SUNDAY through THURSDAY / FEBRUARY 11, 12, 13, 14, and 15, 2018

2018 IEEE INTERNATIONAL

SOLID-STATE CIRCUITS CONFERENCE

IEEE SOLID-STATE CIRCUITS SOCIETY

2018 DIGEST OF TECHNICAL PAPERS

65

ISSCC®

INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS • UNIVERSITY OF PENNSYLVANIA •

IEEE

SSCS

VOLUME SIXTY-ONE
ISSN 0193-6530

2018 IEEE INTERNATIONAL SOLID-STATE CIRCUITS CONFERENCE

DIGEST OF TECHNICAL PAPERS

First Edition

February 2018

IEEE Catalog Number CFP18ISS-PRT

2018 IEEE International Solid-State Circuits Conference

DIGEST OF TECHNICAL PAPERS

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 61

IEEE Cat. No. CFP18ISS-PRT
ISBN Softbound 978-1-5090-4939-4
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

PAPER SESSIONS

TUTORIALS

FORUMS

EVENING EVENTS

SHORT COURSE

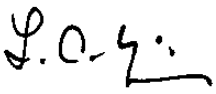
Reflections

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. 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.

Again, this year, we have a technical editorial group (listed below) under the direction of a managing editor (Laura Chizuko Fujino). Once again, we emphasize nearly 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: Alison Burdett, Eugenio Cantatore, 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 (Kapik Integration), 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!



Laura Chizuko Fujino
ISSCC Director of Publications and Presentations

February 2018

 978-1-5090-4940-0/18/$31.00

Foreword: Silicon Engineering a Social World

Welcome to the 65th International Solid-State Circuits Conference! The Conference continues its tradition of showcasing the most-advanced and innovative work from industry and academe around the world, in integrated circuits and systems. The geographical distribution of the accepted technical papers reflects the truly international character of the Conference: 44% of the accepted papers are from North America, 39% are from the Far East, and 17% are from Europe. Of these papers, 60% are from academe, 27% are from industry, 1% are from research institutions/labs, and 12% were joint submissions from industry and academe. Continuing the practice introduced last year, the ISSCC paper selection followed a double-blind review process with anonymized manuscripts, which has emerged as the best-known practice within the broader technical community.

The 2018 Conference theme is "***Silicon Engineering a Social World***." This theme reflects how continued advances in solid-state circuits and systems have brought evermore powerful communication and computational capabilities into mobile form factors. Such ubiquitous smart devices lie at the heart of a revolution shaping how we connect, collaborate, build relationships, and share information. These social technologies allow people to maintain connections and support networks that otherwise would not be possible; they provide the ability to access information instantaneously and from any location, thereby helping to shape the world's events and culture, empowering citizens of all nations, providing social networks allowing worldwide communities to develop and bond with common interests. ISSCC, as the flagship conference in solid-state circuits, will cover the latest advancements surrounding these trends.

The Conference will begin with the Plenary Session, where four innovators will share their views on challenges and opportunities in our field. The first talk by Vince Roche of Analog Devices will discuss challenges for semiconductor innovation within a highly competitive global environment, where business pressures drive companies to avoid risk and to develop new products following a model of incrementalism. Roche argues that such pressures must be resisted if the semiconductor industry is to continue to carry the mantle of technological leadership and maintain a bright future. Technical innovation and new discoveries are the theme of the following talk from Barbara de Salvo of CEA-Leti, which discusses new paradigms for computing based on recent advances in understanding of cognitive and neuro-sciences. By understanding how the human brain processes information through distributed and connected pathways, new models of semiconductor computing are likely to be developed to enable step-change improvements in processing efficiency. The third talk from Yukihiro Kato of Denso, focuses on the role of semiconductor technology in driving, a once in a lifetime transformation in the automotive industry. Technologies once seen to be in the realm of science fiction, such as electric vehicles and autonomous cars, are rapidly becoming a reality. Kato outlines some of the technical challenges on the way to the future 'mobility-enhanced society', and describes semiconductor innovations in progress to overcome them. The final talk from David Patterson of Google and UC Berkeley review 50 years of innovation in computer architecture. From mainframe computers in the 1960s to the dominant RISC architecture of recent times, Patterson gives an entertaining and informative insight into the technical and business constraints which underpinned these rapid developments. With the end of Moore's Law slowing the pace of improvements due to scaling, Patterson describes how this slowdown is actually rejuvenating innovation in computer architectures, as future performance improvements cannot come from scaling alone.

The ISSCC 2018 Technical Program consists of 207 outstanding technical papers distributed over 31 thematic sessions which include, as well, 5 system-focussed invited talks. 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 5 invited talks focus on systems applications linked to silicon innovation: Thomas Cameron from Analog Devices will discuss architectures and technologies for 5G mmWave radio (Monday afternoon); Olivier Temam of Google will present aspects of edge machine-learning processing (Tuesday afternoon); Ashok Krishnamoorthy of Axalume will present insights on the anatomy of a 20Tb/s switch card (Tuesday afternoon); Katsu Watanabe of Fujitsu will outline how IoT sensors and Cloud computing are being used for next-generation food and agriculture production (Tuesday afternoon); Tim Denison of Medtronic will describe research towards the creation of neural co-processors to explore treatments for neurological disorders (Wednesday afternoon).

On Sunday, the 10 Tutorials cover a wide range of topics: *"Low-Jitter PLLs for Wireless Transceivers", "Nonvolatile Circuits for Memory, Logic, and Artificial Intelligence", "Basics of Quantum Computing", "Error-Correcting Codes in 5G/NVM Applications", "Hybrid Design of Analog-to-Digital Converters", "Single-Photon Detection in CMOS", "Basics of Adaptive and Resilient Circuits", "Fundamentals of Switched-Mode Power-Converter Design", "Digital RF Transmitters"* and *"ADC-Based Serial Links: Design and Analysis"*. These Tutorials are given by top experts in the field and offer an opportunity to experience an introduction to and overview of important developments in each topic area.

The goal of our 6 all-day Forums is to update experts on advanced developments in their fields: On Sunday, February 11th, two forums will be held in parallel: *"Intelligent Energy-Efficient Systems at the Edge of IoT"* and *"FinFETs & FDSOI – A Mixed Signal Circuit Designer's Perspective"*.

On Thursday, February 15th, four forums will be offered in parallel: "*Circuits and Architectures for Wireless Sensing, Radar and Imaging", "Circuit and System Techniques for mm-Wave Multi-Antenna Systems", "Advanced Optical Communication: From Devices, Circuits, and Architectures to Algorithms"* and *"Advances in Energy Efficient Analog Design"*. Also on Thursday is an all-day Short Course on "*Hardware Approaches to Machine Learning and Inference*", which will explore design aspects from algorithms to the implementation of efficient architectures for machine learning.

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 "Circuits for Social Good" aims to highlight various ways that circuits designers can help address some of the most important challenges facing society today, ranging from healthcare to energy conservation. The workshop will feature keynote presentations, as well as interactive 'round tables' where attendees can join in group discussions.

On Monday evening, two sessions will be held in parallel. The first panel session "*Figures-of-Merit on Trial"* will probe the weaknesses and strengths of popular analog FOMs in an entertaining and educational way, with one panellist defending the FOM, and another acting as the prosecutor, with the audience acting as the jury. The second session, the *"Industry Showcase"* will highlight how advances in silicon circuits, SoCs, and systems are fueling the most innovative industrial applications and products of the future. The event will feature short presentations, as well as interactive demonstrations from each of the Showcase participants. These participants were chosen through a nomination and voting process by members of the Industry Showcase Committee.

On Tuesday evening, an entertaining and motivating panel discussion *reviews "Lessons Learned – Great Circuits That Didn't Work – (Oops, If Only I Had Known!)"*. As well as provide a forum in which recognized experts share their past mistakes and failures, (and disclose lessons learned), the audience is invited to contribute "learning experiences" (in less than a minute). Also on Tuesday evening, a panel will discuss *"Can Artificial Intelligence Replace My Job? The Dawn of a New IC Industry with AI"*. Diverse experts will share their vision on the daunting new development of AI in our business.

Two Demonstration Sessions, which have become an integral part of the Conference, will be held during the social hours on Monday and Tuesday. These sessions provide an opportunity to witness live demos of devices described in over 50 selected papers, and to engage in face-to-face discussions with the authors. Also, the Student-Research Preview held on Sunday evening will allow graduate students from around the globe to interact with each other and with attendee technical experts from both academe and industry. The students will have the opportunity to briefly introduce their work prior to the subsequent Poster Session, in which they will benefit from attendee feedback.

The high standards that we associate with ISSCC are a result of the diligent volunteer work of the International Technical-Program Committee (ITPC). This year, the ITPC consists of 156 members from industry and academe, organized into eleven technical subcommittees. Each member has spent a significant amount of time reviewing the submitted papers, planning and organizing evening sessions, and educational events, preparing the Advance Program, Press-Kit, and Digest material, and performing session-chair/organizer duties. I am especially indebted to the Subcommittee Chairs for their leadership in overseeing these tasks: Kofi Makinwa (Analog), Un-Ku Moon (Data Converters), Byeong-gyu Nam (Digital Architectures & Systems), Edith Beigné (Digital Circuits), Makoto Ikeda (Imagers/MEMS/Medical/Displays), Leland Chang (Memory), Axel Thomsen (Power Management), Piet Wambacq (RF), Makoto Nagata (Technology Directions), Stefano Pellerano (Wireless), and Frank O'Mahony (Wireline). Also, I must thank the leadership team of the Regional Committees: Marian Verhelst and Kostas Doris from the European Region, and Sungdae Choi and Tai-Cheng Lee from the Far-East Region. Additionally, I have immensely benefited from the help of the ISSCC 2018 Program Vice-Chair, Eugenio Cantatore, and the ISSCC 2017 Program Chair, Boris Murmann.

ISSCC would not be possible without the help of many other individuals that deliver high-quality support work behind the scenes. I want to acknowledge Melissa Widerkehr and Widerkehr and Associates for their invaluable 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. I must also thank 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. Also, special thanks go to Laura Fujino and Kenneth Smith for their unrelenting help with many aspects of the Conference, including the paper submission process, preparation of the Advance Program, the Press Kit, the Digest, and Tutorial and Short Course DVDs, as well as editing and presentation preparation. Also thanks to Denis Daly for his leading role in the Press Kit preparation, 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 SeongHwan Cho for organizing the Student-Research Preview, to Uming Ko for organizing the Demonstration Sessions, to Trudy Stetzler for managing the Conference website and A/V coordination, and to Bryant Griffin for his financial oversight for the Conference.

We are indebted as well to an unusual group of volunteer graduate students from the University of Toronto, who through their individual technical expertise ensure the orderly conduct of the presentations in each session, as well as countless other behind-the-scene activities.

Finally, my special acknowledgement goes to Anantha Chandrakasan, the ISSCC Conference Chair, for his attentive and visionary leadership, and for the many conference improvements that he has initiated and overseen over the past several years. His selfless dedication to maintain excellence in all aspects of ISSCC has been a true inspiration and reward in my years serving on the Technical Program Committee.

I look forward to seeing you at the Conference and wish you an enjoyable ISSCC 2018!

Alison Burdett

ISSCC 2018 International Technical-Program Chair

Session 1 Overview

Plenary Session

Chair: Anantha Chandrakasan
Massachusetts Institute of Technology, Cambridge, MA
ISSCC Conference Chair

Associate Chair: Alison Burdett
Sensium Healthcare, Abingdon, United Kingdom
ISSCC International Technical Program Chair

The Plenary Session will begin with welcome remarks, and an introduction from the Conference Chair, Anantha Chandrakasan. Then, Alison Burdett, the Technical Program Chair, will provide a summary of the 2018 Technical Program. The Plenary Session features four keynote speakers, who are leaders in their respective fields and collectively represent the broad spectrum of our industry. 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 second Plenary talk.

Overall, the four Plenary talks address a wide variety of topical issues:
In today's highly competitive global environment, business pressures are driving the semiconductor industry to avoid risk and to develop new products following a model of incrementalism and consolidation. Vince Roche, CEO of Analog Devices argues that such pressures must be resisted if the semiconductor industry is to continue to carry the mantle of technological leadership and maintain a bright future. He concludes that application challenges such as the spread of pervasive ubiquitous sensing, rapid advances in artificial intelligence, heterogeneous integration, and the continued impact of digitization on virtually every industry on earth, will require more, not less, semiconductor innovation.

