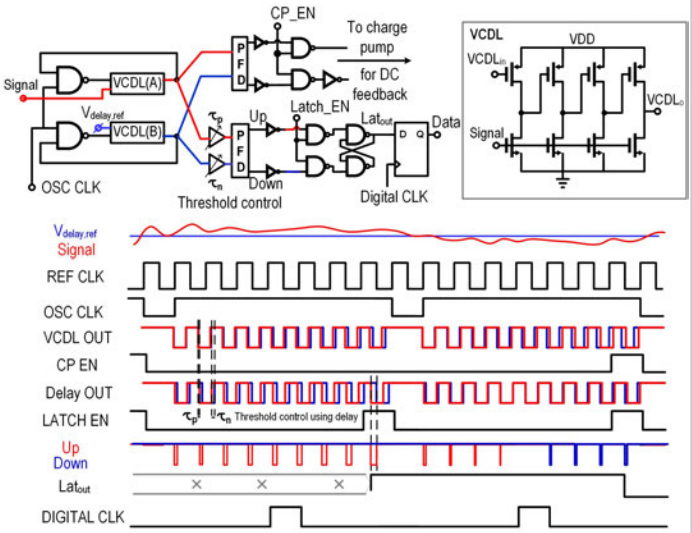


Figure 28.1.3: Operation of the time-domain windowed integrator and comparator.

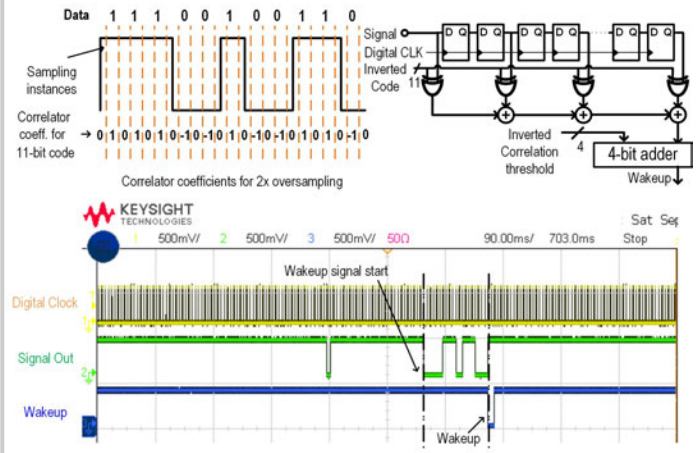


Figure 28.1.4: Implementation of the low-power 11b correlator and it's response for a -79dBm wake-up signal at 434.4MHz.

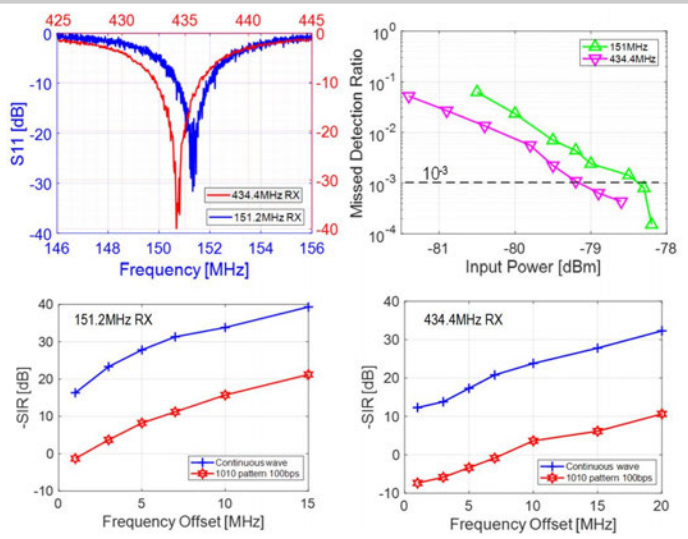


Figure 28.1.5: Receiver RF input reflection S_{11}, missed detection ratio and sensitivity, inverse of SIR for continuous-wave and worst-case AM interferer.

		Wang ESSCIRC'17	Jiang ISSCC'17	Sadagopan RFIC'17	Mangal CICC'17	Moody ISSCC'18		This Work	
Features		Active Rectifier	Active Rectifier	High-Q FE Co-design	Self-Mixer	Passive Rectifier + Auto-threshold compensation		Gate-biased ED + Time-domain Integrator	
Tech Node		180 nm	180 nm	65 nm	130 nm	130 nm		65nm LP	
Frequency	[MHz]	405	113.5	2400	550	151.8	433	**434.4**	**151.25**
Modulation		OOK	OOK	OOK	OOK	OOK		OOK	
Chip Rate	[kbps]	0.3	0.3	2.5	400	0.2		0.1	
Code Length		16-bit	16-bit	32-bit	11-bit	16-bit		11-bit	
Power	[nW]	4.5	4.5	365	222	7.4		**0.42**	**0.37**
Passive Gain	[dB]	18.5	25	N/A	19	NR	NR	23	23
Sensitivity	[dBm]	-63.8	-69	-61.5	-56.4	-76	-71	**-79.1**	**-78.3**
Latency	[mSec]	53.3	53.5	12.8	0.028	82.5	82.5	**110**	**110**
In Band SIR[1] [3MHz off.]									
AM Jammer	[dB]	N.R.	-4	N.R.	-2	-4	N.R.	5.8	-3.6
Cont. Wave	[dB]	N.R.	-11	N.R.	-14	-30	N.R.	-14	-23.3
Latency Norm. Sens.	[dBm]	-70.2	-75.4	-71.2	-79.2	-81.4	-76.4	**-83.9**	**-83.1**
FoM	[dB]	-154	-159	-136	-146	-163	-158	**-178**	**-177**

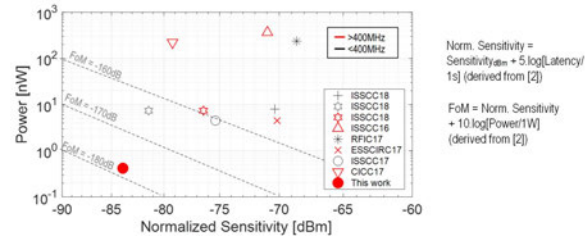


Figure 28.1.6: Summary of performance and comparison with state of the art.

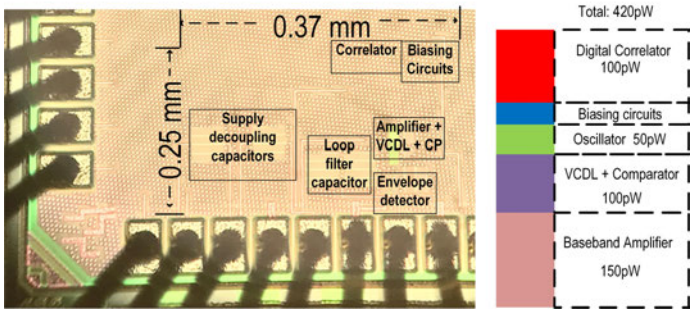


Figure 28.1.7: Die micrograph (active area is 0.093mm²) and breakdown (in part based on simulations) of the measured 0.42nW total power consumption at 434MHz.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

28.2 A 220μW -85dBm Sensitivity BLE-Compliant Wake-up Receiver Achieving -60dB SIR via Single-Die Multi-Channel FBAR-Based Filtering and a 4-Dimentional Wake-Up Signature

Po-Han Peter Wang, Patrick P. Mercier

University of California, San Diego, La Jolla, CA

Wake-up receivers (WuRXs) offer an attractive low-power means to synchronize low-average-throughput wireless devices without requiring energy-expensive periodic synchronization routines between primary radios. To achieve low-power operation, most prior-art WuRXs tend to utilize custom OOK or FSK modulation schemes that are not compliant with any standards, while also forgoing important capabilities present in more powerful radios, such as the ability to operate in the presence of potentially large interferers or dynamically switch between multiple channels – both of which are very important when operating in congested bands. Since Bluetooth Low Energy (BLE) is one of the most popular standards for low-power IoT devices, prior work has endeavored to develop BLE-compliant WuRXs that achieve low-power operation through means of back-channel (BC) communication – signals that are generated by a standard-compliant BLE TX, yet encode information in a modality that can be demodulated by low-power means (e.g., direct energy detection [1,2]). Unfortunately, the wideband energy detectors in [1,2] not only directly demodulate interferers with no channel selectivity, they also introduce significant demodulated RF noise, limiting sensitivity to -56.5dBm at 236nW and -58dBm at 164μW, respectively. To impart channel selectivity, a mixer-first architecture was demonstrated in [3], achieving a sensitivity of -80dBm at 230μW, though this does not include the power of a PLL or FLL to stabilize the free-running single-channel VCO. To facilitate stabilized multichannel operation, a 120μW crystal-stabilized FLL-based LO generator was included in [4] with frequency hopping support, for a total WuRX power of 150μW. While selective, the design achieved a sensitivity of only -57.5dBm.

This paper presents a BLE-compliant WuRX, shown in Fig. 28.2.1, that achieves -85dBm sensitivity at 220μW while supporting multichannel frequency diversity and interference resiliency by: 1) employing a single-die 0.25mm^2 3-channel FBAR for direct filtering of the 3 BLE advertisement channels; 2) integrating the energy in each advertisement channel via a frequency-hopping time-counting majority-voting algorithm for low-cost BLE-compliant energy-detection-based demodulation over four distinct dimensions; 3) mixing each advertisement channel down to low-IF prior to energy-detection via a two-stage heterodyne structure for further channel filtering and adjacent channel rejection; and 4) implementing the 3-channel LO generator via an ultra-low-power crystal-stabilized integer-*N* PLL with frequency tripler.

The block diagram of the WuRX is shown in Fig. 28.2.1. After 9dB of passive voltage gain, RF signals are filtered by the 3-channel single-die FBAR filter, a single channel of which is selected via an integrated SP3T switch that, via an on-resistance of 16.7Ω, degrades the *Q* of the FBAR from 1040 to 370. As shown in Fig. 28.2.2, an RF LO, centered 8MHz lower than the channel, is used to downconvert one of the 3 channels to an 8MHz IF. The LO for the three BLE advertisement channels at 2402, 2426, and 2480MHz are generated in a low-power manner via an integer-*N* PLL, shown in Fig. 28.2.2 (bottom), that takes an 8MHz crystal reference, divides it by 4, and compares to an oscillator that, instead of operating at 2.4GHz, operates at 798, 806, or 824MHz via divider ratios of *N*={399, 403, 412} to save power; the final LO frequency is subsequently generated via an AND-based frequency-tripling edge combiner, enabled via the naturally-occurring multiple phases in the ring oscillator. As part of the tripler, the LO drives a 2-phase 6-switch passive mixer, the output of which passes through an inverter-based IF amplifier. The IF signal is then, thanks to the careful frequency planning, easily mixed to baseband via the 8MHz crystal reference and an *N*-path filtering passive mixer. Fortunately, the 500kHz-spaced FSK tones generated by the BLE standard are, for the purposes of energy detection, self-images of each other when centered around the 8MHz IF LO, and thus a 3-stage lowpass filter and a 27-to-45dB gain PGA are used to condition the 250kHz downconverted signal. Envelope detection is then performed by a cross-coupled self-mixing envelope detector (ED) to provide a pseudo-differential BB signal and forgo the need of a precise voltage reference for the following 25× oversampled 2-stage dynamic comparator, which utilizes capacitive trimming to set the comparison threshold.

An on-chip digital baseband and finite state machine (FSM) are used to parse BB data to decide if a wake-up signature is detected. Similar to [2], two dimensions are used to detect the presence of an advertisement packet at a single frequency: the packet must meet a minimum power threshold defined by both the PGA and the comparator threshold, and be within a pre-specified packet length tolerance between T_L and T_H, which are the lower and upper bound of the targeted packet length (a single BLE packet length can be between approximately 128 and 376μs with 1μs steps). Unlike [2], however, and in a similar manner to [4], though here with majority-voting and with an on-chip digital BB, two additional dimensions are added to the WuRX: channel frequency, and the hopping time between channels. The on-chip FSM begins by detecting energy in Channel 37. If the correct single-channel signature is received, the FSM moves to Channel 38, and then Channel 39 according to desired timing, which is programmable. If Channel 38 has an overlapping Wi-Fi or BLE blocker whose packet is longer than desired, as illustrated in Ex. 1 in Fig. 28.2.1, Channel 38 will miss detection, but if Channels 37 and 39 hit, the 4-dimensional wake-up signature is still detected. The 4D signature also helps prevent false alarms – for example, if Channel 37 hits due to a blocker, yet Channels 38 and 39 do not, no false alarm is observed as illustrated in Ex. 2 in Fig. 28.2.1. There are over 38,000 possible signatures available to decode in this design. The state machine diagram governing the sequencing behavior is shown in Fig. 28.2.3.

The BLE-compatible WuRX was fabricated in 65nm CMOS, and a single-die 3-channel FBAR is stacked and wirebonded on top. The FBAR filters have 3dB bandwidths of 9.7MHz, temperature coefficients of 0.89 ppm/°C [5], and bondwire length variation can alter the center frequency by up to 200kHz/100μm. Figure 28.2.4 (top left) shows the measured S_{11} of the WuRX when configured to operate in each of the advertisement channels, demonstrating the frequency-selectivity of the FBARs. The output of the PLL before the edge combiner is shown in Fig. 28.2.4 (top right). With the entire system operating together, the WuRX achieves a sensitivity of -85dBm for a missed detection rate of 10^{-3} and a false-alarm rate less than once per half hour, as shown in Fig. 28.2.4 (bottom left). The filtering characteristics of the FBAR and the heterodyne structure help a single channel achieve a signal-to-interference ratio of -23dB at a 10MHz offset (Fig. 28.2.4, bottom right). If the FBAR notch were placed at 16 instead of 27MHz offset, it would have improved image rejection by a further ~10dB; this can be easily rectified by FBAR tuning. More importantly, however, the frequency-hopping voting scheme enables extremely large blockers to be located near any single channel. Testing to the limit of the employed signal generator revealed that, with frequency-hopping voting, up to -60dB SIR could be achieved for both BLE and 802.11g Wi-Fi jammers, with that number likely to be higher with availability of more capable test equipment.

To validate realistic use, the WuRX was connected to an antenna in a normal lab environment and was demonstrated to wake up from a commercial mobile phone, yet the WuRX did not wake up when an adjacent phone transmitted packets of a different length, as shown in Fig. 28.2.5 (top). The WuRX also woke up correctly when two phones playing videos over Wi-Fi were placed next to the antenna, while a phone that delivered the BLE advertisement packets was placed upwards of 10m away (1m shown in Fig. 28.2.5, bottom for clarity). During continuous operation, the WuRX consumes 220μW. According to Fig. 28.2.6, the proposed WuRX achieves the best sensitivity and highest amount of interference rejection amongst all BLE-compatible WuRXs at comparable power consumption in the table. A photograph of the die-stacked WuRX is shown in Fig. 28.2.7.

Acknowledgements:
The authors would like to acknowledge Broadcom Inc. for providing the FBAR die.

References:
[1] N. E. Roberts et al., "A 236nW -56.5dBm-Sensitivity Bluetooth Low-Energy Wakeup Receiver with Energy Harvesting in 65nm CMOS," *ISSCC*, pp. 450-451, Feb. 2016.
[2] M. Ding et al., "A 2.4GHz BLE-Compliant Fully-Integrated Wakeup Receiver for Latency-Critical IoT Applications Using a 2-Dimensional Wakeup Pattern in 90nm CMOS," *IEEE RFIC*, pp. 168-171, June 2017.
[3] M. R. Abdelhamid et al., "A –80dBm BLE-Compliant, FSK Wake-up Receiver with System and Within-Bit Duty Cycling for Scalable Power and Latency," *IEEE CICC*, pp. 1-4, April 2018.
[4] A. Alghaihab et al., "A 150 μW -57.5 dBm-Sensitivity Bluetooth Low-Energy Back-Channel Receiver with LO Frequency Hopping," *IEEE RFIC*, pp. 324-327, June 2018.
[5] K. Wang et al., "Design of 1.8-mW PLL-Free 2.4-GHz Receiver Utilizing Temperature-Compensated FBAR Resonator," *IEEE JSSC*, vol. 53, no. 6, pp. 1628-1639, June 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

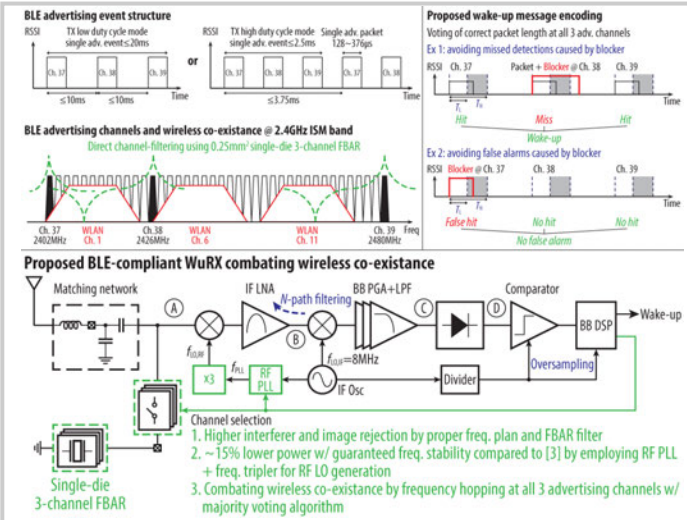


Figure 28.2.1: BLE advertising event structure, channel allocation, and frequency hopping wake-up messages (top); proposed interference-resilient BLE-compliant WuRX architecture (bottom).

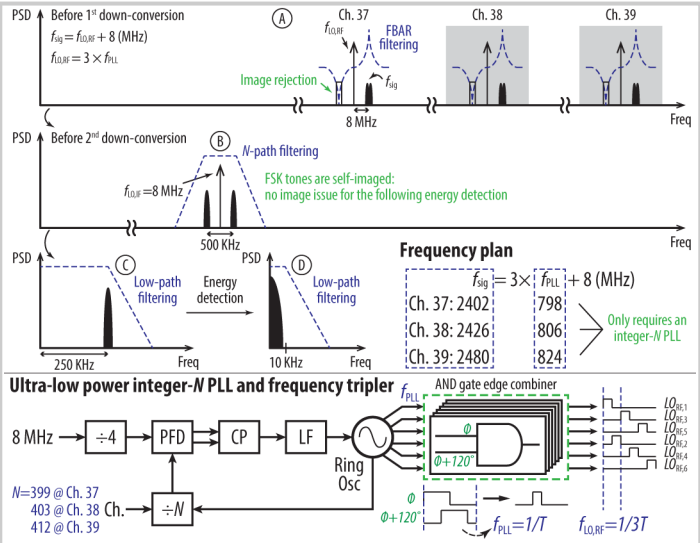


Figure 28.2.2: Frequency plan and multi-stage filtering of the proposed WuRX (top); schematic of the ultra-low-power PLL and frequency tripler (bottom).

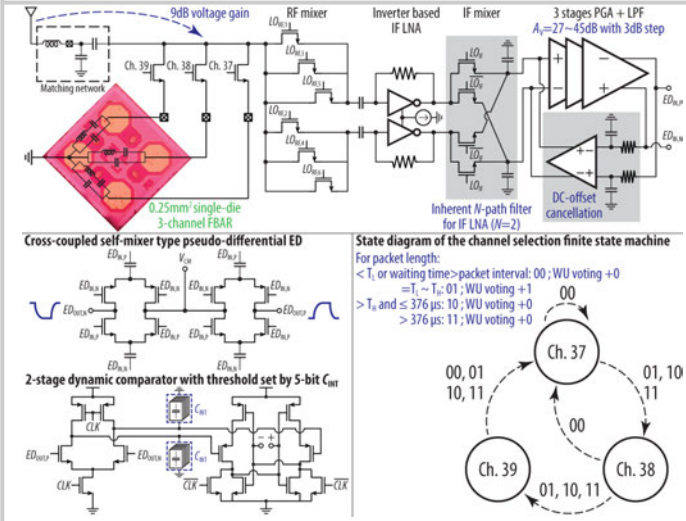


Figure 28.2.3: Schematic of proposed WuRX signal chain (top and bottom left); state diagram of the proposed majority-voting channel selection algorithm for 4D wake-up signature detection (bottom right).

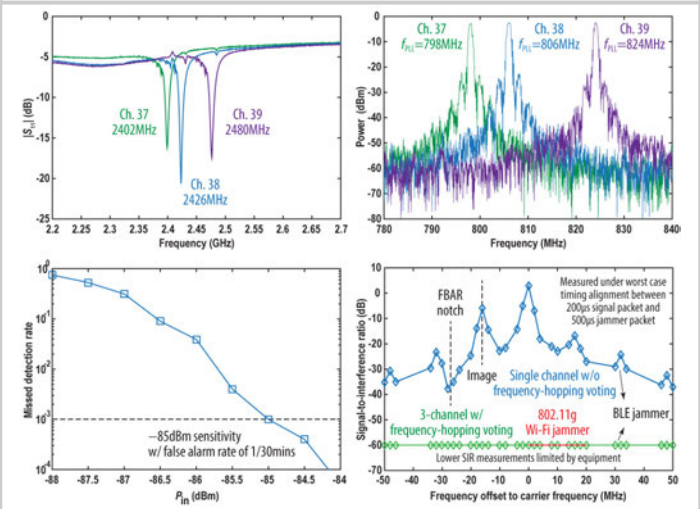


Figure 28.2.4: Measurement results of $|S_{11}|$, PLL locking frequency, missed detection rate waterfall curve, and signal-to-interference ratio using modulated BLE and 802.11g Wi-Fi jammers.

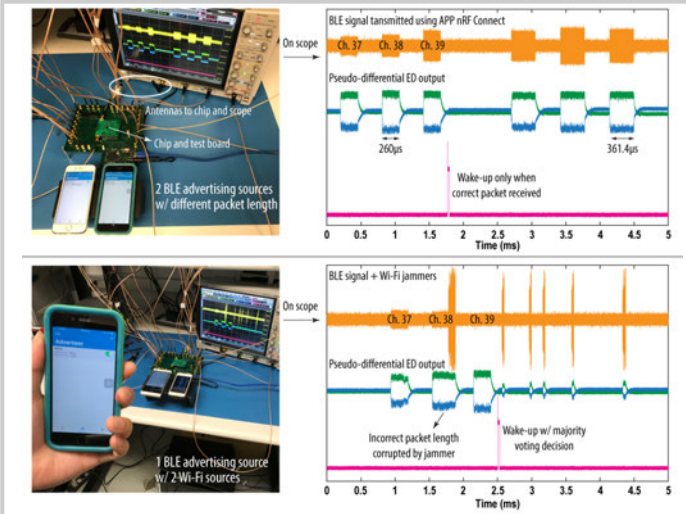


Figure 28.2.5: Demonstration of correct wake-up using mobile phones by sending: 1) two different BLE advertising packets with different packet lengths (top), and 2) BLE packets alongside two proximal Wi-Fi jammers (bottom).

Power breakdown

IF LNA	IF PGA	Comparator	PLL + frequency tripler (including LO buffers)	Digital baseband
22.4µW	22.2µW	2.7µW	166.2µW	6.1µW

	Roberts ISSCC'16	Ding RFIC'17	Abdelhamid CICC'18	Alghaihab RFIC'18	**This Work**
Technology	65nm	90nm	65nm	65nm	**65nm**
Die area	2.25mm²	1.24mm²	4mm²	1.1mm²	**2.4mm²**
Supply voltage	1 / 0.5V	2V	0.75V	1.1 / 0.9V	**0.5V**
WuRX decoding	Interval and duration of adv. events	Packet duration + RSSI	GFSK	3 channels detection + BC symbol correlation	**3 channels voting + packet length + packet interval + power threshold**
BLE compatible frequency hopping	No	No	No	Yes	**Yes**
Frequency locked demodulation	N/A[1]	N/A[1]	No	Yes	**Yes**
Integrated digital baseband	Yes	Yes	Yes	No	**Yes**
Adjacent channel SIR @ 2MHz @ 10MHz	N.R.[2]	N.R.[2]	N.R.[2]	-4dB -20dB	**-6dB[3] / -60dB[4] -23dB[3] / -60dB[4]**
RX power	236nW	164µW	230µW	150µW	**220µW**
Latency	~2s	~100µs	~200µs	1.47ms[5]	**~200µs[3] / 1.47ms[4, 5]**
Sensitivity	-56.5dBm	-58dBm	-80dBm	-57.5dBm	**-85dBm**
FoM[6]	89.3dB	105.9dB	123.4dB	94.1dB	**128.6dB[3] / 119.9dB[4]**

[1] Not applicable.
[2] Not reported.
[3] Single channel w/o frequency hopping mode enabled.
[4] 3 channels w/ frequency hopping mode enabled.
[5] 1 advertising event, use UUID=0x180D here for example.
[6] $FoM\ (\text{dB}) = -P_{BEN} + 10\log\frac{1}{Latency} - 10\log\frac{P_{dc}}{1\text{mW}}$

Figure 28.2.6: Power breakdown of the proposed WuRX (top), and comparison to state of the art (bottom).

28

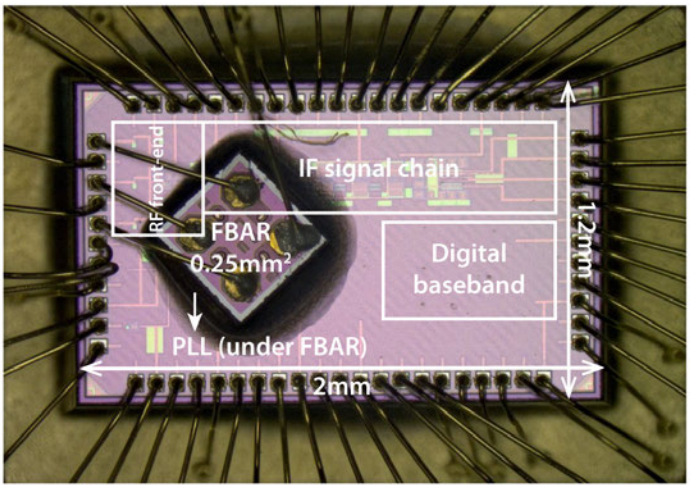


Figure 28.2.7: Die micrograph of the WuRX employing a single-die 3-channel FBAR stacked on top of the WuRX die.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

28.3 A 606μW mm-Scale Bluetooth Low-Energy Transmitter Using Co-Designed 3.5×3.5mm² Loop Antenna and Transformer-Boost Power Oscillator

Yao Shi, Xing Chen, Hun-Seok Kim, David Blaauw, David Wentzloff

University of Michigan, Ann Arbor, MI

Wireless communication has been a limiting factor for achieving millimeter-sized wireless sensor nodes because of the high power consumption, large antenna size and off-chip components typically required. Several mm-scale radios have been proposed [1-3]; however, all use proprietary communication protocols, which afford their designers more flexibility to address the above challenges. For interoperability and ubiquitous adoption, it is important that mm-scale radios use standard protocols, such as Bluetooth Low Energy (BLE). However, implementing a BLE-compliant radio in a mm-scale form factor poses significant additional challenges. These include the requirements for center frequency deviation (<150kHz), FSK modulation accuracy and frequency drift during the packet (<50kHz), which drive up the power consumption of typical PLL+VCO+PA BLE transmitters to >3mW [4-5]. A 0.5mW BLE transmitter was achieved by relaxing these requirements, but suffered from high phase noise due to its low-power ring oscillator [6].

In this work, we propose several techniques to realize a mm-scale BLE transmitter and antenna consuming only 606μW. To achieve this, we employ a front-end based on a power oscillator with a high-Q resonator formed by a printed 3.5×3.5mm² inductive loop antenna (sim. Q~110) and an on-chip digitally switched capacitor array (sim. Q~283). The power oscillator (Fig. 28.3.1) replaces the traditional VCO plus PA, achieving lower power consumption while maintaining phase noise of -118.5dBc/Hz at 1MHz offset that results in low FSK modulation error (2.1%) and low frequency drift during the BLE packet (<10kHz). We propose a transformer-boost technique for the power oscillator that improves output power by 1.2× and power efficiency by 3.16× compared to a conventional Class-B architecture, resulting in an overall transmit efficiency of 23.6%. The center frequency is digitally controlled using an on-chip digitally switched capacitor array resonating with the inductance of the loop antenna, and achieves a 32.2kHz frequency resolution required for BLE standard compliance. The loop antenna is 3.5×3.5mm², which gives an optimal trade-off between antenna gain and the required on-chip capacitance to resonate the antenna. Finally, a compact ADPLL realizes fast center frequency settling (<15μs), operates only during frequency settling and is powered off when transmitting the BLE packet to reduce overall energy consumption. The proposed BLE transmitter, including a CMOS chip and a 3.5×3.5mm² loop antenna, consumes 606μW when transmitting a 368μs advertising packet. It is incorporated in a 4.5×18.6mm² fully self-contained wireless circuit board powered by a mm-scale coin battery that we used to demonstrate BLE advertising to typical mobile phones with >8m distance.

The overall architecture of the proposed BLE transmitter is shown in Fig. 28.3.1. An electrically-small size (0.028λ) magnetic dipole antenna is chosen because it achieves better efficiency than an electric dipole when considering the Q of on-chip passives for resonating [1]. Thus, a 3.5×3.5 mm² loop antenna was implemented on a Rogers 4003C PCB with -15dBi omni-directional gain. Figure 28.3.2 shows the HFSS model and simulation results of the antenna including bonding wires, chip pads and silicon substrate. The simulated capacitance to resonate this antenna at 2.4GHz is ~680fF, which is sufficiently large to implement a digitally controlled capacitor array on-chip with practical tuning range and provides margin for parasitic capacitance.

The Q-factor of the digitally controlled capacitor array is critical to transmit efficiency, and hence the equivalent series resistance (ESR) of the resonant capacitance must be kept at a minimum. At the same time, the capacitance resolution and tuning range must meet the BLE standard. To co-optimize capacitor resolution and Q-factor, we use a coarse capacitor array with 1.02fF unit cap, achieving ~110MHz range to cover the BLE bandwidth. Then, a 6b fine capacitor array with 16.8aF tuning resolution achieves 32.2kHz resolution for BLE channel selection and another 6b modulator capacitor array has 22.6kHz resolution for GFSK modulation. Each unit capacitor consists of MOM capacitors (PDK available) and NMOS switches. LVT transistors with a large W/L ratio are used to reduce unit ESR. The top two metal layers are 2.5μm wide to reduce ESR overhead and are alternatingly used for adjacent parallel wires to minimize parasitic coupling capacitance.

A traditional Class-B power oscillator uses a tail current source, similar to a differential LC oscillator [1]. Figure 28.3.3 (bottom, right) shows that peak transmit efficiency of 28% is reached when the tail current is tuned to the cross-over between current-limited and voltage-limited regions. However, output power is low, only 22μW. To increase output power, tail current can be increased, but this causes efficiency to drop dramatically (13% @ 45μW). Entirely removing the current source results in maximum power of 75μW, limited only by supply voltage, but the efficiency is even lower at 8.8%. To achieve higher output power and efficiency we propose a transformer-boost technique (Fig. 28.3.3, left). A transformer replaces the tail current sources on the top and bottom of the cross-coupled pair, forming two resonant tanks with tunable capacitor arrays at 4.8GHz. The resonant tanks filter out frequency components of current other than 4.8GHz, which contribute to loss. Through the transformer's magnetic coupling, it also boosts the virtual supply (VTOP and VBOT), increasing oscillation swing by 1.3×, resulting in 90μW output power with 28% transmit efficiency. The transformer was implemented on-chip using a top metal layer achieving k=0.75 and Q=12. The measured output power is -8.4dBm with 27% efficiency due to the degraded Q-factor of the loop antenna.

The proposed transmitter utilizes open-loop GFSK modulation for energy saving. When the BLE packet is ready, the ADPLL locks the frequency of the digitally controlled power oscillator (DCPO) within 15μs. An on-the-fly lock detector monitors the DCPO control word and, upon lock, disables the ADPLL and starts the GFSK modulator of the packet with the DCPO operating in open-loop. The open-loop nature eliminates the need to implement a high-resolution TDC; instead we implemented an injection-locked TDC based on a 12-phase pseudo-differential ring oscillator to avoid DCPO period calibration. The GFSK modulator is implemented as an oversampled (8×) interpolative digital filter where the output frequency during the current bit is determined considering both preceding and following bits. The look-up table of the modulator array control word is pre-calibrated for open-loop modulation.

The proposed BLE transmitter was fabricated using 65nm CMOS. Figure 28.3.7 shows the chip micrograph with 0.49mm² active area. The chip is integrated in a 4.5×18.6mm² prototype wireless board with custom loop antenna (Fig. 28.3.4). When operating using a 5.8mm coin battery, the self-contained transmitter generates -23.4dBm EIRP, wirelessly measured using a horn antenna (LB-530-NF). The proposed transmitter chip consumes an average power of 606μW dominated by the power oscillator with packet-level duty cycling. The only necessary off-chip component is a 16MHz crystal oscillator whose power can be as low as 30μW [8].

Figure 28.3.4 shows that measured phase noise of the power oscillator is -118.5dBc/Hz at 1MHz offset consuming 534μW. Measured relative 2nd- and 3rd-harmonic power is -47dBc and -52dBc, respectively (both below BLE requirement of -41.5dBc). Figure 28.3.5 shows measured results of BLE modulation. The BLE spectral mask is met and FSK error is only 2.1%. The measured transient response of the DCPO frequency shows that the frequency drift is below 10kHz during a 368μs BLE packet, which is within the BLE specification (<50kHz). Figure 28.3.6 shows a comparison with recent low-power transmitters.

References:

[1] L.-X. Chuo et al., "A 915MHz Asymmetric Radio Using Q-Enhanced Amplifier for A Fully Integrated 3×3×3mm³ Wireless Sensor Node with 20m Non-Line-of-Sight Communication," *ISSCC*, pp. 132-133, Feb. 2017.

[2] M. Tabesh et al., "A Power-Harvesting Pad-Less Millimeter-Sized Radio," *IEEE JSSC*, vol. 50, pp. 962-977, April 2015.

[3] A, Yakovlev et al., "An 11μW Sub-pJ/bit Reconfigurable Transceiver for mm-Sized Wireless Implants," *IEEE TBioCAS*, vol. 10, no. 1, pp. 175-185, Feb. 2016.

[4] M. Babaie et al., "A Fully Integrated Bluetooth Low-Energy Transmitter in 28 nm CMOS with 36% System Efficiency at 3 dBm," *IEEE JSSC*, vol. 51, no. 7, pp. 1547-1565, July 2016.

[5] Y.-H. Liu et al., "A 3.7mW-RX 4.4mW-TX Fully Integrated Bluetooth Low-Energy/IEEE802.15.4/Proprietary SoC with an ADPLL-Based Fast Frequency-Offset Compensation in 40nm CMOS," *ISSCC*, pp. 236-237, Feb. 2015.

[6] X. Chen et al., "A 486 μW All-Digital Bluetooth Low Energy Transmitter with Ring Oscillator Based ADPLL for IoT Applications," *IEEE RFIC*, pp. 168-171, June 2018.

[7] J.Pandey et al., "A Sub-100 μW MICS/ISM Band Transmitter Based on Injection-Locking and Frequency Multiplication," *IEEE JSSC*, vol. 46, no. 5, pp. 1049-1058, May 2011.

[8] K.-M. Lei et al., "A Regulation-Free Sub-0.5 V 16/24MHz Crystal Oscillator for Energy-Harvesting BLE Radios with 14.2 nJ Startup Energy and 31.8 μW Steady-State Power," *ISSCC*, pp. 52-53, Feb. 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

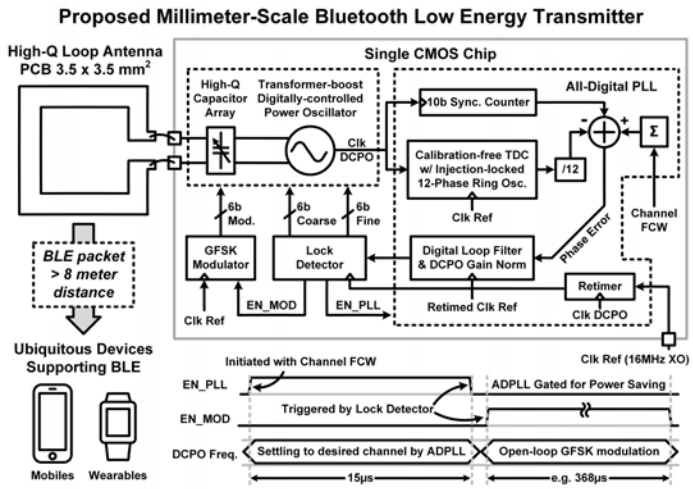


Figure 28.3.1: Architecture and timing of proposed millimeter-scale BLE transmitter consisting of 3.5×3.5mm² high-Q printed loop antenna and single CMOS chip.

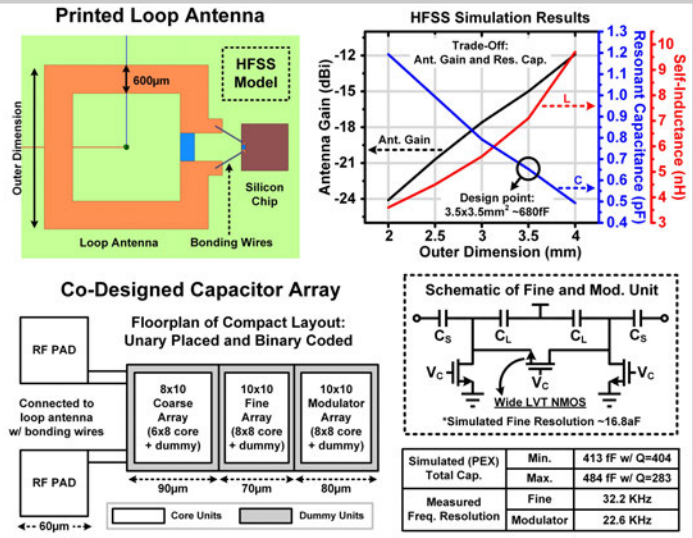


Figure 28.3.2: Co-design of loop antenna and capacitor array for high-Q resonator and BLE compliance.

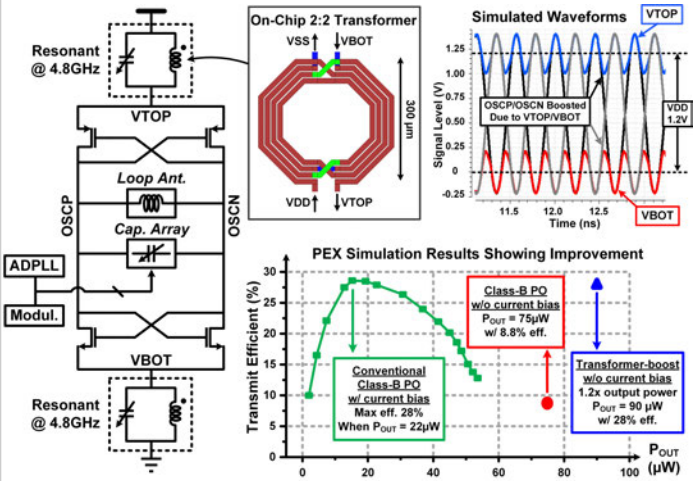


Figure 28.3.3: Proposed transformer-boost power oscillator for boosted output power with enhanced transmit efficiency.

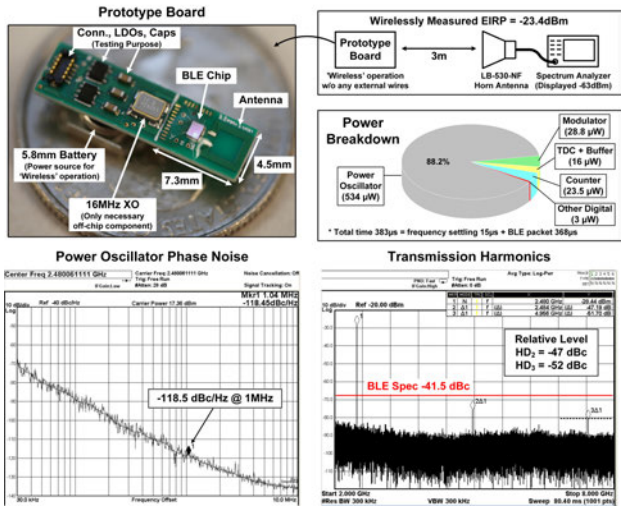


Figure 28.3.4: EIRP measured using the prototype wireless board. Transmitter power breakdown with packet-level duty cycling. Measured phase noise and transmission harmonics of the digitally controlled power oscillator.

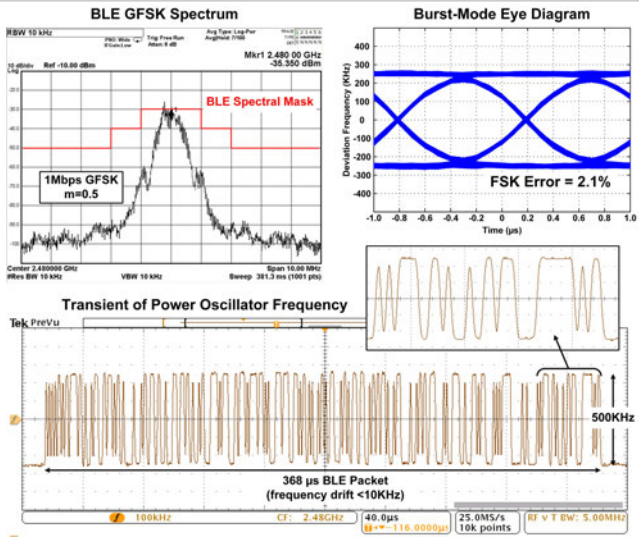


Figure 28.3.5: Measured performance of BLE GFSK modulation: continuous-mode spectrum, eye diagram, and transient.

		This Work	ISSCC 2017 [1]	JSSC 2015 [2]	JSSC 2016 [4]	ISSCC 2015 [5]	RFIC 2018 [6]	JSSC 2011 [7]
Technology		65 nm	180 nm	65 nm	28 nm	40 nm	40 nm	130 nm
Co-Designed Antenna		3.5x3.5 mm² Planar Loop	3x3x2 mm3 3D Loop	2.5 mm On-Chip Dipole	N/A	N/A	N/A	N/A
Carrier Frequency		2.4 GHz	915 MHz	60 GHz	2.4 GHz	2.4 GHz	2.4 GHz	400MHz
Standard		BLE	Propritary	Propritary	BLE	BLE	BLE	MICS
Modulation		GFSK	PPM	PPM	GFSK	GFSK	GFSK	BFSK
Data Rate		1 Mbps	30.3kbps	1.2 Mbps	1 Mbps	1 Mbps	1 Mbps	200 kbps
Supply Voltage		1.2 V	4 V	0.9 V	0.5/1 V	1 V	0.6 V	N/A
Power Consumption		0.61 mW [1]	2 mW	11 mW [4]	3.6 mW	4.2 mW	0.49 mW	0.09 mW
Energy Per Bit		0.61 nJ/b	66 nJ/b	9.1 nJ/b	3.6 nJ/b	4.2 nJ/b	0.49 nJ/b	0.45 nJ/b
TX Output Power		-8.4 dBm [2]	-2 dBm	-3 dBm	0 dBm	-2 dBm	-19 dBm	-17 dBm
Overall Efficiency		23.6%	32.4% [3]	4.5%	28%	15%	2.6%	22%
Modulation Error		2.1%	N/A	N/A	2.7%	4.8%	9%	N/A
Osc. Phase Noise @ 1MHz (dBc/Hz)		-118.5	N/A	N/A	-116	-110	-85	-104 [5]
Spur Level	HD2 (dBc)	-47	N/A	N/A	-50	-47	-42	N/A
	HD3 (dBc)	-52	N/A	N/A	-47	-51	-45	N/A

[1] Average power for transmitting one BLE packet w/ open-loop modulation. Power of external commercial XO is not included.
[2] Calculated based on wireless testing results: EIRP -23.4dBm and simulated antenna gain -15dBi (omni-directional).
[3] Power oscillator only and high-Q off-chip capacitor used.
[4] Calculated based on simulated results: 250pC total charge for 3 pulses (width 680ps) from 0.9V supply.
[5] Measured phase noise of frequency multiplier output.

Figure 28.3.6: Comparison with recent low-power transmitters.

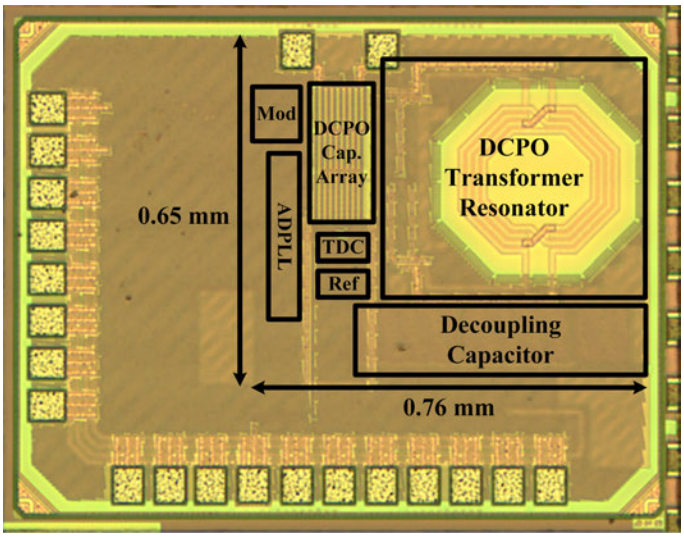


Figure 28.3.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

28.4 A High-Q Resonant Inductive Link Transmit Modulator/Driver for Enhanced Power and FSK/PSK Data Transfer Using Adaptive-Predictive Phase-Continuous Switching Fractional-Capacitance Tuning

Henry Kennedy, Rares Bodnar, Teerasak Lee, William Redman-White

University of Southampton, Southampton, United Kingdom

As well as transferring power, inductively coupled systems such as RFID and wireless charging commonly require a downlink channel to transfer data to the receiving function, for simplicity usually using the same carrier frequency used for the power transfer. A high-Q resonant transmitter coil is highly desirable to create the strong magnetic field required for a practical operating range. However, this not only raises major problems with sensitivity to tolerances and environmental factors, but also seriously restricts the available bandwidth and hence downlink data-rate. Amplitude Shift or On-Off Keying (ASK/OOK) are commonly used to allow simple demodulation, but in addition to the Q factor restricting the data-rate, the average power transfer will be reduced by around 50%. Frequency Shift Keying (FSK) or Phase Shift Keying (PSK) are attractive inasmuch as the nominally constant envelope provides a potentially higher power throughput, but the data-rate issue with a high-Q transmitter still remains. This is obvious for FSK, where by definition operation cannot be maintained away from the transmitter antenna's resonance frequency. Less obviously, for PSK applied to a nominally constant frequency carrier, the stored energy in the transmit tuned circuit will slow the phase transitions making demodulation more difficult; for binary PSK the amplitude will also drop significantly at each symbol transition. Note that the receiver Q factor is usually lower to avoid the need for active tuning in a micropower circuit.

Thus in order to derive benefits from FSK or PSK it is essential to solve the problem of maintaining resonance during frequency or phase deviations. In prior art [1] an array of selectable capacitors has been employed to achieve resonance, with the elements also switched synchronously with the modulation, but the detection of resonance in real time is still a bottleneck, and does not solve the problem for PSK modulation. Furthermore, a large number of capacitors, HV switches and control lines are required for precise tuning to resonance at high-Q to compensate for detuning effects and component tolerances. Adaptive real-time tuning of a driven resonant circuit can be achieved by synchronously switching a capacitor in and out of the circuit within each cycle, where the duty cycle determines the average capacitance [2]. Resonance is maintained with external magnetic loading and across PVT variations by sensing the voltage across the switch each cycle at the closure instant, and determining the direction of error from resonance to close a control loop (Fig. 28.4.1) similar to a conventional PWM system. In an analogue implementation, the excitation source also generates a triangle signal that is compared with tuning voltages $V_{TN+/-}$ to determine the switch timing.

With a single tuning feedback loop, smooth variations in the excitation frequency can be tracked at the cost of a small residual error increasing with deviation rate, but abrupt frequency transitions still result in a finite settling time. However, if the excitation is always switched between two defined FSK symbol frequencies, it is possible to use two separate tuning feedback loops, and select which one has control using the modulation data input. At either of the FSK symbol frequencies the active tuning loop can be updated with error information to compensate for parametric drift, while the inactive loop is held in its last active state pending the next frequency transition. Thus resonance is maintained at all times. To preserve the continuity in the current circulating in the transmit resonant circuit, it is important that the excitation phase is continuous and the changeover between the tuning loops is synchronised precisely with the zero crossing of the antenna voltage waveform V_C. Sensing this condition would be cumbersome, but we can easily infer the correct timing from the phase of the excitation drive, on the basis that at resonance it will be known precisely. Figure 28.4.2 shows the concept using similar hardware to Fig. 28.4.1. At the instant that the frequency is switched from F_L to F_H the tuning control voltages $V_{TR+/-}$ can be seen to jump to new stored values, setting the capacitor switch timing to the previous settled values at resonance. This concept may be extended with more than two control loops to allow multiple defined discrete frequencies to be used.

This continuously resonant FSK arrangement can be adapted to handle PSK [3]. A positive or negative phase rotation may be implemented by a temporary increase or decrease in the excitation frequency. A precise phase rotation can thus be achieved by using a defined offset frequency for a defined number of cycles, and hence an adaptive tuning system with at least three control loops can maintain continuous resonance. Two frequencies F_H and F_L need to be set so that they create the required net phase angle θ_{LEAD} or θ_{LAG} relative to the nominal frequency F_C over the allocated number of cycles N for the phase shift. Equations 1 and 2 give the required F_L for phase lag and F_H for phase lead respectively, in terms of F_C, θ in radians and N.

$$F_L = \frac{2\pi \cdot F_C}{2\pi + \frac{\theta}{N}} \quad (1) \qquad F_H = \frac{2\pi \cdot F_C}{2\pi - \frac{\theta}{N}} \quad (2)$$

In practice, the wide frequency shifts possible with 3 adaptive tuning loops mean QPSK modulation can be achieved with as few as 2 cycles at each offset frequency. In order for resonance control to function and phase to be continuous it is important to control not only the timing of the frequency transitions but also the precise value of the offset frequencies. Implementation of the system in a fast process using digital synthesis would allow precise setting of both frequency and exact phase transition timing. However in the "smart power" process used, manual setting is employed for F_H and F_L, although adaptive frequency trimming to compensate for PVT variation is also quite practical.

Figure 28.4.3 shows the architecture for a combined driver, tuner and modulator integrated using a mixed-signal architecture in a 0.18µm 1.8V/5V/50V CMOS/LDMOS process, operating over 10kHz to 2MHz and intended to drive up to 100mA in a resonant transmit coil, with a die area of 4.8mm^2 (limits set by project time/budget constraints). The fixed and fractional capacitors C_1 and C_2 (Fig. 28.4.1) are realised as equal capacitors C_N and C_P, alternately switched by separate P and N HV devices on the positive and negative excursions of V_C, avoiding the need for an ideal bi-directional switch. This gives a min: max tuning range of 1:1.41, sufficient to permit both compensation of tolerance/environmental effects as well as the required frequency shifting for FSK/PSK data modulation. Capacitive dividers isolate the high voltages on the tuning capacitors from the tuning control input [2]. The timing reference is derived from a current integrator oscillator, creating V_{TR} and V_{DR}, the currents generated by a digital V-to-I function. In PSK mode, an FSM manages transitions between the PSK frequencies to create phase lead/lag as appropriate. The FSM is synchronous with the mid-point crossings of V_{TR}, denoted by V_{TRLIM}, providing ideally synchronous selection of operating frequency and active tuning loops at resonance.

Figure 28.4.4 compares FSK operation with a fixed-tuned inductor and the adaptive-predictive system, using a 30µH inductor at 2MHz with a Q of 50. At the FSK *mark* frequency the fixed-tuned antenna circuit is no longer resonating and hence there is a significant reduction in the circulating current. By contrast, the adaptive-predictive system synchronously changes both the drive and antenna resonant frequencies such that there is a negligible reduction in V_C, hence the magnetic field strength around the inductor is maintained. QPSK operation of the adaptive-predictive system is shown in Fig. 28.4.5 with phase lead and lag at 2MHz on the same antenna circuit. In both cases the transition occurs over 4 cycles due to an appropriate frequency shift (approx. ±6.5% adjustment of nominal frequency), maintaining a large amplitude of oscillation during this time. Figure 28.4.6 shows a comparison between the achievable data-rates of the adaptive-predictive system and OOK/FSK/PSK based on simulations of a fixed-tuned antenna circuit. With a Q of 50 the presented architecture provides an improvement of over 8× (PSK) and 30× (FSK) in the data-rate and induced EMF improvement of up to 4.5× at maximum data-rate compared with a fixed-tuned antenna. Measured FSK and QPSK modulation waveforms are shown at the maximum data-rates.

References:

[1] R. Hash and D. Bowman, "Data communication system harnessing frequency shift keyed magnetic field", *US Patent* US6349116B1, Feb. 2002.

[2] H. Kennedy et al., "A Self-Tuning Resonant Inductive Link Transmit Driver Using Quadrature-Symmetric Phase-Switched Fractional Capacitance", *ISSCC*, Feb. 2017.

[3] W. Redman-White, "Electronic Tuning System", *UK Patent* GB2539113B, Oct. 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

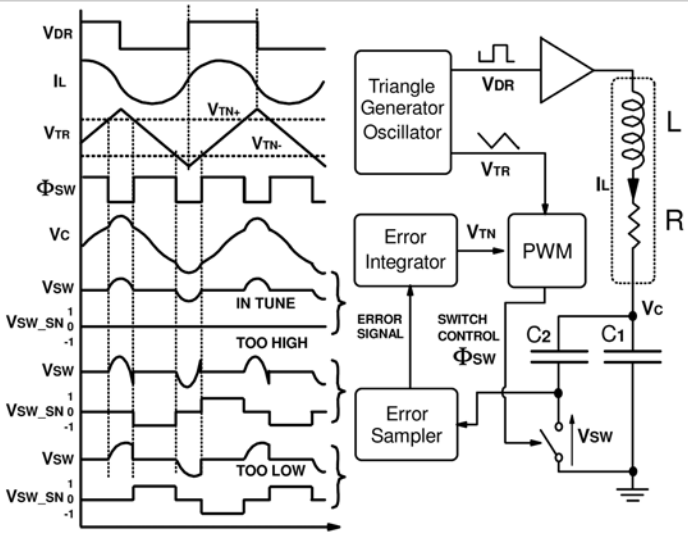


Figure 28.4.1: Left: Timings for adaptive tuning LC circuit using quadrature phase-switched fractional capacitance. Right: tuning system block diagram.

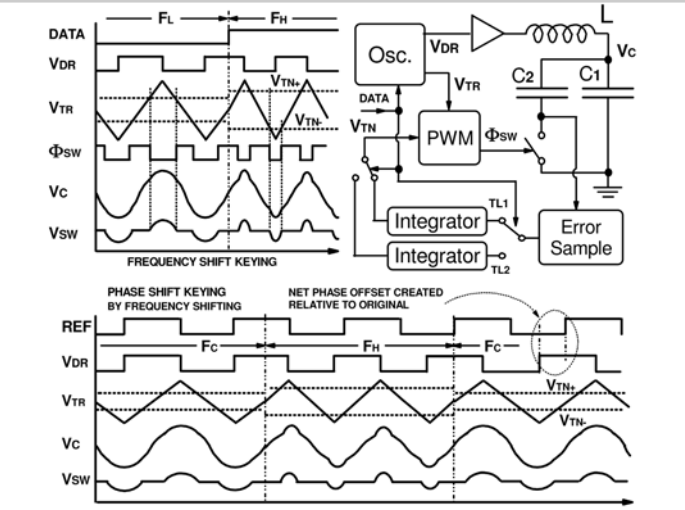


Figure 28.4.2: Top left: FSK with instantaneous frequency and tuning step. Top right: system architecture for adaptive-predictive tuning during FSK. Bottom: PSK phase advance by short-duration frequency offset.

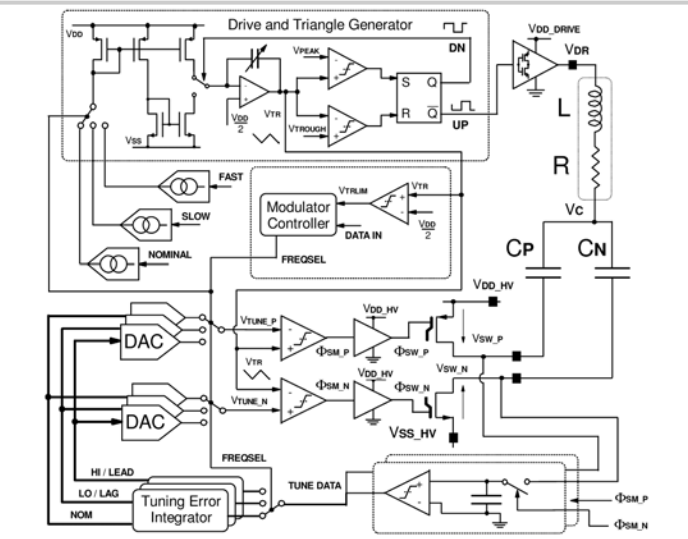


Figure 28.4.3: Circuit architecture of adaptive-predictive self-tuning system for antenna driver/modulator.

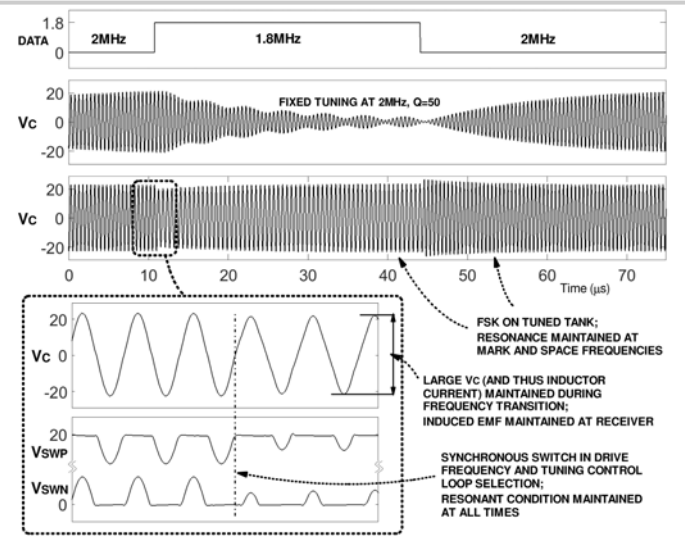


Figure 28.4.4: Wideband FSK measurement, F0=2MHz, F1=1.8MHz, Q = 50. Fixed-tuning circuit, (top, upper), adaptive-predictive tuning, (top, lower).

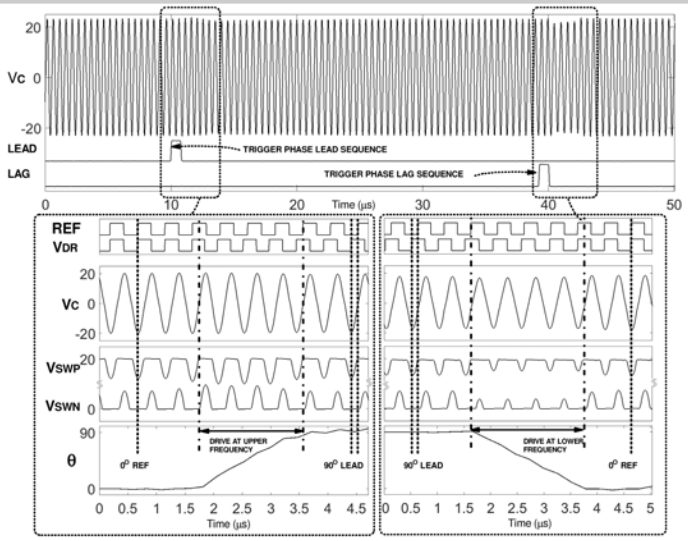


Figure 28.4.5: QPSK modulation measurement, FC= 2MHz, FH = 2.13MHz, FL = 1.88MHz, antenna circuit Q = 50. Phase lead (left) and phase lag (right).

For Q = 50 2MHz carrier frequency	This work		Fixed-tuning Inductor		
	FSK[1]	O-QPSK	OOK	FSK[2]	O-QPSK
Symbol transition time (µs) to within 1% of settled value	0.5	0.875	37.5	66.0	27.5
Maximum data rate (kbps)	500	296	26.6	15.1	36.4
Normalised induced EMF at maximum data rate	1	1	0.607	0.216	0.862

1 : F0 = 1.8MHz, F1 = 2MHz, adaptive-predictive tuning
2 : F0 = 1.8MHz, F1 = 2MHz, fixed tuning at 1.9MHz mean frequency

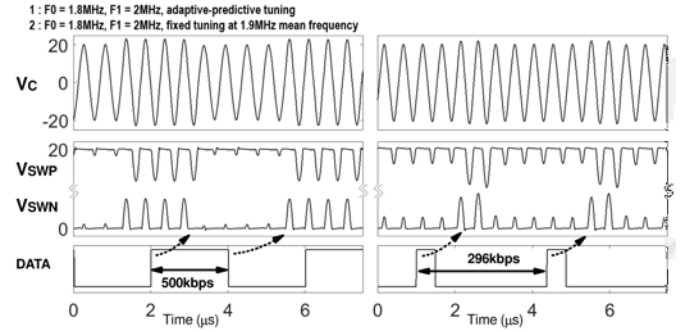


Figure 28.4.6: Top: Achievable data rate and power delivery for adaptive-predictive tuning vs fixed-tuning. Bottom left: Measured FSK modulation (max data rate). Bottom right: Measured PSK modulation (max data rate).

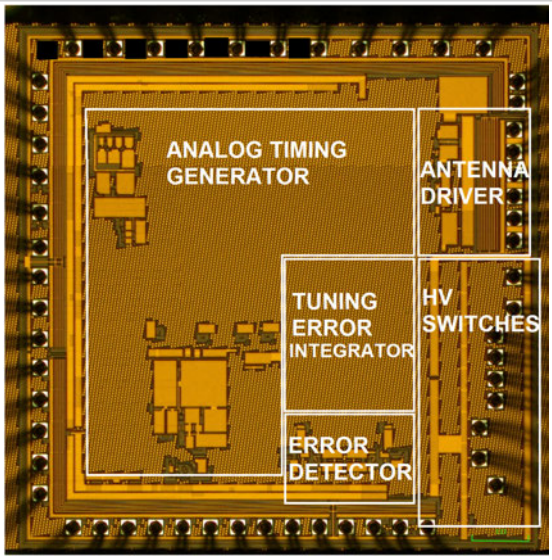


Figure 28.4.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

28.5 Non-Magnetic 60GHz SOI CMOS Circulator Based on Loss/Dispersion-Engineered Switched Bandpass Filters

Aravind Nagulu, Harish Krishnaswamy

Columbia University, New York, NY

There has been significant recent research on non-magnetic non-reciprocal components at RF and mm-waves, such as circulators and isolators, based on spatio-temporal modulation [1-3]. Circulators enable simultaneous transmit and receive (STAR) on a shared antenna for applications such as full-duplex wireless communication and FMCW radar. While there has been exciting initial progress, existing architectures do not scale well to mm-waves, whether they are based on switch-based conductivity modulation [1,2] or varactor-based permittivity modulation [3]. Loss levels increase due to losses in the switches or varactors, isolation is degraded due to reflections produced by parasitics, and power consumption is high due to the relatively high modulation frequencies required. In this work, we present a non-magnetic CMOS 60GHz circulator based on spatio-temporal conductivity modulation (STCM) across a loss/dispersion-engineered bandpass filter. This new architecture improves the insertion loss, isolation, power consumption, and spurious response compared to prior art. The 60GHz circulator achieves 3.6dB/3.1dB insertion loss for TX-to-ANT/ANT-to-RX paths, respectively, TX-to-RX isolation >40dB over 1.3GHz, 3.2dB of ANT-to-RX NF, >+19.5dBm TX-to-ANT/ANT-to-RX IP_{1dB}s and spurious tones lower than -30dBc at both ANT and RX ports at a power consumption of 41mW from 1.2V.

In [1], the concept of STCM across a broadband delay line was proposed, exploiting I/Q Gilbert quad switches on either side of a delay line that is quarter-wave at the modulation frequency (f_m). The quads on either side are clocked with 90° phase shift, resulting in a non-reciprocal 180° phase response (gyrator behavior) arising from the fact that forward-traveling signals pass through the delay line with no net sign inversion, while reverse-traveling signals pick up an extra sign inversion. This gyrator architecture enabled the implementation of a mm-wave (25GHz) circulator as the modulation frequency can be arbitrarily lowered, albeit at the expense of a longer delay line and hence, additional loss. However, there exists a fundamental limit to the loss of this architecture (Fig. 28.5.1), arising from the unfavorable trade-off between switch R_{ON} loss and loss due to dispersion in the delay line (Bragg effect). When viewed in the frequency domain, mixing products between the input frequency f_{in} and f_m ($f_{in}\pm f_m$, $f_{in}\pm 3f_m$, $f_{in}\pm 5f_m$…) are generated by the first Gilbert quad. These travel down the delay line to the second quad mixer, which mixes them back to the input frequency. In order to absorb the switch parasitic capacitance, the delay line must be implemented as a quasi-distributed LC ladder. Large switch devices are needed to minimize R_{ON}, but this results in large parasitic capacitance, and limited Bragg frequency in the LC ladder, which results in filtering of the mixing products and hence, signal loss.

In this work, we recognize that the mixing products are centered around the input frequency f_{in}, and therefore, a bandpass filter (BPF) with quarter-wave group delay is sufficient to preserve them. The bandpass filter exploited here includes shunt-LC resonators at the input and output of an 8-section LC ladder (Fig. 28.5.1). The shunt LC resonators allow for the absorption of larger switch capacitance, allowing for a lower R_{ON}, while still allowing enough bandwidth in the BPF to sustain the dominant mixing products, thus allowing low R_{ON} loss as well as low Bragg loss. Figure 28.5.1 depicts the fundamental loss level achievable across frequency by the architecture of [1] and the architecture of this work, including inductor loss, R_{ON} loss, and filtering loss. Assuming switch $R_{ON}C_{OFF}$=200fs (45nm SOI CMOS), inductor Q of 20 and maximum f_m of 10GHz, the losses of circulators using both switched-delay-line and switched-BPF gyrators were estimated and then optimized for best loss across frequency. The new architecture has an optimal R_{ON} which is relatively constant across frequency, thereby improving the loss of the circulator compared to [1]. Power handling, limited by R_{ON} linearity, is also improved. While the new architecture would enable a circulator with near-2dB loss at 28GHz, here, we chose to explore 60GHz, where it improves loss and power handling by 1.3dB and 9dB respectively.

LPTV circuits must be characterized for their spurious emissions, unlike LTI circuits, as they can violate spectral masks. This issue has not received attention in prior non-magnetic circulator work [1-3]. Filtering of the mixing products in the delay line not only introduces signal loss, but also introduces spurious mixing products at $f_{in}\pm 2f_m$, $f_{in}\pm 4f_m$ … at TX and ANT ports as the mixing products are not perfectly mixed back to the input frequency by the second quad mixer. Unabsorbed parasitic capacitances at the gyrator nodes lead to mismatch and also result in spurious tones at $f_{in}\pm f_m$, $f_{in}\pm 3f_m$ etc. In this new architecture, lower R_{ON} of the switches in the quads results in a mixing operation closer to ideal mixing, dispersion-engineering reduces Bragg filtering, and the shunt inductors absorb parasitic capacitances, all of which result in low spurious tones that are at least 30dBc below the TX signal across a wide range of TX powers.

For 60GHz operation (Fig. 28.5.2), we chose to modulate at 8.6GHz (f_{in}/f_m = 7) after optimizing the trade-off between losses due to gyrator passives and rise/fall time of the modulation clocks, and power consumption. The fully balanced Gilbert quads are designed using 2×16μm/40nm floating-body transistors to have an ON-resistance of 8.6Ω. The bandpass filter between the switches is implemented by cascading 8 π-type CLC sections with single-turn shunt inductors (Q=22) on either side, and is designed to support up to the 5th mixing product. Another pair of shunt inductors is placed on the other side of the Gilbert quads to absorb the parasitics of the switches. The LO path consists of fine-phase-tuning varactors, a 2-stage polyphase filter generating the differential I/Q signals, and self-biased pseudo-differential inverter chains. The gain of the final buffer is boosted using inductive peaking to generate the square clocks.

The 60GHz circulator was implemented in a 45nm SOI CMOS process and occupies an area of 1.72mm² (Fig. 28.5.7). The on-chip baluns are de-embedded through test structures on-chip. Measured insertion losses at 54GHz are 3.6dB and 3.1dB, for TX-ANT and ANT-RX paths respectively, when operated using a 1.2V supply at a clock power consumption of 41mW (Fig. 28.5.3). Their sum degrades by 2dB over a bandwidth of 6.8GHz, and TX-to-RX isolation is >20dB for this entire frequency range. Near the center frequency, the isolation is >40dB over a bandwidth of 1.3GHz. When operated at a lower supply voltage of 1V at a clock power consumption of 24.1mW, the insertion losses are 3.9dB and 3.1dB. Since the degradation in losses due to increased R_{ON} at 1V supply is low relative to the lower power consumption, we performed all subsequent measurements at 1V supply. The measured IP_{1dB}s of TX-to-ANT and ANT-to-RX paths are >+19.7dBm and >+19.1dBm respectively, and are limited by the measurement setup (Fig. 28.5.4). At these power levels, the paths compress by 0.77 and 0.89dB, respectively. TX-to-ANT and ANT-to-RX IIP3s are measured to be +19.4dBm and +19.0dBm respectively. TX-induced ANT-to-RX compression is also measured (Fig. 28.5.5), and a weak ANT signal does not yet compress by 1dB at TX power levels >+20dBm. Compression of TX-RX isolation is also measured, and isolation is maintained to be >20dB/15dB for TX power <+13dBm/17.5dBm. The measured odd-harmonic intermodulation products at ANT and RX ports are very low, as expected, due to enhanced absorption of parasitic capacitance. As shown in Fig. 28.5.6, the conversion to even harmonics has been greatly suppressed by the design of a filter supporting up to the 5th intermodulation product, and among them, the worst-case spur is >30dBc lower than the TX signal tone for a wide range of TX powers. This spur is spaced 17GHz (f_{in}-$2f_m$) away from the center frequency, and satisfies the WiGig spectral mask. The measured ANT-to-RX noise figure is 3.2dB, showing minimal degradation due to LO phase noise. Figure 28.5.6 summarizes the performance. This work is superior in all metrics to active mm-wave circulators [4,5]. When compared to a mm-wave hybrid-transformer circular-polarization duplexer [6], this work achieves similar loss and isolation while not requiring two antennas or orthogonal polarizations. When compared to [1], this work operates at lower power consumption, achieves better TX-to-RX isolation, and comparable losses while operating at double the frequency. When compared with mm-wave circulators and isolators based on varactor modulation [3], it achieves far superior loss and power handling. As shown in [1], using an external ANT tuner for maintaining TX-to-RX isolation across ANT VSWR would limit the isolation bandwidth. Hence a tightly integrated ANT tuner is necessary. Integrating this circulator with on-chip antenna balancing similar to [2] is a topic of future research.

Acknowledgements:
We acknowledge Texas Instruments for financial support and Global Foundries for fabrication donation.

References:

[1] T. Dinc and H. Krishnaswamy, "A 28GHz Magnetic-Free Non-Reciprocal Passive CMOS Circulator Based on Spatio-Temporal Conductance Modulation," *ISSCC*, pp. 294-295, Feb. 2017.

[2] A. Nagulu et al., "Fully-Integrated Non-Magnetic 180nm SOI Circulator with > 1W P1dB, >+50dBm IIP3 and High Isolation Across 1.85 VSWR," *IEEE RFIC,* pp. 104-107, June 2018.

[3] C. Yang and P. Gui, "85–110 GHz CMOS Tunable Nonreciprocal Transmission Line with 45 dB Isolation for Wideband Transceivers," *IEEE RFIC*, pp. 284-287, June 2017.

[4] H.-S. Wu et al., "CMOS Active Quasi-Circulator with Dual Transmission Gains Incorporating Feedforward Technique at *K*-Band," *IEEE TMTT*, vol. 58, no. 8, pp. 2084-2091, Aug. 2010.

[5] J.-F. Chang et al., "Design and analysis of 24-GHz active isolator and quasi-circulator," *IEEE TMTT,* vol. 63, no. 8, pp. 2638-2649, Aug. 2015.

[6] C. Lu et al., "A Millimeter-Wave Tunable Hybrid-Transformer-Based Circular Polarization Duplexer with Sequentially-Rotated Antennas," *IEEE TMTT*, vol. 64, no. 1, pp. 166-177, Jan. 2016.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

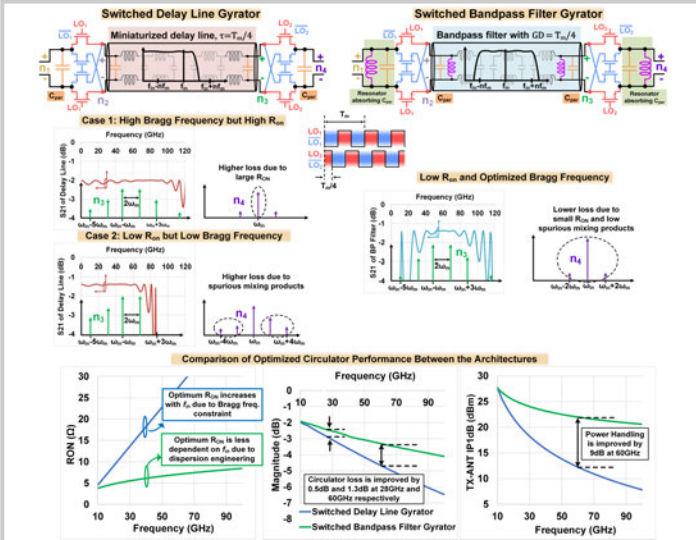


Figure 28.5.1: Millimeter-wave gyrator and circulator based on switched bandpass filters - concept and fundamental performance limits.

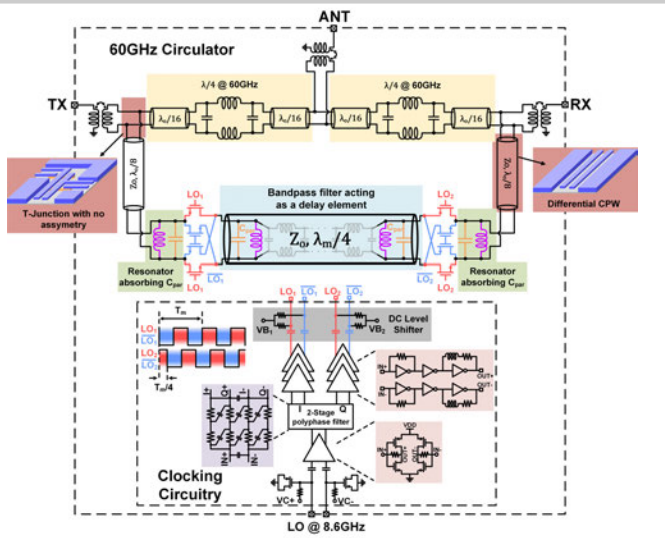


Figure 28.5.2: 60GHz 45nm SOI CMOS non-magnetic circulator block and circuit diagram.

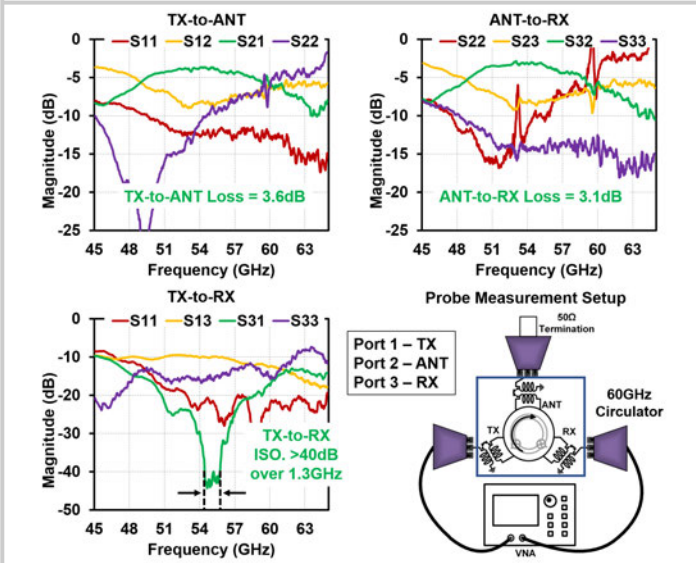


Figure 28.5.3: Measured S-parameters at 1.2V supply.

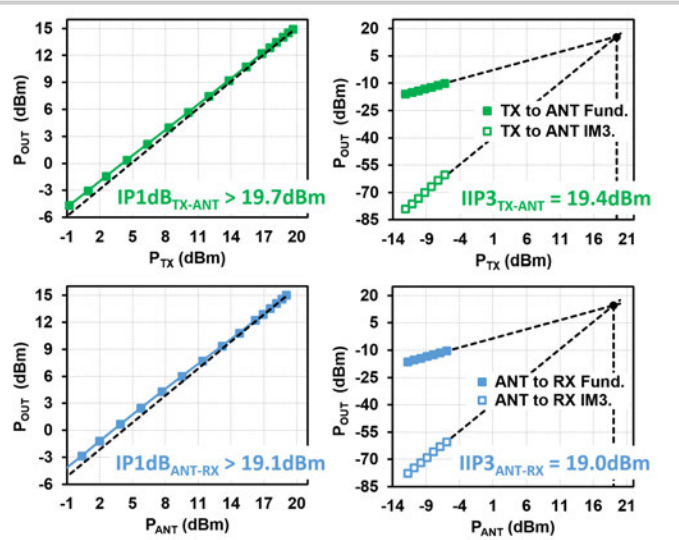


Figure 28.5.4: Measured IIP3s and 1dB compression points of TX-ANT and ANT-RX paths at 1V supply.

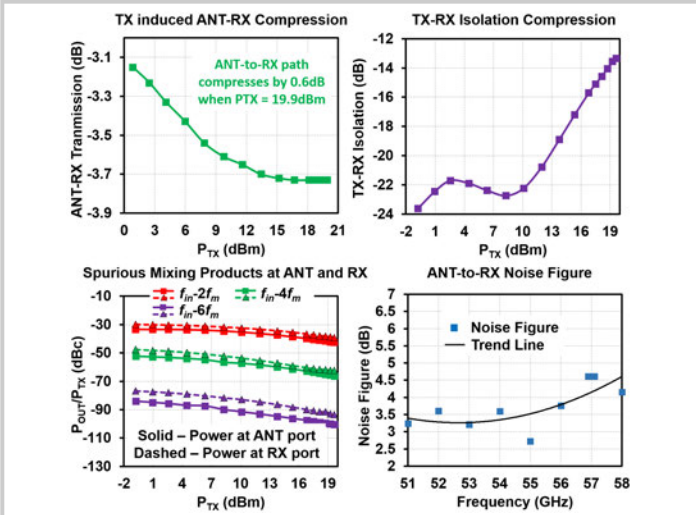


Figure 28.5.5: Measured TX-induced ANT-RX and TX-RX isolation compression, measured spurious mixing products at ANT and RX ports across TX power, and measured NF at 1V supply.

	TMTT 2010 [4]	TMTT 2015 [5]	TMTT 2016 [6]	ISSCC 2017 [1]	RFIC 2017 [3], WMCAS 2018[6]		This work
Technique	Active Circulator	Active Circulator	Hybrid-Transformer-Based Circular Polarization Duplexer	Spatio-Temporal Conductivity Mod. across Delay Line	Parametric Non-reciprocal Transmission Line	Circulator based on parametric Non-reciprocal TL [1]	Spatio-Temporal Conductivity Modulation across a Bandpass Filter
Technology	180nm CMOS	180nm CMOS	0.25μm SiGe BiCMOS	45nm CMOS SOI	65nm CMOS	65nm CMOS	45nm CMOS SOI
Frequency Range (GHz)	24.17 – 24.55	24.17 – 24.55	27 - 34	22.7 – 27.3	85 – 110	85 – 105	50 – 56.8
No. of Antennas	1	1	2	1	N/A	1	1
Supply Voltage	1.8 V	1.2 V	N/R	1.2 V	N/A[2]	N/A[2]	1V / 1.2V (All meas. below are for 1V)
TX-ANT/ANT-RX Transmission	+22.4 dB / +12.3 dB	-5.7 dB / -5.7 dB	-3.1 dB / +18 dB	-3.3 dB / -3.2 dB	-6.5 dB	-9.5 dB / -9.5dB	-3.6 dB / -3.1 dB: VDD=1.2V -3.9 dB / -3.1 dB: VDD=1V
TX-RX Isolation BW	>15 dB (~1%)	>20 dB (~1.6%)	>40 dB (6.5%)	>18.5[2] dB (18%)	>45[2] dB (1.5%)	>45[2] dB (1.6%)	>20 dB (14.6%) >40 dB (2.5%)
Center Freq./Modulation Freq.	N/A	N/A	N/A	3	10	10	7
ANT-RX NF	17 dB	N/R	4 dB	3.3 dB - 4.4 dB	N/R	N/R	3.2 dB
TX-ANT/ANT-RX IP1dBs	-19.18 dBm	+9.5 dBm	>+20[1,3] dBm / N/R	>+21.5 dBm/ >+21dBm [3]	10.9[2] dBm	10.9 dBm / 10.9dBm	>+19.65 dBm/>+19.05dBm [3]
TX-ANT/ANT-RX IIP3s	-11 dBm	N/R	N/R	20.1 dBm / 19.9 dBm	N/R	N/R	+19.43 dBm / +19.03 dBm
TX-RX Compression[5] (20dB / 15dB Levels)	N/R	N/R	>20 dBm / >20dBm [1,5]	N/A / +21 dBm	N/R	N/R	+13 dBm / 17.5 dBm
TX-induced ANT-RX Compression	N/R	N/R	N/R	N/R	N/R	N/R	>+20[2] dBm
Spurious Mixing Products	N/R	N/R	N/R	N/R	N/R	N/R	-30dBc (worst case)
PDC	144.8 mW	7.2 mW	60 mW	78.4 mW	N/R	N/R	41.04 mW @ VDD=1.2V 24.14 mW @ VDD=1V
Area	3.22 mm²	0.715 mm²	0.0625 mm²	2.16 mm²	0.245 mm²	N/A	1.72 mm²

1. Simulation Results 2. Clock Path is not integrated 3. Limited by measurement setup 4. Tunable over the BW 5. Devices breakdown limit
6. P_{TX} at which isolation is compressed to 20 dB/15 dB. N/A: Not Applicable N/R: Not Reported

WMCAS 2018: Chang Yang, Ping Gui, "On-chip 100GHz full-duplex circulator/duplexer design based on nonreciprocal transmission line", *Wireless and Microwave Circuits and Systems (WMCS) 2018 Texas Symposium on*, pp. 1-4, 2018.

Figure 28.5.6: Performance summary and comparison table.

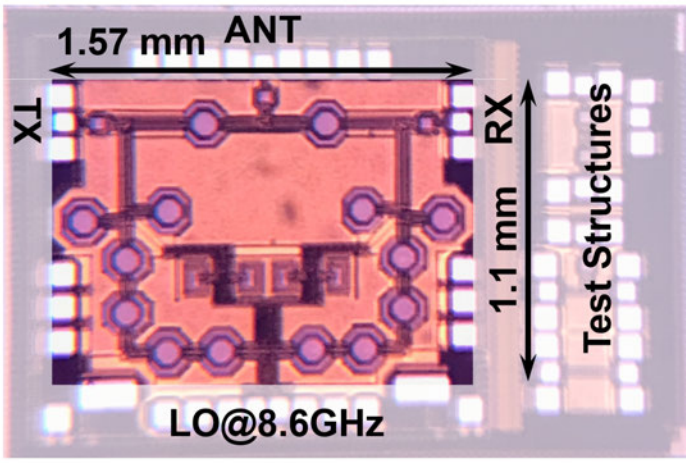


Figure 28.5.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

28.6 Full-Duplex 2×2 MIMO Circulator-Receiver with High TX Power Handling Exploiting MIMO RF and Shared-Delay Baseband Self-Interference Cancellation

Mahmood Baraani Dastjerdi[1], Sanket Jain[2], Negar Reiskarimian[1], Arun Natarajan[2], Harish Krishnaswamy[1]

[1]Columbia University, New York, NY; [2]Oregon State University, Corvallis, OR

Full-duplex (FD) wireless has garnered a lot of attention in recent years due to multiple benefits such as increased spectral efficiency and improved network latency [1-5]. Several challenges remain in the quest for high-performance integrated FD radios. Transmitter (TX) power handling remains an open problem, particularly in FD radios that integrate a shared-antenna interface. Recent integrated shared-antenna interfaces that exhibit high power handling exploit SOI CMOS technologies and occupy substantial area [6]. Secondly, FD operation must be achieved across antenna VSWR variations and a changing EM environment. Finally, FD operation must be extended to multiple-antenna systems, as beamforming and MIMO are an integral part of emerging wireless networks. While the feasibility of combining FD with beamforming in an integrated radio has been recently demonstrated [4], combining FD with MIMO is fundamentally challenging because of cross-talk self-interference (CT-SI) between every pair of antennas, causing the canceller complexity to grow as $O(N^2)$. While there has been work at the system level to investigate these challenges [5], CMOS integrated implementations impose unique challenges and constraints. This work presents a 2×2 MIMO FD Receiver (RX) array with integrated N-path-filter-based non-magnetic circulators that (i) exploits bootstrapping in the circulator N-path filters to enhance TX power handling by 8dB, and (ii) features area- and power-efficient passive RF and active baseband (BB) wideband MIMO cancellation with shared delay elements to address the $O(N^2)$ cancellation challenge in the presence of a changing EM environment. We have demonstrated a 2×2 FD MIMO RX achieving (i) up to 35/45dB average self-interference-cancellation (SIC) across 40/20MHz BW, (ii) more than 42dB/53dB average cross-talk (CT)-SIC across 40/20MHz BW with <2.1dB degradation in RX NF, and (iii) and overall TX power handling of +14dBm enabled by clock bootstrapping.

MIMO RF cancellation is essential to knock down the SI and CT-SI early in the RX chain, and relax RX dynamic range requirements. Passive MIMO RF cancellation directly addresses the $O(N^2)$ power consumption challenge. In this work, we exploit programmable feed capacitor banks from the TX port of each circulator to either side of the N-path filter in the circulator-RX (C-RX) circuit introduced in [6] (Fig. 28.6.1). Since the switches on either side of the two-port N-path filter are driven by quadrature clocks, the capacitor feed banks add quadrature cancellation signals to the BB nodes. The magnitude of these cancellation signals can be controlled by varying the feed capacitance through a digitally-controlled capacitor bank. A differential circulator implementation also allows sign inversion to be incorporated in the differential feed capacitor banks, thus realizing a passive full vector-modulator RF SI canceller. Since vector modulation is achieved without area-hungry inductors, this approach addresses the $O(N^2)$ area challenge as well. As compared to the approach in [5], the proposed scheme (i) reduces the loading on any one side by feeding both sides of the N-path filter and (ii) does not load or impact the antenna interface, which is critical to simultaneously achieve high SIC and CT-SIC. Importantly, in this work, the same concept is extended to MIMO CT-SI cancellation, as feed capacitor banks are incorporated between each TX and RX pair enabling passive CT-SI cancellation at RF, thereby reducing cancellation requirements from the $O(N^2)$ BB cancellers, which is critical to reducing noise figure degradation due to the cancellers (Fig. 28.6.1).