Recent developments in cognitive and neuro-sciences are bringing new understanding as to how the human brain processes information through distributed and connected pathways. Barbara de Salvo, Deputy Director for Science and Long-Term Research for CEA-Leti, discusses how these new discoveries are driving new models of semiconductor computing, which have the potential to enable step-change improvements in processing efficiency. She outlines a research strategy encompassing algorithms, circuits, and components, that aims to develop brain-inspired technologies to meet the needs of 21st-century applications.

The automotive industry is undergoing a once-in-a-lifetime transformation, where technologies once seen to be in the realm of science fiction, such as electric vehicles and autonomous cars, are rapidly becoming a reality. Yukihiro Kato, Executive Director of Denso explains how the semiconductor industry has a vital role to play in solving many of the technical challenges (ranging from efficient energy conversion and smart sensing, to real-time communication and decision making) that must be overcome to enable the reality of this future 'mobility-enhanced society'.

Our modern society has been revolutionised by the development of powerful and ubiquitous computing devices. David Patterson of Google and UC Berkeley reviews 50 years of innovation in computer architectures, from mainframe computers in the 1960s to the dominant RISC architecture of recent times. With the end of Moore's Law slowing the pace of improvements due to scaling, Patterson describes how this slowdown is actually rejuvenating innovation in computer architectures, as future performance improvements cannot come from scaling alone.

With this line-up of speakers, the Plenary Session covers a wide range of topics, ranging from semiconductor industry challenges and brain-inspired computing, to future automotive transportation and the history and future of computer architectures. We hope that you will find the presentations of our distinguished speakers informative and inspiring. Enjoy!

FORMAL OPENING OF THE CONFERENCE **8:30 AM**

1.1 Semiconductor Innovation: Is the party over or just getting started? **8:45 AM**

Vincent Roche, *President & CEO, Analog Devices, Norwood, MA*

The future pace of semiconductor innovation is by no means certain. A little more than a decade ago, Dennard scaling ground to a halt. Symposia and media outlets have been speculating on what comes after Moore's Law for years now. Beyond these technology challenges, business challenges, as well, are putting pressure on traditional semiconductor innovation: semiconductor prices continue their steady decline while small geometry wafer fab facilities now cost close to $10B to build.

In this environment, is there any room for continued innovation or is the future of the semiconductor industry defined by incrementalism, commoditization, and financial engineering? If our future is the latter, how will we meet the demands of a world where businesses, governments, and societies are digitizing at a blazing pace? The spread of pervasive ubiquitous sensing, rapid advances in artificial intelligence, heterogeneous integration, and the continued impact of digitization on virtually every industry on earth will require more, not less, semiconductor innovation.

The physicist and philosopher Thomas Kuhn might describe our situation as the crisis that catalyzes a new paradigm. So what is the next paradigm for semiconductor innovation? What is our path forward as scientists, technologists, and an industry?

 978-1-5090-4940-0/18/$31.00

1.2 Brain-Inspired Technologies: Towards Chips that Think — 9:20 AM

Barbara De Salvo, *Deputy Director for Science and Long Term Research, CEA-Leti, Grenoble, France*

Since the late 50s, brain-inspired computing has been regarded as an interesting alternative to conventional computational paradigms. Today, the omnipresence of "big data" and worldwide social interactions requires technologies capable of analyzing complex objects (such as sounds, images, or videos), in real time, and interact with humans in a cognitive way. The demand for computational efficiency and "intelligent" features has gone well beyond what can be achieved with traditional solutions. The advent of the Internet-of-Things has also introduced a new paradigm that supports a decentralized and hierarchical communication architecture, where a great deal of analytics processing should be done at the edge and at the end-devices instead of in the cloud. Specialized low-power architectures, inspired by the human brain, have thus recently become one of the most active research areas in the computing landscape, offering tremendous opportunities for novel applications. To map the embedded systems requirements, new challenges in brain-inspired technologies should be addressed, in particular automatic sensor fusion, system fault tolerance and data-privacy while achieving high recognition accuracy, low power consumption, and reducing cost.

In this talk, we will illustrate a research strategy – one encompassing algorithms, circuits, and components – to develop brain-inspired technologies and meet the needs of 21st-century applications. To explore the architecture of neural networks, an open software platform has been created and several neural network circuits conceived and fabricated. The use of innovative components, such as emerging resistive memories, advanced CMOS, and 3D technologies has been explored to allow for the implementation of cognitive tasks in neural networks. Those novel components bring memory closer to the processing unit and offer extraordinary potential to implement "intelligent" features, approaching the way knowledge is created and processed in the human brain. Several concrete examples will be given to illustrate how brain-inspired technologies are developed using a holistic research approach, where process development and integration, circuit design, system architecture, and learning algorithms are simultaneously optimized, opening the door to new disruptive applications.

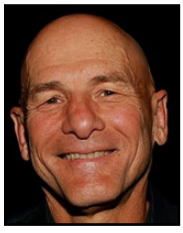
ISSCC, SSCS, IEEE AWARD PRESENTATIONS — **9:55 AM**

BREAK — **10:20 AM**

1.3 Future Mobile Society Enabled by Semiconductor Technology — 10:45 AM

Yukihiro Kato, *Executive Director, DENSO, Aichi, Japan*

The automotive industry is in the midst of a once-in-a-century greatest transformation. The transformation of the automobile is a consequence of three technology trends: (1) electrification, (2) driving automation, and (3) vehicle interconnection. All three of these require dramatic advancement within semiconductor technologies. In this talk, our vision of the future mobile society is presented, focusing especially on automotive semiconductor electronics. To promote the electrification of the car, power semiconductors are a key technology. The energy conversion efficiency of the motor has been increased by IGBTs, and next-generation SiC MOSFETs will further increase efficiency. However, to put automated driving to practical use, both advanced sensors and intelligent SoCs are required. Improved performance of sensors, such as cameras, LIDARs, and millimeter-wave radars, with increased range and resolution are required to precisely monitor the total environment of the car. Path planning for automated driving involves recognizing the vehicle's proximity to nearby objects and the free space available. In this process, deep learning is an exceedingly useful method and highly sophisticated SoCs with GPUs are essential for its implementation. Finally, from the viewpoint of a connected vehicle, cars will shift from lumps of metal into something like a giant smartphone! Of course, in the connected vehicle, you can make a phone call, receive and/or transmit emails, do shopping, make payments, and so on. As well, updated maps are constantly available. But, also, each vehicle must be constantly aware of the status and intent of nearby vehicles. Clearly, to implement the intricacies of such an interconnected vehicle network, we need a myriad of semiconductors including communication ICs.



1.4 50 Years of Computer Architecture: From Mainframe CPUs to Neural-Network TPUs — 11:15 AM

David Patterson, *Google, Mountain View, CA, University of California, Berkeley, CA*

This talk reviews a half-century of computer architecture: We start with the IBM System 360, which in 1964 introduced the concept of "binary compatibility". Next, came the idea of the "dominant microprocessor architecture", for which the early candidate was the Intel 432 which was shortly replaced by the emergency introduction of the Intel 80x86 in 1978. However, for the next 20 years, the Reduced Instruction Set Computers (RISC) became dominant. Then, the Very-Long-Instruction-Word (VLIW) HP/Intel Itanium architecture was heralded as their replacement in 2001, but instead the role was usurped by AMD's introduction of the 64 bit 80x86. Thus, while the 80x86 dominated the PC-Era, RISCs have led thereafter, currently with 20B shipped annually (versus 0.4B 80x86s). Since the ending of Moore's Law and Dennard scaling has stalled performance of general-purpose microprocessors, domain-specific computer architectures are the only option left. An early example of this trend introduced by Google in 2015 is the Tensor Processing Unit (TPU) for cloud-based deep neural networking .

PRESENTATION TO PLENARY SPEAKERS — **11:50 AM**

CONCLUSION — **11:55 AM**

1.1 Semiconductor Innovation: Is the party over, or just getting started?

Vincent Roche, President and CEO

Analog Devices, Norwood, MA

Semiconductor Innovation Trends:
The hardware paradigm underpinning the Information and Communications Technology (ICT) industry has experienced three major shifts over the past 60 years - the mainframe era, the personal computing era, and the Internet of Things era (Figure 1.1.1). During this period, the user-to-device ratio inverted; the mainframe era's many-to-one ratio has become a one-to-many ratio. Today, users are outnumbered by the devices they access, and even more significantly outnumbered by the instrumented nodes on which they rely. This exponential growth in devices and nodes has created a virtuous cycle of pervasive digitization and automation that is driving an even greater need for devices and nodes.

The $4T ICT industry that is changing how people learn, work, and live is both literally and figuratively built upon the $400B semiconductor industry that enables it. Technologies such as Artificial Intelligence, Machine Learning, and Virtual and Augmented Reality, as well as companies such as Facebook and Google, rely upon and gain an order-of-magnitude benefit from the semiconductor industry's innovation (Figure 1.1.2).

The direction and pace of future semiconductor innovation, however, is increasingly uncertain. One of the traditional drivers of innovation, Dennard scaling, ground to a halt a little more than a decade ago, when it was no longer possible to further reduce supply voltages. Now, Moore's Law is reaching its physical limits at the deep submicron level due to technical pressures (e.g., lithography, power, quantum tunneling, etc.) and economic pressures that have driven the cost of state-of-the-art wafer fabrication facilities to more than $10B (Figure 1.1.3). It is clear that the industry can no longer rely solely on these classical drivers of innovation in the future. New paths forward must be charted and different forms of investments made if semiconductor innovation is going to match, or even come close to, the progress and pace of the past half century.

Beyond these technological and economic headwinds, business realities are complicating the path forward. While innovation has always been the lifeblood of the semiconductor industry, growth is slowing and ROI has become an issue (Figure 1.1.4). This environment creates a subtle but steady pressure to abandon traditional innovation as the path to value generation, and shift business models to ones focused on financial engineering, with technical innovation relegated to a more modest supporting role marked by incrementalism.

This pressure must be resisted if the semiconductor industry is to continue to carry the mantle of technological leadership. That continued leadership is critical, not just for its own sake, but for the sake of the ICT industry and even larger aggregation of industries and organizations that increasingly rely on ICT for their own growth and progress.

To feed humanity's ancient hunger to better understand and shape the world around us, businesses, governments, and societies are digitizing at a blazing pace. Doctors are increasingly leveraging bits and bytes to make and keep people healthier (Figure 1.1.5). Computers are transforming transportation and making it safer, greener, and more enjoyable (Figure 1.1.6). People can communicate, collaborate, and create, with virtually anyone, anywhere today thanks to the spread of digital communications. These advances are just the start, as the spread of pervasive, ubiquitous sensing, continuous connectivity, and virtually unlimited processing and storage capability promise to drive digitization deeper into virtually every facet of life. The reality is that the future will require more, not less, semiconductor innovation.

Crisis:
The late physicist and philosopher Thomas Kuhn might describe the current situation as the early stages of a crisis that catalyzes a new paradigm or way forward[1] (Figure 1.1.7). The semiconductor industry's existing innovation model will increasingly struggle to respond to market demand within its current parameters. This is the pre-paradigm shift phase where alternative models begin to compete to become the new model.

For example, on the digital side of the industry, innovation teams began managing thermal and clocking constraints in their processing engines and moving to multicore processing architectures years ago to overcome limitations that could not be overcome through scaling alone. On the analog side, issues are being tackled through a strategy referred to as "More than Moore." Kuhn would say these non-Moore's Law based innovation models mark the beginning of a paradigm shift in how semiconductor innovation happens.

The industry is increasingly approaching the innovation problem from a perspective that supplements the traditional *technology supply-driven* approach with an *application demand-driven* approach (Figure 1.1.8). The technology supply-driven perspective focuses on driving improvements along the primary dimensions of performance, size, cost, and power-efficiency. Now, the industry has begun to adopt a demand-driven perspective as well that starts with the problem to be solved and works backward from there to more efficiently and effectively align innovations to applications. This application demand-driven approach is creating a *Triangle of Innovation* whose sides represent the principle areas in which future semiconductor innovation holds the greatest promise (Figure 1.1.9).

The Path Forward: The Triangle of Semiconductor Innovation:
One side of the triangle is technology innovation. The primary dimensions of innovation in the Moore's Law era have been reducing gate lengths and increasing wafer sizes to achieve scale and convergence around a limited number of nodes; what ITRS refers to as "More Moore." In contrast, "More than Moore" innovation is characterized by divergence of scope and diversification, employing process and package technologies (Figure 1.1.10), and more elements from the periodic table of elements (Figure 1.1.11), for example, to create new passive and active devices, and more complex and complete System on Chip hardware architectures. As we begin to address the application demand-driven requirements of customers, however, we need to look beyond the silicon and hardware domains to innovations along a third "Beyond Moore" axis in which vertical integration of the technology stack enables massive and varied new applications and content.

When you combine these three axes of technology innovation – More Moore, More than Moore, and Beyond Moore – you free yourself to deliver a much wider array of solutions, and more comprehensive solutions, to the market (Figure 1.1.12). Silicon-based software-defined radios (SDR), for example, are able to provide a reconfigurable, future-proof radio platform by combining an RF-to-baseband transceiver PHY and a digital processor. Correspondingly, ADI has been able to revolutionize SDR design by combining radio signal chain hardware, including RF transmitters and receivers, mixers, A-to-D converters, and digital signal processing, with software technology such as a Digital PreDistortion (DPD) algorithm to form a complete SDR platform on a chip (Figure 1.1.13).

The second side of the triangle is focused on driving innovation from the outside in; that is, based on insights and advancements from the world in which semiconductor technologies are being applied. Semiconductor suppliers need to actively seek application domain expertise to better align their technologies with the specific needs of an application. For example, if a provider seeks to provide technology for healthcare, they must go beyond merely understanding the technical problem they are trying to solve to a deeper understanding of why the problem matters, alternative solutions for the problem, and even the market and business context in which that solution will be applied (Figure 1.1.14). Innovations will deliver more impact when they are inspired and informed by the application and its environment.

The final side of the semiconductor innovation triangle addresses the growing complexity of applications by pushing suppliers to go beyond their four walls and begin truly innovating *as an ecosystem*. Suppliers must be clear about the areas where they are unique and should strive for leadership, but also the areas where they should take a federated approach in which they leverage formal and informal partnerships to drive innovation forward (Figure 1.1.15). This ecosystem approach also creates the potential of entirely new revenue streams and business models (e.g., charging for access to data) for ecosystem members.

Conclusion
Semiconductor suppliers' understanding of innovation at the end of Dennard Scaling and Moore's Law must become much more expansive. As suppliers explore the full scope of the triangle of innovation, they will be forced to change themselves – their skills, capabilities, and shapes – and in the process, new market leaders will no doubt emerge. Those who satisfy themselves with incremental technological improvements, and delay their embrace of this broader approach to innovation will be relegated to being yet another footnote in the annals of semiconductor history. The future is wide open and the owners of that future will be those who sense market shifts and move most boldly to adapt their innovation and businesses to the world that is and will be.

[1] *T. Kuhn, The Structure of Scientific Revolutions, Second Ed., Enlarged (Chicago: The University of Chicago Press, 1970)*

 978-1-5090-4940-0/18/$31.00

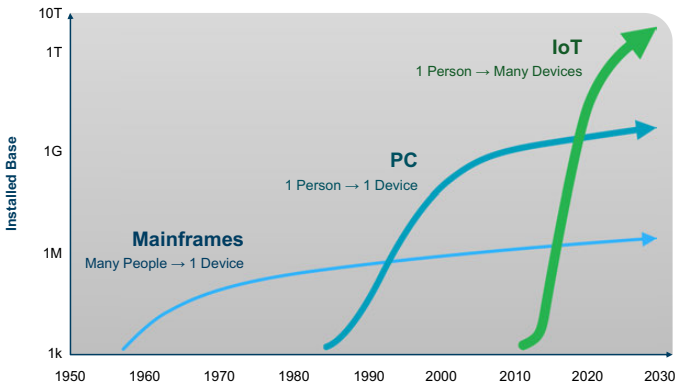


Figure 1.1.1: Three Waves of Information and Communications Technology (ICT).

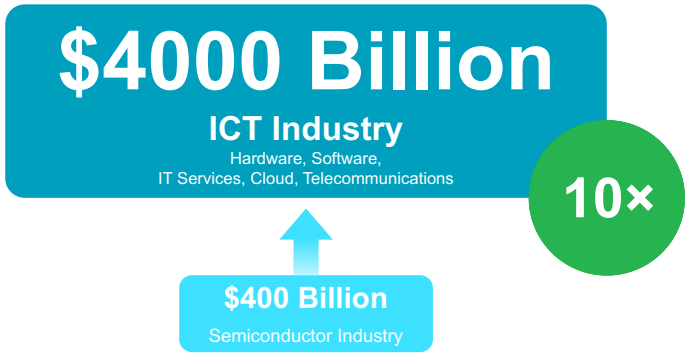


Figure 1.1.2: Semiconductor Industry Impact.

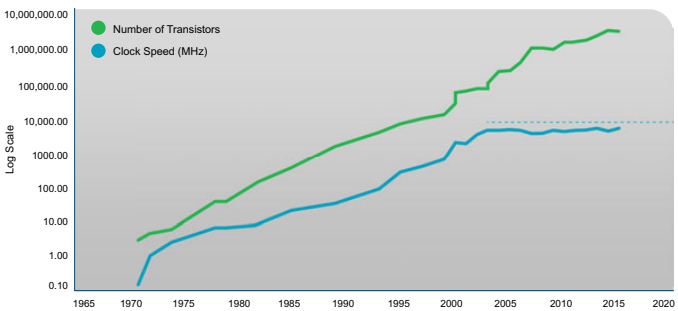


Figure 1.1.3: Physical Limits of Traditional Semiconductor Innovation.

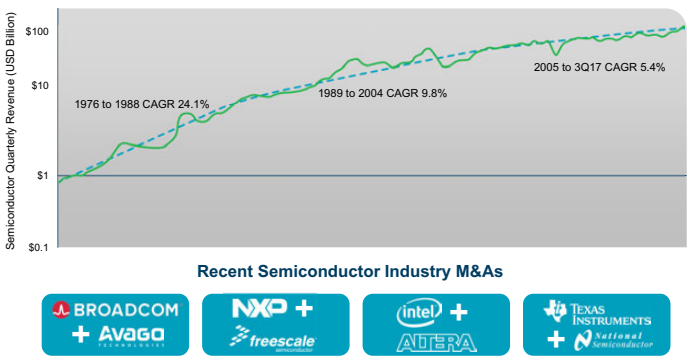
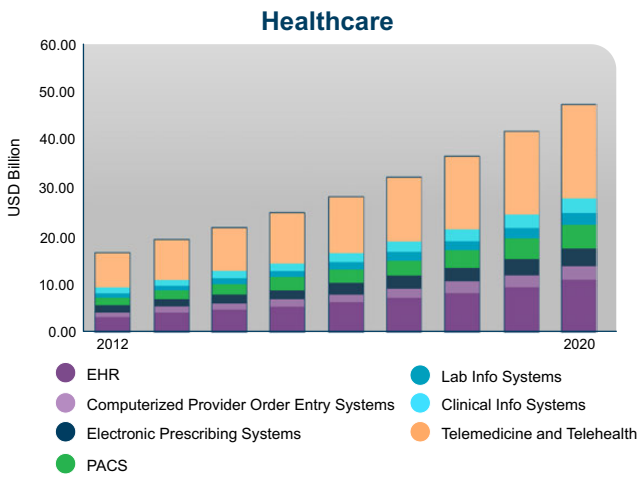
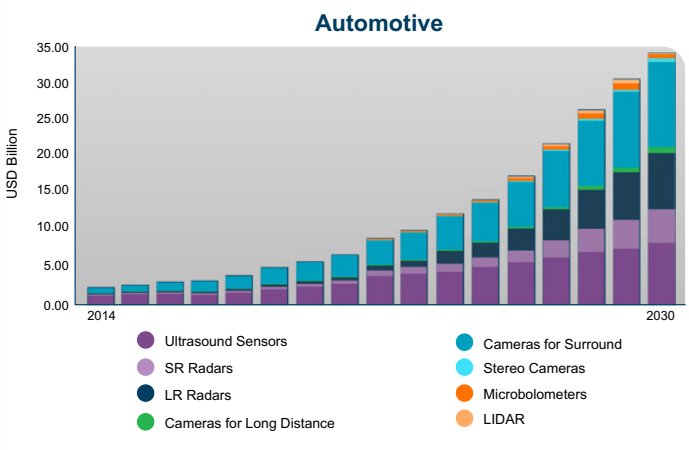


Figure 1.1.4: Maturation of Semiconductor Industry.



Source: Grand View Research, "Healthcare IT Market Analysis by Application (Electronic Health Records, Computerized Provider Order Entry Systems, Electronic Prescribing Systems, PACS, Lab Information Systems, Clinical Information Systems, Telemedicine and Telehealth) and Segment Forecasts to 2020," October 2015.

Figure 1.1.5: Growing Demand for Semiconductor Innovation.



Source: Yole Développement, "Sensors and Data Management for Autonomous Vehicles report 2015," October 2015.

Figure 1.1.6: Growing Demand for Semiconductor Innovation.

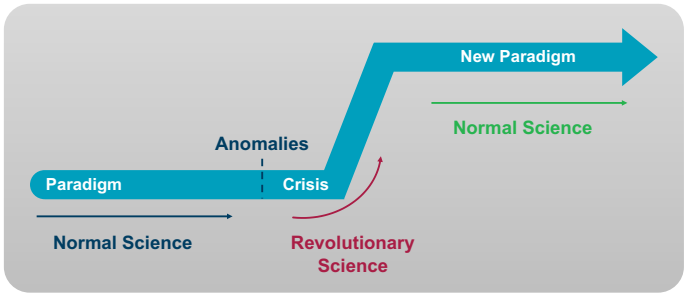


Figure 1.1.7: Crisis as Catalyst.

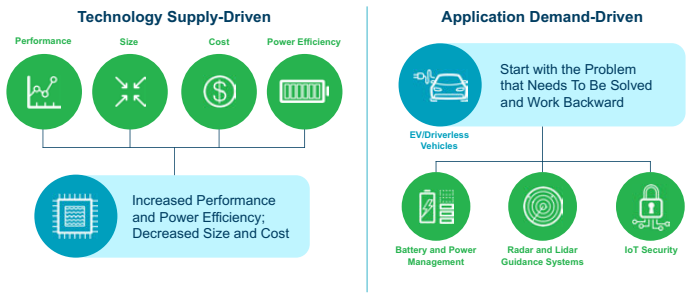


Figure 1.1.8: Approaches to Innovation.

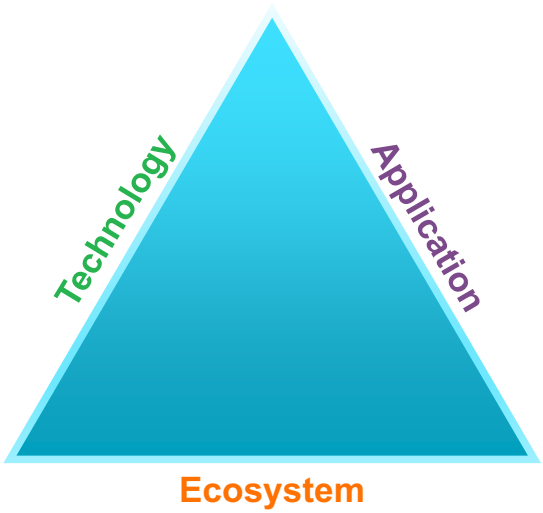


Figure 1.1.9: The Innovation Triangle.

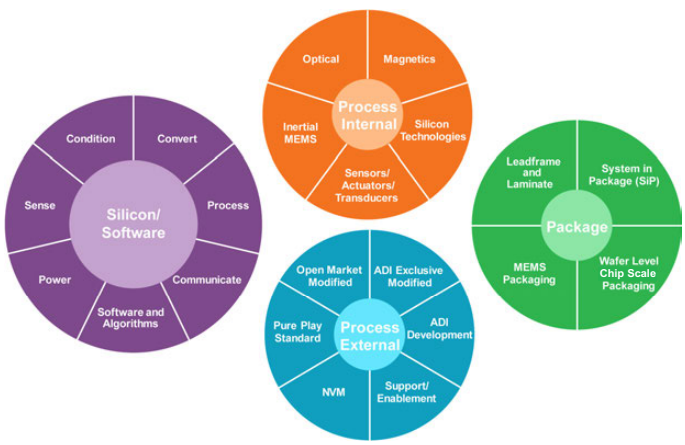


Figure 1.1.10: Integration.

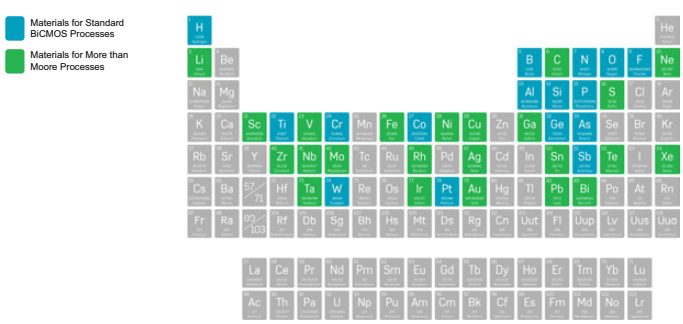


Figure 1.1.11: Materials Innovation.