Two other innovations are introduced in the proposed MIMO C-RX. Prior C-RX circuits exploited off-chip inductors in the 3λ/4 transmission line that was wrapped around the N-path-filter gyrator. In this work, the C-RX operates at 2.2GHz and exploits wirebond inductance for the same purpose (Fig. 28.6.2), enabling a realization with no off-chip components. Secondly, in prior C-RX circuits, the TX power handling is limited by the voltage handling of the N-path filter switches. In this work, bootstrapping akin to [7] is exploited to achieve substantial improvement in TX power handling (Fig. 28.6.1).

To get further SI and CT-SI cancellation, a BB MIMO delay-based FIR filtering canceller is employed (Fig. 28.6.1). The canceller taps at the output of each TX, and consists of IQ TX downmixers, cascaded active g_m-C delay cells acting on I and Q paths, and programmable g_m cells on each delay tap acting on I and Q, effectively realizing vector modulation for programmable amplitude and phase on each delay tap. An analysis of such an active canceller reveals that the dominant power consumption is in the delay cells, and therefore, a single delay chain is used for each TX, with g_m cells injecting into each RX. Therefore, the delay cell complexity only grows as $O(N)$ while g_m cell complexity grows as $O(N^2)$.

The proposed FD MIMO approach is extended to a scalable 65nm CMOS 2.2GHz 2×2-element C-RX MIMO array (Figs. 28.6.2 and 28.6.7). The array uses the N-path-filter-based combined-circulator-RX concept described in [1,3]. The transmission lines are implemented using lumped CLC sections that employ on-chip MIM capacitors and wirebond inductance (Fig. 28.6.2). Each of the feed capacitors for the vector-modulator RF SI cancellers is implemented using 6 bits of control (~0 to 550fF). Two differential capacitor banks (6b) are added to the TX and ANT ports to compensate for package parasitics, and to further improve the antenna VSWR coverage. As mentioned earlier, bootstrapping circuitry is employed to enhance power handling [7], and the BB voltages are level shifted and added to the gate control voltage only during the clock ON-period. Each TX signal is coupled through a small capacitor to a 4-path passive mixer to downconvert the TX signal for BB cancellation. Delay cells are implemented with 4 taps using a topology similar to [3] but with a complementary implementation to enhance the maximum voltage swing at the output, resulting in less noise contribution from the BB canceller and less power consumption in the g_m cells. The delay cells are shared between the SI and CT-SI cancellation paths, and each tap includes 3 bits of capacitance control and 2 bits of gain control to compensate for parasitic losses. The delay taps are designed to cumulatively provide as high as ~120ns of delay over 10MHz of bandwidth. The 8b programmable vector modulator g_m cells tap from the delay cells to inject current into both RX elements. Each element also includes two additional weighted BB I/Q outputs to be used when tiling two ICs on PCB for future 4×4 MIMO implementations. The cancellation signals from all two/four elements are summed up in the current domain and injected at the input of the BB TIA.

Measurements are summarized in Figs. 28.6.3 to 28.6.5. As illustrated in Fig. 28.6.7, the chip is packaged using an 80-pin QFN package. Based on our measurements, the wirebond inductance in the circulator is around 5nH, and our simulations show that 10% variation in wirebond inductance introduces only 0.15dB additional TX-ANT loss. The C-RXs exhibit 3.5dB TX-ANT loss and +34dBm TX-ANT IIP3 without the bootstrapping circuitry. When bootstrapping is engaged, the TX-ANT IIP3 is boosted to +44dBm. The nominal RX conversion gain is 30dB, and the RX NF is 9.5dB. The high NF is primarily a function of the high operating frequency for N-path-like circuits in 65nm CMOS as the cumulative parasitic capacitance on the RF side of the N-path filter introduces substantial loss due to harmonic conversion. The use of a scaled CMOS technology would substantially lower the NF. Exploiting the vector-modulator RF SI cancellers allows complete coverage of ANT VSWR<2 and partial coverage of ANT VSWR<3 (Fig. 28.6.3). Figure 28.6.3 also shows measurements of the BB delay cells, verifying the ability to generate up to 120ns of delay. Additionally, by jointly optimizing both the feed-capacitors and the shared-delay BB canceller, a total SIC of 35/45dB has been achieved across 40/20MHz BW with an NF penalty of less than 1.7dB. Under cancellation, without bootstrapping, TX-induced 1dB RX compression happens at +7dBm, and this is boosted to +14dBm when bootstrapping is engaged. Therefore, +14dBm represents the TX power handling of the system. In 2×2-element FD MIMO configuration, by using the RF passive CT-SI canceller, up to -18.5dB of cross-antenna coupling with arbitrary phase across the complex plane can be cancelled. By engaging both the RF and BB cancellers in conjunction with a slot-loop antenna pair, more than 42/53dB CT-SI suppression across 40/20MHz has been achieved. Thanks to the shared-delay architecture, only 4 delay cells per element are needed and consume 23mW (0.35mm^2), while the cancellation gm-cells per element only consume 2mW (0.375mm^2). The overall area of the IC excluding the pads is 2.15×2.6mm^2. The associated overall NF degradation is less than 2.1dB per element. When compared with prior art (Fig. 28.6.6), the key contribution is the 2×2 FD MIMO design, allowing higher TX power handling by incorporating clock bootstrapping and featuring power and area-efficient wideband SIC and CT-SIC.

References:

[1] N. Reiskarimian et al., "Highly-Linear Integrated Magnetic-Free Circulator-Receiver for Full-Duplex Wireless," *ISSCC*, pp. 316-317, Feb. 2017.

[2] K. Chu et al., "A Broadband and Deep-TX Self-Interference Cancellation Technique for Full-Duplex and Frequency-Domain-Duplex Transceiver Applications," *ISSCC*, pp. 170-172, Feb. 2018.

[3] T. Zhang et al., "Wideband Dual-Injection-Path Self-Interference Cancellation Architecture for Full-Duplex Transceivers," *IEEE JSSC*, vol. 53, no. 6, pp. 1563-1576, June 2018.

[4] M. B. Dastjerdi et al., "Full Duplex Circulator-Receiver Phased Array Employing Self-Interference Cancellation via Beamforming," *IEEE RFIC*, pp. 108-111, June 2018.

[5] D. Bharadia and S. Katti, "Full Duplex MIMO Radios," *USENIX NSDI*, pp. 359-372, 2014.

[6] A. Nagulu et al., "Fully-Integrated Non-Magnetic 180nm SOI Circulator with > 1W P1dB, >+50dBm IIP3 and High Isolation Across 1.85 VSWR," *IEEE RFIC*, pp. 104-107, June 2018.

[7] R. Chen and H. Hashemi, "Passive Coupled-Switched-Capacitor-Resonator-Based Reconfigurable RF Front-End Filters and Duplexers," *IEEE RFIC*, pp. 138-141, June 2016.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

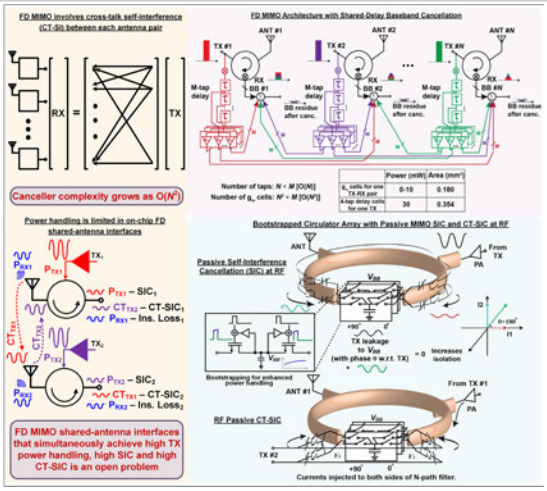


Figure 28.6.1: FD MIMO architecture with RF passive self-interference cancellation (SIC), cross-talk (CT)-SIC, and shared-delay baseband cancellation. Clock bootstrapping is embedded to enable a higher TX power handling.

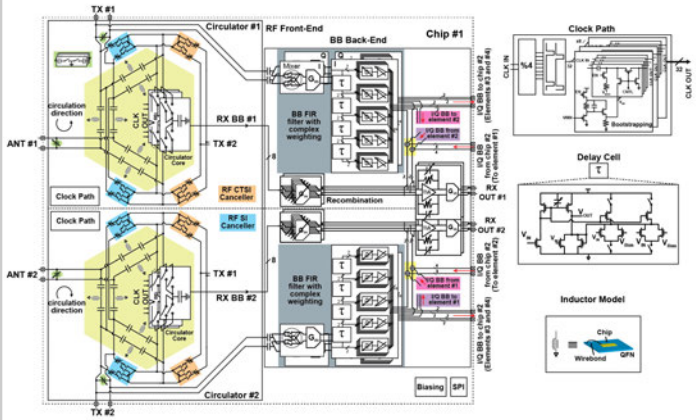


Figure 28.6.2: Block and circuit diagram of the 65nm CMOS 2.2GHz full-duplex 2×2 MIMO circulator-receiver with high TX power handling exploiting MIMO RF and shared-delay baseband self-interference cancellation.

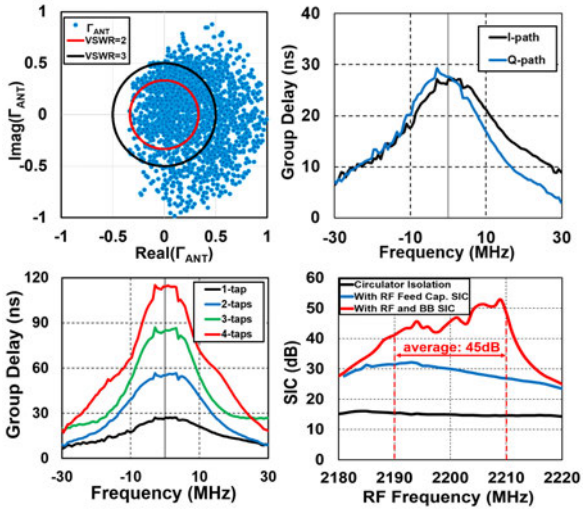


Figure 28.6.3: Single-element measurements - ANT VSWR coverage, delay-cell measurements (I/Q group delay measured for various delay taps), and single-element SIC by enabling both RF and BB cancellers.

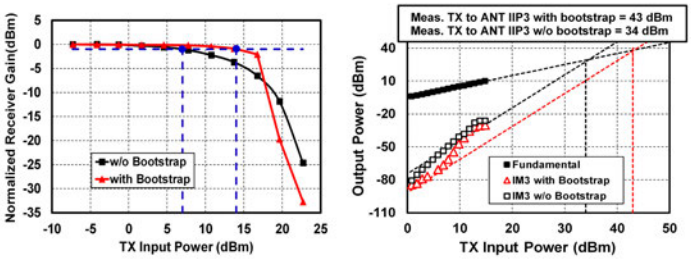


Figure 28.6.4: Single-element measurements - TX-induced ANT-RX gain compression under cancellation and TX-ANT IIP3 measurements with and without clock bootstrapping.

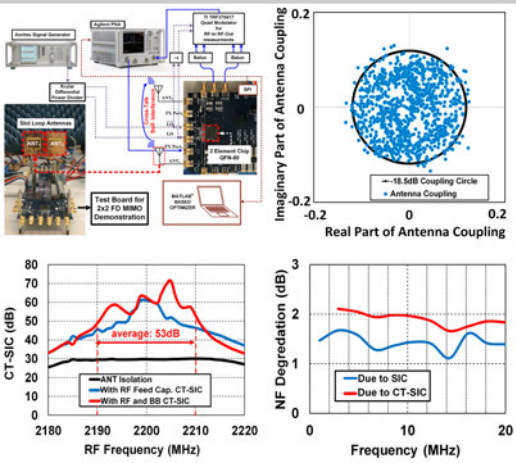


Figure 28.6.5: Measured full-duplex MIMO performance across 2-elements - CT-SI cancellation up to 18.5dB can be achieved across the complex plane by utilizing the RF CT-SI canceler. Total CT-SIC by engaging both RF and shared-delay BB cancellers is shown, along with the NF degradation due to SIC and CT-SIC.

		RFIC 2018 [4]	RFIC 2018 [6]	ISSCC 2018 [2]	This Work
Architecture		Full-duplex phased-array with integrated magnetic-free N-path-filter-based circulator-receivers	High performance circulator	Full-duplex transceiver employing an electrical balance duplexer and dual RF cancellers	Full-duplex MIMO circulator-RX with MIMO RF and Shared-Delay Baseband Self-Interference Cancellation
RX metrics	RX Frequency Range	0.61-0.975GHz	0.86-1.08GHz	1.6-1.9GHz	2.2GHz
	Number of Antenna Paths	4 phased-array elements per IC, 2 ICs tied for FD measurements	1	1	2 MIMO elements per IC (scalable)
	Gain	Max: 51dB Nominal: 41dB (single-element)	N/A	42dB	30dB
	Noise Figure	5dB (single-element)	3.1dB (circ. only)	8.1dB	9.5dB (single-element)
FD Metrics	Integrated SI Suppression Domains	RF + Spatial	RF	RF	RF+BB
	Amount of Integrated SI Suppression	50dB array SIC across 16.25MHz BW (Across 8 elements with slot-loop antennas)	25/40dB isolation across 160/30MHz BW (With 50 ohm termination) 25dB isolation across 18MHz BW (With a dipole antenna)	72.8dB SIC across 20MHz BW 70.1dB SIC across 40MHz BW 65.2dB SIC across 80MHz BW	45dB SIC across 20MHz BW (Using slot-loop antennas)
	Amount of Integrated CT-SI Suppression	N/A	N/A	N/A	53dB
	Overall TX Port Power Handling	+16.5dBm[1] (TX Array Power)	+21.3dBm[1]	+5dBm[2]	+14dBm[1]
	RX NF Degradation in Full-Duplex Mode	3dB + 3.5dB (RX/TX Array Gain Degradation)	N/A	1.6dB	1.7 (SIC) / 2.1dB (SIC+CT-SIC)
Resources	RX Power and Canceller Power	8mW / element	N/A	106mW	46mW / element (21mW Base Band + 23mW delay cells+2mW canc. gm-cells)
	Antenna Interface Power	26.25mW per element	170mW (at 0.95GHz)	N/A	156mW per element
	Technology	65nm CMOS	180nm SOI CMOS	40nm CMOS	65nm CMOS
	Active Area	3.6mm² (0.9mm² / element)	16.5mm²	4mm²	5.6mm² (2.8mm² / element)

1. Limited by ~1dB gain compression induced in the receive signal. 2. SIC under TX Power
N/A: Not Applicable N/R: Not Reported

Figure 28.6.6: Comparison with state-of-the-art FD receivers with an integrated shared antenna interface.

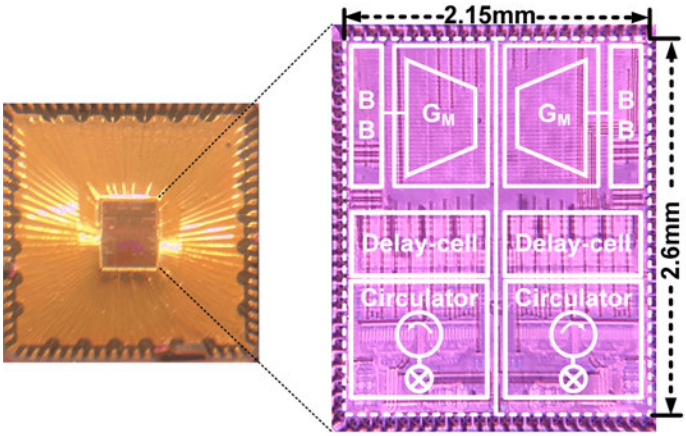


Figure 28.6.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

28.7 A Wideband Blocker-Tolerant Receiver with High-Q RF-Input Selectivity and <-80dBm LO Leakage

Huan Wang, Zisong Wang, Payam Heydari

University of California, Irvine, CA

High-Q selectivity at RF input is strongly desirable for wideband radio receivers (WB-RX) to combat strong interferences present in today's cellular mobile communications. Originally explored in the 1950's, the N-path filtering technique has been revisited in recent years to realize tunable high-Q selectivity at RF input [1,2]. From a generic perspective shown in Fig. 28.7.1, it can serve as a bandpass filter in shunt with the wideband RF LNA or as a mixer-first direct downconversion architecture. The N-path technique, however, mandates mixing switches to be directly connected to the RF port with their on-resistances required to be much smaller than 50Ω This leads to excessively large switches and thus strong LO leakage of >-67dBm [1,2] in clear infraction of the <-76dBm spurious emission requirement for 3GPP2 radios [3]. This work describes a WB-RX architecture that realizes high-Q bandpass (BP) selectivity at RX input while suppressing LO leakage down to <-80dBm. As will be described, unlike conventional N-path-based RXs, this RX avoids direct switch connection to the antenna interface, thereby breaking the fundamental trade-off between the RX input selectivity and the LO leakage.

The LO leakage strength is directly proportional to switch size and mismatch between N-path switches and eventually varies with the square root of switch size (Fig. 28.7.1). Mixer digital-to-analog-converter-based (MxDAC) calibration of switch sizes [4] is limited by mismatch between unit cells and only suppresses LO leakage to -62dBm. Injecting DC cancellation currents at baseband (BB) [5] provides better improvement of LO leakage suppression at the cost of using N high-resolution DACs and degraded noise figure (NF) performance. Moreover, this cancellation relies on amplitude and phase matching which is sensitive to antenna impedance variation during RX operation.

Direct switch connection to the RX input port is inevitable during upconversion of the BB lowpass impedance to RF. A tunable frequency selective input resistance at RF, however, circumvents the need for direct switch connection to achieve impedance frequency translation (Fig. 28.7.1). A regulated common-gate structure provides the desired BP input resistance looking into its source when a notch filter (realized using N-path structure for tunable high-Q response) is inserted in the loop with some wideband gain stages. The LO leakage from the N-path notch filter inside this bandpass common-gate (BPCG) circuit is suppressed in several ways: (1) The switch size can be much smaller than that in conventional N-path-based RXs since the switches no longer interface with any 50Ω impedance. The loop-gain of BPCG further aids the reduction of switch size to reduce LO leakage; (2) NMOS and PMOS transistors with identical size are used in the switch implementation. Even though the mismatch will be increased by a factor of $\sqrt{2}$, the LO leakage amplitude is significantly reduced due to differential LO drive, resulting in an overall LO leakage reduction; (3) The LO leakage due to switch bank1 (SWB1) is greatly suppressed by the reverse isolation of A_1 and the notch filter itself when travelling to the input port. The LO leakage originated from SWB2 is attenuated by 6dB from M1's gate to source when the input port is perfectly matched. Switch size in SWB2 is further reduced to minimize LO leakage while the notch filter loss is kept low by increasing switch size in SWB1.

A simplified RX architecture incorporating the above techniques is shown in Fig. 28.7.1. The complementary BPCG with higher g_m efficiency serves to attenuate blocker voltage swing at the RF input (node X) and thus improves the linearity of the noise cancelling (NC) downconversion path. This comes at the cost of enlarged blocker current flowing through the BPCG stage, which can saturate the following stages. A blocker sink presenting a BP impedance at f_{LO} is added to the drain of the BPCG stage to provide a low impedance path for blocker current and suppress blocker voltage swing on the signal path. The CG downconversion path employs a frequency translational feedback to synthesize a moderate (~200Ω) in-band input impedance such that the BPCG stage can provide voltage gain to suppress noise from the CG downconversion path and the blocker sink. The out-of-band (OB) input impedance of the CG downconversion path rises due to lack of BB gain at OB frequencies. However, the overall OB impedance connected to the drain of BPCG is kept low by the blocker sink, thereby suppressing blocker voltage swing at the drain of BPCG. The noise and IM3 generated in the BPCG circuit are ultimately cancelled at BB output upon optimum gain and phase matching between CG and NC downconversion paths.

Figure 28.7.2 shows the complete block diagram of the WB-RX. The parasitic capacitances along the BPCG loop cause slight shift of BP center frequency seen at RF_in. By feeding amplified BB signals within the notch filter to adjacent branches (e.g., A to B in Fig. 28.7.2), phase lead is introduced to counter the parasitic phase delay and correct the center frequency shift. The amount of phase lead is controlled by the gain of the amplifiers within the notch filter. The blocker sink is realized by an N-path notch filter around an RF gain stage. The voltage swing at the output of the blocker sink is large but does not pose any linearity concern since it is not on the signal path. Downconversions in both CG and NC paths employ 8-phase LOs to reject the 3rd and 5th LO harmonics, thereby improving NF. A harmonic-rejection recombination ratio of 29:12 is employed in BB amplifiers to approximate the ideal ratio of $1+\sqrt{2}$. The downconverted outputs of the CG and NC paths are combined for noise and IM3 cancellation. The best achievable gain and phase matching for noise and IM3 cancellation are obtained by varying BB TIA gain and the relative delay between the downconversion LO signals for the CG and NC paths. The LO generation circuitry is shown in Fig. 28.7.3. The single-ended external signal with 8× the LO frequency is divided by 8 and passed through a tunable delay cell to adjust the relative delay between the CG and NC LO signals. The 8× LO frequency signal is used to sample the delayed and the divided signals to generate 8-phase 50% duty-cycle clocks. These clocks are then combined by AND gates to generate 12.5% duty-cycle non-overlap LO signals for 8-phase downconversions.

A prototype has been fabricated in a 45nm RF SOI process (Fig. 28.7.7). This technology offers lower parasitics to substrate for better RF performance compared to its bulk CMOS counterpart. Furthermore, PMOS transistors with a speed on par with that of NMOS make the complementary switch implementation for LO leakage suppression a viable choice. The measured S_{11} (Fig. 28.7.4) demonstrates tunable highly selective input impedance across the operating frequency range. The measured input impedance in the Smith chart at 1GHz LO is shown in Fig. 28.7.4, clearly demonstrating a low OB input impedance (~8Ω). The LO leakage was measured across the operating frequency (Fig. 28.7.4). The LO leakage spectrum at two frequencies, 200MHz and 2GHz, with -92 and -80dBm leakage power are shown in Fig. 28.7.4. The effect of introducing phase lead in the notch filter to compensate for center frequency shift is shown in Fig. 28.7.5 (top left). The center of S_{11} was shifted by 4MHz for a 1.6GHz LO frequency without phase compensation, degrading the input matching at the center frequency. The S_{11} gradually re-centered as the amplifier gain in the notch filter increased. The measured IIP3 and IIP2 vs. offset frequency at an LO frequency of 1GHz are shown in Fig. 28.7.5. OB-IIP3 and OB-IIP2 are +14dBm and +60dBm at 100MHz offset. NF is plotted for different LO frequencies showing a worst case 2.45dB NF at 2GHz. The blocker NF was measured with 80MHz offset frequency showing 4.5dB degradation under the influence of a 0dBm blocker. Figure 28.7.6 compares this work with prior art. The RX achieves a low LO leakage level with high-Q selectivity at the RF input while maintaining a low NF and good linearity.

Acknowledgements:
The authors would like to thank GLOBALFOUNDRIES for chip fabrication, Fuqiang Wan and Run Xie for their assistance in chip layout, Dave Huh and Keysight Technologies for the help during chip measurement, Bao Nguyen and Professor Ahmed Eltawil for providing measurement equipment.

References:
[1] D. Murphy et al., "A Blocker-Tolerant, Noise-Cancelling Receiver for Wideband Wireless Applications," *IEEE JSSC*, vol. 47, no. 12, pp. 2943-2963, Dec. 2012.
[2] J. W. Park and B. Razavi, "Channel Selection at RF Using Miller Bandpass Filters," *IEEE JSSC*, vol. 49, no. 12, pp. 3063-3078, Dec. 2014.
[3] Digital cellular telecommunication system (Phase2+); Radio transmission and reception (3GPP TS 45.005 version 15.0.0 Release 15), July 2018.
[4] C. Wu et al., "An Interference-Resilient Wideband Mixer-First Receiver with LO Leakage Suppression and I/Q Correlated Orthogonal Calibration," *IEEE TMTT*, vol. 64, no. 4, pp. 1088-1101, Apr. 2016.
[5] S. Jayasuriya et al., "A Baseband Technique for Automated LO Leakage Suppression Achieving < –80dBm in Wideband Passive Mixer-First Receivers," *IEEE CICC*, pp. 1-4, Sept. 2014.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

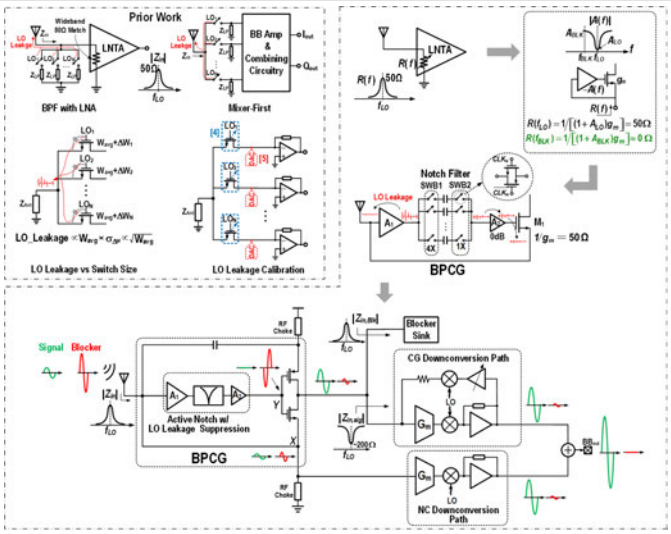


Figure 28.7.1: Summary of prior work (top left) and conceptual evolution of the proposed receiver.

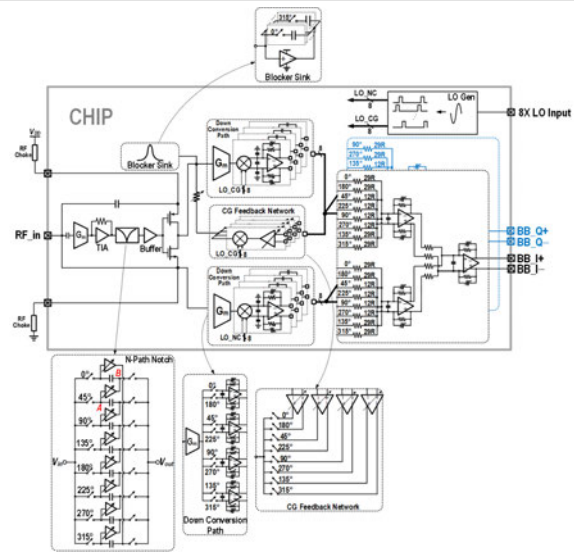


Figure 28.7.2: Complete block diagram of the proposed receiver.

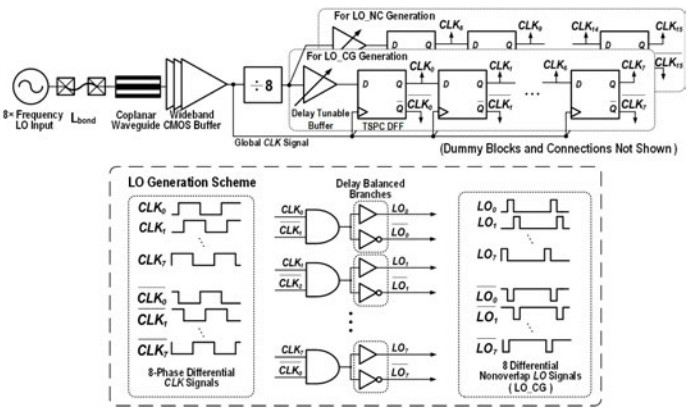


Figure 28.7.3: Eight-phase non-overlap LO generation with tunable delay for optimum noise and IM3 cancellation.

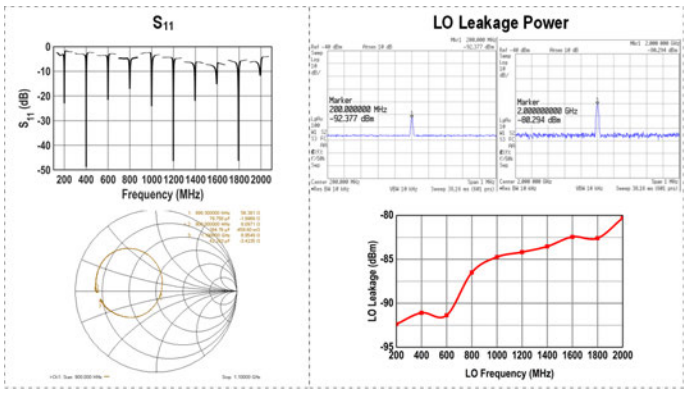


Figure 28.7.4: Measured RX S_{11} and corresponding LO leakage power.

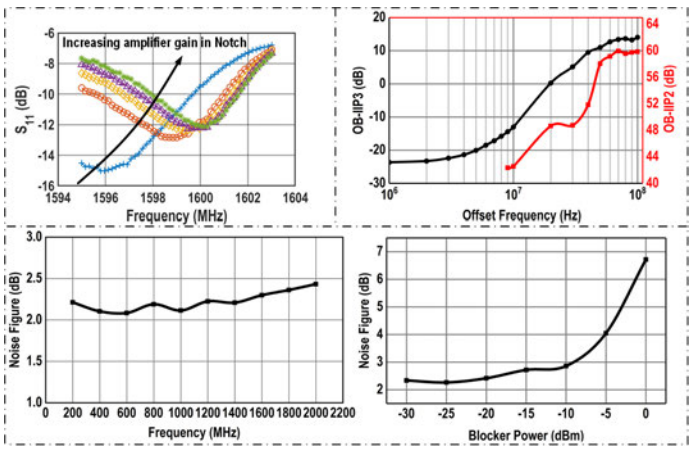


Figure 28.7.5: Measured notch filter phase lead compensation, IIP2, IIP3, small-signal noise figure and blocker noise figure.

	[1]	[2]	[4]	[5]	[6]	[7]	This Work
Architecture	Mixer-First with Noise Cancellation	Miller Band-Pass Filtering	Mixer-First with MxDAC Calibration	Mixer-First with BB LO Leakage Calibration	LNA-First with BB Blocker Filtering	LNA-First with OB IM3 Cancellation	BPCG with LO Leakage Suppression
RF Input Selectivity	Yes	Yes	Yes	Yes	No	No	**Yes**
LO Leakage [dBm]†	-65@2G† Direct Feedthrough	-67@2G† Direct Feedthrough	-62@1.5G† w/ Calibration	-80@0.7G w/ Calibration	-82** w/o Calibration	-92@1G -88@2G w/o Calibration	**-92@0.2G -80@2G w/o Calibration**
Frequency Range [MHz]	80-2700	50-2500	400-3500	100-1400	100-2800	500-2500	200-2000
Gain [dB]	72	38	35	25	50	40	40
NF [dB]	1.5 ~ 1.9	2.9	2.4 ~ 2.6	4.8 ~ 8.1	1.5 ~ 2.2	3.2 ~ 5.3	2.1 ~ 2.5
0dBm Blocker NF [dB]	4.1 (Δf=80MHz)	5.1 (Δf=20MHz)	6.5 (Δf=50MHz)	NR*	14 (Δf=50MHz)	NR*	6.7 (Δf=80MHz)
OB-IIP3 [dBm]	+13.5	+10	+16	+17	+5	+32.5	+14
OB-IIP2 [dBm]	+54	+52	+60	+23	+50	+54	+60
Supply Voltage [V]	1.3	1.2	1.1/1.5	1.2	1.1	1.2	1.2
RX Power [mW]	35.1 ~ 78	20 @ 2GHz	38 ~ 75	57.6 ~ 73.6	27 ~ 40	70	68 ~ 95
Active Area [mm²]	1.2	0.82	0.23	2***	0.8	0.84	1.05
CMOS Proc.	40nm	65nm	28nm	65nm	40nm	65nm	45nm SOI

[6] H. Hedayati, et al., JSSC 2015 [7] Y. Zhang, et al., T-MTT 2018

*NR: Not reported. **Frequency not specified. ***Graphically estimated from the reported total chip area 2.7mm².
†The largest acceptable LO leakage power according to 3GPP2 radios specs is -76dBm.

Figure 28.7.6: Performance summary and comparison with state-of-the-art receivers.

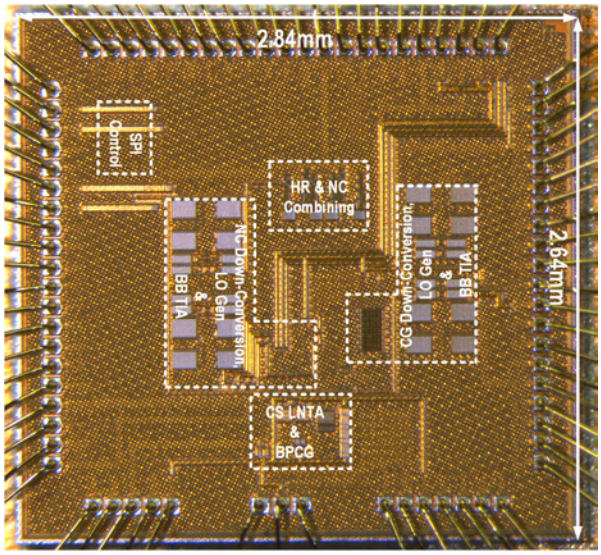


Figure 28.7.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

28.8 A 21dBm-OP_{1dB} 20.3%-Efficiency -131.8dBm/Hz-Noise X-Band Cartesian-Error-Feedback Transmitter with Fully Integrated Power Amplifier in 65nm CMOS

Jinbo Li[1], Zhiwei Xu[2], Qun J. Gu[1]

[1]University of California, Davis, CA
[2]Zhejiang University, Hangzhou, China

The increase of carrier frequency enables the utilization of the abundant high-frequency resource to meet the ceaseless demand for high data-rates. With respect to the transmitter, the design challenges include linear and high-efficiency PAs at high frequencies, low out-of-band (OOB) noise for frequency division duplex (FDD) applications, etc. Moreover, the integration of PAs on the CMOS die with other transmitter blocks greatly reduces the form factor and cuts down the cost. Among various linearization techniques, the Cartesian feedback architecture excels in its robustness against the process, voltage and temperature (PVT) variations as well as aging effects, but suffers from the noise-linearity trade-off caused by the feedback path [1-3]. To mitigate this issue, a Cartesian error-feedback architecture is proposed by adopting a reference path to cancel the signal before the feedback path and only leave the error information in principle. In this work, an *X*-band Cartesian error-feedback transmitter with a fully integrated power amplifier in bulk CMOS is implemented, and delivers maximum power (P_{SAT}) of 21.4dBm and output-referred 1dB compression point (OP_{1dB}) of 21dBm. At P_{SAT} and OP_{1dB}, the PA efficiency reaches 33.3% and 30.3%, with total system efficiency of 22.3% and 20.3%, respectively. Compared with the open-loop condition, the closed-loop transmitter demonstrates up to 13.6dB suppression of ACLR using WCDMA uplink (UL) signals, and improves EVM by over 6dB for 4Ms/s 64-QAM signals and over 3dB up to 16Ms/s signals. The out-of-band noise at 100MHz is -133/-131.8dBm/Hz for the lower and upper sidebands respectively.

Figure 28.8.1 illustrates the noise-linearity trade-off in the conventional Cartesian feedback architecture and the mitigation of the trade-off in the proposed Cartesian error-feedback architecture (the gain annotations are in 'dB' scale for this figure). The trade-off lies in the contradictory preference of the attenuator gain α for noise and linearity performances. The nonlinearities generated by the feedback path, Δ_{FB}, are positively proportional to its input $P_{OUT}+\alpha$. Thus, α needs to be small to lower Δ_{FB}. Nevertheless, this increases the noise floor referred to the system output. To address this noise-linearity trade-off, the error-feedback architecture utilizes a reference path to cancel the signal before the feedback path while only leaving the error information in principle, thus greatly suppressing the nonlinearities generated by the feedback path. In this way, larger α can be used to mitigate noise performance degradation.

Similar to the conventional feedback architecture, the output signal is attenuated by α before feeding back. Moreover, the reference path supports additional signal attention of η. Following the procedure in Fig. 28.8.2, the new ratio between the feedback path input and system output for the error-feedback architecture is derived. The additional term η is determined by the loop gain $H\alpha\beta$ and the mismatch between the reference path output and feedback path input, represented as Δ. $H\alpha\beta$ exists in the denominator of the equation of η and suppresses the mismatch Δ. With $H\alpha\beta$ of 15dB, η of -20dB is conveniently achieved with the allowed amplitude mismatch range of -60% to >100% and phase mismatch range of -60° to 60°. The amplitude and phase mismatch tolerance ranges become larger with a larger $H\alpha\beta$. This phenomenon saves design efforts and costs significantly as accurate amplitude and phase matching at RF frequencies is challenging.

The X-band PA employs a push-pull cascode structure as the output stage to allow large output voltage swings. The bias condition not only affects the static PA's DC current (I_{DC}) but also has dependence upon the input power (P_{IN}), and thus affects the efficiency. At low bias voltages, I_{DC} increases with P_{IN}. During high biases, the instantaneous output current barely changes in the first half cycle, as limited by the current delivering capability of the active device. By contrast, the output current keeps decreasing when P_{IN} increases. Therefore, the overall I_{DC} decreases when P_{IN} increases at high gate-voltage biases. This leads to similar PAE values at peak output powers despite different bias voltages. For the target output power of 21 to 22dBm, the optimal bias voltage of 600mV provides the highest PAE of 33%.

The reference path provides a signal to cancel the signal sampled from the PA output without introducing excess nonlinearities into the system. The reference path incorporates a feedback-based g_m stage and a harmonic-trapping (HT) capacitor-based LO switching stage, as depicted in Fig. 28.8.4. With sufficiently large loop gain, the effective g_m is purely determined by the resistor values, i.e., $1/(R_{1\,or\,2} \parallel R_3/2)$. For the LO switching stage, its linearity is highly affected by the $2f_{LO}$ (f_{LO} is the LO signal frequency) components at the common-source node of the switching pair, V_S. The $2f_{LO}$ components of V_S are downconverted by the $-f_{LO}$ component of the switching pair and impact the output linearity performance. The HT capacitors provide a low-impedance path for the $2f_{LO}$ components of V_S and thus suppress their contributions. The odd harmonics are unaffected as V_S is close to a virtual short node. The simulation verifies the concept with 3.3dB improvement in OIP3 and negligible effects on the noise performance of the reference path. Both linearity and noise performances comply with the system specifications.

Measurement results include using continuous-wave (CW), 2-tone, WCDMA UL and 64-QAM 4Ms/s signals, as depicted in Fig. 28.8.5, for both the open- and closed-loop conditions. In the CW measurement, the transmitter delivers 21.4dBm P_{SAT} with 22.3% total system efficiency and 33.3% PA efficiency, and 21dBm OP_{1dB} with 20.3% system efficiency and 30.3% PA efficiency. 2-tone measurement shows that OIP3 reaches 38dBm for the closed-loop condition and demonstrates around 13.5dB suppression of the 3rd-order modulation distortions (IMD3) at medium output powers. For WCDMA UL signals, the closed-loop transmitter achieves >13dB suppression on ACLR compared with the open-loop condition. EVM is improved from -28.5dB to -34.5dB for 4Ms/s 64-QAM signal and shows more than 3dB improvements up to 16Ms/s. The measured wideband noise spectrum shows -133/-131.8dBm/Hz for lower and upper sidebands, respectively.

Compared with prior Cartesian feedback transmitters listed in Fig. 28.8.6, this work operates at the highest frequency, i.e., 10GHz, integrates all the functional blocks on the same die and delivers the highest output power of 21.4dBm. None of the prior works mentioned their noise performances due to the aforementioned noise-linearity trade-off, and this work achieves the output noise of -133/-131.8dBm/Hz for lower and upper sidebands, with the realization of a low-noise Cartesian feedback transmitter. This PA-included transmitter demonstrates comparable PA and system efficiency with other standalone *X*-band PA works, and exhibits smallest efficiency degradation between P_{SAT} and OP_{1dB}. With the figure of merit (FOM) defined as $10^{OIP3/10}/P_{DC}$, this design achieves the highest FOM of 10.3. Fig. 28.8.7 depicts the power breakdown chart and the die micrograph. The chip is fabricated in 65nm CMOS with dimensions of 2.37mm×1.8mm.

Acknowledgement:
The authors would like to thank ONR and Dr. Deborah Van Vechten for financial support.

References:
[1] S. Ock et al., "A Cartesian Feedback-feedforward Transmitter IC in 130nm CMOS," *IEEE CICC*, pp. 1-4, 2015.
[2] H. Ishihara et al., "A 10MHz Signal Bandwidth Cartesian-loop Transmitter Capable of Off-Chip PA Linearization," *ISSCC*, pp. 66-67, Feb. 2010.
[3] J. Li et al., "A Fully-Integrated Cartesian Feedback Loop Transmitter in 65nm CMOS," *IEEE IMS*, pp. 103-106, 2017.
[4] H. Wang et al., "A CMOS Broadband Power Amplifier With a Transformer-Based High-Order Output Matching Network," *IEEE JSSC*, vol. 45, no. 12, pp. 2709-2722, Dec. 2010.
[5] F. Bohn et al., "Fully Integrated CMOS X-Band Power Amplifier Quad with Current Reuse and Dynamic Digital Feedback (DDF) Capabilities," *IEEE RFIC*, pp. 208-211, 2017.
[6] J. H. Tsai et al., "An X-Band Half-Watt CMOS Power Amplifier Using Interweaved Parallel Combining Transformer," *IEEE MWCL*, vol. 27, no. 5, pp. 491-493, May 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

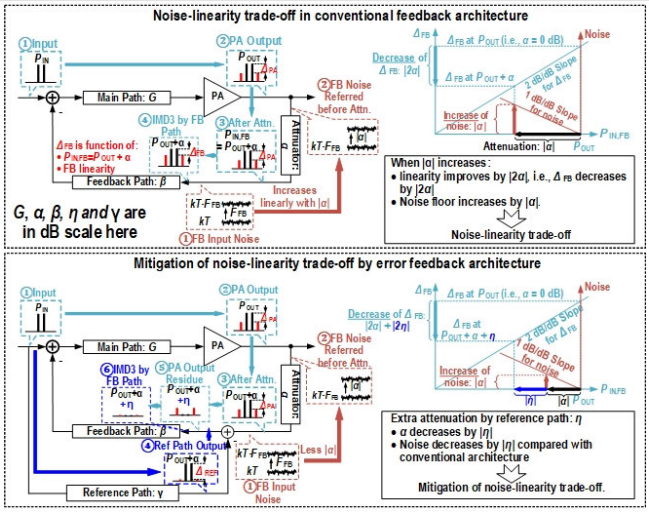


Figure 28.8.1: Noise-linearity trade-off in conventional Cartesian feedback architecture, and mitigation of the trade-off in the new Cartesian error-feedback architecture. The symbols-G, α, β, η and γ-are in dB scale in this figure.

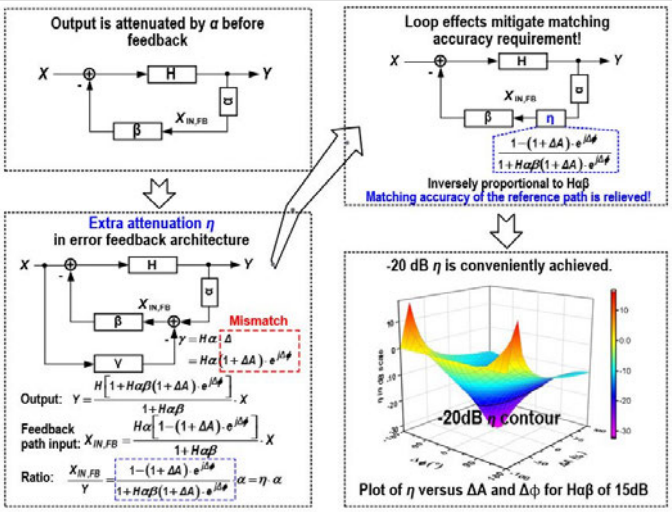


Figure 28.8.2: Mitigation of the matching requirement between the reference path and the main path due to the feedback-loop effect. The gain annotations in this figure are linear by default.

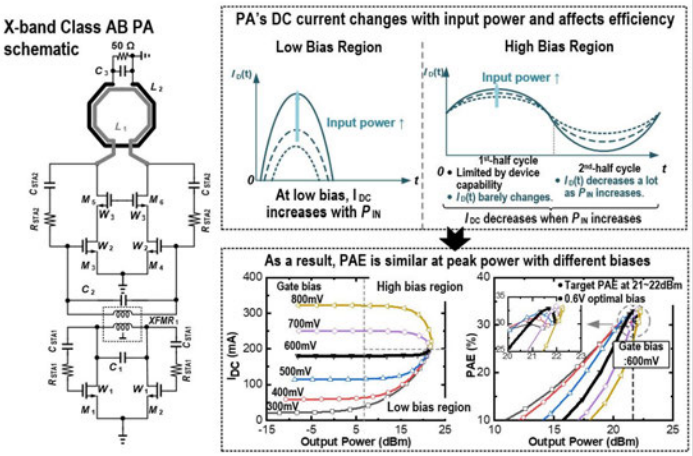


Figure 28.8.3: X-band PA design with efficiency optimization. PA's DC current (I_{DC}) changes with input power and thus affects efficiency. The contradictory effects of the bias and drain effects lead to similar PAE at peak output powers despite different bias voltages.

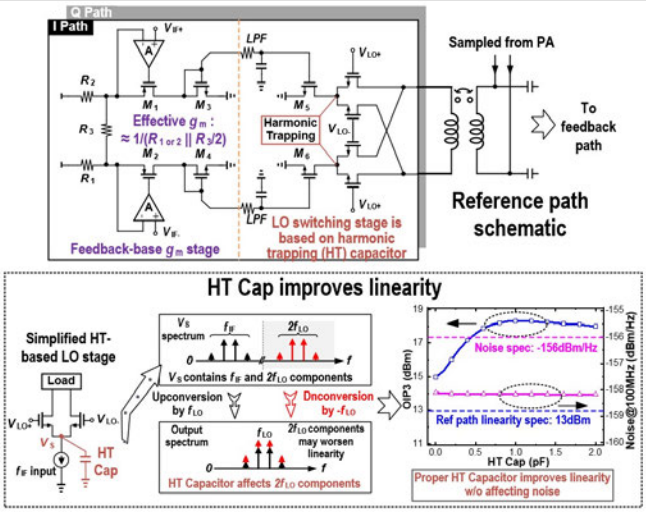


Figure 28.8.4: Implementation of the reference path. The g_m stage is based on a feedback loop, and the LO switching stage utilizes harmonic-trapping (HT) capacitors.

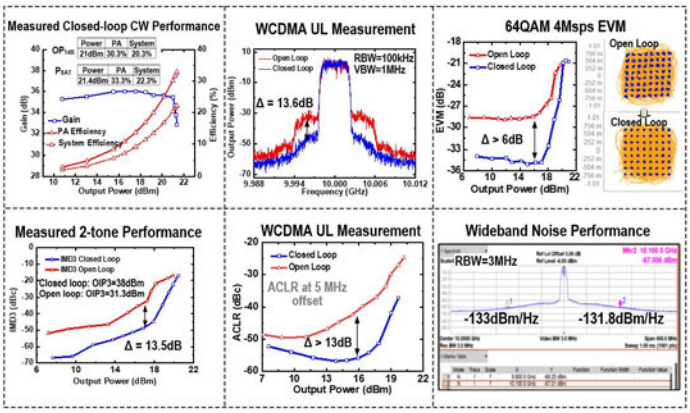


Figure 28.8.5: Measured results for continuous-wave (CW) signal, 2-tone signal, WCDMA uplink (UL) signal, 64-QAM 4Ms/s modulated signal and wideband noise performance.

Performance Comparison with Prior Cartesian Feedback Transmitters

Ref.	CICC'15 [1]	ISSCC'10 [2]	IMS'17 [3]	This work
Frequency	2.4 GHz	2 GHz	9 GHz	10 GHz
P_{OUT}	16.6 dBm	15 dBm	18.5 dBm	21.4 dBm
RF Signal BW	1.4 MHz	10 MHz	2 MHz	4 MHz
ACLR Suppression	8.7 dB	8.2 dB	9 dB	13.6 dB
EVM	N/A	-24.6dB (OL[a]) to -36dB (CL[a])	N/A	-28.5dB (OL) to -34.5dB (CL)
Noise	N/A	N/A	N/A	-133/-131.8dBm/Hz @ 100 MHz
P_{DC}	225 mW	238 mW	N/A	614.4 mW
Integration Level	Off-chip PA, coupler and Attn.	Off-chip PA and coupler	Fully-integ. with PA	Fully-integ. with PA
CMOS Technology	0.13 μm bulk	0.13 μm bulk	65 nm bulk	65 nm bulk

[a] OL is short for open loop; CL is short for closed loop

Performance Comparison with Prior X-band PAs in Bulk CMOS

Ref.	JSSC'10 [4]	RFIC'17 [5]	MWCL'17 [6]	This work
Frequency	10 GHz	10 GHz	9 GHz	10 GHz
P_{SAT}	25.2 dBm	23 dBm	27.1 dBm	21.4 dBm
Peak Efficiency	21.6%	37%	22.7%	33.3% (PA output) /22.3% (Total TX)
OP_{1dB}	22.6 dBm	N/A	24.2 dBm	21 dBm
Efficiency at OP_{1dB}	15%	N/A	7.5% [a]	30.3% (PA output) /20.3% (Total TX)
OIP3	32.2 dBm [b]	N/A	33.8 dBm [b]	38 dBm
Efficiency at -30dBc IMD3	N/A	N/A	2%	22% (PA output) /13.5% (Total TX)
P_{DC}	1533 mW	539.3 mW	2259 mW	614.4 mW [d]
FOM [c]	1.1	N/A	1.1	10.3
Integration Level	Full-integ. PA	Fully-integ. PA	Fully-integ. PA	Full-integ. PA with TX
CMOS Technology	90 nm bulk	65 nm bulk	0.18 μm bulk	65 nm bulk

[a] graphically estimated. [b] estimated by adding 9.6 dB to OP_{1dB}. [c] FOM is defined as $10^{(OIP3/10)}/P_{DC}$
[d] total TX power used for this work.

Figure 28.8.6: Performance summary and comparison with SOAs.

28

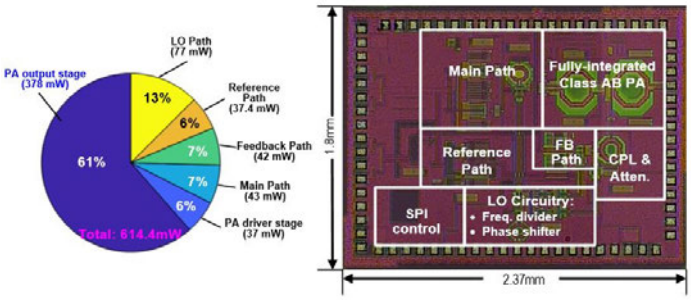


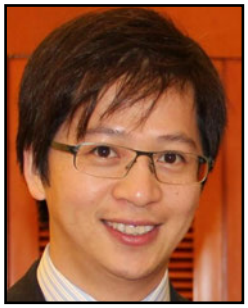
Figure 28.8.7: Power breakdown chart and die micrograph of the Cartesian error-feedback transmitter in a 65nm CMOS process.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 29 Overview: *Quantum & Photonics Technologies*

TECHNOLOGY DIRECTIONS SUBCOMMITTEE

Session Chair: ***Edoardo Charbon***
EPFL, Neuchatel, Switzerland



Associate Chair: ***Pui-In Mak***
University of Macau, Macau, China

Subcommittee Chair: Makoto Nagata, Kobe University, Kobe, Japan

Quantum and photonics technologies are emerging as a major research topic touching several disciplines, including computing, sensing, telecommunications, information technology, security, etc. This session describes important developments in the field and its applications.

1:30 PM

29.1 A 28nm Bulk-CMOS 4-to-8GHz <2mW Cryogenic Pulse Modulator for Scalable Quantum Computing

J. C. Bardin, University of Massachusetts, Amherst, MA; Google, Goleta, CA

In Paper 29.1, the University of Massachusetts at Amherst, Google, and the University of California at Santa Barbara describe a chip designed to control superconductive qubits by means of 4-to-8GHz signal patterns operating at deep-cryogenic temperatures (3K). The 28nm cryo-CMOS chip generates up to sixteen waveforms to perform several gate operations on qubits, dissipating less than 2mW per qubit.

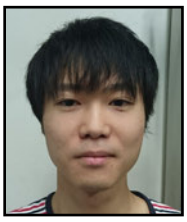
2:00 PM

29.2 A Scalable Quantum Magnetometer in 65nm CMOS with Vector-Field Detection Capability

M. I. Ibrahim, Massachusetts Institute of Technology, Cambridge, MA

In Paper 29.2, the Massachusetts Institute of Technology proposes an integrated system for the control and readout of an ensemble of nitrogen vacancy (NV) centers in diamond. The approach is based on a scalable hybrid NV-CMOS platform enabling spin control and optical readout featuring microwave generation, uniform radiation, and detection.

2:30 PM



29.3 A 48GHz 5.6mW Gate-Level-Pipelined Multiplier Using Single-Flux Quantum Logic

I. Nagaoka, Nagoya University, Nagoya, Japan

In Paper 29.3, Nagoya University proposes an 8×8b superconductor single-flux-quantum logic multiplier integrated in a 1.0μm, 9-layer process featuring 20,251 Josephson junctions. The multiplier operates at 4.2K and reaches a frequency of 48GHz while dissipating 5.6mW.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

3:15 PM

29.4 Ultra-Low-Power Atomic Clock for Satellite Constellation with 2.2×10^{-12} Long-Term Allan Deviation Using Cesium Coherent Population Trapping

H. Zhang, Tokyo Institute of Technology, Tokyo, Japan

In Paper 29.4, the Tokyo Institute of Technology presents a miniaturized atomic clock system based on a cesium-133 gas cell. The system features a vertical-cavity surface-emitting laser (VCSEL), temperature/magnetic controllers, and control circuits; it achieves a long-term Allan deviation of 2.2×10^{-12} at $\tau=10^5$s within a volume of 15.4cm^3 dissipating 59.9mW.

3:45 PM

29.5 A Single-Chip Optical Phased Array in a 3D-Integrated Silicon Photonics/65nm CMOS Technology

T. Kim, University of California, Berkeley, CA

In Paper 29.5, the University of California at Berkeley presents an optical phased array implemented in a 3D platform, where photonics and CMOS, independently optimized, are oxide-bonded and connected through high-density vias. With 18.5°/16° steering range and 0.15°/0.6° beamwidth, the system achieves an effective antenna aperture of 500µm×130µm resulting in a lateral resolution of 0.26m×1m at 100m.

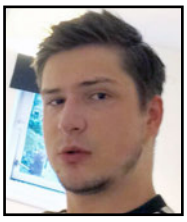
4:15 PM

29.6 A Digital-Type GaN Driver with Current-Pulse-Balancer Technique Achieving Sub-Nanosecond Current Pulse Width for High-Resolution and Dynamic Effective Range LiDAR System

Z-Y. Lin, National Chiao Tung University, Hsinchu, Taiwan

In Paper 29.6, National Chiao Tung University proposes a digital-type GaN driver for sub-nanosecond laser pulses to be used in a LiDAR featuring a resolution of 0.3m and dynamic range higher than 0.75m. The driver achieves a wide range of repetition frequencies up to 200MHz, adjusting the laser current rise and fall times to within hundreds of picoseconds and a pulse width of 0.9ns.

4:45 PM



29.7 A 500Mb/s -46.1dBm CMOS SPAD Receiver for Laser Diode Visible-Light Communications

J. Kosman, University of Edinburgh, Edinburgh, United Kingdom; STMicroelectronics, Edinburgh, United Kingdom

In Paper 29.7, the University of Edinburgh presents a Visible Light Communication receiver fabricated in 0.13µm CMOS technology. Based on a 64×64 SPAD array with 43% fill factor, the receiver achieves a data rate of 500Mb/s 4-PAM with a sensitivity of –46.1dBm and a 2×10^{-3} BER in a 450nm laser diode free-space link at 1m and 1klux background illumination.

5:00 PM

29.8 SHARC: Self-Healing Analog with RRAM and CNFETs

A. G. Amer, Massachusetts Institute of Technology, Cambridge, MA

In Paper 29.8, the Massachusetts Institute of Technology describes a new technique, SHARC (Self-Healing Analog with RRAM and Carbon nanotube), making carbon-nanotube field-effect transistor (CNFET) analog circuits robust to metallic carbon nanotubes (CNTs). The demonstrated circuits in a 3µm technology node include a 4b ADC (2-stage op-amp and switches) and a 4b-SAR ADC (strong-arm latch and switches).

29.1 A 28nm Bulk-CMOS 4-to-8GHz <2mW Cryogenic Pulse Modulator for Scalable Quantum Computing

Joseph C. Bardin[1,2], Evan Jeffrey[2], Erik Lucero[2], Trent Huang[2], Ofer Naaman[2], Rami Barends[2], Ted White[2], Marissa Giustina[2], Daniel Sank[2], Pedram Roushan[2], Kunal Arya[2], Benjamin Chiaro[3], Julian Kelly[2], Jimmy Chen[2], Brian Burkett[2], Yu Chen[2], Andrew Dunsworth[3], Austin Fowler[2], Brooks Foxen[3], Craig Gidney[2], Rob Graff[2], Paul Klimov[2], Josh Mutus[2], Matthew McEwen[3], Anthony Megrant[2], Matthew Neeley[2], Charles Neill[2], Chris Quintana[2], Amit Vainsencher[2], Hartmut Neven[4], John Martinis[2,3]

[1]University of Massachusetts, Amherst, MA
[2]Google, Goleta, CA
[3]University of California, Santa Barbara, CA, [4]Google, Los Angeles, CA

While quantum processors are typically cooled to <25mK to avoid thermal disturbances to their delicate quantum states, all qubits still suffer decoherence and gate errors. As such, quantum error correction is needed to fully harness the power of quantum computing (QC). Current projections indicate that >1,000 physical qubits will be required to encode one error-corrected qubit [1]. Implementing a system with 1,000 error-corrected qubits will likely require moving from the contemporary paradigm where control and readout of the quantum processor is carried out using racks of room temperature electronics to one in which integrated control/readout circuits are located within the cryogenic environment and connected to the quantum processor through superconducting interconnects [2]. This is a major challenge, as the cryo ICs must be high performance and very low power (eventually <1mW/qubit). In this paper, we report the design and system-level characterization of a prototype cryo-CMOS IC for performing XY gate operations on transmon (XMON) qubits.

A XMON qubit [3] is a non-linear resonator comprised of a capacitor in parallel with a Josephson junction (see Fig. 29.1.1). The non-linearity results in anharmonic energy levels, with f_{01} and $f_{01} - f_{12}$ typically 4-to-8GHz and 250MHz. The junction loop (SQUID) makes the XMON frequency-tunable. The XY, Z, and readout ports are used for microwave drive, frequency tuning, and readout, respectively. Due to the anharmonicity, a band-limited drive signal centered at f_{01} applied to the XY port excites only the 0→1 transition. In this case, the qubit is represented by its two lowest energy levels: $|\psi\rangle = \cos(\theta/2)|0\rangle + \exp\{j\phi\}\sin(\theta/2)|1\rangle$, which can be interpreted as a point on the surface of the Bloch sphere (Fig. 29.1.1). The XY drive signal produces a deterministic rotation (gate operation) about an axis in the XY plane of the Bloch sphere, where the axis and angle of rotation are determined by the carrier phase and integrated envelope amplitude of the drive signal, respectively. The finite coherence time of a XMON (~0.1ms) motivates fast gates, but there is a tradeoff: the increased spectral width of fast pulses can drive the f_{12} transition, causing errors. Thus, shaped pulses, typically Gaussian or cosine, are used. Pulse lengths and envelope amplitudes referenced to the XY port of the qubit are typically 10-to-30ns and 10-to-100μV.

A standard XMON control/readout system [4], as shown in Fig. 29.1.2, uses a number of AWGs and ADCs that are fast (1Gsps) and high-resolution (14b and 8b, respectively). We propose replacing the XY AWG with an IC of the architecture shown in Fig. 29.1.2, situated on the 3K stage of the system. The IC was designed to support carriers from 4-to-8GHz and consists of a pair of current-mode DACs that are controlled via a waveform memory and upconverted by a vector modulator. It was designed to operate at a physical temperature of 3K, where PDK models are unavailable: in addition to power draw and envelope accuracy, tolerance to increases in mismatch and threshold voltages [5,6] was considered during design. Fortunately, non-monotonicity and/or non-linearity in the DACs is tolerable for this application, since calibration would be carried out in a practical QC system.

As shown in Fig. 29.1.3, the envelope currents are generated by a DAC architecture optimized to produce symmetric waveforms. Each DAC contains 11 8b sub-DACs. Once triggered, a state-machine sequentially enables each sub-DAC, producing a monotonically increasing staircase of current. Once all DACs have been enabled, they are disabled in the reverse order from which they were enabled, resulting in a 22 clock-cycle long pulse. The nominal weighting of each sub-DAC is found from the derivative of the desired envelope. On-chip memory permits storage of 16 waveforms (instructions). The 8b weightings for each sub-DAC and the value of the reference current for each envelope generator (I_N) is reconfigurable on a waveform-by-waveform basis. The reference current, I_P, has 6b of resolution. The baseband currents are filtered and then upconverted using a pair of passive mixers, whose differential outputs are transformer-coupled and combined as a single-ended signal. Each baseband DAC is connected to the corresponding mixer through a polarity switch, enabling full phase coverage. The transformer-coupled LO signals drive amplifier chains that were optimized as a tradeoff between power draw and frequency range. Tuning capacitors are incorporated on the bondpad side of each transformer to permit optimum coupling from 4-to-8GHz. For simplicity, no provision was made to null LO leakage, as it can be cancelled off-chip and was not viewed as a fundamental limitation. This functionality could be added in a future design.

A die micrograph appears in Fig. 29.1.7, which also shows the wirebonded IC and the module used for testing. The module was mounted on the 3K stage of a dilution refrigerator for testing with a qubit. The setup (Fig. 29.1.4) included provisions to inject a signal from a standard XY AWG, null-out LO leakage, and monitor the IC output. The IC was characterized at 300K using an oscilloscope connected to TP1 and functioned with f_{LO} and f_{CLK} exceeding the range of 4-to-8GHz and 0.5-to-3GHz, respectively. The power required to drive the inputs of the clock and LO hybrids was less than –20dBm at 2GHz and –10dBm at 5.6GHz, respectively. An example waveform for f_{LO}=5.6GHz and f_{CLK}=2GHz appears in Fig. 29.1.4. Here, the chip was initialized with 16 different waveforms, which were stepped through using the select lines.

Once cold, the qubit was tuned to 5.6GHz and settings for AT1 and PH1 were found to minimize idle $|1\rangle$ state occupation (i.e., due to LO leakage). Rabi experiments using the cryo-CMOS IC were next carried out. As described in Fig. 29.1.5, state probabilities were measured as a function of pulse amplitude when the qubit had been initialized to $|0\rangle$ and then driven by either one or two raised cosine pulses of varying amplitude. The amplitude was varied by sweeping the DAC reference current (I_N) for eleven different values of I_P. At each point, state probabilities were computed from 5,000 measurements. At 3K, the DACs producing I_N were non-linear and non-monotonic, so the Rabi data are plotted versus pulse amplitude, measured at the XY monitor port using a spectrum analyzer. The results in Fig. 29.1.5 show the expected behavior: the maxima of the $|0\rangle$ and $|1\rangle$ state probabilities are consistent with separately measured $|0\rangle$ and $|1\rangle$ state readout error rates of 2.4% and 6.8%.

The fast-switching and phase-control features of the IC were evaluated using a three-gate experiment, consisting of (1) initializing the qubit to the $|0\rangle$ state, (2) applying an X-pulse to produce a rotation of θ_A about the X-axis, (3) applying a π-pulse with carrier phase ϕ_B to produce a rotation of π radians about a vector at an angle of ϕ_B from the X-axis in the XY plane, (4) applying a second X-pulse to produce a rotation of θ_A about the X-axis, and (5) measuring the qubit state (see Fig. 29.1.6). This sequence was run for ϕ_B in $(0, 2\pi)$ and A_A such that θ_A was estimated to be in the range of 0 to π. Raised cosine envelopes were used for all pulses. The results, which show the IC can be used to perform coherent quantum control, appear in Fig. 29.1.6 alongside baseline measurements taken using standard qubit control electronics. The RMS error for the IC is <12%, which is excellent, given that no calibration was performed.

The cryo-CMOS IC is compared to a standard system in Fig. 29.1.6. Of particular importance is the low power, which was estimated by measuring the total AC and DC power required to continuously drive π-pulses. While the demonstration of a cryogenic quantum control interface dissipating <2mW is an important step towards the development of a scalable quantum control and measurement system, much research is still required to implement such a system. Future work could focus on calibration algorithms or the development of ICs for control of multiple qubits, qubit readout, or other related applications.

References:
[1] A.G. Fowler, et al., "Surface codes: Towards practical large-scale quantum computation," *Phys. Rev. A*, vol. 86, no. 3, p. 032324, 2012.
[2] E. Charbon, et al., "Cryo-CMOS circuits and systems for scalable quantum computing," *ISSCC Dig. Tech. Papers*, pp. 264-265, Feb. 2017.
[3] J. Koch, et al., "Charge-insensitive qubit design derived from the Cooper pair box," *Phys. Rev. A*, vol. 76, no. 4, p. 042319, 2007.
[4] R. Barends, et al., "Superconducting quantum circuits at the surface code threshold for fault tolerance," *Nature*, no. 508, p. 500, 2014.
[5] K. Das and T. Lehmann, "SOS current mirror matching at 4K: A brief study," *Proc. ISCAS*, pp. 3405-3408, 2010.
[6] A. Beckers, F. Jazaeri, and C. Enz, "Characterization and Modeling of 28-nm Bulk CMOS Technology Down to 4.2 K," *IEEE J. Elec. Dev. Soc.*, vol. 6, pp. 1007-1018, 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

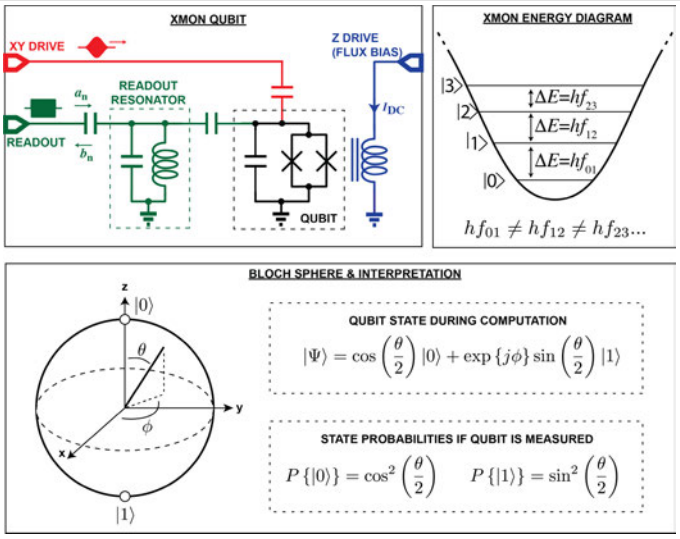


Figure 29.1.1: Transmon qubit schematic, energy diagram, and Bloch sphere representation. "X" schematic symbols represent Josephson junctions.

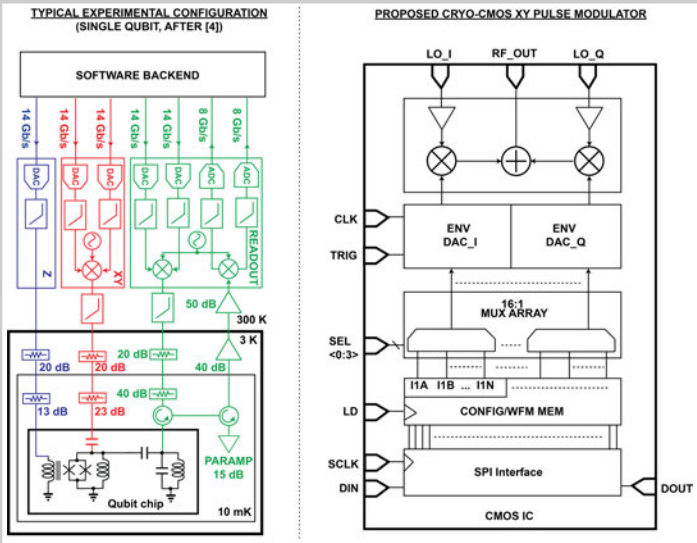


Figure 29.1.2: Standard qubit control/readout hardware and proposed XY Control IC.

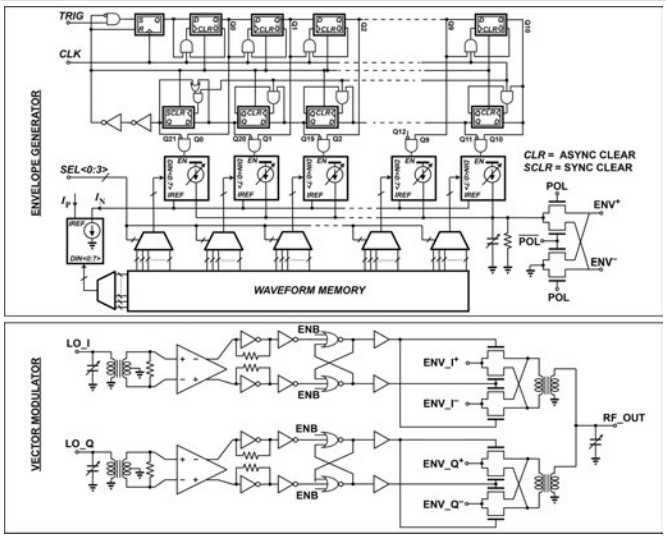


Figure 29.1.3: Schematic diagrams of current-mode envelope generator and vector modulator.

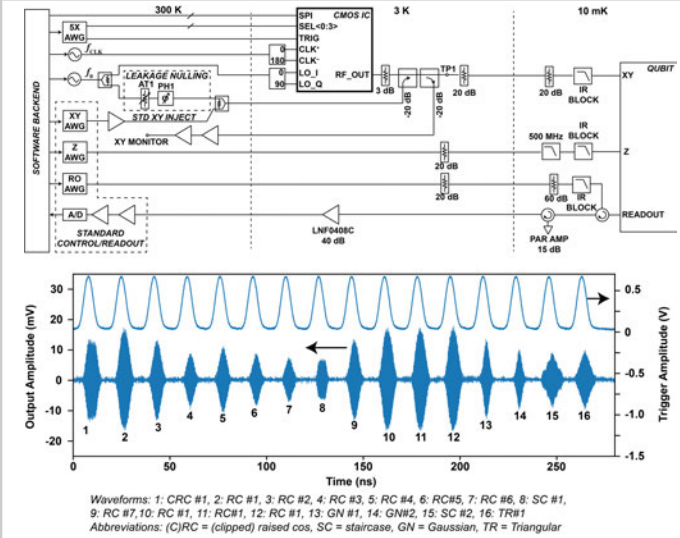


Figure 29.1.4: Experimental configuration and time domain waveforms, measured at room temperature.

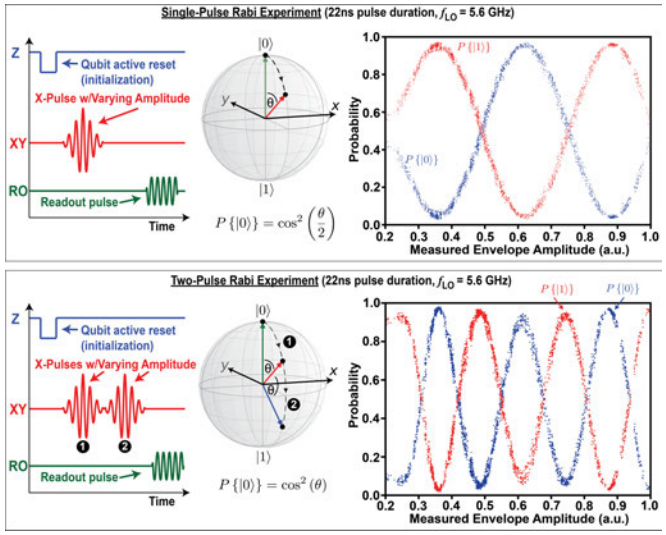
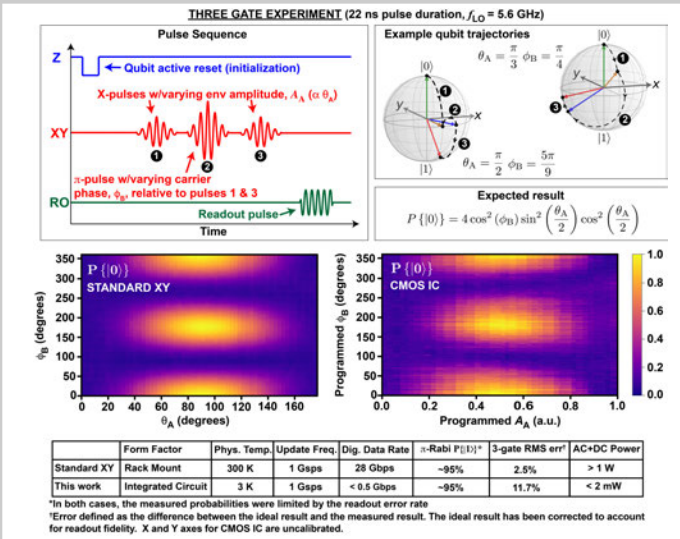


Figure 29.1.5: One- and two-pulse Rabi experiments. The envelope amplitude could not be measured below ~0.25 due to SNR.



	Form Factor	Phys. Temp.	Update Freq.	Dig. Data Rate	π-Rabi P{|0\rangle}*	3-gate RMS err†	AC+DC Power
Standard XY	Rack Mount	300 K	1 Gsps	28 Gbps	~95%	2.5%	> 1 W
This work	Integrated Circuit	3 K	1 Gsps	< 0.5 Gbps	~95%	11.7%	< 2 mW

*In both cases, the measured probabilities were limited by the readout error rate
†Error defined as the difference between the ideal result and the measured result. The ideal result has been corrected to account for readout fidelity. X and Y axes for CMOS IC are uncalibrated.

Figure 29.1.6: Pulse sequence experiment and comparison to standard XY control architecture.

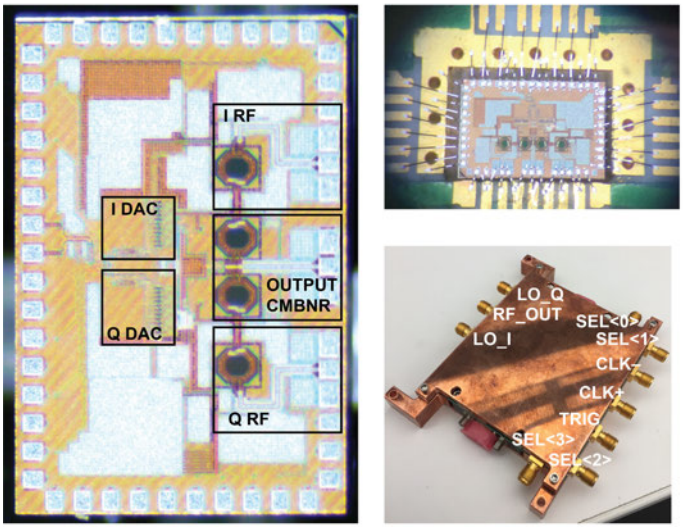


Figure 29.1.7: Die, PCB interface, and module photos. The integrated circuit measures 1mm by 1.6mm.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

29.2 A Scalable Quantum Magnetometer in 65nm CMOS with Vector-Field Detection Capability

Mohamed I. Ibrahim, Christopher Foy, Dirk R. Englund, Ruonan Han

Massachusetts Institute of Technology, Cambridge, MA

Room-temperature control and detection of the nitrogen vacancy (NV) center in diamond's spin-state has enabled magnetic sensing with high sensitivity and spatial resolution [1], [2]. However, current NV sensing apparatuses use bulky off-the-shelf components, which greatly increase the system's scale. In [3], a compact platform, which attaches nanodiamond particles to a CMOS sensor, shrinks this spin-based magnetometer to chip scale; however, the optically detected magnetic resonance (ODMR) curve it generates carries large fluctuation leading to inferior sensitivity. In this paper, we present a CMOS-NV quantum sensor with (i) a highly-scalable microwave-delivering structure and (ii) a Talbot-effect-based photonic filter with enhanced green-to-red suppression ratio. The former enables coherent driving of an increased number of NV centers, and the latter reduces the shot noise of the photo-detector caused by the input green laser. In addition, the usage of a bulk diamond also enables vector magnetometry, which allows for the tracking of magnetic objects and navigation. The prototype sensor provides a measured vector-field sensitivity of 245nT/Hz$^{1/2}$.

The NV center in diamond is an optically active point defect (Fig. 29.2.1), which consists of a nitrogen atom and a vacancy that replace an adjacent pair of carbon atoms in the diamond lattice. Figure 29.2.1 also illustrates the NV energy level diagram. It forms a spin triplet at its ground state, with a $|m_s = 0\rangle$ state at its lowest energy and degenerate $|m_s = \pm1\rangle$ states raised by 2.87GHz. An external magnetic field with a component, B_z, along the N-V axis splits the $|+1\rangle$ and $|-1\rangle$ states by $2\gamma_e B_z$, (γ_e= 28GHz/T). Consequently, the NV spin resonance frequencies (f_- and f_+) required to drive these two spin states split from 2.87GHz. The difference between f_- and f_+ is proportional to B_z and can be measured from optical detection of the magnetic resonances (ODMR). This optical detection is accomplished by exciting the NV centers with green excitation, resulting in red fluorescence minima at f_- and f_+. In a bulk diamond, each of the four crystal lattice axes (Fig. 29.2.1) possesses a population of NV centers. To fully reconstruct an applied magnetic-field's direction and magnitude, its projections along each of the NV's axes are measured.

Shown in Fig. 29.2.2, the CMOS quantum sensor is composed of two sub-systems: (i) generation, switching and uniform delivery of ~ 2.87GHz microwave to coherently excite the spin states of the NV ensemble, and (ii) filtering of green excitation and detection of red NV fluorescence. The 2.87GHz microwave signal is generated by an on-chip phase-locked loop (PLL). In [3], a loop inductor was used to deliver the microwave to diamond, but the magnetic field uniformity was sufficient only in a small area within the loop, leading to a lower number, N, of coherently driven NV centers (hence degraded sensitivity: $S_m \propto \sqrt{N}^{-1}$). In this work, an array of current-driven wires is used to synthesize a highly-uniform field in the lateral direction (B_y in Fig. 29.2.3) of the chip's surface. This design stems from the fact that an infinite and uniform current (I_0) array produces a homogeneous transverse magnetic field (Fig. 29.2.3). However, with a finite number of current-driven wires, the field in the vertical direction B_z is no longer fully cancelled and increases towards the array edges. To recover field homogeneity, a few additional wires carrying current larger than I_0 are placed on the edges of the array, providing a ramp-up profile of B_z that cancels the aforementioned vertical field across a wide range of wire array. Our design consists of 32 uniform-current-driven wires (I_0=0.5mA) and 8 field-compensating wires (I_B=3I_0) at each array edge. As shown in the EM simulations and Ampere's law calculations in Fig. 29.2.3, this simultaneously improves the uniformity of B_y and reduces B_z close to zero. The wire currents are regulated by a bank of current sources gated by the PLL output signal. To prevent phase variation of the microwave currents along the Y-axis, the VCO of the PLL is based on a bank of coupled ring oscillators, which are uniformly placed along the Y-axis. The phase noise is subsequently decreased, too.

The incident green light has a much stronger intensity compared to the red fluorescence and adds noise at the sensor output. Previously, a single metal-grating-layer filter based on wavelength-dependent plasmonic loss was used, but its green-to-red rejection ratio (R_{gr}≈10dB) is insufficient [3]. In this work, a multi-grating-layer filter using the Talbot effect is adopted. An incident light upon a grating generates a diffraction pattern, where the intensity maxima and minima repeat in the vertical direction with a wavelength-dependent Talbot length (Fig. 29.2.4). Previously, on-chip Talbot filters were used for angle-sensitive imaging [4]. Here, we strategically place other grating layers in M6 and M3 in positions aligned with the maxima for green wavelength and minima for red wavelength. As the FDTD simulation shows, such a configuration further rejects the green light, in addition to the plasmonic filtering via thick M8 grating layer. The simulated transmission through the filter is 45% for red and 0.05% for green (R_{gr}≈30dB). The pitch in all grating layers is 800nm. Note that the aforementioned current-driven wires plus interleaved dummy wires in M8 form the first photonic filter layer. The spin-dependent red fluorescence is then detected by an on-chip N-well/P-sub photodiode below the optical filter. The photodiode size is 80μm × 300μm. The measured responsivity of the photodiode is 188mA/W.

The quantum sensor is prototyped using a 65nm CMOS technology. The chip has a die area of 1.5mm^2 and consumes 40mA of power in the experiments. A bulk diamond is placed on the chip surface. We introduce a cut in the diamond's corner to direct the incident vertical green laser horizontally. This enhances the overall green rejection. The setup to obtain the ODMR plots is shown in Fig. 29.2.5. A green laser (λ=532nm) is collimated and coupled into the diamond to excite embedded NV centers. Figure 29.2.5 also shows a photo of the setup with green filter on the camera to show the diamond's red fluorescence. The fluorescence collected by the integrated photodiode is measured by a lock-in amplifier. To filter the experimental noise, a frequency-modulated spectroscopy approach is adopted, in which the instantaneous microwave frequency of the chip is modulated with a modulation depth of 6MHz and repetition rate of 3kHz. In addition, differential measurement is done to cancel the green laser variations. A signal source provides both the FM-modulated clock of the PLL and the 3kHz reference to the lock-in amplifier. The lineshape of the measured ODMR is then close to the first-order derivative of the resonances shown in Fig. 29.2.1. First, without applying an external magnetic field, a strong spin resonance is detected at PLL output frequency of 2.87GHz (Fig. 29.2.5). Then, the noise floor of the sensor (σ) is measured (Fig. 29.2.5). Beyond the corner frequency of 3kHz, the sensor white noise is about 0.1μV/Hz^{½}. A magnet placed near the sensor produces four pairs of ODMR dips, corresponding to the four diamond crystal orientations with respect to the magnetic field. The output signal is reduced due to the breaking of the degeneracy of the resonances. This allows the driving of each NV spin population separately. Lastly, the vector-field sensitivity, calculated using the slope of the ODMR curve, m, the measured noise floor, and the constant γ_e, is estimated at 245nT/Hz$^{1/2}$. Figure 29.2.6 summarizes the chip performance in comparison with the prior work [3]. The present device improves sensitivity by 3 orders of magnitude compared to previous CMOS-diamond quantum sensors, while also providing vector-field detection. The current-driven microwave array and photonic filter designs are readily scalable to larger areas for further improvements in sensitivity, or to larger numbers of detectors for gradient magnetometry, multiplexed analytical NMR spectroscopy [5], and other applications [6].

Acknowledgements:
This work is supported by the National Science Foundation (Grant No. 1839159), MIT Center for Integrated Circuits & Systems, Singaporean-MIT Research Alliance, and Master Dynamic Limited. D. E. acknowledges the support of the Army Research Office MURI on "Imaging and Control of Biological Transduction using NV-Diamond". The authors acknowledge Xiang Yi, Donggyu Kim, David Bono, and Cheng Peng at MIT for their technical assistance during prototyping and testing.

References:
[1] J. Taylor, et al., "High-sensitivity diamond magnetometer with nanoscale resolution," *Nature Physics*, vol. 4, no. 10, p. 810, 2008.
[2] G. Balasubramanian, et al., "Nanoscale imaging magnetometry with diamond spins under ambient conditions," *Nature*, vol. 455, no. 7213, p. 648, 2008.
[3] M. I. Ibrahim, et al., "Room-temperature quantum sensing in CMOS: On-chip detection of electronic spin states in diamond color centers for magnetometry," *IEEE VLSI Circuits Symposium,* pp. 249-250, June 2018.
[4] A. Wang, et al., "A light-field image sensor in 180nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 1, pp. 257-271, Jan. 2012.
[5] D. R. Glenn, et al., "High-resolution magnetic resonance spectroscopy using a solid-state spin sensor," *Nature*, vol. 555, no. 7696, p. 351, 2018.
[6] C. L. Degen, et al., "Quantum sensing," *Reviews of modern physics*, vol. 89, no. 3, p. 035002, 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

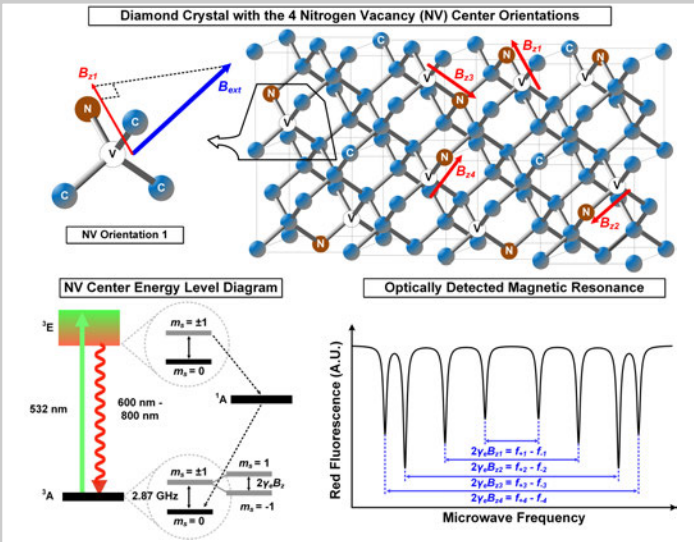


Figure 29.2.1: NV centers in diamond: atomic structure, energy-level diagram and ODMR spectrum with an external field.

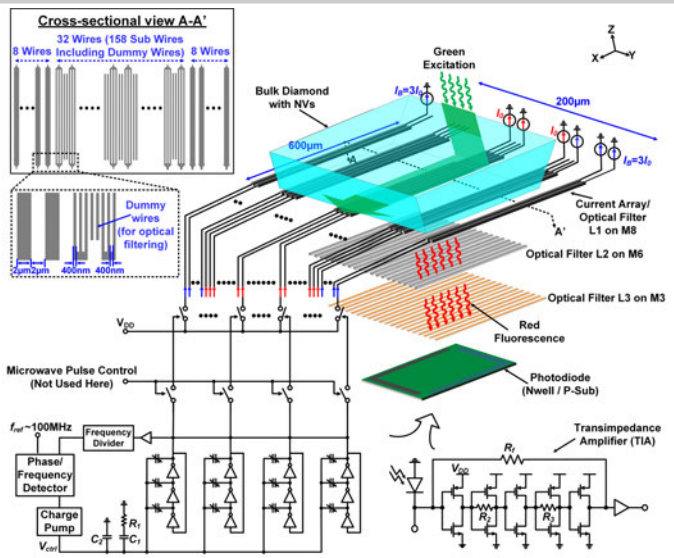


Figure 29.2.2: Schematic of the NV-based magnetic sensor in 65nm CMOS.

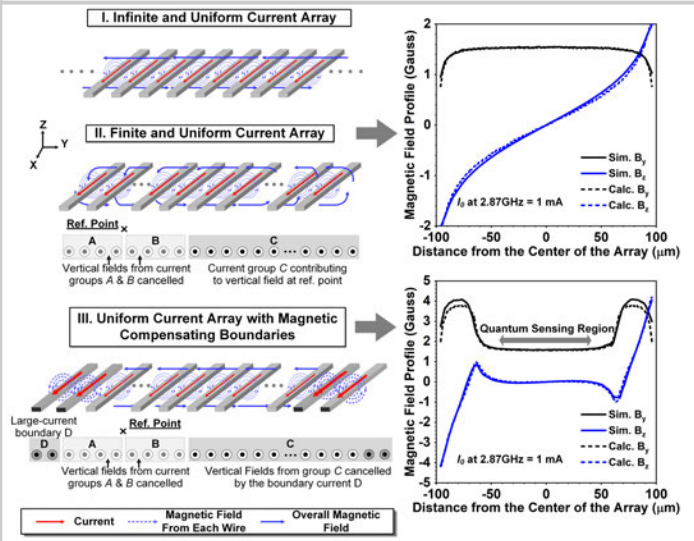


Figure 29.2.3: On-chip current-driven array for high-uniformity magnetic field generation and its field profiles based on EM simulation and calculation.