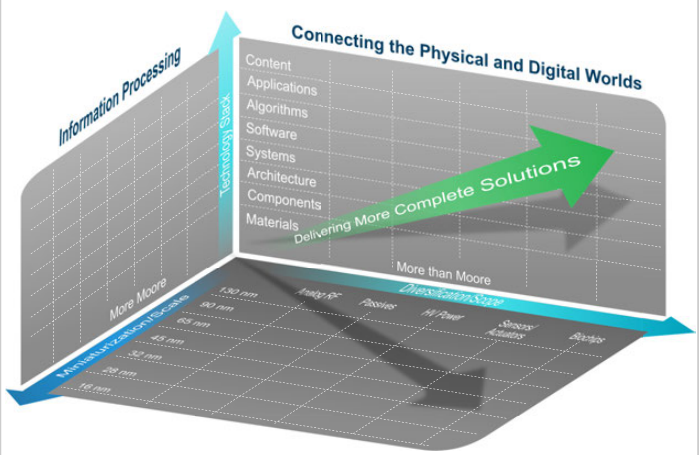


Figure 1.1.12: More Moore, More than Moore, *and* Beyond Moore.

 978-1-5090-4940-0/18/$31.00

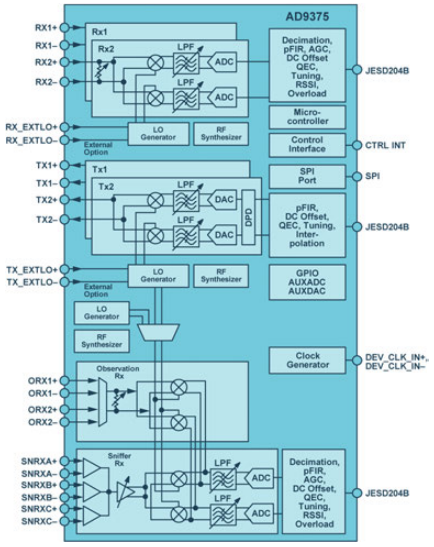


Figure 1.1.13: Vertical Integration.

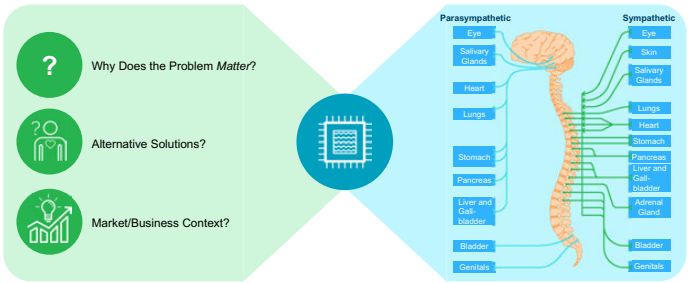


Figure 1.1.14: Healthcare Application Domain Expertise.

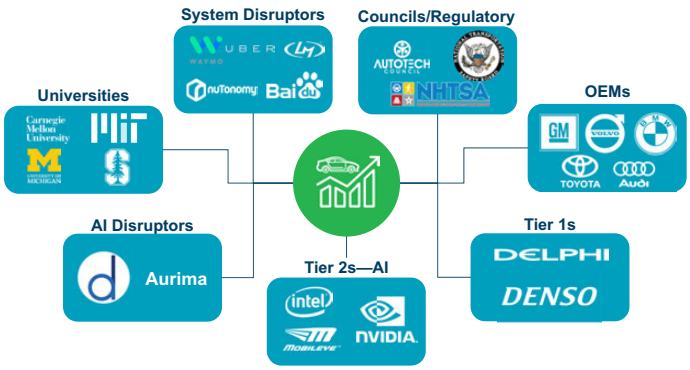


Figure 1.1.15: Automotive Innovation Ecosystem.

1.2 Brain-Inspired Technologies: Towards Chips that Think?

Barbara De Salvo, Chief Scientist and Scientific Director

CEA-Leti, Université Grenoble Alpes, France

Abstract

The advent of the *Internet-of-Things* has introduced a new paradigm that supports a decentralized and hierarchical communication architecture, where a great deal of analytics processing occurs at the *edge* and at the *end-devices* instead of in the *Cloud*. To map the embedded-systems requirements, we present a *holistic research approach* to the development of *low-power architectures inspired by the human brain*, where process development and integration, circuit design, system architecture, and learning algorithms are simultaneously optimized. This paper is organized as follows: We begin with a survey of recent research on the *human brain* and a historical perspective of *cognitive neuroscience*. Then, *artificial intelligence* is introduced, and the challenges of *Deep Learning* systems (in terms of power requirements) are addressed. The key reasons to distribute intelligence over the whole network are discussed. To emphasize the need for low-power solutions, a quantitative *benchmark* of existing *specialized edge platforms* that can execute *machine-learning algorithms* on conventional embedded hardware is presented. The primary focus of this paper will be on the implementation of optimized *neuromorphic hardware* as a highly promising solution for future ultra-low-power cognitive systems. We show that *emerging technologies* (such as advanced CMOS, 3D technologies, emerging resistive memories, and Silicon photonics), coupled with *novel brain-inspired paradigms,* such as spike-coding and spike-time-dependent-plasticity, have extraordinary potential to provide intelligent features in hardware, approaching the way knowledge is created and processed in the human brain. Finally, we conclude with our vision of the *enabled future disruptive applications* and a discussion of the *main challenges* which should be tackled to exploit the full potential of brain-inspired technologies.

1.0 The Brain and Cognitive Science Perspective

More has been learned about the brain in the past decades than in all prior human history. In the quest for understanding the brain, neuroimaging has played a pivotal role, enabling *in-situ*, non-invasive brain mapping [1]. Medically, this has facilitated early diagnosis and treatment of patients with specific neurological or psychiatric diseases. *Magnetic resonance imaging (MRI*) has become the reference technique to investigate the human brain *in-vivo*, making anatomical/functional imaging and cerebral connectivity mapping possible (see **Fig. 1.2.1**). The MRI scanner operating at 11.7Tesla, currently installed at *CEA-Neurospin*, is being used to study the brain on a 100µm scale (addressing volumes corresponding to a few thousand neurons). Recent discoveries suggest that the brain organization has been shaped by a trade-off between a parsimonious principle of minimizing costs and maximizing adaptive values and robustness [2]. The large-scale neuronal networks of the brain are arranged globally as hierarchical modular networks, with dense modules at the local level (cellular circuits, laminar compartments) that are encapsulated in increasingly larger modules (cortical columns, areas and whole lobes), but with very sparse overall connectivity. Such a topology fundamentally enhances the brain's dynamic stability and information-processing abilities. An important research target will be to understand how the three-dimensional organization of brain cells, neurons and glial cells, connected in networks within the layers of the brain cortex, are responsible for the emergence of genetically-determined elementary operations. These operations combined together and interacting with the environment, give rise to higher-order functions, such as language, calculations, and consciousness.

This period of rapid discoveries has also seen the rise of *cognitive science,* a unified science based on interdisciplinary efforts among researchers in various fields (neurosciences, physics, biology, psychology, linguistics, artificial intelligence, robotics, and philosophy) whose aim is to investigate the functional architecture of cognition through computational models. Since its inception in the mid-1950s, cognitive science has moved through a series of different paradigms, which have strongly influenced the evolution of *artificial cognitive* systems. The first shift away from classical *behaviorism* (which asserts the dominant role of environmental factors on mental processes) came with *cognitivism*, under which thinking corresponds to a logical manipulation of symbols representing external phenomena. The rules that specify how symbols are transformed were taken to govern cognitive performance. Since the early 1980s, the important discoveries in the field of brain neurophysiology have led to the emergence of *connectionism. Connectionist* systems rely on parallel processing of non-symbolic distributed activation patterns using statistical properties, rather than logical rules. Their models reflect the concept of "*emergence*" in the brain's organization (cooperative interactions of individual components determining the "emergent" functionalities of the whole entity, given that these functionalities do not exist individually). The most common connectionist models are *neural networks*. By the repeated presentation of a training set and application of the learning rule, networks can learn to produce the correct responses to a set of inputs. In the past decade, connectionist models have strongly evolved and a new class of architectures (such as feedforward, fully-recurrent, simply recurrent) and learning frameworks (such as supervised, unsupervised, reinforcement), with almost no resemblance to biological systems, have been developed in order to implement them in artificial cognitive systems. It is worth mentioning that in recent years, a new approach called *embodied or enactive cognitive science* has emerged [3]. Whereas traditional cognitive science rests on a fixed inside–outside distinction, assuming that the mind is separate from the outside world, the embodied cognition approach views the *mind* as a biological system *rooted in body experience*, and *interacting with the environment and other individuals*. Embodiment refers to both the embedding of cognitive processes in brain circuitry and to the origin of these processes in an organism's sensory–motor experience. Action and perception are no longer interpreted in terms of the classic physical–mental dichotomy, but rather as being closely interlinked. The possible implications of this last paradigm in the design of future cognitive agents that interact with the physical world have yet to be fully explored.

2.0 Hyperconnectivity and Deep-Learning Power Challenges

In the past few decades, the world has experienced great transformations. Enabled by the convergence of miniaturization, wireless connectivity, increased data-storage capacity, and data analytics, the Internet-of-Things (IoT) has been the epicenter of profound social-, business-, and political-changes. With billions of easy-access and low-cost connected devices, the world has entered the era of hyperconnectivity [4], enabling people and machines to interact in a symbiotic way (anytime, anywhere) with both the physical and cyber worlds. *Artificial intelligence (AI)* has been at the center of this revolution. In recent years, we have seen a boost in the performance and applications of *machine learning (ML)*, driven by several factors: (i) the enormous storehouses of *data* (images, video, audio, and text files strewn across the Internet) which have been essential to the dramatic improvement of *learning/training* approaches and algorithms; (ii) the increased *computational power of modern computers* (the advent of parallel computing for neural network processing having compensated the slowing down of Moore's Law below the 10nm node). Among the many fields of ML, *Deep Learning (DL)* is the most popular. Today, for tasks such as image or speech recognition, ML applications are equaling or even surpassing expert human performance. Other tasks considered as extremely difficult in the past, such as natural language comprehension or complex games, have been successfully tackled. The particular case of the AlphaGo program from Google is remarkable in that it demonstrates how to increase performance by refining the algorithm architecture and combining several techniques of ML (DL techniques with reinforcement learning). In the future, new applications will require more and more analysis, understanding of the environment and intelligence. Self-driving cars will have to be able to recognize and analyze their environment through multiple sensors. Personal digital assistants will require voice and context analysis. For ML algorithms to become pervasive, increased computational resources will be needed. However, for the time being, data are transmitted in hierarchical infrastructures, and applications must deal with many different levels of analysis: *Cloud* computing, the *edge* (networked mobile devices), and the *end-devices* (wireless sensor nodes). Most of the data processing for *DL training,* and even for the *inference* phase, happen in the *Cloud* (data are sent to a data center and then processed there, before pushing operational decisions back to the edge platform). But, AI algorithms are not useful in settings where connectivity is sparse. Moreover, training a DL network in the *Cloud* (with conventional processors or GPU) on extremely large datasets involves intensive computing tasks and can take several weeks [5]. As well, the power limitations of servers used for DL are expected to slow down the pace of performance improvements. This poses a great challenge to computing platform designers.

3.0 Towards Distributed Intelligent Systems

Bringing intelligence to the *edge* or to *end-devices* means doing useful processing of the data as close to the collection point as possible, and allowing systems to

 978-1-5090-4940-0/18/$31.00

make some operational decisions *locally*, possibly semi-autonomously. Distributing the intelligence over the network is important for a number of reasons: *Safety* will require local decision making, *in real time*, without having to rely on a connection that could be interrupted for various reasons. Running *real-time* DL locally is essential for many applications, from landing drones to navigating driverless cars. The delay caused by the round-trip to the Cloud could lead to disastrous or even fatal results. *Privacy* will require that key data not leave the user's device, while transmission of high-level information, generated by local neural-network algorithms, will be authorized. Raw videos generated by millions of cameras will have to be locally analyzed to limit *bandwidth issues* and *communication costs.* For all these reasons, new concepts and technologies that can *bring artificial intelligence closer to the edge* and *end-devices* are in high demand. The primary design goal in distributed applications covering several levels of hierarchy (similar to what happens in the brain), is to find a global optimum between *performance* and *energy consumption.* This imperative requires a holistic research approach, where the technology stack (from device to applications) is redesigned. As shown in **Fig. 1.2.2**, to address *embedded applications*, major industrial players and start-ups have developed specialized edge platforms that can execute ML algorithms (*inference*) on embedded hardware (CPU and GPU), such as Movidius Myriad X, MobilEye Eye Q5, Jetson TX2. Impressive power improvements (down to a few Watts) have been achieved by exploiting Moore's Law (pushing the FinFet technology down to the 7nm node) and by hardware-software co-optimization. Since many mobile applications are "*always-on*" (e.g., voice commands), low power is critical for mobile IoT [6]. In this context, several research groups have focused on hardware designs of C*onvolutional Neural Network (CNN)* accelerators. Precision-Scalable Processors (implemented in 40nm LP CMOS) for deep neural networks have shown power consumption in the range of 70mW [7]. The need for off-chip storage devices, such as DRAMs, significantly increases power consumption. Recently, mobile-oriented applications (keyword spotting and face detection) have been demonstrated with a low-power programmable *DL accelerator* [8] (incorporating on-chip weight storage) which consumed less than 300μW.

It is worth mentioning that the challenges of bringing intelligence into *low-power IoT-connected end-devices* (with applications ranging from habitat monitoring to medical surveillance) are much more demanding than those associated with traditional networked mobile devices at the edge [9]. Most connected end-devices are wireless sensor nodes containing microcontrollers, wireless transceivers, sensors, and actuators. The power requirement for these systems is extremely critical (<100μW for normal workloads), as these devices often operate using energy harvesting sources or a single battery for several years. The unreliable, noisy and complex environments where these systems are deployed create difficulties in modeling and predicting this environment (as in the case of energy harvesters and wireless communications). Fixed or non-intelligent communication protocols may dissipate the energy harvested at the nodes. To address this issue, *adaptive mechanisms* have been proposed which reduce energy requirements at the architectural or circuit level (dynamic and frequency scaling approximate computing). However, finding the best system configuration relies on knowledge of the system state [10]. Learning about the changing environment and configuring the system accordingly, using various techniques, is the key to achieving energy savings [11]. Moreover, fast and accurate *decision making* in IoT end-devices can be achieved using learning techniques. Many applications can be foreseen for ML, such as power and reconfiguration management, non-volatility control, and security-countermeasure activation. An example is provided in [12], where *neural cliques* behaving as an associative memory allow for very fast and accurate decision makings. Furthermore, *reinforcement-learning techniques* (where the learner must discover which actions yield the most rewards by trying them) can also be applied for end-device control with good accuracy at a very low cost [13]. When communication with the Cloud or edge devices is not possible, *live in-node data processing* and *classification* are required, and should be optimized so that they consume minimal energy, and preserve the quality of the information. *Genetic machine learning algorithms* have been explored for this purpose [14, 15], but integration of learning algorithms into low-power devices still remains an issue. Evaluations of computational requirements for embedding deep learning in low-power IoT devices (like, for example, a smart glass performing real-time recognition on the video stream that it captures) have shown a large processing efficiency gap between the capabilities of current computing platforms and the requirements imposed by such distributed applications [5]. Finally, to improve implementation efficiency of ML, various approaches have been explored. Nevertheless, given the energy costs related to the memory system, and the constraints on both parallelism and technology scaling, it might seem like there is not much room for additional energy improvements [16]. Finding new affordable, energy efficient ways to implement inference and learning through *new specialized low power and distributed compute engines* is thus key for future intelligent systems.

4.0 Advanced Technologies for Brain-Inspired Computing

Inspired by the *human brain*, whose computing performance and efficiency still remain unmatched (see **Fig. 1.2.2**), a radically different approach is being investigated: It consists in implementing bio-inspired architectures in *optimized neuromorphic hardware* to provide direct one-to-one mapping between the hardware and the learning algorithm running on it. This approach, which originated with the pioneering work of Carver Mead [17], has yet to be fully demonstrated and industrialized. Implementation limitations are linked to several elements [18], such as the difficulties to emulate the behavior of neural network elementary components (*neurons, synapse*s) with standard CMOS technologies, and to achieve a 3D brain-like high-density connectivity with 2D-layer technologies. In the following paragraphs, essential brain-inspired operating principles (such as *spike coding* and *STDP*) will be introduced, followed by a detailed discussion on how *emerging technologies* could lead to new neuromorphic hardware, and thereby change the rules of the game.

4.1 Spike Coding and Spike-Timing-Dependent-Plasticity (STDP)

The first brain-inspired operating principle to consider is the way neuron states are encoded in a system. In the past, neuron values were encoded using analog or digital values. However, a recent trend in neuromorphic computing is to encode neuron values as pulses or *spikes* [17, 24, 25]. This parsimonious signal coding was inspired by the way neurons of the central nervous system interact, leading to higher energy-efficiencies. It differs from the traditional *signal rate-coding* (used in today's main industrial neural network applications), which employs the average frequency of spikes in a given time window. The values manipulated in those networks (inputs and outputs of neurons) are numbers representing the "cumulative" effect of spikes over time. However, if input/output signals are represented as *pulses (spikes)*, the multiplication operation between input signals and synaptic weights is reduced to a gating operation at the synapse level. This typically produces a weighted current at the arrival of the pre-synaptic spike that is integrated by the post-synaptic neuron. The higher the frequency of the input spikes, the larger the value integrated by the neuron. Furthermore, if many synapses receive input spikes in parallel, the weighted sum operation is implemented directly at the input node of the post-synaptic neuron following Kirchhoff's current law. Thereby, power consumption can be reduced by implementing this spike or *event-based signal representation* (called *Address-Event Representation, AER)* using asynchronous schemes. Given these features and because this representation is also optimal for transmitting signals across long distances or chip boundaries, most of the recent state-of-the-art neuromorphic computing approaches are using AER. Moreover, spiking neurons offer the additional advantage of being easily interfaced with *low-power spiking sensors* (e.g., image-, audio- , tactile- or chemical sensors [56-60]). The second brain-inspired principle essential to neuromorphic systems is the *learning paradigm* (i.e. the way the synaptic connections among neurons are created, modified and preserved). The computation schemes to define the synaptic weights can be divided into two types: (1) *supervised learning*, where the inference process is based on training examples (this is the case for most neural-inspired machine learning algorithms, which show impressive performance for solving very specific tasks but at the cost of huge power dissipation,); and (2) *un-supervised learning*, which does not use any feedback from an external teacher, but attempts to classify inputs based on the underlying statistics of the data.

Spike-timing-dependent-plasticity (STDP) is a bio-inspired algorithm that enables unsupervised learning. The assumption underlying STDP is that synapses tend to reinforce causal links. That is, when the presynaptic neuron spikes just before the postsynaptic neuron spikes, the synapse between the two becomes stronger. Therefore, if the presynaptic neuron spikes again, the synapse will allow the postsynaptic neuron to spike faster or with a higher occurrence probability.

We will now present the extraordinary potential of *emerging technologies* which could be coupled to the aforementioned *novel brain-inspired paradigms* to provide intelligent features in hardware.

4.1.1 Fully-Depleted Silicon On Insulator (FDSOI)

For the past decade, FDSOI technology has proven to be a viable solution to satisfy Moore's Law requirements for the next CMOS generations [19]. It has been successfully deployed in many applicative fields (including entry-level application processors for smartphones, system-on-chip devices for autonomous driving and the IoT, and mm-wave applications). Thanks to its suitability for low-power design, FDSOI technology is a great candidate for neuromorphic hardware. In the field of *DL architectures*, *high-performance reconfigurable digital processors* based on 28nm FDSOI have shown power consumption in the range of 50mW. This power efficiency has been achieved by introducing optimized data-movement strategy and exploiting FDSOI back-biasing strategies [20, 21]. Recently, *a large-scale multi-core neuromorphic processor (*named *Dynap-SEL),* also based on 28nm FDSOI, was demonstrated (see **Fig. 1.2.3**) [22, 23, 24, 25]. It occupies an area of 7.28mm^2 and comprises four TCAM-based cores and one plastic core. Each TCAM-based core has 256 neurons and 16k TCAM-based programmable synapses, while the plastic core has 64 neurons with 4k plastic synapses, and 4k programmable synapses. In addition it integrates 8.5k × 18-bit SRAMs as Lookup Tables (LUTs), 3-level hierarchical routers, two temperature compensated bias generator circuits for generating 190 on-chip biases, and one input pre-decoder block. Thanks to the scalable architecture and to the on-chip programmable routers, the routing of all neurons on a 16×16 chip array can be easily configured to implement a wide range of connection schemes, without requiring external mapping, memory, or computing support. In order to minimize power consumption, a *mixed-signal design approach* was chosen and analog circuits were used. In this way, the physics of the device was exploited to implement the desired neural network computational primitives. Because these primitives are mainly composed of exponential and logarithmic functions, using sub-threshold analog circuits is the best choice. Indeed, the mixed-signal accelerator demonstrated in [22, 23] consumes 50pJ per spike, approaching the energy efficiency of biological neurons, which is estimated to be a few pJ per spike. Sub-threshold analog circuits reproduce the synapse and neural dynamics expected from theory. They can be used to provide biologically realistic dynamics or fast rectified linear unit transfer functions. They are also fully compatible with spike-based learning algorithms, and can be readily integrated into the next generation of large multi-neuron multicore neuromorphic architectures.

4.1.2 3D Through Silicon Vias (TSVs) and Monolithic 3D

The human brain's intelligence and efficiency is strongly linked to its extremely dense *3D interconnectivity* (roughly 10,000 synapses per neuron, and billions of neurons in the human brain cortex). The hierarchical structure in the cortex follows specific patterns, through vertical arrangements or *µcolumns* (where local data flow on subcortical specialized structures) and *laminar interconnections* (which foster inter-area communications and to build the hierarchy) [29]. Based on these considerations, it is clear that emerging 3D technologies will be a key enabler of efficient neuromorphic hardware. **Figure 1.2.4** shows the evolution (in terms of connection density) and hardware applications of 3D technologies. *Through Silicon Vias* have enabled heterogeneous system integration and are being increasingly used in devices (such as DRAM memory cubes, passive interposers for FPGA or GPU integration, BSI imagers, heterogeneous integration of MEMS and active interposers for High-Performance Computing [26]). Further scaling of 3D interconnects, to achieve pitches in the 1µm range, will be possible using *hybrid bonding* technology [27]. This approach offers a large architectural perspective and a way to overcome the classical limitations of today's imagers. A two-layer 3D partitioned *CNN architecture* is presented in [28]. Each layer comprises a neuronal compute block and the associated memory. This novel circuit uses fine pitch hybrid bonding and presents a substantial 25% improvement in power consumption when compared to a regular 2D version. Today, *3D Sequential Integration* (3DSI), also called monolithic 3D integration, offers new 3D partitioning options at fine granularities thanks to the ultra-small 3D contact pitch (<100nm) [30]. 3DSI consists in stacking active device layers on top of each other in a sequential manner. It differs from 3D packaging, where the tiers are fabricated in parallel, then stacked by bonding. As the top layer's active patterning is defined by the lithographical process-of-reference, the alignment accuracy and feature size of stacked tiers and inter-tier interconnections are dictated only by stepper resolution. This ultra-dense connectivity between memory arrays and computing logic provides much more parallelism capability for high-energy-efficiency computing [31]. Recently, a 3D monolithic integrated nanosystem, based on beyond-Si nanotechnologies, with vertically interleaved layers of computing and data storage, fine-grained and dense connectivity, was demonstrated [32]. Using 3DSI in neuromorphic computing will allow maximum connectivity and reconfigurability between neurons and synapses, a step forward towards cortical µcolumn-like interconnectivity.