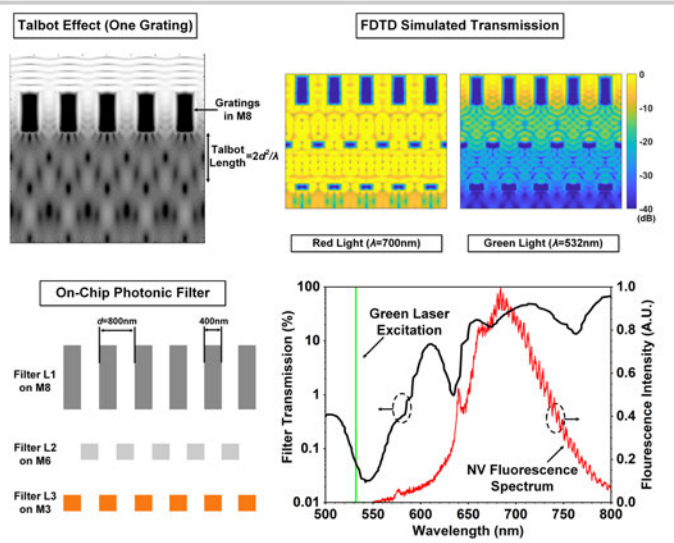


Figure 29.2.4: Design and simulated performance of the multi-layer integrated optical filter based on Talbot effect.

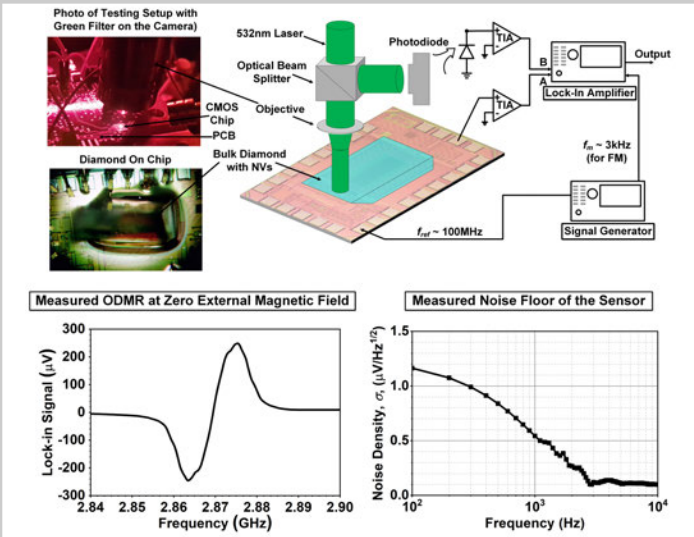
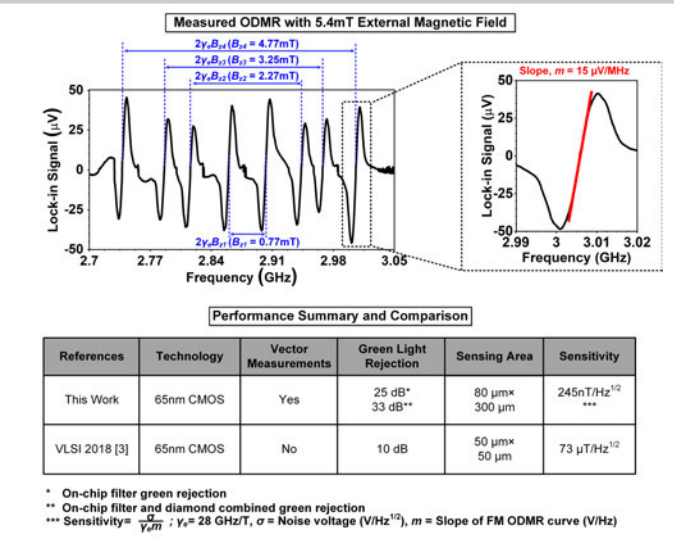


Figure 29.2.5: Test setup of the NV-CMOS magnetometer using frequency modulation, the measured ODMR spectrum and the measured noise floor.



References	Technology	Vector Measurements	Green Light Rejection	Sensing Area	Sensitivity
This Work	65nm CMOS	Yes	25 dB* 33 dB**	80 μm× 300 μm	245nT/Hz$^{1/2}$ ***
VLSI 2018 [3]	65nm CMOS	No	10 dB	50 μm× 50 μm	73 μT/Hz$^{1/2}$

* On-chip filter green rejection
** On-chip filter and diamond combined green rejection
*** Sensitivity= $\frac{\sigma}{\gamma_e m}$; γ_e= 28 GHz/T, σ = Noise voltage (V/Hz$^{1/2}$), m = Slope of FM ODMR curve (V/Hz)

Figure 29.2.6: Measured ODMR spectrum with external magnetic field and a performance summary/comparison table.

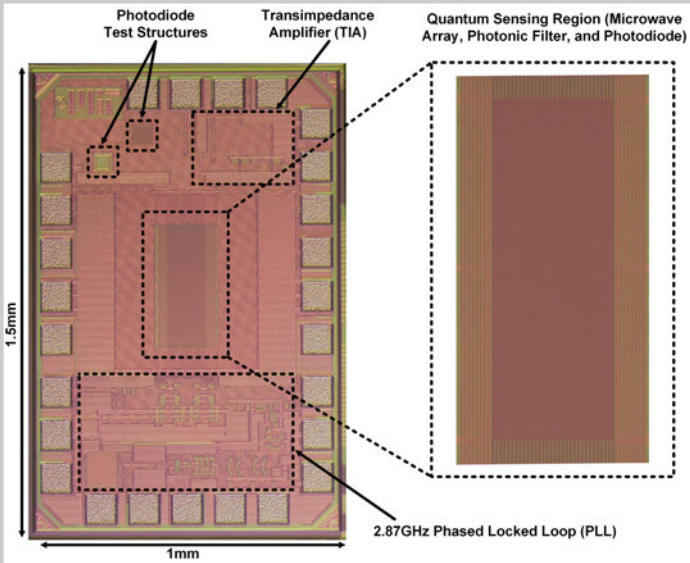


Figure 29.2.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

29.3 A 48GHz 5.6mW Gate-Level-Pipelined Multiplier Using Single-Flux Quantum Logic

Ikki Nagaoka[1], Masamitsu Tanaka[1], Koji Inoue[2], Akira Fujimaki[1]

[1]Nagoya University, Nagoya, Japan
[2]Kyushu University, Fukuoka, Japan

A multiplier based on superconductor single-flux-quantum (SFQ) logic is demonstrated up to 48GHz with the measured power consumption of 5.6 mW. The multiplier performs 8 × 8-bit signed multiplication every clock cycle. The design is based on a bit-parallel, gate-level-pipelined structure that exploits ultimately high-throughput performance of SFQ logic. The test chip fabricated using a 1.0-μm, 9-layer process consists of 20,251 Nb/AlOx/Nb Josephson junctions (JJs). The correctness of operation is verified by on-chip high-speed testing.

The SFQ logic [1] is a superconductor ultrafast digital circuit technology based on magnetic flux quantization and quantum interference in superconductor rings containing JJs. Binary information is represented by absence or presence of an SFQ in a superconductor ring, and JJs are used as switching elements. An impulse-shaped voltage is generated only when an SFQ travels across a JJ, which leads to ultralow power consumption (~10^{-19} J per switching event) in principle. A superconductor stripline wiring behaves like a lossless waveguide and propagates a mass-free SFQ signal at the speed of light for a long distance (> 10mm). Based on these features, several projects on the SFQ-based VLSI technology have been undertaken in the US and Japan toward high-performance computing in the post-Moore's era.

SFQ signals can be split or merged using a few JJs, and easily stored in large-inductance superconducting rings. Unlike voltage level logic, synchronization of an input signal with the reference signal (clock signal) is essential at a storage ring in each SFQ logic gate in order to distinguish between that the input signal has not arrived yet and that a logic value '0' has arrived. In other words, each storage ring intrinsically has latch or memory functionality, and all the SFQ logic gates are clocked gates except for pulse splitters and mergers. Such a unique feature expands an architectural design space of SFQ circuits that would never be reached with conventional CMOS logic circuits. Especially gate-level, ultradeep pipelining, with target clock frequencies in a range of several tens of GHz, is promising for high-performance circuits, because no additional pipeline registers are required and because operating frequencies are not limited by the power-wall problem but simply determined by the sum of setup and hold times in SFQ logic. Figure 29.3.1 summarizes the SFQ circuit elements used in this work [2], by which 50-to-100GHz operation is possible. However, it has been believed that gate-level pipelining is too difficult to be applied to complex SFQ circuits because of timing design. Several approaches including asynchronous design [3] and bit-serial processing [4] have been proposed to avoid the timing design complexity, even though these approaches result in lower clock frequency or poor processing performance.

In this work, the authors demonstrate an ultimately high-throughput multiplier based on a bit-parallel, gate-level-pipelined structure to show the maximum potential of SFQ logic. Figure 29.3.2 illustrates the block diagram of the designed multiplier. The total number of pipelining stages is 15, and the multiplier is divided into three component circuits, i.e., the partial product generator (PPG), partial product accumulator (PPA), and final stage adder (FSA). In PPG, all partial products (PPs) are generated in parallel by AND gates in one clock cycle. Then, the PPs are processed in PPA, which calculates based on carry-save addition using full adders and half adders. The final two PPs are added by a parallel prefix adder. For successful gate-level pipelining, several D flop-flops are inserted into paths to equalize the logic depth, and multiplication is performed every clock cycle. In layout design, clock signals to all the logic gates are carefully distributed to minimize timing jitter, and clock skews are intentionally added to conceal signal propagation delays, in order to achieve the highest clock frequency. The wire lengths are also precisely controlled, with an accuracy of 10μm, because signal propagation at the speed of light is no longer negligible in SFQ circuit design.

Figures 29.3.3 and 29.3.4 display the micrograph of the chip and the cross-sectional device structure fabricated using a 1.0μm, 9-layer process [5]. In addition to the designed multiplier, there are shift-registers (SRs) and a clock generator (CG) for on-chip testing. The CG generates a high-speed clock signal train around 50GHz, depending on the supplied voltage. In total, 20,251 JJs are integrated on a 6.03mm × 5.22mm area.

The chip is tested in a liquid helium bath at 4.2K. The correctness of the multiplier operation is verified using many test vectors. One of the on-chip high-speed test results is shown in Fig. 29.3.5. The six waveforms at the top correspond to input signals, where an SFQ pulse is generated at each rising edge. The others are outputs, whose signal transitions represent the arrival of SFQ signals. In this test sequence, first, four sets of test vectors are written in the SRs at low frequency (~ 1kHz) from room-temperature electronics (i). Then, the CG is triggered to generate high-speed clock signals, and multiplication is performed on the chip (ii). Finally, the results are read out from the SRs at low speed and verified (iii). Figure 29.3.6 shows operating voltage vs. clock frequency for the component circuits. The maximum clock frequency obtained in measurement is 48GHz with 5.6mW power.

Figure 29.3.7 shows the performance summary and comparison to prior works on demonstrated SFQ multipliers whose bit length is longer than four [3, 4]. Thanks to bit-parallel, gate-level pipelining, the measured throughput achieves 48 giga-operations per second (GOPS), and the resultant efficiency is almost 10×10^{12} operations per second per watt (10 TOPS/W). The authors believe that the successful demonstration of the gate-level pipelined multiplier is the starting point toward tera-FLOPS/W-class, ultimately throughput-oriented high-performance cryogenic computing using the SFQ VLSI technology.

Acknowledgements:
This work is supported by JSPS KAKENHI Grant Numbers JP16H02796, JP18H05211 and JP18H01498. The circuits are designed with the support by VDEC of the University of Tokyo, in collaboration with Cadence Design Systems, Inc. and fabricated in the CRAVITY of National Institute of Advanced Industrial Science and Technology.

References:
[1] K. K. Likharev and V. K. Semenov, "RSFQ logic/memory family: a new Josephson-junction technology for sub-terahertz-clock-frequency digital systems," *IEEE Trans. Appl. Supercond.*, vol. 1, pp. 3–28, Mar. 1991.
[2] Y. Yamanashi, et al., "100 GHz demonstrations based on the single-flux-quantum cell library for the 10 kA/cm^2 Nb multi-layer process," *IEICE Trans. Electron.*, vol. E93-C, pp. 440–444, Apr. 2010.
[3] M. Dorojevets, A. K. Kasperek, N. Yoshikawa, and A. Fujimaki, "20-GHz 8 × 8-bit parallel carry-save pipelined RSFQ multiplier," *IEEE Trans. Appl. Supercond.*, vol. 23, p. 1300104, June 2013.
[4] X. Peng, et al., "High-speed demonstration of bit-serial floating-point adders and multipliers using single-flux-quantum circuits," *IEEE Trans. Appl. Supercond.*, vol. 25, p. 1301106, June 2015.
[5] S. Nagasawa, et al., "Nb 9-layer fabrication process for superconducting large-scale SFQ circuits and its process evaluation," *IEICE Trans. Electron.*, vol. E97.C, pp. 132–140, Mar. 2014.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Cell	JJ count	Clocked Gate	Delay, ps	Setup, ps	Hold, ps
Splitter	3	No	4.3	n/a	n/a
Merger	7	No	8.2	n/a	n/a
D flip-flop	6	Yes	5.1	1.2	-0.9
NOT	10	Yes	9.6	1.2	5.0
AND	14	Yes	7.9	-1.4	2.0
OR	12	Yes	5.0	5.8	-0.2
Exclusive OR	11	Yes	6.5	3.7	4.1

Figure 29.3.1: SFQ library developed for 1.0-μm, 9-layer process.

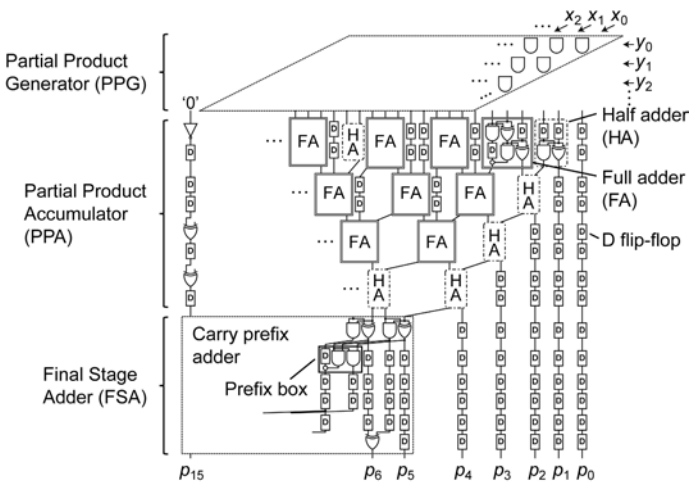


Figure 29.3.2: Schematic diagram of gate-level pipelined multiplier.

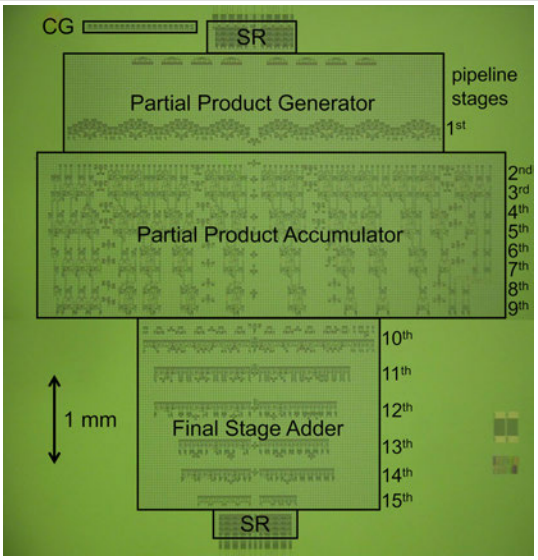


Figure 29.3.3: Die micrograph of multiplier with on-chip test circuitry.

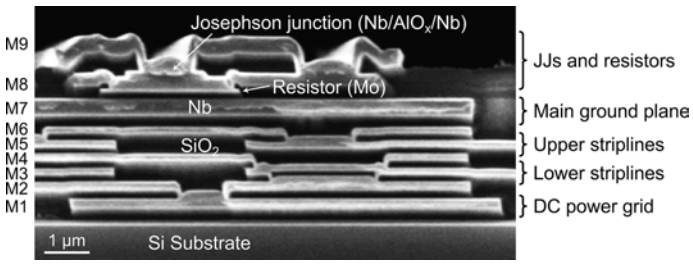


Figure 29.3.4: Cross-sectional scanning electron microscopy image of integrated device.

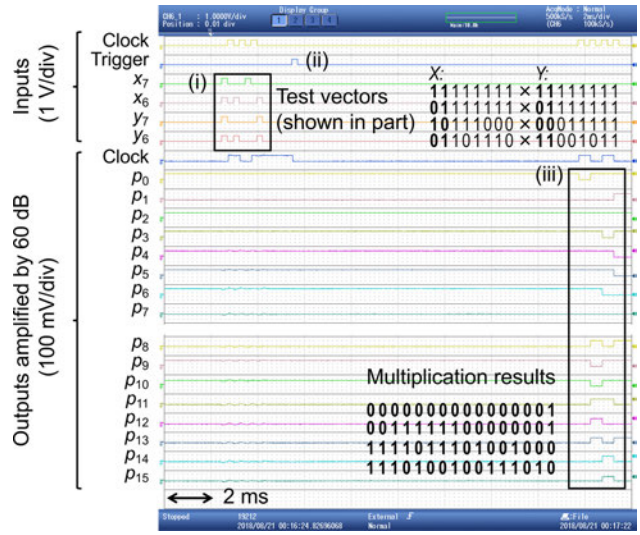


Figure 29.3.5: Oscilloscope waveforms of on-chip high-speed test (oscilloscope screens were captured twice and merged).

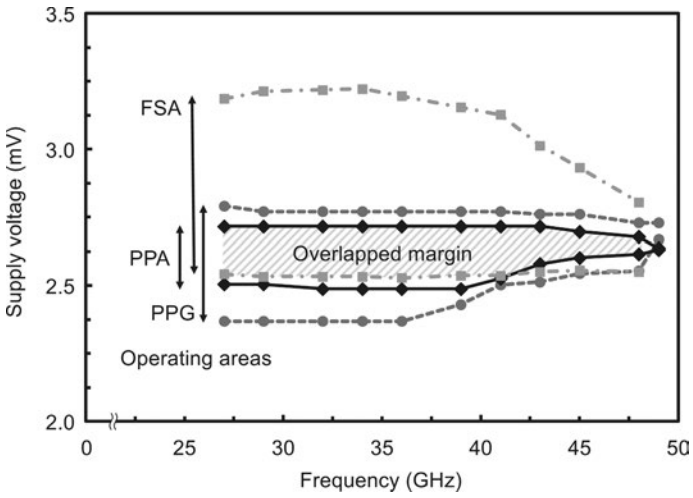


Figure 29.3.6: Frequency dependence of operating margin in supply voltage.

29

	This work	[3] (2013)	[4] (2015)
Process	1.0-μm, 9-layer, Nb/AlO$_x$/Nb JJs		
Bit processing	Parallel	Parallel	Serial
Design scheme	Gate-level pipeline	Asynchronous wave-pipeline	Pipelined systolic array
Operands	8-bit, signed integers	8-bit, unsigned integers	32-bit floating-point numbers
Bit width of product	16 bits	8 bits	32 bits
Clock Frequency, GHz	**48**	20†	59
Throughput, GOPS	**48**	20†	2.4
Power, mW	**5.6**	1.7†	5.8
JJ count	17,488	5,948	18,766
Area, mm^2	5.07 × 5.22	2.50 × 1.45	3.77 × 6.66

† Simulation results

Figure 29.3.7: Experimentally-obtained performance summary and comparison to prior works.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

29.4 Ultra-Low-Power Atomic Clock for Satellite Constellation with 2.2×10^{-12} Long-Term Allan Deviation Using Cesium Coherent Population Trapping

Haosheng Zhang[1], Hans Herdian[1], Aravind Tharayil Narayanan[1], Atsushi Shirane[1], Mitsuru Suzuki[2], Kazuhiro Harasaka[2], Kazuhiko Adachi[2], Shinya Yanagimachi[3], Kenichi Okada[1]

[1]Tokyo Institute of Technology, Tokyo, Japan
[2]Ricoh, Miyagi, Japan
[3]National Institute of Advanced Industrial Science and Technology, Ibaraki, Japan

Nano/micro satellites in low earth orbit (LEO), and unmanned-aerial-vehicle base stations (UAV-BS) in the stratosphere are being considered to be used for increasing the coverage and provision of on-demand high data rates of mobile communication networks all over the globe as beyond 5G technology. One of the most important key technologies for such high-speed and long-distance communication is a very accurate time standard, especially for the LEO satellites constellation [1]. Presently, the best time accuracy can be acquired from atomic clocks. Atomic clock assisted GEO satellites such as GPS can be a primary reference, but they suffer from large path loss and delay, degrading the clock accuracy to 10^{-6} in the receiver part. In addition, GPS is not always available in the space, while the conventional atomic clock has deployment difficulties in the large array due to large volume and huge power consumption. For example, due to the special condition of the atomic cell required for reference frequency locking and probing, even a compact atomic clock ranges from 150cm^3 to 775cm^3 in size and consumes 1.2W-to-10W of power. Thus, a miniaturized, low power and low cost time standard is required for each LEO satellite. Recent developments in photonics and MEMS processes show the potential to realize low-power and small-volume quantum package atomic clock based on a coherent population trapping (CPT) method [2]. With the reference frequency locking and probing techniques realized by advanced CMOS integrated circuits, it is now possible to manufacture a small form-factor atomic clock. This paper presents a complete ultra-low-power and miniaturized atomic clock (ULPAC) system with a cesium-133 gas cell, vertical-cavity surface-emitting laser (VCSEL), temperature/magnetic controllers inside a quantum package and the driving/controlling circuitry required for complete atomic clock operation. The prototype of ULPAC achieves a long-term Allan deviation of 2.2×10^{-12} at $\tau=10^5$s while realizing only 15.4cm^3 volume.

Figure 29.4.1 shows the system block diagram of the proposed ULPAC prototype. ULPAC outputs an accurate 10MHz clock generated by a 10MHz VCXO, which is calibrated by the cesium resonance frequency. 1PPS (1 pulse per second) signal is also generated from the 10MHz clock. The quantum package integrates cesium-133 gas cell, VCSEL, and photo-detector (PD). The D1-line CPT transition of cesium-133 vapor cell exhibits the half of the ground-state hyperfine resonance frequency $\Delta hf/2$ of 4.596315885GHz, which is served as a stable reference. By using a fractional-N PLL, the 10MHz clock is multiplied by 459.6315885 times, and the multiplied LO signal (f_{LO}) is compared with $\Delta hf/2$. The frequency error, $\Delta hf/2 - f_{LO}$, is feedbacked to the VCXO control signal for obtaining the very accurate 10MHz clock.

The LO signal is fed to the VCSEL after amplification. And the VCSEL produces two first-order sidebands separated by $2\times f_{LO}$ around the VCSEL wavelength (λ_{VCSEL}). The bichromatic light from the VCSEL is passed through the $\lambda/4$ waveplate and neutral-density (ND) filter. Then it is input to the cesium gas cell. When the λ_{VCSEL} and $2xf_{LO}$ are exactly matched to the cesium D1 line resonance energy levels, the PD will detect the very sharp and stable peak caused by the CPT. To search the peak, f_{LO} and λ_{VCSEL} are modulated through the PLL ($\pm\Delta f$) and bias current ($+\Delta I$), respectively. The demodulation of this searching and locking loop based on the frequency/current is done by the MCU in this prototype.

Figure 29.4.2 shows the building blocks for the VCSEL biasing and modulation circuitry, and 1PPS generation/calibration circuitry. To search the cesium resonance frequency, a very-fine fractional-N PLL is used. 3×10^{-4}Hz frequency step is realized by 3rd–order DSM and dithering. One of four banks of 35b frequency control word (FCW) is selected for fast frequency modulation with Clk1=71.4Hz. The VCSEL wavelength (λ_{VCSEL}) is controlled through the bias current to capture 894.6nm so a higher SNR of CPT resonance can be obtained. The 16b low-noise and PVT stable current source for VCSEL biasing realizes 0-to-3.2mA range with 0.1μA resolution for -187MHz/μA wavelength sensitivity, and the measured current noise is below 0.07μA at the VCSEL operating point of 1.5mA. The bias current is also modulated with Clk2=9.7kHz. The VCSEL buffer is programmable with 0.2dB step ranging from -5dBm to 5dBm to fine control the VCSEL output power, which is used to obtain an optimum full-width half-maximum (FWHM) of CPT resonance. 1PPS signal is generated from the 10MHz with programmable duty cycle. By using a 0.8ns resolution TDC, the initial phase/frequency offset is calibrated with the primary time standard or GPS signal.

Figure 29.4.3 shows the proposed ultra-small quantum package, integrating cesium-133 gas cell, VCSEL as well as optical filter, PD, heater/thermistor for the temperature control, and magnetic coil for electromagnetic control. During the fabrication of VCSEL, the wavelength is set to 894.6nm for the CPT absorption. However, the VCSEL has a temperature coefficient of -19GHz/K for wavelength, so the temperature control loop by the heater and thermistor keeps gas cell and VCSEL temperature at 80°C degrees within ±1mK resolution in spec. The thermal isolation for the gas cell and VCSEL contributes to reducing the heater power consumption, which can be reduced to only 9.1mW by the proposed quantum package structure. The thermal isolation is realized by a polyimide-film suspension and Au coating for higher thermal resistance. Pressure variation of inert buffer gas used in the gas cell for reducing the collision between the cesium atom and the wall due to temperature changes is mitigated by using a mixed buffer gas of N_2 and Ar with 2.55 ratio, which improves the long-term stability. The CPT frequency also has a magnetic-field sensitivity of ~3.5 kHz/μT, so the magnetic control loop keeps a highly-stable unidirectional 10μT magnetic field for avoiding undesired resonance due to the Zeeman degeneracy.

Figure 29.4.4 shows the measurement results about low-threshold VCSEL, low-power PLL, CPT resonance line width, and frequency stability. The fabricated VCSEL can output 0.5mW light power with 3.1mW power consumption. The measured FWHM of CPT resonance using the D1 line of cesium atom is around 1.5kHz, which corresponds to a qualify factor Q of 6.1×10^6. The low phase noise and low jitter performance are realized by 2.0mW PLL, which contribute to a better short-term stability. The measured short-term (τ =1s) Allan deviation is below 8.4×10^{-11}, and the long-term ($\tau=10^5$s) Allan deviation is below 2.2×10^{-12}.

Figure 29.4.5 plots a performance comparison about Allan deviation and power consumption, and the prototype of the proposed ULPAC achieves very low power and very stable time accuracy satisfying the satellite communication requirement of 10^{-11} long-term Allan deviation.

Figure 29.4.6 summaries the key performances of the state-of-the-art atomic/molecular clocks [3, 4, 5]. Thanks to the advanced quantum packaging, the volume of this work is reduced to 15.4cm^3. The frequency synthesizer only consumes 2.0mW including 1PPS generation, while the VCSEL buffer gives 0dBm output at 4.5mW power consumption. The long-term stability is achieved by the well-controlled electromagnetic filed, minimized pressure variation due to temperature variation of gas cell and VCSEL buffer power variation. The power consumption of VCXO, heater, and regulator are 2.9mW, 9.1mW, and 7.9mW, respectively. The total power consumption is 59.9mW, which is mainly dominated by the MCU in this prototype and can be further reduced by using a custom logic circuitry.

Figure 29.4.7 shows the prototype of this work with the quantum package and the CMOS die micrograph. The volume of the quantum package is 16mm^3. The core area of the chip is 2.55mm^2, which is fabricated in a 65nm CMOS process.

Acknowledgment:
This work is partly supported by New Energy and Industrial Technology Development Organization (NEDO).

References:
[1] A. Guidotti, et al., "Satellite-enabled LTE systems in LEO constellations," *IEEE ICC Workshops*, pp. 876-881, 2017.
[2] C.T.C. Nguyen and J. Kitching, "Towards chip-scale atomic clocks," *ISSCC Dig. Tech. Papers*, pp.84–85, Feb. 2005.
[3] D. Ruffieux, et al., "A low-power fully integrated RF locked loop for miniature atomic clock," *ISSCC Dig. Tech. Papers*, pp. 48–50, Feb. 2011.
[4] C. Wang, et al., "A CMOS molecular clock probing 231.061-GHz rotational line of OCS with sub-ppb long-term stability and 66-mW DC power," *IEEE Symp. VLSI Circuits*, pp. 113-114, June, 2018.
[5] Microsemi, "Chip Scale Atomic Clock (CSAC)," *datasheet*, 2018.
[6] SiTime, "SiT5356 Precision Super-TCXO," *datasheet*, 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

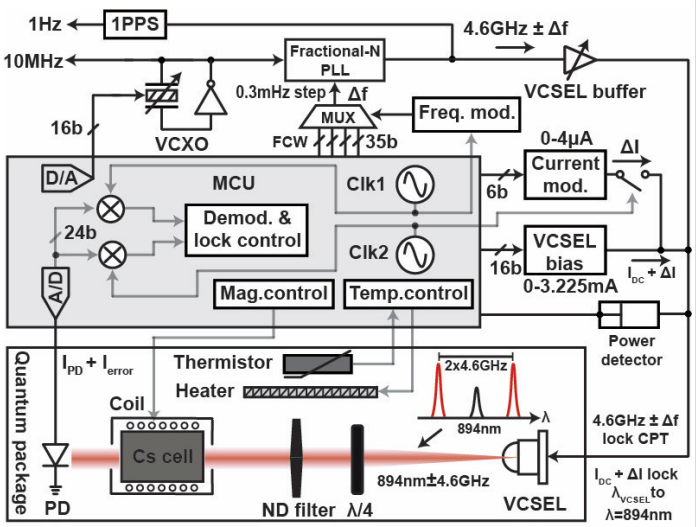


Figure 29.4.1: ULPAC system diagram.

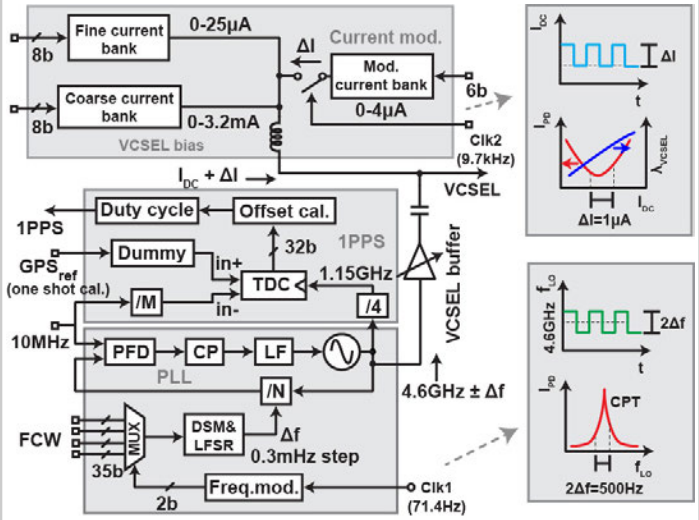


Figure 29.4.2: Block diagram of VCSEL modulator and CPT locking loops.

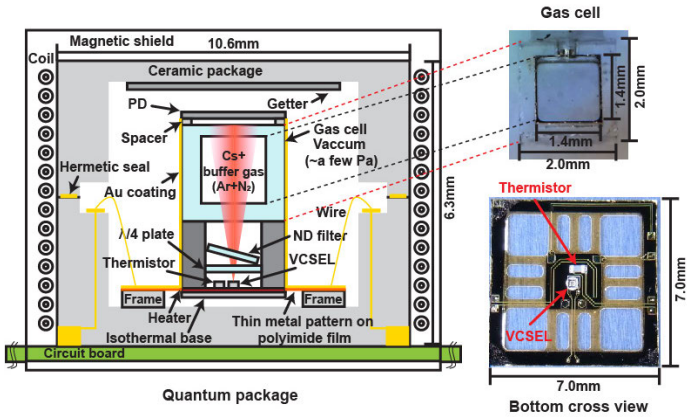


Figure 29.4.3: Integrated optical quantum package consisting of low-threshold VCSEL, optical filter, PD, and cesium-133 gas cell with high thermal isolation.

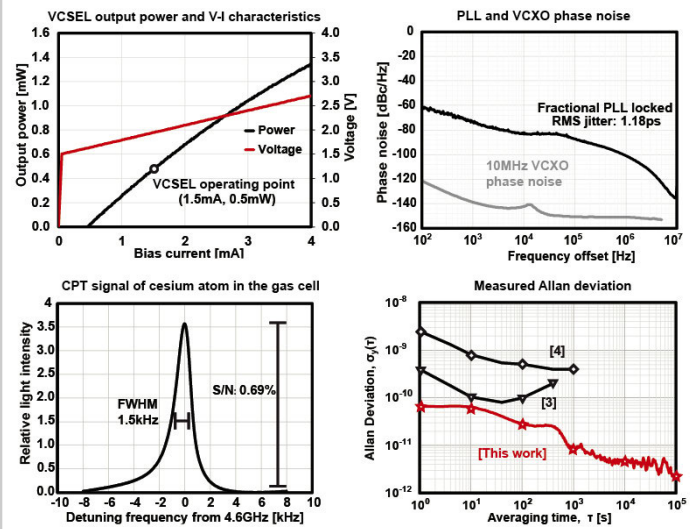


Figure 29.4.4: Measurement of VCSEL characteristics, PLL phase noise, CPT signal and Allan deviation.

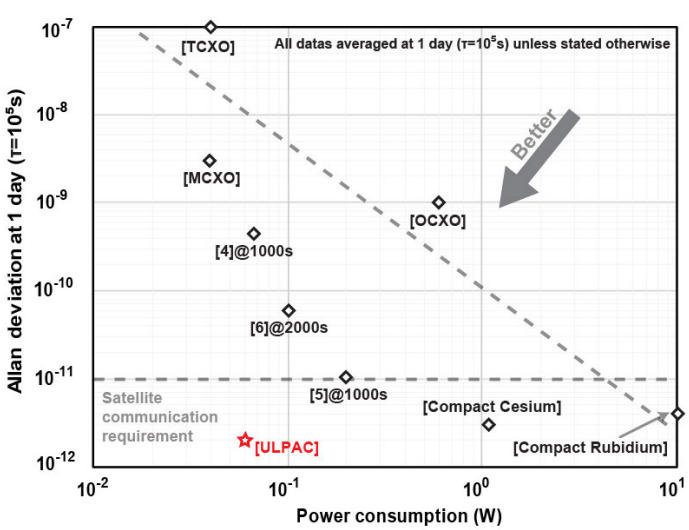


Figure 29.4.5: Long-term stability and power consumption.

Parameter	[3]	[4]	[5]	This work
Integration level &Type	PLL+PA+ Freq. mod. (CSAC)	TX + RX (MC)	PLL+PA+ Freq. mod.+ VCSEL driver (CSAC)	PLL+PA+ Freq. mod.+ VCSEL driver (ULPAC)
Volume [cm³]	nd	50	16	15
Reference freq. [MHz]	40	80	10	10
Atom & Line	Rb & D1	$^{16}O^{12}C^{32}S$ & N/A	Cs & D1	Cs & D1
PLL output freq. [GHz]	3.4	231.0	4.6	4.6
PLL power cons. [mW]	7.8	66.0	52.0	2.0
VCSEL driver power cons. [mW]	14.7@0dBm	N/A	N/A	4.5mW@0dBm 3.1mW@-5dBm
Allan dev. (τ=10⁵s) [10⁻¹¹]	40.0	250.0	30.0	8.4
Allan dev. (τ=10⁵s) [10⁻¹¹]	N/A	N/A	N/A	0.22

Figure 29.4.6: Performance comparison table.

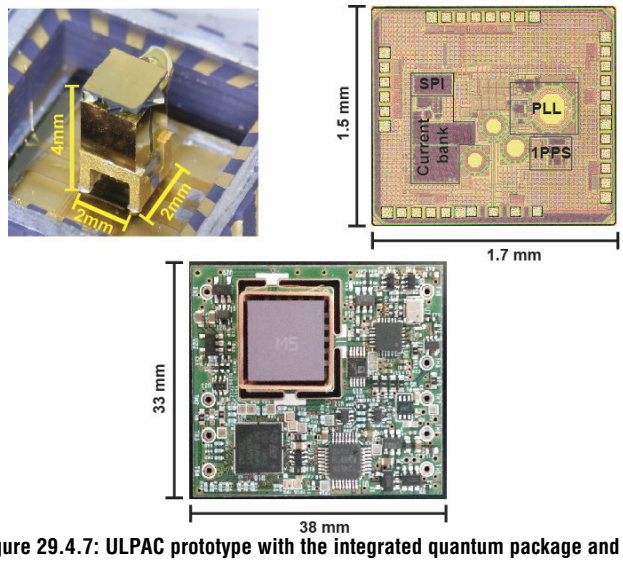


Figure 29.4.7: ULPAC prototype with the integrated quantum package and the CMOS die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

29.5 A Single-Chip Optical Phased Array in a 3D-Integrated Silicon Photonics/65nm CMOS Technology

Taehwan Kim[1], Pavan Bhargava[1], Christopher V. Poulton[2], Jelena Notaros[2], Ami Yaacobi[2], Erman Timurdogan[2], Christopher Baiocco[3], Nicholas Fahrenkopf[3], Seth Kruger[3], Tat Ngai[3], Yukta Timalsina[3], Michael R. Watts[2], Vladimir Stojanovic[1]

[1]University of California, Berkeley, CA
[2]Massachusetts Institute of Technology, Cambridge, MA
[3]Colleges of Nanoscale Science and Engineering, Albany, NY

The realization of a low-cost and robust optical beam-steering platform is a key enabler for a number of applications, including light detection and ranging (LIDAR) and free-space optical communications (FSO). Optical phased arrays (OPAs) have emerged as a promising solution, due to advancements in photonic integrated circuits (PIC) foundry processes, which have enabled high-precision fabrication of PICs with a large number of components [1-3]. In order to meet steering range and directivity requirements in systems such as autonomous vehicles, a clear path to scaling OPAs to millimeter-scale apertures with thousands of tight-pitched antenna elements is critical. As the element count grows, independent phase control for each element becomes crucial since maintaining coherence between elements becomes more difficult due to process variations. Moreover, independent control allows for unique system capabilities, such as multi-beam formation and converging/adaptive beams, which makes OPAs a particularly attractive solid-state beamforming technology.

In the majority of prior OPA demonstrations, only the optical components have been implemented on a PIC chip, with phase-shifting elements driven by discrete electronics [1-3]. Even with electronics integrated on a separate chip, the number and density of I/O connections required for large-element-count OPAs limits the scalability of multi-chip solutions. This issue can be addressed by monolithic integration [4-5], where both photonics and electronics are integrated on a common SOI CMOS substrate. However, monolithic integration limits the photonic device design due to CMOS design rules, available materials and processing steps. Moreover, placement/routing constraints result in a large die footprint for small active array apertures, especially in older process nodes with feature sizes above 100nm [4-5].

In this work, these challenges are overcome by implementing an OPA in a wafer-scale 3D integration platform, which allows for photonics and CMOS electronics to be independently optimized while enabling flexible, dense connections between them. In the designed OPA, a cascaded architecture is used (Fig. 29.5.1). Input light is uniformly distributed to antenna elements from a bus waveguide via cascaded directional couplers, and the relative phase between adjacent elements is thermally adjusted using silicon heaters embedded into the waveguide sections between couplers. By controlling the element phases and tuning the laser wavelength, full two-dimensional beam-steering along antenna (Φ) and lateral (Θ) directions is achieved. Each thermal phase shifter is independently driven by a CMOS controller with digital-to-analog converters (DACs) and on-chip lookup tables (LUTs) storing the beam position codes, enabling rapid steering along arbitrary trajectories.

Figure 29.5.2 shows the 3D heterogeneous platform used to implement the OPA. Photonic devices are fabricated with 193nm immersion lithography on a 300mm SOI wafer, while electronics are implemented using a 65nm bulk CMOS technology. The two wafers are face-to-face oxide-bonded, the silicon handle on the photonic wafer is globally etched away to expose the devices. Through-oxide vias (TOV) are formed to establish the connection between CMOS and photonics and back-metal is placed for pads and TOV-pad connections. TOVs can be densely placed at arbitrary locations with a pitch as small as 7μm, which allows the placement of CMOS drivers directly underneath the heaters and eliminates wiring overhead.

Figure 29.5.3 shows the details of the optical components used in our OPA. The thermal phase shifter is designed to have a resistance of 270Ω and achieves 2π phase shift at 3.3V, making it compatible with thick-oxide devices in the CMOS without cascoding. The optical antenna element is a long waveguide grating formed by fully-etched sidewall perturbations at ~700nm pitch. Long length (500μm) of the antenna enables small beam divergence in Φ and wide steering range by changing the laser wavelength, unlike the short 3.55μm-long gratings with large beam divergence used in a previous monolithic OPA [2]. Furthermore, the antenna perturbation magnitude is apodized to produce a uniform emission pattern and maximize the effective aperture. To further facilitate beam-steering with wavelength tuning, broadband inverse-taper edge-couplers are used to couple light from the fiber to the input waveguide on the side of the chip. Edge couplers have much wider (>>100nm) bandwidth than standard vertical grating couplers (often limited to ~60nm) due to the adiabatic transition.

Figure 29.5.4 shows our 8-bit DAC design for the thermal phase shifters. We utilized 1st order delta-sigma modulator driving a thick-oxide NMOS device as a switch for the heater. The power from the heater following the pulse density modulated (PDM) pattern is then low-pass filtered by the thermal response of the silicon waveguide (~50kHz bandwidth), which is more than 3 orders of magnitude slower than the clock rate (400MHz). As a result, a linear relationship between the input code and the actual temperature of the waveguide was achieved. Since the driver is composed of only a level shifter and a switch, it is compact (22μm×22μm) and can be pitch-matched to the heater (30μm length). The highly digital nature of our DAC design and negligible parasitics of TOV connections enable efficient utilization of the chip area with a floorplan matched to the size of the OPA (Fig. 29.5.7). Furthermore, the CMOS power overhead (~300μW/DAC) is minimal compared to the current-DACs [4-5] where multiple levels of current mirror tree for distributing the current reference amount to high static power consumption.

The far-field intensity measurement results from our 32-element OPA chip are shown in Fig. 29.5.5. Calibration of the DAC codes was achieved by monitoring the beam pattern and updating the LUTs using an optimizer running on a PC. After calibration, the sidelobe suppression was 8.5dB, and the beam full-width at half-maximum (FWHM) was 0.15/0.6° in Φ/Θ. Phase control enabled 16° steering range in Θ direction. Steering of 18.5°in Φ direction was demonstrated by tuning the input laser wavelength across a 120nm range centered around 1550nm. Steering range can easily be further extended by tighter placement of grating antennas, as shown in the previous works with similar antenna design [2-3]. Figure 29.5.6 shows the comparison with the state-of-the-art. Among existing single-chip OPAs, this is the first demonstration of wide-range (>10°) 2D beam-steering, and with the largest active antenna aperture (500μm×130μm = 0.065mm^2) in a compact die area (~0.5mm^2). This enables high, diffraction-limited, lateral resolution (0.26m×1m at 100m range), critical for automotive LIDARs.

Our work is the first single-chip OPA implemented with 3D integration, which allows the photonics to be customized for target LIDAR/FSO applications (e.g. special layers/etches required for antennas, Ge photodetectors and extensions such as hybrid laser/optical gain), while achieving tight connections to underlying CMOS and enabling further OPA scaling.

Acknowledgements:
This work was supported by the Defense Advanced Research Projects Agency (DARPA) of the United States under the E-PHI project, grant no. HR0011-12-2-0007 and DARPA MTO Program Manager Dr. Josh Conway and also in part by W911NF-12-1-0210 under Dr. Jag Shah. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the US Government. We thank the Berkeley Wireless Research Center for continued support.

References:
[1] C. V. Poulton, et al., "High-Performance Integrated Optical Phased Arrays for Chip-Scale Beam Steering and LiDAR," *CLEO*, ATu3R.2, May 2018.
[2] D. N. Hutchison, et al., "High-Resolution Aliasing-Free Optical Beam Steering," *Optica*, vol. 3, no. 8, pp. 887-890, Aug. 2016.
[3] S. A. Miller, et al., "512-Element Actively Steered Silicon Phased Array for Low-Power LIDAR," *CLEO*, JTh5C.2, May 2018.
[4] S. Chung, et al., "A 1024-Element Scalable Optical Phased Array in 0.18μm SOI CMOS," *ISSCC Dig. Tech. Papers*, pp. 262–263, Feb. 2017.
[5] H. Abediasl, et al., "Monolithic Optical Phased-Array Transceiver in a Standard SOI CMOS Process," *Opt. Exp.*, vol. 23, no. 5, pp. 6509-6519, Mar. 2015.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

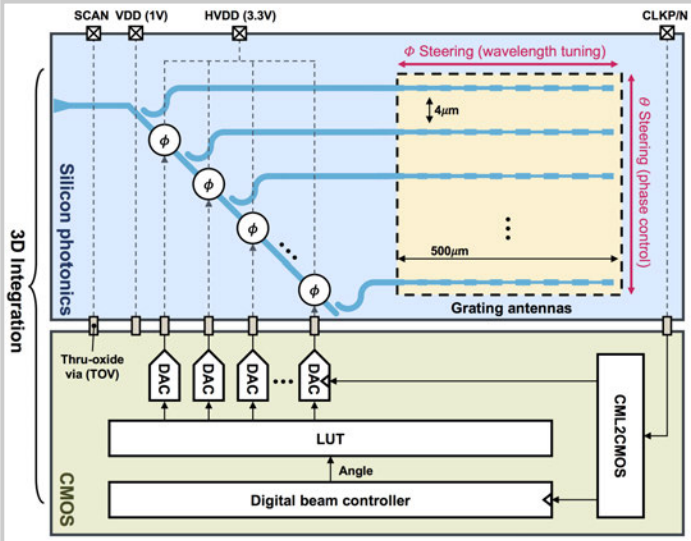


Figure 29.5.1: Optical phased array with cascaded architecture and integrated electronics.

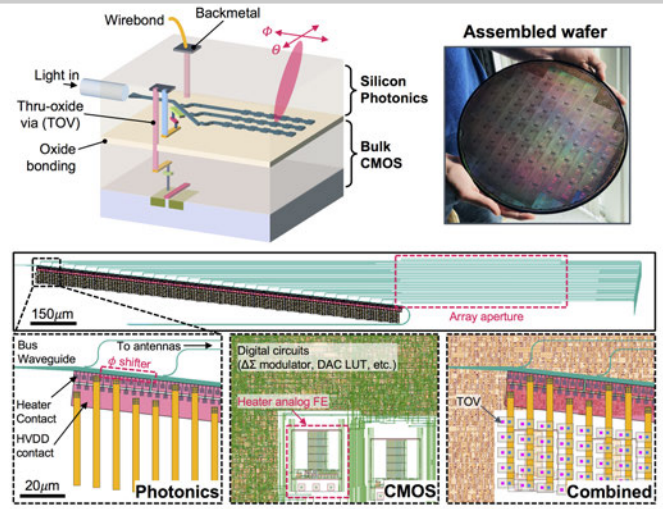


Figure 29.5.2: Overview of the 3D heterogeneous integration platform and vertical placement of CMOS and optical devices for area-efficient implementation.

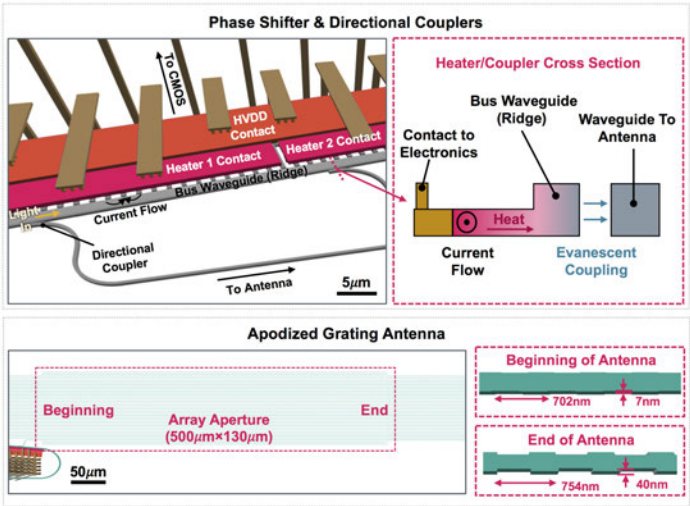


Figure 29.5.3: Details of the optical components, including the thermal phase shifter, directional coupler and waveguide grating antenna.

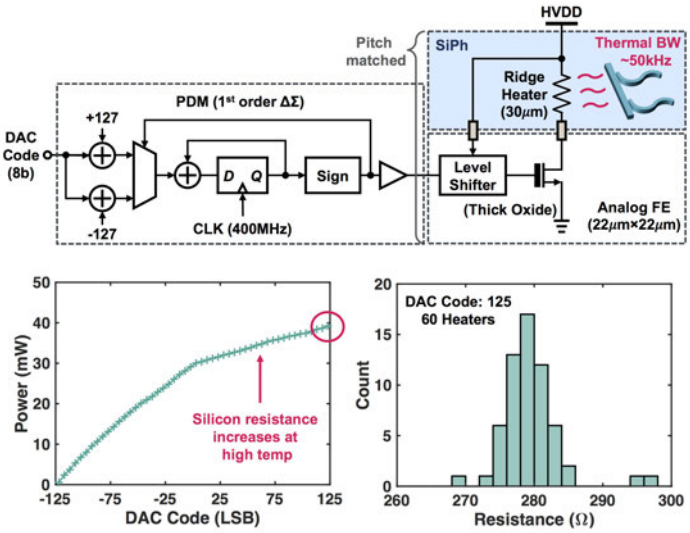


Figure 29.5.4: Compact, linear ΔΣ DAC designed for phase shifter control leveraging low-parasitic through-oxide via 3D connections.

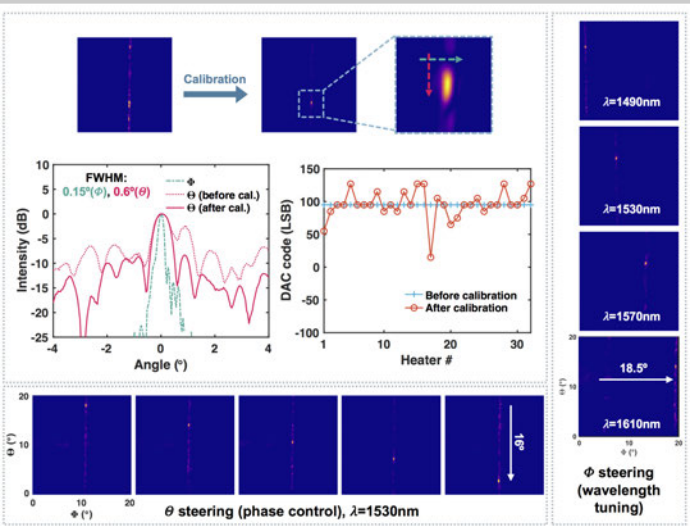


Figure 29.5.5: Beam-steering performance of the implemented optical phased array.

	This work	ISSCC 2017 [4]	Optics Express 2015 [5]	CLEO 2018 [1]	Optica 2016 [2]	CLEO 2018 [3]
Control Electronics	Integrated DACs			External		
Technology	3D-integrated Silicon photonics & 65nm bulk CMOS	0.18μm SOI CMOS	0.18μm SOI CMOS	Silicon photonics	Silicon photonics	Silicon photonics
Steering Dimension	2D (phase/wavelength)	1D (phase)	2D (phase)	2D (phase/wavelength)	2D (phase/wavelength)	2D (phase/wavelength)
Beamwidth (Φ/θ)	0.15°/0.6°	0.03°/17°	0.45°/0.45°	0.04°/0.04°	0.14°/0.14°	0.15°/0.15°
Array Aperture	500μm×130μm	3.55μm×2046μm	23.1μm×23.1μm	2000μm×2000μm	910μm×910μm	1000μm×665μm
Steering Range (Φ/θ)	18.5°/16°	0°/45°*	1.6°/1.6°	15°/20°	17°/80°	14°/70°
Sidelobe Suppression	8.5dB	9dB	12dB	13dB	9dB	8dB
# of Elements/ Independent Phase Controls	32/32	1024/136	64/64**	512/512	128/128	512/512
Phase Shifter Type	Thermal	Thermal	Thermal	Electro-optic	Thermal	Thermal
Required Voltage @ 2π	3.3V	4.9V	5.2V	N/A	N/A	N/A
Tuning Efficiency	20mW/π	34.7mW/π	13.6mW/π	1μW/π	80mW/π	2.6mW/π

*Demonstrated with a separate smaller array implementation of 1/32 size **Excluding amplitude control

Figure 29.5.6: Performance summary and comparison with the state-of-the-art integrated optical phased arrays.

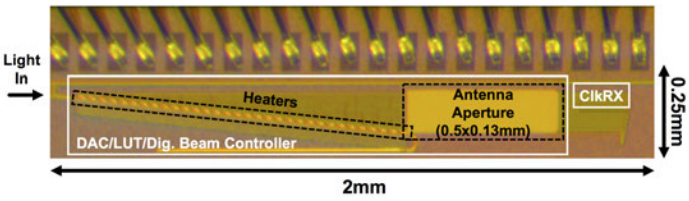


Figure 29.5.7: Die micrograph of the optical phased array fabricated in a 3D-integrated platform.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

29.6 A Digital-Type GaN Driver with Current-Pulse-Balancer Technique Achieving Sub-Nanosecond Current Pulse Width for High-Resolution and Dynamic Effective Range LiDAR System

Yu-Sheng Ma[1], Zong-Yi Lin[1], Yen-Ting Lin[1], Cheng-Yen Lee[1], Tzu-Ping Huang[1], Ke-Horng Chen[1], Yin-Hsi Lin[2], Shian-Ru Lin[2], Tsung-Yen Tsai[2]

[1]National Chiao Tung University, Hsinchu, Taiwan
[2]Realtek Semiconductor, Hsinchu, Taiwan

In automotive applications, short laser pulse I_{LASER} widths through GaN FET control methods and effective assurance of accurate I_{LASER} are major challenges in high-resolution light detection and ranging (LiDAR) systems (top of Fig. 29.6.1). Figure 29.6.1 shows two main laser-diode driver circuits for pulsed LiDAR applications [1]. A well-controlled pulse that accepts the parasitic inductance L_{stray} simplifies control in capacitor discharge driver circuit [2]. But, the fixed pulse shape of the I_{LASER} results in an ineffective modulation of the pulse width t_P and limits the distance resolution l_P (left of Fig. 29.6.1). Under a high pulse repetition frequency (PRF), incomplete pulses and insufficient power will reduce the effective detection range since the V_{BUS} is near hundred volts during the off time of each command cycle $T_{command}$. Although the FET control driver [3] with a low supply voltage V_{supply} allows complex command sequences, the structure needs a large inductor current in the front-end DC/DC converter throughout the $T_{command}$ to ensure sufficient power and high dv/dt transition at V_{BUS}, which means extremely low driving efficiency. To further increase the resolution, a lower V_{BUS} is applied and the parasitic capacitance gain at the drain V_D of the GaN FET is smaller. Meanwhile, chip scale packaging or laser diode mounting minimizes L_{stray} and significantly reduces rise time t_R and fall time t_F, achieving sub-nanoseconds t_P and tens of centimeters distance resolution (right of Fig. 29.6.1). Under high speed $V_{command}$, the gate resistor prevents the device being damaged [4], but it limits the switching frequency and increases losses. The adaptive triple-slope gate driver [5] is not free from process, voltage, and temperature (PVT) variations and the falling slope is indeterminate. If a short pulse width is applied to [4] [5], a slight imbalance between t_R and t_F will cause the average laser current $I_{LASER,\ Avg}$ to change, which changes V_{BUS} and distorts I_{LASER}, and then DC/DC takes a long time to regulate V_{BUS}. Therefore, this paper proposes a digital-type GaN driver with a laser-diode-peak-current-correction (LDPCC) loop and a current-pulse-balancer (CPB) loop for LiDAR systems. Even with parasitic resistance changes in the power path, the LDPCC loop ensures a constant I_{LASER} by adjusting the V_{BUS} over a wide range of PRF (maximum 200MHz) to enhance pulse-to-pulse reliability. In addition, with the proposed asynchronous-binary-driver (ABD), the CPB loop optimizes driver speed, and reduces pulse width to 0.9ns. To prevent V_{BUS} transient, the CPB balances I_{LASER} rise and fall times in tens of nanoseconds.

Figure 29.6.2 depicts the proposed GaN driver with the LDPCC and CPB dual control loops. If the path resistance is R_{path} (summation of laser diode resistance R_{LD}, GaN series resistance R_{ON} and parasitic bonding resistance $R_{bonding}$) and varies with PVT, the I_{LASER} will deviate from its rated value at a fixed V_{BUS}. Therefore, the LDPCC adjusts the V_{BUS} to ensure a constant peak current I_{peak} (left center of Fig. 29.6.2). However, the pulse width of laser current is around a nanosecond, which is difficult to implement by a current sensor. Instead, the I_{peak} is evaluated with I_{target} by $V_{command}$ and I_{sen} (current sensing information of boost converter) in the proposed LDPCC technique. The relationship between I_{peak} and the average current $I_{sen,DC}$ is illustrated by the equation $I_{peak}*t_{command}=I_{sen,DC}*T_{command}$ (center of Fig. 29.6.2), where $I_{sen,DC}$ is the average of peak current $I_{sen,H}$ and valley current $I_{sen,L}$, which is obtained by sampling I_{sen} when the power MOSFET MP and MN turn on respectively (lower of Fig. 29.6.2). Furthermore, the '$t_{command}/T_{command}$' represents the duty D_C of $V_{command}$, and the D_C is derived from the ratio of V_d and V_{IN} by the equation $D_C=V_d/V_{IN}$, where V_d is the DC voltage of $V_{command}$. When the boost converter is in steady state, I_{peak} compares with the reference current I_{target}. The compared result U/D decides the up-tuning or down-tuning of the reference voltage V_{ref} in the reference correction module. For example, with a large R_{path}, I_{peak} is below the I_{target}, and U/D is set to logic 'L' to up-tune V_{ref}. Conversely, with a small R_{path}, U/D is set to logic 'H' for down-correction until I_{peak} is close to I_{target} (right of Fig. 29.6.2). Without any extra analog-to-digital converters, the peak of I_{LASER} is regulated by the proposed LDPCC technique.

Figure 29.6.3 shows the implementation of the CPB technique. Under short t_P operation for high distance resolution, it is necessary to minimize t_R and t_F. Simultaneously, the CPB balances t_R and t_F to avoid large I_{LASER} variations according to $\Delta I_{LASER,Avg} \propto (t_F - t_R)/t_P$ as shown in the lower left of Fig. 29.6.3 ($\Delta I_{LASER,Avg}$ represents variations of average laser current). The imbalance detection stage applied in the CPB checks t_R and t_F by charging and discharging V_{comp} in each $V_{command}$ cycle. When t_R is longer, the discharging time decided by the gate voltage V_G is shorter than the charging time controlled by $V_{command}$, making V_{comp} higher than initial state V_{set}. The comparator output Q_0 becomes a logic 'L' to set Rise[2:0] to a maximum value, where Rise[2:0] indicates the driving capability. In the next $V_{command}$, Rise[2:0] is sent to the driver stage to drives GaN with large driving current I_G. To prevent reverse current caused by the V_S spike during GaN turn-on transition, the V_D of GaN is checked at the rising edge of V_G. If a voltage spike occurs at V_D, the signal R_max is triggered to decrease Rise[2:0] for laser diode protection and optimized t_R. Moreover, the signal Fall[2:0] increases according to logic 'H' of Q_0 to speed up falling transition, and both Rise[2:0] and Fall[2:0] are locked till the V_{comp} is close to V_{set}. Likewise, under fast turn-off transition, the parasitic inductance L_S induces ringing at V_{GS} and causes unintended turn-on of GaN, so the Fall[2:0] is adjusted by F_max to prevent large under-damping at V_G.

Figure 29.6.4 shows the proposed ABD circuit with rising-driving-cell[7:0] and falling-driving-cell[7:0]. The parasitic inductance L_G and L_S representing the bond-wire components on the printed circuit board (PCB) circuit, while C_{GD}, C_{GS}, and C_{DS} are the parasitic capacitances of GaN. In the turn-on transition, the small Rise[2:0] drives the V_G with a smooth I_G, causing the V_G to rise slowly and have a long t_R. In contrast, the large Rise[2:0] drives V_G with a sharp I_G, while the dv/dt couples V_S through the C_{GS}, inducing a reverse I_{LASER}. Similarly, in turn-off transition, the small Fall[2:0] takes a long time to pull the V_G down below the GaN threshold voltage, so the turn-off is delayed. However, di/dt caused by the large Fall[2:0] produces a negative voltage on L_S, and the underdamped ringing of V_{GS} above the threshold voltage causes shoot-through I_{LASER}. To optimize turn-on and turn-off transition, the CPB adjusts the driving capability by asynchronously turning on the PMOSFETs in binary ascending order at 2^3 units in a short period of time. When $V_{commamd}$ is triggered, Rise[2:0] and Fall[2:0] from the CPB are decoded by signals V_H and V_L into signals V_{R_drive}[7:0] and V_{F_drive}[7:0], respectively. The order of the logic "H" in V_{R_drive}[7:0] and V_{F_drive}[7:0] determines which driving-cell should be turned on first. For example, if V_{R_drive}[7:0] is 7b'00000111, the rising-driving-cell[2:0] are set first, and the output V_G[2:0] turns on the PMOSFETs at 2^0, 2^1, and 2^2 units. After that, the signal V_{GB}[2] triggers the next driving-cell[3] at 2^3 units. Finally, the last cell at 2^7 units fully turn on GaN.

The test chip was fabricated in a 55nm process. In the top of Fig. 29.6.5, the proposed LDPCC technique demonstrates V_{BUS} against R_{path} variations to ensure a constant 28A I_{peak}. With the proposed CPB, the t_R and t_F of laser current is balanced in 6 cycles under a 200MHz PRF operation (bottom of Fig. 29.6.5). Fig. 29.6.6 shows the measured gate voltage V_G of GaN and laser current I_{LASER} with different driving capability. The proposed ABD drives GaN with a minimum value of Rise [2:0] and Fall [2:0], the long t_R delays the pulse and t_F prolong the pulse width. Conversely, a maximum value of Rise[2:0] induces reverse current and a maximum value of Fall[2:0] causes a shoot-through current. With the help of CPB, both turn-on and turn-off transitions are optimized to around 0.2ns, and the pulse width is around 0.9ns. Compared to [2] and [3], the proposed digital-type GaN driver has the ability against path resistance variations under a wide range of PRF by the LDPCC, and a high current pulse can be realized under low bus voltage. Furthermore, the adjustable driving capability controlled by the CPB ensures a high slew rate and balances the rising and falling transition compared to [4] and [5]. The die micrograph is shown in Fig. 29.6.7.

Acknowledgements:
The authors would like to thank Hann-Huei Tsai and Ying-Zong Juang for many useful discussions of high voltage process. This work was supported and funded by the National Chip Implementation Center (CIC), Taiwan, ROC.

References:
[1] J. Glaser, "How GaN power transistors drive high-performance Lidar: generating ultrafast pulsed power with GaN FETs," *IEEE Power Electronics Magazine,* vol. 4, no. 1, pp. 25-35, March 2017.
[2] J. Glaser, "High Power Nanosecond Pulse Laser Driver Using an GaN FET," *PCIM Europe*, pp. 662-669, June 2018.
[3] E. Abramov, et al., "Low voltage sub-nanosecond pulsed current driver IC for high-resolution LIDAR applications," *APEC*, pp. 708-715, 2018.
[4] X. Ke, et al., "A 3-to-40V 10-to-30MHz automotive-use GaN driver with active BST balancing and V_{SW} dual-edge dead-time modulation achieving 8.3% efficiency improvement and 3.4ns constant propagation delay," *ISSCC Dig. Tech. Papers*, pp. 302-303, Feb. 2016.
[5] X. Ke, et al., "A 10MHz 3-to-40V VIN Tri-Slope Gate Driving GaN DC-DC Converter with 40.5dBμV Spurious Noise Compression and 79.3% Ringing Suppression for Automotive Applications," *ISSCC Dig. Tech. Papers*, pp. 430-431, Feb. 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

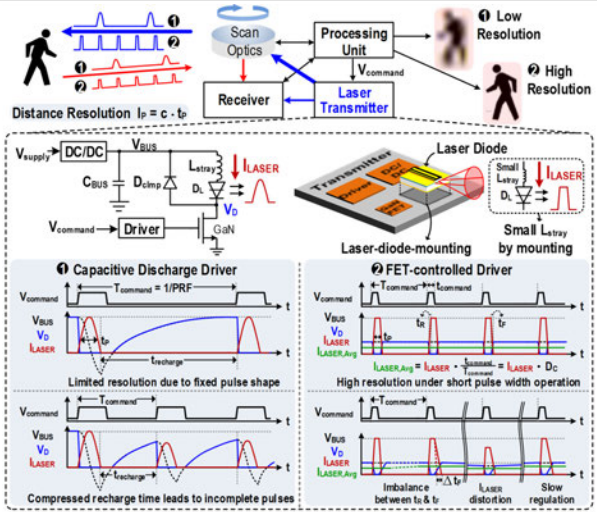


Figure 29.6.1: Structure of capacitor discharge driver circuit and FET control driver circuit in LiDAR system and the corresponding operation waveforms (bottom).

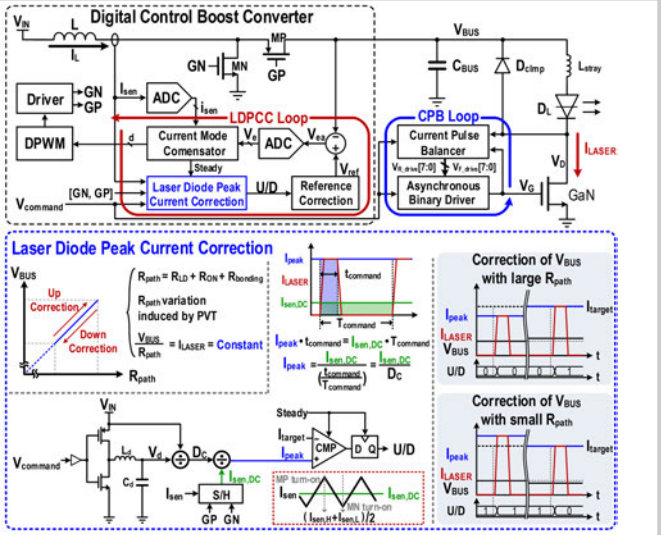


Figure 29.6.2: Proposed LDPCC and CPB loops, and implementation of the LDPCC circuit with operation.

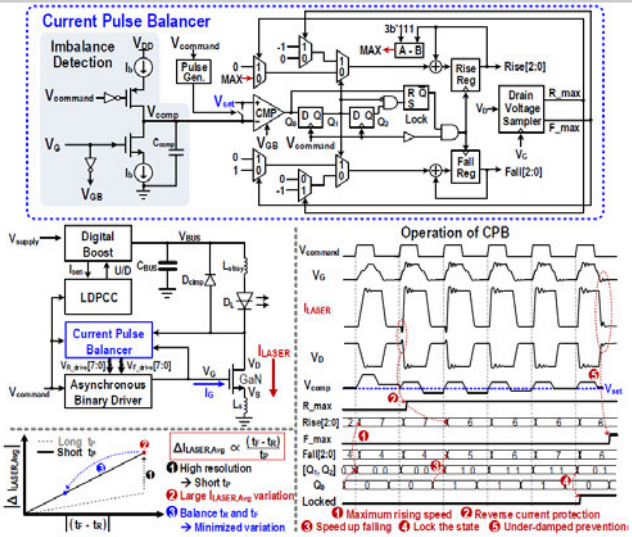


Figure 29.6.3: Circuit Implementation of the CPB with key operation waveforms and timing diagram.

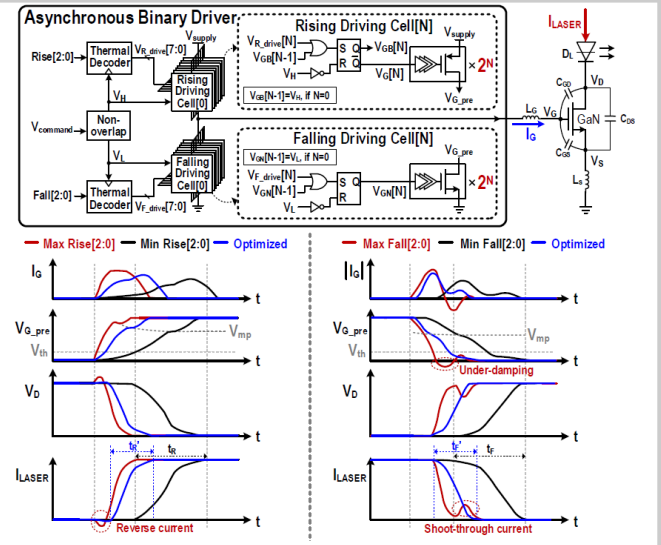


Figure 29.6.4: Implementation of the proposed asynchronous-binary-driver during GaN turn-on and turn-off transition.

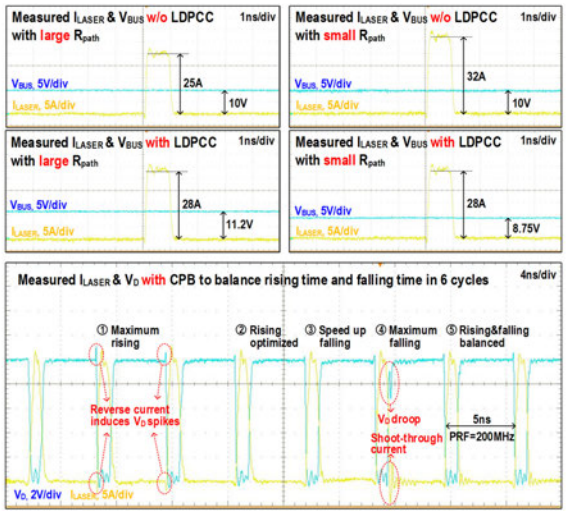
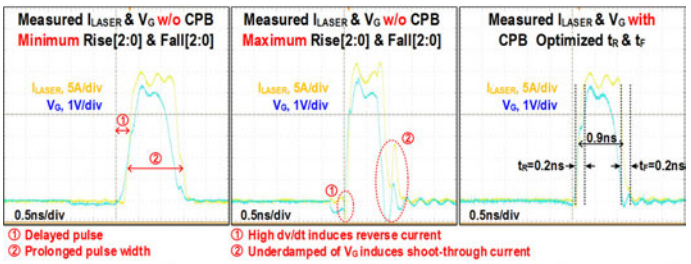


Figure 29.6.5: Measured results of the adjusted V_{BUS} for achieving constant peak current with the LDPCC, and transient performances with the CPB under a 200MHz PRF.



	[2] PCIM Europe'18	[3] APEC'18	[4] ISSCC'16	[5] ISSCC'17	This Work
Process	N/A	180nm	0.35um HV	0.35um HV	55nm
Transistor Type	GaN	Customized LDMOS	GaN	GaN	GaN (EPC2040)
Laser Diode Driver Type	Capacitive Discharge	FET Controlled	N/A	N/A	FET Controlled
Pulse Repetition Frequency (PRF)	10MHz(1)	20MHz	10MHz-30MHz(2)	10MHz± Δ fr/2(2)	200MHz(3)
Bus Voltage	75V	18V	3V-40V(4)	3V-40V(4)	10V
Peak of Laser Current Pulse	26A	5A	N/A	N/A	28A
Width of Current Pulse	1.8ns	3.5ns	N/A	N/A	0.9ns
Rising Slew Rate	75V/1.3ns	18V/1.4ns	40V/1.5ns	40V/2.2ns (Controlled)	10V/0.2ns
Falling Slew Rate	75V/1.4ns	18V/2.5ns	40V/1.2ns	N/A	10V/0.2ns
Gate Driving Control	N/A	N/A	Dual SR Control	Adaptive Tri-Slope Gate Control	Asynchronous Binary Driving Control

(1) 10MHz is mentioned in reference [1]. (2) Switching frequency in reference [4] and [5].
(3) Dynamic PRF is allowable in this work, 200MHz is maximum frequency. (4) Input voltage in reference [4] and [5].

Figure 29.6.6: Measurement of V_G and I_{LASER} with different driving capability and performance comparison.

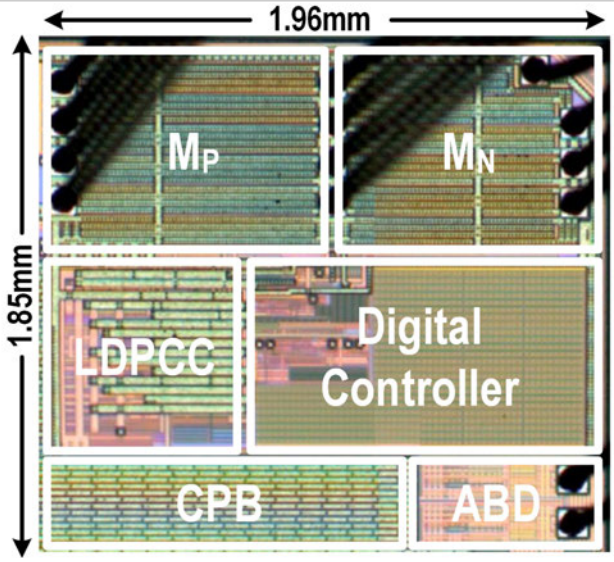


Figure 29.6.7: Die micrograph in 55nm process.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

29.7 A 500Mb/s -46.1dBm CMOS SPAD Receiver for Laser Diode Visible-Light Communications

John Kosman[1,2], Oscar Almer[1], Tarek Al Abbas[1], Neale Dutton[2], Richard Walker[3], Stefan Videv[1], Kevin Moore[2], Harald Haas[1], Robert Henderson[1]

[1]University of Edinburgh, Edinburgh, United Kingdom
[2]STMicroelectronics, Edinburgh, United Kingdom
[3]Photon Force, Edinburgh, United Kingdom

III-nitride laser diodes (LDs) are promising sources for light fidelity (LiFi) networks, underwater wireless optical communications (UWOC), and plastic optical fiber (POF) communications [1] due to their high modulation bandwidths (>5GHz) compared to LEDs (<20MHz). Their narrow linewidths also enable robust free-space Gb/s LiFi links in the presence of high intensities of sunlight through selective spectral filtering. Fully integrated CMOS visible light communications (VLC) receivers have recently been developed to enable miniaturised and low cost Gb/s LD-based links [2]. Sensitivity of these devices is constrained by electrical noise sources such as thermal, shot or excess noise related to the employment of PIN photodiodes (PDs) or linear avalanche photodiodes (APDs) and their amplification circuits. The extremely high gain of SPADs operating in Geiger mode allows quantum sensitivity limits to be approached [3, 4]. A SPAD receiver (RX) for fiber optic applications achieved 200Mb/s at 6.5×10^{-3} BER within 24dB of the quantum limit [3]. In this paper, we demonstrate a fully integrated CMOS SPAD RX SoC extending this data rate by 2.5× to 500Mb/s, whilst improving sensitivity to -46.1dBm, reducing the margin to the quantum limit to only 15dB. Our RX architecture permits massively parallel (4096) photon event summation to be achieved at a high fill-factor (43%) and sample rate (800MHz). Detector redundancy obviates the requirement on current RX implementations that the SPAD dead time be matched to the symbol period to achieve the maximum data rate [3, 4, 5]. Another advantage is that complex modulation schemes such as OFDM or PAM can be applied for high spectral efficiency and multipath interference mitigation. We demonstrate the RX in a practical, background insensitive VLC link at 1m in 1klx ambient conditions using a 450nm LD. The higher power consumption and dead time pile-up nonlinearity of the SPADs at high signal levels, leads us to conclude that the practical use of this device to be in assistance (rather than replacement) of existing APD or PIN RXs for LiFi, UWOC and POF applications towards extended link range or maintaining a low rate link in highly scattering environments.

The 2.8mm by 2.6mm sensor is integrated in 130nm CMOS imaging technology and incorporates 64 × 64 receiver elements at 21μm pitch. Each receiver element contains a single p-well/deep-n-well SPAD biased at 15.2V (1.3V excess bias) with a dead time of 12ns, a median dark count rate of 6kHz at room temperature and a photon detection probability of 37% at 450nm. The 1.34 × 1.34mm active area is chosen for ease of alignment to POF or VLC optics but can be adjusted electronically by disabling areas of receiver elements. Figure 29.7.1 shows a simplified schematic of the chip architecture. A receiver element (RE) comprises a SPAD interfaced to a NMOS passive quench, enable SRAM and toggle-flop. The toggling RE output encodes photon events on both rising and falling edges. A row XOR tree combines 32 RE outputs in an asynchronous DDR sequence at a maximal rate of around 900MHz (1.8Gphotons/s) limited only by wiring parasitics (not SPAD dead time). The RE array divides into two sets of 64-row XOR trees feeding digital readout chains placed on the left and right of the active area. Figure 29.7.2 shows the row parallel interface circuit, consisting of three 8b ripple counters, sampling and converting the asynchronous DDR row input signal to a synchronous binary count without dead time. The three counters are operated in a round robin fashion, so that in every sampling clock cycle one counter is reset, one is being read out and one is counting. A local state machine rotates continuously around the counters. The 128 sets of row counters operate in parallel and their outputs are added through a pipelined 7-stage adder tree to give an overall, 16b synchronous sum of SPAD events. The entire digital readout operates from a sample clock generated by an on-chip PLL with programmable frequency up to 800MHz and distributed through clock trees to the pipelined adder. This digital readout replaces the TIA, analog signal conditioning and ADC chain of conventional APD/PD based receivers and will scale favorably to advanced nanometer process nodes. It also lends itself to the integration of the DSP required for complex modulation schemes.

After digital gain control and formatting, the 16b sum is communicated off-chip using two D-PHY transmission blocks, each dual lane, with line rates of 400Mb/s, for a total off-chip bandwidth of 1.6Gb/s. In addition to the D-PHY blocks, an 8b 400MHz bandwidth current-steering DAC regenerates an analog output compatible with existing ADC-based receive paths. The total RX power consumption is 115mW comprising 15mW of SPAD power and 100mW of DSP. We generate PAM and DC biased optical OFDM (DCO-OFDM) waveforms with random data and transmit these using an Agilent 81180A Arbitrary Waveform Generator (AWG) driving an LD (Thorlabs PL450B, 450nm). The AWG sampling rate is 4GS/s and the LD is collimated onto the SPAD RX at 1m. All experiments are conducted in a background illuminance of 1klx with a 450nm optical bandpass filter (10nm full width half maximum (FWHM)) and two neutral density filters (31% and 10% transmission) placed above the array to attenuate the ambient light and LD power respectively. After synchronization, decoding of the received waveform is performed offline.

At 400Mb/s OOK, 1.8×10^{-3} BER is reached at -49.9dBm optical power. The detected signal is processed by a recursive least squares equalizer in Matlab, which is tuned with a training sequence at the start of the data frame. Equalization (EQ) has marginal effect on performance at this rate, decreasing the BER from 3×10^{-3} to 1.8×10^{-3}. The OOK histogram (Fig. 29.7.3a) shows approximate Poisson distributions with variances equal to their mean. The lowest peak is the actual '0' probability density function and the secondary peak is caused by ISI due to spurious counts from SPAD afterpulsing (<1%) and XOR tree jitter (0.5ns FWHM). The '1' distribution has an average of 23.1 photons detected per bit and so the RX is operating at around 10% of peak capacity of 230Gphotons/s. Figure 29.7.3a also shows a raw eye diagram captured from the DAC output signal with a LeCroy 735Zi oscilloscope via a ZS4000 active probe. The received 4-PAM histogram (Fig. 29.7.3b) has overlapping between adjacent symbols, which was negligible in the OOK histogram. BER degradation occurs earlier (at -45dBm) than the SPAD transfer curve (Fig. 29.7.4a) would indicate because of transient dead time distortion of the signal at high photon rates, which reduces SNDR. We attain 500Mb/s 4-PAM at 2×10^{-3} BER and -46.1dBm sensitivity (Fig. 29.7.4b). EQ improves BER from 7.6×10^{-3} to 2×10^{-3}. We also apply adaptive modulation DCO-OFDM with 512 subcarriers and bit loading in each of the narrowband channels to attain 350Mb/s and 3.7b/s/Hz mean spectral efficiency. Figure 29.7.5 is the frequency response of the SPAD RX determined through channel estimation. At low frequencies, where there is less input signal attenuation, the effective SNR is more impaired from the total power. This distortion is due to dead time nonlinearity at higher incident photon rates. SNR (18dB peak near DC) is sufficient to convey 32-QAM and 16-QAM with the lower frequency subcarriers, down to BPSK at subcarrier index 416. We attain a peak spectral efficiency of 5.6b/s/Hz at a reduced rate of 200Mb/s. Figure 29.7.6 shows a comparison of visible light RXs and quantum limits for 450nm and 635nm wavelengths at 2×10^{-3} BER. Our sensitivities from 100Mb/s OOK to 500Mb/s 4-PAM all outperform that of APD and SPAD RXs and our maximum bit rate is improved by a factor of 2.5 compared to [3] whilst maintaining 2×10^{-3} BER suitable for FEC with minimal redundancy. Although the RX energy per bit is not lower than PD/APD RXs, these do not include an ADC function, which is inherently realised by this SPAD RX architecture. Our value of 230pJ/b is the best of SPAD RXs and could readily be reduced by over an order of magnitude by adopting advanced CMOS nodes. Figure 29.7.7 is an annotated micrograph of the manufactured chip.

Acknowledgements:
This work was supported by the Engineering and Physical Sciences Research Council through the Ultra-Parallel Visible Light Communications Project under Grant EP/K00042X/1. We are grateful to STMicroelectronics for chip fabrication.

References:
[1] C. Lee, et al., "4Gb/s Direct Modulation of 450nm GaN Laser for High-speed Visible Light Communication", *Opt. Express*, vol. 23, no. 12, pp. 16232-16237, 2015.
[2] B. Fahs, et al., "A 12m 2.5Gb/s Lighting Compatible Integrated Receiver for OOK Visible Light Communication," *Journal of Lightwave Tech.*, vol. 34, no. 16, pp. 3768-3775, Aug. 2016.
[3] B. Steindl, et al., "Single-Photon Avalanche Photodiode Based Fiber Optic Receiver for Up to 200Mb/s," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, no. 2, pp. 1-8, Mar. 2018.
[4] E. Fisher, I. Underwood and R. Henderson, "A Reconfigurable Single Photon-counting Integrated Receiver for Optical Communications," *IEEE J. Solid-State Circuits*, vol. 48, no. 7, pp. 1638-1650, July 2013.
[5] L. Zhang, et al., "A Comparison of APD and SPAD-Based Receivers for Visible Light Communications," *Journal of Lightwave Tech.*, vol. 36, no. 12, pp. 2435-2442, June 2018.
[6] D. O'Brien, et al., "High-speed optical wireless demonstrators: Conclusions and future directions," *Journal of Lightwave Tech.*, vol. 30, no. 13, pp. 2181-2187, July 2012.
[7] T. Jukic, B. Steindl and H. Zimmermann, "400μm Diameter APD OEIC in 0.35μm BiCMOS," *IEEE Photon. Tech. Lett.*, vol. 28, no. 18, pp. 2004-2007, Sept. 2016.
[8] P. Brandl, R. Enne, T. Jukic, and H. Zimmermann, "OWC using a fully integrated, highly sensitive optical receiver with large-diameter APD," *IEEE Photon. Tech. Lett.*, vol. 27, no. 5, pp. 482-485, Mar. 2015.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

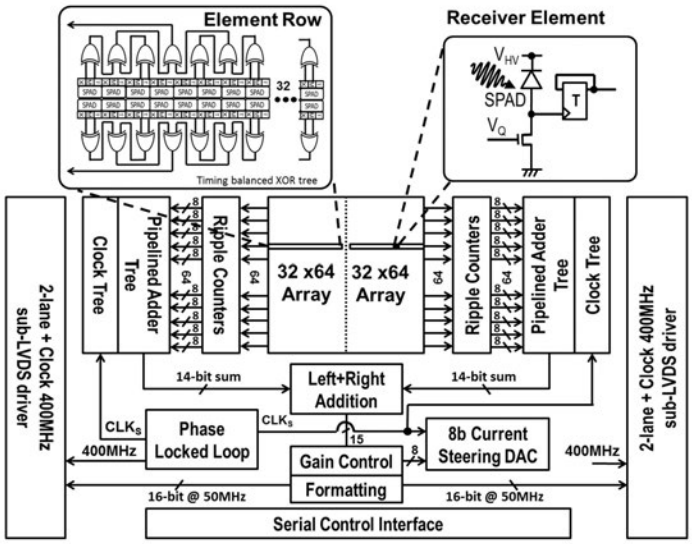


Figure 29.7.1: Simplified chip block diagram.

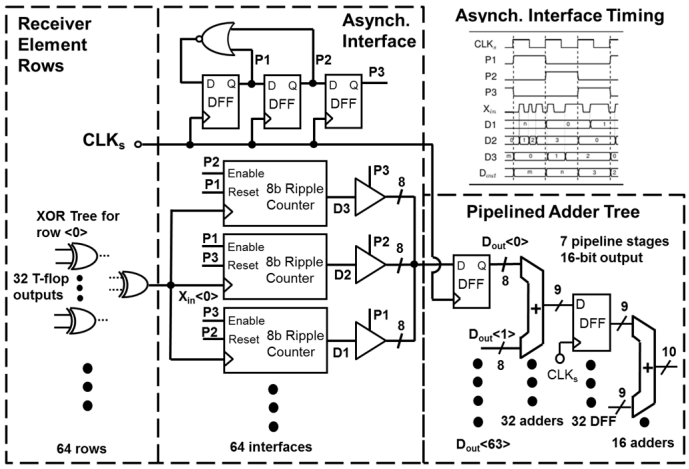


Figure 29.7.2: Asynchronous interface between (right) SPAD XOR tree and pipelined added tree.

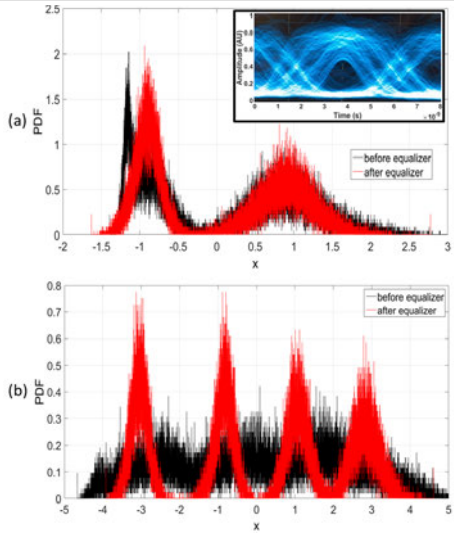


Figure 29.7.3: (a) OOK histogram before and after EQ at 400Mb/s and eye diagram (before EQ) (b) 4-PAM histogram before and after EQ at 500Mb/s.