4.1.3 Resistive Memories (ReRAM)

Several large-scale neuromorphic systems have been proposed in the last years, taking advantage of the enormous potential of current Silicon technologies. Examples include the Heidelberg's HICANN [35], IBM's TrueNorth [36, 37], and ETH's ROLLS [38] chips. These approaches use standard CMOS technologies to implement both neurons and synapses. The synaptic weights are stored in analog or digital devices such as capacitors or SRAM. Nevertheless, SRAM-based synapses are affected by the problems of area consumption and data volatility. When the network is turned off, the synaptic weights stored in the SRAM are lost, stressing the need for storage in nonvolatile memories (NVMs) during or after the learning process; but NVMs come with additional power and area consumption. Recently, new memory technologies, called ReRAM (such as phase-change memory (PCM), spin-transfer magnetic memory (STT-MRAM), metal-oxide resistive-switching memory (OxRAM), conductive-bridge memory (CBRAM) and Vertical Resistive Memories (VRAM)) have appeared. These memories offer several key features, such as: low voltages (ranging from 1V to 3V), fast programming and reading time (few 10s of ns, even <1ns), long data retention, single-bit alterability, execution in place, good cycling performance (higher than Flash), density and ease of integration in the Back-End-Of-Line of advanced CMOS. ReRAM are currently developed for applications such as microcontrollers [33], servers and high performance computers. Bringing memory close to the processing unit will revolutionize traditional memory hierarchy [34] and facilitate the implementation of in-memory computing architectures. Due to their low power consumption, multi-value properties, and non-volatility, ReRAM memories are also promising for implementing energy-efficient bio-inspired synapses in complex neural network systems [39-41] (see **Fig. 1.2.5)**. In [42], a CNN spike-based architecture for pattern recognition, using HfO_2-OxRAM devices as synapses for convolution kernels has been presented. It was inspired from the mammalian visual cortex organization and consists of two cascaded convolutional layers and a classification module. The CNN was simulated using an in-house special purpose C^{++} event-based simulator (*Xnet*) [43]. Kernels were defined using a backpropagation supervised learning algorithm. The OxRAM based CNN demonstrated high accuracy (recognition rate > 98%) for complex visual pattern recognition applications. This result is in agreement with the state-of-the-art recognition success rate obtained with formal CNN models, implemented with floating–point precision synapses. Thanks to the use of ReRAM synapses to implement the kernel, the convolution operations are performed directly in memory, reducing the latency per image recognition with respect to software implementations on GPU. The use of ReRAM synapses also opens a path towards *online real time unsupervised learning* (through continuous weight updating performed on local synaptic weights) and biological brain *life-long learning* abilities (i.e. once learned, it is almost impossible to train the same algorithm or network on a different task without completely re-learning all parameters). Plasticity will play an important role in achieving these goals. Two main approaches to emulate synaptic conductance modulation have been successfully demonstrated. In the *analog approach*, multiple low-resistance states for emulating long-term potentiation (cumulative increase of conductance, LTP) and multiple high resistance states for long-term depression (cumulative and gradual decrease of conductance, LTD) are used. In the *binary approach,* only two distinct resistive states (LRS and HRS) are used per device, with probabilistic STDP bio-inspired learning rules. This approach is also motivated by biological studies which suggest that STDP learning might be a partially stochastic process in nature. In the case of the binary approach, in order to improve performance, a single synapse could be composed of *n multiple binary cells* in parallel. Several ideas have been proposed to implement STDP with memory devices. A simplified version of STDP is presented in [43], where the analog time dependence of biological STDP is neglected, and only two conditions (increasing or decreasing synaptic weight) are considered. This model requires technologies with multilevel capability. *Phase-change memories* show a strong asymmetry between the SET and RESET process: whereas the SET process is extremely gradual and resembles learning in neural networks, the RESET process is abrupt. In [43, 44] a 2-PCM synapse that recreates artificial symmetry between SET and RESET by employing two devices per synapse has been proposed. This strategy has been shown to achieve unsupervised learning in a fully-connected neural network for automobile tracking. An average detection rate of 92%, and

 978-1-5090-4940-0/18/$31.00

a system power consumption for learning of 112μW have been demonstrated by means of system-level simulations. In [45], an original methodology that uses *conductive-bridge RAM devices* as easy-to-program and low-power binary synapses with stochastic learning rules, is proposed. This learning scheme has been demonstrated on a *fully-connected neural network* able to process asynchronous analog data streams for recognition and extraction of repetitive patterns in a fully-unsupervised way. These demonstrated applications exhibit very good performance (auditory pattern sensitivity >2) and ultra-low synaptic power dissipation (0.55μW) in the learning mode. Low-power neuromorphic computing systems can also be coupled with *Brain-Computer Interfaces* (BCI) to enable the design of *autonomous implantable devices* for rehabilitation purposes, capable of making decisions based on real-time on-line processing of in-vivo recorded biological signals. In [46], a ReRAM-based *two-layer fully-connected neural network* able to identify, learn, recognize, and distinguish between different spike shapes of measured biological signals without any supervision, has been proposed.

Figure 1.2.5 shows the topological view of the network architecture: The biological signal is encoded by 32 frequency band-pass filters. The 32 filtered signals are then full-wave rectified and presented to the input layer of 32 neurons where the analog continuous signals are converted into spikes which are then propagated along the synapses to the five output neurons. To solve one of the main challenges of *biological signal treatment in BCI* (the high background-noise level), a synaptic compound using HfO_2-based OxRAM cells, able to implement two different flavors of spike-based synaptic plasticity, the long-term and the short-term learning rules, has been presented [47]. Thanks to long-term plasticity, the system is capable of learning based on an unsupervised paradigm, while the short-term plasticity allows for improved accuracy despite the significant background noise in the input data. Biology teaches us that noise can improve the performance of biological sensory systems. Inspired by this assessment, several studies have been devoted to leveraging *intrinsic device noise* for neuromorphic computing. For example, the stochastic switching behavior of ReRAM under weak programming conditions was used to implement synapses with probabilistic STDP learning rules [45-48], and neuron circuits with stochastic firing [49].

4.1.4 Silicon Photonics

Silicon (Si) photonic technologies are used today in datacenters for high-bandwidth multi-user communication networks. The recent advent of *hybrid platforms* that integrate photonic components on Si wafers in a cost-effective way [50, 51] opens new application fields. Optical interposers to stack and connect computing and memory chiplets together for very fast processing and high energy efficiency have been recently demonstrated [52]. Si photonics has also been explored for application in neuromorphic hardware. Photonic platforms offer an alternative approach to microelectronics, potentially overcoming the fundamental limit of highly-interconnected networks (the bandwidth connection-density tradeoff). The *high speeds, high bandwidth, and low cross-tal*k achievable in photonics seem very well-suited for ultra-fast spike-based information schemes with high interconnection densities. In [53], the use of electro-optic modulators as *photonic neurons* has been proposed. A reconfigurable 49-node Si photonic neural network able to perform emulation tasks has been presented. The results predict a *1960× speed-up over a CPU benchmark*. In [54], photonic hardware is proposed for the implementation of a *Reservoir Computer or Echo-State Networks*. This is a new paradigm in artificial Recurrent Neural Network (RNN) training, where an RNN, the *reservoir,* is generated randomly, and only a readout is trained [55]. Aside from the many potential advantages of photonics in general, it should be noted that photonic neuromorphic computing still remains a very exploratory field and more studies are needed to validate the promises.

5.0 Future Opportunities and Challenges

We are entering a new era where *artificial-intelligence systems* are becoming key players, shaping the future world. With the end of Moore's Law in sight, transformative approaches are needed to address the enduring power efficiency issues of traditional computing architectures. *Brain-inspired hardware,* coupled to new computing paradigms and algorithms, will exploit the full potential of new disruptive technologies and will allow for distributed intelligence over the whole IoT network, all-the-way down to ultra-low power end-devices. This will also open the way to unforeseen new applications. Nevertheless, to make this happen in a way that brings growth to society and benefits to individuals, several challenges still need to be tackled:

a) Despite the tremendous success of connectionist models (such as deep learning) in many important applications, our theoretical *understanding* of these systems is still far from complete. The complexity of the resulting systems makes it difficult to say which of their properties is most responsible for improved performance. Generalization in learning, abstraction, and reasoning abilities remains extremely limited, compared to human general intelligence. *Prediction* remains one of the fundamental problems in neural computation. Recently, neural networks were shown to fail while performing easy tasks where a human would never have failed (e.g., recognizing "*fooling images*", or images changed in a way imperceptible to humans [61]). Indeed, this threat limits market expansion, betrays user confidence, and gives rise to serious *ethical questions*. For these reasons, we believe that more understandable models should be developed, and more efforts should be put into the study of *neural network information and learning theories* [62, 63]. The biological plausibility of artificial systems should not be a burden for engineers' creativity. Nonetheless, we believe that more interactions between AI engineers, neuroscientists, and biologists will be strongly beneficial from a fundamental point of view.

b) The conceptual basis of the *embodied or enactive cognition paradigm* could be highly inspiring when defining new artificial systems suitable for the hyperconnected world. Future artificial cognitive systems will be autonomous physical systems which will need to interact in real time with the environment and individuals everywhere. Physical constraints will shape the dynamics of these interactions: In such systems, as with biological organisms, *the link between the low-level sensory-motor processes, control systems, and cognition will play a key role*. Bio-inspired approaches will force us to think differently. Simpler biological systems, rather than the human brain, will be highly inspirational and instructive. For example, the use of *insects as templates for artificial intelligent systems* [64] highlights the need to think in a systemic way, as *organisms do not decouple sensors and signal treatment*. Future autonomous systems will be required to perform intelligent tasks well beyond the possibilities of current ML systems (designed with a traditional input-output scheme and optimized to address classification tasks). The way they learn autonomously will be essential to define their predictive and interactive capabilities. Moreover, to account for the complexity of the world and the whole spectrum of future demands, it is probable that *a plurality of representational and cognitive architectural approaches* (based on cognitivist connectionist embodied-mind theories) will be needed, leading to *a world of heterogeneous and interconnected mixed-systems solutions*. Each approach will succeed in addressing different classes of empirical behaviors or will be more suitable for specific tasks.

c) Finally, *bio-hybrid interfaces between biological systems and VLSI neuromorphic systems* of varying complexity will play an important role in the future. Primarily intended as a computational tool for investigating fundamental questions related to neural dynamics, the sophistication of current neuromorphic systems makes direct interfacing with large neuronal networks and circuits possible, giving rise to interesting *clinical applications for neuroengineering systems, neuroprosthetics, and neurorehabilitation* [65, 66]. Leti's biomedical research center (Clinatec), dedicated to preclinical and clinical trials [4], is equipped with a cutting edge surgical operating room and medical facilities that are specifically and exclusively used for the qualification of advanced therapies and new prototypes based on micro-technologies. Here*, physicians, biologists, and engineers* work together to provide efficient and rapid validation of diagnostic and therapeutic tools using regulation-based evaluation processes. A high level of miniaturization and real-time data analysis were necessary to develop a *BCI* used in patients' rehabilitation [67] (see **Fig. 1.2.6**).

In the future, we will also witness the introduction of stimulation strategies based on real closed-loop systems, with signals emerging from wearable sensors (for example, sensing gloves). New materials to interface devices with living cells and tissues, new design architectures for lowering power consumption, data extraction and management at the system level, and secured communications are the next domains that will experience intense development. Brain-inspired implantable microdevices, acting as *intelligent neuroprostheses,* and *bio-hybrid systems* represent the new era of cross-disciplinary *brain-repair strategies*, where biological and engineered solutions will complement each-other, probably mediated by artificial intelligence [68, 69].

Acknowledgments - E. Beigne, S. Catrou, M. Belleville, A. Valantian, C. Reita, D. Dutoit, A. Hihi, E. Vianello, T. Dalgaty, S. La Barbera, P. Vivet, M. Causo, D. Morche, G. Sicard, A. Molnos, F. Heitzmann, L. Poupinet, S. Bonnetier from CEA-Leti, M. Duranton, C. Gamrat, A. Dupret, O. Bichler from CEA-List, A. Jerraya from CEA-DRT, B. Yvert from INSERM, Prof. G. Indiveri from ETH Zurich, Prof. J. Casas from University of Tours, Prof. S. Mitra from Stanford University.