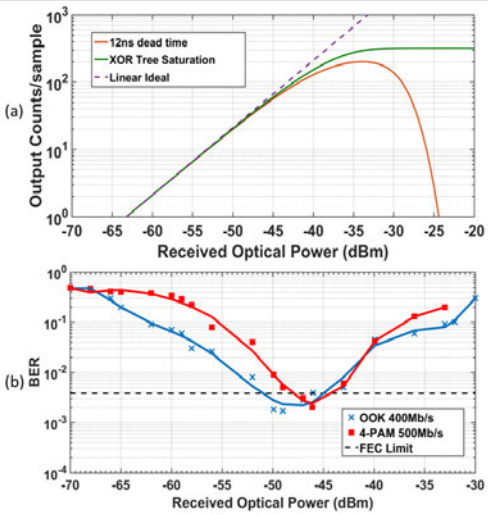


Figure 29.7.4: (a) Simulated photon transfer curve at 400MHz RX sampling rate with received power (b) BER against received power for 400Mb/s OOK and 500Mb/s 4-PAM.

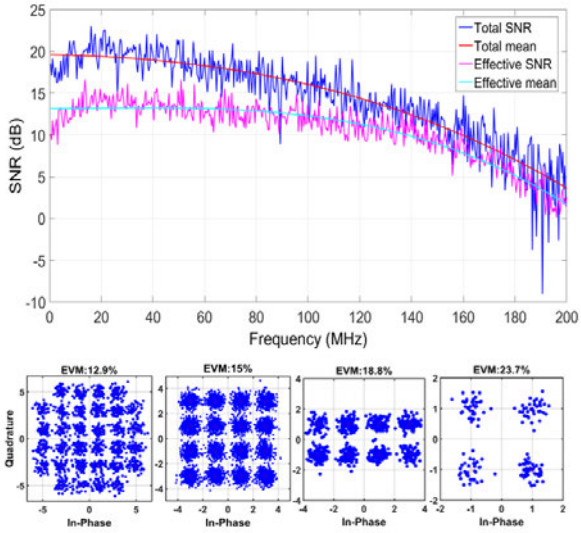
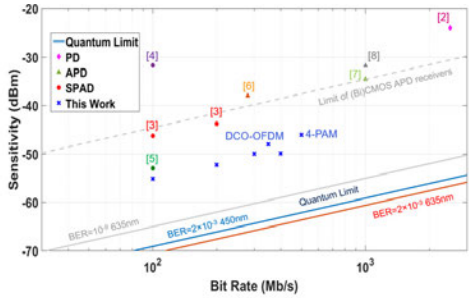


Figure 29.7.5: SNR measured at each subcarrier from channel estimation. 32-QAM, 16-QAM, 8-QAM and 4-QAM received constellations.



Ref.	[2]	[6]	[7]	[4]	[5]	[3]	This Work
Type	PD	APD	APD	SPAD	SPAD	SPAD	SPAD
Technology	0.35μm	N/A	0.35μm	130nm	180nm	0.35μm	130nm
Elements	1	1	1	1024	60	4	4096
Fill Factor	N/A	N/A	N/A	2.42%	3.2%	48%	43%
Sensitivity (dBm)	-23	-38	-34.6	-31.7	-53	-43.8	-46.1
Modulation Type	OOK	OOK	OOK	OOK	OOK	RZ-OOK	4PAM, OFDM
Bit Rate	2.5Gb/s	280Mb/s	1Gb/s	100Mb/s	100Mb/s	200Mb/s	500Mb/s, 350Mb/s
BER	10^{-3}	10^{-9}	10^{-9}	10^{-9}	10^{-3}	6.5×10^{-3}	2×10^{-3}
Consumption (pJ/bit)	86	N/A	244	800	N/A	248	230

Figure 29.7.6: Comparison of PD, APD and SPAD visible light RX devices.

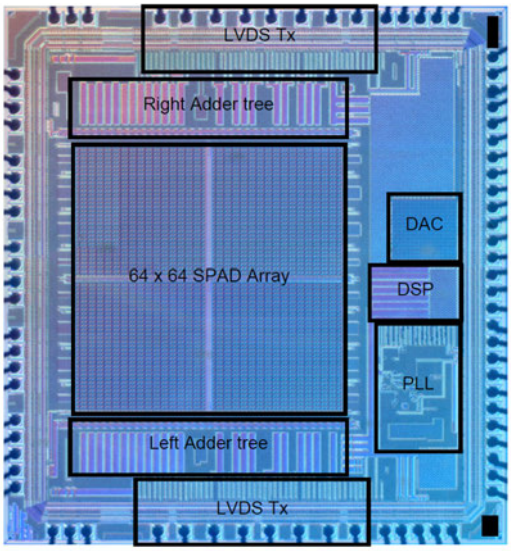


Figure 29.7.7: Annotated chip micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

29.8 SHARC: Self-Healing Analog with RRAM and CNFETs

Aya G. Amer, Rebecca Ho, Gage Hills, Anantha P. Chandrakasan, Max M. Shulaker

Massachusetts Institute of Technology, Cambridge, MA

Carbon nanotube (CNT) field-effect transistors (CNFETs) are a promising emerging technology for energy-efficient electronics (Fig. 29.8.1). Despite this promise, CNTs are subject to substantial inherent imperfections; every ensemble of CNTs includes some percentage of metallic CNTs (m-CNTs). m-CNTs result in conductive shorts between CNFET source and drain, resulting in excessive leakage and degraded (potentially incorrect) circuit functionality (Fig. 29.8.1). Several techniques have been developed to remove the majority of m-CNTs (no technique today removes 100% of m-CNTs). While these techniques enabled the first digital CNFET circuits, it is still not possible to realize large-scale CNFET analog or mixed-signal CNFET circuits due to m-CNTs. As shown in Fig. 29.8.1, while a digital logic gate can still function correctly in the presence of a small fraction of m-CNTs (but with degraded resilience to noise) [1], a single m-CNT in an analog circuit can result in catastrophic failure (e.g., degrading amplifier gain resulting in functional failure of circuit blocks such as ADCs and DACs)[1]. This paper presents a circuit design technique, Self-Healing Analog with RRAM and CNFETs (SHARC), that leverages the programmability of non-volatile resistive RAM (RRAM) to automatically "self-heal" analog circuits in the presence of m-CNTs. Using SHARC, we experimentally demonstrate analog CNFET circuits robust to m-CNTs as well as the first mixed-signals CNFET sub-system (4-bit DAC and SAR ADC; these are the largest reported complementary (CMOS) CNFET circuit demonstrations to-date).

The working principle of SHARC is as follows (Fig. 29.8.2): (1) Fabrication: each CNFET is split into multiple minimum width FETs (*i.e., "sub-CNFETs"*), with a RRAM cell fabricated directly under (or over) the source or drain contact of each sub-CNFET. (2) Initialization: a positive voltage is applied across the RRAM electrodes (top-to-bottom) to program the RRAM into a low resistance state (LRS) (this is the RRAM "FORM" operation;the RRAM stack is engineered so that the subsequent SET voltages are above the operating voltage of the circuit so that during operation the RRAM is not inadvertently programmed). (3) Self-Healing: the gate terminals of the CNFETs are biased to turn the CNFETs "OFF" and a negative voltage (i.e., opposite polarity used during the previous FORM operation) is applied across the CNFET and RRAM. CNFETs with only semiconducting CNTs (s-CNTs) in the channel do not conduct significant current, and thus the RRAMs in series with those CNFETs remain in a low resistance state. However, CNFETs containing m-CNTs in the channel conduct current (m-CNTs are not controlled by the gate and conduct current regardless of gate bias), and with sufficient current under negative bias the RRAM is re-programmed into a high-resistance state (HRS) (the RRAM "RESET" operation). To ensure RRAM RESET, the RRAM stack should be designed to require less reset current than the off-state leakage current through an m-CNT (<100μA for a scaled technology node). (4) Final circuit fabrication: Perform final chip fabrication (e.g., additional metal layers). After performing SHARC, sub-CNFETs containing m-CNTs are in series with the RESET high-resistance RRAM (effectively removing those sub-CNFETs from the circuit), while CNFETs containing only s-CNTs are in series with SET low-resistance RRAM. Figure 29.8.2 shows measured current-voltage (IV) transfer curves for typical CNFET prior- and post- performing the SHARC self-healing. Prior to the self-healing, the CNFET exhibits substantial off-state leakage current due to the presence of m-CNTs (typical measured I_{ON}/I_{OFF} is <10 for a CNFET containing m-CNTs). Post self-healing, the CNFETs exhibit I_{ON}/I_{OFF} >1000 and I_{OFF} reduces by >800× due to the series RRAM resetting to a high-resistance state during the self-healing.

Importantly, SHARC is a broad technique that can be applied to a wide range of analog circuit blocks. Additional factors should also be taken into consideration to minimize impacting circuit performance. For example, the absolute values and ratios between the RRAM high and low resistance states impact the leakage power (e.g., increased RRAM high-resistance decreases leakage current through sub-CNFETs containing m-CNTs) and output signal swing (e.g., increased RRAM low-resistance decreases available voltage headroom). Moreover, the location of the RRAMs in the circuit may impact gain and input referred noise (e.g., for the case of input transistors to amplifiers, the RRAM should be placed in series with the drain terminal of the sub-CNFETs rather than the source terminal). Figure 29.8.3 illustrates the benefit of SHARC on various amplifiers configurations. Simulations (based on experimentally calibrated CNFET and RRAM compact models for 10nm technology nodes [2]) confirm that all of these circuits retain correct functionality despite an m-CNT (within any CNFET in the circuit). As an example, without SHARC, a single m-CNT results in a worst-case reduction of gain by 106dB for a cascode amplifier, while with SHARC, a single m-CNT results in a worst-case reduction of gain of only 3dB. The measured transfer characteristics of a 2-stage op-amp is also shown indicating a gain of >800.

The key benefits of SHARC are: (1) the self-healing process (e.g., RRAM programming) is automatically performed by the m-CNTs themselves (as the m-CNTs provide the current path for resetting the RRAM during the self-healing process), (2) CNFETs and RRAMs are fabricated at low temperatures (<300°C), and therefore can be fabricated directly vertically overlapping one-another in a monolithic three-dimensional fashion; this allows each RRAM cell to be fabricated either above or below the source or drain contact of the CNFET, minimizing area penalty associated with SHARC (the area overhead due to splitting the initial CNFET into N sub-CNFETs is ~(N-1)*S/W (subject to other considerations such as signal routing), where S is the minimum lithographic pitch and W is the original CNFET width), and (3) the circuit retains its healed configuration due to the non-volatility of the RRAM (though due to their programmability, the RRAM can potentially be reprogrammed repeatedly in the future).

Importantly, SHARC scales to larger-scale circuits since SHARC can be performed once for all CNFETs in parallel. An example of how RRAMs are formed and reset for the op-amp as well as the process flow is shown in Fig. 29.8.4. Bipolar RRAMs are initially fabricated on the first layer of the chip, and a sacrificial layer of metal can be defined to connect the RRAMs for the FORM operation. The sacrificial layer of metal is then removed, followed by the CNFET fabrication (CNFET source/ drain contacts are patterned directly vertically overlapping the RRAM cells). The CNFETs are fabricated by depositing solution-based purified (~99.99% s-CNTs) CNTs on the wafer, followed by lithographically defining the source and drain regions. Another sacrificial layer of metal can be used to configure the circuit so all sub-CNFET gates are shorted, and all CNFET-RRAM pairs are connected in parallel. This enables all sub-CNFETs gates to be biased (e.g., $|V_{GS}| = 0V$) simultaneously, as well as for all the RRAMs to undergo the self-healing RESET operation in parallel (importantly, all the CNFETs are PMOS as-fabricated, allowing a single gate bias to turn off all the CNFETs simultaneously). The sacrificial layer of metal can then be removed, and the CNFETs are doped to form the PMOS and NMOS followed by the final metal routing to complete the circuit (the NMOS CNFETs are formed through electrostatic doping, whereby a non-stoichiometric HfO_X dielectric is deposited over exposed CNTs in the channel of the FETs, converting PMOS CNFETS to NMOS CNFETs [3]). Importantly, all the experimental circuits fabricated and reported in this paper are complementary (using both PMOS and NMOS CNFETs).

As an experimental demonstration, we use SHARC to fabricate the following analog and mixed-signal circuits: 4-bit DAC (SHARC implemented in the 2-stage op-amp and switches), and a 4-bit SAR ADC (SHARC implemented in the strong-arm latch and switches). All circuits are fabricated with ~2μm channel length CNFETs. Figure 29.8.5 shows the schematic and die image of a 4-bit DAC. The measured transfer characteristics show monotonic behavior with non-linearity and gain error (due to large parasitic caps and routing resistance). Figure 29.8.6 shows the SAR ADC schematic, its die micrograph and its transfer characteristics. These circuits are the first mixed-signal CNFET circuits demonstrated, and the largest reported CMOS CNFET circuits to-date (see the comparison table in Fig. 29.8.4 [4], [5], [6]). Furthermore, since SHARC is implemented at the CNFET-level, it can be combined with additional existing circuit techniques to further improve performance, such as technology node scaling to improve energy efficiency and improved circuit topologies to improve linearity. Thus, SHARC offers a feasible path forward to enabling large-scale, energy-efficient analog and mixed signal circuits using CNFETs.

[1] Analog circuits sensitivity to m-CNTs is caused by reducing the circuit output impedance and changing the DC operating point which drives the CNFETs out of saturation regime.

References:

[1] Jie Zhang, et al., "Probabilistic analysis and design of metallic-carbon-nanotube-tolerant digital logic circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 28, no. 9, pp. 1307-1320, 2009.

[2] Chi-Shuen Lee and H.-S. Philip Wong, Stanford Virtual-Source Carbon Nanotube Field-Effect Transistors Model 1.0.1, *nanoHUB*, 2015, doi:10.4231/D3BK16Q68.

[3] Christian Lau, et al., "Tunable n-Type Doping of Carbon Nanotubes Through Engineered Atomic Layer Deposition HfOX Films," *ACS nano*, 2018, doi:10.1021/acsnoano.8b04208..

[4] Yingjun Yang, et al., "High-performance complementary transistors and medium-scale integrated circuits based on carbon nanotube thin films," *ACS nano*, 11.4, pp. 4124-4132, 2017.

[5] Michael L. Geier, et al., "Solution-processed carbon nanotube thin-film complementary static random-access memory," *Nature nanotechnology*, 10.11, p.944, 2015.

[6] Shu-Jen Han, et al., "High-speed logic integrated circuits with solution-processed self-assembled carbon nanotubes," *Nature nanotechnology*, 12.9 , p.861, 2017.

We acknowledge Analog Devices, Inc., and DARPA 3DSoC for support.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

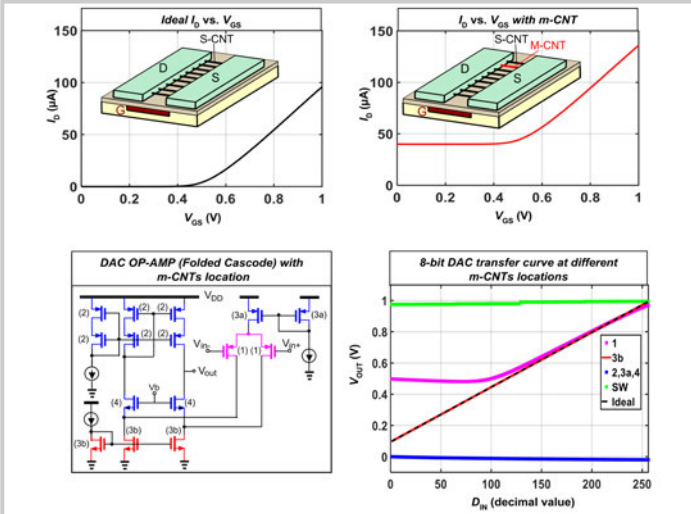


Figure 29.8.1: CNFET technology and impact of m-CNTs at the device and circuit level. The 10nm compact model used for simulations is calibrated to experimental data.

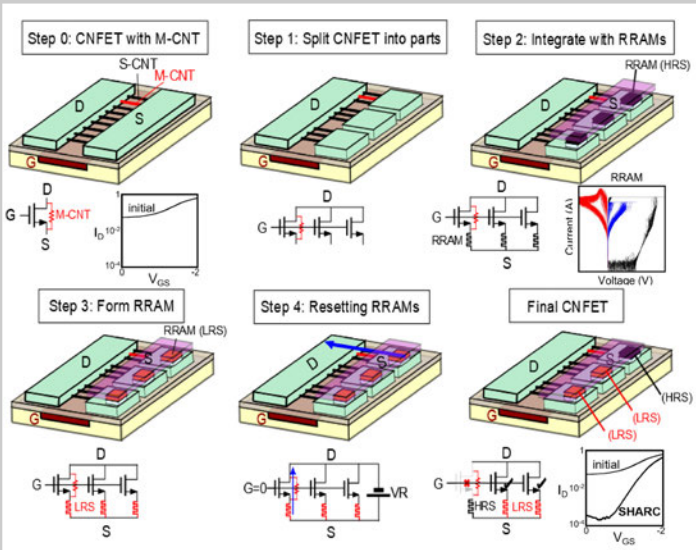
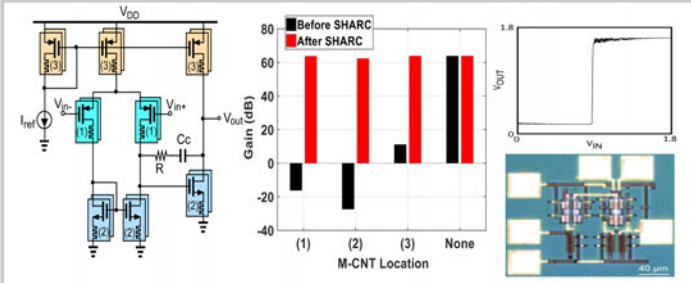


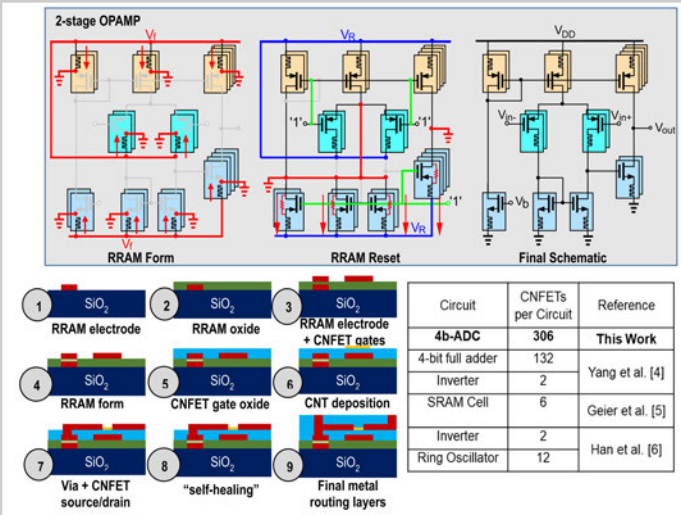
Figure 29.8.2: Principles of SHARC with programming steps.



M-CNT location	Common Source		Source Follower		Common Gate		Diff. Pair		Folded Cascode	
	Before	After	Before	After	Before	After	Before	After	Before	After
None (All S.C)	43 dB	42.89 dB	-50 mdB	-48 mdB	40.5 dB	40.7 dB	33.8 dB	35.4 dB	67.7 dB	67.7 dB
Input (1)	-3.4 dB	9.3 dB *	-108 dB	-75 mdB	2.13 dB	7.8 dB *	-130 dB	32 dB **	40.5 dB	66 dB
Active Load (2)	2.8 dB	40.8 dB	-1.2 dB	-57 mdB	2.76 dB	38.6 dB	-3.6 dB	34.4 dB	1.2 dB	65 dB
Current Source (3)	-		-		-		31.9 dB	35.4 dB	-38.8 dB	65 dB
									45.8 dB	67.5 dB
Cascode (4)	-		-		-		-		22.7	65.5

*Gain reduction due to high impedance node → can be solved in –ve FB systems, ** 3dB reduction as gm reduced to half

Figure 29.8.3: M-CNTs impact on analog circuits before and after SHARC (experimental OPAMP transfer curve (top right) and simulations (table and bar chart)).



Circuit	CNFETs per Circuit	Reference
4b-ADC	**306**	**This Work**
4-bit full adder	132	Yang et al. [4]
Inverter	2	
SRAM Cell	6	Geier et al. [5]
Inverter	2	Han et al. [6]
Ring Oscillator	12	

Figure 29.8.4: SHARC implementation (top) with fabrication process flow (bottom), as well as comparison to state-of-the-art CNFET CMOS demonstrations.

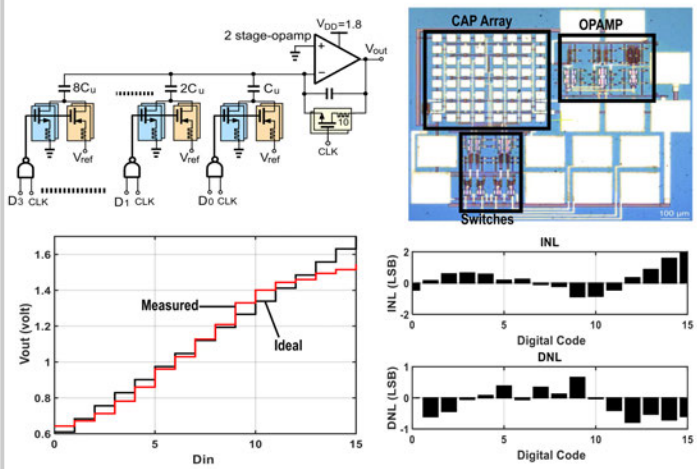


Figure 29.8.5: 4-Bit DAC with SHARC. Measured response is monotonic (|DNL|<0.8 LSB) (note: INL is larger in the last 2 codes due to saturation).

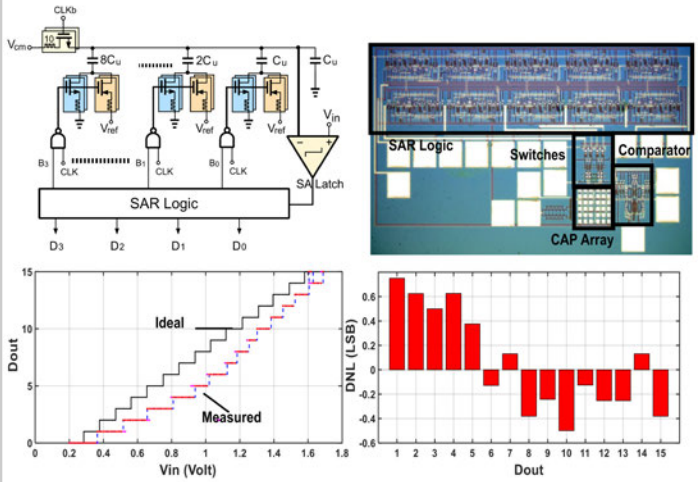


Figure 29.8.6: 4-Bit SAR DAC with SHARC. Measured characteristics show the ADC behavior with offset (35mV), gain error, and non-linearity (-0.5 LSB <DNL< 0.75 LSB).

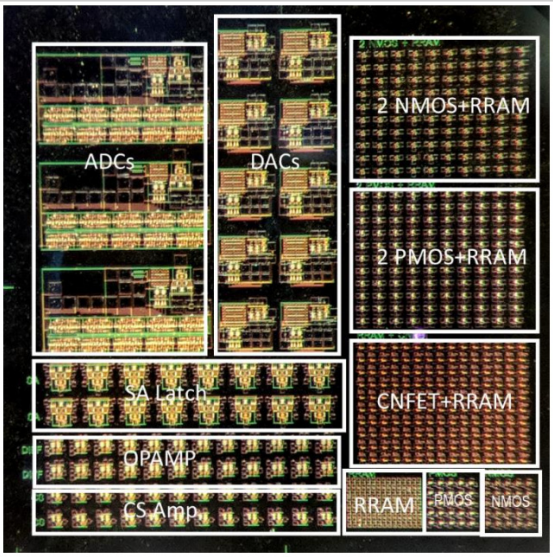


Figure 29.8.7: Die micrograph of the chip showing different circuits.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 30 Overview: *Advanced Wireline Techniques*

WIRELINE SUBCOMMITTEE

Session Chair: ***Yohan Frans***
Xilinx, San Jose, CA

Associate Chair: ***Friedel Gerfers***
Technische Universitaet Berlin, Berlin, Germany

Subcommittee Chair: Frank O'Mahony, Intel, Hillsboro, OR

As wireline transceivers become widely adopted in more systems and applications, challenges such as signaling over channels with high-loss/high-reflection and limited link power budget continue to arise. Since these challenges cannot be addressed by process technology advances alone, new design innovations at circuit level as well as link/system level continue to be critical.

This session starts with an invited paper describing the need and challenges of deploying high-speed links for automotive application. The next four papers propose new ideas on link modulation, architecture, coding, and adaption: a 56Gb/s link using discrete multi-tone (DMT) modulation, a ring topology uplink/downlink interface for large-capacity storage systems using cascaded CDRs, a sequence-coded PAM-4 link using a reduced-state trellis, and the use of a genetic algorithm to perform link adaptation. The last three papers in the session highlight innovations at building-block levels: a highly linear silicon photonic driver, a low-power multi-phase clock generator using injection-locked oscillator, a high-performance PLL that operates at 0.65V, and a low-power ring-oscillator-based injection-locked clock multiplier.

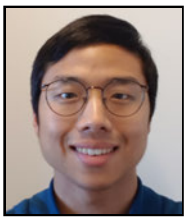
INVITED PAPER **1:30 PM**

30.1 Single-Pair Automotive PHY Solutions from 10Mb/s to 10Gb/s and Beyond

G. W. den Besten, NXP Semiconductors, Eindhoven, The Netherlands

In Paper 30.1, NXP Semiconductors describes emerging automotive Ethernet network standards and solutions enabling future driver-assistance as well as autonomous-driving technologies. Weight-space-cost constraints lead to a new class of high-bandwidth full-duplex interfaces operating over 15m single twisted-pair cabling meeting the stringent automotive requirements in terms of emissions, robustness, reliability, and fail safety.

2:00 PM



30.2 A 161mW 56Gb/s ADC-Based Discrete Multitone Wireline Receiver Data-Path in 14nm FinFET

G. Kim, EPFL, Lausanne, Switzerland; IBM Zurich Research Laboratory, Rueschlikon, Switzerland
now with KAIST, Daejeon, Korea

In Paper 30.2, École Polytechnique Fédérale de Lausanne demonstrates a 56Gb/s wireline RX employing discrete multi-tone (DMT) modulation that subdivides a data stream into multiples of frequency-domain sub-channels. The RX is implemented in 14nm FinFET technology and achieves an energy efficiency of 2.9pJ/b (1.2pJ/b for the DSP).

2:30 PM

30.3 A 25.6Gb/s Uplink-Downlink Interface Employing PAM-4-Based 4-Channel Multiplexing and Cascaded CDR Circuits in Ring Topology for High-Bandwidth and Large-Capacity Storage Systems

T. Toi, Toshiba Memory, Kawasaki, Japan

In Paper 30.3, Toshiba Memory Corporation presents an uplink-downlink interface for large-capacity storage systems employing four-channel multiplexing. The interface utilizes 25.6Gb/s PAM-4 signaling with cascaded CDRs in a ring topology, where each TX uses the RX recovered clock. Fabricated in 28nm CMOS technology, it achieves 3.69pJ/b at 25.6Gb/s with a BER < 10^{-15} over a channel loss of 1.84dB.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

3:15 PM

30.4 A 32Gb/s 2.9pJ/b Transceiver for Sequence-Coded PAM-4 Signalling with 4-to-6dB SNR Gain in 28nm FDSOI CMOS

Aurangozeb, University of Alberta, Edmonton, Canada

In Paper 30.4, the University of Alberta describes a 32Gb/s sequence-coded PAM4 transceiver using a self-converging, reduced-state trellis structure. Compared to conventional PAM-4, it provides up to 6dB SNR gain for a 30dB loss channel with 12.5% bandwidth overhead. The prototype, implemented in 28nm FDSOI technology, consumes 2.9pJ/bit at 32Gb/s over a channel loss of 30dB.

3:45 PM

30.5 A 1.41pJ/b 56Gb/s PAM-4 Wireline Receiver Employing Enhanced Pattern Utilization CDR and Genetic Adaptation Algorithms in 7nm CMOS

S. Shahramian, Huawei Technologies, Toronto, Canada

In Paper 30.5, Huawei Technologies presents a 56.25Gb/s analog-mixed signal PAM-4 receiver in 7nm CMOS that employs a genetic algorithm to perform adaptation of its analog front-end frequency response. The RX achieves an energy efficiency of 1.41pJ/b over a channel loss of 22.3dB without a DFE.

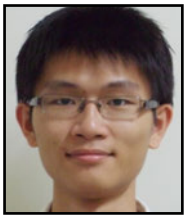
4:15 PM

30.6 A 6V Swing 3.6% THD >40GHz Driver with 4.5× Bandwidth Extension for a 272Gb/s Dual-Polarization 16-QAM Silicon Photonic Transmitter

A. H. Ahmed, University of British Columbia, Vancouver, Canada; Elenion Technologies, New York, NY

In Paper 30.6, the University of British Columbia presents a 0.13μm SiGe driver using resistor-based capacitor splitting, zero-peaking as well as source degeneration to obtain 40GHz of bandwidth. The proposed silicon photonic modulator driver achieves a signal swing of $6V_{ppd}$ and a THD of 3.6%.

4:30 PM

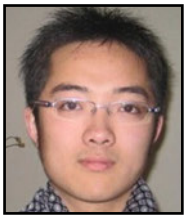


30.7 An 8b Injection-Locked Phase Rotator with Dynamic Multiphase Injection for 28/56/112Gb/s Serdes Application

Y-C. Huang, MediaTek, Hsinchu, Taiwan

In Paper 30.7, MediaTek describes an 8b injection-locked phase rotator that provides multiphase outputs with a peak-to-peak INL of 1.14LSB without the need of a multiphase input clock. The silicon area and power consumption are $0.0043mm^2$ and 11.4mW, respectively.

4:45 PM



30.8 A 0.65V 12-to-16GHz Sub-Sampling PLL with $56.4fs_{rms}$ Integrated Jitter and -256.4dB FoM

Z. Zhang, Hong Kong University of Science and Technology, Hong Kong, China

In Paper 30.8, the Hong Kong University of Science and Technology presents a 0.65V sub-sampling PLL that uses a hybrid dual-path structure. Fabricated in a 40nm CMOS process, the PLL achieves a frequency range from 12 to 16GHz with $56.4fs_{rms}$ jitter at 14GHz output, resulting in a FoM of -256.4dB.

5:00 PM

30.9 A $140fs_{rms}$-Jitter and -72dBc-Reference-Spur Ring-VCO-Based Injection-Locked Clock Multiplier Using a Background Triple-Point Frequency/Phase/Slope Calibrator

S. Yoo, Ulsan National Institute of Science and Technology, Ulsan, Korea

In Paper 30.9, the Ulsan National Institute of Science and Technology describes an ultra-low reference spur and low-jitter ring-VCO-based injection-locked clock multiplier. Dissipating only 11mW, the measured reference spur level and jitter are -72dBc and $140fs_{rms}$, respectively.

30.1 Single-Pair Automotive PHY Solutions from 10Mb/s to 10Gb/s and Beyond

Gerrit W. den Besten

NXP Semiconductors, Eindhoven, The Netherlands

Car communication networks are rapidly evolving from a collection of sub-MHz control busses to a high-performance data network for connecting sensors, processors, and actuators to enable autonomous driving [1]. A plurality of high-bandwidth nodes like cameras, displays, radars, and wireless transceivers, demand a network architecture revision and drive the need-for-speed on individual links [1,2]. Harsh automotive requirements on emissions, robustness, and reliability apply [3]. Size and weight of cables are critical factors too. These challenges inspired a new class of single-pair physical layer solutions, with 100Mb/s currently in production, 1Gb/s ramping up soon, and new standards for 2.5-10Gb/s in development.

Figure 30.1.1 shows a typical domain-based network as used in today's cars. Domains are separated for safety reasons, with a secure gateway for cross-domain connections. Data traffic within a domain cannot be hampered by other domains. However, this architecture has also disadvantages. Nodes of a domain are typically not grouped in physical location, leading to parallel networks per domain, implying more and longer cables due to non-optimal usage of the infrastructure. An alternative approach is to sub-divide the network in physical zones with a high-speed backbone. This can reduce cabling, but mixed data traffic of different domains on the backbone comes with safety challenges and worsens congestion problems. Next to this, there is also centralized computing needed for data fusion from different sensors, and redundancy when it comes to safety-critical functionality. This is an area of investigation, and the ultimate network architecture is probably unknown yet. In any case, future cars will undoubtedly contain more and much higher speed links than today's cars.

Unlike conventional multi-drop car data busses like CAN and LIN, the high-speed network is Ethernet-based with point-to-point PHY connections and switch functions to branch out. Link segments can be 15m long and may include up to 4 in-line connectors to facilitate car assembly (Fig. 30.1.2a). Arbitrary connector locations worsen return loss, which poses a challenge as insertion loss increases with speed. Practical routing distance increases substantially due to cable tray limitations implied by mechanical car construction. The increasing number of links makes network weight a concern too, as it impacts fuel consumption and tax. High-end cars already contain a few km of wire today and more than 30kg of network-related hardware. Furthermore, the cross-section of cable trays is limited, so thin and lightweight cables are beneficial, preferably unshielded-twisted pair (UTP).

For this reason, new full-duplex single-pair Ethernet PHY technology (BASE-T1) was developed, starting at 100Mb/s over single-pair UTP cables [4,5], as applied in cars today. The next generation single-pair PHYs, achieving full-duplex 1Gb/s over UTP [6], is currently in development by semiconductor manufacturers and will be introduced in next generation car platforms. On top of data, power can be transported over the same twisted pair [7] to eliminate separate power wiring to end-nodes.

For bi-directional traffic over a single pair, a 'hybrid' is used to discriminate received from transmitted signals and reduce required ADC performance. Both signals get super-positioned on the channel as shown in Fig. 30.1.3a. Note that for lossy channels, the transmitted signal can be much stronger than the received signal. In combination with multi-level modulation, this leads to challenging noise, linearity, and timing requirements for the front-end.

BASE-T1 PHY technology is based on the hybrid subtraction method (3b), enabling usage of the full channel bandwidth in each direction simultaneously. Figure 30.1.4 shows a generic block diagram of a BASE-T1 Ethernet transceiver. The analog front-end (AFE) receive path is used for analog signal conditioning and digitization, and includes gain control and filtering before the ADC, while the transmit path consists of a DAC, low-pass filtering and a line driver. An additional DAC feedback path can be used for digitally assisted signal corrections in the receiver front-end. Clock-and-data recovery, equalization, and residual echo cancellation are accomplished by digital signal processing (DSP). DSP-based transceiver architectures are beneficial for functional safety, as they can monitor state-of-health and flag warnings before failures occur. Equalization consists of feed-forward (FFE) and decision-feedback (DFE) parts. Both equalization and echo cancellation coefficients are background calibrated for actual time-variant channel characteristics using least-mean square algorithms. Clock-and-data recovery at higher data rates is typically based on Müller-Mueller algorithms to allow the ADC to run at Nyquist rate. The physical coding sublayer (PCS) includes the symbol codec and scramblers. Scramblers are used to preserve pseudo-random symbol patterns with an almost white spectrum, and to ensure uncorrelated ingress and egress data streams. The media-independent interface (MII) exists in different parallel CMOS rail-to-rail and SERDES flavors and serves at data bus to a switch or processor. The PHYs are controlled by registers via a serial management interface (SMI).

The 100/1000BASE-T1 PHYs use PAM-3 signal modulation, providing 1½ bits per symbol. This makes a 100BASE-T1 PHY run at a baud rate of 66⅔Mbd and a 1000BASE-T1 PHY at 750Mbd, including 9/8 overhead for Reed-Solomon (RS) coding to combat burst errors.

Currently, multi-Gb/s PHYs are being standardized by the IEEE for 2½, 5, and 10Gb/s. These PHYs will be conceptually similar to the 1000BASE-T1 PHY described above, but use PAM-4 at 1.4Gbd, 2.8Gbd, and 5.6Gbd respectively, including RS-coding. TX precoding is added to reduce burst error length. Figure 30.1.2 shows draft insertion and return loss limits, showing that multi-Gb/s will be much more challenging. Note that the echo can become larger than the received signal power, which imposes tight requirement on AFE and DSP accuracy. Low power consumption is essential. Full transceiver IC power consumption is targeted at $< (\sqrt{\text{DataRate}})/10^5$W. For the highest baud rates, link segment specs are based on shielded twisted pair cabling.

Shielded cabling reduces insertion loss, emissions, and robustness requirements for transceivers, with a penalty on weight, size, and cost of cables and connectors. The emission challenge is illustrated in Fig. 30.1.5 for 100/1000BASE-T1 PHYs over UTP cable, which acts as antenna for common-mode (CM) signals. This sets tight limits on transceiver CM signals, and mode-conversion of all link components. The transceiver emission test set-up and limit are shown in Fig. 30.1.5. The μV limit beyond 76MHz necessitates a CM choke (Fig. 30.1.4) and requires extreme care for symmetry in IC and board design. For multi-gig PHYs, emission limits are currently considered not feasible with UTP, mainly due to cable mode-conversion. Cable shielding strongly reduces emissions, improves immunity, and reduces CMC requirements or may even enable their elimination.

Immunity to RF interference is critical. Figure 30.1.6 shows two robustness tests: bulk current injection (BCI) and direct power injection (DPI), which are performed on running links and shall not cause link failures. DPI is tested on UTP up to 39dBm, injecting $>100V_{pp}$ CM signal on the link. Despite the CMC or shielding attenuation, hundreds of mVs of CM signal can occur at the transceivers.

For any transceiver type, automotive reliability requirements apply, including <1ppm failures in 3yrs as illustrated in Fig. 30.1.7, an operating temperature range from -40C to 125C ambient (grade-1), temperature cycle and mechanical stress robustness, system ESD tolerance, and moisture resilience. These aspects have large impact and drive architecture, design, testability, package, and manufacturing choices.

In the network with a steadily growing number of nodes and increasing link speeds, there is also an even lower speed PHY solution desired for cost effectively connecting nodes with less bandwidth demand. This is addressed by 10BASE-T1S: a 10Mb/s multi-drop bus within the Ethernet framework.

Besides network communication addressed by Ethernet, there is need for an efficient raw data streaming solution from high-resolution sensors to data processors too. This type of communication is typically highly asymmetrical with speeds up to 16Gb/s, and different BER requirements. For chip-to-chip connections MIPI CSI-2 is utilized [8], but there is no standardized PHY solution yet to efficiently span larger distances. Multiple standard bodies are investigating solutions [9]. Strong data asymmetry allows exploitation of frequency multiplexing to achieve a simpler low-cost single-pair SERDES solution as shown in Fig. 30.1.3c.

Single-pair PHY technologies are steadily gaining momentum, with growing interest from industrial and data center applications. Enabling (high-speed) data and power over compact low-cost channels in hostile environments makes these solutions widely applicable. Although 10Gb/s isn't finished yet, even higher speed grades will undoubtedly follow in the future, as bandwidth demand is never satisfied.

References:

[1] L. Reger, "The Road Ahead for Securely-Connected Cars", *ISSCC,* pp. 29-33, Feb. 2016.
[2] K. Matheus and T. Konigseder, "Automotive Ethernet", Cambridge University Press, ISBN 978-1-107-05728-9, 2015
[3] IEEE 100/1000BASE-T1 EMC Test Specification for Transceivers, Open Alliance
[4] IEEE 802.3bw, Physical Coding Sublayer (PCS), Physical Medium Attachment (PMA) sublayer and baseband medium, type 100BASE-T1
[5] H. Pan, J. Tan, et al., "An analog front-end for 100BASE-T1 automotive Ethernet in 28nm CMOS", *ISSCC,* pp. 186-187, Feb. 2016.
[6] IEEE 802.3bp, Physical Coding Sublayer (PCS), Physical Medium Attachment (PMA) sublayer, and baseband medium, type 1000BASE-T1
[7] IEEE 802.3bu, Physical Layer and Management Parameters for Power over Data Lines (PoDL) of Single Balanced Twisted-Pair Ethernet
[8] MIPI CSI-2: https://mipi.org/specifications/csi-2
[9] MIPI A-PHY: https://mipi.org/specifications/a-phy

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

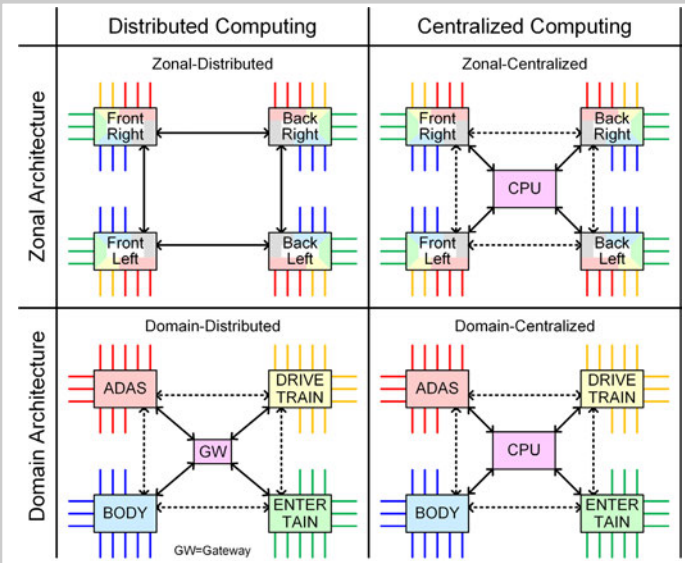


Figure 30.1.1: Network architectures.

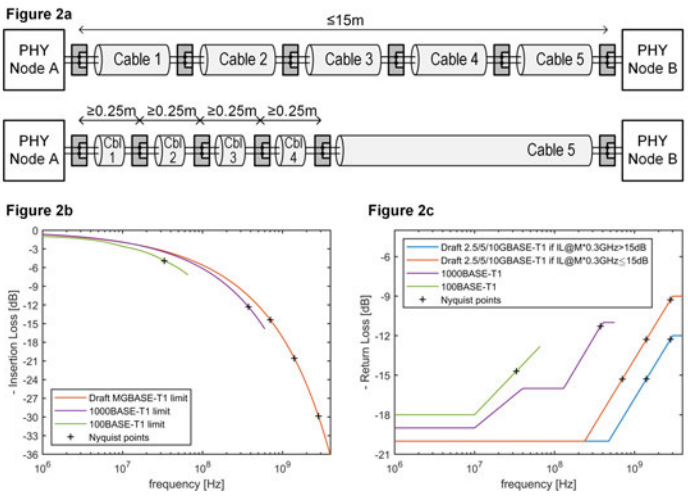


Figure 30.1.2: Link segment topology and specifications a) topology and connector positions b) insertion loss specifications c) return loss specifications.

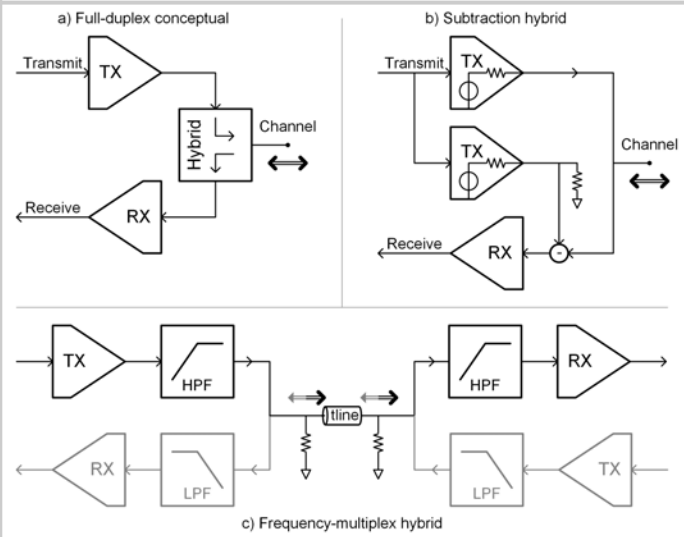


Figure 30.1.3: Full-duplex single-pair/coax transmission a) Full-duplex conceptual b) Subtraction hybrid c) Frequency-multiplex hybrid.

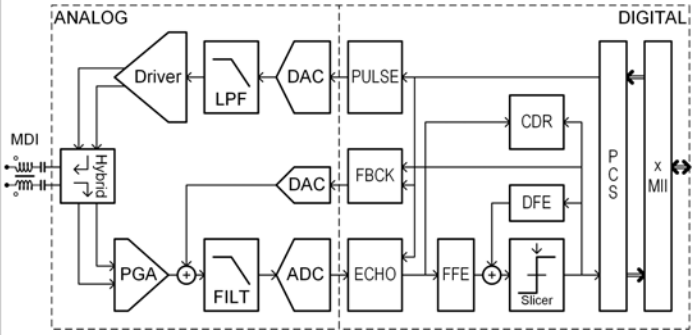


Figure 30.1.4: Ethernet transceiver block diagram.

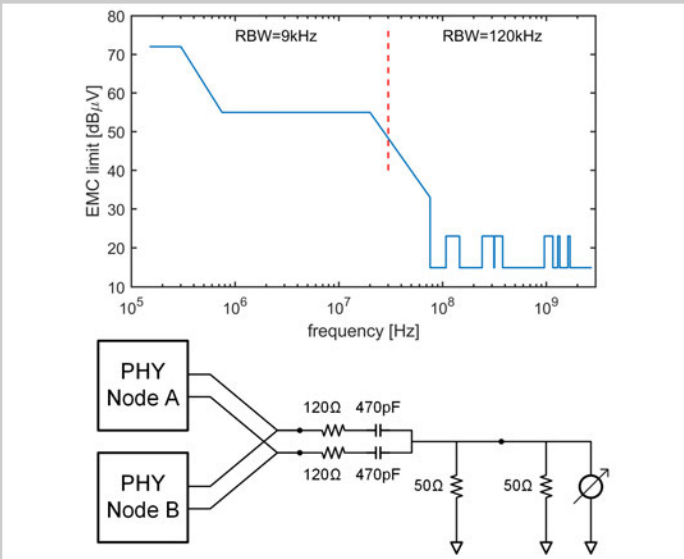


Figure 30.1.5: EMC specifications (100/1000BASE-T1).

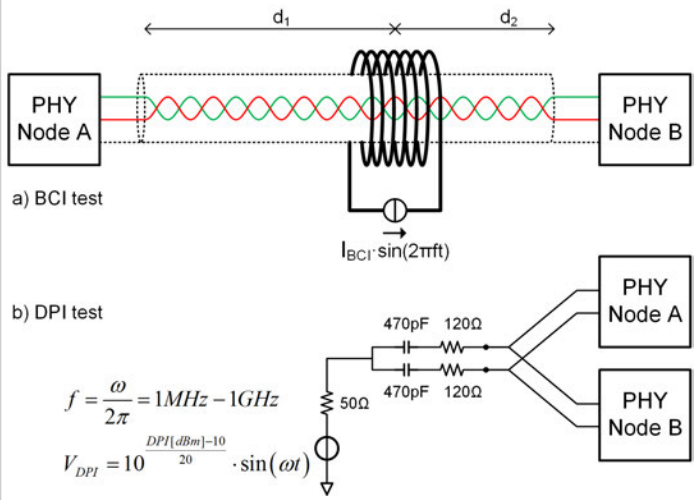


Figure 30.1.6: BCI and DPI immunity tests.

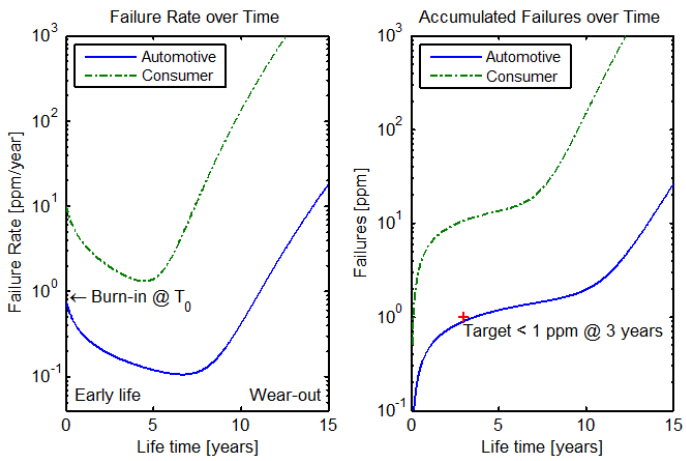


Figure 30.1.7: Reliability.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

30.2 A 161mW 56Gb/s ADC-Based Discrete Multitone Wireline Receiver Data-Path in 14nm FinFET

Gain Kim[1,2,3], Lukas Kull[2], Danny Luu[2,4], Matthias Braendli[2], Christian Menolfi[2], Pier-Andrea Francese[2], Hazar Yueksel[5], Cosimo Aprile[1,2], Thomas Morf[2], Marcel Kossel[2], Alessandro Cevrero[2], Ilter Ozkaya[1,2], Andreas Burg[1], Thomas Toifl[2], Yusuf Leblebici[1]

[1]EPFL, Lausanne, Switzerland
[2]IBM Zurich Research Laboratory, Rueschlikon, Switzerland
[3]*now with KAIST, Daejeon, Korea
[4]ETH Zürich, Zurich, Switzerland
[5]IBM T. J. Watson Research Center, Yorktown Heights, NY

The increasing demand on bandwidth for communicating among processors through wired interconnects in large-scale servers motivates the increase in the lane-data-rate from the current 28Gb/s to 56Gb/s or further. Recently published works [1-3] demonstrated ADC-based receiver (RX) prototypes equalizing >56Gb/s PAM-4 symbols for legacy channels with pre-FEC BERs of less than 2E-4 satisfying IEEE p802.bj/bs pre-FEC BER requirements. While the ADC-based >56Gb/s PAM-4 RXs provide strong equalization performance using a large number of feed-forward equalization (FFE) taps and a few decision-feedback equalization (DFE) taps [1,2] implemented in digital, their power consumption remains excessive due to heavy arithmetic operations in the DSP.

Operating in the frequency-domain, a discrete multi-tone (DMT) modulation enables simple equalization without a feedback loop by subdividing a data stream into multiples of frequency-domain sub-channels (sub-CH), and provides strong immunity to inter-symbol interference by the insertion of a cyclic prefix (CP) [4]. This work demonstrates a 56Gb/s wireline RX data-path employing a DMT modulation. An energy efficiency of 1.2pJ/b for the DSP and 2.9pJ/b for the entire RX was achieved which is considerably lower in terms of energy-per-bit as compared with recently published ADC-based PAM-4 RXs with similar data-rates for lossy transmission lines.

A simplified block diagram of the implemented RX architecture is shown in Fig. 30.2.1. A 10b pipelined-SAR ADC [5] is used as the RX front-end. The RX receives a differential half-rate clock (C2) as a reference clock to generate 1/48-rate clocks (C48) for sub-ADC sampling. Differential data input and C2 clocks are on-die-terminated with 100Ω parallel resistors and T-coil. Eight phases of C48 clocks, spaced by 45 degrees, synthesize a 1/36-rate clock (C36) using multiply-by-four and divide-by-three circuits for clocking the main data-path and the on-chip measurement circuitries of the DSP. The ADC outputs 48 time-interleaved digital codes operating at 22.4GS/s to achieve a 56Gb/s raw data-rate. The time-interleaved digital codes are then synchronized to a single C48-domain before being flushed into the DSP. The interface of the DSP clocked by a C48-domain clock and the main data-path of the DSP clocked by a C36-domain clock, are synchronized by a FIFO. The on-die test circuitry can be configured as a DMT constellation scanner, a bit-error counter, and an equivalent-time digital oscilloscope to observe any internal signal of the DSP from off-chip. A bi-directional serial interface (BIDI) configures the chip from a personnel computer (PC) and pulls-out the on-chip measured data to the PC.

The input logic in the DSP firstly converts the raw 11b ADC output codes including 1b of redundancy into 10b codes. Then the offset calibration is digitally applied to 10b codes while only 8 offset-calibrated higher-order bits are fed to a 48-in/36-out FIFO for further processing. The removal of the DMT CP is performed by simply disregarding the first four samples out of 36 samples that form one DMT symbol. A 32-point discrete-time Fourier transform (DFT) processor transforms the time-domain CP-removed DMT symbol into the frequency-domain. The DFT processor is implemented by two 16-point Winograd DFT processors followed by a twiddling network for best area-efficiency. Since the input of the 32-point DFT is purely real-valued, the second-half of its output taps are the complex conjugate symmetric values of the first-half of the DFT output taps. Hence, the transmitted data can be recovered by demodulating only the first-half of the DFT outputs. Moreover, as no bit is allocated at the DC sub-CH, 15 complex-numbered output taps of the DFT processor are fed to the frequency-domain equalizer (FDE).

The upper-inset of Fig. 30.2.2 depicts the frequency-domain equalization process that is performed by 15 complex multipliers for 15 non-DC sub-CHs. The FDE for each sub-CH receives 11b Re/Im DFT output and 9b Re/Im coefficients. The absence of an analog equalizer, such as a CTLE, makes the magnitude of the FDE coefficients cover a wide range amplification levels with respect to the sub-CH indices for making the resulting channel frequency response flat. Hence, the gain-adjustment technique is applied to the FDE for precisely equalizing the frequency-dependent phase delay and attenuation by maximizing the usage of the bit-width of the complex multipliers regardless of the coefficient magnitudes, as shown in Fig. 30.2.3.

The RX is implemented in 14nm FinFET technology. The DSP including the test circuitry is fully synthesized using standard cell libraries, placed and routed with back-end design flow, operating at 622MHz. The RX prototype is characterized using wafer probe needles on the probe station. The active area of the ADC and the DSP are 300x400um^2 and 180x400um^2, respectively. Including the ADC-DSP interface and C36 clock synthesizer circuits, total RX area is ~0.26mm^2.

To characterize the implemented RX, a discrete 8b DAC [6] operating at the same sampling rate as the RX ADC (22.4GS/s) is used as a DMT transmitter (TX) with 1V_{ppd} output dynamic range. A 16-sub-CH DMT symbol stream (with an empty DC sub-CH) is constructed off-chip based on three independent PRBS11 patterns. The inverse Fourier transform and CP insertion, followed by an 8-tap TX FIR filtering for de-emphasizing the low-frequency components, are performed in software. In order to minimize the relative DAC/ADC quantization SNR degradation of near-DC components as compared with high-frequency components, balanced gains are applied to the TX sub-CHs, as shown in the bottom insets of Fig. 30.2.2. The resulting time-domain stream is loaded into a DAC memory to be repeatedly transmitted. The 56Gb/s DMT signal, including 4 samples of CP per symbol, occupies 11.2GHz BW with all 64-QAM-modulated sub-CHs.

The RX was tested with transmission lines exhibiting 11dB (CH_A), 18dB (CH_B), and 28dB (CH_C) of attenuation at 14GHz, including losses from limited TX output BW, stripline, connectors, and cables. The constellation diagram scanning, BER counting, and internal signal observation for debugging purposes are performed by the on-chip test circuitry. The received and equalized 56Gb/s all-64-QAM DMT constellation diagrams and I/Q bathtub curves for channel CH_A, CH_B, and CH_C are shown in Fig. 30.2.4. The RX achieves pre-FEC BER less than 1e-6 for $CH_{A/B}$, and 2e-4 for CH_C. At nominal supplies of 0.75V, 0.8V, 0.7V, and 0.7V for DSP, clock buffer, TI-ADC interleaver, and ADC, respectively, the power consumption of the total RX is 161mW (68mW consumed in the DSP), corresponding to 2.9pJ/b (1.2pJ/b in the DSP).

Figure 30.2.5 shows the constellation diagrams and bit-error count results with 58.67Gb/s (64-QAM for 1st – 14th sub-CHs, 16-QAM for the 15th sub-CH) for CH_A, versus ADC sampling skew. After the sampling phase is skewed, the FDE coefficients are adapted off-chip and are loaded to coefficient table in the RX before the new bit-error counting starts. The data converter sampling rate is increased from 22.4GS/s to 24GS/s for this measurement, hence nominal voltage supplies are all increased to 0.8V for a reliable operation, resulting in a total RX energy efficiency of 3.66pJ/b for the 58.67Gb/s configuration. The pre-FEC BER is lower than 1e-6 over a measured range of ±40 degrees of the sampling phase skew. Comparisons with state-of-the-art ADC-based RXs with similar data-rate are shown in Fig. 30.2.6. A test chip die photo and the layout view are shown in Fig. 30.2.7.

Acknowledgements:
The authors would like to thank GlobalFoundries for chip fabrication.

References:
[1] Y. Frans, et al., "A 56-Gb/s PAM4 Wireline Transceiver Using a 32-way Time-Interleaved SAR ADC in 16-nm FinFET," *IEEE JSSC*, vol.54, no.4, pp. 1101-1110, Apr. 2017.
[2] P. Upadhyaya, et al., "A fully adaptive 19-to-56Gb/s PAM-4 wireline transceiver with a configurable ADC in 16nm FinFET," *ISSCC*, pp. 108-110, Feb. 2018.
[3] L. Wang, et al., "A 64Gb/s PAM-4 Transceiver Utilizing an Adaptive Threshold ADC in 16nm FinFET," *ISSCC*, pp 110-112, Feb. 2018.
[4] J. Armstrong, "OFDM for Optical Communications," *IEEE J. Lightw. Technol.,* vol.27, no.3, pp.189-204, Feb. 2009.
[5] L. Kull, et al., "A 10-bit 20-40GS/s ADC with 37dB SNDR at 40GHz Input using First Order Sampling Bandwidth Calibration," *IEEE Symp. VLSI Circuits*, pp. 275-276, June 2018.
[6] C. Menolfi, et al., "A 112Gb/s 2.6pJ/b 8-Tap FFE PAM-4 SST TX in 14nm CMOS," *ISSCC*, pp. 104-106, Feb. 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

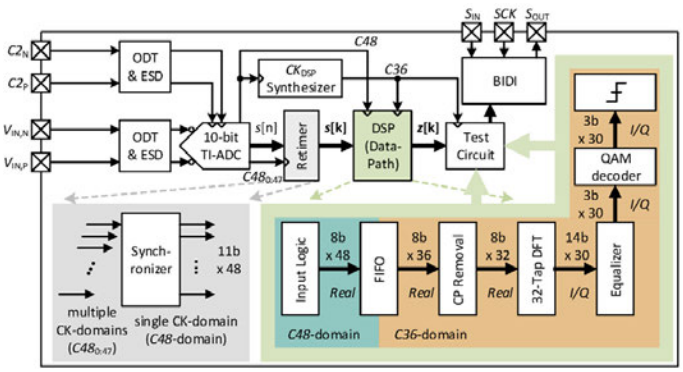


Figure 30.2.1: Receiver block diagram.

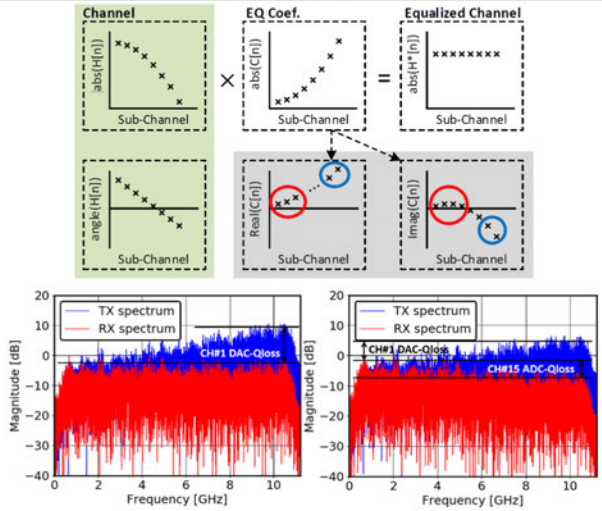


Figure 30.2.2: Frequency-domain equalization process (upper inset) and the TX/RX spectrum without/with the TX sub-channel gain balancing for balanced quantization SNR at near-DC and high-frequency sub-channels (bottom insets).

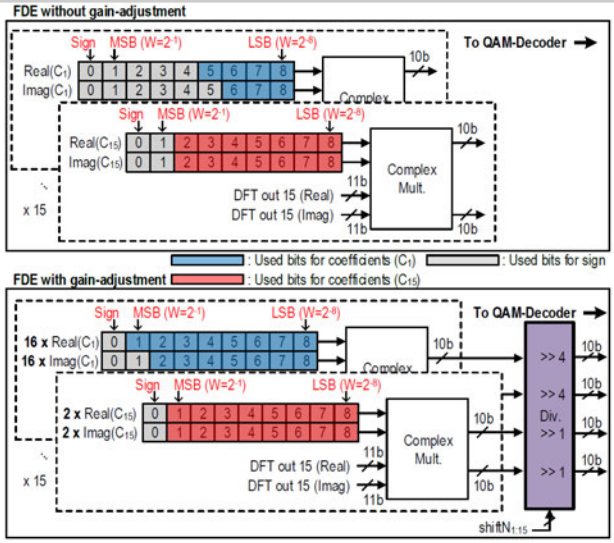


Figure 30.2.3: FDE details with (bottom inset) and without (upper inset) coefficient gain adjustment.

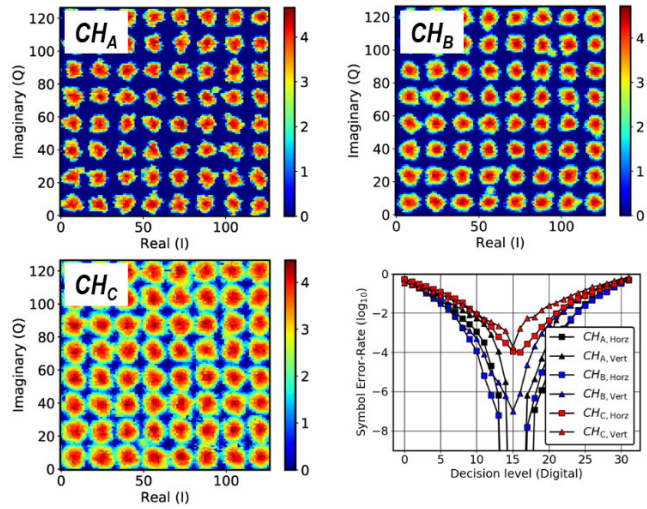


Figure 30.2.4: Constellation diagrams measured with the on-chip constellation scanner and symbol error-rate bathtub curves for channels A, B, and C at 56Gb/s data-rate.

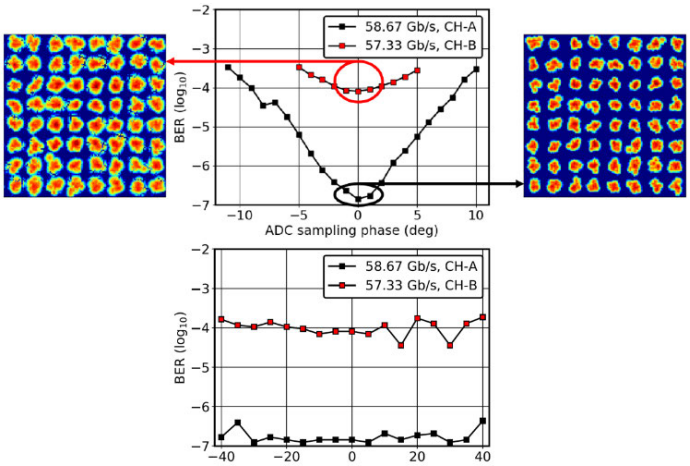


Figure 30.2.5: A constellation diagram and BER curves measured by on-chip BER counter, with respect to the ADC sampling phase; upper-left inset: without FDE adaptation, bottom inset: with FDE adaptation (64-QAM for 1st to 14th sub-CHs, 16-QAM for the 15th sub-CH).

	JSSC2017[1]	ISSCC18[2]	ISSCC18[3]	**This work**
Data Rate (Gb/s)	56	56	64.375	**56**
Modulation	PAM-4	PAM-4	PAM-4	**DMT**
Analog EQ	CTLE/VGA	CTLE/VGA	CTLE/VGA	**–**
ADC Res. (bits)	8	7	6	**10/8***
Fs (GS/s)	28	28	32.1875	**22.4**
DSP Power (mW)	N.A.	220	N.A.	**68.22**
RX Tot. Power (mW)	280**	545	283.9**	**161.46**
Energy (pJ/b)	5**	9.7	4.41**	**2.88**
Supplies (V)	0.9/1.2/1.8	0.85/0.9/1.2/1.8	0.9/1.2	**0.8/0.7/0.7/0.75**
RX Area (mm2)	N.A.**	1.76***	0.1625**	**0.26**
Ch. Loss (dB)	31	32	29.5	**CH-A:11**** CH-B:18**** CH-C:28******
Pre-FEC BER	<1e-8	<1e-12	<1e-4	**CH-A:<1e-6 CH-B:<1e-6 CH-C:<2e-4**
Technology	16nm FinFET	16nm FinFET	16nm FinFET	**14nm FinFET**

*DSP receives 8 MSBs for equalization out of 10 bits.
**DSP is not implemented in silicon; hence DSP area and power are not included.
***Estimated from die photo reported in [2].
****Loss at f_{NYQ} of PAM-4 equivalent system for comparison purpose. Actual losses in DMT system are 8 dB (CH-A), 13 dB (CH-B) and 20 dB (CH-C).

Figure 30.2.6: A comparison table with state-of-the-art ADC-based RXs with similar data-rates.

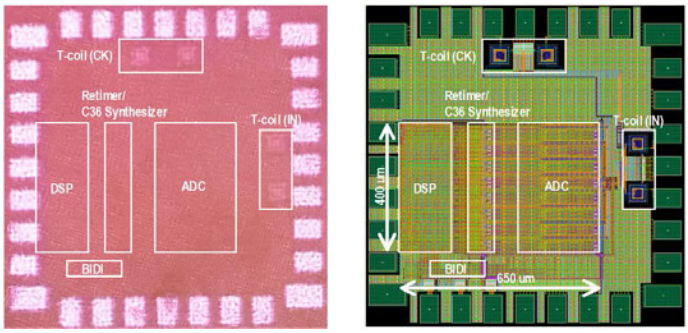


Figure 30.2.7: Die photo and layout view of the implemented RX.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

30.3 A 25.6Gb/s Uplink-Downlink Interface Employing PAM-4-Based 4-Channel Multiplexing and Cascaded CDR Circuits in Ring Topology for High-Bandwidth and Large-Capacity Storage Systems

Takashi Toi, Junji Wadatsumi, Hiroyuki Kobayashi, Yutaka Shimizu, Yuji Satoh, Makoto Morimoto, Rui Ito, Mitsuyuki Ashida, Yuta Tsubouchi, Mai Nozawa, Go Urakawa, Jun Deguchi

Toshiba Memory, Kawasaki, Japan

High-bandwidth (BW) and large-capacity storage systems with NAND Flash memory (hereinafter referred to as "NAND") have been increasingly required for big data applications, such as the field of advanced biomedical science [1]. However, a conventional NAND interface (I/F), e.g. Toggle DDR, with multi-drop bus topology has a tradeoff between BW and capacity due to the large load capacitance of NAND packages (PKGs). Although increasing the number of parallelized lanes of Toggle DDR improves both BW and capacity, it costs a large number of pins/wires on a controller/PCB. In order to overcome these problems, a daisy-chained serial I/F has been proposed [2]. In the I/F, bridge chips mask large load capacitance of NAND PKGs seen from a controller's transmitter (TX) so that a 12.8Gb/s downlink is realized. However, the multi-band multiplexing technique employed in [2] has a drawback in the difficulty in implementing an uplink because severe timing control is required for cumulatively multiplexing multiple bands (i.e. channels) in each bridge chip. In order to realize both a downlink and an uplink with lower power consumption, this paper presents a newly developed serial I/F with three key techniques: (1) PAM-4-based 4-channel (4-ch) multiplexing, (2) cascaded CDR circuits in (3) ring topology. The fabricated transceiver (TRX) for the proposed I/F achieves 3.69pJ/b with a BER lower than 10^{-15} at 25.Gb/s with PRBS31 through 1.84dB of channel loss at 6.4GHz. The proposed I/F can achieve a state-of-the-art FoM (defined as "# of packages × Data Rate / power consumption") of 1.80PKG·Gb/s/mW.

Figure 30.3.1 shows an overview of the proposed uplink-downlink I/F with the three key techniques. By employing the PAM-4-based 4-ch multiplexing, 4-ch data with 6.4Gb/s each are allocated to 4-amplitude levels in even and odd symbols of the PAM-4 signal. Each bridge chip identifies the 4-ch and extracts/inserts the data of one particular channel from/to the PAM-4 signal in order to write/read one channel data to/from NAND. This happens via Toggle DDR for downlink/uplink with tapered bandwidth using an analog architecture without a complex digital procedure such as ID addressing of bridge chips. Additionally, the cascaded CDR circuits enable TXs in bridge chips to reuse the RX recovered clock so as to omit high frequency PLLs for the TX, leading to a reduction in the power consumption. Although the cascaded CDR circuits accumulate a portion of jitter within the CDR BW, the amount of accumulated jitter is negligible for the function of the I/F because a phase-error-dependent bang-bang phase detector (PED-BBPD) autonomously improves the tracking ability of the CDR as described later. Finally, the ring topology, consisting of a controller and four bridge chips, realizes both an uplink and downlink, where one TRX in a bridge chip is shared for both of them.

Figure 30.3.2 shows the TRX architecture of the prototype for a controller and bridge chips. The PAM-4 signal, properly shaped by a continuous-time linear equalizer (CTLE) and a variable gain amplifier (VGA), is converted into 3b thermometer code by three slicers with thresholds of V_H, V_C and V_L in the four data/edge samplers with the phases of ck000 (even edge), ck090 (even data), ck180 (odd edge), and ck270 (odd data). The phases of the 4 sampling clocks are adjusted by a half-rate all-digital CDR circuit. However, the 4-phase clocks can be locked in two different states of correct and pseudo-lock, which have 180-degree phase difference as shown in the top left of Fig. 30.3.2. For PAM-4-based 4-ch multiplexing, the data/edge sampling with the pseudo lock causes reception errors. In order to have the CDR locked in the proper state, a training scheme using a pattern generator/checker (PAT GEN/CHK) for even/odd identification is implemented as described later. The 12b output of the data/edge samplers is fed to the CDR logic including PED-BBPD and digital loop filter (DLF). The 6b output of the data samplers is Gray coded into the four data streams corresponding to the 4-ch. The Gray-coded data is parallelized by 1:16 DEMUXs, and data for one particular channel is extracted/inserted by the digital block to be written/read to/from NANDs. The inserted data is sent to 16:4 MUXs together with the other 3-ch data. The 16b data streams of the 4-ch are serialized by 16:4 MUXs and subsequent 4:1 Gray decoders. Then, they are retimed and multiplexed into the MSB and LSB of the output PAM-4 signal (TXP and TXN) generated by a segmented source-series terminated (SST) driver. The SST driver is connected to T-coils compensating parasitic capacitances of ESD and pads. The clocks used for the TX are generated by either the CDR or PLL; the controller TX utilizes the clocks from the PLL and the bridge chips reuse the recovered clocks.

Figure 30.3.3 shows the training scheme for even/odd identification between adjacent chips. The TX sends the specific training pattern to the subsequent RX, which is a 3.2GHz clock pattern generated by the digital block. The CDR adjusts its own phase by using only ck270, ck000, and ck090 for CDR logic in synchronous mode. By omitting ck180, the rising edge of ck000 is locked to the data edge rising from odd to even as the correct state. After the training, a communication with an actual data pattern is started.

Figure 30.3.4 shows a function of the proposed PED-BBPD, which uses different numbers of transition states according to the relative phase error between the data and the recovered clock. When the relative phase error is smaller than α, it is only detected by the 8 transitions that have proper edge-crossing points, resulting in suppressing the PAM-4 transition jitter. When the relative phase error is larger than α, it is detected by the 12 transitions including the additional 4 transitions that have undesired edge-crossing points. As jitter within CDR BW is accumulated, then the PED-BBPD captures more transitions, leading to autonomous improvement of CDR tracking ability.

The prototype of the proposed I/F is fabricated in a 28nm CMOS process. The TRX for a controller and four bridge chips are connected via stacking connectors to configure the ring topology. The insertion loss between adjacent chips is 1.84dB at 6.4GHz. The top left of Fig. 30.3.5 shows the measured phase noise of the PLL clock in a controller and recovered clocks in four bridge chips. The CDR BW is around 200MHz, and therefore, the part of jitter lower than that is accumulated. However, the jitter accumulation is adequately suppressed and the CDR circuits autonomously track the accumulated jitter using the PED-BBPD so that 10^{-10} eye openings at the even phase of Bridge-0's RX and the controller's RX are sufficiently large as shown in the bottom of Fig. 30.3.5. As shown in the top right of Fig. 30.3.5, a BER better than 10^{-12} is measured at the low-level horizontal eye openings at the even phase of Bridge-0's RX and the controller's RX .

Figure 30.3.6 shows the performance summary: comparisons with the state-of-the-art I/F system and PAM-4 TRXs, and power breakdown. The proposed I/F system can achieve the FoM, defined as "# of packages × Data Rate / power consumption," of 1.80 PKG·Gb/s/mW. The power consumption is 455.3mW including 103.7mW from the controller and 4 × 87.9 mW from bridge chips, where removal of high-frequency PLLs for the TX saves 11.7mW per bridge chip. The fabricated TRX realizes good energy efficiency of 3.69pJ/b (TX: 1.47pJ/b and RX: 2.22pJ/b) with BER lower than 10^{-15} at 25.6Gb/s PRBS31. Figure 30.3.7 shows the die micrograph.

References:
[1] S. Mikula, "Progress Towards Mammalian Whole-Brain Cellular Connectomics," *Frontiers in Neuroanatomy*, 10:62, June 2016.
[2] Y. Tsubouchi, et al., "A 12.8 Gb/s Daisy Chain-Based Downlink I/F Employing Spectrally Compressed Multi-Band Multiplexing for High-Bandwidth and Large-Capacity Storage Systems," *IEEE Symp. VLSI Circuits*, pp. 149-150, June 2018.
[3] A. Nazemi, et al., "A 36Gb/s PAM4 Transmitter Using an 8b 18GS/s DAC in 28nm CMOS," *ISSCC*, pp. 58-60, Feb. 2015.
[4] T. O. Dickson, et al., "A 1.8pJ/b 56Gb/s PAM-4 Transmitter with Fractionally Spaced FFE in 14nm CMOS," *ISSCC*, pp. 118-120, Feb. 2017.
[5] G. Zhu, et al., "A Low-Power PAM4 Receiver Using 1/4-Rate Sampling Decoder with Adaptive Variable-Gain Rectification," *IEEE ASSCC*, pp. 81-84, Nov. 2017.
[6] J. Im, et al., "A 40-to-56Gb/s PAM-4 Receiver with 10-Tap Direct Decision-Feedback Equalization in 16nm FinFET," *ISSCC*, pp. 114-116, Feb. 2017.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

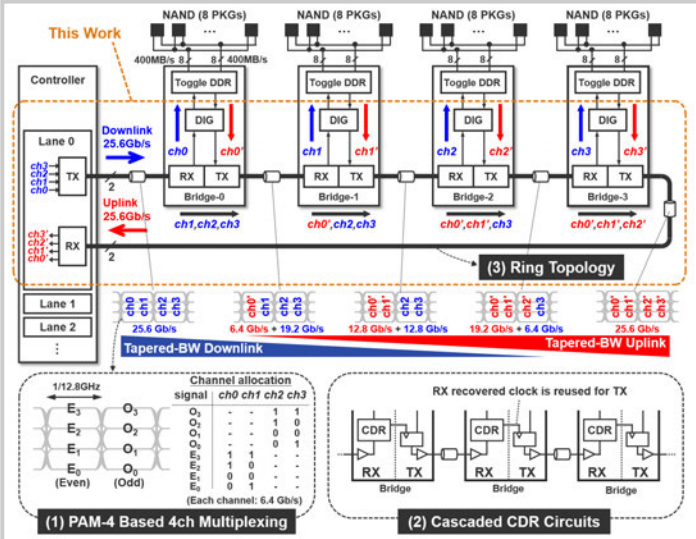


Figure 30.3.1: The proposed uplink-downlink I/F for high-BW and large-capacity storage systems.

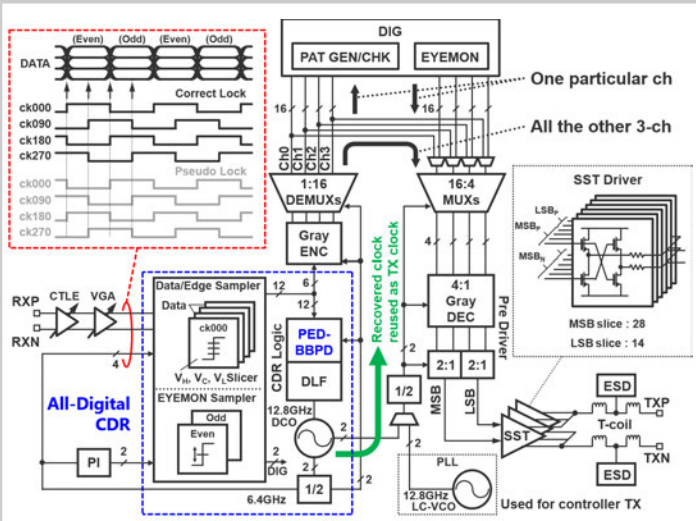


Figure 30.3.2: 25.6Gb/s transceiver architecture for a controller and bridge chips in the proposed I/F.

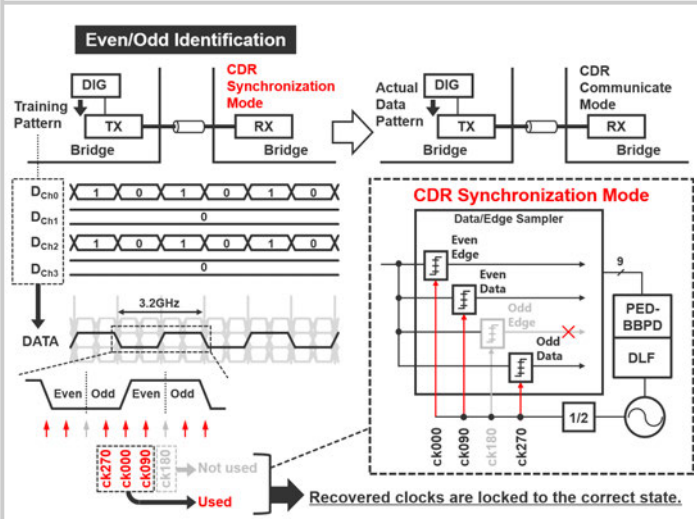


Figure 30.3.3: Training scheme with pattern generator/checker (PAT GEN/CHK) and CDR for even/odd identification.

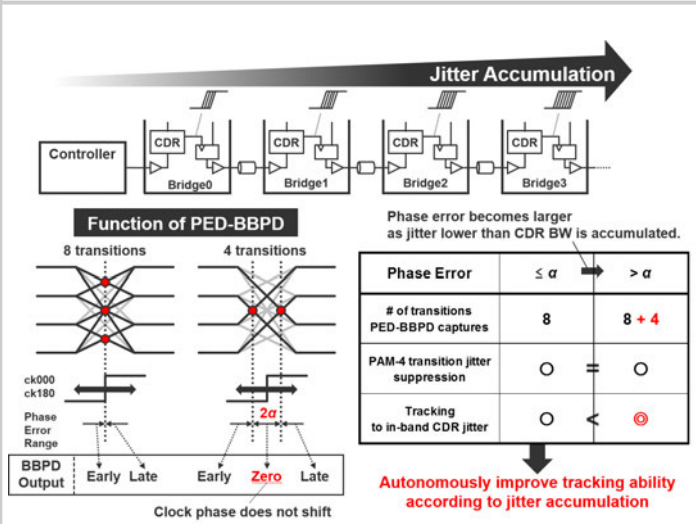


Figure 30.3.4: Function of phase-error-dependent BBPD (PED-BBPD).

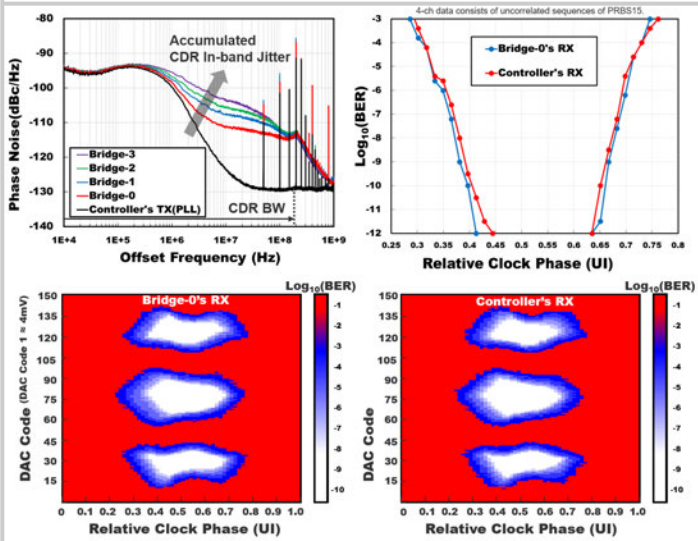


Figure 30.3.5: Measurement results of jitter accumulation of cascaded CDR circuits, eye diagrams and BER curves.

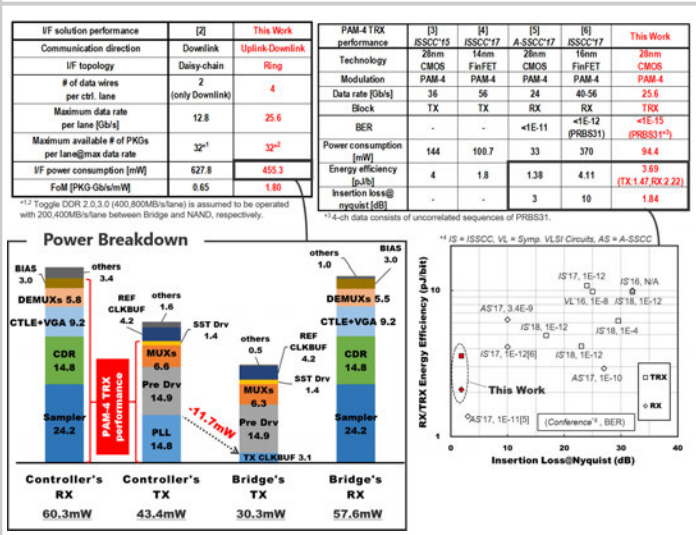


Figure 30.3.6: Comparisons with the recent state-of-the-art I/F system and PAM-4 transceivers, and the power breakdown.

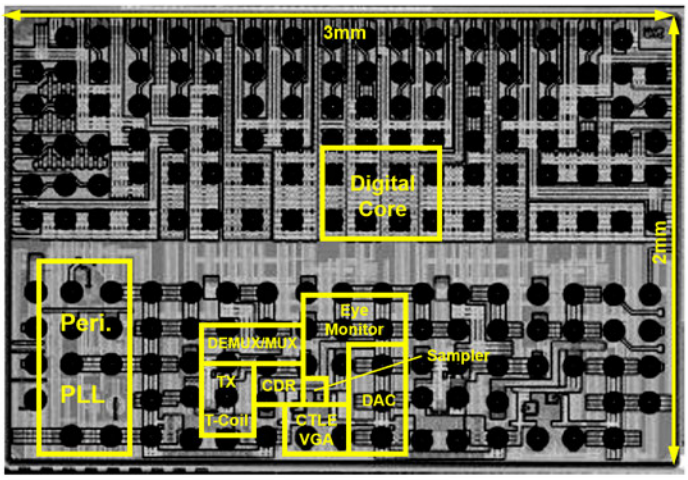


Figure 30.3.7: Die micrograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

30.4 A 32Gb/s 2.9pJ/b Transceiver for Sequence-Coded PAM-4 Signalling with 4-to-6dB SNR Gain in 28nm FDSOI CMOS

Aurangozeb[1], Carson Dick[1], Maruf Mohammad[2], Masum Hossain[1]

[1]University of Alberta, Edmonton, Canada
[2]Intel, Santa Clara, CA

Multilevel signaling such as PAM-4 makes more efficient use of the better part of the channel response by mapping multiple bits in the same time interval (i.e. UI). At the same time, to achieve lower power consumption simple equalization techniques such as TX-FIR and RX peaking equalization are often preferred. For example, in Fig. 30.4.1, a channel with 30dB of loss is equalized using these linear equalization techniques. However, the increased spectrum efficiency comes at the cost of SNR that is captured in the single bit response (SBR). Here, the equalized symbol response can be decomposed into MSB and LSB SBRs. In PAM-4 signaling, individual eye heights are set by the LSB SBR, whereas, the total signal amplitude is set by the MSB and LSB combined SBR. This 1/3 ratio suggests a 9dB SNR penalty, but in reality, crosstalk and residual ISI noise also get enhanced. Since the residual ISI is proportional to the pulse amplitude, the LSB will experience both MSB and LSB residual ISI that is also 3× worse compared to NRZ, and the same is also true for crosstalk. Obviously, this reduction in signal-to-noise-and-crosstalk ratio is compensated by improved channel characteristics to make PAM-4 signaling justified and advantageous over NRZ. However, as the channel loss increases beyond 20dB, the peak power constraint of the transmitter causes further SNR loss. Even with high frequency boost and gain provided by the linear equalizer, eye opening is less than 30mV at the sampler input and this limits the achievable symbol error rate in the link [1].

Sequence encoding is an effective way to improve the SNR, however, its decoding complexity (i.e. MLSE) and latency have often limited its usage in high-speed links. Here, we introduce a 3b sequence encoding that is simple, yet provides significant SNR improvement. The sequence consists of current MSB (B_0^M), current LSB (B_0^L) and previous LSB (B_{-1}^L) e.g., $B_0^M B_0^L B_{-1}^L$. Since N bits result in 2^N states, a reduced state encoding scheme is implemented. Firstly, by taking advantage of the channel ISI we set LSB main (h_0^L) and post cursor (h_{+1}^L) to have equal tap values. As a result, the combination of B_0^L and B_{-1}^L creates three levels – 11, 10 or 01, and 00 (Fig. 30.4.1). Secondly, the MSB is fully equalized to a single tap response and its weighting factor is the same as the LSB. Therefore, the combined eye diagram is two duobinary LSB eyes offset by the MSB forming a 4-level signal. However, in this case, each level represents a combination of a 3b sequence and therefore, these signal levels can be considered 'states' in the trellis. For simplicity, we refer to these states as -3, -1, +1 and +3 and corresponding 3b sequences are labelled in Fig. 30.4.1. The SNR benefits become clearer from the SBRs and eye diagrams when the transmit signal swing and analog front-end (AFE) response are identical (Fig. 30.4.1). Firstly, from a signalling point of view, the ratio between the LSB (or MSB) is only half of MSB + LSB. That means both signal-to-ISI and signal-to-crosstalk ratios will improved by 3dB compared to conventional PAM-4. Second, constructive use of the ISI translates to signal gain as evident from the LSB SBR amplitude. Thirdly, elimination of the transition between -3 and +3 provides higher timing margin for the outer eye that compensates the nonlinearity of the AFE. All these factors together contribute 6dB of SNR improvement as evident from 2× eye opening improvement from 30mV to 60mV. Note that part of this SNR benefit takes advantage of channel ISI – therefore, SNR benefit scales with loss. However, even for a 10dB loss channel, 4dB SNR improvement is achievable.

The 3-tap FIR-based transmitter architecture is shown in Fig. 30.4.2. The 32Gb/s half-rate PAM-4 transmitter requires an 8GHz differential clock, which is generated from the 8GHz four-stage ring injection-locked oscillator (ILO). In the proposed encoding, the LSB eye needs to be ½ UI shifted compared to the MSB eye. Fortunately, the required quadrature phase is readily available from the ring ILO. For reconfigurability purposes, the MSB serializer can choose between 0 and 90 degree clocks. The output driver is segmented in current DACs, where MSB and LSB tap weights are independently set. Sequence encoding is done using those tap weights – no additional encoder is needed. Valid sequence sets are pre-selected at the PRBS output. Formation of the sequence-encoded eye is better understood from MSB/LSB eyes at the TX and AFE outputs. Since the AFE output is an internal node, visibility of the eye is impacted by an unavoidable buffer. To minimize its impact here we are showing the eye at 16Gb/s in Fig. 30.4.2 with 32Gb/s results to follow in later sections.

The goal of the proposed encoding is to shift equalization in the digital domain and simplify the AFE. Therefore, the analog front-end is simplified to a simple passive filter followed by feedback equalizer. Although the output of the equalizer is an open eye, those 4 states need to be decoded back to MSB and LSB since the levels represent multiple sequences of 3 bits. For example, state +1 represents 110,101,011 and it is not obvious which of these three sequences is the correct one. Fortunately, construction of the trellis is self-converging, +3 and -3 represents a unique 3b sequence. In addition, the previous LSB is also part of the sequence that can be utilized in the decoding process shown in Fig. 30.4.3, and the hardware implementation is shown in Fig. 30.4.4. The decoding process is described below.

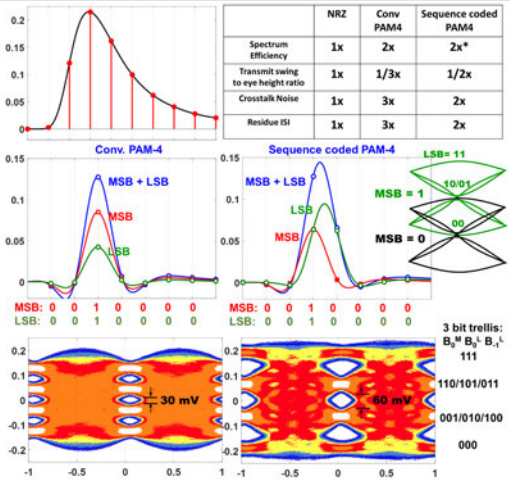
First, the output of the AFE is digitized to 2b state values by comparing to 3 reference levels and flagging the start and end points of the trellis decoding, which are +3 or -3 states. The number of states between two convergence points is referred to as sequence length. Second, each of these states are translated to 3b sequences using 8 to 3 selectors. Third, the forward selection path selects 2 of 3 possible sequences by correlating B_0^L and B_{-1}^L between consecutive states. This is functionally equivalent to a DFE without timing constraints since the states are decoded based on threshold. Decision propagation is implemented in a feed-forward manner (Fig. 30.4.4). For guaranteed convergence, trellis construction is constrained: 1) from states -1 or +1, to +1 transitions should not include 011, and 2) from +1 or -1, to -1 should not include 100. With these constrains there are two survivors from each state. Finally, once we reach the last state, B_{-1}^L propagates backward (i.e. traceback) to select the final sequence. Note that the sequence length between start and end-point is unknown. In this prototype we designed for a maximum 32-UI sequence length. The transition constraint translates to a bandwidth penalty – in the worst-case scenario it is 12.5%. However, some of sequences naturally converge without any additional constraint. As shown in Fig. 30.4.3 the sequence -3 to +3 does not require any constraint for convergence. Therefore, it is possible to further reduce the bandwidth penalty, especially when combined with forward error correction (FEC) encoding.

Performance of the proposed signaling scheme is compared to conventional PAM-4 over different losses in Fig. 30.4.5. At 10dB loss, conventional PAM-4 achieves a BER better than 10^{-12}. However, at higher loss, reduced SNR significantly degrades the eye opening (Fig. 30.4.5). Although the internal eye at the sampler input is expected to be better, it only marginally meets a BER of 10^{-6}. However, SNR improvement in the proposed PAM-4 provides sufficient eye opening to achieve BER < 10^{-12}. In Fig. 30.4.6 when compared to the analog mixed signal solutions [1-3], we can compensate significantly higher loss with better energy efficiency. In addition, the scalable digital architecture eliminates DFE latency constraints. The loss compensation capability of this work is comparable to ADC-DSP solutions [4-5] with 5× to 10× lower power consumption (2.9pJ/bit). The implemented 28nm FDSOI prototype in Fig. 30.4.7 occupies 0.4mm^2 dominated by the inductor. Note that the 12.5% bandwidth penalty reduces the 32Gb/s raw data rate to a 28Gb/s (32 × 0.875) effective information rate. However, the SNR and energy efficiency benefit remain compelling with further opportunity to improve data rate.

Acknowledgements:
The authors would like to acknowledge CMC Microsystems, Canada for their support. This work was supported by the Computing Hardware for Emerging Intelligent Sensory Applications Project through the National Sciences and Engineering Research Council of Canada Strategic Networks under Grant NETGP485577-15.

References:
[1] L. Tang, et al., "A 32Gb/s 133mW PAM-4 transceiver with DFE based on adaptive clock phase and threshold voltage in 65nm CMOS," *ISSCC*, pp. 114-116, Feb. 2018.
[2] P. Peng, et al., "A 56Gb/s PAM-4/NRZ transceiver in 40nm CMOS," *ISSCC*, pp. 110-111, Feb. 2017.
[3] A. Roshan-Zamir, et al., "A Reconfigurable 16/32 Gb/s Dual-Mode NRZ/PAM-4 SerDes in 65-nm CMOS," *IEEE JSSC*, vol. 52, no. 9, pp. 2430-2447, Sept. 2017.
[4] D. Cui, et al., "A 320mW 32Gb/s 8b ADC-based PAM-4 analog front-end with programmable gain control and analog peaking in 28nm CMOS," *ISSCC*, pp. 58-59, Feb. 2016.
[5] Aurangozeb, et al., "Channel-Adaptive ADC and TDC for 28 Gb/s PAM-4 Digital Receiver," *IEEE JSSC*, vol. 53, no. 3, pp. 772-788, March 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE



	NRZ	Conv PAM4	Sequence coded PAM4
Spectrum Efficiency	1x	2x	2x*
Transmit swing to eye height ratio	1x	1/3x	1/2x
Crosstalk Noise	1x	3x	2x
Residue ISI	1x	3x	2x

Figure 30.4.1: Comparison of sequence coded PAM-4 with NRZ and conventional PAM-4 in terms of relative signal and noise amplitude, SBR and equalized eye. (Ch. loss: 30 dB, loss compensation by Tx-FFE: 12 dB, AFE boost: 12-to-14 dB).

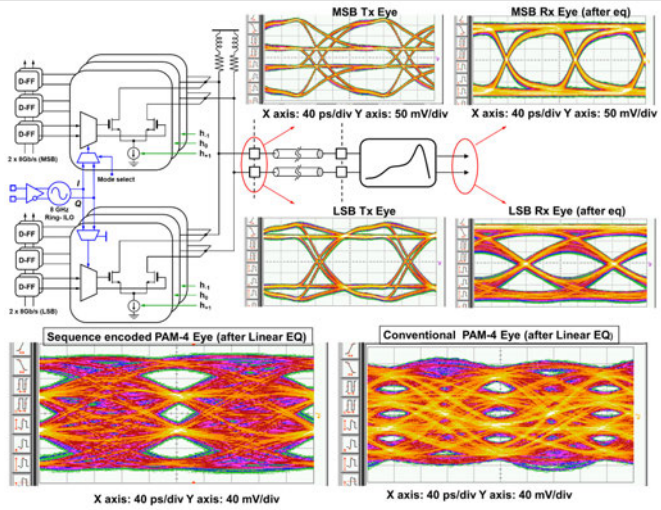


Figure 30.4.2: Sequence-coded Tx-FFE architecture. Measured MSB and LSB eyes are at the TX output and at the RX-AFE output at 8Gb/s. 16Gb/s PAM-4 eye is measured by enabling both MSB and LSB DACs with appropriate tap values.

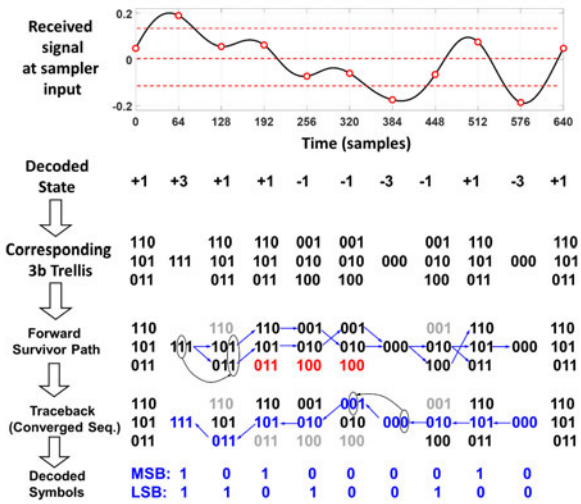


Figure 30.4.3: Sequential steps showing proposed trellis decoding and MSB, LSB detection. Here the arrows Indicate the direction of decision propagation. Red sequences are forbidden sequences.

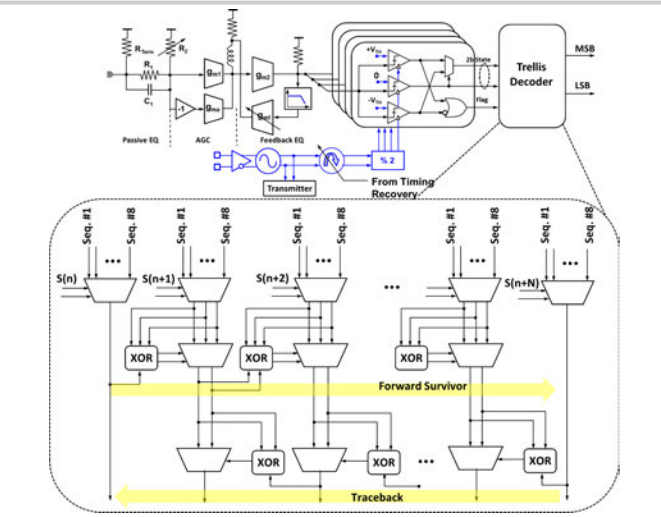


Figure 30.4.4: Detail of the receiver implementation including analog front-end. Drawing is single-ended for simplicity, actual implementation is differential. Digital trellis decoder is shown in detail in the bottom figure.

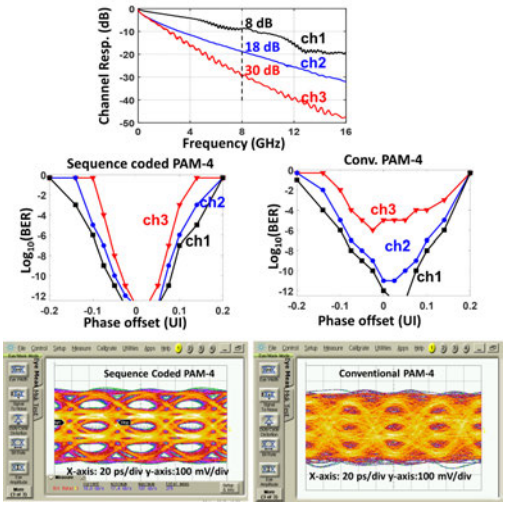


Figure 30.4.5: Performance comparison over different ch. losses. Measured BER vs. phase offset for corresponding ch. losses. Measured eye diagrams at the sampler inputs are observed through an additional amplifier and 50Ω buffer.

	This Work	ISSCC 2017 [2]	ISSCC 2016 [4]	JSSC 2017 [3]		ISSCC 2018 [1]	JSSC 2018 [5]
Max Data Rate Effective Rate	32 Gb/s 28 Gb/s	56 Gb/s ---	32 Gb/s ---	32 Gb/s ---	16 Gb/s ---	32 Gb/s ---	28 Gb/s ---
Modulation	Sequence Encoded PAM4	PAM4	PAM4	PAM4	NRZ	PAM4	PAM4
Process	28 nm FDSOI	40 nm	28 nm	65 nm		65 nm	65 nm
Supply (V)	0.9 (Analog) 0.75 (Digital)	1	NA	TX: 1.2 RX: 1.0		1.2	1.2
Channel Loss (dB)	30 dB @ 8 GHz	24 dB @ 14 GHz	32 dB @ 8 GHz	13.5 dB @ 8 GHz	27.6 dB @ 8 GHz	23 dB @ 8 GHz	30 dB @ 7 GHz
Equalization	TX: 3-tap FFE + Seq. Encoding RX: CTLE +Trellis decoding	TX: 3-tap FFE RX: CTLE + 3-tap DFE	CTLE-DSP	TX: 2-tap FFE RX: 1-tap FFE, 2-tap IIR DFE	TX: 4-tap FFE RX: 1-tap FFE, 2-tap IIR DFE	TX: 2-tap FFE RX: CTLE + 1-tap DFE	CTLE-DSP
Power Consumption (mw)	TX: 38 RX: 55 TX+RX : 93	TX: 200 RX: 382 TX+RX: 582	RX: 320	TX: 158.6 RX: 17.7 TX+RX: 176.3	TX+RX: 173.7	TX: 53 RX: 80 TX+RX: 133	RX: 130
FoM (pJ/bit)	TX: 1.18 RX: 1.72 TX+RX: 2.90	TX: 3.57 RX: 6.82 TX+RX: 10.39	RX: 10	TX: 4.96 RX: 0.55 TX+RX: 5.51	TX+RX: 10.86	TX: 1.66 RX: 2.5 TX+RX: 4.16	RX: 4.6
Area (mm^2)	TX: 0.11 RX: 0.29	TX: 1.14 RX: 1.26	0.89	0.074		TX: 0.028 RX: 0.16	0.2025

Figure 30.4.6: Comparison with state-of-the-art solutions.

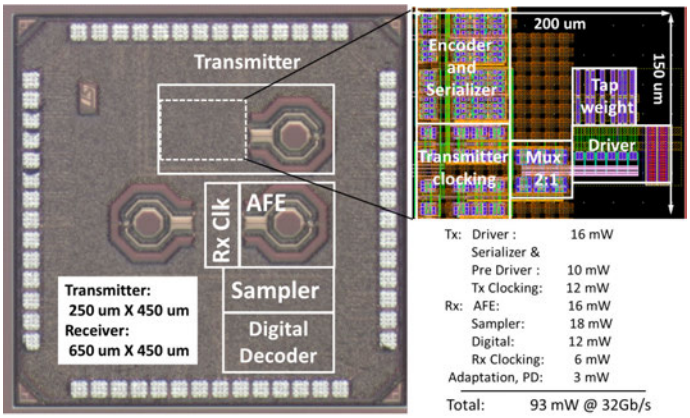


Figure 30.4.7: Implemented prototype in 28nm FDSOI process with power and area breakdown.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

30.5 A 1.41pJ/b 56Gb/s PAM-4 Wireline Receiver Employing Enhanced Pattern Utilization CDR and Genetic Adaptation Algorithms in 7nm CMOS

Shayan Shahramian*[1], Behzad Dehlaghi*[1], Joshua Liang[1], Ryan Bespalko[1], Dustin Dunwell[1], James Bailey[1], Bo Wang[1], Alireza Sharif-Bakhtiar[1], Michael O'Farrell[1], Kerry Tang[1], Anthony Chan Carusone[2], David Cassan[1], Davide Tonietto[3]

[1]Huawei Technologies, Toronto, Canada
[2]University of Toronto, Toronto, Canada
[3]Huawei Technologies, Ottawa, Canada
*Equally-Credited Authors (ECAs)

Analog mixed-signal (AMS) receivers for 50+Gb/s PAM-4 offer lower power than ADC-DSP receivers [1-3]. Those using DFEs [2-3] suffer from relatively high power consumption due to the large number of latches needed in PAM-4 speculative DFEs. Better power efficiency can be achieved using only a CTLE [1]. However, analog front-ends (AFEs) are sensitive to variations in process, supply voltage and temperature. To combat this while accommodating links with loss exceeding 20dB, an AFE with extensive programmability is combined with an efficient genetic adaptation algorithm to select a setting that minimizes BER thus equalizing a 22dB-loss channel. The lack of a DFE, combined with a novel PAM-4 clock recovery scheme greatly reduces the number of latches required compared to previous works, resulting in 1.41pJ/bit power consumption in 7nm CMOS technology.

Figure 30.5.1 shows the quarter-rate receiver block diagram with the termination and CTLE schematics inset. Twelve quarter-rate latches sample the top/middle/bottom eyes of the PAM-4 signal, and only 4 quarter-rate edge latches are used for clock recovery. The receiver termination employs shunt inductive peaking in series with the 50Ω termination (15 adjustable settings) to extend the bandwidth. The CTLE is composed of three CML stages with source degeneration. The first stage uses resistive (6 settings) and capacitive (15 settings) source degeneration to allow adjustment of the low- and high-frequency gain. There is also midband shaping included in the first stage via a series-connected degeneration resistance and capacitance (25 adjustable settings). The second stage includes source resistance degeneration (6 settings) to allow for further adjustment of the low-frequency gain. The third stage acts as a buffer to drive the latches. All three stages employ squelch capacitances at their outputs to further fine-tune the high-frequency gain (7 settings total) as well as load resistance trim (8 settings total). In total, the AFE has over 10 million different combinations that can be chosen. While the large number of settings allow for flexibility to equalize different channels, the large number and interdependence of the different controls makes it difficult to adapt to the optimal setting.

Figure 30.5.2 explains the adaptation algorithm used to adapt the front-end settings. It is loosely based on a class of genetic optimization algorithms [4]. Genetic algorithms have been extensively studied for automated design of analog circuits [5] but here we use it for the adaptation of analog circuits over channel, process, voltage and temperature variations. The algorithm relies partly on randomness in choosing coefficients to help it avoid local minima in the performance surface that could cause sub-optimal convergence in traditional gradient descent algorithms. Figure 30.5.2 illustrates the procedure on a 2-D performance surface (i.e. AFE with only 2 knobs) to allow for better visualization, but it is applied to the AFE's 7-D performance surface. The cost function being optimized is the vertical eye opening measured using the on-chip eye monitor. For PAM-4 operation, either the bottom or top eye opening are used since these outer eyes see more impairment from non-linearity. Initially, in step 1, 15 random children (AFE settings) are chosen as an initial population for the algorithm. The eye opening is measured at each of these settings (children) and the top 3 are chosen to become parents. These are then combined with additional new children created in steps 2 to 4, to form the population for the next iteration of the algorithm. In step 2 of the algorithm, the 3 parents "evolve" into 3 new children by taking the arithmetic mean of each pair of surviving parents. In step 3, 3 additional children are created by applying controlled mutations to the parents from step 2 by choosing new random children within some set distance away from the parents; these help the population descend the performance surface quickly while preventing the children from wandering too far from the optimal value. Step 4 involves adding 6 completely random mutations (random settings) to help explore other parts of the performance surface and help avoid getting trapped in local minima. The children generated in steps 2, 3, and 4 are combined with the parents and are used in the next iteration of the algorithm in step 5. Steps 2-5 are then repeated. The number of required iterations depends on the number of settings and the performance surface. More iterations allow the algorithm to do a better job of scanning the performance surface. Over time, the randomness of the algorithm can be reduced to only track variations due to supply voltage and temperature.

The clock recovery method is understood considering the PAM-4 transitions shown in Fig. 30.5.3 where they are decomposed into good and bad transitions. The location of the good transitions is the optimal lock point for the edge clock. To utilize all good transitions, 3 sets of edge latches are required (12 latches for a quarter-rate system) which increases the power and area of the receiver. Alternatively, only the center transitions can be used with pattern filtering to ignore the bad transitions [1-2], however, this lowers tracking bandwidth due to the attendant reduction in transition-density. In the proposed receiver, only 1 set of edge latches (4 latches for a quarter-rate system) is used, however, instead of ignoring the bad transitions, they are used advantageously to help both the CDR lock time and to increase the tracking bandwidth. The phase detector (PD) logic is shown in Fig. 30.5.3. For the good transitions, the phase detector behaves like a conventional bang-bang PD. For the bad transitions, the edge sample is still used to indicate when the clock is "very" early/late. If the sampling point falls within the very early/late region, a larger shift in the sampling point is applied by the PD than when detecting good transitions. Using this PD logic reduces the receiver area/power (uses 4 edge latches vs. 12 edge latches) and reduces the capacitive loading on the CTLE which translates into additional power savings.

Figure 30.5.4 shows the measured system performance for two different channels. The channels are comprised of cables, connectors, an ISI board, an evaluation board, and the chip package. Channel A has a loss of 17.8dB and Channel B has a loss of 22.3dB at 14GHz as shown in Fig. 30.5.6. A Keysight (M8045A) Pattern generator is used as the PAM-4 transmitter using 5dB of pre-emphasis. Figure 30.5.4a shows the measured contours captured using the on-chip eye monitor for a PRBS31 pattern and for channel B; the top and bottom eyes have 33mV of opening while the middle eye has an opening of 38mV at a BER of 1E-6. Figure 30.5.4b shows the measured bathtub curve and shows an opening of 0.17UI for channel A and 0.12UI for channel B at a BER of 1E-6. There is an upper bound in the measurable BER of 1E-2 due the counter size used in the on chip BER checker.

Figure 30.5.5(left) shows the measured receiver jitter tolerance at a BER of 1E-6 for channels A and B. The receiver meets the CEI-56G-VSR-PAM4 jitter tolerance mask for both channels. Figure 30.5.5(right) shows the measured jitter histogram from a Keysight sampling oscilloscope (DCA-X 86100D). The RJ is 240fs when the CDR is frozen and is 600fs when the CDR is in tracking mode.

Figure 30.5.6 shows the power breakdown of the receiver. Figure 30.5.6 also shows a table comparing this work to previous receivers at similar data-rates, including clock distribution but not clock generation power. This work consumes 50% less power and 40% less area than other AMS receivers operating at similar data rates and channel losses. For a fair comparison to previous work, we report the receiver power both with/without regulators. The power efficiency with regulators is 1.87pJ/bit (reported from the regulator supply), however, for a fair comparison to previous works, the energy efficiency of 1.41pJ/bit should be used (reported from the regulated supply voltage). A chip photo with area breakdown is shown in Fig. 30.5.7.

References:

[1] E. Depaoli, et al. "A 4.9pJ/b 16-to-64Gb/s PAM-4 VSR Transceiver in 28nm FDSOI CMOS", *ISSCC*, pp. 112-114, Feb. 2018.
[2] P.-J. Peng, et al. "A 56Gb/s PAM-4/NRZ Transceiver in 40nm CMOS", *ISSCC*, pp.110-111, Feb. 2017.
[3] J. Im, et al. "A 40-to-56Gb/s PAM-4 Receiver with 10-Tap Direct Decision-Feedback Equalization in 16nm FinFET", *ISSCC*, pp. 114-115, Feb. 2017.
[4] D. E. Goldberg, "Genetic Algorithms in Search, Optimization and Machine Learning", Addison-Wesley, Reading MA USA 1989.
[5] H. Shibata, et al. "Analog circuit synthesis by superimposing of sub-circuits," *IEEE ISCAS*, pp.427-430, 2001.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

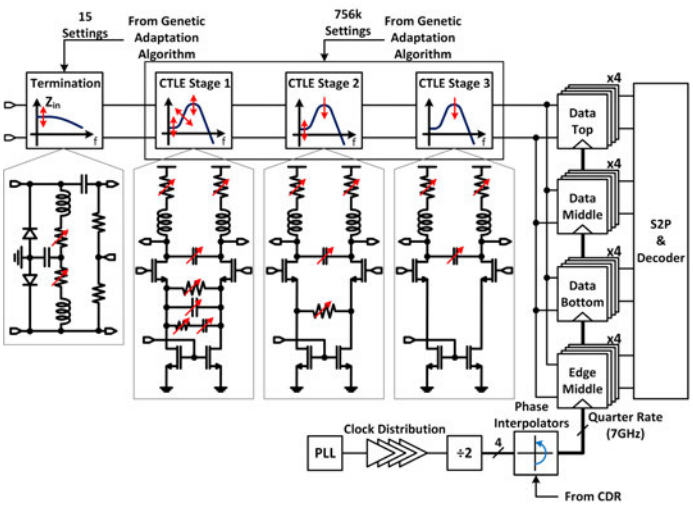


Figure 30.5.1: Block diagram of the receiver with termination and CTLE schematic (inset).

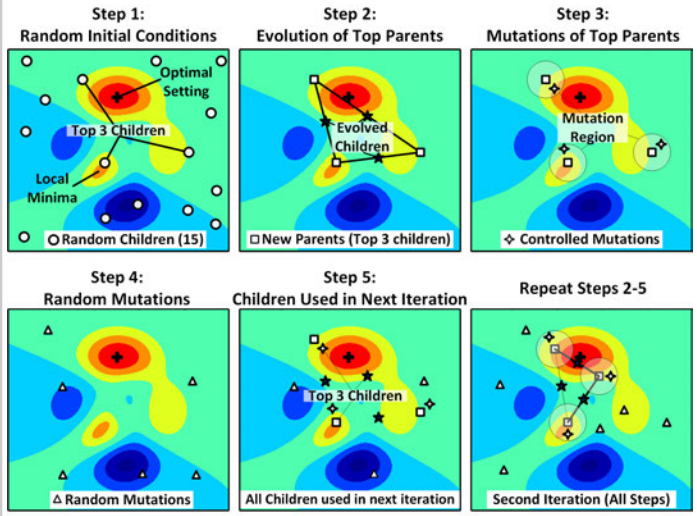


Figure 30.5.2: Genetic adaptation algorithm procedure for optimization of the analog front-end parameters.

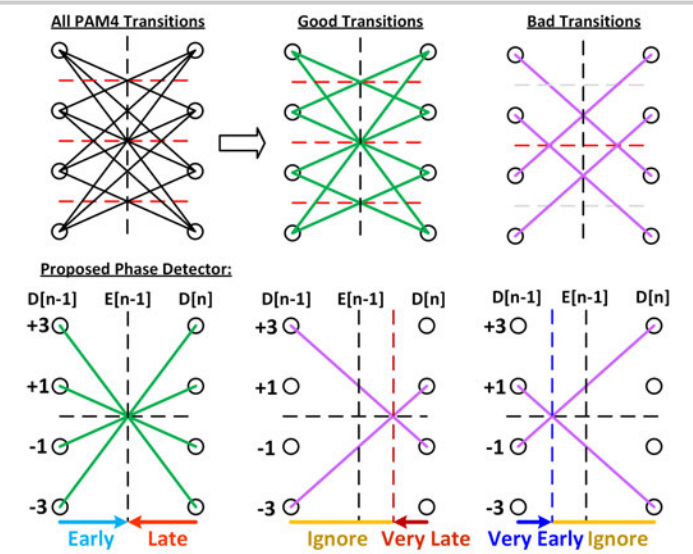


Figure 30.5.3: PAM-4 transitions and proposed phase detector logic.

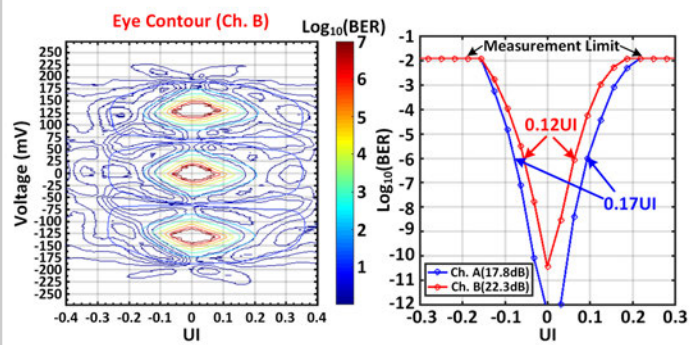


Figure 30.5.4: Measured RX eye diagram and bathtub curve at 56.25Gb/s using 17.8dB and 22.3dB loss channels with a PRBS31 Pattern.

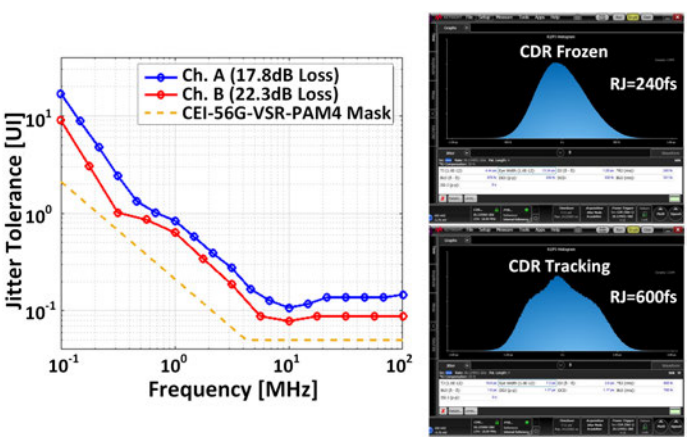
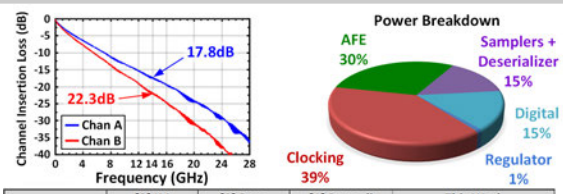


Figure 30.5.5: Measured receiver jitter tolerance and jitter histograms.



	[3] IM ISSCC 2017	[2] Peng ISSCC 2017	[1] Depaoli ISSCC 2018	This Work	
Technology	16nm FinFet	40nm CMOS	28nm FDSOI	7nm FinFet	
Data Rate (Gb/s)	56	56	64	56.25	
Modulation	4PAM	4PAM	4PAM	4PAM	
Loss (dB)	10	24	16.8	17.8	22.3
Equalization	TX FIR CTLE 10-tap DFE	TX FIR CTLE 3-tap DFE	TX FIR CTLE	TX FIR CTLE	
Min BER (pre-FEC)	~ 1E-12	< 1E-12	~ 1E-12	~ 1E-12	~ 1E-10
Horizontal eye @ 10^{-6} BER [UI]	0.2	0.36	0.19	0.17	0.12
Area [mm^2]	0.36	1.26	0.32	0.13	
Supply [V]	0.9/1.2/1.8	1/1.5	1	0.75/0.85/1.3 (Digital/Analog/Regulator)	
Power [mW]	230	382	180	104 / 79 (with / without regulator)	
Efficiency [pj/bit]	4.1	6.8	2.8	1.87 / 1.41 (with / without regulator)	

Figure 30.5.6: Channel insertion loss, power breakdown and comparison table.

Breakdown of Active Area

Block	Dimensions
AFE	97 µm x 205 µm
Latches & Deserializer	50 µm x 190 µm
Phase Interpolators	140 µm x 127 µm
Reference DACs	90 µm x 90 µm
Regulator Core	98 µm x 40 µm

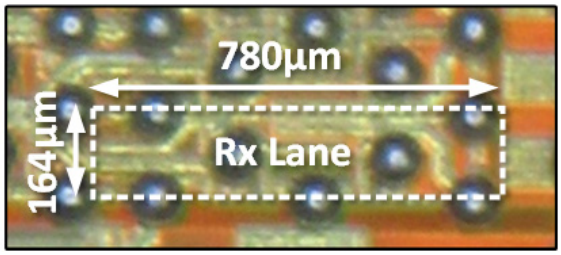


Figure 30.5.7: 7nm FinFET chip micrograph and area breakdown.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

30.6 A 6V Swing 3.6% THD >40GHz Driver with 4.5× Bandwidth Extension for a 272Gb/s Dual-Polarization 16-QAM Silicon Photonic Transmitter

Abdelrahman H. Ahmed[1,2], Daihyun Lim[2], Abdellatif Elmoznine[2], Yangjin Ma[2], Tam Huynh[2], Christopher Williams[2], Leonardo Vera[2], Yang Liu[2], Ruizhi Shi[2], Matthew Streshinsky[2], Ari Novack[2], Ran Ding[2], Rick Younce[2], Rafid Sukkar[2], Jose Roman[2], Michael Hochberg[2], Sudip Shekhar[1], Alexander Rylyakov[2]

[1]University of British Columbia, Vancouver, BC, Canada
[2]Elenion Technologies, New York, NY

Traditionally, the high-performance optical coherent communication TX has been a discrete assembly based on $LiNbO_3$ modulators and III-V drivers. While delivering high bandwidth (BW) and linearity, such a platform is bulky and does not work for high-volume or intra-datacenter applications. Silicon photonics (SiPh) offers a Si-based platform for next-generation transceivers by integrating all required optical functions. But even in existing SiPh-based commercial modules, the driver generally remains a III-V-based chipset, impeding the path to an all-silicon solution. The challenge for an all-silicon-based coherent optical TX is in the simultaneous requirement of high differential voltage swing (V_{ppd}), linearity and BW. In this work, we present a 130nm SiGe driver achieving a $6V_{ppd}$ swing, 3.6% THD and small-signal BW over 40GHz. Co-packaged with a SiPh transceiver, the driver enables the same level of performance as $LiNbO_3$ modulators with III-V drivers and demonstrates 272Gb/s dual-polarization (DP)-16QAM transmission. This is enabled using (1) circuit techniques that achieve a BW extension ratio (BWER) of 4.5× for the Mach-Zehnder modulator (MZM) driver while simultaneously achieving large swing (V_{ppd}=6V), high linearity (THD=3.6%) and mitigating breakdown voltage (BV) and reliability concerns with large V_{ppd}; (2) pre-emphasis control in the driver output stage and gain control in the pre-driver VGA to compensate for the electro-optic (E/O) BW of the TX over corners; and (3) a monolithic integration of MZMs with a polarization rotator (PR) optimized for high-BW, low-crosstalk and co-integration with drivers for DP-QAM and DP-QPSK operation.

New coherent standards seek to expand the existing orthogonality of amplitude (NRZ/PAM-4) and wavelength (WDM) with phase (QPSK/QAM) and polarization (DP). Figure 30.6.1 shows the block diagram of a DP-16QAM TX comprising 4 MZMs, optical splitters and combiners, PR and phase shifters, and 4 drivers. An MZM with two arms of equal length (L) works on the principle of interference – for 0 (π) differential phase shift across the two arms, the output intensity is constructive (destructive). When DC-biased at the null-intensity operating point of the MZM, the phase shift can be switched between $0 \leftrightarrow \pi$ by modulating the p-n junctions on these arms by a push-pull electrical driver sweeping the differential voltage over π radians (V_π). The electrodes are designed as Z_0=44Ω transmission lines with L≈3.5mm, requiring $6V_{ppd}$ for an extinction ratio of 15.8dB. With a PAM-4 output, the driver enables 4 *I*-phase constellation points out of the MZM. Another MZM in parallel but concatenated with a $\pi/2$ optical phase shift creates 4 *Q*-phase points. A 16-QAM signal is thus generated, and a PR in the upper branch leads to DP-16QAM in X and Y planes. A DP-QPSK is similarly generated through NRZ outputs from the 4 drivers.

The driver has significant design constraints. It must: (1) support $6V_{ppd}$ without BV/reliability concerns in a 1.65V-BV_{CEo} BiCMOS process; (2) drive a low-Z_0 T-line electrode, usually double-terminated with $R_T = R_L = Z_0$; (3) support a nominal electrical BW of 50GHz and an electro-optical (E/O) BW of 40GHz; (4) have a linearity of < 6% THD to permit amplitude modulation and retain pulse shaping; (5) have sufficient slew-rate for large-signal swings; and (6) minimize power consumption. Prior-art in high-swing driver design has mostly been limited to non-linear NRZ switching stages in lower-cost BiCMOS processes [1-3] or reduced-V_{ppd} for a linear high-BW design using a state-of-the-art 55nm BiCMOS process (high f_T MOSFETS and HBTs) [4]. Furthermore, E/O measurements are seldom shown.

Figure 30.6.2 shows the block diagram of the driver comprised of 3 stages of pre-driver including VGA, the output stage and an offset-cancellation loop. The pre-driver has sufficient pre-emphasis using RC degeneration to account for cable and measurement losses at its input. The VGA provides 10dB of gain tuning. As the driver and the SiPh chip are flip-chipped to a ceramic substrate permitting short connections, the driver on-chip termination is made larger (R_L=200Ω, R_T=Z_0=44Ω) to reduce power without significant near-end reflections. The driver output stage (Fig. 30.6.2) consists of a main and an auxiliary path. The main path must have adequate BW for required data rates, and adequate gain for the input from the pre-driver. The output branch current I_1 is governed by V_{ppd} across the load, $R_{eff} \approx R_L || R_T$. Thus, a large V_{ppd} requires a large I_1, resulting in a large Q_1 sized to operate near its f_T for high-speed. Large Q_1 also leads to large capacitance C_{BE1} restricting BW. To divide the large V_{ppd} equally on Q_1 and Q_2, and recalling V_{BE2} ~ constant, the base of Q_2 is fed a signal using an auxiliary path to ensure $V_{CE1} \approx V_{CE2}$ < BV [1]. The auxiliary path has low gain but large BW due to a highly degenerated common-emitter Q_2 branch. C_H and R_H are added to ensure the two paths have equal BW/delay, but the additive action of two paths driven by the same input node leads to BW shrinkage compared to a regular cascoded output branch [1]. Note that small-signal gain/BW and linearity considerations reflect the overall performance even for a large-swing driver [2].

However, as Fig. 30.6.3 shows, the pole (RC) at the input node can be split to shield the input capacitance of the auxiliary path. Capacitor splitting is often done using series-peaked inductors. Herein, we propose resistor-based capacitor splitting, which although not so effective for an amplifier with single output node, can lead up to BWER=2 for our two-path additive design. This resistor-based capacitor splitting technique creates a pole-zero pair at higher frequency, but more importantly, also pushes the dominant pole further. A higher split-ratio (r/R) leads to higher BWER, but also reduces the signal into the auxiliary path that must be compensated by I_3R_H. A split-ratio of r/R = 0.5 leads to BWER=1.5.

Figure 30.6.4 shows the design evolution. After optimizing C_H for large-signal slew-rate considerations, a BWER=1.3 is attained with capacitor splitting without any power, gain or area overhead. Next, resistor-degeneration (R_{E1}) is implemented for Q_1 to increase linearity and BW, as g_{m1} and C_{BE1} are decreased by $1+g_{m1}(R_{E1}/2)$. It is important to note that BW extension in the linear high-V_{ppd} driver is more important than gain reduction, as the gain can be easily compensated in the pre-driver with more relaxed design constraints. To maintain delay/BW/linearity matching, emitter degeneration (R_{E3}) is also implemented in the auxiliary path with $1+g_{m3}(R_{E3}/2) \approx 1+g_{m1}(R_{E1}/2)$. Thus, overall BWER=2.3 and THD is improved from 6.7% to 5.1%.

Finally, emitter degeneration (R_{E5}) is used to further improve the linearity for an overall THD of 3.6%, and a pre-emphasis capacitor C_{E5} is used to introduce a low-frequency zero and push the overall BWER to 4.54. Overall, the combination of these techniques enhances the electrical BW from 11GHz to 50GHz and reduces THD by 47%, from 6.7% to 3.6% for the high-V_{ppd} driver. A tunable C_{E5} along with the VGA ensures an E/O BW of 40GHz across corners.

Die micrographs are shown in Fig. 30.6.7. Two 2-channel driver dies and a SiPh IC are flip-chipped to a ceramic substrate and co-packaged. Figure 30.6.5 shows the measured TX E/O response, along with constellations for 64Gbaud DP-QPSK (BER<1E-6) and 34Gbaud DP-16QAM (BER<1E-3) with aggregate data rates of 256Gb/s and 272Gb/s, respectively. In both cases the BER is well below FEC threshold. Single-polarization 45Gbaud 16QAM has also been measured. Measurements are done using a hybrid-Si laser and BiCMOS TIAs, detailed in previous work [5]. Required optical SNR (ROSNR) at the RX is an overall metric to characterize the coherent link performance. For a target BER of 3E-2 (pre-FEC) and an RX input optical power range from -25 to 0dBm, the 34Gbaud DP-16QAM performance of our TX matches that of a commercial fixture comprising a 32Gbaud III-V linear driver and 35GHz $LiNbO_3$ MZM. Figure 30.6.6 shows the comparison table for the driver, and summarizes the TX performance. To the best of our knowledge, this is the first publication describing high-swing linear BiCMOS drivers co-integrated with a SiPh IC to support >200Gb/s DP-16QAM.

References:

[1] S. Mandegaran and A. Hajimiri, "A Breakdown Voltage Multiplier for High Voltage Swing Drivers," *IEEE JSSC*, vol. 42, no. 2, pp. 302-312, Feb. 2007.
[2] C. Knochenhauer, et al., "A Compact, Low-Power 40-GBit/s Modulator Driver With 6-V Differential Output Swing in 0.25µm SiGe BiCMOS," *IEEE JSSC,* vol. 46, no. 5, pp. 1137-1146, May 2011.
[3] L. Vera and J. R. Long, "A 40-Gb/s SiGe-BiCMOS MZM Driver With 6-V_{p-p} Output and On-Chip Digital Calibration," *IEEE JSSC*, vol. 52, no. 2, pp. 460-471, Feb. 2017.
[4] A. Zandieh, et al., "Linear Large-Swing Push–Pull SiGe BiCMOS Drivers for Silicon Photonics Modulators," *IEEE Trans. on Microw. Theory and Tech.*, vol. 65, no. 12, pp. 5355-5366, Dec. 2017.
[5] A. Novack, et al., "A Silicon Photonic Transceiver and Hybrid Tunable Laser for 64 Gbaud Coherent Communication," *Optical Fiber Comm. Conf.*, pp. 1-3, Mar. 2018.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

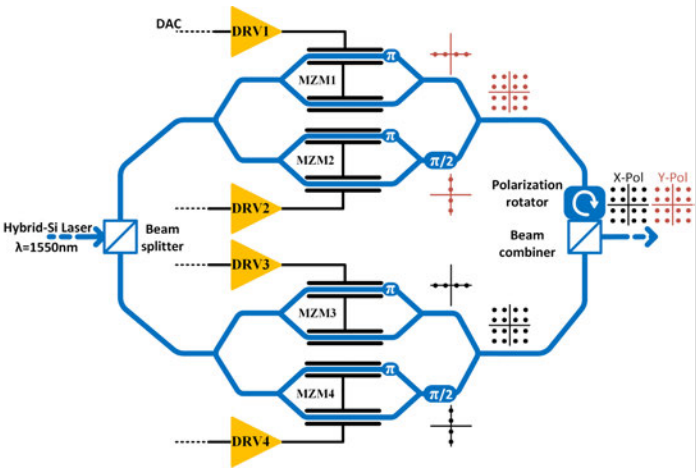


Figure 30.6.1: A dual-polarization 16-QAM TX: 4 high-swing linear differential push-pull drivers co-integrated to the silicon-photonic IC.

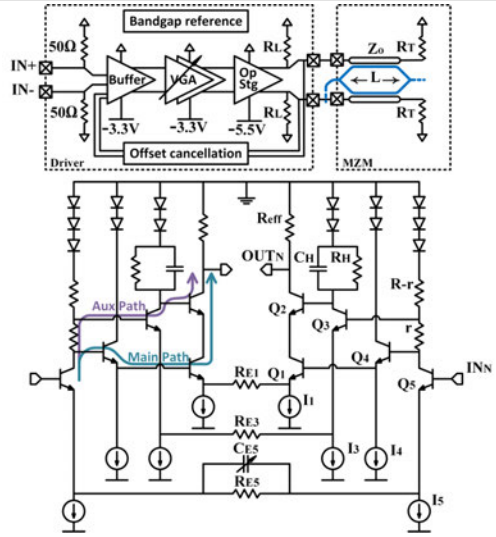


Figure 30.6.2: Block-level diagram for the driver (top) and schematic for the driver output stage (bottom).

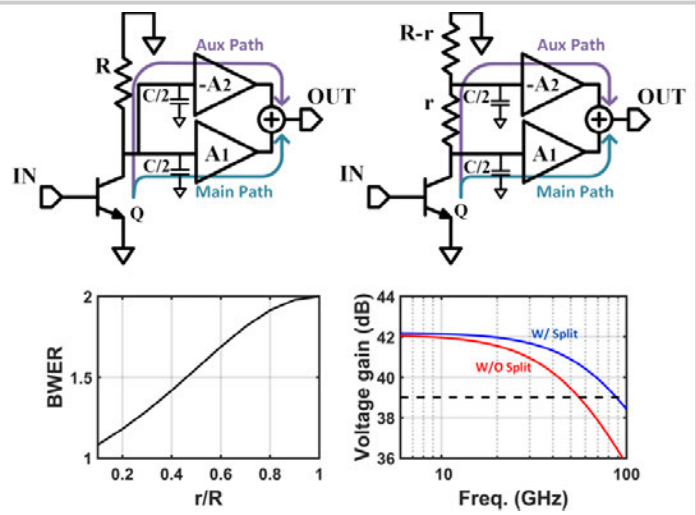


Figure 30.6.3: A resistor-based capacitor-splitting BW extension technique, where the split-ratio (r/R) dictates the BWER, and simulated VOUT/VIN with and without split.

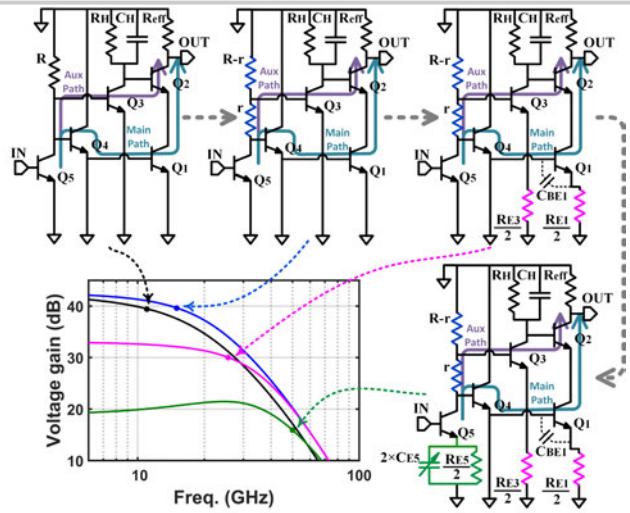


Figure 30.6.4: Driver design for improving BWER and THD: Breakdown-voltage doubler → capacitor-splitting → emitter-degeneration → emitter-degeneration and pre-emphasis.

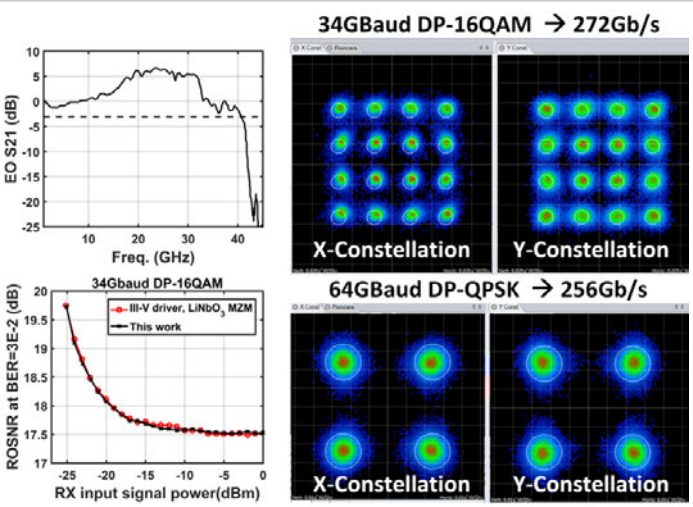


Figure 30.6.5: Measured E/O S21, constellation for 272Gb/s DP-16QAM, 256Gb/s DP-QPSK, and required optical SNR vs. RX input signal power for the TX.

Performance comparison: High-swing driver

	[1]	[2]	[3]	[4]	This work
Gbaud/s	10	40	40	42.6	45/64
Signaling	NRZ	NRZ	NRZ	8-PAM	4-PAM/NRZ
Driver Type	Switching	Switching	Switching	Linear	Linear
Elec. BW (GHz)	N/R	33.7	N/R	57.5	>40[1]
E/O-BW (GHz)	N/R	N/R	N/R	N/R	40
Gain (dB)	19	13	26	14.5	20~30 (w/ VGA)
Supply[2] (V)	6.5	5.5	5	6	5.5
$V_{PP,d}$ (V)	7.6	6	6	4.8	6
THD	N/A	N/A	N/A	-12.5dB ≈ 6% (@10GHz 6$V_{pp\text{-}diff}$)	3.6% (@1GHz 6$V_{pp\text{-}diff}$)
P_{DC} (W)	3.7	1.35	1.92	0.82	1
Area (mm²)	1.2	0.72	3	0.6	1.6
BiCMOS Tech.	N/R	0.25um	0.13um	55nm	0.13um
f_T (GHz)	120 (HBTs only)	180 (HBTs only)	200 (HBTs only)	300 (HBTs, MOSFETs)	250 (HBTs only)

[1] Measured on a 40GHz Vector Network Analyzer [2] Of output stage

TX performance summary for this work

Modulation	QPSK	DP-QPSK	16QAM	DP-16QAM
Gbaud/s	64	64	45	34
Data Rate (Gb/s)	128	256	180	272
# Driver modules	2	4	2	4
P_{DC} (W)	2	4	2	4
Energy/bit (pJ/b)	15.6	15.6	11.11	14.7

Figure 30.6.6: Driver performance summary and comparison (top), TX performance summary (bottom).

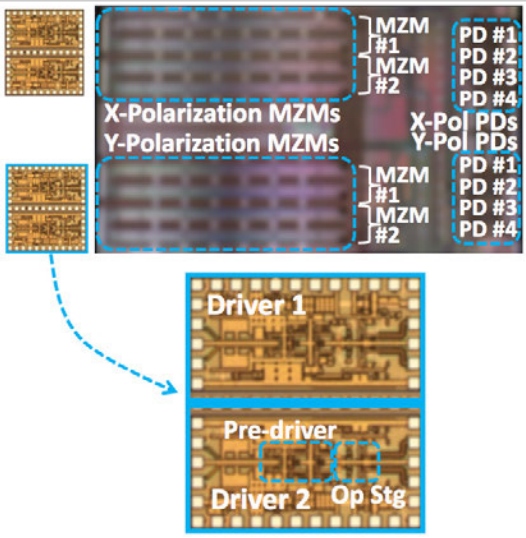


Figure 30.6.7: Die micrographs of the SiPh IC along with 2×2 driver ICs (top), and a zoom-in on the driver IC (bottom).

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

30.7 An 8b Injection-Locked Phase Rotator with Dynamic Multiphase Injection for 28/56/112Gb/s Serdes Application

Yi-Chieh Huang, Bo-Jiun Chen

MediaTek, Hsinchu, Taiwan

Growing traffic in data centers imposes multiple challenges on serial link design with higher speed and stringent power requirements. Clocking is one major challenge due to the significant portion of total power that clock distribution consumes in a serial link. Recent publications [1-4] deploy either a delay-locked loop (DLL), or an Injection-locked ring oscillator (ILRO) combined with phase interpolators (PIs) to generate multiphase clocks. For example, a 28Gb/s quarter-rate NRZ receiver requires an 8-phase 7GHz clock while a 56/112Gb/s DSP-based PAM-4 receiver requires a 4/8-phase 7GHz clock. In this work, the proposed injection-locked phase rotator (ILPR) can provide multiphase clocks without multiple PIs or a multiphase input clock.

As shown in Fig. 30.7.1, the ILPR consists of an input buffer, a phase selector, an injection NMOS switch array, a 64-phase ICO, a VCO-based ADC and a frequency-locked loop (FLL). A phase rotator (PR) code EN[31:0] controls the phase selector to connect input clocks CK_INJ and CKB_INJ to the target injection NMOS switches. As shown in the figure, VCM is sampled from ICO differential clocks by 2 NMOS switches during injection. The following VCO-based ADC and FLL are used to background calibrate the frequency of the 64-phase ICO by monitoring VCM. A 16-phase oscillator in the VCO-based ADC operates at around 500MHz to save power. This 16-phase oscillator is used to enhance the resolution of the ADC and minimize the quantization noise.

Figure 30.7.2 shows the proposed 64-phase ICO. This ICO is realized by coupling multiple 3-stage ring oscillators with NMOS resistors. Therefore, its maximum frequency is close to that of a 3-stage oscillator. Similar to [5], a phase coupler is employed to ensure phase relation. However, the pull-down NMOS as a phase coupler in [5] induces short current and imbalanced rise/fall time compared to an NMOS resistor. Unlike conventional PIs, The intrinsic phase balance ensures high linearity if the layout is well taken care of. Eight clocks with the phases uniformly spaced by $\pi/4$ are generated by the AC-coupled buffers. The rest of the unconnected nodes are tied with dummy capacitors to maintain the linearity of ILPR. Besides, in order to support a wide frequency range, the inverter strength and the NMOS resistors are programmable by EN_HB. In the 5-to7GHz high-frequency band, EN_HB is tied high to increase the transconductance of the inverter and decrease the NMOS resistance to enhance the coupling factor of the adjacent phase to desensitize phase mismatch. In the 2-to-5GHz low-frequency band, EN_HB is tied low to slow down the ring oscillator frequency. Figure 30.7.2 shows the layout floorplan. The minimal gate length of 16nm and 2 poly pitch per slice are chosen to minimize the routing parasitic resistors and capacitors. The core area occupies only 26μm by 19μm.

As illustrated in Fig. 30.7.3, the phase selector connects CK_INJ and CKB_INJ to the selected injection NMOS switches by digital logic. A NOR gate and a pull-down NMOS are used to avoid the floating gate when both injection NMOS switches are off. The injection strength is programmable by tuning the bias current I_LR through a diode-connected load. To digitally rotate the output phase of the 64-phase ICO, a dynamic multiphase injection method is shown in Fig. 30.7.3. The arrows represent the injection pulse from either CK_INJ or CKB_INJ. For example, when EN[4,5,6,7] = 1, four successive differential clocks P[4,5,6,7] and P[36,37,38,39] are injected by CK_INJ while four counterparts with $\pi/2$ phase shift are injected by CKB_INJ. The phase relationship is illustrated in Fig. 30.7.3. The center of the injection pulse CK_INJ will align with the average phase of four differential clocks P[4,5,6,7] and P[36,37,38,39]. In general, four PR codes among EN[31:0] are enabled to select four differential clocks. The locked phase by CK_INJ is defined as the average phase of these four differential clocks. In this case, the locked phase is (4+5+6+7)/4 = 5 + 2/4. By changing EN[4,5,6,8] = 1, the locked phase becomes (4+5+6+8)/4 = 5+3/4. Therefore, a 1/4 fractional step can be created. Finally, the 64-phase ICO can be digitally rotated with 6b integer steps and 2b fractional steps.

Figure 30.7.4 shows the proposed FLL and the measured VCM curve and correlation slope. If $\Delta f \neq 0$, the static phase error is created to compensate the frequency error, where $\Delta f = f_{ICO} - f_{CK_INJ}/2$. Because of the injection NMOS switch, VCM will be determined by the lower voltage of P[n] or P[n+32]. When $\Delta f = 0$, the injection pulse is aligned with the zero crossing of the selected differential phases and VCM will reach the peak. Therefore, the transfer curve VCM versus Δf is shown in Fig. 30.7.4. To correct Δf, a current with PRBS x[t] is injected into a 64-phase ICO. For the transfer curve VCM versus Δf, the slope is positive if $\Delta f<0$ and vice versa. The VCM is modulated with slope x[t] with an additional noise n[t] due to PR rotation. The following mixer multiplies x[t] by the digitized VCM to recover the slope. By using an FIR filter to filter out the noise, the slope correlation can be obtained. The polarity of the slope can be easily determined by a slicer. Then, the FLL_CODE adjusts the current into the 64-stage ICO until Δf converges to zero. The measured VCM and slope correlation are also shown in Fig. 30.7.4. The slope correlation is proportional to the derivative of VCM curve. The locking range can be calculated by the corresponding FLL codes of maximal and minimal slope, which are -80 and 73, respectively. It is equivalent to an 8.6% locking range.

Figure 30.7.5 shows INL and DNL measurement results at 7GHz for typical and four other extreme corners. The typical case shows INLpp = 1.14 LSB, while other corners range from 1.06 to 1.84LSB. The measured spectrum with -1301ppm rotation shows many additional fractional spurs compared to the case without rotation due to nonlinearity. The integrated additional spurious tone is -39.4dBc, equivalent to 243fs or 0.43LSB rms deterministic jitter.

The measured phase noise is shown in Fig. 30.7.6. The integrated rms jitter of the ILPR at 7GHz from 4MHz to 3GHz is 254fs, where the 4-to-100MHz profile follows the input clock of 14GHz CK_PLL from LCPLL with a -6dB shift. The noise floor partly comes from testing buffer chain to the output pad. The noise power below 4MHz is negligible given that the CDR has a bandwidth larger than 4MHz. The comparison table is shown in Fig. 30.7.6. This work shows superior power and area performance thanks to fewer high-speed analog blocks. The total power consumption is 11.4mW, including 3.6mW for the input buffer and phase selector, 6.9mW for the 64-phase ICO and 0.9mW for the VCO-based ADC. The INLpp can achieve 1.14LSB using only a differential clock. Figure 30.7.7 shows the die photograph. The total area occupies 0.0043mm^2.

Acknowledgements:
The authors would like to thank Ching-San Wu, Jason Hsu and Yan-Bin Luo for resource support. We also appreciate Prof. Shen-Iuan Liu for valuable discussion and review.

References:
[1] B. Zhang, et al., “A 28Gb/s Multi-Standard Serial-Link Transceiver for Backplane Applications in 28nm CMOS” *ISSCC*, pp. 52-54, Feb. 2015.
[2] S. Chen, et al., “A 4-to-16GHz Inverter-Based Injection-Locked Quadrature Clock Generator with Phase Interpolators for Multi-Standard I/Os in 7nm FinFET” *ISSCC*, pp. 390-392, Feb. 2018.
[3] E. Monaco, et al., "A 2–11 GHz 7-Bit High-Linearity Phase Rotator Based on Wideband Injection-Locking Multi-Phase Generation for High-Speed Serial Links in 28-nm CMOS FDSOI," *IEEE JSSC*, vol.52, no.7, pp. 1739-1752, July 2017.
[4] E. Depaoli, et al., "A 4.9pJ/b 16-to-64Gb/s PAM-4 VSR Transceiver in 28nm FDSOI CMOS," *ISSCC*, pp. 110-112, Feb. 2018.
[5] A. Matsumoto, et al., “A Design Method and Developments of a Low-Power and High-Resolution Multiphase Generation System,” *IEEE JSSC*, vol.43, no.4, pp. 831-843, July 2008.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

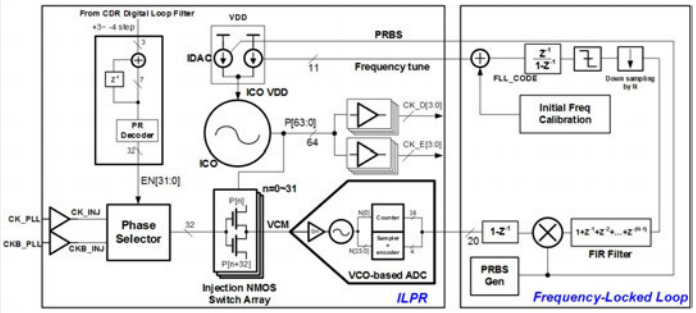


Figure 30.7.1: Block diagram of the injection-locked phase rotator and frequency-locked loop.

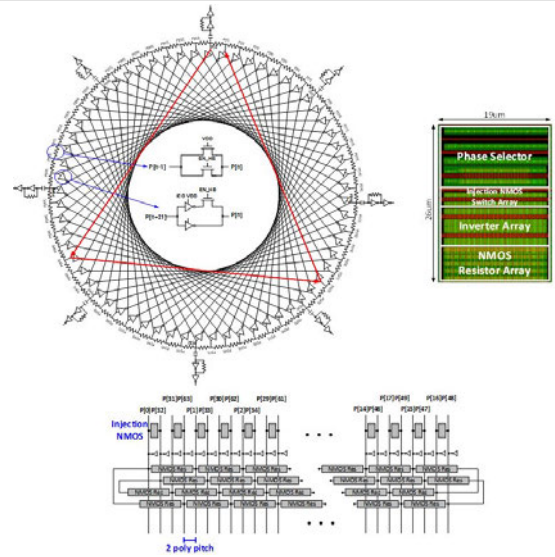


Figure 30.7.2: Proposed 64-phase ICO and its layout floorplan.

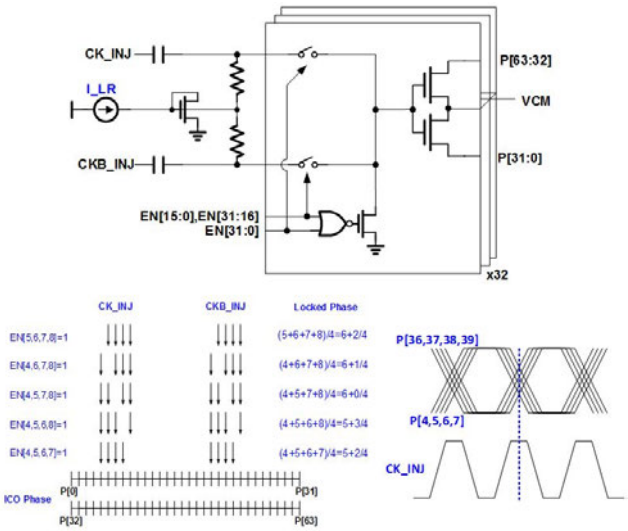


Figure 30.7.3: Phase selector, injection NMOS switch array and principle of dynamic multiphase injection.

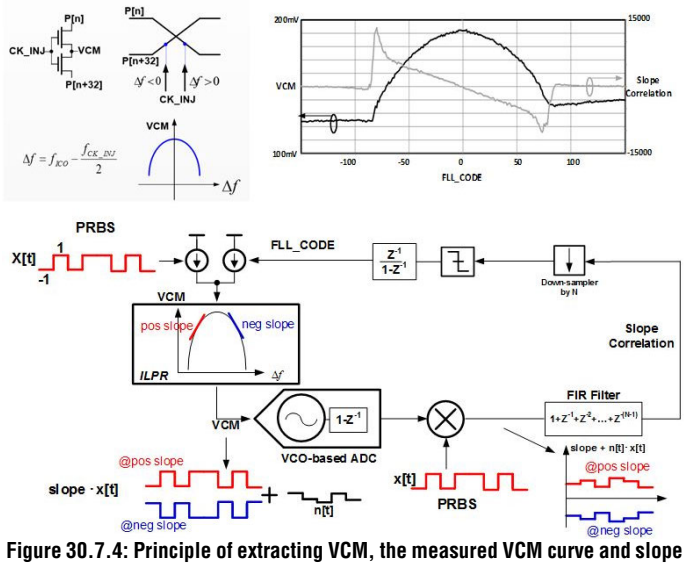


Figure 30.7.4: Principle of extracting VCM, the measured VCM curve and slope correlation versus FLL_CODE, and conceptual diagram of FLL.

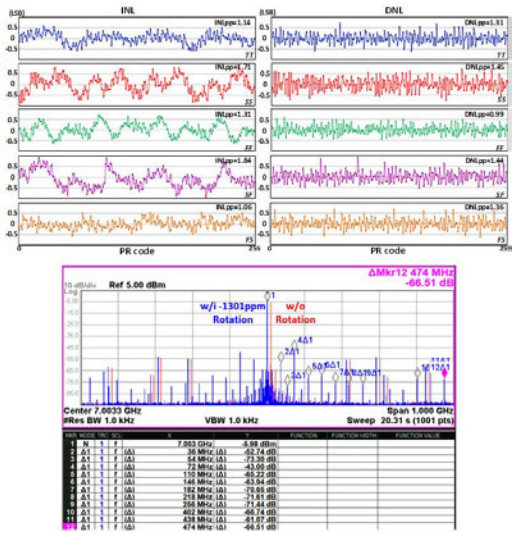


Figure 30.7.5: Measured peak-to-peak INL and DNL of 5 corners and spectrum w/o and w/i -1301ppm rotation.

	[2]	[3]	This Work
Technology	7nm FinFET	28nm FDSOI	16nm FinFET
RX Clocking Architecture	ILRO + CMOS PI x 2	ILRO + CML PI x 2	ILPR x 1
RX Clocking Speed	Half Rate	Half Rate	Quarter Rate
INLpp/DNLpp @TT corner	1.44/0.87	1.1/0.5	1.14/1.31
INLpp/DNLpp @extreme corner	N/A	N/A	1.06-1.84/ 0.99-1.45
Random Jitter	143fs (0.1M-8GHz)	139fs* (0.01M-1GHz)	254fs** (4M-3GHz)
Resolution	7 bits	7 bits	8 bits
Frequency Range	4-16GHz	2-11GHz	2-7 GHz
# of Output Phase	4	4	8
Supply	1.2/0.88v	1.0v	1.2v
Power @ supported data rate	34.8mW @ 32Gbps	22.2mW @ 22Gbps	11.4mW @ 28Gbps
Area	0.105mm²***	0.022mm²	0.0043mm²

*ILRO only **Noise is dominated by LCPLL & Testing buffer *** Include LCVCO

Figure 30.7.6: Measured LCPLL and ILPR phase noise and comparison to prior arts.

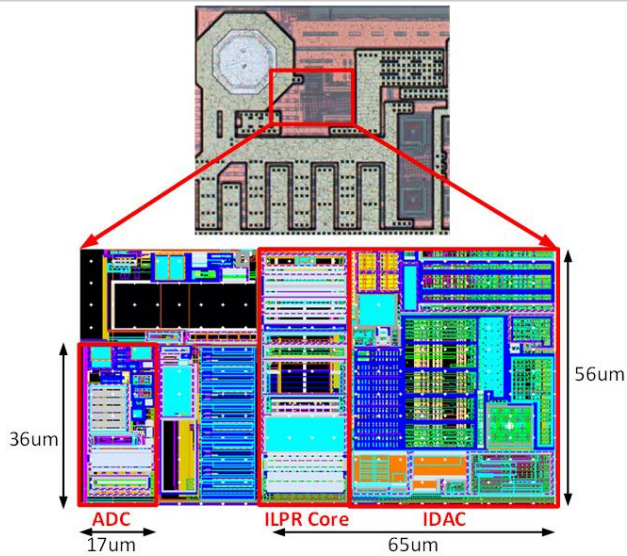


Figure 30.7.7: Die photograph.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

30.8 A 0.65V 12-to-16GHz Sub-Sampling PLL with 56.4fs$_{rms}$ Integrated Jitter and -256.4dB FoM

Zhao Zhang, Guang Zhu, C. Patrick Yue

Hong Kong University of Science and Technology, Hong Kong, China

Lowering supply voltage is an effective way to reduce circuit power consumption, especially in digital-centric system-on-chips (SoC). For low-jitter phase-locked loops (PLL) required in high-speed serial links and data converters, operation under low voltage is highly desirable such that the PLL can be readily integrated in a low-voltage (LV) SoC without a dedicated high-voltage supply. Among the various PLL architectures, the sub-sampling PLL (SSPLL) [1-3] offers low jitter with a superior jitter-power product figure-of-merit (FoM) because of its inherent rejection of N^2 amplification of in-band phase noise induced by the charge pump (CP) and phase detector (PD), as reported in [1]. However, the design of an LV SSPLL (LVSSPLL) must overcome several daunting issues. First, the CP design suffers from limited voltage headroom, degraded current noise, and limited output voltage range. Although the type-I SSPLL [3] can avoid this issue by eliminating the CP, it suffers from limited phase-noise suppression of the voltage-controlled oscillator (VCO). Second, the high on-resistance of the sub-sampling PD (SSPD) under low voltage causes attenuation of the sampled clock swing and thus degrades the in-band phase noise of the SSPLL induced by the SSPD.

This paper presents a 0.65V LVSSPLL with 56.4fs$_{rms}$ jitter and -256.4dB FoM. It mitigates the issue of limited CP output voltage range by adopting a hybrid dual-path loop architecture. By employing a combination of LV circuit techniques in designing the sub-sampling CP (SSCP), the SSPD, and a digitally controlled capacitor array (DCCA) for the class-C VCO, the LVSSPLL jitter and voltage headroom requirement are reduced simultaneously.

Figure 30.8.1 shows the block diagram of the LVSSPLL. The hybrid dual-path topology includes a proportional path (P-path) with a discrete-time switched capacitor low-pass filter (DTSCF) and an integral path (I-path) with a continuous-time passive RC low-pass filter (CTRCF). The P-path is responsible for phase tracking and thus must have high linearity and low noise. The I-path is for frequency acquisition and needs to produce a wide control voltage (V_I) tuning range to cover the desirable frequency locking range, but its nonlinearity can be tolerated. The P-path gain $G_p(s)$ dictates the loop bandwidth if it is much larger than the I-path gain $G_I(s)$. So, the overall loop stability is not sensitive to the nonlinearity in $G_I(s)$, which is caused by the variation of transconductance $g_{m,I}$ over V_I variation due to the limited output voltage range of the I-path SSCP (I-SSCP). With $G_p(s)$ set to be much higher than $G_I(s)$, the in-band phase noise of the PLL is dominated by the noise of the P-path SSCP (P-SSCP), which is suppressed by using the structure of the P-SSCP with the DTSCF. When the PLL is in lock, the P-path control voltages, V_{P+} and V_{P-}, are kept nearly constant regardless of the output frequency. Therefore, the P-SSCP only requires a small output voltage range, which in turn improves its linearity. The issue of limited CP output voltage range is relaxed by the hybrid dual-path loop structure.

Figure 30.8.1 also shows the frequency-locked loop (FLL) that is needed for the initial frequency acquisition [1]. Two current-mode logic (CML) based divide-by-2 dividers (DIV2) are used as the prescaler to operate at up to 16GHz under only 0.65V. This limits the SSPLL output frequency resolution to $4F_{REF_FLL}$. To improve the output frequency resolution from $4F_{REF}$ to F_{REF}, we set the FLL reference frequency F_{REF_FLL} to $0.25F_{REF}$ by using a divide-by-4 divider (DIVIN/4).

Figure 30.8.2 shows the P-SSCP. Compared with the conventional SSCP [1], the P-SSCP only has two stacked MOSFETs (M1/M2 and the current source) that operate in the saturation region. Only M1 and M2 contribute noise to the P-SSCP output because the output is differential. So, not only is the voltage headroom relaxed, but also the noise is reduced. Since the P-SSCP only has a current discharge path, its output is reset and pre-charged to V_{DD} every clock cycle. Hence the DTSCF, instead of a CTRCF, is required in the P-path. The key waveforms in the P-path are presented in Fig. 30.8.2. First, C_S is pre-charged to V_{DD} when RESET is high. Next, the voltage difference as well as the phase of the sampled differential clock is converted to a voltage (V_{CP+}-V_{CP-}) by discharging C_S when PUL is high. Finally, when HOLD is high, the voltage of C_S is transferred to C_H to create the P-path VCO control voltage (V_{P+}-V_{P-}). The I-SSCP utilizes a folded input stage to relax voltage headroom requirements. When the PLL is in lock, the output current of the I-SSCP is zero, which means the sampled input differential voltage (V_{SAM+}-V_{SAM-}) is around 0V. At the same time, (V_{P+}-V_{P-}) is also around 0V, and both V_{P+} and V_{P-} are kept almost constant regardless of the output frequency.

To solve the issue of the conventional SSPD mentioned above, the LV SSPD (Fig. 30.8.3) is adopted. It adopts two high-level boosted inverters (HBINV) to enhance the turn-on voltage of the gate of M1 and M2 in the SSPD. The HBINV is modified based on the delay cell in [5]. The working principle of the HBINV is shown in Fig. 30.8.3. The post-layout simulation results of the HBINV (Fig. 30.8.3) verify that the output of the HBINV is boosted to around 1V with a 0.65V supply. As a result, the on-resistance of M1/M2 is significantly reduced even using a 0.65V supply. This prevents the attenuation of the sampled clock swing and lowers the in-band phase noise of the SSPLL induced by SSPD. M3 and M4 are inserted to compensate for the clock feedthrough from M1 and M2, respectively. Simulation results show that the sampled clock swing has a weak dependency on the SSPD input common-mode voltage V_{CM} (from 0.1 to 0.5V). This helps to reduce the design complexity of the clock isolation buffer (ISOBUF) between the SSPD input and the VCO output (Fig. 30.8.1).

A class-C VCO is used due to its low phase noise and high power efficiency [2] (Fig. 30.8.4). A differential varactor bank is used for P-path tuning whereas a single-ended varactor bank is used for I-path tuning. A 6b DCCA is adopted for coarse tuning. At 0.65V supply, the high on-resistance of the switches in the DCCA degrades the VCO Q and thus worsens the phase noise. To relax this issue, we adopt an LV DCCA. The control logic of the DCCA is supplied by the peak detector output V_{PkD} [4]. The level shifter (LS) converts the low-swing input control signals to high-swing output signals. So, the switches of the DCCA (M_{SW}) can be turned on with higher gate voltage (>V_{DD}), and thus suppress the Q degradation. Two thick-oxide PMOS devices are used in the LS to make the static current of LS negligible when the input P[i] is at V_{DD} (0.65V), and thus avoid degrading V_{PkD}. When M_{SW} turns off, its drain and source are biased to V_{DD} instead of V_{PkD} to increase the off-state Q while reducing voltage stress on M_{SW}. Figure 30.8.4 also shows the cascode stage adopted for the ISOBUF to improve the isolation between the VCO and the SSPD. Thanks to the large-swing VCO output, the ISOBUF can generate a buffered clock with improved swing, and the first-stage CML DIV2 (Fig. 30.8.4) in the FLL can operate at up to 16GHz under a 0.65V supply.

Figure 30.8.5 shows the measurement results of the 0.65V SSPLL with the output F_{PLL} at 14GHz using a 200MHz F_{REF}. The measured rms jitter integrated from 1kHz to 100MHz is 56.4fs, the reference spur is -64.6dBc and the FoM is -256.4dB. The measured phase noise curve and spectrum are performed on the divide-by-2 signal of the VCO output, as shown in Fig. 30.8.1. The integrated rms jitter, reference spur, and FoM remains <61fs, <-63dBc, and <-256dB, respectively, across the frequency range from 12 to 16GHz with resolution of 200MHz. Thanks to the P-SSCP with reduced noise, the LV SSPD and the LV DCCA, low-jitter performance can be achieved under a 0.65V supply.

Figure 30.8.6 shows that among the state-of-the-art low-jitter PLLs, this work achieves the best FoM and FoM_T under the lowest supply voltage reported to date. The 40nm CMOS LVSSPLL occupies 0.234mm^2 and consumes 7.2mW, as shown in Fig. 30.8.7.

Acknowledgements:
This work was sponsored by the HKUST-Qualcomm Joint Innovation and Research Laboratory.

References:
[1] X. Gao, et al., "Sub-sampling PLL Techniques", *IEEE CICC*, pp. 1-8, Sept. 2015.
[2] A. T. Narayanan, et al., "A Fractional-N Sub-Sampling PLL using a Pipelined Phase-Interpolator With a FoM of -250 dB", *IEEE JSSC*, vol. 51, no. 7, pp. 1630-1640, July 2016.
[3] A. Sharkia, et al., "A 0.01 mm^2 4.6-to-5.6GHz Sub-Sampling Type-I Frequency Synthesizer with -254dB FOM", *ISSCC*, pp. 256-257, Feb. 2018.
[4] W. Deng, et al., "Class-C VCO With Amplitude Feedback Loop for Robust Start-Up and Enhanced Oscillation Swing", *IEEE JSSC*, vol. 48, no. 2, pp. 429-440, Sept. 2013.
[5] Y. Ho, et al., "A Near-Threshold 480MHz 78μW All-digital PLL With a Bootstrapped DCO", *IEEE JSSC*, vol. 48, no. 11, pp. 2805-2814, Nov. 2013.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

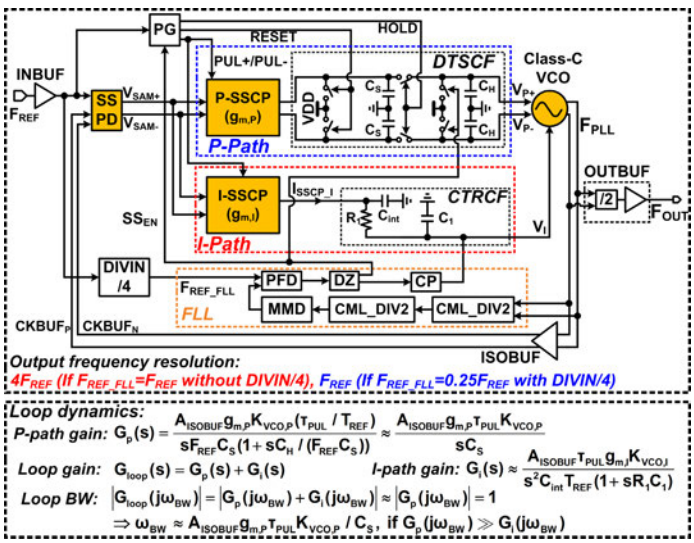


Figure 30.8.1: Block diagram of the proposed LVSSPLL.

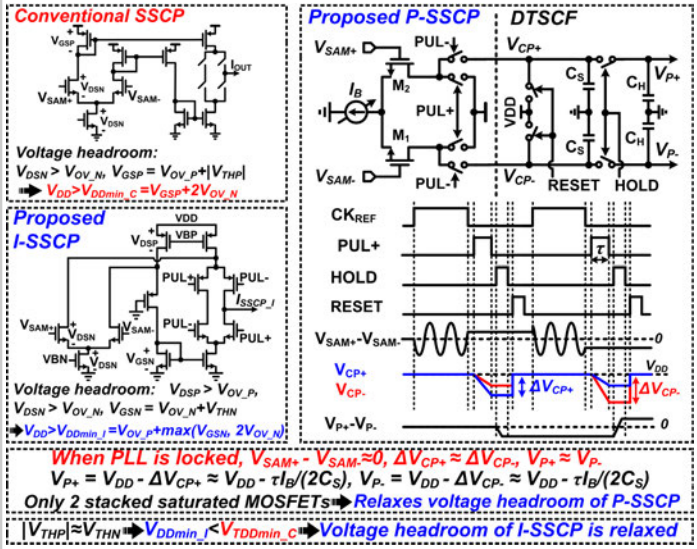


Figure 30.8.2: The P-SSCP, I-SSCP, and waveforms of the P-path.

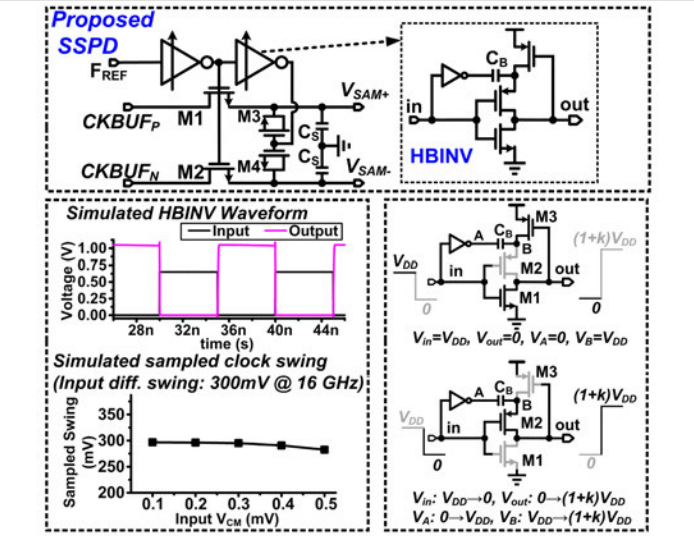


Figure 30.8.3: The SSPD and HBINV with its operation principle and simulation results.

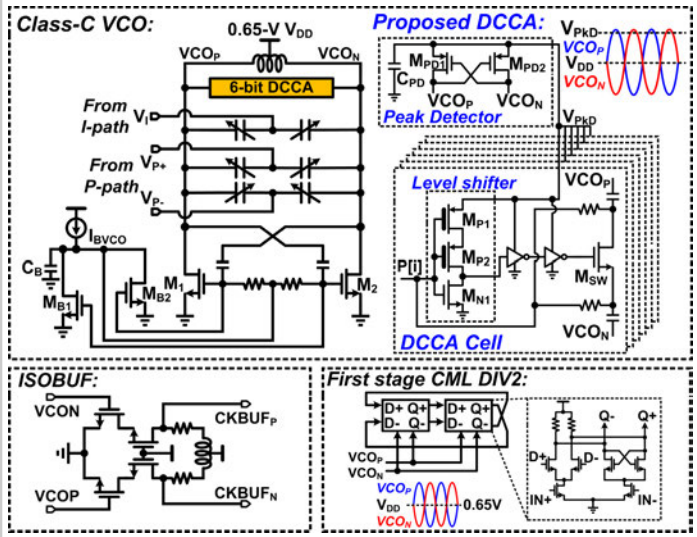


Figure 30.8.4: The Class-C VCO, DCCA, ISOBUF, and first-stage CML DIV2.

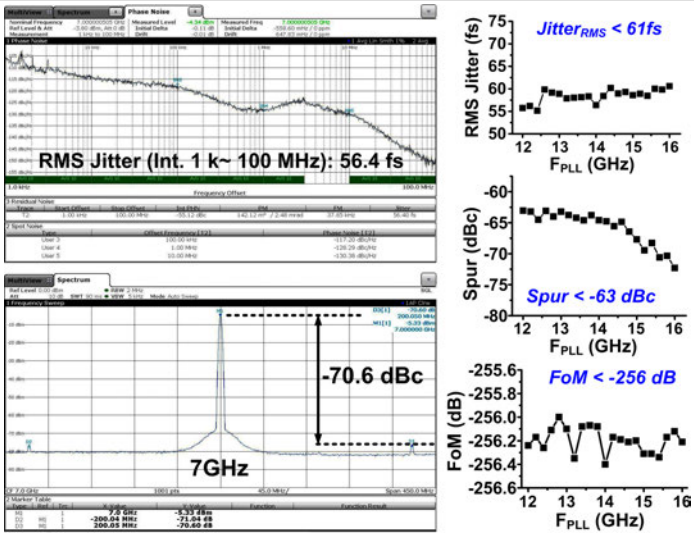


Figure 30.8.5: Measured phase noise, spectrum, rms jitter, spur level, and FoM.

	This work	A. Sharkia ISSCC'18 [3]	D. Turker ISSCC'18	M. Raj VLSI'17	A. Elkholy JSSC'18	K. H. Zeng TCAS-II'15	A. T. Narayanan JSSC'16 [2]
Architecture	SSPLL	SSPLL	CPPLL	SSPLL	ILCM (Integer-N)	ILPLL	SSPLL (Integer-N)
Process	40nm	65nm	16nm	16nm	65nm	40nm	65nm
Supply (V)	0.65	0.8	NA	0.9/1.8	0.9	1.1	1
Osc. Type	LC	LC	LC	LC	LC	LC	LC
Osc. Number	1	1	2	2	1	1	1
Out. Freq (GHz) /FTR (GHz)	14 /12-16	5 /4.6-5.6	12.5 /7.4-14	18 /9-18	7.36 /6.75-8.25	4.8 /NA	NA /4.34-4.94
Ref. Freq (MHz)	200	100	500	450	105-129	400	40
RMS jitter (fs) /Int. range	56.4 (1k~100MHz)	185.3 (10k~50MHz)	53.6 (10k~10MHz)	164 (1k~10MHz)	109 (10k~30MHz)	109.6 (1k~100M)	116 (10k~10MHz)
Ref. Spur (dBc)	-64.6	-64.1	-75.5	NA	-52	-61.59	-70.8
Power (mW)	7.2	1.1	45	29.2	2.65	3.04	2.82
FoM (dB)	-256.4	-254	-248.9	-241	-255	-254.4	-254
FoMT (dB)	-251	-247.2	-246.8	-239.3	-248	NA	-245.3
Core Area (mm²)	0.234	0.01	0.35	0.39	0.27	0.105	0.2

$FoM = 10\log((\frac{\sigma_{RMS}}{1s}) \times (\frac{P_{DC}}{1mW}))$ $FoM_T = FoM + 10\log(\frac{f_{Mid}}{f_{Max} - f_{Min}})$

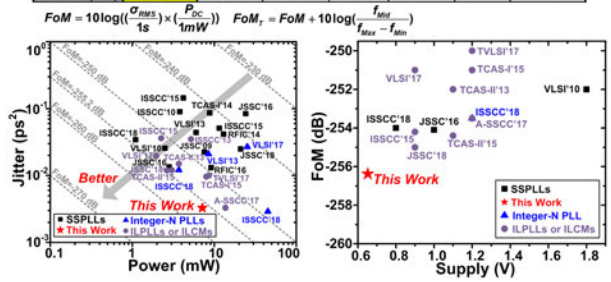
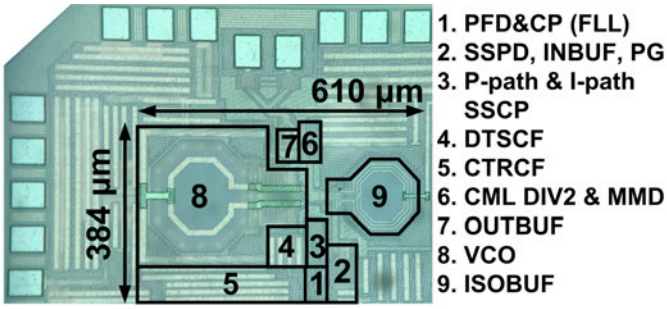


Figure 30.8.6: Performance summary and comparison.



Power Consumption						
Class-C VCO	SSPD, PG	ISOBUF	INBUF	SSCP	FLL	*Total*
3.05 mW	0.11 mW	1.4 mW	0.8 mW	0.29 mW	1.55 mW	*7.2 mW*

Figure 30.8.7: Chip photograph and power consumption breakdown.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

30.9 A 140fs$_{rms}$-Jitter and -72dBc-Reference-Spur Ring-VCO-Based Injection-Locked Clock Multiplier Using a Background Triple-Point Frequency/Phase/Slope Calibrator

Seyeon Yoo, Seojin Choi, Yongsun Lee, Taeho Seong, Younghyun Lim, Jaehyouk Choi

Ulsan National Institute of Science and Technology, Ulsan, Korea

An injection-locked clock multiplier (ILCM) is one of the best options to generate low-jitter high-frequency signals, while using a ring VCO. In the sense that the VCO jitter is removed periodically by the reference clock, S_{REF}, an MDLL also can be considered to be an ILCM. However, the most critical problem of ILCMs is that their jitter performance can degrade easily due to PVT. To ensure ultra-low jitter robustly, ILCMs must be equipped with background calibration that continuously adjusts the free-running frequency of the VCO, f_{VCO}, to stay close to the target frequency, Nf_{REF}, where N is the multiplication factor and f_{REF} is the frequency of S_{REF}, thereby minimizing the frequency error. To date, various calibrating methods have been developed, and many of them [1-4] successfully tracked f_{VCO} and prevented the degradation in random jitter, but none of the ILCMs succeeded in reducing reference spurs to a level comparable to that of PLLs. This is because even a slight phase error of an edge of the output signal, S_{OUT}, could result in large reference spurs [1]. There are three major causes of the phase error, the first of which is the *frequency error* due to drifts in f_{VCO} away from Nf_{REF}, which is the main target of prior calibrators. The second cause is the *phase offset*, which is generated due to any systematic errors of calibrators, such as mismatches in delay cells, input offsets of phase detectors, and limited resolution of digital circuits. Recently, state-of-the-art ILCMs [3, 4] successfully addressed these two causes, but they still could not reduce reference spurs to less than –65dBc. This is because, to date, none of the ILCMs has considered the third cause, i.e., the *slope modulation* of the edges of S_{OUT}, which occurs due to the periodic injection of S_{REF}.