References:

[1] "Human Brain MRI at 500MHz, Scientific Perspectives and Technological Challenges", Denis Le Bihan et al., Supercond. Sci. Technol., 30, 2017.
[2] "The Economy of Brain Network Organization", Ed Bullmore and at., Nature, 13, 2012.
[3] "The Embodied Mind", F.Varela et al, MIT Press, 1991.
[4] "Symbiotic Low-Power, Smart and Secure Technologies In the age of Hyperconnectivity", M.N. Semeria, IEDM 2016.
[5] "Efficient Embedded Learning for the IoT Devices", S. Venkataramani, IEEE 2016.
[6] "Can Deep Learning Revolutionize Mobile Sensing?", N.D. Lane et al., ACM International Workshop on Mobile Computing Systems and Applications, 2015.
[7] "A 0.3-2.6 TOPS/W Precision-Scalable Processor for Real-Time Large-Scale ConvNets", B.Moons et al., VLSI 2016.
[8] "A 288μW Programmable Deep-Learning Processor with 270KB On-Chip Weight Storage Using Non-Uniform Memory Hierarchy for Mobile Intelligence," S. Bang et al., ISSCC 2017.
[9] "Ultra-Low-Power Networked Systems", A. Chandrakasan, Nano Tera Workshop 2015.
[10] "Spendthrift: Machine Learning Based Resource and Frequency Scaling for Ambient Energy Harvesting Nonvolatile Processors," K. Ma et al., 22nd Asia and South Pacific Design Automation Conf., p.678, 2017.
[11] "A Learning Theoretic Approach to Energy Harvesting Communication System Optimization", P. Blasco et al., IEEE Trans. on Wireless Comm., 12, 4, p. 1872, 2013.
[12] "Twin Neurons for Efficient Real-World Data Distribution in Networks of Neural Cliques: Applications in Power Management in Electronic Circuits", B. Boguslawski et al., IEEE Trans. on Neural Networks and Learning Systems, 27, 2, p.375, 2016.
[13] "A 55nm Time-Domain Mixed-Signal Neuromorphic Accelerator with Stochastic Synapses and Embedded Reinforcement Learning for Autonomous Micro-Robots", A. Amravati et al., ISSCC 2018.
[14] "An Approach to Implement Data Fusion Techniques in Wireless Sensor Networks Using Genetic Machine Learning aAlgorithms", A.R. Pinto et al., Information Fusion, 15, p.90, 2014.
[15] "Machine Learning Methods in Data Fusion Systems", R. Nowak et al., 13th International Radar Symposium, p.400, 2012.
[16] "Computing's Energy Problem (and what we can do about it)", M. Horowitz, pp. 10-14, ISSCC 2014.
[17] "Analog VLSI and Neural Systems", C. Mead, Addison-Wesley VLSI Systems Series, 1989.
[18] "Advanced Technologies for Brain-Inspired Computing", F. Clermidy et al., IEEE 2014.
[19] "Planar Fully-Depleted-Silicon-On-Insulator Technologies: Past Research, Current Status and Future Directions", B.Doris et al., Solid State Electronics, 117, p.37, 2016.
[20] "ENVISION: A 0.26-to-10TOPS/W Subword-Parallel Dynamic-Voltage-Accuracy-Frequency-Scalable Convolutional Neural Network Processor in 28nm FDSOI", B.Moons, et al, p. 246, ISSCC 2017.
[21] "A 2.9TOPS/W Deep Convolutional Neural Network SoC in FD-SOI 28nm for Intelligent Embedded Systems", G.Desoli et al., p. 238, ISSCC 2017.
[22] "Scaling Mixed-Signal Neuromorphic Processors to 28 nm FD-SOI Technologies", N. Qiao et al., IEEE Biomedical Circuits and Systems Conf., 2016.
[23] "Analog Circuits for Mixed-Signal Neuromorphic Computing Architectures in 28 nm FD-SOI Technology", N. Qiao et al., IEEE S3S, 2017.
[24] "Neuromorphic Architectures for Spiking Deep Neural Networks", G.Indiveri et al., IEDM 2015.
[25] "Philosophy of the Spike: Rate-Based vs. Spike-Based Theories of the Brain", R. Brette, Frontiers in Systems Neuroscience, 9:151, 2015.
[26] "ITAC: A Complete 3D Integration Test Platform", D. Lattard et al., 3DIC 2016.
[27] "New Perspectives for Multicore Architectures Using Advanced Technologies", F. Clermidy et al., IEEE IEDM 2016.
[28] B. Belhadj et al. CASSES'2014.
[29] "The Neocortical Circuit: Themes and Variations", Harris KD et al., Nat Neurosci., 2, p.170, 2015.
[30] "3D Sequential Integration: Application-Driven Technological Achievements and Guidelines", P. Batude et al., IEDM 2017.
[31] "Energy-Efficient Abundant-Data Computing: The N3XT 1,000x", M. M. Sabry Aly et al., IEEE Computer, 48, 12 2015.
[32] "Three-Dimensional Integration of Nanotechnologies for Computing and Data Storage on a Single Chip", M. M. Shulaker et al., Nature, 547, p.74, 2017.
[33] "Universal Signatures from Non-Universal Memories: Clues for the Future...", L; Perniola, IEEE IMW, 2016.
[34] "Non-Volatile Memory Evolution and Revolution", P. Cappelletti, IEDM 2016.
[35] "A Wafer-Scale Neuromorphic Hardware System for Large-Scale Neural Modeling", J. Schemmel et al., IEEE Int. Symp. on Circuits and Systems, 2010.
[36] "A Million Spiking-Neuron Integrated Circuit with a Scalable Communication Network and Interface", P. A. Merolla et al., Science, 345, 6197, p.668, 2014.
[37] "TrueNorth Ecosystem for Brain-Inspired Computing: Scalable Systems, Software, and Applications", J. Sawada et al., ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis, 2016.
[38] "A Reconfigurable On-Line Learning Spiking Neuromorphic Processor Comprising 256 Neurons and 128K Synapses", N. Qiao et al., Frontiers Neuroscience, 9, p.141, 2015.
[39] "From Memory in our Brain to Emerging Resistive Memories in Neuromorphic Systems", B. DeSalvo, IEEE IMW, 2015.
[40] "Oxide Based Nanoscale Analog Synapse Device for Neural Signal Recognition System", D. Lee et al., IEEE IEDM 2015.
[41] "Large-Scale Neural Networks Implemented with Non-Volatile Memory as the Synaptic Weight Element: Comparative Performance Analysis (Accuracy, Speed, and Power)", G.W. Burr et al., IEEE IEDM 2015.
[42] "HfO2-based OxRAM Devices as Synapses for Convolutional Neural Networks", D. Garbin et al., IEEE Tr. On El. Devices, 2015. [43] "Synapses Made by Two Phase-Change Memory Devices for Efficient Spiking Neural Networks", O. Bichler et al., IEEE Tr. On El. Devices, 2012.
[44] "Phase Change Memory as Synapse for Ultra-Dense Neuromorphic Systems: Application to Complex Visual Pattern Extraction", M.Suri et al., IEDM 2011.
[45] "Bio-Inspired Stochastic Computing Using Binary CBRAM Synapses", M. Suri et al., IEEE Trans. on Electron Device, 60, 7, 2402, 2013.
[46] "Spiking Neural Networks Based on OxRAM Synapses for Real-Time Unsupervised Spike Sorting", T. Werner et al. Frontiers in Neuroscience, 10, 474, 2016.
[47] "Resistive Memories for Spike-Based Neuromorphic Circuits", E. Vianello et al., IMW 2017.
[48] "Spintronic Devices as Key Elements for Energy-Efficient Neuroinspired Architectures", N. Locatelli et al, Design, Automation & Test in Europe Conference & Exhibition, p. 994, 2015.
[49] "Stochastic Neuron Design Using Conductive Bridge RAM", G. Palma et al., IEEE/ACM International Symposium on Nanoscale Architectures, 2013.
[50] "Light Is the Ultimate Medium for High-Speed Communications", C. Kopp et al., EuroPhotonics, 2017.
[51] "Low-Temperature Crack-Free Si3N4 Nonlinear Photonic Circuits for CMOS-Compatible Optoelectronic Cointegration", M. Casale et al., SPIE Photonics West, OE109, 2017.
[52] "10 Gbps, 560 fJ/b TIA and Modulator Driver for Optical Networkson-Chip in CMOS 65nm", J.L. Gonzalez et al., 14th IEEE Intern. NEWCAS Conference, 2016.
[53] "Neuromorphic Photonic Networks Using Silicon Photonic Weight Banks", A.N. Tait et al., Scientific Reports 7, 7430, 2017. [54] "High-Speed Photonic Reservoir Computing Using a Time-Delay-Based Architecture: Million Words per Second Classification", L. Larger et al., Phys. Rev. X 7, 2017.
[55] "Reservoir Computing Approaches to Recurrent Neural Network Training", M. Lukoševicius et al., Computer Science Review, 3, 3, p.127, 2009.
[56] "A 128 x128 120dB 30mW Asynchronous Vision Sensor that Responds to Relative Intensity Change", P. Lichtsteiner et al., IEEE ISSCC 2006.
[57] "A QVGA 143dB Dynamic Range Asynchronous Address-Event PWM Dynamic Image Sensor with Lossless Pixel-Level Video Compression", C. Posch et al., IEEE ISSCC 2010.

[58] "Low-Power Spiking Chemical Pixel Sensor", P. Georgiou et al., Electronics Letters 42, 23, p.1331, 2006.
[59] "A 0.5V 55µW 64×2-Channel Binaural Silicon Cochlea for Event-Driven Stereo-Audio Sensing", M. Yang et al., IEEE ISSCC, p. 388, 2016.
[60] "Spike-Based Readout of POSFET Tactile Sensors", S. Caviglia et al., IEEE Trans. on Circuits and Systems, 64, 6, p.1421, 2017.
[61] "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images", A. Nguyen et al., IEEE Computer Vision and Pattern Recognition (CVPR), 2015.
[62] "Theory of Deep Learning I, II, III", C. Zhang et al., Tech. Rep., MIT Center for Brains, Minds and Machines, 2017.
[63] "Deep Learning", Y. LeCun et al., Nature, p.436, 2015.
[64] "Biomimetic Flow Sensors", J.Casas et al., Encyclopedia of Nanotechnology, Springer Verlag, 264, 2012.
[65] "Generation of Locomotor-Like Activity in the Isolated Rat Spinal Cord Using Intraspinal Electrical Microstimulation Driven by a Digital Neuromorphic CPG", S. Joucla et al., Frontiers in Neuroscience, 10, 67, 2016.
[66] "Real-Time Control of An Articulatory-Based Speech Synthesizer for Brain-Computer-Interfaces", F. Bocquelet et al., PLoS Comput. Biol., 12, 11, 2016.
[67] "WIMAGINE®: Wireless 64-Channel ECoG Recording Implant for Long Term Clinical Applications", C. Mestais et al., IEEE Trans. on Neural Systems and Rehabilitation Engineering, 23, 1, 2015.
[68] "Intelligent Biohybrid Systems for Functional Brain Repair", G. Panuccio et al., New Horizons in Translational Medicine, 3, p.162, 2016.
[69] "Trends and Challenges in Neuroengineering: Toward "Intelligent" Neuroprostheses Through Brain-"Brain Inspired Systems", S. Vassanelli et al., Communication. Front. Neurosci. 10, 438, 2016.

Figure 1.2.1: Atlas of brain connectivity, showing the long white matter fiber bundles (connections are visualized using diffusion MRI). A specific color is attributed to each fiber bundle [1].

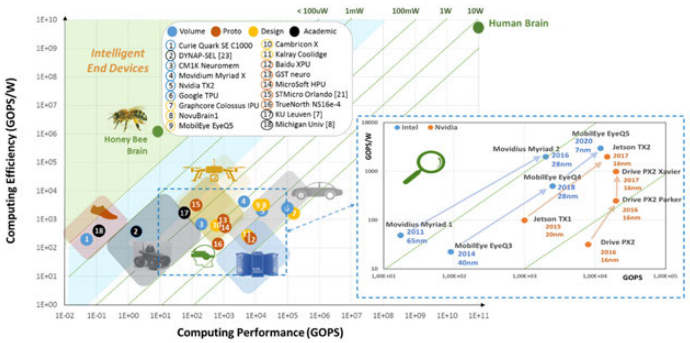


Figure 1.2.2: Comparison of computing efficiency (GOPS/W) during the inference phase versus computing performance (GOPS) of several intelligent chips from literature and the web, showing the gap between the intelligent end-device requirements and existing solutions. *Note that we took the very coarse approximation of a 1:1 correspondence between OPS, FLOPS, IPS, SOPS (SOPS = firing rate × average active synapses).*

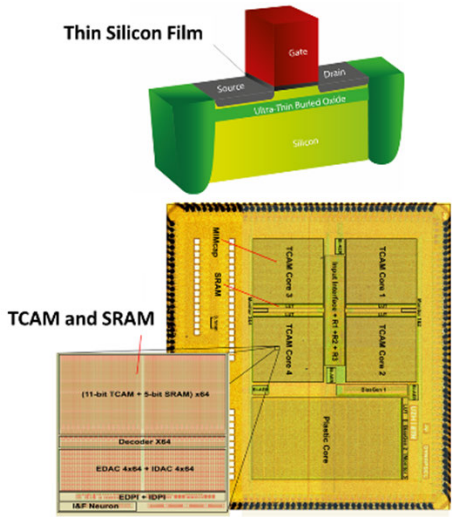


Figure 1.2.3: Dynap-SEL neuromorphic chip, based on a 28nm FDSOI process [22, 23].