In this work, we present a ring-VCO-based ILCM that can achieve a low reference spur as well as low rms jitter by using a background triple-point frequency/phase/slope calibrator (TP-FPSC). Since the proposed TP-FPSC provides three orthogonal mechanisms of calibration, it can remove all three causes of the generation of the phase errors in S_{OUT}, i.e., *frequency error*, *phase offset*, and *slope modulation*. So, all edges of S_{OUT} can be aligned at their ideal positions, minimizing the reference spur. In addition, since the calibration of the *frequency error* is designed to have a wide bandwidth, it can help in further suppressing the flicker noise of the VCO. Consequently, the ILCM in this work can achieve both low rms jitter and a low reference spur, while using little power and a small area.

The top of Fig. 30.9.1 shows how the unequal slopes of S_{OUT} edges, i.e., SL_{INJ} (when an injected S_{REF} edge occurs) and $SL_{NO\text{-}INJ}$ (when the VCO is free-running), cause a phase error on the kth edge, $1 \le k \le N$. Since the magnitude of the current of an inverter increases as the slope of the entering edge is steeper, the falling edge falls faster, thereby generating a smaller delay [5]. The change in the slope fades while passing through the next inverters, but the delay remains as a phase shift. So, when SL_{INJ} is steeper than $SL_{NO\text{-}INJ}$, the positions of all subsequent edges are shifted forward, causing phase errors and degrading reference spurs (second diagram of the top left). Conventional frequency calibrations consider these phase errors as a frequency error and try to equalize the distances between edges (third diagram). This can help in suppressing phase errors but cannot remove them sufficiently, since the phase shift keeps occurring at every injection. The significance of the effect of the *slope modulation* is shown in the simulation results of a phase error at $k = 2$ and the level of reference spur according to the ratios of SL_{INJ} to $SL_{NO\text{-}INJ}$. The bottom of Fig. 30.9.1 shows the overall architecture of the proposed ILCM using the TP-FPSC that includes the three independent control voltages for the three calibration mechanisms. First, V_{C1} of the VCO is to remove the *frequency error*. Second, V_{C2} of the voltage-controlled delay line (VCDL) is to remove the *phase offset*. Third, and most importantly, V_{C3} of the slope controller is to eliminate the *slope modulation*. In the TP-FPSC, all three V_{Cj}s ($1 \le j \le 3$) are generated by the multi-V_C generator (MVG) that extracts information for each V_{Cj} from the output of one bang-bang PD (BBPD), PD_{OUT}.

Figure 30.9.2 shows that the VCDL generates S_{VCDL} by delaying S_{OUT} by τ_{VCDL} (to be two periods of the free-running VCO, $2T_{VCO}$). The BBPD compares the phases of the corresponding edges of S_{OUT} and S_{VCDL}. Then, PD_{OUT} is sampled sequentially by FE_1, FE_4, and FE_2, generating error codes, D_{ER1}, D_{ER2}, and D_{ER3}, respectively. Each error code, D_{ERj}, enters the jth analog accumulator (A-ACC) and determines whether charges will be sourced to or extracted from C_{Lj}. Finally, during PS_{UD} from FE_5, all three V_{Cj}s are updated at the same time. Compared to digital accumulators requiring a large number of bits or a DSM, the proposed A-ACC can easily achieve a high resolution by decreasing the pulse width of PS_{UD}. Different from conventional charge pumps, the pulse width of PS_{UD} is fixed so that the digital information from the BBPD is accumulated in C_{Lj} in a discrete manner. The bottom of Fig. 30.9.2 shows how each cause of the phase errors can be detected at the same PD_{OUT}. First, when τ_{VCDL} is adjusted to $2T_{VCO}$, the effect of the *frequency error* of f_{VCO} appears in the difference between the phases, ϕ_{PD}, of the first edges of S_{OUT} and S_{VCDL}, so that it can be detected by sampling PD_{OUT} by FE_1. Second, when all *phase offsets* of the calibrator are referred to the inputs of PD, ϕ_{PD} becomes constant for all edges of S_{OUT} and S_{VCDL}. In this work, it is detected by FE_4. Third, when SL_{INJ} and $SL_{NO\text{-}INJ}$ are different, the effect of this *slope modulation* is detected by sampling PD_{OUT} by FE_2. (The *frequency calibration* has a much wider bandwidth than the *slope calibration*.) This mechanism also can detect any phase errors, caused at the moment of injection. In the diagrams of Fig. 30.9.2, each cause is assumed to be singled out, but in real situations their effects would be mingled with each other, and the three calibrations would be required to correct the detected errors concurrently. To secure the stability, the three calibrating loops have different bandwidths, such that the *frequency calibration* has the largest, whereas the *slope calibration* has the smallest. Figure 30.9.3 shows the schematics of the slope controller, the VCO, and the VCDL. The slope controller generates only a slight delay, and it has no impact on the jitter.

The ILCM of this work uses 11.0mW power and 0.055mm^2 area in 65nm CMOS. Figure 30.9.4 shows the phase noise (PN) of the 2.4GHz output signal with 100MHz f_{REF}. Due to the wide injection bandwidth, the PN of the VCO was reduced significantly. As the frequency-calibrating loop suppressed the VCO in-band noise, the ILCM achieved an ultra-low jitter of 136fs, when the TP-FPSC was enabled. The bottom of Fig. 30.9.4 shows this low-jitter performance was maintained over output frequencies and supply voltages in five different chips. The spectrums in Fig. 30.9.5 show that, as the TP-FPSC removed all three causes of phase errors, the reference spur was decreased to –72dBc. The bottom of Fig. 30.9.5 shows that the TP-FPSC can also ensure low reference spurs over output frequencies and supply voltages. The comparison table and the graph in Fig. 30.9.6 show that the proposed ILCM concurrently achieved the lowest reference spur and the lowest rms jitter, thereby resulting in excellent FoM$_{JIT}$. This work achieved the best FoM$_{JIT2}$ and FoM$_{REF}$ including the reference spur and the value of f_{REF}, respectively.

Acknowledgement:
This work was supported by Future Interconnect Technology Cluster Program of Samsung Electronics.

References:
[1] Y. Lee, et al., "A PVT-Robust –59-dBc Reference Spur and 450-fs$_{RMS}$ Jitter Injection-Locked Clock Multiplier Using a Voltage-Domain Period-Calibrating Loop," *IEEE Symp. VLSI Circuits*, pp. C236–C239, June 2016.
[2] S. Choi, et al., "A 185fs$_{rms}$-Integrated-Jitter and –245dB FOM PVT-Robust Ring-VCO-Based Injection-Locked Clock Multiplier with a Continuous Frequency Tracking Loop Using a Replica-Delay Cell and a Dual-Edge Phase Detector," *ISSCC*, pp. 194–195, Feb. 2016.
[3] S. Kim, et al., "A 2.5GHz Injection-Locked ADPLL with 197fs$_{rms}$ Integrated Jitter and -65dBc Reference Spur Using Time-Division Dual Calibration," *ISSCC*, pp. 494–495, Feb. 2017.
[4] S. Yang, et al., "A 0.0056mm^2 All-Digital MDLL Using Edge Re-Extraction, Dual-Ring VCOs and a 0.3mW Block-Sharing Frequency Tracking Loop Achieving 292fs$_{rms}$ Jitter and -249dB FOM," *ISSCC*, pp. 118–119, Feb. 2018.
[5] D. Auvergne, et al., "Signal Transition Time Effect on CMOS Delay Evaluation," *IEEE TCAS-I*, vol. 47, no. 9, pp. 1362-1369, Sept. 2000.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

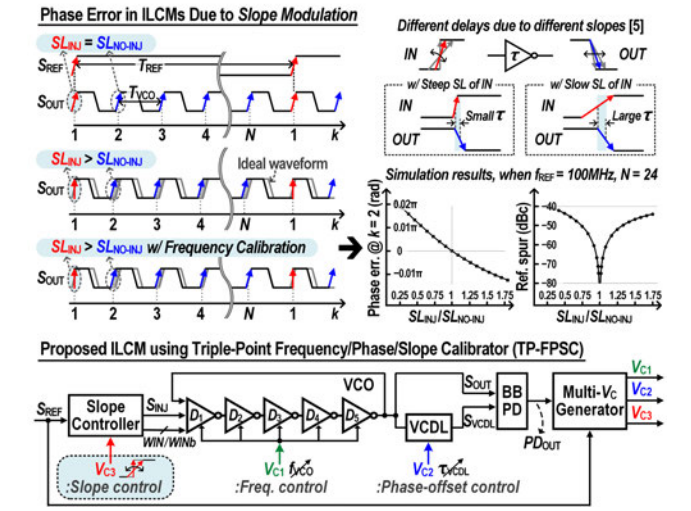


Figure 30.9.1: Phase errors due to the slope modulation in ILCMs, and simulation results of phase errors and reference spurs (top) and the architecture of the proposed ILCM using TP-FPSC (bottom).

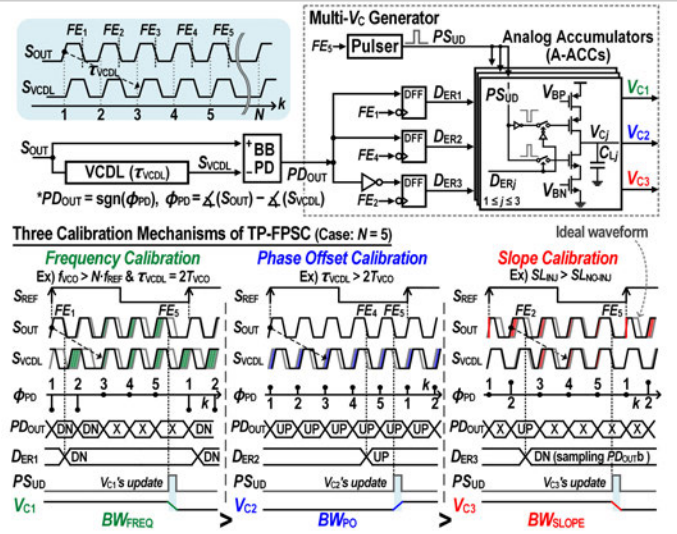


Figure 30.9.2: Block diagram of the multi-V_C generator (top) and three calibration mechanisms of the TP-FPSC (bottom).

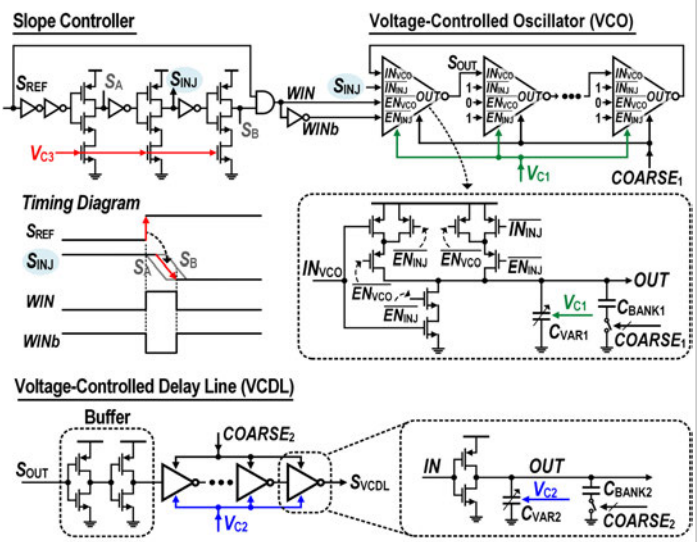


Figure 30.9.3: Schematics of the slope controller, the VCO, and the VCDL.

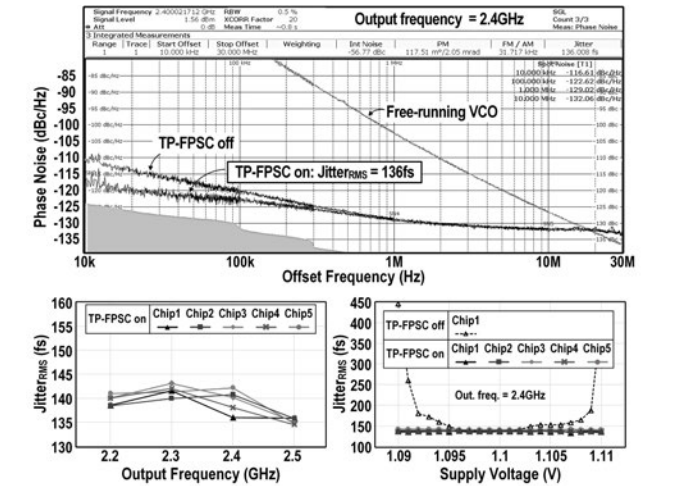


Figure 30.9.4: Measured phase noises of the 2.4GHz output signal, when the TP-FPSC was turned on and off (top), and variations of integrated rms jitter over output frequencies and supply voltages in five different sample chips (bottom).

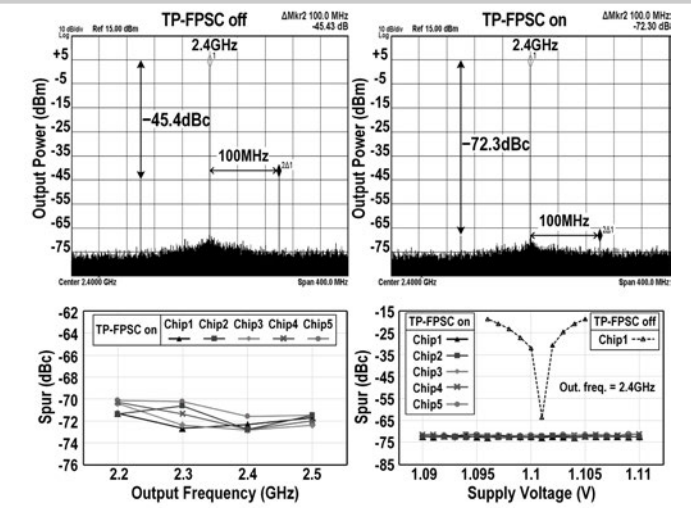


Figure 30.9.5: Measured spectrums of the 2.4GHz output signal, when the TP-FPSC was turned on and off (top), and variations of reference spurs over output frequencies and supply voltages in five different sample chips (bottom).

Comparison with state-of-the-art ring-VCO-based ILCMs

	This work	ISSCC'16 [2] S. Choi	ISSCC'17 H. C. Ngo	ISSCC'17 D. Coombs	ISSCC'17 [3] S. Kim	ISSCC'18 K.M. Megawer	ISSCC'18 [4] S. Yang
Process (CMOS)	65 nm	65 nm	65 nm	65 nm	65 nm	65 nm	28 nm
Calibrator	Frequency/ Offset/ Slope	Frequency	Frequency	Frequency/ Offset	Frequency/ Offset	Frequency/ Offset	Frequency/ Offset
Out. freq., f_0 (GHz)	2.4 (2.2 – 2.5)	1.2	0.9 (0.52 – 1.15)	5.0 (2.5 – 5.75)	2.5	4.752 (2.6 – 5.2)	3.0 (1.55 – 3.35)
Ref. freq., f_{REF} (MHz)	100	120	150	125	156.25	54	200
1MHz PN (dBc/Hz) @ f_0 (GHz)	–129.0 @ 2.4	–134.4 @ 1.2	–124.0 @ 0.9	–115.9 @ 5.0	-	–113.7 @ 4.752	–118.0 @ 3.0
Ref. spur, L_{spur} (dBc)	–72	–53	N/A	–45	–65	–53	–44
Out. jitter$_{RMS}$, σ_t (fs) (Integ. Range)	140 (10k – 30MHz)	185 (10k – 40MHz)	420 (10k – 10MHz)	340 (10k – 40MHz)	198 -	370 (10k – 30MHz)	292 (10k – 40MHz)
Total jitter$_{RMS}$, σ_{tot} (fs) (= $\sqrt{(\sigma_t^2 + {}^*\sigma_{spur}^2)}$)	142	459	N/A	424	204	385	556
Power, P_{DC} (mW)	11.0	9.5	3.8	5.3	13.5	6.5	1.45
Active area (mm²)	0.055	0.06	0.028	0.09	0.064	0.16	0.0056
**FoM_{JIT} (dB)	–246.7	–244.9	–241.7	–242.4	–242.8	–240.5	–249.1
***FoM_{JIT2} (dB)	–246.5	–237.0	N/A	–240.2	–242.5	–240.2	–243.5
****FoM_{REF} (dB)	–246.7	–244.1	–240.0	–241.2	–240.8	–243.2	–246.1

*$\sigma_{Lspur} = 10^{L_{spur}/20}/\sqrt{2}\pi f_0$ **FoM_{JIT}: $10\log_{10}(\sigma_t^2 \cdot P_{DC}(mW))$ ***FoM_{JIT2}: $10\log_{10}(\sigma_{tot}^2 \cdot P_{DC}(mW))$ ****FoM_{REF}: $10\log_{10}(\sigma_t^2 \cdot P_{DC}(mW) \cdot \frac{f_{REF}}{100MHz})$ [4]

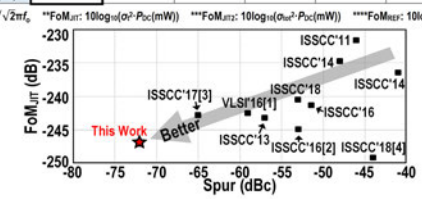
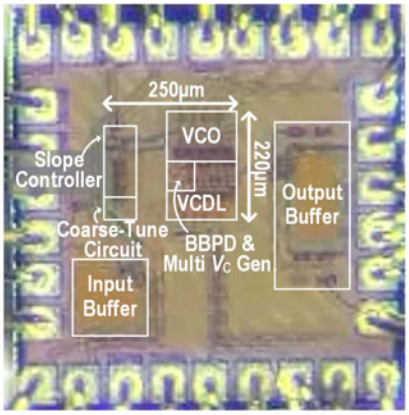


Figure 30.9.6: Performance comparisons with state-of-the-art ring-VCO-based ILCMs.



Power Consumption (mW)	
Ring VCO	7.5
VCDL	3.0
PG & BBPD & Multi-V_C Generator	0.5
Total	11.0

Figure 30.9.7: Die photograph and power breakdown.

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

1

1P1M	1-Polysilicon layer 1-Metal layer
1T1C	1-Transistor 1-Capacitor

3

3D	3-Dimensional
3G	Third-Generation (Wireless)
3T	3-Transistor

6

6T	6-Transistor

8

8T	8-Transistor

ΔΣ

ΔΣ	Delta-Sigma
ΔΣM	Delta-Sigma Modulator

ΣΔ

ΣΔ	ΔΣ is preferred.
ΣΔM	ΔΣM is preferred.

A

a-Si	Amorphous Silicon
AC	Alternating Current
A/D	Analog-to-Digital Converter
AAC	Advanced Audio Coding
ACI	Adjacent-Channel Interface
ACL	Access Control List
ACLR	Adjacent-Channel Leakage Power Ratio
ACPR	Adjacent-Channel Power Ratio
ADC	Analog-to-Digital Converter
ADDLL	All-Digital DLL
ADPLL	All-Digital PLL
ADSL	Asynchronous Digital Subscriber line
ADU	Analog-to-Digital Unit
AES	Advanced Encryption Standard
AFC	Automatic Frequency Control
AFE	Analog Front End
AFM	Adaptive Flash Management ™
AGC	Automatic Gain Control
AGU	Address-Generation Unit
AIP	Artificial-Intelligent Partner
ALU	Arithmetic Logic Unit
AM	Amplitude Modulation
AMI	Advanced Metering Infrastructures
AMLCD	Active-Matrix LCD
AMOLED	Active-Matrix OLED
AMP	Asymmetric Multi-Processing
AMPS	Advanced Mobile-Phone Service
AMS	Analog Mixed-Signal {System}
APD	Avalanche Photo-Diode
APG	Algorithmic Pattern Generator
API	Application-Programming Interface
APSK	Amplitude Phase-Shift Keying
ARM	Advanced RISC Machine
ASIC	Application-Specific Integrated Circuit
ASK	Amplitude Shift Keying
ASP	Average Selling Price
ASP	Advanced Simple Profile {MPEG-4 Video}
ATA	Advanced Technology Attachment
ATD	Address-Transition Detection
ATE	Automatic Test Equipment
ATM	Asynchronous Transfer Mode
ATSC	Advanced Television Systems Committee
AVC	Audio-Visual CODEC
AVC	Automatic Volume Control {use AGC}
AWG	Arrayed-Waveguide Grating

B

BAN	Business Area Network
BAW	Bulk Acoustic Wave
BB	Baseband
BBT	Band-to-Band Tunneling
BCD	Bipolar-CMOS-DMOS Process
BCD	Binary-Coded Decimal
BCH	Bose-Chaudhuri-Hocquenghem {a type of error-correcting code}
BD	Blu-ray disc
BER	Bit-Error Rate
BGA	Ball-Grid Array
BGR	Band-Gap Reference
BiCMOS	Bipolar Complementary-MOS
BIOS	Basic Input/Output System
BIST	Built-in Self-Test
BJT	Bipolar Junction Transistor
BL	Bitline
BLE	Bluetooth Low Energy
BOM	Bill of Materials
BPF	Bandpass Filter
BPSK	Binary Phase-Shift Keying
BSI	BackSide Illumination
B-VOP	Bidirectional-Video Object Planes
BW	Bandwidth

C

C4	Controlled-Collapse Chip Connection
CAD	Computer-Aided Design
CAM	Content-Addressable Memory
CAN	Controller Area Network
CAS	Column-Address Strobe
CCCS	Current-Controlled Current Source
CCD	Charge-coupled Device
CCK	Complementary Code Keying
CCO	Current-Controlled Oscillator
CCVS	Current-Controlled Voltage Source
CDAC	Capacitor DAC
CDMA	Code-Division Multiple Access
CDR	Clock and Data Recovery
CDS	Correlated Double Sampling
CF	Compact Flash
CFA	Color Filter Array
CFL	Compact Fluorescent Lamp
CHE	Channel Hot-Electron (Injection)
CIS	Complex-Instruction-Set Computer
CIS	CMOS Image Sensor
CLI	Command Line Interface
CML	Current-Mode Logic
CMOS	Complementary Metal-Oxide Semiconductor
CMRR	Common-Mode Rejection Ratio
CMU	Clock Multiplier Unit
CMUT:	Capacitive micromachined ultrasonic transducer
CNFET	Carbon Nanotube NFET
CO	Central-Office {hardware}
CODEC	Coder-Decoder
COFDM	Coded FDM
CoMP	Coordinated MultiPoint
CPE	Customer Premises Equipment
CPU	Central Processing Unit
CPW	CoPlanar Waveguide
CRC	Cyclic Redundancy Check
CSMA	Carrier-Sense Multiple Access
CT	Continuous Time {system}
CUI	Command User Interface
CVD	Chemical Vapor Deposition

D

D/A	Digital-to-Analog Converter
DAB	Digital-Audio Broadcasting
DAC	Digital-to-Analog Converter
dBFS	dB relative to Full Scale
DBS	Direct-Broadcast Satellite
DCC	Duty-Cycle Corrector
DCO	Digitally Controlled Oscillator
DCT	Discrete Cosine Transform
DCVS	Differential Cascode Voltage Switch
DCVS	Digitally-Controlled Voltage Source
DCXO	Digitally Controlled Crystal Oscillator
DDFS	Direct-Digital Frequency Synthesis (or synthesizer)
DDR	Dual Data Rate
DDS	Direct Digital Synthesis
DECT	Digitally Enhanced Cordless Communication
DEM	DEModulator
DEM	Dynamic Element Matching
DEMOS	Depletion MOS
DEMUX	Demultiplexer
DES	Data-Encryption Standard
DEVM	Differential Error Vector Magnitude
DFE	Decision-Feedback Equalizer
DFF	D-type Flip Flop
DFT	Design for Testability
DFT	Discrete Fourier Transform
DfY	Design-for-Yield
DIMM	Dual In-Line Memory Module
DIP	Dual In-line Package
DLL	Delay-Locked Loop
DMA	Direct Memory Access
DMB	Digital Multimedia Broadcasting
DMIPS	Dhrystone Million Instructions Per Second
DMOS	(Double-)Diffused MOS
DNA	Deoxyribonucleic Acid
DNL	Differential Non-Linearity
DNR	Dynamic Range (DR preferred)
DPT	Double Patterning Technology
DOM	Digital Optical Module
DR	Dynamic Range (see also DNR)
DRAM	Dynamic Random-Access Memory
DRC	Design-Rule Check
DSB	Double Side Band
DSC	Digital Still Camera
DSL	Digital Subscriber Line
DS-OFDM	Direct-Sequence Orthogonal Frequency Division Multiplexing
DSP	Digital Signal Processing
DSSS	Direct-Sequence Spread-Spectrum
DT	Discrete Time
DTL	Diode-Transistor Logic
DTV	Digital Television
DUT	Device Under Test
DVB	Digital-Video Broadcasting
DVB-C	Digital-Video Broadcasting - Cable
DVB-H	Digital-Video Broadcasting - Handhelds
DVB-S	Digital-Video Broadcasting - Satellite
DVB-T	Digital-Video Broadcasting - Terrestrial
DVD	Digital Video Disc
DVFS	Dynamic Voltage Frequency Scaling
DVS	Dynamic Voltage Scaling
DWA	Data Weighted Averaging
DWDM	Dense-Wavelength-Division Multiplexing
DWMT	Discrete Wavelet Multi-Tone

E

ECC	Error-Correcting Code
ECG	Electro-CardioGram
ECL	Emitter-Coupled Logic
ECP	Emitter-Coupled Pair
EDGE	Enhanced Data rates for Global Evolution
EDR	Enhanced Data Rate
EEG	Electro-EncephaloGram
EEPROM	Electrically Erasable Programmable Read-Only Memory
EFR	Enhanced Full-Rate {GSM}
EIRP	Effective Isotropic Radiated Power
EKG	Electro-CardioGram (see ECG)
EMG	ElectroMyoGram
EMI	Electro Magnetic Interference
eICIC	enhancement in Inter-Cell Interference Coordination
ENOB	Effective Number of Bits
EOT	Electrical Oxide Thickness
EPON	Ethernet-based Passive Optical Network
EPROM	Erasable Programmable Read-Only Memory
ERBW	Effective-Resolution Bandwidth
ESD	ElectroStatic Discharge
EUV	Extreme UltraViolet
EVDO	EVolution Data Optimized (in the context of CDMA)
EVM	Error-Vector Magnitude
EWC	Enhanced Wireless Consortium

F

f_{max}	Unity power gain frequency.
f_s	Sampling frequency
f_t	Transit frequency
FAMOS	Floating-gate Avalanche-injection MOS Transistor
FAN	Field Area Network
FBAR	Film Bulk Acoustic Resonator
FBDIMM	Fully Buffered DIMM
FCC	Federal Communications Commission (U.S.)
FDM	Frequency-Division Multiplexing
FDMA	Frequency-Division Multiple-Access
FDNR	Frequency-Dependent Negative Resistor
FDSOI	Fully Depleted Silicon-on-Insulator
FEC	Forward Error Checking
FEM	Front End Module
FeRAM	Ferro-electric Random Access Memory
FET	Field-Effect Transistor
FF	Flip-Flop
FFC	Flexible-Flat Cable
FFE	Feed-Forward Equalizer
FFT	Fast Fourier Transform
FIB	Focused Ion Beam
FIFO	First In-First Out
FinFET	A MOSFET with the gate on two sides {acronym describes the physical shape}
FIR	Finite Impulse Response {filter}
FLOPS	Floating-Point Operations Per Second
FLOTOX	Floating-gate Tunnel OXide
FM	Frequency Modulation
FMCW	Frequency Modulated Continuous Wave
FN	Fowler-Nordheim
FO4	Fan-Out of 4
FOM	Figure Of Merit
FPGA	Field-Programmable Gate Array
FPN	Fixed Pattern Noise
FPU	Floating Point Unit
FSG	FluoroSilicate glass {dielectric}
FSG	Fluorine-doped Silicate Glass
FSK	Frequency-Shift Keying
FSM	Finite-State Machine

G

GAA	Gate-All-Around
GaN	Gallium Nitride
GBW	Gain-BandWidth {product}
GCA	Gain-Controlled Amplifier
GDDR	Graphics Double-Data-Rate
GDfM	Generalized Design-for-Manufacturability
GE-PHY	Gigabit-Ethernet Physical
GFSK	Gaussian Frequency-Shift Keying
GFLOPS	Giga Floating-Point Operations Per Second
GIDL	Gate-Induced Drain Leakage
GMSK	Gaussian Minimum-Shift Keying
GOPS	Giga-Operations Per Second
GPRS	General Packet-Radio Service
GPS	Global Positioning System
GPU	Graphic Processing Unit
GSM	Global Standard for Mobile Communication
GUI	Graphical User Interface
GVCO	Gated VCO

H

HBT	Hetero-junction Bipolar Transistor
HBM	High-Bandwidth Memory
HBM	Human Body Model
HCI	Host-Controller Interface
HCI	Human-to-Computer Interface
HD	High-Density
HDD	Hard-Disk-Drive
HDL	Hardware-Description Language
HDTV	High-Definition TeleVision
HD2	2nd order Harmonic Distortion
HD3	3rd order Harmonic Distortion
HiFi	High Fidelity
Hk	High-k dielectric
HMD	Head Mounted Display
HPF	High-Pass Filter
HSPDA	High-Speed Downlink Packet Access
HT3	Hypertransport 3 (I/O standard)
HTM	Hierarchical Temporal Memory
HV	High Voltage
HVAC	Heating, Ventilation and Air Conditioning
HVCMOS	High-Voltage Complementary MOS
HVMOS	High-Voltage MOS

I

I/O	Input-Output
I/Q	In Phase and Quadrature
IAN	Industrial Area Network
IBOC	In-Band Out-of-Channel
IC	Integrated Circuit
ID	Identification
IF	Intermediate Frequency
IGBT	Insulated Gate Bipolar Transistor
IIP2	Input-referred Input 2nd-order Intercept Point
IIP3	Input-referred Input 3rd-order Intercept Point
IIR	Infinite Impulse Response Filter
IMD	Inter-Modulation Distortion
IM2	2nd-order InterModulation distortion
IM3	3rd-order InterModulation distortion
INL	Integral Non-Linearity
InP	Indium Phosphide
IoT	Internet of Things
IP	Intellectual Property
IPC	Inter-Process Communication
IPSEC	Internet (Network) Protocol for Security
IR	Image Rejection
IR	InfraRed
ISDB	Integrated Services Digital Broadcasting
ISDB-C	Integrated Services Digital Broadcasting - Cable
ISDB-S	Integrated Services Digital Broadcasting - Satellite
ISDB-T	Integrated Services Digital Broadcasting - Terrestrial
ISFET	Ion-Sensitive Field-Effect Transistor
ISI	Inter-Symbol Interference
ISM	Industrial, Scientific and Medicine Band
ITU-T	International Telecommunications Union

J

JPEG	Joint Photographic Expert Group
JTAG	Joint Test-Automation Group

K

KGD	Known Good Die

L

LAGS	Locally-Asynchronous Globally-Synchronous
LAN	Local-Area Network
LBS	Location-Based Services
LCD	Liquid-Crystal Display
LCOS	Liquid-Crystal On Silicon
LDI	LVDS Display Interface
LDCMOS	Laterally Diffused Complementary Metal-Oxide Silicon
LDMOS	Laterally Diffused Metal-Oxide Silicon
LDO	Low Drop-Out {regulator}
LDPC	Low-Density Parity Check
LED	Light-Emitting Diode
LEP	Light-Emitting Polymer
LFCSP	LeadFrame Chip-Scale Package
LFSR	Linear-Feedback Shift Register
LHC	Large Hadron Collider
LIN	Local Interconnect Network
LMS	Least Mean Square
LNA	Low-Noise Amplifier
LNB	Low-Noise Blockers
LO	Local Oscillator
LPCVD	Low-Pressure Chemical-Vapor Deposition
LPF	Low-Pass Filter
LRU	Least-Recently Used
LSB	Least-Significant Bit
LSI	Large-Scale Integration
LTE	Long Term Evolution
LTPS	Low-Temperature-Poly Silicon
LV	Low Voltage
LVDS	Low-Voltage Differential Signaling
LVS	Layout Verification to Schematic

M

MAC	Media-Access Controller
MAC	Mutiply-Accumulate
MASH	Multi-stAge noise SHaping
MBOA	Multi-Band OFDM Alliance
MB-OFDM	Multi-Band OFDM
MCM	Multi-Chip Module
MCP	Multi-Chip Package
MCU	MicroController Unit
MCU	Multipoint Conferencing Unit
MDAC	Multiplying DAC
MEMS	Micro-Electro-Mechanical System
MER	Modulation Error Ratio
MfD	Manufacturing-for-Design
MPU	Microprocessor Units
microSD	micro Secure Digital
MICS	Medical Implant Communication Band
MIM	Metal-Insulator-Metal
MIMO	Multiple- Input, Multiple-Output
MIPI	Mobile Industry Processor Interface
MIPS	Million Instructions Per Second
MISR	Multiple-Input Signature Register
ML	Matchline

MLC	Multi-Level Cell
MLSD	Maximum-Likelihood-Sequence Detection
MLSE	Maximum-Likelihood-Sequence Estimation
MMAC	Mega Mutliply-Accumulate
MMIC	Monolithic Microwave Integrated Circuit
MMIO	Memory-Mapped I/O
mmW	millimeter-Wave
MODEM	Modulator-Demodulator
MOS	Metal-Oxide-Semiconductor (Silicon)
MOSFET	Metal Oxide Semiconductor Field-Effect Transistor
MOST	MOS Transistor
MOST	Media-Oriented Systems Transport
MP	Multi-Processor
MP3	MPeg-1 audio layer 3 (lossy compression algorithm)
MPEG	Motion-Picture Expert Group
Mpps	Million packets per second {VoIP}
MPPT	Maximum-Power-Point Tracker
MPU	Microprocessors
MSB	Most Significant Bit
MRAM	Magnetic Random-Access Memory
MRAM	Magnetoresistive Random-Access Memory
MRC	Maximum-Ratio Combining
MSB	Most-Significant Bit
MTJ	Magnetic Tunnel Junction
MTPR	Multi-Tone Power Ratio
MUX	Multiplexer
MWPC	Multi-Wire Proportional Chamber

N

NAICS	Network Assisted Interference Cancellation and Suppression
NAN	Neighbourhood Area Network
NBTI	Negative-Bias Temperature Instability
NEF	Noise Efficiency Factor
NEM	Nano-Electro-Mechanical system
NF	Noise Figure
NFV	Network Function Virtualization
NMOS	N-channel MOS
NMOST	NMOS Transistor
NoC	Network on (a) Chip
NPN	N-type-P-type-N-type bipolar (transistor)
NRTZ	Non Return-To-Zero (see also NRZ)
NRZ	Non-Retun- to-Zero (see also NRTZ)
NTF	Noise Transfer Function
NVM	Non-Volatile Memory
NVRAM	Non-Volatile Random-Access Memory

O

ODT	On-die Termination
OEM	Original-Equipment Manufacturer
OFDM	Orthogonal Frequency-Division Multiplexing
OIF	Optical-Internetworking Forum
OIP2	Output-referred intercept point for 2nd-order distortion
OIP3	Output-referred intercept point for 3rd-order distortion
OLED	Organic LED
ONO	Oxide-Nitride-Oxide
OOK	On-Off Keying
OSR	Over-Sampling Ratio
OTA	Operational Transconductance Amplifier
OTP	One Time Programmable

P

P_{1dB}	1dB gain-compression Point
PA	Power Amplifier
PA	Public-Address {system}
PAE	Power-Added Efficiency
PAM	Pulse-Amplitude Modulation
PAN	Personal-Area Network
PCB	Printed-Circuit Board
PCH	Platform Controller Hub
PCI	Peripheral-Component Interconnect
PCI-X	PCI Express
PCM	Pulse-Code Modulation
PCS	Personal Communication Services
PCU	Power-management Control Unit
PD	Phase Detector
PD-SOI	Partially-Depleted SOI
PDA	Personal Data Assistant
PDN	Power-Delivery Network
PDP	Power-Delay Product
PFD	Phase and Frequency Detector
PGA	Programmable-Gain Amplifier
PGA	Programmable Gate Array
PHEMT	Pseudomorphic High-Electron-Mobility Transistor
PHEVs	Plug-in Hybrid Electric Vehicles
PHY	PHYsical layer {of a communications protocol}
PID	Proportional, Integral, Derivative (a type of control loop)
PLA	Programmable Logic Array
PLC	Power-Line Communication
PLD	Programmable Logic Device
PLL	Phase-Locked Loop
PMOS	P-channel MOS
PMOST	PMOS transistor
PMU	Power Management Unit
PNP	P-type-N-type-P-type bipolar (transistor)
PON	Passive Optical Network
PoP	Package-on-Package
POTS	Plain-Old Telephone Service
ppm	parts per million
PPM	Pulse-Position Modulation
PR	Pseudo Random
PR	Partial Response
PRAM	Phase-Change RAM
PRBS	Psuedo-Random Binary Sequence
PRML	Partial-Response, Maximum-Likelihood
PROM	Programmable Read-Only Memory
PSD	Power Spectral Density
PSK	Phase-Shift Keying
PSNR	Peak SNR
PSRR	Power-Supply Rejection Ratio
PTAT	Proportional To Absolute Temperature
PV	PhotoVoltaics
PVD	Physical Vapor Deposition
PVT	Process, Voltage, Temperature
PWM	Pulse-Width Modulation

Q

QPT	Quadruple Patterning Technology
QAM	Quadrature Amplitude Modulation
QDR	Quad Data Rate
QoS	Quality of Service
QPSK	Quadrature Phase-Shift Keying
QVCO	Quadrature Voltage-Controlled Oscillator
QVGA	Quarter Video Graphics Array

R

R/W	ReadWrite
RAM	Random-Access Memory
RAT	Radio Access Technologies
RBW	Resolution BandWidth
RDAC	Resistor DAC
RDF	Random Dopant Fluctuation
RF	Radio Frequency
RFID	RF ID (tag)
RISC	Reduced-Instruction-Set Computer
ROM	Read-Only Memory
rms	root mean square

ROM	Read-Only Memory
RSA	A public-key cryptographic system, named after: Ron Rivest, Adi Shamir, and Leonard Adleman
RSSI	Received-Signal-Strength Indicator
RTL	Resistor-Transistor Logic
RTS	Random Telegraph Signal
RTZ	Return-To-Zero
RX	Receiver
RZ	Return-to-Zero (see also RTZ)

S

SAL	Service Abstraction Layer
SAR	Successive-Approximation-Register
SAW	Surface Acoustic Wave
SATA	Serial Advanced-Technology Attachment
SC	Switched-Capacitor
SCL	Source-Coupled Logic
SCP	Source-Coupled Pair
SCR	Silicon Controlled Rectifier
S-DMB	Satellite Digital-Multimedia Broadcasting
SD	Secure Digital (package type)
SDI	Software-Defined Infrastructure
SDN	Software-Defined Networking
SDR	Software Defined Radio
SDRAM	Synchronous Dynamic Random-Access Memory
SEM	Scanning Electron Microscope
SER	Soft-Error Rate
SER	Symbol-Error Rate
SerDes	Serializer/Deserializer
SFDR	Spurious-Free Dynamic Range
SFI	Serdes Framer Interface
SFP	Small Form-factor Pluggable
S/H	Sample-and-Hold
SHA	Sample-and-Hold Amplifier
SiC	Silicon Carbide
SiGe	Silicon Germanium
SiGe:C	Silicon, Germanium, Carbon
SIL	Safety Integrity Level
SILC	Stress-Induced Leakage Current
SIMD	Single-Instruction Multiple-Data
SINAD	SIgnal-to-Noise And Distortion (ratio)
SIO	Synchronous I/O
SIP	Single-Inline Package
SiP	System in (a) Package
SL	Searchline
SMP	Symmetric Multi-Processing
SMS	Short-Messaging Service
SNDR	Signal-to-Noise and Distortion Ratio
SNM	Static Noise Margin
SNR	Signal-to-Noise Ratio
SNS	Social-Network Service
SoC	System on (a) Chip
SOI	Semiconductor on Insulator
SONET	Synchronous Optical NETwork
SONOS	Silicon-Oxide-Nitride-Oxide-Silicon
SOS	Silicon On Saphire
SP	Simple Profile
SPAD	Single-Photon Avalanche Diode
SPDT	Single Pole Double Throw
SPI	System Packet Interface
SRAM	Static Random-Access Memory
SSB	Single Side-Band
SSC	Spread Spectrum Clocking
SSCG	Spread Spectrum Clock Generator
SSD	Sold State Disk
SSN	Simultaneous Switching Noise
SSO	Simultaneous Switching Output
SSTL	Stub Series Terminated Logic
SXGA	Super Extended Graphics Array

T

TC	Temperature Coefficient
TCAM	Ternary Content-Addressable Memory
TCON	Timing Controller
TCP	Transmission Control Protocol
TDC	Time-to-Digital Converter
TDDB	Time-Dependent Dielectric Breakdown
TDM	Time-Division Multiplexing
TDMA	Time-Division Multiple-Access
TD-SCDMA	Time-Division Synchronous Code-Division Multiple Access
TEM	Tunneling-Electron Microscope
TFLOPS	Tera Floating-Point Operations Per Second
TFT	Thin-Film Transistor
TIA	TransImpedance Amplifier
T/H	Track and Hold
THA	Track-and-Hold Amplifier
THD	Total Harmonic Distortion
THD+N	THD plus Noise
TLB	Translation Lookaside Buffer
ToF	Time of flight
TOPS	Tera-Operations Per Second
TSV	Through-Silicon Via
TTL	Transistor-Transistor Logic
TV	Television
TWh	Tera-Watt hour
TX	Transmitter

U

UD	Ultra-high Definition
UDTV	Ultra-high-Definition TeleVision
UGC	User-Generated Contents
UHDV	Ultra-High-Definition Video
UHF	Ultra-High Frequency
UI	Unit Interval
UI	User Interface
UIPP	UI_{pp} (peak-to-peak)
U-NII	Unlicensed National Information Infrastructure
UMPC	Ultra-Mobile-PC
UMTS	Universal Mobile-Telecommunication System
UPROM	Unerasable Programmable Read-Only Memory
USB	Universal Serial Bus
UTP	Unshielded Twisted Pair
UWB	Ultra-WideBand
UXGA	Ultra-eXtended Graphics Array
UPF	Unified Power Format

V

V_t	MOS transistor threshold voltage
V_T	thermal voltage
VCCS	Voltage-Controlled Current Source
VCDL	Voltage-Controlled Delay Line
VCO	Voltage-Controlled Oscillator
VCVS	Voltage-Controlled Voltage-Source
VCXO	Voltage-Controlled Crystal Oscillator
VCSEL	Vertical-Cavity Surface-Emitting Laser
VDMOS	Vertically Diffused MOS
VDSL	Very high bit-rate Digital Subscriber Line
VGA	Variable-Gain Amplifier
VGA	Video Graphics Array
VLIW	Very Long Instruction Word
VLF	Very Low Frequency
VLSI	Very Large-Scale Integration
VoD	Vision on Demand
VoIP	Voice over IP
VR	Voltage Regulator
VRM	Voltage Regulator Module
VRT	Variable-Retention Time
VSB	Vestigial Side Band

VSoC	Virtual System-on-Chip
VSWR	Voltage Standing-Wave Ratio
VTG	Vertical-Transfer Gate

W

WAN	Wide-Area Network
WCDMA	Wideband Code-Division Multiple- Access
WDM	Wavelength-Division Multiplexing
WebRTC	Web Real-Time Communication
WEP	Wired-Equivalent Privacy
WiFi	Wireless Fidelity; {an interoperability certification for IEEE 802.11 WLAN products}
WiMax	Worldwide Interoperability for Microwave Access (IEEE802.16)
WL	Wordline
WLAN	Wireless Local-Area Network
WLCG	Worldwide LHC Computing Grid
WSN	Wireless Sensor Node

X

XAUI	(10 Gigabit) eXtended Attachment Unit Interface
XDR	Extreme Data Rate
XENPAK	10 Gb Ethernet-compatible fiber-optic standard
XFMR	TransForMeR
XFP	Small form-factor pluggable
XGA	Extended Graphics Array
XO	Crystal Oscillator

Z

ZB	Zettabyte

There are a total of 10 tutorials this year on 10 different topics. Each tutorial, selected through a competitive process within each subcommittee of the ISSCC, presents the basic concepts and working principles of a single topic. These tutorials are intended for non-experts, graduate students and practicing engineers who wish to explore and understand a new topic.

The speakers for these tutorials have spent significant amount of time and effort in preparing these tutorials. In addition, the tutorials have been vetted carefully by their respective coordinators (one coordinator for each tutorial). In some cases, the tutorials have gone through several revisions to bring them to the high quality you will experience at ISSCC. I would like to thank all the speakers and the coordinators for their great effort and dedication, and I would like to invite you to take pleasure in attending one or more of these tutorials.

The coordinators for 2019 ISSCC Tutorials are (in the same order as the tutorials): Andrea Bevilacqua, Jason Stauth, Sriram Vangal, Mahesh Mehendale, Bob Verbruggen, Sam Palermo, Keith Bawman, Byungsub Kim, Renaldi Winoto, Bruce Rae.

Ali Sheikholeslami
ISSCC Education Chair

8:30 AM

T1: Fundamentals of Integrated Radars

Brian Ginsburg, *Texas Instruments, Dallas, TX*

Radar is essential to modern aviation, meteorology, and automobiles, among others. This tutorial will cover the basic radar parameters including resolution, accuracy, and maximum range, and show how those are linked to fundamental system specifications. Modern highly integrated radars primarily use continuous-wave modulations, including FMCW and PMCW. High-level architectures and key circuits for a mm-wave implementation will be presented, along with a look at detection signal processing.

Brian Ginsburg received his S.B., M.Eng., and Ph.D. degrees from the Massachusetts Institute of Technology. He joined Texas Instruments, Dallas, Texas in 2007 working in its wireless terminals business unit and Kilby research labs. He is the systems manager of TI's Radar and Analytics Processors organization, developing mm-wave sensors for automotive and industrial applications. He serves on the ISSCC and VLSI Circuits Symposium Technical Program Committees.

8:30 AM

T2: Fundamentals of Power-Conversion Topologies

Robert Pilawa, *UC Berkeley, Berkeley, CA*

This tutorial will provide an overview of various power-converter topologies, with a focus on techniques to achieve high-efficiency and high-power density. A brief summary of common power converters, such as buck, boost, single-input multiple-output (SIMO) and multi-phase converters will be presented followed by a detailed analysis and comparison to more recent topologies, such as switched-capacitor (SC) power converters, hybrid and resonant SC converters, and multi-level converters.

Robert Pilawa-Podgurski is currently an Associate Professor in the Electrical Engineering and Computer Sciences Department at the University of California, Berkeley. He received his B.S., M.Eng., and Ph.D. degrees from MIT. He performs research in the area of power electronics. Dr. Pilawa-Podgurski received the Google Faculty Research Award in 2013, and the 2014 Richard M. Bass Outstanding Young Power Electronics Engineer Award of the IEEE Power Electronics Society, given annually to one individual for outstanding contributions to the field of power electronics before the age of 35. He is co-author of nine IEEE prize papers.

8:30 AM

T3: Advances and Prospects for In-Memory Computing

Naveen Verma, *Princeton University, Princeton, NJ*

As applications become more data centric, the cost of accessing data from memory dominates energy and throughput. An emerging paradigm to address this is in-memory computing. Early approaches show the potential to reduce energy-delay product (EDP) by 2-to-3 orders of magnitude. However, architectures for enabling computation within the constrained structure of memory raise new challenges and opportunities. This tutorial will map out advances in these areas and examine recent demonstrations.

Naveen Verma received the B.A.Sc. degree in Electrical and Computer Engineering from the UBC, Vancouver, Canada, in 2003, and the M.S. and Ph.D. degrees in Electrical Engineering from MIT in 2005 and 2009, respectively. Since July 2009, he has been with the Department of Electrical Engineering at Princeton, where he is currently an Associate Professor. His research focuses on advanced sensing systems and on exploring how systems for learning, inference, and action planning can be enhanced by algorithms that exploit new sensing and computing technologies. This includes research on large-area, flexible sensors, energy-efficient statistical-computing architectures and circuits, and machine-learning and statistical-signal-processing algorithms.

10:30 AM

T4: Fundamentals of Efficient IoT Microcontrollers

James Myers, *ARM, Cambridge, United Kingdom*

IOT microcontrollers are complex constrained SOCs with analog peripherals, DCDC converters, CPUs, accelerators, volatile/non-volatile memories, and radios – all well represented at ISSCC.

However state-of-the-art circuits do not automatically translate to a state-of-the-art system. This tutorial presents an overview of system components and their system efficiency considerations, using published systems as examples. Future trends, such as battery-free energy harvesting, will also be covered.

James Myers leads the Devices, Circuits and Systems team within Arm Research, which seeks to push compute to the furthest extremes of cost, efficiency, performance, and integration. He joined Arm in 2007 initially responsible for reference implementation flows for the whole processor family. Joining R&D full time from 2009, he has since focused on deployable techniques for reduction of CPU and SoC power, embodied across more than a dozen tapeouts. James holds 20 granted and 16 pending US patents, co-authored one book chapter, and has presented at ISSCC and VLSI Circuits Symposium. His current interests include energy harvesting, printed electronics, sub-threshold circuits, and better-than-worst-case design.

10:30 AM

T5: Noise-Shaping in Data Converters

Venkatesh Srinivasan, *Texas Instruments, Dallas, TX*

Noise shaping is a powerful technique that helps improve the overall accuracy of a data converter. While this technique is widely used in the context of $\Delta\Sigma$ modulators, noise shaping is also gaining popularity in other kinds of converters such as SARs. This tutorial will present the fundamentals of noise shaping along with the architectural view of their application in different data converters such as $\Delta\Sigma$ modulators and SARs. Circuit implementation of key building blocks along with challenges in realization and their mitigation strategies will be discussed as well.

Venkatesh Srinivasan received his B.E (Hons) in Electrical and Electronic Engineering from Birla Institute of Technology and Science, Pilani, India in 1999 and his M.S and Ph.D. degrees in Electrical and Computer Engineering from the University of Tennessee, Knoxville and Georgia Institute of Technology, Atlanta in 2002 and 2006 respectively. Since 2006, he has been with Texas Instruments leading the architecture, design and development of multiple RF, Analog and high-speed $\Delta\Sigma$ modulators for wireless transceivers. He is the author/co-author of 27 publications in IEEE Journals and Conferences and has been issued 14 US patents. He is the recipient of the Best Student Paper Award from IEEE CICC in 2005. He currently serves on the Technical Program Committee of ISSCC.

10:30 AM

T6: Basics of Clock-and-Data-Recovery Circuits

Amir Amirkhany, *Samsung Electronics, San Jose, CA*

The choice of the clock-and-data-recovery (CDR) architecture in serial links dictates many of the block-level circuit specs. In the first half of this tutorial we will discuss the basics of CDR operation, CDR main performance metrics, and the relationship between circuit-level parameters and system-level performance metrics. In the second half, we will review various CDR architectures and their associated design trade-offs. We will conclude with the review of a few practical CDR design considerations.

Amir Amirkhany is a Director of Engineering at Samsung Display America Lab, San Jose, CA. He received his B.S. from Sharif University of Technology, M.S. from University of California, Los Angeles, and Ph.D. from Stanford University, all in Electrical Engineering. At Samsung Electronics, Dr. Amirkhany is in charge of the development of future generations of high-speed interfaces for Samsung displays. Prior to Samsung, he was an engineering manager at Rambus Inc., where he led the development of DDR and proprietary high-speed memory interfaces. Dr. Amirkhany has authored or co-authored over 25 IEEE conference and journal papers, including a best paper award, and has over 50 issued US patents.

1:30 PM

T7: Hardware Security – from Basics to ASICs

Massimo Alioto, *National University of Singapore, Singapore*

Many applications require security to be rooted in hardware, due to constrained resources or demanding security levels (e.g., IoT, automotive, e-wallets). This tutorial introduces state-of-the-art techniques to create security primitives (e.g., PUFs, TRNGs, ciphers), and counteract hardware attacks.

At the end of the tutorial, attendees will understand basic principles of HW security, its state of the art and the challenges ahead, with emphasis on low-cost techniques for pervasive adoption.

Massimo Alioto directs the Integrated Circuits and Embedded Systems area, and leads the Green IC group at the National University of Singapore. He held visiting positions at the University of California Berkeley (BWRC), University of Michigan Ann Arbor, Intel Labs, EPFL, and positions at the University of Siena.

He is author of 250+ publications and three books, focusing on ultra-low power integrated system design, widely energy-scalable circuits, and hardware security, among others.

He is Editor in Chief of IEEE Transactions on VLSI Systems, and was Deputy Editor in Chief of IEEE JETCAS, Guest Editor of various IEEE journal special issues, Program Chair of several IEEE conferences, and an IEEE Distinguished Lecturer. He is an IEEE Fellow.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

1:30 PM

T8: Current-Sensing Techniques

Mahdi Kashmiri, *Robert Bosch, Sunnyvale, CA*

Current-sensing is an essential part of a wide range of applications from hand-held products to the electro-mobility drivetrains. The state of charge in a battery and the control of an electromotor or a power converter have specific demands for dynamic-range, precision, speed, isolation, size, and cost. This tutorial provides an overview of the principles of operation and readout techniques for the two current-sensing approaches: Contacted (shunt-based) and contactless (Hall, fluxgate, transformer, AMR/GMR based).

Mahdi Kashmiri received his B.Sc. from Tehran University and his M.Sc. and Ph.D. in Microelectronics from Delft University of Technology, Delft, The Netherlands. From 2010 to 2016 he was with Texas Instruments Inc. in Delft and Santa Clara locations, where he worked on precision products including AFEs for low-power ECG, shunt-based current and voltage monitor, contactless current sensor based on integrated Fluxgate and automotive and industrial ultrasound. Since 2016 he has been with the Mixed-signal ASICs group of Robert Bosch Research and Technology Center in Sunnyvale, CA, where he works on ASICs for automotive Lidar. He received the ESSCIRC 2009 Young Scientist Award and is an ISSCC member of the Analog technical program committee.

3:30 PM

T9: Calibration Techniques for Wireless Transceivers

David McLaurin, *Analog Devices, Raleigh, NC*

Modern wireless transceivers are expected to meet aggressive performance, cost, and power consumption targets with adequate yield in short design cycles. To address these challenges, RFICs are relying extensively on calibration techniques. This tutorial presents the impact of analog/RF non-idealities on key radio metrics and will cover techniques to address topics such as IQ imbalance, LO feed-through, IIP2 nonlinearities, analog filter tuning, full duplex radio and phased array calibrations.

David McLaurin joined Analog Devices in 2000, designing high speed pipeline ADCs. In 2002 he joined an RF IC team in Raleigh, NC, where he architects and designs wideband reconfigurable wireless transceivers. David graduated from the Georgia Institute of Technology (MSEE 2000), and North Carolina State University (BSEE 1998). He has been serving as a member of the technical program committee of ISSCC since 2016.

3:30 PM

T10: Low-Noise Sensor Interfaces

Pedram Lajevardi, *Robert Bosch, Sunnyvale, CA*

Applications, such as autonomous driving, rely on the perception of the environment around us based on extensive sensing. While most signals in the nature are not changing quickly and can be captured with relatively low-bandwidth sensors, most require low-noise sensors. This tutorial goes over the main sources of noise with a focus on capacitive and image sensors. Advanced noise-reduction circuit techniques, such as chopping, correlated-double sampling, and Boxcar sampling, are explained.

Pedram Lajevardi received B.A.Sc. degree in Electrical and Computer Engineering from the University of British Columbia and MS and Ph.D. in Electrical Engineering from Stanford University. He joined Bosch Research and Technology Center, Sunnyvale, CA in 2012 working on CMOS image sensors. Since 2015 he has been serving as a member of the technical program committee of ISSCC. His research interests are low-power and high-precision sensor interfaces.

F1: Sub-6GHz 5G Radio Circuits and Systems: From Concepts to Silicon

Organizer: **Theodoros Georgantas,** *Broadcom, Athens, Greece*

Committee: **Chan-Hong Chern,** *TSMC, Hsinchu, Taiwan*
Xin He, *NXP, Eindhoven, The Netherlands*
Yen-Hsun Hsu, *MediaTek, Hsinchu, Taiwan*
Harish Krishnaswamy, *Columbia University, New York, NY*
Hua Wang, *Georgia Institute of Technology, Atlanta, GA*

The next-generation cellular (5G) will offer Gb/s data rates with high spectral efficiency and minimal latency. To meet these performance targets and provide an optimal balance between coverage and implementation cost, 5G deployments in sub-6GHz bands will coexist with LTE and employ broadband operation along with massive MIMO and beamforming. This forum will describe how 5G requirements affect radio designs below 6GHz, current state-of-the-art of 5G systems, and what the future directions of key circuits and system architectures are. In addition, power efficiency along with various challenges and solutions on silicon integration of 5G radios will be also addressed.

Overview of Radio RF Systems Requirements and Challenges for sub-6GHz 5G NR Commercialization

Walid Ali-Ahmad, *Facebook, Menlo Park, CA*

The current need for eMBB, URLLC, and mMTC led to the introduction of 5G NR in the sub-6GHz range, along with existing LTE-A radio. This presentation focuses on overviewing key RF systems requirements and issues for sub-6GHz 5G NR UE design with respect to the current 3GPP standard; it also discusses challenges with the increased complexity in the radio RF section, due to the use of CA, MIMO, wider channels, and coexistence scenarios with other radio access technologies in the sub-6GHz range.

Walid Ali-Ahmad is currently RF/Analog Systems lead with Facebook in Menlo Park, CA, with focus on 60GHz and 5G NR radio systems design for connectivity and VR applications. From July 2014 until December 2017, he was a VP of Technology at Qualcomm and was involved in the architecture and RF systems design of advanced RF front-ends and transceiver systems for 4G and sub-6GHz 5G UE. Before July 2014, he was a senior director of Technology at Mediatek, and led the architecture and RF systems design of low-cost 3G/4G transceivers and SoCs for China market feature phones and smartphones. He is a Senior member of IEEE and an IEEE Distinguished Microwave Lecturer for 2017–2019. He was the general chair of the IEEE RFIC conference held in Philadelphia in 2018.

Multiband Receivers for LTE-A/5G Systems

Thomas Byunghak Cho, *Samsung Electronics, Hwaseong, Korea*

Designing the next generation LTE-A/5G(Sub-6G) cellular receiver faces several new stringent requirements. To increase the throughput over many fragmented frequency bands, a sophisticated carrier aggregation operation is a must. Supporting legacy 2/3/4G operation is also critical for global roaming. In addition, the RF front-end design must be approached differently, depending on the target market segment. In this talk, we will discuss these requirements and examine recently commercialized CMOS RFICs.

Thomas Byunghak Cho received his B.S. degree from UCLA, and his M.S. and Ph.D. degrees from UC, Berkeley, all in electrical engineering. For the past 20+ years, he has been working on many CMOS RFIC products for various cellular and connectivity applications. Since 2012, he has been with Samsung Electronics in Korea, and is currently the Senior VP of the Infra & Design Technology Team in charge of cellular/connectivity RFIC and mixed-signal IP development, digital physical implementation, design verification, and other topics.

Digital Transmitters for Sub-6GHz Wireless Applications

Leo de Vreede, *Delft University of Technology, Delft, The Netherlands*

Digital-intensive transmitter (TX) concepts are rapidly gaining interest, since they provide conversion of the digital baseband data directly into a RF signal, while facilitating compact and cost-effective integration. Their digital (switching) nature opens new possibilities in achieving high efficiency, spectral purity and frequency-agile operation. This forum contribution describes the strengths, challenges and promises of these digital-inspired TX solutions for sub-6GHz wireless applications.

Leo de Vreede received his Ph.D. degree from the Delft University of Technology in 1996, where he is currently heading the Electronic Research Laboratory (ERL/ELCA) as full professor. He worked on solutions for improved linearity and RF performance at the device, circuit and system level. He is co-founder of Anteverta-mw, a company specialized in RF device characterization, and was recipient of the IEEE Microwave prize in 2008. He has guided several students that won paper awards at the BCTM, PRORISC, GAAS, ESSDERC, IMS, RFIT and RFIC conferences. He has authored more than 120 IEEE refereed conference and journal papers and holds several patents. His current interests include RF technology optimization and (digital-intensive) concepts for wireless applications.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

High-Power, High-Bandwidth, and High-Efficiency Supply Modulator for 5G Systems

Paul Liang, *MediaTek, Hsinchu, Taiwan*

To achieve multi-Gb/s data-rate for enhanced mobile broadband, the next generation 5G cellular system adopts wider channel BW up to 100MHz. The power transient time mask is greatly reduced due to shorter symbol time. Furthermore, to accommodate UE power Class 2 (26dBm) and complex RF front-end, PA output power is expected to increase to an unprecedented level. As a result, the entire TX system designs are being challenged. In this talk, the challenge of envelope tracking in 5G will be presented and discussed.

Paul Liang received the Ph.D. degree in EE from the University of Michigan in 2001, with a focus on communication system design. From 2000 to 2006, Dr. Liang joined a start-up, Tropian, in California, working on high-speed digital system design. In 2006, he joined Panasonic Wireless Research Lab through acquisition. In 2010, Dr. Liang joined MediaTek where he holds the Director position, and supervised the Advanced System and Technology team on mixed-signal architectural design for 2/3/4/5G MODEMs and RF, envelope tracking system design, 5G prototypes, etc. Dr. Liang also served on the ISSCC technical program committee until 2017. He holds 40+ US patents and has authored/co-authored tens of technical articles in these areas.

Massive MIMO for sub-6GHz Wireless Communications

Ove Edfors, *Lund University, Lund, Sweden*

Starting from fundamental Massive-MIMO theory and associated performance predictions, we compare these theoretical results with those obtained from Massive-MIMO propagation channel measurements and real-time communication tests using a testbed with an LTE-like physical layer, in sub-6GHz frequency bands. Massive-MIMO properties, derived from theory and observed in experiments, are discussed and their influence on Massive-MIMO implementation is elaborated.

Ove Edfors received the M.Sc. degree in Computer Science and Electrical Engineering in 1990 and the Ph.D. degree in Signal Processing in 1996, both from Luleå University of Technology, Sweden. In 1997 he joined the staff at the Department of Electrical and Information Technology, Lund University, Sweden, where he is professor of Radio Systems since 2002 and director of the Excellence Center for System Design on Silicon since 2017. His research interests include statistical signal processing and low-complexity algorithms with applications in wireless communications. He has a long experience with multicarrier technology and was an early contributor to Massive-MIMO technology.

RF Transceivers for 5G Basestations

Tony Montalvo, *Analog Devices, Raleigh, NC*

The promise of 5G comes with a long list of challenges that impact the radio. These include wider bandwidth, more bands and higher frequency bands. Perhaps most important, 5G relies on Massive MIMO to improve spectral efficiency. Massive MIMO arrays – which have as many as 64 or even 128 antennas – multiply the above mentioned challenges. Reusing existing macro base-station architectures is clearly not practical. The state of the art in RF transceivers that address these challenges is described in this talk.

Tony Montalvo joined Analog Devices in 2000, has been a Fellow since 2012 and has been the Director of Technology of the Communications Business Group since 2017. Before joining Analog Devices, Tony developed RF ICs for cellular terminals at Ericsson, Inc. and was involved with the design of Flash memories at Advanced Micro Devices. He received a B.S. in Physics from Loyola University in 1985, an M.S.E.E. from Columbia University in 1987 and a Ph.D. from North Carolina State University in 1995.

Power Amplifier Linearity and Efficiency Challenges for Sub-6GHz 5G Systems

Christian Fager, *Chalmers University of Technology, Gothenburg, Sweden*

This talk will survey some of the key challenges that drive the development of PAs for 5G mobile communication. Focusing on sub-6GHz applications, promising PA and transmitter architectures will be reviewed in the context of 5G transmitter needs. A new simulation framework will be introduced for accurate prediction of the joint nonlinear effects of circuit-, antenna-, and thermal design. Different phased-array and Massive-MIMO transmitter examples will illustrate typical applications.

Christian Fager received his Ph.D. degree from Chalmers University of Technology, Sweden, in 2003. Since 2015, he has been a Professor at the Microwave Electronics Laboratory at the same university. He was a visiting scholar with the University of California, San Diego, in 2014. Dr. Fager has authored and co-authored more than 120 publications in international journals and conferences. His research is focused in the area of energy-efficient and linear transmitters for future wireless communication systems. Dr. Fager received the Best Student Paper Award at the IEEE International Microwave Symposium in 2003. He serves as a TPC member of the IEEE IMS and INMMiC technical conferences and is currently an Associate Editor of IEEE Microwave Magazine.

Silicon Technologies for Sub-6GHz Front-End and Radio Integrations

Alvin Joseph, *GLOBALFOUNDRIES, Essex Junction, VT*

Silicon technologies have shown great promise in penetrating the front-end for switch, LNA, and to some extent PA integration for 4G LTE advanced cellular handsets and in displacing the more costly option of GaAs based ICs. With continued bandwidth needs, next generation phones are adopting sub-6GHz LTE-U / eLAA standards which would drive more silicon content, but, with a focus on integration. We will present the current and future potential of various RFSOI and SiGe BiCMOS technologies for these applications.

Alvin Joseph joined the SiGe BiCMOS technology development team at the IBM Microelectronics Division, Essex Junction, Vermont after receiving his Ph.D. degree in electrical engineering from Auburn University, Auburn, AL, in 1997. He is currently a GLOBALFOUNDRIES fellow. In his leadership role, Alvin is driving the RF technology roadmap with a focus on broad market adoption of the silicon-based RF front-end and mm-wave solutions. He has over 100 technical journal papers and conference publications as well as over 100 patents issued. In 2016, he was appointed an IEEE Fellow for contributions to silicon-germanium bipolar-CMOS and RF silicon-on-insulator technology.

F2: Memory-Centric Computing from IoT to Artificial Intelligence and Machine Learning

Organizer: **Fatih Hamzaoglu,** *Intel, Hillsboro, OR*

Committee: **Meng-Fan Chang,** *National Tsing Hua University, Hsinchu, Taiwan*
Ki-Tae Park, *Samsung, Hwaseong, Korea*
Yasuhiko Taito, *Renesas, Kodaira, Tokyo, Japan*
Alicia Klinefelter, *Nvidia, Durham, NC*
Naveen Verma, *Princeton University, Princeton, NJ*

This forum will present state-of-the-art memory-centric architectures, as well as future innovative solutions to enable energy-efficient, high-performance AI/ML applications. It will also describe the challenges and solutions from edge processors to cloud applications, such as algorithmic accuracy, cost, security and practicality, including technology readiness to areas where further technology development is needed. At the high-performance and machine-learning end, emerging and storage-class memories are going to change the memory hierarchy. Meanwhile, low-power high-bandwidth DRAMs and SRAMs continue to be innovated around, to remain the workhorses of the latest process nodes (HBM, GDDR6, 7nm-FinFET SRAM etc.). Furthermore, with the explosive growth of memory intensive workloads like machine learning, video capture/playback, language translation, etc. there is a tremendous interest in preforming some compute near memory, by placing logic inside the DRAM/NVM main-memory die (AKA near-memory compute), or even doing the compute within the SRAM/STTRAM/RRAM array embedded within the compute die (AKA in-memory compute). In either case, the motivation is to reduce the significant data movement between main/embedded memory and compute units, as well as to reduce latency by preforming many operations in parallel, inside the array. Many challenges to productize these ideas remain to be addressed, including area-cost trade-offs for adding logic in the memory die for near-memory compute or augmenting embedded arrays with mixed-signal circuits to enable in-memory compute. Noticeable degradation of S/N ratio, especially in the in-memory compute case, is lurking and may direct its use to specific applications.

Hardware Enabled AI

Bill Dally, *Nvidia, Santa Clara, CA*

The current resurgence of artificial intelligence is due to advances in deep learning. Systems based on deep learning now exceed human capability in speech recognition, object classification, and playing games like Go. Deep learning has been enabled by powerful, efficient computing hardware. The algorithms used have been around since the 1980s, but it has only been in the last few years - when powerful GPUs became available to train networks - that the technology has become practical. This talk will describe current hardware for deep learning and research to make this hardware more efficient. Dedicated accelerators, special instructions, data representation, sparsity, and analog methods will be discussed.

Bill Dally is chief scientist at NVIDIA and senior vice president of NVIDIA Research, the company's world-class research organization, which is chartered with developing the strategic technologies that will help drive the company's future growth and success.

Dally joined NVIDIA in 2009 after spending 12 years at Stanford University, where he was chairman of the computer science department and the Willard R. and Inez Kerr Bell Professor of Engineering. Dally and his Stanford team developed the system architecture, network architecture, signaling, routing and synchronization technology that is found in most large parallel computers today. Dally was at the Massachusetts Institute of Technology from 1986 to 1997, where he and his team built the J-Machine and M-Machine, experimental parallel computer systems that pioneered the separation of mechanism from programming models and demonstrated very low overhead synchronization and communication mechanisms. From 1983 to 1986, he was at the California Institute of Technology, where he designed the MOSSIM Simulation Engine and the Torus Routing chip, which pioneered wormhole routing and virtual-channel flow control. Dally is a cofounder of Velio Communications and Stream Processors. He is a member of the National Academy of Engineering, a Fellow of the American Academy of Arts & Sciences, a Fellow of the IEEE and the ACM. He received the 2015 Funai Achievement Award from the Information Processing Society of Japan, the 2010 Eckert-Mauchly Award, considered the highest prize in computer architecture, as well as the 2004 IEEE Computer Society Seymour Cray Computer Engineering Award and the 2000 ACM Maurice Wilkes Award.

Dally has published more than 200 papers, holds more than 100 issued patents and is the author of three textbooks, "Digital Systems Engineering," "Principles and Practices of Interconnection Networks" and "Digital Design, A Systems Approach." Dally received a Bachelor of Science degree in electrical engineering from Virginia Tech, a Master of Science degree in electrical engineering from Stanford University, and a Ph.D. in computer science from the California Institute of Technology.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Embedded Memory Solutions for AI, ML and IoT

Masanori Hayashikoshi, *Renesas Electronics, Tokyo, Japan*

The extensive sensor nodes have a significant role in obtaining a large amount of real-time information for IoT, and power reduction for accumulating the huge information data without degrading processing performance is becoming more important. In this presentation, non-volatile memory technologies for low-power operation and power management schemes for low-power computing suitable for IoT will be mainly described. In addition, the embedded non-volatile memory solution for endpoint AI applications will be discussed.

Masanori Hayashikoshi received the BS and MS degrees in electronic engineering from Kobe University, Hyogo, Japan, in 1984 and 1986, respectively, and the Ph.D. degree in electrical engineering and computer science from Kanazawa University, Ishikawa, Japan, in 2018.
He is a Senior Principal Specialist of Shared R&D Division 1 in Renesas Electronics Corporation. He has been engaged in the research and development of Flash memory, high density DRAM, low power SDRAM, embedded MRAM for MCUs, and Normally-off computing architecture for further low-power solution with NVRAMs. He is now engaged in a feasibility study of in-memory computing for future neural network application.

High-Bandwidth Memory (HBM) DRAM for Energy-Efficient Near-Memory Computing

Kyomin Sohn, *Samsung Electronics, Hwaseong, Korea*

HBM DRAM is the best memory solution for HPC, high-end graphics and AI applications. It provides unparalleled high bandwidth by wide-IO with stacked DRAMs using TSV technology. However, it has also many challenges like power, thermal, and testability.
As one of the new solutions for the data-intensive computing era, HBM DRAM has additional possibilities and chances for more efficient and powerful near-memory computing. In this talk, HBM as a near-data-processing platform will be presented in various aspects.

Kyomin Sohn received the B.S. and M.S. degrees in Electrical Engineering in 1994 and 1996, respectively, from Yonsei University, Seoul. From 1996 to 2003, he was with Samsung Electronics, Korea, involved in SRAM Design Team. He received the Ph.D. degree in Electrical Engineering and Computer Science in 2007 from KAIST, Daejeon, Korea. He rejoined Samsung Electronics in 2007, where he has been involved in the DRAM Design Team. He is a Master (Technical VP) in Samsung and he is responsible for HBM DRAM Design and Future Technology of DRAM Design. His interests include the next generation 3D-DRAM, robust memory design, and processing-in-memory for artificial intelligence applications.

Novel Memory/Storage Solutions for Memory-Centric Computing

Mohamed Arafa, *Intel, Chandler, AZ*

The exponential growth in connected devices is generating a staggering number of digital records. This era of big data is driving fundamental changes in both memory and storage hierarchy; data and compute need to be brought closer together to avoid protocol inefficiencies. In this presentation, novel solutions for memory-centric architecture will be discussed with a focus on their value, performance, and power.

Mohamed Arafa is a Sr. Principal Engineer with the Data Center Group Intel Corporation. He recently has been focusing on the definition and evaluation of novel server memory architectures. He has more than 20 years of industry experience and seven years of academic/research experience. He authored/co-authored more than 30 technical papers and holds 10 US patents. He also has been an adjunct professor in the Electrical Engineering department at Arizona State University. Dr. Arafa holds a Ph.D. in Electrical Engineering from the University of Illinois at Urbana-Champaign and an MBA from the W.P. Carey School of Business.

The Deep In-memory Architecture for Energy Efficient Machine Learning

Naresh Shanbhag, *University of Illinois at Urbana-Champaign, Urbana, IL*

The *Deep In-Memory Architecture* (DIMA) is a non-von Neumann approach that breaches the memory wall by reading functions of data, embedding analog operations around the bitcell array, and leveraging the learning capabilities of inference algorithms. DIMA reads generate an inference leading to 50X-to-100X reduction in the energy-latency product. This talk will describe key DIMA principles, design challenges, their solutions realized via IC prototypes, and discuss future prospects.

Naresh R. Shanbhag is the Jack S. Kilby Professor of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign. He worked at AT&T Bell Laboratories at Murray Hill (1993-95) on VDSL chip-sets. Dr. Shanbhag's research interests are in the design of energy-efficient systems, architectures, and integrated circuits for machine learning, signal processing, and communications. He is an IEEE Fellow and served as the Director of the Systems On Nanoscale Information fabriCs (SONIC) Center (2013-17). He was a co-founder and Chief Technology Officer of Intersymbol Communications, Inc., (now part of Finisar Corporation, Inc.,), a start-up providing DSP-enhanced ICs for dispersion compensation of OC-192 links.

Low Power SRAM for IoT and ML Applications

YK Chong, *ARM, Austin, TX*

The Internet of Things (IoT) and Machine Learning (ML) involve a tremendous amount of data movement. IoT and ML bandwidths are often limited by the high energy cost of data communication. Consequently, it is important to have low-power memories to transfer the data.

In this presentation, I discuss several low power memory solutions for IoT and ML applications. I further examine reliability and safety memory techniques that are essential for designs in the rapidly growing automotive sector. Finally, I address the design and testability challenges of embedded MRAM and other non-volatile replacements for the current ML, IoT, and automotive memories.

YK Chong received the B.S. and M.S. degrees from Mississippi State University, Starkville, Mississippi in 1993 and 1995, respectively. He is currently a Senior Technical Director and Distinguished Engineer at Advanced Product Development, ARM Physical Design Group. He leads the memory technology development of Client, Infrastructure, and Machine Learning applications at ARM. He is also responsible for leading the Design Technology Co-Optimization (DTCO) with major foundries. He holds more than 30 U.S. patents in integrated circuit design and technology.

Advanced Memory, Logic and 3D Technologies for In-Memory Computing and Machine Learning

Stefan Cosemans, *imec, Leuven, Belgium*

Artificial neural network (ANN) inference is now key to many applications. Some run in the cloud, others on wireless IoT sensors, but they all require an unprecedented amount of computation and storage. This talk first discusses novel memory, logic and 3D technologies that can improve power, performance and cost of digital inference. Second, it examines the potential of analog in-memory computing for inference, with attention to the challenges and opportunities at device, circuit, architecture and algorithm levels.

Stefan Cosemans is a Principal Member of Technical Staff at imec, Leuven, Belgium. His current focus is on circuits and memory devices for machine-learning accelerators. He received his Ph.D. degree in 2009 from KU Leuven for his work on low-power SRAM circuits. From 2010 to 2014, he worked at imec on circuit design for RRAM, STT-MRAM and other novel memory devices. From 2015 to 2017, he led the development of single-rail near-threshold SRAM compilers at sureCore Ltd. In 2017, he joined imec's Machine Learning program.

Deep-Learning Hardware Acceleration: Opportunities in Memory Design

Leland Chang, *IBM T. J. Watson Research Center, Yorktown Heights, NY*

Deep-learning hardware acceleration has become a critical need across many application domains. While reduced precision arithmetic enables energy-efficient compute, feeding data to many parallel engines remains a key bottleneck – one that strongly depends on memory innovation to improve core, chip, and system architectures. This talk describes on-chip/external memory, disk requirements to support a diversity of neural networks and considers the impact of coming developments in algorithms and technology.

Leland Chang is a Principal Research Staff Member and Senior Manager of AI Hardware Acceleration at the IBM T. J. Watson Research Center. His work has spanned technology, circuits, and systems, including contributions such as the development of FinFET devices, SRAM and register file arrays, and integrated voltage regulators as well as the management of teams that have driven CMOS technology roadmaps, IBM server and mainframe microprocessor products, and AI hardware accelerators. He received the B.S., M. S. and Ph.D. degrees in electrical engineering and computer sciences from UC Berkeley and has authored more than 80 papers and 120 patents. He is a former program committee member and memory subcommittee chair of the ISSCC.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

F3: The Complete Advanced Driver-Assistance System (ADAS) Sensing Network

Organizer: **Bruce Rae,** *STMicroelectronics, Edinburgh, United Kingdom*

Co-Organizers: **Pedram Lajevardi,** *Robert Bosch, Sunnyvale, CA*
Tim Piessens, *ICsense, Leuven, Belgium*
Gerard Villar Pique, *NXP Semiconductor, Eindhoven, The Netherlands*

Committee: **Chris Van Hoof,** *IMEC, Leuven, Belgium*
Kofi Makinwa, *Delft University of Technology, Delft, The Netherlands*
Yogesh Ramadass, *Texas Instruments, San Jose, CA*

As the automotive industry continues its intensive commercial and R&D drive towards ever more sophisticated advanced driver-assistance systems, the focus on sensor quality, reliability and security has never been greater. Such systems comprise a multi-sensor network and decision engine capable of processing high volumes of data, in real-time to take safety critical, vehicle actions. In this forum, we will review the ADAS system and the requirements of OEMs and system integrators before presentations on state-of-the-art automotive sensors covering CIS, inertial sensors, LiDAR, RADAR and ultrasonics.

Overview of ADAS Systems

Patrick Denny, *Valeo, Galway, Ireland*

Automotive sensing systems are the building blocks of advanced driving assistance systems and pave the way for the future of vehicles. The new applications that are being realized drive the novelty in the performance, implementation, fusion and architectures for modern vehicles and Tier 1 development links the solid-state circuit innovation with the end-user applications. This talk provides an overview of the sensing technologies and their co-ordination from a Tier 1 perspective, prefacing the subsequent deep dives in the automotive forum.

Patrick Denny is a senior technologist at Valeo where he has developed automotive sensing systems for 16 years for the major automotive/tech companies, ranging from image and RF sensor and system design to manufacturing to algorithmic development. He is Adjunct Professor of Automotive Electronics at the National University of Ireland, Galway, and has co-founded and chaired conferences and standards bodies such as AutoSens, IS Auto, IEEE P2020 and the AVM conference at Electronic Imaging. Patrick has published and consulted widely, has over 100 patent filings and holds an M.Sc in Mathematics and a Ph.D. in Physics.

ADAS - AV Central Processing Hub

Richard Gavin Bramley, *Nvidia, Santa Clara, CA*

The talk will present the philosophy behind NVIDIA's Drive processor family and ways of seamlessly seamlessly scalling perfromance through multiple Xavier processors and TensorCore GPUs. High-performance AI computing and functional safety must coexist to safely handle highly automated and fully autonomous driving. The talk will explore the use of multiple diverse processing paths to enable sensor redundancy and efficient sensor fusion algorithms.

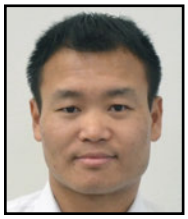
After obtaining his Ph.D. in "Signal Processing Architecture" at the University of Birmingham, UK, Richard spent the first part of his career in microelectronics for consumer applications. He went on to work in low-power devices for mobile telephony and was chief architect of the world's first multi-core processor for mobile phone. Following this period he moved into automotive microelectronics at ST Microelectronics working on the development of a low-power radiation-resilient silicon FD-SOI and ADAS processors. Richard is now driving automotive-functional safety for GPUs at NVIDIA, based in Santa Clara, USA.

Intelligent Sensors for Autonomous Driving

Sergey Velichko, *ON Semiconductor, Meridian, ID*

Autonomous driving is emerging at the higher-end of ADAS requiring heterogeneous intelligent sensing that would provide capabilities for robust object detection including vulnerable road users. The wide acceptance of LED signs and lights, illumination and weather conditions, and automotive range of temperatures are addressed by latest automotive intelligent sensors, such as high-dynamic-range imaging sensors, RADARs, and LIDARs. A potpourri of automotive imaging-sensor advancements is highlighted.

Sergey Velichko received the M.Sc. degree in electrical engineering with honors from the Kishinev Polytechnic Institute (KPI), Kishinev, Moldova, in 1985. He is a Senior Member of IEEE. Since 1985, he worked in various teaching and R&D positions at the KPI, Micron Technology, HP, and, since 2007, at the ON Semiconductor (former Micron Imaging and Aptina). He authored and co-authored more than 50 publications and 45 issued and pending patents. His interests include intelligent sensor technologies and product strategic development; novel imaging pixel architectures, characterization techniques, and pixel modeling; and innovations in automotive sensor designs.



Achieving High-Performance and Robustness: Technologies for Automotive MEMS Gyroscopes

Deyou Fang, *STMicroelectronics, Coppell, TX*

This presentation describes a digital-intensive ASIC architecture for automotive MEMS gyroscopes targeting autonomous driving applications. With a system clock that is independent of MEMS drive F_d, it avoids using BP-$\Delta\Sigma$ ADCs to digitize drive motion and Coriolis acceleration signals, and it guarantees system performance over temperature and life time. Gyros based on this architecture have achieved less than 6mDPS/ noise and less than ±1DPS ZRO under various mechanical shocks and vibrations.

Deyou Fang received his M.S. and Ph.D. degrees in Electrical and Computer Engineering from the University of Florida, Gainesville, FL, in 2003 and 2006, respectively.

He has been with the Analog MEMS & Sensor (AMS) Group of STMicroelectronics since 2013,where he is in charge of architecture and system design for automotive inertial sensors. From 2006 to 2012, he was with Freescale Semiconductor, Tempe, AZ, where he designed ASICs for automotive inertial sensors. During his research at the University of Florida, he focused on ASIC design for CMOS-MEMS accelerometers and gyroscopes. Before 2001, he was with Huawei Technologies Co. Ltd., Shenzhen, China, where he worked on communication systems, such as telecommunication switches, GSM, and CDMA wireless systems.

LiDAR Sensor Systems

Ming Wu, *UC Berkeley, Berkeley, CA*

This presentation will cover the principle and the current state of the art of automotive LiDAR (light detection and ranging) sensor systems. Fundamental design considerations for flash and scanning LiDARs with pulsed time of flight, frequency-modulated continuous-wave ranging methods will be presented. Long-range solid state LiDARs with micro-electro-mechanical-systems (MEMS), optical phased array, silicon photonic, electro-optic and laser-based scanners will be discussed.

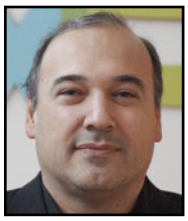
Ming C. Wu received the B.S. degree in electrical engineering from National Taiwan University in 1983 and the M.S. and Ph.D. degrees in electrical engineering and computer sciences from the University of California, Berkeley, in 1986 and 1988, respectively.

From 1988 to 1992, he was a Mmeber of Technical Staff with AT&T Bell Laboratories, Murray Hill. From 1993 to 2004, he was a professor with the University of California, Los Angeles (UCLA). Since 2005, he has been a faculty member at the University of California, Berkeley, where he is currently Nortel Distinguished Professor. He is also Co-Director of the Berkeley Sensor and Actuator Center (BSAC) and Faculty Director of UC Berkeley Marvell Nanolab.

Prof. Wu pioneered the development of micro-optical systems enabled by micro-electro-mechanical systems (MEMS). He demonstrated the first free-space integrated optics using the surface-micromachining technology. He co-founded OMM Inc. to commercialize 16x16 fiber optic switches realized by this technology. Recently, he has combined silicon photonics and MEMS to create 128x128 switches, the largest integrated switch reported to date. He has also contributed to the optofluidics field. He invented optoelectronic tweezers that enables massively parallel manipulation of individual biological cells by projectd optical images. It is being commercialized by a spin-off company, Berkeley Lights Inc., for single cell biology applications. He is the autor of more than 500 papers in journals and conferences and 25 U.S. patents.

Prof. Wu is an IEEE Fellow and a Packard Foundation Fellow (1992 – 1997). He was a member of the IEEE Photonics Society Board of Governors from 2013 to 2016. He was a recipient of the IEEE Photonics Society William Streifer Scientific Achievement Award in 2016, the Paul F. Forman Engineering Excellence Award in 2007, and the C.E.K. Mees Medal in 2017, both from the Optical Society of America.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE



The Evolving Landscape in Automotive Radars: Waveform, System Implementation, and IC Technologies

Cicero S. Vaucher, *NXP Semiconductors, Eindhoven, The Netherlands*

In this talk, we review the latest signaling waveforms, system implementation aspects, and IC technology options for next-generation car-radar products. We then discuss a fully integrated 76-to-81GHz band radar front-end in a 40nm CMOS silicon technology. Link budget and phase noise considerations are presented and discussed and followed by results of a sensor-level MIMO implementation demonstrating increased angular resolution with the IC.

Cicero S. Vaucher received the Ph.D. degree in Electrical Engineering from the University of Twente, The Netherlands, in 2001. He is presently with NXP Automotive Business Unit working as a Radar Product Architect. He is also a part-time professor at TU Delft, The Netherlands.

His research activities and interests include micro-wave and mm-wave transceiver architectures, radar system implementations, and signal processing. He is the author of Architectures for RF Frequency Synthesizers and a co-author of Circuit Design for RF Transceivers. He is (co-) inventor of 28 unique patent families.

Dr. Vaucher is acting on the IEEE-MTT Wireless-Enabled Automotive and Vehicular Applications Technical Committee 27. He is a Senior Member of the IEEE, and an NXP Technical Fellow.

Ultrasonic Sensors and Their Role in ADAS

Ajit Sharma, *Texas Instruments, Dallas, TX*

We will discuss value proposition, trends, and challenges in the design of mixed-signal SoCs for ultrasonic park assists in the context of 3 generations of TI ultrasonic park-assist SoCs. Forum participants will gain an in-depth understanding of the trade-offs involved in developing algorithms and silicon for this space. This will provide an opportunity for transducer vendors, chip designers, software and signal processing experts, and OEM system integrators to interact and exchange ideas.

Ajit Sharma received the B.E. (Hons.) degree from BITS Pilani,India (2001), M.S. degree from Oregon State, Corvallis (2003), and the Ph.D. degree from the Georgia Tech, Atlanta (2007). He joined TI in 2007 to work on automotive sensor interface SoCs for pressure, occupant-weight sensors, and MEMS accelerometers and gyroscopes. Subsequently he joined TI's Kilby Labs to work on low-nergy, multi-modal AFEs for biomedical applications and led the system definition and design of ultra-low-energy authentication platforms for TI's Medical Product Line. Ajit currently manages Applications for Audio ADC/Codec and Voice Recognition products.

His primary research interests are in the areas of smart microsystems, sensor transduction, and low-power sensing SoCs.

Battery Management IC Design Complying with Functional Safety

Pieter De Muyter, *ICsense, Heverlee, Belgium*

As today's ADAS networks are expanding rapidly, there is an increasing demand for automotive Battery Management Systems (BMSs) that are able to balance fluctuations in battery loading conditions. These trends toward electrification and car autonomy require a high degree of quality and safety in BMS ICs in order to reduce the risk for hazardous malfunctioning. This presentation shows how automotive IC designs can integrate the functions of the BMS together with functional safety requirements.

Pieter De Muyter is VP engineering at ICsense. He received the M.S. degree in microelectronics (cum magna laude) from the Katholieke Universiteit Leuven, Belgium, in 2005. Since 2005, he has been working at ICsense as an analog design engineer and project leader in several high-voltage and mixed-signal state-of-the-art sensor ASIC developments. His key areas of expertise are automotive magnetic sensor ASICs and functional safety. From 2018, he is leading the ICsense engineering team.

F4: Intelligence at the Edge: How Can We Make Machine Learning More Energy Efficient?

Organizer: **Masato Motomura,** *Hokkaido University, Sapporo, Japan*

Committee: **Vivek De,** *Intel, Hilsboro, Oregon*
Long Yan, *Samsung Electronics, Hwaseong, Korea*
Dong Uk Lee, *SK hynix, Icheon, Korea*
Kyu-hyoun Kim, *IBM T. J .Watson, Yoktown Heights, NY*

Deep learning algorithms mainly used today in Machine Learning applications are very costly in terms of energy consumption, due to their large amount of required computations and large model sizes. Many issues like connectivity to the cloud, latency, privacy, and public safety could be resolved by establishing intelligent computing at the edge. This forum brings together architecture, circuit, and integrated design solutions to minimize the energy consumption of edge learning systems. Various approaches from different perspectives will be covered: neuro-inspired architectures, in memory computation, embedded sensing, and spiking neural networks.

Enabling Embedded Intelligence: Application, Architecture and Design Solutions

Bert Moons, *Synopsys, Leuven, Belgium*

Embedding neural intelligence at the edge remains challenging today, due to the huge computational and memory requirements and the large algorithmic diversity of modern vision, sensing, and sequence-modeling tasks. These challenges can be overcome by simultaneously co-optimizing smarter applications, leaner neural networks, optimized computer architectures and more efficient circuits. This talk is an overview of recent advances in the field and of Synopsys's solutions in its DesignWare EV6x IP.

Bert Moons received the Ph.D. degree in Electrical Engineering cum ultima laude from KU Leuven in 2018. He performed his Ph.D. research at ESAT-MICAS, focusing on energy-scalable and run-time adaptable digital architectures and circuits for embedded deep learning applications. Bert authored 15+ conference and journal publications, was a Visiting Research Student at Stanford University and received the SSCS pre-doctoral achievement award in 2018. Currently, he is with Synopsys, as a hardware design architect for the DesignWare EV6x Embedded Vision and AI processors.

Mixed-Signal Circuits for Inference at the Edge

Daniel Joel Bankman, *Stanford University, Stanford, CA*

Deep neural network (DNN) inference systems capable of operating autonomously at the edge have numerous applications, such as microrobotic vision and low-latency speech interfaces. In extremely resource-constrained platforms, where battery life is of primary importance, mixed-signal processing can minimize energy consumption by exploiting the noise tolerance of DNNs. This talk will focus on the increasingly popular crossbar architecture and its implementation with switched-capacitor circuits.

Daniel Bankman received the S.B. degree in electrical engineering from the Massachusetts Institute of Technology, Cambridge, MA in 2012 and the M.S. degree from Stanford University, Stanford, CA in 2015. Since 2012, he has been working toward the Ph.D. degree at Stanford University, focusing on mixed-signal processing for machine learning. He has held internship positions with Analog Devices and Intel. His research interests include algorithms, architectures, and circuits for energy-efficient learning and inference in smart devices. He was a recipient of the Texas Instruments Stanford Graduate Fellowship in 2012, the Numerical Technologies Founders Prize in 2013, and the SONIC John von Neumann Student Research Award in 2015 and 2017.

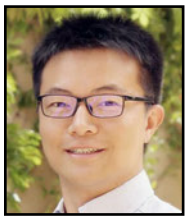
 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Processing Data Where It Makes Sense in Modern Computing Systems: Enabling In-Memory Computation

Onur Mutlu, *ETH Zürich, Zürich, Switzerland*

In this talk, I will discuss some recent research that aims to practically enable computation close to data. After motivating trends in applications, as well as technology, we will discuss at least two promising directions: 1) performing massively parallel bulk operations in memory by exploiting the analog operational properties of DRAM, with low-cost changes, 2) exploiting the logic layer in 3D-stacked memory technology in various ways to accelerate important data-intensive applications. In both approaches, we will discuss relevant cross-layer research, design, and adoption challenges in devices, architecture, systems, applications, and programming models. Our focus will be the development of in-memory processing designs that can be adopted in real computing platforms and real data-intensive applications, spanning machine learning, graph processing and genome analysis, at low cost. If time permits, we will also discuss and describe simulation and evaluation infrastructures that can enable exciting and forward-looking research in future memory systems, including Ramulator and SoftMC.

Onur Mutlu is a Professor of Computer Science at ETH Zurich. He is also a faculty member at Carnegie Mellon University, where he previously held Strecker Early Career Professorship. His current broader research interests are in computer architecture, systems, and bioinformatics. He obtained his Ph.D. and M.S. in ECE from the University of Texas at Austin and B.S. degrees in Computer Engineering and Psychology from the University of Michigan, Ann Arbor. His started the Computer Architecture Group at Microsoft Research (2006-2009), and held various product and research positions at Intel Corporation, Advanced Micro Devices, VMware, and Google. He received the inaugural IEEE Computer Society Young Computer Architect Award, the inaugural Intel Early Career Faculty Award, faculty partnership awards from various companies, a healthy number of best paper or "Top Pick" paper recognitions at various computer systems and architecture venues, and the ACM Fellow recognition "for contributions to computer architecture research, especially in memory systems."



Bandwidth-Efficient Deep Learning with Algorithm and Hardware Co-Design

Song Han, *Massachusetts Institute of Technology, Cambridge, MA*

Computation is cheap, bandwidth is expensive for efficient deep-learning accelerator design. This talk first describes efficient algorithms that: 1) reduce memory bandwidth by deep compression with pruning and quantization; 2) reduce networking bandwidth by deep gradient compression; and, 3) reduce engineer bandwidth by AutoML, which offers design automation for neural networks. Next, this talk describes efficient deep-learning accelerators that incorporate emerging algorithms.

Song Han is an Assistant Professor in the EECS Department of Massachusetts Institute of Technology and PI for HAN lab: Hardware for AI and Neural-nets. Dr. Han received the Ph.D. degree in Electrical Engineering from Stanford advised by Prof. Bill Dally. Dr. Han's research focuses on energy-efficient deep learning. He proposed “deep compression” that widely impacted the industry. His work on deep compression and hardware acceleration received best paper awards at ICLR 2016 and FPGA 2017, and these technologies led to DeePhi Tech that Dr. Han co-founded in 2016 providing efficient solutions for deep-learning computing.

On-Chip Epilepsy Detection: Where Machine Learning Enables Patient-Specific Pervasive Healthcare

Jerald Yoo, *National University of Singapore, Singapore*

Epilepsy is a severe and chronic neurological disorder that affects over 65 million people worldwide. We explore the challenges, limitations, and requirements of patient-specific on-chip seizure/epilepsy detection, and design strategies to overcome the outstanding issues. For patient-specific detection, embedded machine learning is essential and choosing the correct feature extraction and classification scheme is a must. We will see in detail the criteria and metrics to consider and trade-off. Finally, an on-chip epilepsy detection and recording SoC will be presented, which integrates all the components covered during the talk. The talk will conclude with interesting the aspects and opportunities that lie ahead.

Jerald Yoo received the B.S., M.S., and Ph.D. degrees in Electrical Engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2002, 2007, and 2010, respectively. Since 2017, he has been with the Department of Electrical and Computer Engineering, National University of Singapore, where he is currently an Associate Professor. From 2010 to 2016, he was with the Department of Electrical Engineering and Computer Science, Masdar Institute, Abu Dhabi, United Arab Emirates, as an Associate Professor. From 2010 to 2011, he was also with the Microsystems Technology Laboratories (MTL), Massachusetts Institute of Technology, MA as a Visiting Scholar. He pioneered research on low-energy Body-Area Network (BAN) transceivers, wearable healthcare using embedded machine learning and planar-fashionable circuit boards. He received several awards, including the ISCAS 2015 Best Paper Award (BioCAS Track), the Masdar Institute Best Research Award in 2015, and the A-SSCC Outstanding Design Award in 2005. He was the Vice Chair of IEEE Solid-State Circuits Society (SSCS), United Arab Emirates (UAE) Chapter. He served as the IEEE Solid-State Circuits Society’s Distinguished Lecturer (2017-2018).

Applying Principles of Neural Computation for Efficient Learning in Silicon

Mike Davies, *Intel, Hillsboro, OR*

In contrast to today's deep-learning models, the neurons of living brains learn continuously, effortlessly, with very few labelled data samples. Modern neuroscience is beginning to reveal the underlying computational principles that give rise to such remarkable behavior. This talk introduces Intel's Loihi neuromorphic research chip, describes its application of those principles to silicon, and shares results towards our vision of low power, real-time, on-chip learning.

Mike Davies leads Intel's Neuromorphic Computing Lab. Since joining Intel Labs in 2014, he has researched neuromorphic prototype architectures and is responsible for Intel's Loihi neuromorphic research chip. Mike began his career in 2000 as a founding employee of Fulcrum Microsystems. As Fulcrum's Director of Silicon Engineering, he pioneered high performance asynchronous design methodologies, as applied to several generations of industry-leading Ethernet switch products. Mike joined Intel in 2011 by Intel's acquisition of Fulcrum. He holds B.S. and M.S. degrees from Caltech in Computer Engineering, Mathematics, and Electrical Engineering.

Energy-Efficient Deep Learning for Intelligent Embedded Systems

Nitin Chawla, *STMicroelectronics, Greater Noida, India*

Deep Convolutional Neural Networks (DCNNs) or ConvNets achieve state-of-the art results in many applications involving recognition, identification and/or classification tasks. However, DCNNs are associated with a high computational cost, hindering their adoption in embedded and IoT domains. This talk focuses on SoC implementations of deep learning at the edge where performance requirements co-exist with tight constraints in terms of power and energy consumption. The talk will discuss design solutions involving configurable dataflow architectures and near-threshold voltage design, and will present measured results of state-of-the-art implementations of low-power AI edge nodes.

Nitin Chawla is a Fellow at the Central Research and Development organization at STMicroelectronics. Nitin's current area of work includes Low Power Neural Computing Architectures for Embedded Systems and Ultra Low Power Signal Processing Systems. Nitin graduated with a degree in Electronics and Telecommunication and has served in different R&D and product organizations over the last 20 years. Before joining STMicroelectronics he was the Chief Scientist of the HLS Product Group at Mentor Graphics Corporation. Nitin has 9 granted and several pending US patents. He has over a dozen peer-reviewed IEEE conference and journal publications including a Best Paper Award in the area of System On Chip Design.

Energy-Efficient Intelligent Vision Sensor for Next-Generation Mobile and Autonomous Applications

Eric Ryu, *Samsung Electronics, Hwasung, Korea*

Unlike a conventional camera, a Dynamic Vision Sensor (DVS) mimicking the human visual system outputs a stream of events at a rate of thousands of frames per second enabling energy-efficient machine vision. This talk will share broad industry development experience on DVSs, beginning with the working principle of the DVS, and including how to reduce data volume to achieve lower energy consumption, but higher speed performance, with improved dynamic range. The talk will then highlight two practical system examples: object recognition and SLAM.

Hyunsurk (Eric) Ryu received the Ph.D. degree in Physics from POSTECH in 1998, where his Ph.D. research focused on statistical interferometric image and signal processing. From 1998, he was with Samsung Advanced Institute of Technology (SAIT), where he led neuromorphic sensor and computing system research and development. In 2016, he moved to System LSI in Samsung Electronics. He is now a Master (Research VP), and leads machine vision research and product development. He also serves as a technology strategy leader and has executed several new sensor business projects.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

F5: 56Gb/s to 112Gb/s and Beyond – Design Challenges and Solutions in Wireline Communications

Organizer: **Andrew Joy**, *Cavium, Northampton, United Kingdom*

Committee: **Amir Amirkhany**, *Samsung, San Jose, CA*
Mounir Meghelli, *IBM, Yorktown Heights, NY*
Samuel Palermo, *Texas A&M University, College Station, TX*
Hyeon-min Bae, *KAIST, Daejong, Korea*
Takayuki Shibasaki, *Fujitsu Laboratories, Kanagawa, Japan*

This forum brings together experts from the forefront of all areas in the design of high-speed wireline SerDes. The talks will take you through a detailed review of current and next generation serial link designs and the associated challenges in each area, as well as where the industry and roadmap maybe heading beyond 112Gb/s.

The forum is a circuit- and algorithm-oriented review of high-performance tranceivers, focussing on all aspects of SerDes design including analog front-end circuits, high-speed ADC design for receivers, high-speed DAC design for transmitters, clock-and-data recovery techniques, wide-range PLLs and clocking, and adaptive equalization. Talks include the Forward Error Correction techniques that are essential for the future and those that are currently employed, as well as the error statistics and how solutions can be tailored around them. Short-reach links and their challenges are included, as well as longer-reach backplane links, together with the modeling techniques used with IBIS AMI for the verification of how well they will operate in practise. Finally, some non-standard techniques are also discussed to solving the bandwidth limitation problem.

Signaling Techniques and Trends for 56+Gb/s Electrical Links

Brian Ji, *Avera Semi, Hopewell Junction, NY*

The performance of 56Gb/s serial links will be reviewed, focusing on channel and package characteristics, and equalization architecture. Extending serial link performance to 112Gb/s will require improved channels with lower loss, but also extreme focus on controlling reflections and crosstalk within the package and channel. The competing demands of performance, power and area may lead to a divergence in SerDes architecture for applications. Some observations on moving beyond 112Gb/s will be presented.

Brian L. Ji received the B.S. degree from University of Science and Technology of China in 1984, and the Ph.D. degree in physics from Harvard University in 1991.

From 1996 to 2005, he worked at IBM Microelectronics on VLSI circuit design of 256Mb, 512Mb, 1Gb DRAMs, embedded memory, and high-speed serial link products. From 2005 to 2009, he was a research staff member at IBM T. J. Watson Research Center, researching in exploratory devices and circuits. From 2010 to 2015, he was a researcher at SUNY CNSE, working on CMOS derivatives, graphene and spintronics. In 2015, he joined GlobalFoundries, where he has been working on various high-speed SerDes products including 30Gb/s, 56Gb/s and 112Gb/s designs. He holds 38 U.S. patents.

ADCs for High-Speed Communication - Architectures and Circuit Implementations

Filip Tavernier, *KU Leuven, Leuven, Belgium*

This talk will focus on the ADC in high-speed wireline receivers. Depending on the signaling scheme, this ADC should combine a medium resolution, a multi-GHz sampling rate, a high analog bandwidth and a good power efficiency. Therefore, the focus will be on SAR ADCs. Techniques will be discussed that can maximise their speed. Next, the design of a recent single-stage high-speed SAR will be discussed. Finally, an new architecture will be introduced that will enable the next big increase in speed.

Filip Tavernier received the M.Sc. degree in Electrical Engineering and the Ph.D. degree in Engineering Science from KU Leuven, Belgium, in 2005 and 2011, respectively. From 2011 to 2014, he was Senior Fellow with the Microelectronics Group at CERN, Switzerland. From 2014 to 2015, he was a Postdoctoral Researcher with the Department of Electrical Engineering, ESAT-MICAS, KU Leuven, where he is Assistant Professor since 2015. His main research interests include circuits for optical communication, data converters, power converters, and chips for radiation environments. He is Treasurer of the IEEE SSCS Benelux Chapter, SSCS Webinar Coordinator for Europe, member of the TPC of ESSCIRC and tutorial chair of ESSCIRC 2017.

Design Techniques for a 112Gb/s PAM-4 Transmitter

Jihwan Kim, *Intel, Hillsboro, OR*

In this talk, we provide design techniques to overcome challenges in implementing an energy-efficient 112Gb/s transmitter using a cutting-edge CMOS technology. High-speed clock distribution techniques along with jitter analysis are covered. Implementations of the data serializers and output drivers are thoroughly reviewed with comparison to other state-of-the-art architectures. Designing a high-bandwidth output pad network is also thoroughly investigated.

Jihwan Kim (S'07-M'11) received his B.S. degree in electrical and computer engineering from Hanyang University, Seoul, Korea in 2005 and the M.S. and the Ph.D. degrees in electrical and computer engineering from Georgia Institute of Technology, Atlanta, GA, in 2007 and 2011, respectively. His doctoral research focused on design techniques for RF and mm-wave front-end integrated circuits, including power amplifiers (PAs), mixers, low-noise amplifiers (LNAs), and voltage-controlled oscillators (VCOs) using CMOS/SiGe technologies.

Since 2011, he has been with Intel's Advanced Design, Hillsboro, OR, working on designing integrated circuits and systems for ultra high-speed and low-power wireline data communications.

Equalization Techniques/Trends for 56+Gb/s and the Efficacy of Forward Error Correction

Cathy Liu, *Broadcom, San Jose, CA*

An overview of the evolution from 100GE to 400GE systems with 56-to-112Gb/s serial link designs and the associated challenges and potential solutions will initially be presented. The detailed modulation, equalization and forward error correction (FEC) evolution in terms of their architecture, modeling, performance and design implementation trade-offs will be discussed. In order to achieve joint optimal system performance, both SerDes Pre-FEC performance and system post-FEC performance will be studied.

Cathy Ye Liu, the distinguished engineer/director, heads up Broadcom SerDes architecture and modeling group. Previously she worked as director and distinguished engineer in Avago/LSI. Since 2002, she has been working on high-speed transceiver solutions. Previously she has developed read channel and mobile digital TV receiver solutions. Her technical interests are signal processing, FEC, and modeling in high-speed optical and electrical transceiver solutions. She has published many journal and conference papers and holds 20+ US patents. Currently she serves as a member of the board director of Optical Internetworking Forum (OIF) and the co-chair of the DesignCon technical track of high speed signal processing, equalization and coding. She received her B.S. degree in Electronic Engineering from Tsinghua University, China, and received her M.S. and Ph.D. degrees in Electrical Engineering from University of Hawaii.

High-Speed Clock-and-Data Recovery Circuit: Architecture, Modeling, Specifications, and Consideration

Jri Lee, *National Taiwan University, Taipei, Taiwan*

This talk presents design and analysis of modern clock-and-data recovery circuits dedicated to very high-speed data links. It briefly reviews different phase locking topologies, including PLL-based linear and binary CDRs, DLL and PI-based structures, oversampling architecture, all-digital realization, and burst-mode design. Capture range, frequency detection techniques, and jitter characteristics are incorporated as well to complete the design consideration. A design example is presented as a case study.

Jri Lee received the B.Sc. degree in electrical engineering from National Taiwan University (NTU), Taipei, Taiwan, in 1995, and the Ph.D. degree in electrical engineering from UCLA, CA, in 2003. He joined National Taiwan University in 2004, where he is currently Professor of electrical engineering. Prof. Lee has received the Beatrice Winner Award and the Takuo Sugano Award at ISSCC. He has served on the Technical Program Committee of ISSCC, Symposium on VLSI Circuits, and A-SSCC. He has also served as a guest editor of the JSSC, and a Distinguished Lecturer of the IEEE Solid-State Circuits Society from 2011 to 2013.

Challenges and Solutions for High-Bandwidth Density, Energy-Efficient, Short-Reach Signaling that Enables Massively Scalable Parallelism

John Wilson, *Nvidia, Durham, NC*

To maximize computational performance per watt we must realize communication is a tax on performance. In this talk, I'll cover solutions to enable Tera-Byte/s communication for on- and off-chip links, including: packaging limitations, signaling integrity and the design of energy-efficient high-speed links. I'll discuss the benefits and details of Ground Referenced Signaling for silicon interposers, organic packages & PCBs, and Balanced Charge-Recycling Signaling for on-chip communication.

John Wilson joined Nvidia Research in 2012 as a Sr. Research Scientist. Recently, he led a team in the design of a 25Gb/s Ground Referenced Signaling link, in 16nm FinFET CMOS, enabling Tera-Byte/s communication for on-package and off-package channels. He received the BS, MS and Ph.D. degrees in EE from NCSU. His interests include: high-speed I/O circuit design, on-chip signaling, signal integrity, packaging, & chip/package co-design. From 2003-to-2006 he was a Research Professor at NCSU leading projects in advanced packaging, transceivers for 3D-ICs, and circuits for on-chip signaling. From 2006-2012, while with Rambus, Inc. he worked on low-power, high-speed I/O circuit design. He has 66+ pubs., 30+ patents, and is a Senior Member of the IEEE.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Clock Generation and Distribution for 56Gb/s, 112Gb/s and Beyond Serial Links

Didem Turker, *Xilinx, San Jose, CA*

As serial link data rates increase to 56Gb/s and beyond, clock generation and distribution becomes ever more challenging. Jitter requirements are more stringent due to reduced symbol-time. For power efficiency, 56Gb/s to 112Gb/s transceivers employ multi-phase clocking which introduces additional area and jitter trade-offs and a need for multi-phase error and DCD correction schemes. This talk covers clock generation and distribution methods that address these challenges for 56Gb/s, 112Gb/s and beyond serial links.

Didem Turker received her Ph.D. degree in electrical engineering from Texas A&M University, in 2010. During her doctoral studies, she worked on frequency synthesizers and all-digital PLLs. She's been with the SerDes Technology Group at Xilinx, USA, since 2011, where she is a senior design manager and leads the PLL and clocking team that develops clock generation for high speed transceivers and RF SoCs. Her research interests include wireline and wireless transceiver circuits.

Link Modelling and the Challenges and Limitations for IBIS AMI for 56+Gb/s in Matching Circuit Performance

Mike L. Steinberger, *Signal Integrity Software, Chippewa Falls, WI*

Since the mid 1990's, high-speed serial links have evolved from data rates of hundreds of megabits per second to the point where individual channels with net data rates of 56Gb/s to 112Gb/s and beyond are a realistic expectation. While there have been modest improvements to the passive wireline channels over which these high speed serial links operate, most of the increase in data rate has come from more and more effective circuit designs to cope with the intersymbol interference these passive channels generate. Feed forward equalization (FFE) was augmented by increasingly effective circuits for decision feedback equalization (DFE) and continuous-time linear equalization (CTLE). Within the last several years, these designs have been further augmented through the use of pulse amplitude modulation (especially four level PAM4) and forward error correction (FEC). This evolution has been supported by increasingly detailed and accurate performance modeling, focused primarily on estimating the residual effects of intersymbol interference. This talk will present some of the performance modeling techniques currently under development, especially for modeling the effects of nonlinear channels on PAM4 and the response of FEC to the error patterns produced by intersymbol interference.

Michael Steinberger, Ph.D., has over 30 years experience designing very high speed electronic circuits. Dr. Steinberger holds a Ph.D. from the University of Southern California and has been awarded 14 patents.

He is currently responsible for the architecture of SiSoft's Quantum Channel Designer tool for high speed serial channel analysis. Before joining SiSoft, Dr. Steinberger led a group at Cray, Inc. performing SerDes design, high speed channel analysis, PCB design and custom RAM design.

Non-Differential Techniques: Multi-Wire Multi-level I/O

Armin Tajalli, *University of Utah, Salt Lake City, UT*

Demand for growing data transfer bandwidth over wireline copper links with denser routing channels and more stringent energy budget, underlines the need to develop proper signaling schemes for these applications. This presentation explores advanced signaling methods for low-power and extremely wideband data communication over ISI-limited channels. Exploiting proper spatial and temporal signaling, it is possible to improve spectral efficiency, and relax the required amount of equalization..

Armin Tajalli received his B.S. from Sharif University of Technology, Tehran, Iran, and the Ph.D. from Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. He has 20 years of industrial experience in the field of analog and mixed-signal design. He was a Senior Analog Architect with Kandou Bus, Lausanne, Switzerland (2010-2017). Since December 2017, he has been an assistant professor at the University of Utah, Salt Lake City.

F6: The Right Tool for the Job: Application-Optimized Data Converters

Organizers: **Kostas Doris,** *NXP Semiconductors, Eindhoven, The Netherlands*
Seng-Pan (Ben) U, *University of Macau and Synopsys Macau, Macau*
Jan Westra, *Broadcom, Bunnik, The Netherlands*

Committee: **Yiannos Manoli,** *University of Freiburg - IMTEK, Germany*
David Robertson, *Analog Devices, Wilmington, MA*
Michael Flynn, *University of Michigan, Ann Arbor, MI*

Moderator: **David Robertson,** *Analog Devices, Wilmington, MA*

Conditioning the data converter function to the properties of evolving CMOS technologies through architectural, algorithmic and circuit innovation enabled the last decades making radical steps in accuracy, speed and power efficiency. In analogy to this, optimizing the data converter for the unique requirements of emerging applications such as 5G mobile, biomedical and IoT, imaging, automotive radars and wireline communications offers the designer new degrees of freedom to improve performance and power efficiency of the data converter.

This forum addresses the unique challenges of application-specific data converters and presents the best tradeoffs and optimum architectures for each application.

Application-Optimized Data Converters: An Introduction

Boris Murmann, *Stanford University, Stanford, CA*

Over the past several decades, the symbiotic interplay between technology push and application pull has led to remarkable performance gains in data converters. While there appears to be no weakening in application pull, bottom-up innovation has slowed. A promising remedy is to explore application-specific optimizations based on system and signal insight. In this talk, we will review performance trajectories of data converters and discuss the rationale for such application-optimized design.

Boris Murmann is a Professor of Electrical Engineering at Stanford University. He joined Stanford in 2004 after completing his Ph.D. degree in electrical engineering at the University of California, Berkeley in 2003. Dr. Murmann's research interests are in mixed-signal integrated circuit design, with special emphasis on sensor interfaces, data converters and custom circuits for embedded machine learning. He has served as an Associate Editor of the IEEE Journal of Solid-State Circuits, an AdCom member and Distinguished Lecturer of the IEEE Solid-State Circuits Society, as well as the Data Converter Subcommittee Chair and the Technical Program Chair of the IEEE International Solid-State Circuits Conference (ISSCC). He is a Fellow of the IEEE.

ADCs for Biomedical and IoT

ChihCheng Hsieh, *National Tsing Hua University, Hsinchu, Taiwan*

For IoT and biomedical systems, ultra-low power analog-to-digital interfacing circuits are highly demanded. SAR ADCs have proven to be an efficient solution in advanced technology nodes. This talk will explore the application-driven techniques of low-power SAR ADCs. The power reduction techniques of the main building blocks will be introduced first. Then, the data-oriented low-power conversion techniques will be examined with examples from literature. Finally, we conclude with future perspectives.

Chih-Cheng Hsieh received the Ph.D. degree from National Chiao-Tung University, Taiwan, in 1997. From 1999-2007, he was with Pixart Imaging Inc., Taiwan, where he led the Mixed-Mode IC department. In 2007, he joined National Tsing-Hua University, Taiwan, where he is currently a Full Professor. His research interests include low-voltage, low-power ADC and CMOS image sensor IC design.

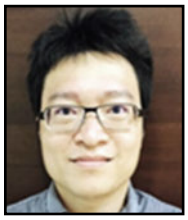
 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Filtering Delta-Sigma ADCs for Mobile Receivers

Maurits Ortmanns, *University of Ulm, Ulm, Germany*

Mobile receivers are highly challenged by interferers. The $\Delta\Sigma$ ADC offers the advantage to implicitly realize signal filtering and thus improved interferer robustness. Most power efficient $\Delta\Sigma$ ADCs however have been realized with out-of-band peaking, posing even higher challenges to the dynamic range and phase noise. This talk reviews the filter properties of $\Delta\Sigma$ ADCs, highlights the possibilities of embedding additional filtering into the loop, and extends this to bandpass $\Delta\Sigma$ for superior interferer robustness.

Maurits Ortmanns (M'04–SM'11) received Dr.-Ing. degree in 2004. From 2004 - 2005, he was with sci-worx GmbH as technical project leader. From 2006-2008 he was Assistant Professor for Integrated Interface Circuits at the University of Freiburg. Since 2008, Prof. Ortmanns is full professor at the University of Ulm, Germany, where he heads the Institute of Microelectronics. Prof. Ortmanns' main research interests include mixed-signal integrated circuit design with special emphasis on data converters and biomedical applications. He holds several patents, is author of the book "Continuous-Time Sigma-Delta A/D Conversion", and published several other book chapters and more than 250 IEEE journal and conference papers.



Data Converter Design Considerations for Mobile Transceivers: Benchmark and Tends from 4G LTE to 5G NR

Tien-Yu Lo, *MediaTek, Hsinchu, Taiwan*

Mobile phones have played an important role in the past decade, and the cellular communication system is the key block in the device. There are several data converter architectures to implement cellular transceiver chips in industry. In this talk, the concept based on the transceiver system view for data converters will be discussed, and then the potential topology to achieve the target in link budget will be shown. Finally, the challenge of the data converter from 4G LTE to 5G NR will be presented.

Tien-Yu Lo received the B.S., M.S., and Ph.D. degrees in electrical and computer engineering from National Chiao Tung University, Hsinchu, Taiwan, in 2001, 2003, and 2007. Since 2008, he joined MediaTek Inc. and currently serves in the position of Senior Technical Manager. In 2009, he authored one book 1-V CMOS Gm-C Filters: Design and Applications (Springer). He published 35 international journal/conference papers and holds several US Patents. His work focuses on 4G/5G multi-mode wireless baseband, audio signal recording, mobile fuel gauge, and high precision sensor interfaces. Dr. Lo served on the Editorial Board Member of Journal of Analog Integrated Circuits and Signal Processing since 2013 and is a Senior Member of IEEE.

Data Converters for Imaging / Lidar Sensors

Ron Kapusta, *Analog Devices, Wilmington, MA*

Digital imaging is pervasive, from security cameras to autonomous vehicles, and the ADCs in these systems vary dramatically. This talk will explore the state-of-the-art in converter design for each of these applications, covering a wide variety of converters, including: highly parallel ADCs, giga-sample rate time-interleaved ADCs for lidar, and even TDCs for direct time-of-flight. The relevant trade-offs for each application will also be discussed, such as area, linearity, throughput, and power.

Ron Kapusta received the B.S. and M.Eng. degrees from the Massachusetts Institute of Technology in 2002. Upon graduation, he joined Analog Devices, designing data converters and sensor interface circuits for digital imaging systems. More recently, he has worked on electrical and photonic technologies as an architect for Lidar. Ron has presented papers and invited educational talks at multiple IEEE conferences. He holds 30 issued U.S. patents and won the 2013 JSSC Best Paper award. He has served on the technical program committees for CICC and VLSI Circuits.

Analog-to-Digital Converters for Automotive Radar

Erwin Janssen, *NXP Semiconductors, Eindhoven, The Netherlands*

In this presentation we will take a look at analog-to-digital converters (ADCs) for automotive frequency modulated continuous wave (FMCW) radar. After a brief introduction of the application we will extract requirements of the ADC. These include not only raw signal conversion performance, but also system-level tradeoffs and functional safety. A number of general-purpose state-of-the-art data converters will be analyzed for usage in an FMCW radar system. Finally, some radar specific ADC designs will be presented.

Erwin Janssen was born in The Netherlands in 1976. He received the M.Sc degree cum laude in electrical engineering, with additional degree in computer science from the University of Twente in 2001, and the Ph.D. degree in electrical engineering from the Technical University of Eindhoven in 2010. From 2001 to 2006 he was with Philips Research Laboratories. In 2006 he joined NXP Semiconductors, where he has been involved in mixed-signal IC design.

Multi GS/s Data Converters for Wireline and Optical Links

Jun Cao, *Broadcom, Irvine, CA*

This talk focuses on the ADC/DACs with sampling rates higher than 10GS/s for DSP-based wireline transceivers. Data converters discussed include ones for PAM4 links inside the data centers with data rate up to 100Gb/s and phase-aligned quad-ADC/DACs in optical coherent DSP chips, which can support data rates well beyond 100Gb/s for long-haul, metro and data center interconnect (DCI) applications. Various mismatch calibration approaches and speed enhancement techniques are also discussed.

Jun Cao received B.S. degree in Physics from Peking University, Beijing, China, MSEE from the University of Michigan, Ann Arbor and Ph.D. degree in electrical engineering from UC Irvine. He joined NewPort Communications in 1999, which was acquired by Broadcom in 2000, where he is now a director and distinguished engineer, responsible for analog/RF and mixed-signal circuits for wireline applications. He has published more than 30 circuit papers, most of them in ISSCC and JSSC. He also holds over 60 US and international patents, in the area of high speed transceivers and data converters. Dr. Cao has served as a member of technical program committee in CICC since 2014 and was also a guest lecturer at UC Irvine.

EE1: Student Research Preview (SRP)

The Student Research Preview (SRP) will highlight selected student research projects in progress. The SRP consists of 20 ninety-second presentations followed by a Poster Session, by graduate students from around the world, which have been selected on the basis of a short submission concerning their on-going research. Selection is based on the technical quality and innovation of the work. This year, the SRP will be presented in three theme sections: Digital and Machine Learning Circuits and Systems; Data Converters and Clocking; Analog and Wireless Circuits.

The Student Research Preview will include the talk "Machine Learning: The Next Big Opportunity for Chip Designers" by Professor Marian Verhelst, KU Leuven. The SRP begins at 7:30 pm on Sunday, Febuary 17th. SRP is open to all ISSCC registrants.

Denis Daly

Makoto Ikeda

Co-Chair:	Denis Daly	Omni Design Technologies, MA
Co-Chair:	Makoto Ikeda	University of Tokyo, Japan
Secretary:	Jerald Yoo	National University of Singapore, Singapore
Advisor:	Jan Van der Spiegel	University of Pennsylvania, PA
Advisor:	Anantha Chandrakasan	Massachusetts Institute of Technology, MA
Media/Publications:	Laura Fujino	University of Toronto, Canada
A/V:	Trudy Stetzler	Houston, TX

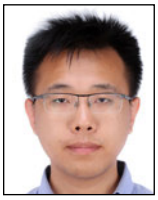
Committee Members

Jason Anderson	University of Toronto, Canada
Masoud Babaie	Delft University of Technology, Netherlands
Andrea Baschirotto	University of Milan-Bicocca, Italy
Ben Calhoun	University of Virginia, VA
Hayun Chung	Korea University, Korea
Denis Daly	Omni Design Technologies, MA
Shidhartha Das	ARM, United Kingdom
Andreas Demosthenous	University College London, United Kingdom
Makoto Ikeda	University of Tokyo, Japan
Seulki Lee	IMEC-NL, Netherlands
Yoonmyung Lee	SungKyunKwan University, Korea
Salvatore Levantino	Politecnico di Milano, Italy
Qiang Li	University of Electronic Science & Tech., China
Shih-Chii Liu	University of Zurich / ETH Zurich, Switzerland
Shahriar Mirabbasi	University of British Columbia, Canada
Tinoosh Mohsenin	University of Maryland, MD
Cormac O'Connell	TSMC, Canada
Mondira Pant	Intel, MA
Shanthi Pavan	Indian Institute of Technology, India
Jae-sun Seo	Arizona State University, AZ
Mingoo Seok	Columbia University, NY
Farhana Sheikh	Intel, OR
Bing Sheu	Chang Gung University, Taiwan
Atsushi Shirane	Tokyo Institute of Technology, Japan
GuoXing Wang	Shanghai Jiao Tong University, China
Jeffrey Weldon	University of Hawaii, HI
Chia-Hsiang Yang	National Taiwan University, Taiwan
Rabia Tugce Yazicigil	Boston University, MA
Jerald Yoo	National University of Singapore, Singapore
Samira Zaliasl	Ferric, New York, NY
Milin Zhang	Tsinghua University, China

Session 1: Digital and Machine Learning Circuits and Systems

Session Co-Chair: ***Chia-Hsiang Yang***
National Taiwan University, Taiwan

Session Co-Chair: ***Tinoosh Mohsenin***
University of Maryland, MD



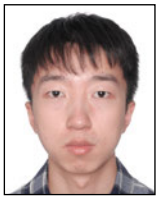
EE1.1
Zhe Yuan
Tsinghua University

EE1.2
Deepak Kadetotad
Arizona State University

EE1.3
Hyunjoon Kim
Nanyang Technological University

EE1.4
Muya Chang
Georgia Institute of Technology

EE1.5
Van Loi Le
Nanyang Technological University



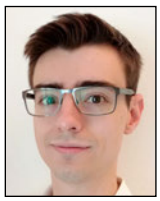
EE1.6
Jiaming Xu
Southeast University

EE1.7
Aminah Hina
Lahore University of Management Sciences

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Session 2: Data Converters and Clocking

Session Co-Chair: ***Samira Zaliasl***
Ferric, New York, NY



EE2.1
Alessio Santiccioli
Politecnico di Milano

EE2.2
Kwanseo Park
Seoul National University

EE2.3
Chaerin Hong
Ewha Womans University

Session Co-Chair: ***Shanthi Pavan***
Indian Institute of Technology, India

EE2.4
Mingqiang Guo
University of Macau

EE2.5
Baibhab Chatterjee
Purdue University

EE2.6
Hamidreza Maghami
Oregon State University

Session 3: Analog and Wireless Circuits

Session Co-Chair:
Rabia Tugce Yazicigil
Boston University, MA

EE3.1
Yun Wang
Tokyo Institute of Technology

EE3.2
Tsung-Ching Tsai
National Taiwan University

EE3.3
Hesam Sadeghi Gougheri
Pennsylvania State University

Session Co-Chair:
Milin Zhang
Tsinghua University, China

EE3.4
Muhammad Bilawal Khan
SungKyunKwan University

EE3.5
Beomsoo Park
University of Florida

EE3.6
Bangan Liu
Tokyo Institute of Technology

EE3.7
Chao Fan
University of Macau

Poster Session

Co-Chair: Yoonmyung Lee
SungKyunKwan University, Korea

Co-Chair: Mondira Pant
Intel, MA

EE2: How to Save Lives with Circuits

Chair: **Kathy Wilcox,** AMD, Boxborough, MA

Organizers: **Edith Beigne,** Facebook, Menlo Park, CA
Dina El-Damak, University of Southern California, Los Angeles, CA
Azita Emami, California Institute of Technology, Pasadena, CA
Ulkuhan Guler, Worcester Polytechnic Institute, Worcester, MA
Rikky Muller, University of California, Berkeley, Berkeley, CA
Mandy Pant, Intel, Hudson, MA
Negar Reiskarimian, Columbia University, New York, NY
Abira Sengupta, IEEE SSCS, Piscataway, NJ
Farhana Sheikh, Intel, Hillsboro, OR
Trudy Stetzler, Halliburton, Houston, TX
Vivienne Sze, Massachusetts Institute of Technology, Cambridge, MA
Ingrid Verbauwhede, KU Leuven, Leuven, Belgium
Alice Wang, PsiKick, Santa Clara, CA
Rabia Tugce Yazicigil, Boston University, Boston, MA

The workshop highlights circuits and their impact on healthcare-related industries. The goal of the panel is to provide perspectives from system architects, security experts and circuit designers on where we should be heading with the large amount of data that is being generated from more-advanced tests and increased monitoring of our current health status.

Distinguished Speakers

6:00 - 6:30 PM

Catalyzing Growth and Innovation

Sue Siegel, *GE Chief Innovation Officer and CEO, GE Business Innovations*

6:30 - 7:00 PM

Better Circuits for a Better World

Jen Lloyd, *Vice President, Healthcare and Consumer Systems at Analog Devices, San Jose, CA*

From healthcare products for an aging population to reliable communications for first responders, circuits are at the heart of products that address pervasive challenges in our world. Circuit innovation can make our lives better, but to truly tap into the power of circuits, we need to make circuit-based technology accessible and easy to implement in designs. Doing so will empower inventors and entrepreneurs to explore new use cases and leverage technology across domain areas, extending and amplifying the impact of circuit-level innovation. The talk will include some career insights related to this topic.

Invited Talks

7:00 - 7:20 PM

Unravelling the Brain with High-Density CMOS Neural Probes

Carolina Mora Lopez, *Team Leader Circuits for Neural Interfaces at imec, Belgium*

The human brain is considered the most complex biological structure known, with tens of billions of neurons controlling our thinking, feelings and behaviour. To unravel this complexity, scientists need to measure as many individual neurons as possible, all over the brain and simultaneously. By combining IC design and advanced micro-fabrication, it is possible to develop high-density neural probes that can record the activity of hundreds of individual neurons from multiple brain regions, in real time. This talk provides an overview of such devices and their impact on neuroscience research.

7:20 - 7:40 PM

Neurotechnology: where engineering meets neuroscience

Hyunjoo J. Lee, *Assistant Professor Electrical Engineering, KAIST*

"The science of today is the technology of tomorrow." - Edward Teller. Interestingly, past decades of rapid developments in technology have shown quite the opposite; the technology of today is now paving new paths for scientific breakthroughs. Neuroscience is a great example where neurotechnology is boosting the advances in the fundamental understanding of our brain. This talk will discuss exciting new tools that are changing the way we are studying our brain. Specifically, the talk will cover the use of multifunctional neural probes and non-invasive brain neuromodulation in neuroscience, where we are not just passively monitoring the brain, but stimulating the brain for interactive investigation.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Panel: What Can Circuit Designers do to Bolster Security in AI-driven Healthcare

7:45 - 9:00 PM

Moderator: ***Ingrid Verbauwhede,*** *KU Leuven, Leuven, Belgium*

Security and privacy problems with medical devices and IOT devices in general are in the news on an almost daily basis. One example from 2017 stated: "FDA issues recall of 465,000 pacemakers to patch security holes." Once medical data is obtained reliably and securely, it is stored on remote servers and in remote databases where there are risks of leaks and data breaches of private medical records. However, it is difficult to put the genie back into the bottle! We have asked our distinguished panel to discuss how circuit designers can contribute to bolster our trust in medical devices and in electronic healthcare systems that manage private medical records. We also encourage the audience to propose attacks and countermeasures.

Distinguished Speaker & Panelist

7:45 -8:15 PM

Perspectives on Machine Learning and Cryptography by Turing Award Winner Shafi Goldwasser

Shafi Goldwasser, *Professor Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA*

Panelist's Statements

Alison Burdett, *Sensium Healthcare, Abingdon, United Kingdom*

Medical device development is regulated following principles of safety engineering – assessing the likelihood that a hazardous event (non-malicious fault) will occur and the severity of the resulting incident or harm. As medical devices embrace connectivity through open standards, the threat of malicious attacks becomes of equal importance during the design process. Although considering cybersecurity resilience is important in all areas of circuit design, in medical devices the stakes are high – in some instances, quite literally a matter of life and death. Real-world examples of security weaknesses in connected medical devices will be described to stimulate discussion on how circuit designers could offer robust solutions to protect patient safety.

Rikky Muller, *University of California at Berkeley, Berkeley, CA*

Medical devices not only monitor and help inform patient care, and in many cases provide treatment (e.g. drug delivery, neurostimulation). Intelligent, closed-loop devices can make treatment decisions autonomously and in real time. Giving devices the ability to process data and perform this decision-making will alleviate the need to transmit data off of the device, thereby improving privacy. To enable this vision, energy efficient algorithms and circuits must be developed, particularly in the case of implantable devices, which have stringent power and volume constraints.

Sugako Otani, *Renesas Electronics Corporation, Tokyo, Japan*

Security is becoming a real concern with the rapidly changing IoT environment. Recent security attacks, like Meltdown and Spectre, have awaken the public to the need for security-aware microarchitecture. On the other hand, security is now becoming a core part of the regulatory process in healthcare. Standards landscape (UL 2900, etc.) is starting to mature and align with regulatory & business goals. While security is very broad, to avoid the Internet of "Insecure" Medical Things, circuit designers will be in a mutual learning process with security community.

Vivienne Sze, *Massachusetts Institute of Technology, Cambridge, MA*

Artificial intelligence (AI) has the promise of making a significant impact on health care by improving diagnostics and tailoring treatment for individual patients. However, AI relies on significant data collection, which opens up privacy vulnerabilities on very personal data. One way to reduce the distribution and thus the exposure of private data is to perform the analysis of the data locally near the sensors where it is collected rather than in the cloud. Circuit designers play an important role in enabling this processing under the stringent energy, area, and cost requirements that exist at the edge.

Wenyuan Xu, *Zhejiang University, Hangzhou, China*

Modularity and abstraction have been widely used to reduce the complexity of software and hardware design. Naturally, security experts either focus on the hardware or software aspects as the result of abstraction. However, when it comes to security, existing abstraction that considers little security could become the barrier to solve the system-wise security problems. It is time to examine the cross-hardware-software security vulnerabilities and re-design the abstraction such that we can address the overall security better.

EE3: Making a Career Choice

Organizing Committee: **Emre Ayranci,** *pSemi, CA*
Anantha Chandrakasan *Massachusetts Institute of Technology, MA*
Denis Daly, *Omni Design Technologies, MA*
Jan Van der Spiegel, *University of Pennsylvania, PA*

Moderator: **Negar Reiskarimian,** *Columbia University, NY*

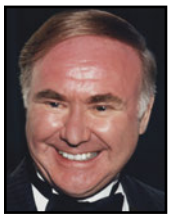
Speakers: **Edith Beigné,** *Facebook, CA*
Robert Brodersen, *University of California, Berkeley, CA*
Barrie Gilbert, *Analog Devices, OR*
Nicky Lu, *Etron Technology, Taiwan*
Marian Verhelst, *KU Leuven, Belgium*

This evening interactive event will include several distinguished panelists representing a broad variety of career choices in the areas of start-ups, industry, research, and academia available to graduates in electrical and computer engineering. Following short introductory remarks by each panelist, this forum will open for audience interaction in which the audience is invited to express a broad range of questions to the panelists.

Bios

Edith Beigné
Facebook, CA
Edith Beigné joined Facebook in November 2018 where she is currently Silicon Research Manager working on integrated circuits for AR/VR applications. From 1998 to 2018 she worked at CEA-LETI, Grenoble, France, where she was the Research Director of the design division. She visited Stanford University in 2018. Her research focuses on low power digital and mixed-signal integrated circuits and systems, adaptive and intelligent circuits with a specific focus on emerging technologies applications like 3D integration or resistive memories.

Robert Brodersen
University of California, Berkeley, CA
Robert Brodersen is Professor Emeritus in EECS and a founder of the Berkeley Wireless Research Center (BWRC) at the University of California, Berkeley. He joined the faculty of the Electrical Engineering and Computer Science Department at Berkeley in 1976, where his research focused on low power design and wireless communications. Professor Brodersen has also co-founded several companies including SiBEAM and BEEcube.



Barrie Gilbert
Analog Devices, OR
Barrie Gilbert (Life Fellow, IEEE) arrived on Earth a day after Sir James Barrie died, the same year the magical cavity magnetron was invented. Almost a decade later, the Battle of Britain was narrowly won because of this milepost in electronics. Though the war was over, his father had been killed during a bombing raid. Not long after, this young electronics tinkerer was inundated with a cornucopia of superb - yet very cheap - surplus equipment. Even before the age of ten, he'd made his first short wave radio; by eleven had built his first oscilloscope and a TV set from these precious war-spared parts. He has been tinkering with enigmatic electrons and mischievous transistors ever since, resulting in over 100 patents, some 50 papers and contributions to 6 books. Along the way, he has learned some important lessons, and taught them to young audiences, about living the life of an inventor and engineer.

Nicky Lu
Etron Technology, Taiwan
As a researcher, design architect, entrepreneur and chief executive, Dr. Nicky Lu has dedicated his career to the worldwide IC design and semiconductor industry. He is Chairman, CEO and Founder of Etron Technology and co-founded several other high-tech companies including Ardentec and Global Unichip Corp. Dr. Lu received 35 US Patents, IBM Corporate Award and IEEE Solid-State Circuits Field Award, and is an IEEE Fellow and a member of NAE.

Negar Reiskarimian
Columbia University, NY
Negar Reiskarimian is a PhD student at Columbia University. She will join Massachusetts Institute of Technology's department of Electrical Engineering and Computer Science (EECS) as an Assistant Professor in July 2019. Her research interests span integrated circuits and systems, applied electromagnetics and nanophotonics, with a focus on theory, design and experimental validation of analog, radio-frequency, millimeter-wave and photonic integrated circuits, metamaterials and systems.

Marian Verhelst
KU Leuven, Belgium
Marian Verhelst is a professor in the MICAS group of KU Leuven since 2012. Her research focuses on processing architectures for ubiquitous sensing and machine learning, processor design, and self-adaptive sensor interfaces for embedded devices. Before that, she received a PhD from KU Leuven, was a visiting scholar at the Berkeley Wireless Research Center (BWRC) of UC Berkeley and worked for 3 years in the research labs of Intel Corporation in Portland, OR, USA.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

EE4: Industry Showcase

Chair: ***Eugenio Cantatore,*** *Eindhoven University of Technology, Eindhoven, The Netherlands*

Organizers: ***Alison Burdett,*** *Sensium Healthcare, Abingdom, United Kingdom*
Kush Gulati, *Omni Design Technologies, Milpitas, CA*
Un-Ku Moon, *Oregon State University Corvallis, OR*

Committee

Ahmed Ali, *Analog Devices, Greensboro, NC*
Min Chen, *Analog Devices, Milpitas, CA*
Muhammad Khellah, *Intel, Hillsboro, OR*
Yan Li, *Western Digital, Milpitas, CA*
Yao-Hong Liu, *imec, Eindhoven, The Netherlands*
Mike Perrott, *TDK-Invensense, Boston, MA*
Bruce Rae, *ST Microelectronics, Edinburgh, United Kingdom*
Jiayoon Ru, *Broadcom, Irvine, CA*
Takayuki Shibasaki, *Fujitsu Laboratories, Kawasaki, Japan*
Youngmin Shin, *Samsung, Hwansung, Korea*

The event will begin with a Showcase Keynote "Accelerated Platforms: The Future of Computing" by Bill Dally, Chief Scientist and Senior Vice President of Research, NVIDIA, Santa Clara, CA and Professor (Research) of EE and CS, Stanford University, Stanford, CA.

Following the keynote, the Industry Showcase will feature short presentations as well as interactive demonstrations from each of the Industry Showcase participants, and represent an exciting introduction to the next generation of applications and products enabled by advances in solid-state integrated circuits.

Bill Dally

Bill is Chief Scientist and Senior Vice President of Research at NVIDIA Corporation and a Professor (Research) and former chair of Computer Science at Stanford University. Bill is currently working on developing hardware and software to accelerate demanding applications including machine learning, bioinformatics, and logical inference. He has a history of designing innovative and efficient experimental computing systems. While at Bell Labs Bill contributed to the BELLMAC32 microprocessor and designed the MARS hardware accelerator. At Caltech he designed the MOSSIM Simulation Engine and the Torus Routing Chip which pioneered wormhole routing and virtual-channel flow control. At the Massachusetts Institute of Technology his group built the J-Machine and the M-Machine, experimental parallel computer systems that pioneered the separation of mechanisms from programming models and demonstrated very low overhead synchronization and communication mechanisms. At Stanford University his group has developed the Imagine processor, which introduced the concepts of stream processing and partitioned register organizations, the Merrimac supercomputer, which led to GPU computing, and the ELM low-power processor. Bill is a Member of the National Academy of Engineering, a Fellow of the IEEE, a Fellow of the ACM, and a Fellow of the American Academy of Arts and Sciences. He has received the ACM Eckert-Mauchly Award, the IEEE Seymour Cray Award, the ACM Maurice Wilkes award, the IEEE-CS Charles Babbage Award, and the IPSJ FUNAI Achievement Award. He currently leads projects on computer architecture, network architecture, circuit design, and programming systems. He has published over 250 papers in these areas, holds over 160 issued patents, and is an author of the textbooks, Digital Design: A Systems Approach, Digital Systems Engineering, and Principles and Practices of Interconnection Networks.

Abstracts

180° Real-time Immersive, Intelligent Vision Systems

Aurangzeb Khan, *Altia Systems, Cupertino, CA*

180° real-time immersive intelligent vision systems, producing 2D Panoramic-4K and 3D 4K video, with HDR and people detection information, will be demonstrated. System performance is enabled with a newly-developed real-time dynamic stitching and video optimization architecture implemented in on-board FPGAs (300M – 2.3B pixels/s). In one system, three 3 megapixel cameras are synchronized, dynamically stitched and optimized in real-time at 30 FPS to produce 3840 x 1080 pixels per frame. In a second system, video outputs from six 13 megapixel cameras are similarly processed in real-time to produce a stacked 4K video frame with 3840 x 2160 pixels per frame, at 60 FPS, and output via 12G SDI.

Direct Time-of-Flight Module in CMOS 55nm HV for Mobile Applications

Robert Kappel, *ams AG, Premstätten, Austria*

This contribution presents a D-ToF micro-system module featuring a 2x2-zone SPAD-based SoC sensor manufactured in CMOS 55nm HV, pulsed light emitter VCSEL and its driver integrated into a compact single package developed for mobile proximity/presence detection applications. The proposed interactive demonstration will show the micro-module in action behind typical smartphone cover glass in the range detection of single and multiple targets at 50fps, between 200-2000 mm, and different levels of ambient light. Raw data processing and target distance extraction are implemented on-chip exploiting an embedded histogramming memory and a CPU.

TeraPHY: A high-density electronic-photonic chiplet for co-packaged optical I/O

Chen Sun, *Ayar Labs, Emeryville, CA*

Despite chip bandwidth requirements growing with no end in sight, electrical I/O races towards an imminent scaling wall beyond the 112Gbps SerDes generation. Usage of optics for chip-to-chip I/O is a promising alternative to electrical SerDes. However, contemporary optics technologies, such as those used for pluggable optical transceivers, do not fit within the density and power requirements of a chip I/O technology. In this work, Ayar Labs, Inc. proposes to host an industry demonstration of its TeraPHY, an electronic- photonic chip with microring resonator based optical SerDes to enable optical I/O for co-packaged SoC applications.

Impedance-based Sensor Chip for Non-invasive Blood Presssure Monitoring

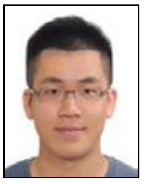
Taehwan Roh, *Healthrian, Daejeon, Korea*

A common non-invasive blood pressure (NIBP) measurement device is based on an oscillometric method with cuff, which is not continuous and uncomfortable. Wrist bands using photoplethysmography (PPG) have failed because of inaccuracy. The tonometric method for NIBP uses a tiny pressure sensor based on MEMS technology. Since it requires applying constant pressure on the device, it is not easy for a long time. Healthrian®'s sensor chip exploits an impedance method that supports both comfortable and continuous monitoring of NIBP. The chip, embedded in our band-type system, was verified and compared with other NIBP methods. With the Healthrian®'s monitoring software, clean signals can be continuously monitored.

Fast channel sounding and beam optimization for 5G using software defined phased array radios

Bodhisatwa Sadhu, *IBM T. J. Watson Research Center, Yorktown Heights, NY*

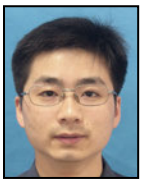
In this mmWave 5G demonstration, we propose to install two software-defined phased array radios featuring 28-GHz 64-element dual-polarized phased arrays creating a 28-GHz link in an indoor environment. We will demonstrate precise 3-D channel map data capture at 28 GHz in the presence of interferers and reflectors. The channel sounding experiments will utilize ultra-fast electronic beam steering and live EVM measurements to capture a new scene in <90 seconds. In addition, we will also demonstrate fast automatic beam finding algorithms that optimize link EVM.



Future Driving with Better Safety Inside-Out

ChiaYu Lin, *Mediatek, HsinChu, Taiwan*

Advanced driver assistance system (ADAS) has come into real life. The safety features cover not only the outside of vehicles but also drivers and passengers inside vehicles. This showcase demonstrates a brand new parking assistance system (PAS) and a smart driver monitoring system (DMS) built into an infotainment system running multiple operating systems. The PAS features the smallest complete 79 GHz radar sensor in the world, and the DMS embeds deep learning capability running on a hypervisor.



The world's first deployed URLLC wireless communication system and SoCs

Li Yang, *NUFRONT, Beijing, China*

The EUHT (Enhanced Ultra High Throughput) Wireless Communication System, which was developed in 2007 by NUFRONT, is the world's first URLLC (Ultra Reliable & Low Latency Communication) wireless system, and it is based on dedicated SoCs. EUHT has been announced as standard and massively deployed in high speed rail, Industry Internet and V2X applications in China. Taking advantage from the joint optimization of System, RF, analog and baseband design, the EUHT SoCs designed by NUFRONT can achieve reliability of 99.999% and end-to-end latency of 0.8 millisecond at a speed of 300km/h. The URLLC performance will be shown in this demo based on the 3rd generation commercial EUHT SoCs.

Deep Learning Inference on NVIDIA GPUs and Accelerators

NVIDIA, Santa Clara, CA, Presenter: TBD

In this demo we will show deep learning inference for both embedded devices and data centers. For embedded devices we will show the NVIDIA Xavier chip simultaneously processing 36 HD video streams using a combination of the NVDLA open-source deep learning accelerator and its Volta CUDA cores. For data center inference we will demonstrate a server with 32 Turing T4 GPUs classifying 50,000 images of flowers each second.

Native camera imaging on lidar and deep learning enablement

Angus Pacala, *Ouster, San Francisco, CA*

The emergence of lidar as a critical sensing modality for real-time robotics has resulted in a need for computer vision scientists to develop new algorithms to segment, track, and classify point clouds. Progress has been limited by the inability to apply decades of methodologies from camera-based vision due to the novel data formats and structures that conventional lidar output. Our breakthrough in lidar hardware using an all-ASIC approach and hardware triggered firing enable camera-like imagery of both ambient and signal data in a rectilinear camera-like grid. This demo will feature applications of GPU-enabled camera deep learning algorithms on lidar for the first time.

PsiKick's Batteryless Sensor Network Platform for the Industrial IoT

Nathan Roberts, *PsiKick, Santa Clara, CA*

PsiKick, Inc., is a venture-backed startup pioneering wireless, batteryless IoT systems for industrial environments. Leveraging silicon technology developed at the University of Virginia and the University of Michigan, PsiKick approaches self-powered systems from multiple angles, combining energy harvesting nodes with differentiated wireless communication and ultra-low power radios. In this industrial showcase PsiKick will demonstrate its first ultra low power sensing platform which monitors steam trap health, providing real time status alerts and will focus on the underlying silicon technology that enables this system to operate without batteries.

A Fully Integrated 10nm Multi-Mode 5G Modem Chipset Compatible with 3GPP NR Standards

Hui Won Je, *Samsung Electronics, Hwaseong, Korea*

Samsung S.LSI exhibits here the world first multi-mode 5G modem chipset. This state-of-the-art technology product, named Exynos Modem 5100, is compliant with the latest 3GPP 5G New Radio (NR) standards supporting both sub-6GHz and mmWave transmissions. It also provides fully integrated solutions of all the global cellular Radio Access Technologies. At the demo, Samsung S.LSI will disclose the device/chipset, detailed figures, and video clips of the Exynos Modem 5100. Devices equipped with this modem are planned to appear on the Korean market at the end of 2018 commercializing NR sub-6GHz services at first.

Motion Artifact Free Dynamic Vision Sensor for Machine Vision

Yunjae Suh, *Samsung Electronics, Hwaseong, Korea*

Samsung S.LSI will showcase here our new Dynamic Vision Sensor (DVS). Succeeding to our VGA-resolution DVS with a 9-um pixel introduced at ISSCC 2017, we have developed a new DVS chip and system that tolerates motion artifacts for the next generation machine vision applications. In this showcase, we will demonstrate the key improvements implemented in the silicon chip and in a practical system. In addition, we will prepare multiple live demos based on the DVS such as a visual odometry (VO) system that the participants can enjoy hands-on.

Unleashed Acceleration: AI on RISC-V

Rajesh Varadharajan, *SiFive, San Mateo, CA*

SiFive will be demonstrating a flexible AI application. The YOLO (You Only Look Once) image recognition application was constructed with the open-source NVDLA framework, demonstrating the flexibility of the hardware and software stack. The demo consists of the NVDLA accelerator running on an FPGA connected via ChipLink to SiFive's HiFive Unleashed board powered by the Freedom U540.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

EE5: "Moving to The Dark Side"

Organizers: **Jan Westra,** *Broadcom, Bunnik, The Netherlands*
Matt Straayer, *Maxim, North Chelmsford, MA*

Many practicing engineers view management and business as 'The Dark Side', and executive decisions often baffle them. Can engineers do a good job on 'The Dark Side' in management and business? Is it possible to return to engineering from 'The Dark Side'?

The relation between business and management on one side and engineering on the other side has always been complex. On one hand, innovative engineers often don't understand the limitations imposed on them by more financially driven business managers. On the other hand, business needs solid engineering, within their budgetary boundaries, in order to be successful. Is it possible to transform from a good engineer to a good business manager? Can business managers return to engineering and be successful engineers after having spent time on 'The Dark Side'? Can engineers and business managers peacefully coexist, and are they able to understand each other? Is that a situation we should work towards or, is a world where each side sees the other as a necessary evil perfectly balanced and sustainable?

In an entertaining and educational discussion between the panelists, the audience will get a closer look into the hearts and minds of Captains of Industry, sharing their views on the subject. How do they view the world after moving to 'The Dark Side' and have they truly moved to 'The Dark Side'?

Join this panel discussion with active audience interaction, take a seat in the tribunal and make your vote count!

Abstracts

Moderator: Matt Straayer, *Maxim, North Chelmsford, MA*
Matt Straayer is Managing Director of IC Design in the Advanced Research & Development group at Maxim. He is responsible for corporate IP and technology development in next-generation, adjacent, and nascent markets.

Matt received the B.S. and M.S. degrees from the University of Michigan in 2000 and 2001, respectively, and the Ph.D. degree from the Massachusetts Institute of Technology in 2008. From 2001 to 2003 he worked at Integrated Sensing Systems designing custom CMOS ICs for capacitive and resonant MEMS sensors, and from 2003-2008 he designed mixed-signal circuits for high-speed analog and RF applications at MIT Lincoln Laboratory. In 2008 he co-founded Cambridge Analog Technologies (CAT) where he led the commercialization of zero-crossing circuits for high performance analog-to-digital converters, and has been with Maxim since CAT was acquired in 2011. Dr. Straayer is the author of numerous publications and patents, is an IEEE SSCS Distinguished Lecturer, and serves on the ISSCC data-converter technical program committee.

Anthony Collins, *Xilinx, Dublin, Ireland*
Engineers can do a "good job" championing a business opportunity and demonstrating the leadership required to make it happen. I say leadership and not management as I want to differentiate between the role of manager and leader. Management responsibility is not a requirement for leadership, however leadership is an essential aspect of business success. I believe most business opportunities are about analyzing and solving problems and that's a skill engineers develop and excel at. Any good business idea should start by asking "what problem are we solving?" and then continuously revisiting and questioning any assumptions. An analytical and data-driven approach can also help to overcome many of the biases and flaws in our decision making. However, the business perspective is also important in assessing issues like strategic alignment and the benefits for the company. Moving to "The Dark Side", even on a temporary assignment, can facilitate better collaboration between the engineering and business teams. And sometimes we may not go back.

Anthony Collins holds a Hons. Dip. EE from the Dublin Institute of Technology and a B.Sc.Eng. degree from Trinity College Dublin ('94). He received an MBA from the Smurfit Post Graduate School of Business, University College Dublin, in 2008. He is the Marketing Director for Radio at Xilinx Inc. where he focuses on the company's RF Data Converter product portfolio. He has over 20 years' experience in product definition, applications, and marketing roles. Anthony also holds several US patents in the areas of mixed signal systems and programmable logic.

Kavé Kianush, *Catena, Delft, The Netherlands*
There is no denying that engineering and business communities are often suspicious of each other's practices. Over the last 2 decades, the gap in understanding has widened. This, I suspect, is mostly due to commoditization of our industry. Engineers stepping into business can reverse the trend. They would end up in the "dark side" only if they don't carry their own guiding torch with them. Their training based on universal physical laws and "common sense" is an invaluable tool in analyzing business situations and making informed and appropriate decisions. Having said that, it is much easier to execute on the "right decisions" in a privately owned company. Quarterly reporting to the Stock Exchange often leads to decisions that don't comply with the laws of nature! Stepping back from The Dark Side into engineering has its challenges. Although physical laws don't change, the tools of the trade are evolving very rapidly, making it difficult to catch up. Psychologically, being used to making your own decisions, it can be quite difficult to blindly execute on the orders of others from The Dark Side.

Kavé Kianush has more than 30 years of experience in the semiconductor industry. His main areas of interest are Automotive and Wireless Connectivity IC design, where he holds several patents and paper publications. After graduating from London University, he started his career at Philips Research Labs in 1982 and moved on to Philips Semiconductors in 1990. He joined Catena in the year 2000, where he is the Vice president and Chief Technology Officer.

Curtis Ling, *MaxLinear, Carlsbad, CA*
Good engineering is done by teams of people working together coherently irrespective of their titles. The blanket categories frequently applied to these people (the term "management" being one example) are rarely helpful, though they are broadly useful for Dark Side activities such as sowing division, casting blame, and stoking an unfounded sense of superiority. At my company we see that management and engineering are neither mutually exclusive, nor orthogonal activities; at each stage in their careers, engineers are developing their (in)competence in each, unwittingly or otherwise. On this panel I will represent my own perspective of MaxLinear's evolving philosophy and praxis of these as we grew from a group of founders to our present form.

Curtis Ling received a B.S. from Caltech, and M.S. & Ph.D. from the University of Michigan, Ann Arbor, in Electrical Engineering. He joined Hong Kong University of Science and Technology in 1993 as an assistant professor doing research in circuits and systems for wireless communications. Fortunately he demonstrated a lack of aptitude in negotiating academia and subsequently joined a fabless startup, Silicon Wave to design Bluetooth and cable modem chips from 1999 to 2003. Curtis and several colleagues quit in 2003 to start MaxLinear, a fabless company which designs communication systems.
As cofounder of MaxLinear, Curtis has served in several capacities, switching roles whenever the Peter Principle caught up with him. These include human resources, finance, accounting & payroll, IC design, RF systems engineering, applications engineering, marketing & business development, communication systems engineering, and Chief Technology Officer.

Jen Lloyd, *Analog Devices, San Jose, CA*
We all know that engineers love to hate management. But we need to fight the cliché! If engineers, managers and business leaders work together, truly impactful innovation can happen. Successful engineers have either spent time on 'The Dark Side', or at least embrace it; they make business constraints part of their problem statement. Similarly, the best business leaders on 'The Dark Side' are either engineers themselves, or at least know how to provide 'The Light Side', i.e. an environment that spurs great innovation to fuel business success.

Jen Lloyd received S.B., S.M., and Ph.D. degrees in EECS from MIT in 1989, 1992 and 1997, respectively. She started her career as an analog designer at Analog Devices in the High Speed Converter group, and has contributed to various ADC, DAC and SerDes products for which she holds 9 US patents, and several papers and publications. She served on the TPC for both the IEEE Custom Integrated Circuits Conference and the IEEE Symposium on VLSI Circuits. For several years, she led ADI's core technology business in amplifiers and precision converters, along with the Instrumentation Market Group. She is currently Vice President, Healthcare and Consumer Systems, where she leads global engineering and business teams to build innovative solutions for those markets.

Tyson Tuttle, *Silicon Labs, Austin, TX*
We all want to work on interesting problems that matter, and there is nothing worse than putting your heart and soul into a project that goes nowhere. Building a successful technology business requires both world-class engineering and solid management, where technical innovation is put into the context of a long-term vision and a winning strategy. When moving to 'The Dark Side', it is essential engineers understand markets and the competition, optimize financial targets and constraints, think both from the top-down as well as the bottoms-up, appreciate the importance of people and culture, and be able to effectively communicate with all stakeholders. Having people move back and forth across these disciplines builds transparency and understanding as well as producing more successful business outcomes.

Tyson Tuttle serves as President and Chief Executive Officer of Silicon Labs, a leading provider of silicon, software and solutions for a smarter, more connected world. Tyson joined Silicon Labs in 1997 as a design engineer working on the company's first breakthrough product, which achieved market share leadership in PC modems and enabled the company to go public in 2000. Tyson subsequently led the marketing effort behind Silicon Labs' first CMOS RF transceiver products for mobile handsets, spearheaded the development and market success of the company's radio and TV tuner ICs, and served as Chief Technology Officer and Chief Operating Officer before being named CEO in 2012. Tyson has more than 25 years of semiconductor experience with industry leaders like Broadcom and Cirrus Logic / Crystal Semiconductor, holds more than 70 patents in RF and mixed-signal IC design, served on the ISSCC technical program committees from 1996 to 2002, and has presented several papers and one tutorial at the ISSCC conference. Tyson received the B.S. degree in Electrical Engineering in 1989 from Johns Hopkins University and the M.S. degree in Electrical Engineering in 1992 from UCLA. He currently serves on the board of the Global Semiconductor Alliance, the advisory board for the Johns Hopkins Whiting School of Engineering, and is a respected leader in the Austin community, serving on the board for KLRU-TV, home of PBS and Austin City Limits.

Patrick Yue, *Hong Kong University of Science and Technology (HKUST), Hong Kong*
Only the bright engineers can turn to "The Dark Side" and excel at it. Advancing to management positions should be considered a move to "the Bright Side" especially in the chip business. I have never met a good IC design team manager who was not an outstanding chip designer. While effective design engineering managers need to be exceptional at team management and resource allocation to consistently deliver silicon success before deadline, they may not necessarily be good at business development (BD). That takes a whole extra set of qualities like market demand trend, technology insight, and financial implications. Most of the iconic executives in our industry excelled at both engineering management and BD. Nevertheless, we must be mindful that management and BD is not the right cup of tea for every brilliant engineer. It is not just about quality, but also personality. The successful companies need leadership that facilitates engineers and business managers to peacefully coexist and cultivate their understanding of the necessity of each other to sustain a successful chip business.

Patrick Yue is a Professor in the ECE Department at Hong Kong University of Science and Technology (HKUST). He is the founding Director of the HKUST-Qualcomm Joint Innovation and Research Lab and served as the Associate Provost for Knowledge Transfer. He co-founded Atheros in 1998 and was a professor at CMU and UC Santa Barbara between 2003 to 2010. Prof. Yue has contributed to more than 150 peer-reviewed technical papers and two book chapters, and holds 15 US patents in the field of RFIC design. He was a co-recipient of the 2003 International Solid-State Circuits Conference (ISSCC) Best Student Paper Award and the 2017 IEEE CAS Society Outstanding Young Author Award. He is the lead author of an all-time top-10 most cited paper in the IEEE Journal of Solid-State Circuits regarding on-chip inductors with patterned ground shields (Google citation index over 1500). He is an IEEE Fellow, an Editor for the Proceedings of IEEE, an Optical Society of America Senior Member, and the Membership Chair of the IEEE Solid-State Circuit Society AdCom.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

EE6: "How Can Hardware Designers Reclaim the Spotlight?"

Organizers: **Sudhakar Pamarti,** *UCLA, Los Angeles, CA*
Mike Shuo-Wei Chen, *USC, Los Angeles, CA*
Nagendra Krishnapura, *IIT Madras, Chennai, India*

Gone seem the heady days when hardware innovation drove entrepreneurship in communications, computing, and other domains. The bar for circuit innovation and design has been inexorably climbing higher, even as circuit components are becoming increasingly commoditized and seemingly losing relevance in the larger scheme of things.

Is this trend here to stay? Should we do something about it? What should we do? Are long hardware design cycles to blame? Can we learn something from software design practices? Or, are there new opportunities where hardware innovation can still play a central role? If so, how far out is the horizon? How do we prepare for these opportunities?

This evening panel assembles a galaxy of international experts from academia, government, and both the hardware and software industries to discuss these issues. The experts will diagnose the situation, offer insightful prognoses, and engage in a lively discussion with each other, and with the audience. Come join us to see if you agree with our experts and for an opportunity to enlighten us with your opinion.

Abstracts

Moderator: Mike Shuo-Wei Chen, *University of Southern California, Los Angeles, CA*
Mike Shuo-Wei Chen is an Associate Professor in Electrical Engineering Department at USC and holds the Colleen and Roberto Padovani Early Career Chair position. He received the B.S. degree from National Taiwan University, Taipei, Taiwan, in 1998 and the M.S. and Ph.D. degrees from the University of California, Berkeley, in 2002 and 2006, respectively, all in electrical engineering. As a graduate student researcher, he proposed and demonstrated the asynchronous SAR ADC architecture, which has been adopted today for low-power high-speed analog-to-digital conversion products in industry. At USC, he leads an analog mixed-signal circuit group, focusing on high-speed low-power data converters, bio-inspired/biomedical electronics, RF frequency synthesizers, power amplifiers, DSP-enabled analog circuits and systems. From 2006 to 2010, he worked on mixed-signal and RF circuits for various wireless communication products at Atheros Communications (now Qualcomm). He was the recipient of an NSF Faculty Early Career Development (CAREER) Award, and a DARPA Young Faculty Award (YFA) both in 2014.

Luca Benini, *ETH Zurich, Zurich, Switzerland, and Universita di Bologna, Bologna, Italy*
The slowdown of Moore's law and the relentless hunger for AI capabilities, from the datacenter to the IoT edge, is creating a new "renaissance of hardware design". We have today the best opportunity in decades to regain the center stage in research and in the job market, Strategic directions to ride the wave are: getting open-source hardware right, understanding and exploiting heterogeneous (more-than-Moore) integration and leaving options open on AI, i.e. avoiding the bandwagon effect on deep learning.

Luca Benini holds the chair of Digital Circuits and Systems at ETH Zurich and is a Full Professor at the Universita di Bologna. He received a Ph.D. from Stanford University. He has been a Visiting Professor at Stanford University, Imec, and EPFL. In 2009-2012, he served as Chief Architect in STmicroelectronics, France. Dr. Benini's research interests are in energy-efficient computing systems design, from embedded to high-performance. He is also active in the design of ultra-low-power VLSI circuits and smart-sensing microsystems. He has published more than 900 peer-reviewed papers and five books. He is, a Fellow of the IEEE, of the ACM and a member of the Academia Europaea. He is the recipient of the 2016 IEEE CAS Mac Van Valkenburg award.

Bill Dally, *NVIDIA, Santa Clara, CA*

With the end of Moore's Law and Dennard Scaling, domain-specific architecture remains the only viable path to continue scaling computing performance. In this era of domain-specific hardware, differentiation will come primarily from efficient hardware design and hardware designers, and the system level will play a key role. Domain specific hardware achieves its performance from massive parallelism, specialized data types and operations, fine-grained memory systems, and efficient interconnection networks. The latter two, memory and interconnect, are domain independent and will be provided by general-purpose platforms, such as GPUs. Hardware system designers will differentiate their products via algorithm-hardware co-design, and the design of special purpose units that provide the data-type and operation specialization – like the TensorCores and RTCores in a Turing GPU.

Bill Dally is Chief Scientist and Senior Vice President of Research at NVIDIA Corporation and a Professor (Research) and former Chair of Computer Science at Stanford University. Bill is currently working on developing hardware and software to accelerate demanding applications, including machine learning, bioinformatics, and logical inference. He has a history of designing innovative and efficient experimental computing systems. While at Bell Labs, Bill contributed to the BELLMAC32 microprocessor and designed the MARS hardware accelerator. At Caltech, he designed the MOSSIM Simulation Engine and the Torus Routing Chip, which pioneered wormhole routing and virtual-channel flow control. At the Massachusetts Institute of Technology, his group built the J-Machine and the M-Machine, experimental parallel computer systems that pioneered the separation of mechanisms from programming models and demonstrated very low overhead synchronization and communication mechanisms. At Stanford University, his group has developed the Imagine processor, which introduced the concepts of stream processing and partitioned register organizations, the Merrimac supercomputer, which led to GPU computing, and the ELM low-power processor. Bill is a Member of the National Academy of Engineering, a Fellow of the IEEE, a Fellow of the ACM, and a Fellow of the American Academy of Arts and Sciences. He has received the ACM Eckert-Mauchly Award, the IEEE Seymour Cray Award, the ACM Maurice Wilkes award, the IEEE-CS Charles Babbage Award, and the IPSJ FUNAI Achievement Award. He currently leads projects on computer architecture, network architecture, circuit design, and programming systems. He has published over 250 papers in these areas, holds over 160 issued patents, and is an author of the textbooks, Digital Design: A Systems Approach, Digital Systems Engineering, and Principles and Practices of Interconnection Networks.

Varada Gopalakrishnan, *Amazon Lab126, Sunnyvale, CA*

"Reclaim" necessarily implies something was lost in hardware design! Growth in personal compute and mobile compute over the last three decades could not have happened without hardware playing a central role. One of the side effects of this growth is the implicit assumption that increases in compute power will offset the loss of efficiency (power, performance, pick a performance index). New applications "at edge" will challenge this assumption. More compute does not necessarily imply better systems in these new applications. Form is just as important as function and different metrics such as compute/mW and compute/mg will drive new products and services. Application-aware circuit design, and co-design of energy-efficient software and hardware, are but a few examples of what lies ahead.

Varada Gopalakrishnan is a Senior Principal Engineer in Amazon Lab126. He received the B.E degree in Electronics and Communication Engineering from Anna University, Chennai, India in 1991 and the M.S in Electrical Engineering, specializing in Digital Signal Processing from the Indian Institute of Technology, Kanpur, India. He started his career in Motorola designing wireline communication systems. Later in Motorola, he designed and developed the physical layer of 2G, 3G wireless handsets. At Amazon, he has been wireless and product architect for all the devices designed and developed by Amazon, including Fire Tablets, Fire Phone, Alexa Devices and Fire TV.

Colin Lyden, *Analog Devices, Limerick, Ireland*

Like beauty, 'spotlight' and 'cool' lie in the eye of the beholder, but the need for ever better hardware will persist for a long time to come. There may be fewer rock star hardware designers these days, but they are essential for the orchestra. We need more and more hardware in order to instrument the real world, if only to allow software people treat it as part of their virtual world. The challenge here is that the real-world-interface hardware is too fragmented, and so hard to justify on the smallest geometry processes.

Colin Lyden is Director of Technology for precision products in Analog Devices. After receiving his Ph.D. in 1984, he joined Ireland's National Micorelectronics Research Centre in Cork. His technical interests progressed there from device modeling, to circuit simulation and behavioral modelling of analog circuits. In 1999, he joined the precision converter group of Analog Devices, working on architectures for data converter and sensor interfaces. Colin was recognized as an Analog Devices Fellow in 2007.

Liam Madden, *Xilinx, San Jose, CA*

Hardware is cool again, thanks to the emergence of domain specific architectures, particularly in machine learning and AI. The critical question is how we avoid the trap of looking at short-term architectural gains at the expense of tolerance to changes in data types and algorithms. This is particularly important, given the rate at which machine learning algorithms are advancing. Having an adaptable heterogeneous hardware solution allows the best of all worlds, hardware that is tuned to the application, but not tied to the two-year ASIC development cycle. As always, we hardware designers are only as good as the software our customers use to program it, so getting the right abstraction level to optimize the hardware is perhaps the most difficult job of all.

Liam Madden is Executive Vice President of Hardware and Systems Product Development at Xilinx. He has contributed to a range of industry leading products, including: high-performance and low-power microprocessors (Alpha and StrongArm at DEC), embedded processors and IP (MIPS) and consumer devices (Xbox 360 at Microsoft). While at Xilinx, he was responsible for the introduction of stacking technology at 28nm, leading the industry in integrating multiple devices on a silicon interposer, for which he received the 2013 Semi Award. He is also responsible for integrating GS/s data converters into Xilinx products, as part of the RFSoC product line, which is revolutionizing 5G radio solutions and related products. He holds a B.E. degree from University College Dublin, an M.Eng. degree from Cornell University, is an Adjunct Professor at UCD and is a member of the board of Science Foundation, Ireland.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Andreas Olofsson, *DARPA, Arlington, VA*

Engineering productivity has not kept pace with Moore's Law, leading to prohibitive increases in development costs and team sizes for leading edge SoCs. The main strategy for managing complexity in today's SoCs is through design reuse of proprietary licensed intellectual property (IP) modules. The current IP reuse approach has markedly improved productivity, but the single layer point-to-point approach has limited the scope of reuse and abstraction. A pathway to revolutionizing SoC design is to copy the software design community, where open source has enabled a deep software hierarchy with many abstraction layers, significantly increasing productivity. Open source technology forms the foundation of fields like machine learning that are moving forward at an astounding rate. The adoption of an open source culture within the circuit design community would speed up the circuit innovation cycle, while democratizing access to specialized circuits needed to create state-of-the-art mixed-signal systems.

Andreas Olofsson is currently serving as a DARPA Program Manager in the Microsystems Technology Office. His research interests include intelligent design automation, system optimization, and open hardware. Prior to his arrival at DARPA, Mr. Olofsson devoted 20 years to designing and testing low-power processors and mixed-signal circuits at Texas Instruments, Analog Devices, and Adapteva. Chip products designed by Mr. Olofsson include low-power digital signal processors (DSPs), charge-coupled device (CCD) readout circuits, and massively parallel reduced instruction set computing (RISC) processors. From 2008 to 2016, Mr. Olofsson served as the CEO of Adapteva, where he developed the Epiphany architecture and Parallella open-source computer. The Parallella democratized access to parallel computing and catalyzed the growth of a community of 10,000 developers and 200 universities across the globe. Mr. Olofsson received his B.S. in Physics and Electrical Engineering and M.S. in Electrical Engineering from the University of Pennsylvania. Mr. Olofsson is a Member of the IEEE and holds nine U.S. patents.

Shaojun Wei*, Tsinghua University, Beijing, China*

Admittedly, people's interest in hardware is diminishing, and the oft-heard excuse is that Moore's Law is coming to an end. In the past 20 years, on the one hand, Moore's law has given the development of hardware a high degree of certainty, less destructive innovation challenges, and resulted in a declining interest in hardware research. On the other hand, Moore's law has also raised the entry barrier for hardware, so that people who want to engage in hardware development are daunted. This is changing. As a matter of fact, when scaling down is no longer available, we are ushering in the return of hardware. AI training and inference, brain-like computing with new hardware architecture, or the Internet of Things (IoT) characterized by perception, transmission, processing, decision-making and execution, all call for the emergence of new hardware. Therefore, although Moore's Law based on existing CMOS technology will come to an end sooner or later, it does not mean the end of integrated circuits, let alone the end of hardware.

Shaojun Wei is Professor and Dean of the Department of Micro- and Nano-Electronics, Tsinghua University, China. He is also Chief Scientist of a state key science and technology project, member of the National Integrated Circuit Industry Development Advisory Committee, Fellow of the Chinese Institute of Electronics (CIE) and a Senior member of IEEE. Dr. Wei was the President and CEO of Datang Telecom Technology Co., Ltd., and the CTO of Datang Telecom Industry Group. He received the Master in Engineering degree from Tsinghua University, China in 1984 and a Doctor of Applied Science from the Service d'Electronique, Faculté Polytechnique de Mons, Belgium in 1991. Dr. Wei has been working on VLSI design methodologies research and reconfigurable computing technology research. He has published more than 200 papers and 3 monographs. Dr. Wei had won many awards, including the National Second Award for Technology Invention, the National Second Award for Technology Progress, and the SIPO and WIPO Patent Golden Award.

SC: Integrated Phased Arrays: Theory, Practice, and Implementation for 5G and Beyond

Organizer: ***Daniel Friedman,*** *IBM Thomas J. Watson Research Center, Yorktown Heights, NY*

Time:	Topic:
8:00 AM	Breakfast
8:20 AM	Introduction by Chair, ***Daniel Friedman*** *IBM Thomas J. Watson Research Center, Yorktown Heights, NY*
8:30 AM	**Introduction to Phased-Array Technology from mm-Wave to Optics** ***Hossein Hashemi,*** *University of Southern California, Los Angeles, CA*
10:00 AM	Break
10:30 AM	**Scalable mm-Wave Phased Arrays for Imaging and 5G Communications** ***Bodhisatwa Sadhu,*** *IBM T. J. Watson Research Center, Yorktown Heights, NY*
12:15 PM	Lunch
1:20 PM	**Challenges and Architectures for Large-Scale Next-Generation Phased Arrays and 5G Systems** ***Shahriar Shahramian,*** *Nokia, Murray Hill, NJ*
2:50 PM	Break
3:20 PM	**Emerging Topics in Phased Arrays and the Path to THz** ***Ehsan Afshari,*** *University of Michigan, Ann Arbor, MI*
4:50 PM	Conclusion

Introduction

Phased array technology is a staple in specialized radar and long reach communications applications. With the rise of mmWave circuits and associated dense packaging, integrated, scalable phased arrays are key technology elements in a new wave of emerging broad-based commercial applications in the communication and imaging spaces. In this short course, we first introduce the concept of electronically steered arrays, describe how to analyze such designs, discuss analog, digital, and hybrid beamforming phased array tradeoffs, and introduce both mmWave and optical examples. We next discuss some of the key circuits enabling these designs, scalability of such designs, and emerging opportunities for systems leveraging these designs. We continue with a discussion of paths to realizing very large scale arrays, including physical design, packaging, verification, and test. Finally, we conclude with a discussion of emerging topics in phased arrays and the opportunities for even higher frequency phased array solutions.

OUTLINE

8:30 AM

SC1: Introduction to Phased-Array Technology from mm-Wave to Optics
Hossein Hashemi, *University of Southern California, Los Angeles, CA*

This talk will cover the basic principles, architectures, applications, and associated building blocks of phased arrays from radio frequencies up to optical frequencies.

Hossein Hashemi is a Professor of Electrical Engineering at the University of Southern California. His research interests include analog, mixed-signal, radio-frequency and photonic integrated circuits. He received the B.S. and M.S. degrees in Electronics Engineering from the Sharif University of Technology, Tehran, Iran, in 1997 and 1999, respectively, and the M.S. and Ph.D. degrees in Electrical Engineering from the California Institute of Technology, Pasadena, in 2001 and 2003, respectively. Hossein is an Associate Editor for the IEEE Journal of Solid state Circuits. He was a co-recipient of the 2004 IEEE Journal of Solid-State Circuits Best Paper Award and the 2007 IEEE ISSCC Lewis Winner Award for Outstanding Paper.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

OUTLINE

10:30 AM

SC2: Scalable mm-Wave Phased Arrays for Imaging and 5G Communications

Bodhisatwa Sadhu, *IBM T. J. Watson Research Center, Yorktown Heights, NY*

Large numbers of millimeter-wave antenna elements (typically from 10s to 100s) can be supported using multiple integrated circuits (ICs) that are further integrated at the package level to create a phased-array module. This talk motivates the need for millimeter-wave phased-array scaling to realize such large apertures and outlines key challenges toward their implementation. Various aspects of phased array scaling, including beamforming architectures and functional partitioning, signal distribution among multiple ICs, module-level implementation of antennas and packaging, and digital control will be discussed. Two recently introduced scaled 64-element dual-polarized phased array designs will be reviewed as implementation examples: 1) A phased-array transceiver module operating at 28GHz supporting 5G applications, and 2) TX and RX phased arrays operating at 94GHz, suitable for backhaul and imaging applications. Design techniques implementing requirements relevant to these applications will be discussed, and measurement results validating the designs will be presented. Finally, an emerging approach to managing the rich control space associated with highly integrated phased-array designs will be discussed.

Bodhisatwa Sadhu is currently a Research Staff Member at IBM T. J. Watson Research Center. He received the Ph.D. degree in Electrical Engineering from University of Minnesota, Minneapolis, in 2012, working on cognitive radio circuits. Since 2012, Dr. Sadhu has been working on RF and mm-wave transceivers at IBM T. J. Watson Research, where he has led the design and demonstration of a self-healing frequency synthesizer, a low-power 60GHz transceiver IC, and a mm-wave 5G phased array base-station IC. In the last five years, he has authored and co-authored 10+ journal papers, 20+ conference papers, and a book on cognitive radio circuits, and holds 20 issued US patents with 20+ pending. He is currently an Adjunct Assistant Professor at Columbia University, serves as a TPC member of the IEEE RFIC Symposium and IEEE BCICTS, and a guest editor of IEEE JSSC. Dr. Sadhu is the recipient of the 2017 ISSCC Lewis Winner Award for Outstanding Paper, the 2017 Pat Goldberg Memorial Award for the best paper in computer science, electrical engineering, and mathematics published by IBM Research, two IBM A-level Accomplishment awards, six IBM Patent Plateau Awards, the University of Minnesota Graduate School Fellowship in 2007, the 3M Science and Technology Fellowship in 2009, the University of Minnesota Doctoral Dissertation Fellowship in 2011, and stood 2nd in India in the Indian School Certificate (ISC) examination in 2003. He was recognized as an IBM Master Inventor in 2017.

1:20 PM

SC3: Challenges and Architectures for Large-Scale Next-Generation Phased Arrays and 5G Systems

Shahriar Shahramian, *Nokia, Murray Hill, NJ*

The wireless communications industry is at the verge of an exciting transformation: a shift from broadcast-mode to directed-beam communication in the form of large-scale mm-wave phased arrays. In order to enable mass deployment of phased array systems, every layer of our future networks must undergo a transformation. This workshop discusses the challenges associated with the design and architecture of mm-wave phased-array systems up to and beyond the W-band. The impact of technology, circuit topology and chipset architecture on the overall system is examined alongside state-of-the-art reported communication platforms. Furthermore, packaging, calibration and cost reductions techniques are explored.

Shahriar Shahramian received his Ph.D. degree from University of Toronto in 2010, where he focused on the design of mm-wave data converters and transceivers. Dr. Shahramian has been with the Bell Laboratories division of Alcatel-Lucent (now Nokia), Murray Hill, NJ, since 2009 and is currently the Director of the mm-Wave ASIC Research Department. He is also a member of the technical program committee of the IEEE Compound Semiconductor Integrated Circuits Symposium (CSICS) and the IEEE BiCMOS and Compound Semiconductor Integrated Circuits and Technology Symposium (BCICTS). He is also a guest Editor of the IEEE Journal of Solid-State Circuits (JSSC). His research focus includes the design of mm-wave wireless and wireline integrated circuits. He is the lead designer of several state-of-the-art ASICs for optical coherent and wireless backhaul products at Bell Laboratories.

Dr. Shahramian has been the recipient of Ontario Graduate Scholarship, University of Toronto Fellowship and the Best Paper Award at the CSICS Symposium in 2005, 2015, and the RFIC Symposium in 2015. He also holds an Adjunct Associate Professor position at Columbia University, has received several teaching awards, and is the founder and host of The Signal Path educational video series.

3:20 PM

SC4: Emerging Topics in Phased Arrays and the Path to THz

Ehsan Afshari, *University of Michigan, Ann Arbor, MI*

The design and implementation of phased arrays at mmWave and THz frequencies has several challenges, including loss of passives, phase mismatch, limited efficiency of active devices, and the antenna implementation. In this short course, we demonstrate that as we get closer to the cut-off frequency of transistors, new approaches that combine the device physics, circuit techniques, and EM theory are needed. We will discuss several scalable phased arrays for communication and radar applications.

Ehsan Afshari received his Ph.D. in Electrical Engineering from Caltech in 2006, and joined the ECE Department of Cornell University. Ten years later, he joined the EECS department of the University of Michigan, Ann Arbor. His team is engaged in the theoretical foundations, design and experimental validation of analog, RF, mm-wave, and THz integrated devices, circuits and systems for a variety of applications, including communications, imaging and sensing. His work is funded by federal agencies, such the NSF, DARPA, ONR, and the ARL, as well as companies like Intel, TI, Raytheon, and Qualcomm. He has been the recipient of several awards and honors, including a 2008 DARPA Young Faculty Award, a 2010 NSF CAREER Award, a first place at the Stanford-Berkeley-Caltech Innovation Challenge in 2005, and several best paper awards at the leading conferences in his field. He has also served as a Distinguished Lecturer for the IEEE Solid-State Circuits Society. He was selected as one of the 50 most distinguished alumni of Sharif University. His doctoral students have also received several prestigious awards and fellowships, including: 2018, 2017, 2012, 2011, and 2010 Solid-State Circuit Society Predoctoral Achievement Award; 2011, 2013, and 2017 IEEE MTT-S Microwave Engineering Graduate Fellowships; the Cornell Best Ph.D. Thesis Award in 2011 and 2014; as well as many best paper awards. The Ph.D. graduates of his group are leaders in the field, including faculty members at MIT, UC Davis and the University of Minnesota, and companies such as IBM, Bell Labs, Qualcomm and Broadcom.

INDEX TO AUTHORS

A

Abbas, Tarek Al 106, 468
Abdelhalim, K. 120
Abdelhalim, Karim 374
Abdulslam, Abdullah 148
Abidi, Asad 266
Adachi, Kazuhiko 462
Agrawal, Abhishek 174
Ahmed, Abdelrahman H. 484
Ahmed, Khondker Z. 234
Ahmed, Mostafa G. 302
Ahn, Hyo-Joo 378
Ahn, Suseob 354
Ainspan, Herschel A. 122
Akasawa, Fumika 204
Alamouti, Sina Faraji 284
Ali, Tamer 118
Alie, Stephen 114
Almer, Oscar 468
Alvin, Joshua 344
Aly, Mohamed M. Sabry 226
Alzate-vinasco, Juan 212
Alzate, Juan G. 214
Amadjikpe, Arnaud 174
Amakawa, Shuhei 170
Amarnath, Anirudh 218
Ambalathankandy, Prasoon 108
Amer, Aya G. 470
Anantula, Pradeep 218
Aoki, Takeshi 204
Appel, Brad 316
Aprile, Cosimo 476
Ariki, Takuya 218
Arslan, Umut 212, 214
Arya, Kunal 456
Aseron, Paolo 48
Ashida, Mitsuyuki 478
Atallah, Francois 316
Aurangozeb 480
Ayers, James S. 320
Aygun, Levent E. 280
Azam, Ali 78
Azarenkov, Leonid 48

B

Babaie, Masoud 166
Bachmann, Christian 166
Bae, Jeongyeol 354
Baeck, Sangyeop 392
Baek, Jongbeom 238
Baert, Maarten 328
Bagherbeik, Mohammad 126
Bai, Zhidong 78
Bailey, James 482
Baiocco, Christopher 464
Bakkaloglu, Bertan 428
Balaga, Muralikrishna 218
Balakrishnan, Aishwarya 352
Ballantyne, Wayne 352
Banerjee, Utsav 46
Bang, Jun-Suk 238
Baoxing, Chen 244
Bardin, Joseph C. 456
Barends, Rami 456
Bartling, Steven 282
Bartolo, Andrew 226
Bassett, Paul 164
Beakes, Michael P. 122
Beigné, Edith 226
Bellary, Suman 282
Bernardinis, Fernando De 116
Bertulessi, Luca 268
Bespalko, Ryan 482
Besten, Gerrit W. den 474
Bhargava, Pavan 464
Bhartiya, Mukesh 48
Bhat, Abhishek 412
Bhoja, S. 120
Bietti, Ivan 116
Biswas, Dwaipayan 360
Blaauw, David 96, 134, 224, 278, 314, 422, 442
Bodnar, Rares 444
Bol, David 322
Bollati, Giacomino 116
Boora, Akshat 230, 312
Bosi, Alessandro 116
Boukhayma, Assim 290
Bourdoux, Andre 168
Bowman, Keith 316
Braendli, Matthias 476
Brandli, Matthias 112
Breeschoten, Arjan 166, 360, 368
Breslin, James 272
Brockman, Justin 214
Brookes, Stephen 272
Brown, Thomas 352
Bryant, Carl 350
Bui, Quang-Diep 354
Bulzacchelli, John F. 122
Bunsen, Keigo 408
Burg, Andreas 476
Burkett, Brian 456
Byeon, Dae Seok 216
Byun, Jindo 216

C

Caizzone, Antonino 290
Calhoun, Benton H. 232
Callender, Steven 174
Cao, Changhua 350
Cao, Ningyuan 222
Cao, Peng 420
Carlen, Peter 374
Carlton, Brent 174
Carmena, Jose M. 284
Carta, Nicola 116
Carusone, Anthony Chan 482
Carvalho, Carlos 114
Casper, Bryan 172
Cassan, David 482
Cassier, Amer 316
Ceballos, Jose Luis 352
Cesura, Giovanni 116
Cevrero, Alessandro 112, 476
Cha, Young San 216
Chae, Youngcheol 100, 194
Chakrabarti, Anandaroop 172
Chakravorti, Mark 174
Chan, Chi-Hang 60, 66, 260, 340
Chan, Susan 106
Chance, Greg 352
Chandrakasan, Anantha P. 46, 470
Chang, Chin-Ho 306
Chang, Jonathan 390
Chang, Meng-Fan 138, 388, 396, 402
Chang, Muya 222
Chang, Soobong 378
Chang, Ting-Wei 388, 402
Chang, Tung-Cheng 388
Chen, Albert B. 212
Chen, Baoxing 246
Chen, Bo-Jiun 486
Chen, Chao-Hsuan 124
Chen, Ehung 118
Chen, Hsin-Hung 350
Chen, Hsinhua 350
Chen, Huan-Sheng 118
Chen, Jhy-Rong 338
Chen, Jia-Jing 396
Chen, Jimmy 456
Chen, Ke-Horng 242, 434, 466
Chen, Kehou 162
Chen, Kuan-Dar 294
Chen, Kuang-Ren 118
Chen, Mark 306
Chen, Mike Shuo-Wei 74
Chen, Min 158
Chen, Pierce 162
Chen, S. 210
Chen, Ting-Sheng 280
Chen, Wei-Hao 388, 402
Chen, Xing 442
Chen, Ya-Chi 294
Chen, Yan-Jiun 162
Chen, Yan-Ting 124
Chen, YangChuan 350
Chen, Yen-Huei 390
Chen, Yi-Fu 162
Chen, Yi-Ju 138
Chen, Yi-Wen 306
Chen, Yingping 248, 250
Chen, Yong 270, 410
Chen, Yu 278, 456
Chen, Yu-Lun 162
Chen, Zhengyu 324
Chen, Zhiyuan 424
Chen, Ziyun Li, Yu 134
Cheng, Po-Hao 372
Cheon, Clifford DY 80
Cheon, Jimin 100
Cherniak, Dmytro 268
Chhabra, Saransh 48
Chi, Hankyu 380
Chi, Hyungjoon 378
Chiaro, Benjamin 456
Chibvongodze, Hardwell 218
Chiu, Pao-Cheng 334
Chiu, Yen-Cheng 388, 396
Cho, Hyunyoon 378
Cho, Joo-Hwan 380
Cho, Minchang 278
Cho, SeongHwan 366
Cho, Thomas Byunghak 238, 240, 354
Cho, Yoonseo 258
Cho, Yunkyo 130
Choi, Beomseok 146
Choi, Byeongchan 380
Choi, Geunho 380
Choi, Heonjin 194
Choi, Jaebin 198
Choi, Jaehyouk 258, 490
Choi, Jaeseung 392
Choi, Jaesung 392
Choi, Jeonghyun 238
Choi, Jonghyuck 380
Choi, Jung-Hwan 378
Choi, Minho 432
Choi, Jonghyuck 382
Choi, Seojin 490
Choi, Seouk-Kyu 378
Choi, Steve 218
Choi, Sungpill 136
Choi, Sunmyung 380
Choi, YeonKyu 378
Choi, Youngdon 216
Chong, Euhan 114
Chong, S. 120
Choo, David Kyojin 422
Choo, Gyosoo 216
Choo, Kyojin D. 96
Choo, Min-Seong 432
Choo, Younghwan 238
Choudhury, Debabani 172
Chow, Chung-Kai 164
Chu, Hsiang-Yun 162
Chu, Li-Cheng 434
Chueh, Tzu-Chan 350
Chun, Junhyun 380, 384
Chung, Jinil 380
Codega, Nicola 116
Collados, Manel 350
Connor, Chris 212
Craninckx, Jan 58, 68, 408
Cressler, John D. 80
Curtin, Ciarán 272
Cutrupi, Massimo 116

D

Danesh, Seyed 180
Darabi, Hooman 266
Darne, Siddhesh 218
Das, Nilanjan 214
Das, Reetuparna 224
Dastjerdi, Mahmood Baraani 448
Date, H. 210
Dattawadkar, Shreela 48
Davis, Brandon 352
Davis, Curtis 164
Davoodi, M. 120
Dazzi, Martino 112
De, Vivek 48, 146, 154, 234, 404
Deguchi, Jun 478
Dehaene, Wim 206, 328
Dehlaghi, Behzad 482
Dehng, Guang-Kaai 350
Dekimpe, Rémi 322
Deng, Wei 256
Dermit, Davide 58, 68
Desai, Nachiket 154
Deshmukh, Pranjali 48
Desset, Claude 168
Dhama, Praveen 48
Dick, Carson 480
Dickson, Timothy O. 122
Dijkhuis, Johan 166
Ding, Ran 484
Do, Huong 154
Doherty, Mark 80
Dong, Qing 314
Dong, Ruibing 170
Donnelly, Yann 272
Dornbusch, Andrew 164
Dreps, Dan 42
Duan, Y. 120
Dufrêne, Krzysztof 356
Dunsworth, Andrew 456
Dunwell, Dustin 482
Dupuis, Tim 164
Dutta, Gairik 102
Dutton, Neale 106, 468

E

Eckert, Charles 224
Eissa, Mohamed Hussein 82
El-Aassar, Omar 86, 416
Elies, Brian 282
Elkholy, Ahmed 302
Elmoznine, Abdellatif 484
ElShater, Ahmed 70, 118
Endoh, Tetsuo 202
Englund, Dirk R. 458
Enz, Christian 290
Eovino, Benjamin E. 190
Eshraghi, Aria 164
Esmaeelzadeh, Hani 300
Everson, Luke R. 50

F

Fabiano, Ivan 116
Fadila, Ashbir Aviat 344
Fahrenkopf, Nicholas 464
Fan, Yongping 320
Farhoodfar, A. 120
Fathima, Shirin 228
Fischer, Kevin 212, 214
Flandre, Denis 322
Flory, Robert 174
Flynn, Michael P. 332
Fong, Kevin 162
Forey, S. 120
Fowler, Austin 456
Foxen, Brooks 456
Foy, Christopher 458
Francese, Pier-Andrea 112, 476
Freiman, Amit 352
Frenkel, Charlotte 322
Fujimaki, Akira 460
Fujishima, Minoru 170
Fujita, Masahiro 276
Fujiwara, Hidehiro 390
Fukai, Shuji 204
Fukuda, R. 210
Fukuda, Ryo 218
Furubayashi, Yuki 182

G

Gamini, Nader 192
Gao, Bin 402
Gao, Yue 370
Garampazzi, Marco 116
Garcia, Manuel Moreno 102
Garghetti, Alessandro 268
Garudadri, Harinath 362
Genoe, Jan 206
Genov, Roman 102, 374
Geren, Bruce 352
Ghanbari, Mohammad Meraj 284
Ghani, Tahir 214
Ghittori, Nicola 116
Ghosh, Dibyendu 48
Giannini, Vito 164
Gidney, Craig 456
Giunco, Fabio 116
Giustina, Marissa 456
Glassee, Miguel 168
Goedemé, Toon 44
Goldenberg, Marius 164
Golonzka, Oleg 212, 214
Gong, Luyao 134
Gopalakrishnan, K. 120
Graff, Rob 456
Griffin, Patrick 272
Griffith, Danielle 302
Grimaldi, Luigi 268
Groppe, David 374
Gu, Jie 318, 324
Gu, Qun J. 452
Guidry, Michael 180
Gulve, Rahul 102
Guo, Hao 410
Guo, Liang 162
Guo, Xiaochuan 350
Gupta, Ankit 48
Gupta, Siddharth 392
Gürleyük, Çağrı 184
Gusev, Nikita 102
Gweon, Surin 288, 364
Gyongy, Istvan 106

H

Ha, Hyunsoo 360, 368
Ha, Kyung-Soo 378
Haas, Harald 468
Hada, Ryuji 132
Hahn, Wook-ghee 216
Haim, Harel 102
Haine, Thomas 322
Hall, Drew A. 196
Hall, Neal 296
Hamada, Toshiki 204
Hamashita, Koichi 70
Hamzaoglu, Fatih 212, 214
Han, Donghyeon 142
Han, Jaeyeol 238, 240
Han, Minsik 380
Han, Ruonan 458
Han, Sangshin 392
Han, Sangwook 354
Han, Seokhwan 384
Han, Yuntack 384
Hanumolu, Pavan Kumar 302
Hanyu, Takahiro 202
Hara, Shinsuke 170
Harasaka, Kazuhiro 462
Hardy, Casey 150
Harinath, Mahesh 316
Harpe, Pieter 166
Hashiguchi, T. 210
Hashimoto, T. 210
Hassan, Mohammed 350
Hassibi, Arjang 192
Hattori, Toshihiro 54
Hausmann, Kurt 352
Hayashi, Koichiro 218
Hayashi, Masato 52
He, Jie 420
Hegde, Manju 164
Helal, B. 120

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Helleputte, Nick Van 360, 368
Henderson, Robert 468
Henderson, Robert K. 106
Heng, Chun Huat 370
Hentges, Patrick 212, 214
Heo, Deukhyoun 152
Heo, Seungchan 354
Herdian, Hans 462
Hershberg, Benjamin 58, 68
Heydari, Payam 450
Hills, Gage 470
Hiramatsu, Shota 108
Hirose, Yutaka 104
Hisada, T. 210
Hisada, Toshiki 218
Hishida, Tomoo 218
Ho, Huong 114
Ho, Rebecca 470
Hochberg, Michael 484
Holland, William 180
Hong, Songcheol 176
Hong, Wookpyo 152
Hong, Yunseok 384
Hong, Zhiliang 420
Honjo, Hiroaki 202
Honkote, Vinayak 48
Hoof, Chris Van 360, 368
Hori, Shinichi 344
Horng, Jaw-Juinn 306
Hosoda, Soichiro 132
Hosono, K. 210
Hossain, Masum 480
Hou, Jinjian 94
Hsiao, Chuan-Hung 294
Hsiao, PC 162
Hsieh, Chia-Hsiang 372
Hsieh, Chih-Cheng 388, 396
Hsieh, Hung-Yi 334
Hsieh, Yi-Jhan 162
Hsieh, Yu-Lin 350
Hsin, Wei-Kuo 162
Hsu, Chen-Kai 64
Hsu, Hua-Ling 218
Hsu, Jui-Lin 162
Hsu, Weishow 162
Huang, Chia-Ming 434
Huang, Chun-Wen Paul 80
Huang, Hongye 256, 344
Huang, Hsiang-En 124
Huang, Min-Yu 346
Huang, Peng-Wei 372
Huang, Po-Chun 162
Huang, Po-Yi 390
Huang, Shih-Hao 118
Huang, Trent 456
Huang, Tsung-Yuan 388
Huang, Tzu-Ping 466
Huang, Wei-Hsing 396
Huang, Yen-ling 352
Huang, Yi-Chieh 118, 486
Huang, Yi-Wei 430
Huemer, Mario 356
Huijsing, Johan H. 298
Hull, Chris 174
Hung, Chao-Ching 162
Hung, Chih-Ming 162
Hung, Szu-Yao 262
Hurwitz, J.E.D. 180
Hutchings, Sam W. 106
Huynh, Tam 484
Hwang, Hyejin 392
Hwang, Hyong-Ryol 378
Hwang, Sangjoon 216
Hwang, William 226
Hyuga, Fumihiko 132
Hyun, Seok-Hun 378

I

Ibrahim, Ahmed 426
Ibrahim, Mohamed I. 458
Ihm, Jeongdon 216
Ikebe, Masayuki 108
Ikeda, Makoto 204
Ikeda, Shoji 202
Ikeda, Takayuki 204
Ikegami, Takahiro 108
Ilderem, Vida 35
Im, Dain 380
Inoue, Akito 104
Inoue, Hiroki 204
Inoue, Koji 460
Ishigaki, Yutaro 132
Ishii, Motonori 104
Ishizu, Takahiko 204
Isobe, Atsushi 182
Isobe, Katsuaki 218
Ito, Masao 54
Ito, Rui 478

J

Jackson, Bradley 48
Jahan, Rawshan 214
Jain, Ankesh 336
Jain, Kartik 48
Jain, Pulkit 212, 214
Jain, Sanket 448
Jang, Jaehyuk 354
Jang, Jin-Hun 378
Jang, Jun-Woo 130
Jang, Moonhyung 194
Jang, Seong-Jin 378
Jang, Sungchun 380
Jang, Sunyung 392
Jang, Taekwang 422
Jann, Benjamin 352
Jaussi, James 172
Jeffrey, Evan 456
Jeon, Byungdeuk 384
Jeon, Dongsuk 140
Jeon, Su Chang 216
Jeong, Chunseok 384
Jeong, Daechul 354
Jeong, Deog-Kyoon 432
Jeong, Hanwool 392
Jeong, Jihoon 316
Jeong, Seokhyeon 278, 314
Jeong, Seungbin 226
Jeong, Stanley 218
Ji, Youngwoo 308
Jia, Tianyu 318
Jiang, Dongyang 336
Jiang, Wenning 60
Jiang, Yang 424
Jie, Lu 332
Jin, Dapeng 162
Jo, Youngmin 216
Johnson, Benjamin C. 284
Johnston, Nick 106
Jong, Ilyong 354
Joseph, Russ 318
Ju, Inchan 80
Juan, Brian 162
Juang, Ying-Zong 434
Jung, Doohwan 90
Jung, Wontaeck 216
Jung, Yeonwook 216

K

Ka, Dongyoon 384
Kabe, Tatsuya 104
Kadavakollu, Sureshbabu 48
Kahle, James A. 42
Kajitani, Y. 210
Kajitani, Yasuyuki 218
Kamada, Yudai 182
Kanazawa, Yuri 108
Kanda, K. 210
Kaneko, T. 210
Kaneko, Tomoya 344
Kang, Byoungjoong 354
Kang, Dongku 216
Kang, Inyup 130, 238, 240, 354
Kang, Jang Yong 218
Kang, Jeonseung 392
Kang, Jungmyung 392
Kang, Kyeongpil 384
Kang, Kyung-Min 216
Kang, Mankeun 384
Kang, Sanghoon 136, 354
Kanumalli, Ram Sunil 356
Kao, Hui-Yao 388
Kar, Monodeep 404
Karandikar, Niranjan 352
Karman, Saleh 268
Kasamatsu, Akifumi 170
Kasuga, Shigetaka 104
Katic, Nikola 102
Kato, Yosuke 218
Kaufmann, Maik 156
Kavala, Anil 216
Kawabe, T. 210
Kawaguchi, K. 210
Kawaguchi, Kazuaki 218
Kawai, Shusuke 252
Kawakami, K. 210
Kawaminami, Masaru 314
Kazama, Hideto 182
Ke, Xugang 158
Keaveney, Michael 272
Kelly, Julian 456
Kennedy, Henry 444
Kennedy, Michael Peter 272
Khanna, Sudhanshu 282
Khashaba, Amr 302
Kiaei, Sayfe 428
Kiani, Mehdi 426
Killips, B. 120
Kim, Byungsoo 384
Kim, Byungsub 308
Kim, Changhyeon 136
Kim, Changhyun 380
Kim, Chris H. 50
Kim, Chulwoo 382
Kim, Daesuk 380
Kim, Dongkyun 380
Kim, Dongsu 238, 240
Kim, Gain 476
Kim, Haksong 380
Kim, Hong 164
Kim, Hongjung 380
Kim, Hoon 392
Kim, Hun-Seok 134, 278, 442
Kim, Hyunjin 216
Kim, Hyunseok 384
Kim, Ik-Hwan 238, 240
Kim, Jaeil 380
Kim, Jaeyoung 392
Kim, James C. 216
Kim, Ji-Hoon 364
Kim, Jongsam 380
Kim, Juyeop 258
Kim, K. 120
Kim, Ki-sung 216
Kim, Kiwon 216
Kim, Kukjoo 198
Kim, Kwandong 384
Kim, Kwang-Ho 218
Kim, Kwangwon 216
Kim, Kwantae 288, 364
Kim, Kyungmin 216
Kim, Kyungwhan 380
Kim, Minseo 288, 364
Kim, Minsu 216
Kim, Nahyun 216
Kim, Nayeon 384
Kim, Sang-Yun 378
Kim, Sanghui 380
Kim, Seung-Bum 216
Kim, Soyeon 364
Kim, Sunghoon 216
Kim, Taehwan 464
Kim, Taehyung 392
Kim, Won-Tae 216
Kim, Yejoong 96, 278, 314, 422
Kim, Yeonuk 380
Kim, Yongju 384
Kim, Yongmi 380
Kim, Yongseop 384
Kim, Yongsun 384
Kim, Young-Hwa 378
Kim, Youngcho 216
Kim, Youngjoo 384
Kim, Youngwoo 136
Kimura, Katsuyuki 132
King, Ya-Chin 388
Kinget, Peter R. 438
Kissinger, Dietmar 82
Klimov, Paul 456
Klotchkov, Ilya 48
Ko, Kuihan 216
Kobayashi, Hiroyuki 478
Kobayashi, N. 210
Koh, Y. L. 210
Koike, Hiroki 202
Koike, Takao 54
Koizumi, Tomohiro 132
Kojima, M. 210
Kollikunnel1, Jisna 48
Komai, Hiromitsu 218
Kondo, Hiroyuki 54
Kong, Yuyao 394
Konijnenburg, Mario 360, 368
Koo, Kibong 380
Koo, Nahmil 366
Kosman, John 468
Kossel, Marcel 112, 476
Koyama, Shinzo 104
Kozuma, Munehiro 204
Krishnamurthy, Harish 154
Krishnamurthy, Harish K. 234
Krishnapura, Nagendra 412
Krishnasawamy, Nagaraj 282
Krishnaswamy, Harish 446, 448
Kronemeijer, Auke J. 206
Krotnev, Petar 114
Kruger, Seth 464
Kruiskamp, Wim 304
Ku, Sanghyun 380
Kubozoe, Ryo 344
Kuimelis, Robert G. 192
Kulkani, Swaroop 218
Kulkarni, Jaydeep P. 228
Kull, Lukas 112, 476
Kumar, Prakhar 280
Kundu, Amit 306
Kundu, Somnath 174
Kunihiro, Kazuaki 344
Kuo, Tai-Haur 430
Kurian, Dileep 48
Kutulakos, Kiriakos N. 102
Kwak, Dong-Hun 216
Kwon, Hyuck-Joon 378
Kwon, Kihun 380
Kwon, Ohwon 218
Kwon, Tae-Hong 216
Kye, Chan-Ho 432
Kyung, Kye Hyun 216

L

Lacaita, Andrea L. 268
LaCroix, Marc-Andre 114
Lagos, Jorge 58, 68
Lai, Sheng-Tsung 124
Lajevardi, Pedram 98
Lambert, William 146, 154, 174
Law, Man-Kay 424
Le, Binh Q. 226
Le, Hanh-Phuc 150
Le, M. 120
Leach, Jonathan 106
Lebedev, Semyon 114
Leblebici, Yusuf 476
LeCun, Yann 12
Lee, Calvin Yoji 70
Lee, Chang-Kyo 378
Lee, Changhyuk 198
Lee, Changhyun 380
Lee, Changwoo 384
Lee, Cheng-Yen 466
Lee, Chun-Ying 388
Lee, Dongho 176
Lee, Dongkeon 378
Lee, Doobock 384
Lee, Eungsuk 216
Lee, Hao-Yun 372
Lee, Hyesoo 194
Lee, Hyeyeon 194
Lee, Inhak 392
Lee, Inhee 422
Lee, Inryoul 216
Lee, J. 210
Lee, Jae-Gon 130
Lee, Jaein 380
Lee, Jeongsup 314
Lee, Jihee 190
Lee, Jinmook 142
Lee, Jinsu 142
Lee, Jiwon 288, 364
Lee, Jiyoung 216
Lee, Jongwoo 238, 240, 354
Lee, Joonhee 354
Lee, Juhyoung 142
Lee, Jungho 308
Lee, Juyun 140
Lee, Kyoung-Rog 190, 288
Lee, Mingyu 384
Lee, Moo Kyung 216
Lee, Dong Uk 380
Lee, Sam 162
Lee, Sangyeop 170
Lee, Sehwan 130
Lee, Seongju 384
Lee, Seungchan 176
Lee, Seungpil 218
Lee, Seungseob 378
Lee, Shuenn-Yuh 372
Lee, Sungjun 238, 240
Lee, Teerasak 444
Lee, Yeonho 382
Lee, Yi-Bin 350
Lee, Yongsu 288, 364
Lee, Yongsun 258, 490
Lee, Yongsung 380
Lee, Yumi 240
Leng, Weiyu 118
Leuenberger, Spencer 70
Levantino, Salvatore 268
Li, Dai 400
Li, Daniel 350
Li, Haitong 226
Li, Heguang 218
Li, Hui 370
Li, Jason 218
Li, Jia-Fang 388
Li, Jinbo 452
Li, Qiang 396
Li, Shaolan 64, 296
Li, Shuo 232
Li, Tong 76
Li, Tso-Wei 88
Li, Xi 352
Li, Xueqing 138
Li, Yongfu 370
Li, Zheng 344
Lian, Yong 370
Liang, Joshua 482
Liao, Chi-Chung 372
Liao, Hung-Jen 390
Liao, Y. 120
Liao, Zhan-Xien 372
Liaw, Jhon-Jhy 390
Liempd, Barend van 58, 68
Liempd, Chris van 360
Liew, W. 120
Lim, Daihyun 484
Lim, Hyungsun 354
Lim, Jaehyun 392
Lim, Jongyup 278, 314, 422
Lim, Lysander 164
Lim, Sukhyun 378
Lim, Wooteak 314
Lim, Yongbeom 194
Lim, Younghyun 258, 490

Lin, Alex TC 162
Lin, Blake 214
Lin, Blake C. 212
Lin, Cheng-Chieh 350
Lin, Cheng-Han 390
Lin, Chih-Yu 390
Lin, Chin-Yu 330
Lin, Chorng-Jung 388
Lin, Chou-Ching 372
Lin, Chungshi 118
Lin, Fa-Yi 242
Lin, Kao-Cheng 390
Lin, Lei 218
Lin, Leon 264
Lin, Liwei 190
Lin, Shian-Ru 434, 466
Lin, Wei-En 388, 402
Lin, Wei-Tin 434
Lin, Wei-Yu 388
Lin, Yen-An 242
Lin, Yen-Ting 466
Lin, Yin-Hsi 242, 434, 466
Lin, Ying-Zu 330
Lin, Zong-Yi 434, 466
Liu, Bangan 344
Liu, Boxiao 370
Liu, Chia-Hao 434
Liu, Chun-Cheng 118
Liu, Fang 246
Liu, Hanfei 162
Liu, Hanli 256, 344
Liu, Je-Syu 388
Liu, Jiaxin 296
Liu, Kuo-Chi 242
Liu, Lu 134
Liu, Qi 402
Liu, Ren-Shuo 388, 396
Liu, Rui 396
Liu, Ruoyang 138
Liu, Xiaosen 234
Liu, Yan 394
Liu, Yang 192, 484
Liu, Yao-Hong 166
Liu, Yongpan 138
Liu, Yun Hua 114
Lo, Chung-Chuan 388
Lo, Tien-Yu 334
Lobo, Ryan 164
Loi, C. 120
Lok, Chi-Hang 162
Lou, Chih-Hong 338
Lu, Chao-Hsin 330
Lu, Danzhu 420
Lu, Shen-Fu 242
Lu, Yen-Ju 162
Lucero, Erik 456
Luebbe, Juergen 282
Lueders, Michael 156
Luo, Xueting 344
Luu, Danny 112, 476
Lyu, Yifan 62

M

Ma, D. Brian 248, 250
Ma, Shaoyu 246
Ma, Weijian 94
Ma, Yangjin 484
Ma, Yitao 202
Ma, Yixin 370
Ma, Yu-Sheng 466
Madhukumar, Akhila 48
Maeda, Shohei 54
Magni, Federico 116
Maharbiz, Michel M. 284
Maher, Monier 164
Mahmoudidaryan, Parisa 428
Mak, Pui-In 270, 410, 424
Makinwa, Kofi A. A. 184, 186, 298
Maligeorgos, James 164
Mandal, Debashis 428
Mangal, Vivek 438
Manickam, Arun 192
Manjunath, Pai K. 218
Markulic, Nereo 58, 68, 408
Martens, Ewout 58, 68
Martinis, John 456
Martins, Rui P. 60, 66, 260, 270, 336, 340, 410, 424
Massolini, Roberto 116
Masuda, Atsushi 132
Masuda, M. 210
Mateman, Paul 166
Mathew, Sanu 404
Matsuda, Aki 218
Matsumoto, Akira 182
Matsumoto, Daiki 344
Matsumoto, Y. 210
Matsuzaki, Takanori 204
Mazhab Jafari, Hamed 374
McEwen, Matthew 456
McPartlin, Mike 80
Megawer, Karim M. 302
Meghelli, Mounir 122
Megrant, Anthony 456
Mellati, A. 120
Meng, Miao 426
Menolfi, Christian 112, 476
Mercier, Patrick P. 148, 286, 362, 440
Mercuri, Marco 166
Meterelliyoz, Mesut 212, 214
Mezhiba, Andrey 320
Miakashi, M. 210
Milani, Antonio 116
Min, Byung-Jun 216
Min, Hao 76
Min, Jinyong 384
Min, Youngsun 216
Minamide, HIroaki 108
Minamoto, T. 210
Minamoto, Takatoshi 218
Minuti, Alberto 116
Mitra, Subhasish 226
Mittal, Anurag 44
Miwa, Hitoshi 218
Miwa, Toru 218
Miyoshi, Satoru 314
Mo, Yaowu 94
Mohammad, Maruf 480
Mohanavelu, R. 120
Moiseev, Mikhail 48
Mondal, Susnata 348
Moon, Daesik 378
Moon, Daeyoung 392
Moon, Seung hyun 216
Moon, Un-Ku 70
Moore, Kevin 468
Moore, Thomas 316
Moreau, Ludovic 322
Moreaux, Laurent 198
Morelli, Joshua 316
Moreno, Jaime 42
Morf, Thomas 112, 476
Mori, Keijiro 182
Mori, Mitsuyoshi 104
Mori, Naoki 182
Morimoto, Makoto 478
Moriya, Akira 132
Morozumi, N. 210
Motoi, Keiichi 344
Muhlestein, Jason 70
Mukherjee, Abhishek 64
Mukhopadhyay, Saibal 404
Muller, Rikky 284
Murakami, Ryusuke 132
Murmann, Boris 98
Murphy, David 266
Musa, Faisal Ahmed 114
Musha, J. 210
Muthukumar, Sriram 48
Mutus, Josh 456
Myny, Kris 206

N

Na, Byongwook 378
Na, Jukwan 194
Na, Taesik 234
Naaman, Ofer 456
Nagadomi, Y. 210
Nagao, O. 210
Nagaoka, Ikki 460
Nagulu, Aravind 446
Najar, Hadi 282
Nakagawa, T. 210
Nakai, J. 210
Nakamura, D. 210
Nakamura, H. 210
Nakamura, Hiroshi 218
Nakamura, Takeshi 344
Nakanishi, Kentaro 104
Nakano, T. 210
Nandakumar, Ganeshram 48
Nani, Claudio 116
Narayanan, Aravind Tharayil 344, 462
Narayanan, Karthik 48
Nasu, H. 210
Nasuno, Takashi 202
Natarajan, Arun 448
Natsui, Masanori 202
Neeley, Matthew 456
Nei, Kousei 204
Neill, Charles 456
Netto, S. 120
Neven, Hartmut 456
Ngai, Tat 464
Nguyen, Bai 152
Nguyen, Hoan 316
Nguyen, Huy Thong 84, 90
Nguyen, Qui 218
Nguyen, Xuan-Thuan 374
Ni, Ronghua 354
Nibhanupudi, S. S. Teja 228
Nicolescu, Dorin Alexandru 114
Nimmagadda, Mallikarjuna Rao 48
Noguchi, Yasuo 202
Nomiyama, Takahiro 238, 240
Nose, Yugo 104
Notaros, Jelena 464
Novack, Ari 484
Nowak, Etienne 226
Nozawa, Mai 478

O

O'Farrell, Michael 482
O'Leary, Gerard 374
O'Donnell, Jim 214
Ochi, Y. 210
Ocket, Ilja 168
Odagawa, Akihiro 104
Ogawa, M. 210
Ogawa, T. 210
Oh, Jonghoon 380, 384
Oh, Jonghyun 432
Oh, Sechang 278, 422
Oh, Seunghyun 354
Oh, Seungwook 380
Oh, Tae-Young 378
Ohno, Hideo 202
Okada, Kenichi 256, 344, 462
Okino, Toru 104
Okumura, Masatoshi 218
Okumura, Naoto 54
Omid-Zohoor, Alex 98
Onizuka, Kohei 252
Ookuma, Naoki 218
Ortmanns, Maurits 336
Oshima, Naoki 344
Oshima, Takashi 182
Otani, Sugako 54
Otsuki, Norimasa 54
Ou-Yang, You-Zheng 242
Ouyang, Yuxuan 362
Ozawa, Susumu 218
Ozkaya, Ilter 112, 476

P

Paek, Ji-Seon 238, 240
Pai, Anil 218
Pal, Nilanjan 302
Pali, Ravi 48
Pamarti, Sudhakar 262, 300
Pamula, Rajesh 312
Pamula, Venkata Rajesh 230
Pan, Hsien-Yu 390
Pan, Sining 184, 186
Pang, Jian 256, 344
Pang, Yachuan 402
Paramesh, Jeyanandh 348
Park, Changnam 392
Park, Chanmin 100
Park, Dukha 378
Park, Euiyoung 238, 240
Park, Gwangtae 142
Park, Henry 118
Park, Hong-June 308
Park, Hyun-Wook 216
Park, Hyunsu 382
Park, Il 384
Park, Injun 100
Park, Jeong Hoan 190
Park, Jeongwoo 140
Park, Ji-hyun 216
Park, Jinseok 176
Park, Jiwoong 286
Park, Joo-Yong 216
Park, Jooyong 216
Park, Jun-Seok 130
Park, Junghwan 378
Park, Kiwon 378
Park, Kwang-Il 378
Park, Minsu 380
Park, Youn-Sik 378
Park, Younghwan 392
Pascale, Paolo 116
Pathak, Abhishek 46
Patil, Sanganagouda 272
Pawliuk, Peter 352
Pazhouhandeh, M. Reza 374
Peguvandla, Ramesh 48
Pei, Lei 192
Pelgrom, Marcel J. M. 62
Pellerano, Stefano 174
Pellicone, Demetrio 116
Peng, BC 162
Peng, Pen-Jui 124
Peng, Yimai 422
Peng, Yung-Chow 306
Pernillo, J. 120
Petrov, Dmitry 114
Piech, David K. 284
Pimenta, Matheus F. 184
Pisati, Matteo 116
Polomik, Anthony 316
Popov, Roman 48
Post, Rein 304
Poulton, Christopher V. 464
Pozzati, Enrico 116
Prabha, P. 120
Pretl, Harald 356
Proesel, Jonathan E. 122
Prummel, Jan 304
Puls, Conor 214

Q

Qi, Liang 336
Qian, He 402
Qian, Yao 420
Qin, Wenhui 246
Qiu, Junjun 344
Quintana, Chris 456
Quintero, Pedro 214
Quintero, Pedro A. 212

R

Rachineni, Raghavendra 218
Rad, F. 120
Radhakrishnan, Arun 48
Radhakrishnan, Kaladhar 146, 154
Radway, Robert 226
Raghunathan, N. 210
Rahman, Fahim ur 312
Rahman, Wahid 126
Rajan, Anand 404
Rajendra, Srinivas 218
Ramachandra, Venky 218
Ramkaj, Athanasios 62
Ramos, Joao 350
Ramos, Juan Carlos Pena 62
Ravichandran, Krishnan 146, 154, 234
Raychowdhury, Arijit 222, 316
Razavi, Behzad 264
Rebeiz, Gabriel M. 86, 416
Redman-White, William 444
Reeves, Nathaniel 316
Reiskarimian, Negar 448
Ren, Ao 138
Ren, Guanjing 94
Renz, Peter 156
Riani, J. 120
Robinson, Nancy 48
Roman, Jose 484
Rooijers, Thije 298
Roose, Florian De 206
Rossini, Andrea 116
Roukes, Michael L. 198
Roundy, Shad 426
Roushan, Pedram 456
Roy, Ankur Guha 352
Ruiz, Alberto Del Rio 48
Rush, Fred 164
Rylyakov, Alexander 484

S

Sadjina, Silvester 356
Saengchan, Rattanan 344
Sagazio, Peter 174
Sahu, Kamakhya P. 48
Sahu, Tanaya 212, 214
Saito, Shigeru 104
Sako, M. 210
Sakurai, K. 210
Saligane, Mehdi 314
Samori, Carlo 268
Sank, Daniel 456
Sano, Eiichi 108
Sano, Toru 132
Sapatnekar, Sachin S. 50
Sarhangnejad, Navid 102
Sathe, Visvesh 230, 312
Sato, Hideo 202
Sato, J. 210
Sato, Junji 170
Sato, M. 210
Sato, Nobuhiro 132
Satoh, Yuji 478
Sayal, Aseem 228
Schaef, Christopher 146, 154
Schmid, Otto 164
Schramme, Maxime 322
Sekhar, Meenakshi 212, 214
Seki, Takako 204
Sekiguchi, Tomonori 182
Seo, Dongwook 392
Seo, Sungwhan 216
Seo, Youngsuk 380
Seol, Ji-Hwan 96
Seong, Taeho 258, 490
Shahramian, Shayan 482
Shahrjerdi, Davood 192
Shao, Ke 94
Shao, Zexu 94
Shaoyu, Ma 244
Sharif-Bakhtiar, Alireza 482
Sheelavant, Sunil 166
Sheikh, Farhana 174
Sheikholeslami, Ali 126
Shekhar, Sudip 484

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

Shelly, Brian 272
Shen, Konlin 284
Shen, Kuan-Yueh James 320
Shen, Linxiao 64, 296
Shen, Yi 64
Shepard, Kenneth L. 198
Shi, C.-J. Richard 414
Shi, Longxin 394
Shi, Qixian 408
Shi, Ruizhi 484
Shi, Wei 64
Shi, Wenjie 94
Shi, Zhan 278
Shi, Yao 442
Shibasaki, Takayuki 126
Shibata, N. 210
Shibata, Noboru 218
Shibuya, T. 210
Shih, Ted 124
Shih, Wei-Yan 282
Shim, Dong-kyo 216
Shim, Youngbo 384
Shimazaki, Yasuhisa 54
Shimizu, T. 210
Shimizu, Yutaka 478
Shin, Dongjoo 136
Shin, Junghyun 384
Shin, Seungjun 378
Shin, Woorim 174
Shirane, Atsushi 256, 344, 462
Shopov, Stefan 172
Shtargot, Leonard 158
Shu, Yun-Shiang 334
Shulaker, Max M. 470
Si, Xin 396
Siau, Chang 218
Sijbers, Wim 360, 368
Sim, Jae-Yoon 308
Sim, Jincheol 382
Sin, Sai-Weng 336
Singh, Arvind 404
Singh, Rahul 348
Singh, Rituraj 192
Sinha, Siddhartha 168
Siriburanon, Teerachot 256
Sivaraj, Vijay Deepak 48
Smets, Sander 44
Smit, Meint 29
Sobue, Kazuki 70
Sohn, Young-Soo 378
Soliman, Mazen 118
Someya, T. 210
Someya, Teruki 256
Son, Kiyong 354
Son, Younghoon 378
Song, Fei 264
Song, Houk 384
Song, Jinook 130
Song, Joon-Ho 130
Song, Jun-Yong 380
Song, Junyong 384
Song, Junyoung 382
Song, Ki-Whan 216
Song, Min Kyu 158
Song, Seonghwi 380
Song, Seungwoo 194
Song, Shuang 360, 368
Song, Taejoong 392
Soni, Saksham 48
Sosio, Marco 116
Spalding, Joe 180
Spelgatti, Giorgio 116
Sreenivasulu, N. 48
Srinivasan, Anuradha 48
Stanzione, Stefano 360
Stas, François 322
Steyaert, Michiel 62
Stojanovic, Vladimir 464
Stoppa, David 102
Strackx, Maarten 62
Strange, Jon 350
Streshinsky, Matthew 484
Strutt, Nathan 212
Sturm, James C. 280
Su, Sheng-Yuan 162, 338
Subramaniyan, Arun 224
Sugawara, H. 210
Sugimoto, T. 210
Sukkar, Rafid 484
Summerfelt, Scott 282
Sun, Chih-Hao Eric 350
Sun, J. 120
Sun, Nan 64, 296
Sun, Wenyu 138
Sun, Xiaoyu 396, 402
Sun, Xun 230, 312
Sun, Yu 316
Sun, Zheng 256, 344
Suzuki, Daisuke 202
Suzuki, Mitsuru 462
Suzuki, Satoshi 174
Suzuki, Yasufumi 54
Sveiven, Michael 196
Sylvester, Dennis 96, 134, 224, 278, 314, 422
Syu, Jin-Siang 350

T

Taal, Adriaan J. 198
Tachibana, R. 210
Tajima, Ryota 204
Takagiwa, T. 210
Takahashi, Kei 204
Takamoto, H. 210
Takehara, Masahito 218
Takemoto, Takashi 52
Takeuchi, Toshihiko 204
Takida, Yuma 108
Talegaonkar, M. 120
Tamai, Takanori 132
Tamakoshi, Akira 202
Tamura, Hirotaka 126
Tan, A. 120
Tan, Kai-Wen 118
Tan, Zhichao 296
Tanabe, Jun 132
Tanabe, Yasuki 132
Tanaka, Masamitsu 460
Tanaka, Tsuyoshi 104
Tandon, Pulkit 226
Tang, Chih-Chun 350
Tang, Kea-Tiong 388, 396
Tang, Kerry 482
Tang, Litian 264
Tang, Nghia 152
Tang, Xiyuan 64, 296
Tang, Yangyang 152
Tanigawa, Takaho 202
Tanzi, Nebil 352
Tavernier, Filip 62
Thakkar, Chintan 172
Theilmann, Paul 362
Tianting, Zhao 244
Tien, Kevin 198
Timalsina, Yukta 464
Timbadiya, Jaykant 48
Timurdogan, Erman 464
Tiruvur, A. 120
Toi, Takashi 478
Toifl, Thomas 112, 476
Tokgoz, Korkut Kaan 344
Tokiwa, Naoya 218
Tol, Jos van der 29
Tonietto, Davide 114, 482
Toprak-Deniz, Zeynep 122
Torf, Tom 166
Traferro, Stefano 166
Trolier-McKinstry, Susan 426
Tsai, Chih-Hou 330
Tsai, Hann-Huei 434
Tsai, Jen-Che 338
Tsai, Jen-Huan 338
Tsai, Kun-Hung 118
Tsai, Tsung-Yen 434, 466
Tsai, Tzung-Chuen 162
Tsai, Wei-Hao 118
Tschanz, James 146, 154, 234
Tse, L. 120
Tseng, Tai-yuan 218
Tsou, Chieh 372
Tsou, Shan-Chih 330
Tsubouchi, uta 478
Tu, Yung-Ning 138, 396
Tulisi, Stefano 272
Tung, Chun-Chao 162
Tyler, Max 106

U

Uchiyama, Masato 132
Uda, M. 210
Uemura, Minoru 54
Ueno, Takeshi 252
Uggetti, Paola 116
Unno, Masaki 218
Urakawa, Go 478
Usuda, Manabu 104

V

Vaesen, Kristof 168
Vainsencher, Amit 456
Valiante, Taufik A. 374
Vangapaty, Ajay 212, 214
Vaz, Struan 164
Vemparala, Divya Shree 174
Venkatachala, Praveen Kumar 70
Vera, Leonardo 484
Verhelst, Marian 44, 62
Verhoef, Bram 304
Verma, Kapil 218
Verma, Naveen 280, 374
Vianello, Elisa 226
Videv, Stefan 468
Vis, Peter 360, 368
Visweswaran, Akshay 168
Vivet, Pascal 226

W

Wadatsumi, Junji 478
Wagner, Sigurd 280
Walker, Richard 468
Walling, Jeffrey Sean 78
Wambacq, Piet 168
Wang, Bo 394, 482
Wang, Caiyi 350
Wang, Chi-Yun 338
Wang, Dixiong 426
Wang, Fei 88
Wang, Guoxing 370
Wang, Haojie 94
Wang, Hua 88, 90, 162, 346
Wang, Huan 450
Wang, Jing-Hong 388, 396
Wang, Jingcheng 224
Wang, Pai-Yi 430
Wang, Po-Han Peter 440
Wang, Sensen Li, Hua 84
Wang, Ting-Yang 334
Wang, Wei 340
Wang, Xiaowei 224
Wang, Xiaoyong 94
Wang, Xin 94
Wang, Yanjie 174
Wang, Yanzhi 138
Wang, Yun 256, 344
Wang, Zhaowen 266
Wang, Zhen 394
Wang, Zhibo 138
Wang, Zisong 450
Warchall, Julian 362
Watanabe, Keiki 182
Watanabe, Mitsuyuki 218
Watanabe, Toshinari 202
Watts, Michael R. 464
Wegberg, Roland van 360, 368
Wei, Liqiong 212, 214
Wei, Mark 162
Wei, Mian 102
Wei, Shih-Ying 388
Wei, Wei-Chen 388, 396
Weisspapir, Iliya 374
Welland, Dave 164
Wen, Shon-Hang 294
Weng, Chan-Hsiang 334
Weng, Po-Shuan 118
Weng, Sheldon 146, 154, 234
Wenhui, Qin 244
Wentzloff, David 442
White, Ted 456
Wicht, Bernhard 156
Wiegand, Chris 214
Williams, Christopher 484
Williams, Kevin 29
Williams, Randy 296
Wong, H.-S. Philip 226
Wong, Henry 114
Wong, J. 120
Wood, Nicholas 192
Wootters, Mary K. 226
Wu, Bart 264
Wu, Chien-Hua 118
Wu, Dong 402
Wu, Hao 266
Wu, Huaqiang 402
Wu, Ke-Chung 118
Wu, Min-Hua 350
Wu, Rui 256, 344
Wu, Ssu-Yen 396
Wu, Tony F. 226
Wu, Xiao 96

X

Xia, Zhengfan 102
Xiang, Bo 320
Xiao, Pengge 94
Xie, Cheng-Yu 242
Xie, Xiao 94
Xiong, Liang 76
Xu, Chen 94
Xu, Hao 266
Xu, Hongtao 76
Xu, Jiawei 360, 368
Xu, Li 96
Xu, Pengcheng 322
Xu, Rongjin 414
Xu, Zhiwei 452
Xue, Cheng-Xin 388
Xue, Jing 158
Xue, Pan 420
Xue, Tiancheng 426

Y

Yaacobi, Ami 464
Yada, Satish 48
Yakubo, Yuto 204
Yalcin, Cem 284
Yamada, Shuhei 172
Yamada, Yutaka 132
Yamahira, Seiji 104
Yamaoka, Masanao 52
Yamaoka, Masashi 218
Yamashita, Minoru 218
Yamashita, T. 210
Yamauchi, Tadaaki 54
Yamawaki, Taizo 182
Yamazaki, Shunpei 204
Yan, Na 76
Yanagidaira, K. 210
Yanagimachi, hinya 462
Yanagita, Tomonori 54
Yang, Dihang 266
Yang, Huazhong 138
Yang, Jun 394
Yang, Kaiyuan 400
Yang, Shang-Hsien 242
Yang, Shiheng 270
Yang, Xiangxing 296
Yang, Xiaofeng 260
Yang, Xin 246
Yang, Zunsong 270
Yao, Chih-Wei 354
Yap, Alex 218
Yasuda, Makoto 314
Yasuhira, Mitsuo 202
Ye, Dawei 414
Yeo, Hong Goo 426
Yi, Shengyu 402
Yin, Shouyi 394
Yin, Yun 76
Yingjie, Guo 244
Yingling, Daniel 316
Yiu, Chris 94
Yokoyama, Sayuri 108
Yoo, Danny 126
Yoo, Hoi-Jun 20, 136, 142, 190, 288, 364
Yoo, Jerald 190
Yoo, Seyeon 490
Yoon, Chiweon 216
Yoon, Euisang 216
Yoon, Heein 258
Yoon, Hyun 352
Yoshida, Takeshi 170
Yoshikawa, Takashi 132
Yoshimura, Chihiro 52
You, Dongwon 344
You, Kae-Dyi 192
Youn, Yongsik 238
Younce, Rick 484
Young, Christopher 98
Yousry, Ramy 118
Yu, Changyeon 216
Yu, Shimeng 396, 402
Yuan, Wen 78
Yuan, Zhe 138
Yuanyuan, hao 244
Yue, C. Patrick 488
Yue, Jinshan 138
Yue, Zhuo 244
Yueksel, Hazar 476
Yuh, Jong 218
Yun, Jaewoong 384
Yun, Sewon 216

Z

Zeng, Lei 370
Zeng, Xiaoyang 424
Zhan, Jing-Hong Conan 162
Zhan, Yan 166
Zhang, Aoyang 74
Zhang, Dan 320
Zhang, Haosheng 462
Zhang, M. 120
Zhang, Minglei 60, 66
Zhang, Wenbing 230
Zhang, Yiqun 314
Zhang, Zhao 488
Zhao, Tianting 246
Zhao, Wenda 64, 296
Zhao, Yu 264
Zhao, Yuanyuan 246
Zheng, Boyi 332
Zheng, Zhiwu 280
Zhou, J. 210
Zhou, Qiang 162
Zhou, Xiahan 196
Zhou, Zhiyuan 152
Zhu, Guang 488
Zhu, Yan 60, 66, 260, 340
Zomagboguelou, Wilfried 166
Zuo, Yongrong 354
Zwerg, Michael 282

CONFERENCE CHAIR
Jan van der Spiegel
University of Pennsylvania
Philadelphia, PA

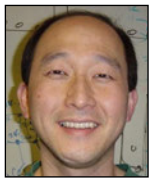
SENIOR TECHNICAL ADVISOR
Anantha Chandrakasan
Massachusetts Institute of Technology
Cambridge, MA

EXECUTIVE COMMITTEE SECRETARY, DATA TEAM, SRP CHAIR
PRESS COORDINATOR
Denis Daly
Omni Design Technologies

WEB SITE AND A/V CHAIR
Trudy Stetzler
Houston, TX

DIRECTOR OF FINANCE
John Weinmann
Webster, NY

PROGRAM CHAIR
Eugenio Cantatore
Eindhoven University of Technology
Eindhoven, The Netherlands



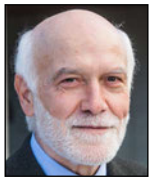
PROGRAM VICE-CHAIR
Un-Ku Moon
Oregon State University
Corvallis, OR

DEMONSTRATION SESSION CHAIR
Uming Ko
MediaTek
Austin, TX

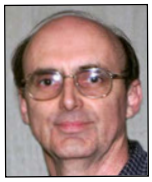
ITPC FAR-EAST REGIONAL CHAIR
Tai-Cheng Lee
National Taiwan University
Taipei, Taiwan

ITPC FAR-EAST REGIONAL VICE-CHAIR
Makoto Takamiya
University of Tokyo
Tokyo, Japan

ITPC EUROPEAN REGIONAL CHAIR
Kostas Doris
NXP
Eindhoven, The Netherlands



ITPC EUROPEAN REGIONAL VICE CHAIR
Yiannos Manoli
University of Freiburg - IMTEK
Freiburg, Germany



ADCOM REPRESENTATIVE
Bryan Ackland
Old Bridge, NJ

DIRECTOR OF PUBLICATIONS & PRESENTATIONS
Laura Fujino
University of Toronto
Toronto, Canada

PRESS LIAISON AND ARC CHAIR
Kenneth C. Smith
University of Toronto
Toronto, Cananda

FORUMS CHAIR
Andreia Cathelin
STMicroelectronics
Crolles, France

EDUCATION CHAIR (SHORT COURSE-TUTORIALS)
Ali Sheikholeslami
University of Toronto
Toronto, Canada

DIRECTOR OF OPERATIONS
Melissa Widerkehr
Widerkehr and Associates
Montgomery Village, MD

TECHNICAL EDITORS

Jason H. Anderson, University of Toronto, Toronto, Canada
Leonid Belostotski, The University of Calgary, Calgary, Canada
Dustin Dunwell, Huawei Technologies, Markham, Canada
Vincent Gaudet, University of Waterloo, Waterloo, Canada
Glenn Gulak, University of Toronto, Toronto, Canada
James W. Haslett, The University of Calgary, Calgary, Canada
David Halupka, Intel, Toronto, Canada
Kenneth C. Smith, University of Toronto, Toronto, Canada

MULTI-MEDIA COORDINATOR

David Halupka, Intel, Toronto, Canada

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

PROGRAM CHAIR: ***Eugenio Cantatore,*** Eindhoven University of Technology, Eindhoven, The Netherlands
PROGRAM VICE CHAIR: ***Un-Ku Moon,*** Oregon State University Corvallis, OR

Analog Subcommittee

Chair: Kofi Makinwa, Delft University of Technology, Delft, The Netherlands
David Blaauw, University of Michigan, Ann Arbor, MI
Youngcheol Chae, Yonsei University, Seoul, Korea
Mahdi Kashmiri, Robert Bosch Research and Technology Center, Sunnyvale, CA
Byungsub Kim, POSTECH, Pohang, Korea
Taeik Kim, Samsung Electronics, Hwaseong, Korea
Man-Kay Law, University of Macau, Macau, China
Yiannos Manoli, University of Freiburg - IMTEK, Freiburg, Germany
Mike Perrott, Nashua, NH
Tim Piessens, ICsense, Leuven, Belgium
Wen-Chieh Wang, MediaTek, Hsinchu, Taiwan

Data Converters Subcommittee

Chair: Michael Flynn, University of Michigan at Ann Arbor, Ann Arbor, MI
Ahmed Ali, Analog Devices, Greensboro, NC
Yun Chiu, University of Texas at Dallas, Ricardson, TX
Kostas Doris, NXP, Eindhoven, The Netherlands
John Keane, Keysight Technologies, Santa Clara, CA
Tai-Cheng Lee, National Taiwan University, Taipei, Taiwan
Dominique Morche, MINATEC Campus, Grenoble, France
Takashi Oshima, Hitachi, Tokyo, Japan
Yun-Shiang Shu, Mediatek, Hsinchu, Taiwan
Venkatesh Srinivasan, Texas Instruments, Dallas, TX
Matt Straayer, Maxim Integrated Products, North Chelmsford, MA
Seng-Pan (Ben) U, University of Macau, Macau, China
Bob Verbruggen, Xilinx, Dublin, Ireland
Jan Westra, Broadcom, Bunnik, The Netherlands

Digital Architectures & Systems Subcommittee

Chair: Byeong-Gyu Nam, Chungnam National University, Daejeon, Korea
Interim Chair: Thomas Burd, Advanced Micro Devices, Santa Clara, CA
Hsie-Chia Chang, National Chiao Tung University, HsinChu, Taiwan
Christopher Gonzalez, IBM, Yorktown Heights, NY
Wookyeong Jeong, Samsung, Hwasung, Korea
Muhammad Khellah, Intel, Hillsboro, OR
Dejan Markovic, University of California, Los Angeles, Los Angeles, CA
Mahesh Mehendale, Texas Instruments, Dallas, TX
Masato Motomura, Hokkaido University, Sapporo, Japan
James Myers, ARM, Cambridge, United Kingdom
Ingrid Verbauwhede, KU Leuven, Leuven, Belgium

Digital Circuits Subcommittee

Chair: Edith Beigné, Facebook, Menlo Park, CA
Massimo Alioto, National University of Singapore, Singapore
Keith Bowman, Qualcomm, Raleigh, NC
Vivek De, Intel, Hillsboro, OR
Koji Hirairi, Sony, Atsugi, Japan
Alicia Klinefelter, NVidia, Durham, NC
John Maneatis, True Circuits, Los Altos, CA
Phillip Restle, IBM T. J. Watson Research Center, Yorktown Heights, NY
Mingoo Seok, Columbia University, New York, NY
Youngmin Shin, Samsung, Hwansung, Korea
Hirofumi Shinohara, Waseda University, Kitakyushu, Japan
Dennis Sylvester, University of Michigan, Ann Arbor, MI
Ping-Ying Wang, CMOS-Crystal, Hsinchu, Taiwan

IMMD Subcommittee

Chair: Chris Van Hoof, imec, Leuven, Belgium
Joonsung Bae, Kangwon National University, Chuncheon, Korea
Gert Cauwenberghs, University of California, San Diego, La Jolla, CA
Calvin Yi-Ping Chao, TSMC, Zhubei, Taiwan
Yoon-Kyung Choi, Samsung, Hwaseong, Korea
Jun Deguchi, Toshiba Memory Corporation, Kawasaki, Japan
Keith Fife, 4Catalyzer, Palo Alto, CA
Morris (Ming-Dou) Ker, National Chiao-Tung University, Hsinchu, Taiwan
Michael Kraft, KULeuven, Leuven, Belgium
Pedram Lajevardi, Robert Bosch Research and Technology Center, Sunnyvale, CA
Masayuki Miyamoto, Wacom, Tokyo, Japan
Pedram Mohseni, Case Western Reserve University, Cleveland, OH
Rikky Muller, University of California, Berkeley, Berkeley, CA
Kazuko Nishimura, Panasonic, Moriguchi, Japan
Matteo Perenzoni, Fondazione Bruno Kessler (FBK), Trento, Italy
Bruce Rae, ST Microelectronics, Edinburgh, United Kingdom
Esther Rodriguez-Villegas, Imperial College London, London, United Kingdom
Hayato Wakabayashi, Sony Semiconductor Solutions, Atsugi, Japan

Memory Subcommittee

Chair: Jonathan Chang, TSMC, Hsinchu, Taiwan
Seung-Jun Bae, Samsung, Hwasung, Korea
Meng-Fan Chang, National Tsing Hua University, Hsinchu, Taiwan
Fatih Hamzaoglu, Intel, Hillsboro, OR
Kyu-Hyoun (KH) Kim, IBM T. J. Watson, Yorktown Heights, NY
Dong Uk Lee, SK hynix, Icheon, Korea
Yan Li, Western Digital, Milpitas, CA
Jongmin Park, SK hynix, Icheon, Korea
Ki-Tae Park, Samsung, Hwasung, Korea
Chun Shiah, Etron, Hsinchu, Taiwan
Shinichiro Shiratake, Toshiba Memory, Yokohama, Japan
Wolfgang Spirkl, Micron Semiconductor Deutschland GmbH, Munich, Germany
Yasuhiko Taito, Renesas, Tokyo, Japan

Power Management Subcommittee

Chair: Yogesh Ramadass, Texas Instruments, San Jose, CA
Min Chen, Analog Devices, Milpitas, CA
Chan-Hong Chern, TSMC, Hsinchu, Taiwan
Yuan Gao, IME, A*STAR, Singapore, Singapore
Zhiliang Hong, Fudan University, Shanghai, China
Yen-Hsun Hsu, Mediatek, HsinChu, Taiwan
Johan Janssens, ON Semiconductor, Mechelen, Belgium
Hoi Lee, University of Texas, Dallas, Richardson, TX
Robert Pilawa, University of California, Berkeley, Berkeley, CA
Jason Stauth, Dartmouth College, Hanover, NH
Makoto Takamiya, University of Tokyo, Tokyo, Japan
Chen-Kong Teh, Toshiba Semiconductor, Kawasaki, Japan
Gerard Villar-Pique, NXP Semiconductor, Eindhoven, The Netherlands
Bernhard Wicht, University of Hannover, Hannover, Germany

RF Subcommittee

Chair: Piet Wambacq, imec, Heverlee, Belgium
Andrea Bevilacqua, University of Padova, Padova, Italy
Jaehyouk Choi, Ulsan National Institute of Science Technology, Ulsan, Korea
Krzysztof Dufrêne, Intel, Linz, Austria
Minoru Fujishima, Hiroshima University, Higashihiroshima, Japan
Xiang Gao, Zhejiang University, Hangzhou, China
Brian Ginsburg, Texas Instruments, Dallas, TX
Giuseppe Gramegna, Radio Frequency Systems (RFS), Golfe-Juan, France
Ramesh Harjani, University of Minnesota, Minneapolis, MN
Payam Heydari, UCI, Irvine, CA
Abbas Komijani, Apple, Mountain View, CA
Harish Krishnaswamy, Columbia University, New York, NY
John Long, University of Waterloo, Waterloo, Canada
Andrea Mazzanti, Università degli Studi di Pavia, Pavia, Italy
Kohei Onizuka, Toshiba, Kawasaki, Japan
Jiayoon Ru, Broadcom, Irvine, CA
Hua Wang, Georgia Institute of Technology, Atlanta, GA
Conan Zhan, MediaTek, HsinChu, Taiwan

Technology Directions Subcommittee

Chair: Makoto Nagata, Kobe University, Kobe, Japan
Edoardo Charbon, EPFL, Neuchâtel, Switzerland
Wim Dehaene, Kuleuven-MICAS, Leuven, Belgium
Antoine Dupret, CEA, Gif-sur-Yvette, France
Frederic Gianesello, STMicroelectronics, Crolles, France
Kush Gulati, Omni Design Technologies, Milpitas, CA
Pui-In Mak, University of Macau, Macau, China
Patrick Mercier, University of California, San Diego, San Diego, CA
Shahriar Mirabbasi, University of British Columbia, Vancouver, Canada
Shuichi Nagai, Panasonic, Moriguchi, Japan
Sriram Vangal, Intel, Hillsboro, OR
Nick Van Helleputte, imec, Leuven, Belgium
Naveen Verma, Princeton University, Princeton, NJ
Long Yan, Samsung Electronics, Hwaseong, Korea

Wireless Subcommittee

Chair: Stefano Pellerano, Intel, Hillsboro, OR
Matteo Bassi, Infineon Technologies, Villach, Austria
Theodoros Georgantas, Broadcom, Athens, Greece
Danielle Griffith, Texas Instruments, Dallas, TX
Xin He, NXP, Eindhoven, The Netherlands
Nagendra Krishnapura, Indian Institute of Technology Madras, Chennai, India
Kyoo Hyun Lim, FCI, Seongnam, Korea
Yao-Hong Liu, imec, Eindhoven, The Netherlands
Howard C. Luong, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
Hideaki Majima, Toshiba, Kawasaki, Japan
David McLaurin, Analog Devices, Raleigh, NC
Arun Natarajan, Oregon State University, Corvallis, OR
Sudhakar Pamarti, University of California, Los Angeles (UCLA), Los Angeles, CA
Yiwu Tang, Qualcomm Technologies, San Diego, CA
Yuu Watanabe, Waseda University, Atsugi, Japan
Renaldi Winoto, Tectus, Saratoga, CA
Alan Chi-Wai Wong, EnSilica, Oxford, United Kingdom
Ken Yamamoto, Sony, Atsugi, Japan

Wireline Subcommittee

Chair: Frank O'Mahony, Intel, Hillsboro, OR
Amir Amirkhany, Samsung Electronics, San Jose, CA
Hyeon-Min Bae, KAIST, Daejong, Korea
Tony Chan Carusone, University of Toronto, Toronto, Canada
Mike Shuo-Wei Chen, University of Southern California, Los Angeles, CA
Yohan Frans, Xilinx, San Jose, CA
Friedel Gerfers, Technische Universität Berlin, Berlin, Germany
Andrew Joy, Marvell, Irvine, CA
Jaeha Kim, Seoul National University, Seoul, Korea
Mounir Meghelli, IBM Thomas J Watson Research Center, Yorktown Heights, NY
Sam Palermo, Texas A&M University, College Station, TX
Takayuki Shibasaki, Fujitsu Laboratories, Kawasaki, Japan
Bo Zhang, Broadcom, Irvine, CA

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

EUROPEAN REGIONAL SUBCOMMITTEE

ITPC EUROPEAN REGIONAL CHAIR
Kostas Doris, NXP, Eindhoven, The Netherlands

ITPC EUROPEAN REGIONAL VICE CHAIR
Yiannos Manoli, University of Freiburg - IMTEK, Freiburg, Germany

ITPC EUROPEAN REGIONAL SECRETARY
Tim Piessens, ICsense, Leuven, Belgium

MEMBERS

Matteo Bassi, Infineon Technologies, Villach, Austria
Edith Beigné, Facebook, Menlo Park, CA
Andrea Bevilacqua, University of Padova, Padova, Italy
Edoardo Charbon, EPFL, Neuchâtel, Switzerland
Wim Dehaene, Kuleuven-MICAS, Leuven, Belgium
Krzysztof Dufrêne, Intel, Linz, Austria
Antoine Dupret, CEA, Gif-sur-Yvette, France
Theodoros Georgantas, Broadcom, Athens, Greece
Friedel Gerfers, Technische Universität Berlin, Berlin, Germany
Frederic Gianesello, STMicroelectronics, Crolles, France
Giuseppe Gramegna, Radio Frequency Systems (RFS), Golfe-Juan, France
Xin He, NXP, Eindhoven, The Netherlands
Johan Janssens, ON Semiconductor, Mechelen, Belgium
Andrew Joy, Marvell, Irvine, CA
Michael Kraft, KULeuven, Leuven, Belgium
Yao-Hong Liu, imec, Eindhoven, The Netherlands
Kofi Makinwa, Delft University of Technology, Delft, The Netherlands
Andrea Mazzanti, Università degli Studi di Pavia, Pavia, Italy
Dominique Morche, MINATEC Campus, Grenoble, France
James Myers, ARM, Cambridge, United Kingdom
Matteo Perenzoni, Fondazione Bruno Kessler (FBK), Trento, Italy
Bruce Rae, ST Microelectronics, Edinburgh, United Kingdom
Esther Rodriguez-Villegas, Imperial College London, London, United Kingdom
Wolfgang Spirkl, Micron Semiconductor Deutschland GmbH, Munich, Germany
Nick Van Helleputte, imec, Leuven, Belgium
Chris Van Hoof, imec, Leuven, Belgium
Ingrid Verbauwhede, KU Leuven, Leuven, Belgium
Bob Verbruggen, Xilinx, Dublin, Ireland
Gerard Villar-Pique, NXP Semiconductor, Eindhoven, The Netherlands
Piet Wambacq, imec, Heverlee, Belgium
Jan Westra, Broadcom, Bunnik, The Netherlands
Bernhard Wicht, University of Hannover, Hannover, Germany
Alan Chi-Wai Wong, EnSilica, Oxford, United Kingdom

FAR EAST REGIONAL SUBCOMMITTEE

ITPC FAR EAST REGIONAL CHAIR
Tai-Cheng Lee, National Taiwan University, Taipei, Taiwan

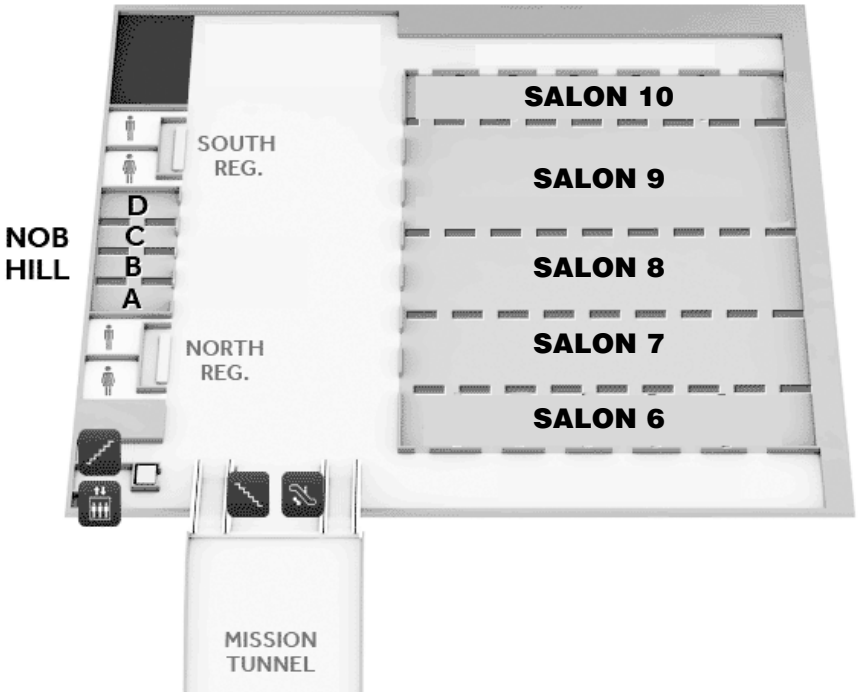
ITPC FAR EAST REGIONAL VICE-CHAIR
Makoto Takamiya, University of Tokyo, Tokyo, Japan

ITPC FAR EAST REGIONAL SECRETARY
Long Yan, Samsung Electronics, Hwaseong, Korea

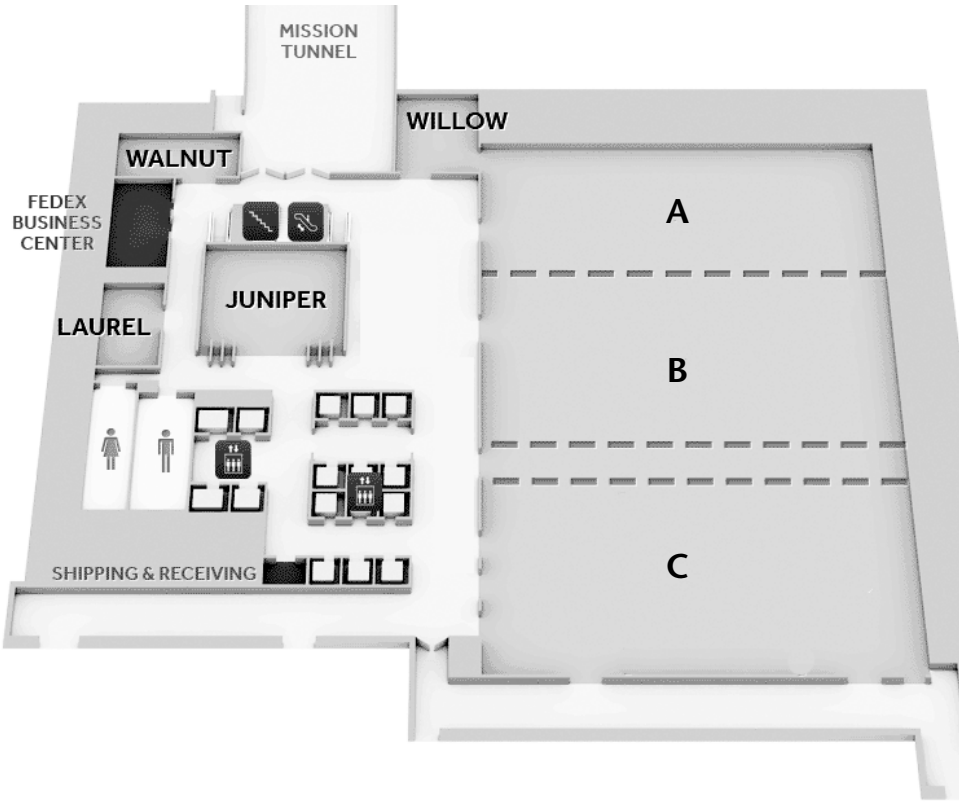
MEMBERS

Massimo Alioto, National University of Singapore, Singapore
Hyeon-Min Bae, KAIST, Daejong, Korea
Joonsung Bae, Kangwon National University, Chuncheon, Korea
Seung-Jun Bae, Samsung, Hwasung, Korea
Youngcheol Chae, Yonsei University, Seoul, Korea
Hsie-Chia Chang, National Chiao Tung University, HsinChu,Taiwan
Jonathan Chang, TSMC, Hsinchu, Taiwan
Meng-Fan Chang, National Tsing Hua University, Hsinchu, Taiwan
Calvin Yi-Ping Chao, TSMC, Zhubei, Taiwan
Chan-Hong Chern, TSMC, Hsinchu, Taiwan
Jaehyouk Choi, Ulsan National Institute of Science Technology, Ulsan, Korea
Yoon-Kyung Choi, Samsung, Hwaseong, Korea
Jun Deguchi, Toshiba Memory Corporation, Kawasaki, Japan
Minoru Fujishima, Hiroshima University, Higashihiroshima, Japan
Yuan Gao, IME, A*STAR, Singapore, Singapore
Koji Hirairi, Sony, Atsugi, Japan
Zhiliang Hong, Fudan University, Shanghai, China
Yen-Hsun Hsu, Mediatek, HsinChu, Taiwan
Wookyeong Jeong, Samsung, Hwasung, Korea
Morris (Ming-Dou) Ker, National Chiao-Tung University, Hsinchu, Taiwan
Byungsub Kim, POSTECH, Pohang, Korea
Jaeha Kim, Seoul National University, Seoul, Korea
Taeik Kim, Samsung Electronics, Hwaseong, Korea
Nagendra Krishnapura, Indian Institute of Technology Madras, Chennai, India
Man-Kay Law, University of Macau, Macau, China
Dong Uk Lee, SK hynix, Icheon, Korea
Kyoo Hyun Lim, FCI, Seongnam, Korea
Howard C. Luong, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
Hideaki Majima, Toshiba, Kawasaki, Japan
Pui-In Mak, University of Macau, Macau, China
Masayuki Miyamoto, Wacom, Tokyo, Japan
Masato Motomura, Hokkaido University, Sapporo, Japan
Shuichi Nagai, Panasonic, Moriguchi, Japan
Makoto Nagata, Kobe University, Kobe, Japan
Byeong-Gyu Nam, Chungnam National University, Daejeon, Korea
Kazuko Nishimura, Panasonic, Moriguchi, Japan
Kohei Onizuka, Toshiba, Kawasaki, Japan
Takashi Oshima, Hitachi, Tokyo, Japan
Jongmin Park, SK hynix, Icheon, Korea
Ki-Tae Park, Samsung, Hwasung, Korea
Chun Shiah, Etron, Hsinchu, Taiwan
Takayuki Shibasaki, Fujitsu Laboratories, Kawasaki, Japan
Youngmin Shin, Samsung, Hwansung, Korea
Hirofumi Shinohara, Waseda University, Kitakyushu, Japan
Shinichiro Shiratake, Toshiba Memory, Yokohama, Japan
Yun-Shiang Shu, Mediatek, Hsinchu, Taiwan
Yasuhiko Taito, Renesas, Tokyo, Japan
Chen-Kong Teh, Toshiba Semiconductor, Kawasaki, Japan
Seng-Pan (Ben) U, University of Macau, Macau, China
Hayato Wakabayashi, Sony Semiconductor Solutions, Atsugi, Japan
Ping-Ying Wang, CMOS-Crystal, Hsinchu, Taiwan
Wen-Chieh Wang, MediaTek, Hsinchu, Taiwan
Yuu Watanabe, Waseda University, Atsugi, Japan
Ken Yamamoto, Sony, Atsugi, Japan
Conan Zhan, MediaTek, HsinChu, Taiwan

Lower B2 Level - Yerba Buena Ballroom



B2 Level - Golden Gate Hall



 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

NEW: Subcommittee Focus on Machine Learning

ISSCC 2020 Call for Papers

IEEE International Solid-State Circuits Conference

Sunday – Thursday, February 16 - 20, 2020

San Francisco Marriott Marquis Hotel, San Francisco, CA

ISSCC 2020 CONFERENCE THEME:

INTEGRATED CIRCUITS POWERING THE AI ERA

Innovative and original papers are solicited in subject areas including (but not limited to) the following:

ANALOG: Amplifiers, comparators, oscillators, filters, references; nonlinear analog circuits; digitally-assisted analog circuits; MEMS/sensor interface circuits.

DATA CONVERTERS: Nyquist-rate and oversampling A/D and D/A converters; embedded and application-specific A/D and D/A converters; analog to information conversion; time-to-digital converters.

DIGITAL ARCHITECTURES & SYSTEMS: Microprocessors, micro-controllers, applications processors, graphics processors; digital systems for communications, video and multimedia, machine-learning, deep-learning, neuromorphism, cryptography, security and trusted computing, special-function acceleration, processing-in-memory; reconfigurable systems, near- and sub-threshold systems, digital architectures and systems for emerging applications (e.g. virtual reality – AR/VR and autonomous vehicles).

DIGITAL CIRCUITS: Building blocks for 2D/3D SoC such as intra-chip communication circuits, clock distribution techniques, soft-error and variation-tolerant circuits. Circuits for power management in digital applications: including voltage regulators, adaptive digital circuits, digital sensors; Near- and sub-threshold circuits; PLLs for digital clocking applications. Circuits for neuro-computing; Hardware security circuits including PUFs, TRNG, and attack-detection sensors.

IMAGERS, MEMS, MEDICAL, & DISPLAY: Image sensors and SoCs; automotive, LIDAR, and ultrasonic sensors; MEMS sensor systems; wearable, implantable, ingestible electronics, biomedical SoCs, neural interfaces and closed-loop systems; biosensors, microarrays, and lab-on-a-chip; display electronics, displays with sensing functionality; sensing for AR/VR.

MEMORY: Static, dynamic, and non-volatile memories for stand-alone and embedded applications; memory/SSD controllers; high-bandwidth I/O interfaces; memories based on phase-change, magnetic, spin-transfer-torque, ferroelectric, and resistive materials; array architectures and circuits to improve low-voltage operation, power reduction, reliability, and fault tolerance; memory-subsystem enhancements, including in-memory logic functions, machine learning, artificial intelligence, and AR/VR.

POWER MANAGEMENT: Power control and management circuits, regulators; switched-mode power converter ICs using inductive, capacitive, and hybrid techniques; energy-harvesting circuits and systems; wide-bandgap topologies and gate-drivers; power and signal isolators; circuits for lighting, wireless power.

RF CIRCUITS and WIRELESS SYSTEMS*: Building blocks and complete solutions at RF, mm-Wave and THz frequencies for receivers, transmitters, frequency synthesizers, transceivers, SoCs, and SiPs. Innovative circuit-level and system-architecture solutions for established wireless standards and future systems or applications such as radar, sensing, and imaging.

TECHNOLOGY DIRECTIONS: Emerging IC and system solutions for: biomedical, sensor interfaces, analog signal processing, power management, computation, data storage, and communication; analog/mixed-signal techniques for security and machine learning; non-silicon-, carbon-, organic-, metal-oxide-, compound-semiconductor- and new-device-based circuits; nano, flexible, large-area, stretchable, printable, quantum, optical, and 3D-integrated electronics; spintronics.

WIRELINE: Receivers/transmitters/transceivers for wireline systems, including backplane transceivers, optical links, chip-to-chip communications, 2.5/3D interconnect, copper-cable links, and equalizing on-chip links; exploratory I/O circuits for advancing data rates, power efficiency, and equalization; wireline transceiver building blocks (such as AGCs, front ends, equalizers, clock-generation and distribution circuits including PLLs, clock-and-data recovery, line drivers, and hybrids).

**Papers submitted to this category will be reviewed by either the RF or Wireless Subcommittee.*

Submission Deadline is Monday, September 9, 2019 • 3:00PM Eastern Daylight Time (19:00 GMT)

STUDENT INITIATIVES

Graduate students are invited to participate in opportunities to showcase ongoing work and exchange experiences with other students and researchers from academia and industry. These include the Student Research Preview and the Silkroad Award (to a first-time student presenting author of a regular paper from an emerging region in the Far East).

Further information including submission procedures, formats, student initiatives and deadlines can be found at http://www.isscc.org

ISSCC 2019 • Sunday February 17th

Tutorials

8:30 AM	**T1:** Fundamentals of Integrated Radars	**T2:** Fundamentals of Power-Conversion Topologies	**T3:** Advances and Prospects for In-Memory Computing
10:30 AM	**T4:** Fundamentals of Efficient IoT Microcontrollers	**T5:** Noise-Shaping in Data Converters	**T6:** Basics of Clock-and-Data-Recovery Circuits
1:30 PM	**T7:** Hardware Security – from Basics to ASICs	**T8:** Current-Sensing Techniques	
3:30 PM	**T9:** Calibration Techniques for Wireless Transceivers	**T10:** Low-Noise Sensor Interfaces	

Forums

8:00 AM	**F1:** Sub-6GHz 5G Radio Circuits and Systems: From Concepts to Silicon	**F2:** Memory-Centric Computing from IoT to Artificial Intelligence and Machine Learning

Events Below in Bold Box are Included with your Conference Registration

Evening Events

	7:30 PM EE1: Student Research Preview: Short Presentations with Poster Session	**6:00 PM EE2:** How to Save Lives With Circuits

ISSCC 2019 • Monday February 18th • Paper Sessions

8:30 AM	**Session 1:** Plenary Session				
1:30 PM	**Session 2:** Processors	**Session 3:** Nyquist-Rate ADCs	**Session 4:** Power Amplifiers	**Session 5:** Image Sensors	**Session 6:** Ultra-High-Speed Wireline

12noon to 7:00 PM – Book Displays • 5:00 PM to 7:00 PM – Demonstration Session • 5:15 PM – Author Interviews • Social Hour

Evening Events

	7:00 PM EE3: Making a Career Choice	**8:00 PM EE4:** Industry Showcase

ISSCC 2019 • Tuesday February 19th • Paper Sessions

8:30 AM	**Session 7:** Machine Learning	**Session 8:** DC-DC Converters	**Session 9:** High-Frequency Transceivers for Radar and Communications	**Session 10:** Sensor Interfaces **Session 11:** Diagnostics	**Session 12:** Emerging Technologies **Session 13:** Non-Volatile Memories
1:30 PM	**Session 14:** Machine Learning & Digital LDO Circuits	**Session 15:** Power for 5G, Wireless Power, and GaN Converters	**Session 16:** Frequency Synthesizers	**Session 17:** Technologies for Human Interaction & Health	**Session 18:** Analog Techniques

10:00 AM to 7:00 PM – Book Displays • 5:00 PM to 7:00 PM – Demonstration Session • 5:15 PM – Author Interviews • Social Hour

Evening Events

8:00 PM	**EE5:** Moving to 'The Dark Side'!	**EE6:** How Can Hardware Designers Reclaim the Spotlight?

ISSCC 2019 • Wednesday February 20th • Paper Sessions

8:30 AM	**Session 19:** Adaptive Digital & Clocking Techniques	**Session 20:** Noise-Shaped & VCO-Based ADCs	**Session 21:** 4G/5G Transceivers	**Session 22:** Physiological Monitoring	**Session 23:** DRAM **Session 24:** SRAM & Computation-in-Memory
1:30 PM	**Session 25:** Circuits Enabling Security **Session 26:** Frequency Generation & Interference Mitigation	**Session 27:** Energy Harvesting & DC/DC Control Techniques	**Session 28:** Techniques for Low-Power & High-Performance Wireless	**Session 29:** Quantum & Photonics Technologies	**Session 30:** Advanced Wireline Techniques

10:00 AM to 3:00 PM – Book Displays • 5:15 PM – Author Interviews

ISSCC 2019 • Thursday February 21st

8:00 AM	**Short Course:** Integrated Phased Arrays: Theory, Practice, and Implementation for 5G and Beyond	**F3:** The Complete Advanced Driver-Assistance System (ADAS) Sensing Network	**F4:** Intelligence at the Edge: How Can We Make Machine Learning More Energy Efficient?	**F5:** 56Gb/s to 112Gb/s and Beyond – Design Challenges and Solutions in Wireline Communications	**F6:** The Right Tool for the Job: Application-Optimized Data Converters

978-1-5386-8531-0/19/$31.00 ©2019 IEEE

INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS

ISSCC®

UNIVERSITY OF PENNSYLVANIA

IEEE®

SSCS