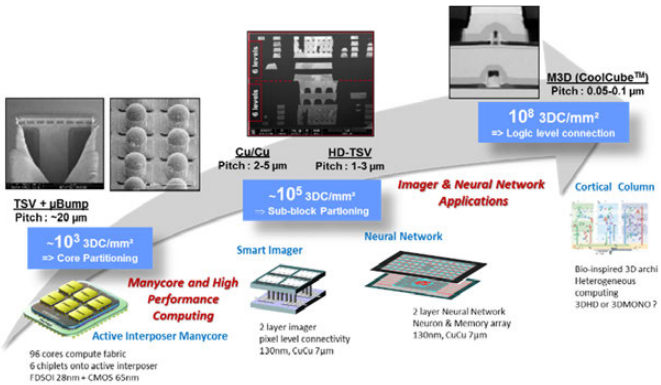


Figure 1.2.4: CEA-Leti roadmap of 3D technologies, showing the connection-density evolution and corresponding hardware applications.

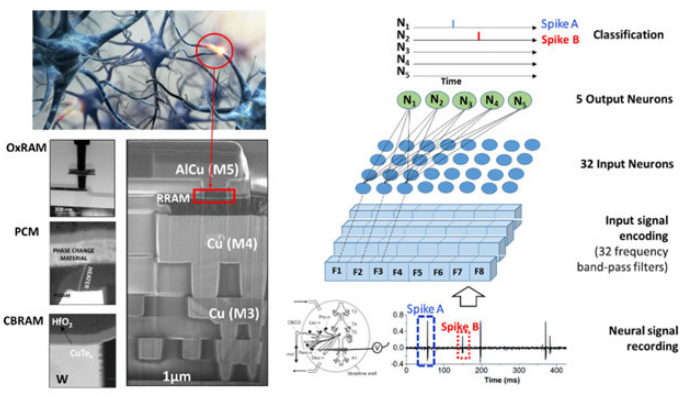


Figure 1.2.5: Left: Illustration of a biological synapse and the concept of using ReRAM as synapses. Right: Functional schematic of a spiking neural network for real-time unsupervised spike sorting [46].

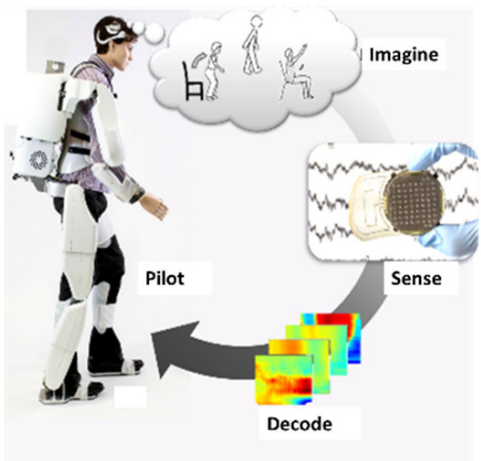


Figure 1.2.6: Illustration of the BCI project, showing functional substitution for tetraplegic subjects via a driven 4-limb exoskeleton [67]. In the future, intelligent neuroprostheses and biohybrid systems for therapeutic purposes are foreseen [68, 69].

 978-1-5090-4940-0/18/$31.00

ISSCC AWARDS

2017 Lewis Winner Award for Outstanding Paper

"A 28GHz 32-Element Phased-Array Transceiver IC with Concurrent Dual Polarized Beams and 1.4 Degree Beam-Steering Resolution for 5G Communication"

Bodhisatwa Sadhu[1], Yahya Tousi[1], Joakim Hallin[2], Stefan Sahl[3], Scott 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], 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

2017 Distinguished-Technical-Paper Award

"**A 12b 10GS/s Interleaved Pipeline ADC in 28nm CMOS Technology**"

Siddharth Devarajan[1], Larry Singer[1], Dan Kelly[1], Steve Kosic[2], Tao Pan[1], Jose Silva[1], Janet Brunsilius[2], Daniel Rey-Losada[2], Frank Murden[3], Carroll Speir[3], Jeff Bray[2], Eric Otte[1], Nevena Rakuljic[2], Phil Brown[3], Todd Weigandt[2], Qicheng Yu[1], Donald Paterson[1], Corey Petersen[2], Jeffrey Gealow[1]

[1]Analog Devices, Wilmington, MA; [2]Analog Devices, San Diego, CA
[3]Analog Devices, Greensboro, NC

2017 Jan Van Vessem Award for Outstanding European Paper

"A 1.1W/mm²-Power-Density 82%-Efficiency Fully Integrated 3:1 Switched-Capacitor DC-DC Converter in Baseline 28nm CMOS Using Stage Outphasing and Multiphase Soft-Charging"

Nicolas Butzen, Michiel Steyaert

KU Leuven, Leuven, Belgium

2017 Takuo Sugano Award for Outstanding Far-East Paper

"A Reconfigurable Bidirectional Wireless Power Transceiver with Maximum-Current Charging Mode and 58.6% Battery-to-Battery Efficiency"

Mo Huang[1,*], Yan Lu[1], Seng-Pan U[1,2], Rui P. Martins[1,3]

[1]University of Macau, Macau, China; [2]Synopsys Macau, Macau, China
[3]Instituto Superior Tecnico, Universidade de Lisboa, Portugal
[*]now with South China University of Technology, Guangzhou, China

ISSCC 2017 Jack Kilby Award for Outstanding Student Paper

"All-Wireless 64-Channel 0.013mm2/ch Closed-Loop Neurostimulator with Rail-to-Rail DC Offset Removal"

Hossein Kassiri[1]*, M. Reza Pazhouhandeh[1], Nima Soltani[2], M. Tariqus Salam[2], Peter Carlen[1,3], Jose Luiz Perez Velazquez[1], Roman Genov[1]

[1]*University of Toronto, Toronto, Canada
[2]GSK (GlaxoSmithKline), Stevenage, United Kingdom
[3]Toronto Western Hospital, Toronto, Canada
*now at York University, Toronto, Canada

2017 ISSCC Award for Outstanding Forum Presenter

"Pushing the Boundaries of Performance – A Technology Perspective"

Marcel Pelgrom

Pelgrom Consult, Helmond, the Netherlands

2017 ISSCC Award for Outstanding Forum Presenter

"High-Performance Clock Generation and Distribution in Very-High-Speed Wireline Transceivers"

Nicola Da Dalt

Intel, San Jose, CA

2017 Evening Session Award

***"Return of* Survey Says!"**

Organizer: **Harry Lee**, MIT, Cambridge, MA
Co-Organizer: **Matt Straayer**. Maxim Integrated, North Chelmsford, MA
Moderator: **Chris Mangelsdorf**, Analog Devices, San Diego, CA

Panelists:
Robert Adams, *Analog Devices, Wilmington, MA*
Lucien Breems, *NXP Semiconductors, Eindhoven, The Netherlands*
Yun Chiu, *University of Texas, Dallas, TX*
Michael Choi, *Samsung Electronics, Yongin, Korea*
Ian Galton, *University of California San Diego, La Jolla, CA*
Shanthi Pavan, *IIT Madras, Chennai, India*
Kathleen Philips, *imec/Holst Centre, Eindhoven, The Netherlands*
Ken Poulton, *Keysight Labs, Santa Clara, CA*

2017 Demonstration Session Certificate of Recognition

"A 1ms High-Speed Vision Chip with 3D Stacked 140GOPS Column-Parallel Pes for Spatio-Temporal Image Processing"

Tomohiro Yamazaki[1], Hironobu Katayama[1], Shuji Uehara[1], Atsushi Nose[1], Masatsugu Kobayashi[1], Sayaka Shida[1], Masaki Odahara[2], Kenichi Takamiya[2], Yasuaki Hisamatsu[2], Shizunori Matsumoto[2], Leo Miyashita[3], Yoshihiro Watanabe[3], Takashi Izawa[1], Yoshinori Muramatsu[1], Masatoshi Ishikawa[3]

[1]Sony Semiconductor Solutions, Atsugi, Japan
[2]Sony LSI Design, Atsugi, Japan; [3]University of Tokyo, Bunkyo, Japan

2017 Demonstration Session Certificate of Recognition

"A 0.1-to-3.1GHz 4-Element MIMO Receiver Array Supporting Analog/RF Arbitrary Spatial Filtering"

Linxiao Zhang, Harish Krishnaswamy

Columbia University, New York, NY

2018 Silkroad Award

"A 22.8-to-43.2GHz Tuning-Less Injection-Locked Frequency Tripler Using Injection Current Boosting with 76.4% Locking Range for Multi-band 5G Applications"

Jingzhi Zhang

University of Electronic Science & Technology of China, Chengdu, China

"QUEST: A 7.49-TOPS Multi-Purpose Log-Quantized DNN Inference Engine Stacked on 96MB 3D SRAM Using Inductive-Coupling Technology in 40nm CMOS"

Kodai Ueyoshi

Hokkaido University, Hokkaido, Japan

ISSCC 2017 Student-Research Preview (SRP) Award

"High Dynamic Range SPAD Pixel for Time of Flight 3D Imaging"

Francescopaolo Mattioli Della Rocca

University of Edinburgh

"A 310-Fs RMS Jitter Injection-Locked Multi-Frequency Generator Using a Time-Interleaved Multi-DCO Calibrator"

Heein Yoon

Ulsan National Institute of Science and Technology

"A 0.55V 1.5mW 2.4GHz All-Digital Fractional-N PLL with PVT-Insensitive TDC Using Switched-Capacitor Doubler in 28nm CMOS"

Seyednaser Pourmousavian

University College Dublin

IEEE SOLID-STATE CIRCUITS SOCIETY AWARDS

2016 Journal of Solid-State Circuits Best Paper Award

A 2.2 GHz Continuous-Time ΔΣ ADC With –102 dBc THD and 25 MHz Bandwidth

Lucien Breems[1], Muhammed Bolatkale[1], Hans Brekelmans[1], Shagun Bajoria[1], Jan Niehof[1], Robert Rutten[1], Bert Oude-Essink[2], Franco Fritschij[2], Jagdip Singh[2], Gerard Lassche[2]

[1]NXP Semiconductors, Eindhoven,The Netherlands
[2]Catena Microelectronics, Delft,The Netherlands

IEEE Technical Field Awards

2018 IEEE Daniel E. Noble Award for Emerging Technologies

Rajiv V. Joshi
IBM T. J. Watson Research Center, Yorktown Heights, NY

"For Contributions to Predictive Failure Analytics, VLSI Memory Design, and Technology"

2018 IEEE Donald O. Pederson Award in Solid-State Circuits (co-recipients)

William S. Carter, Xilinx, San Jose, CA

and

Stephen M. Trimberger, Xilinx, San Jose, CA

"For Contributions to Field-Programmable Gate Array Technology."

2018 IEEE Frederik Philips Award

Ian A. Young
Intel, Hillsboro, OR

"For Leadership in Research and Development on Circuits and Processes for the Evolution of Microprocessors"

2018 IEEE FELLOWS

Pamela Ann Abshire
Silver Spring, MD

"For Contributions to CMOS Biosensors"

Pietro Andreani
Lund University, Lund, Sweden

"For Contributions to CMOS Integrated Voltage-Controlled Oscillators"

Bertan Bakkaloglu
Arizona State University-Tempe, Scottsdale, AZ

"For Contributions to Radio Frequency Circuits"

Kun-Yung Chang
Los Altos Hills, CA

For Contributions to Transceivers for High-Performance Networking and High-Density Memories"

Tsung-Yung Chang
Taiwan Semiconductor Manufacturing Company, Hsinchu, Taiwan

"For Application of SRAM Technology to Low-Power and High-Performance Computing"

Andreas Demosthenous
University College London, London, UK

"For Contributions to Integrated Circuits for Active Medical Devices"

Isaac Lagnado
San Diego, CA

"For Leadership in the Development of Silicon-on-Sapphire Technology"

Sanu Mathew
Intel Corporation, Hillsboro, OR

"For Leadership in Computer Arithmetic Datapath and Security Circuits"

Earl Mc Cune
RF Communications Consulting, Santa Clara, CA

"For Leadership in Polar Modulation Circuits and Signals"

Saibal Mukhopadhyay
Georgia Institue of Technology, Atlanta, GA

"For Contributions to Energy-Efficient and Robust Computing Systems Design"

Hidetoshi Onodera
Kyoto University, Kyoto, Japan

"For Contributions to Variation-Aware Design and Analysis of Integrated Circuits"

Shanthi Pavan
Indian Institute of Technology, Chennai, India

"For Contributions to Delta Sigma Modulators and Analog Filters"

Rahul Sarpeshkar
Thayer School of Engineering at Dartmouth, Hanover, NH

"For Contributions to Ultra Low-Power Biomedical Electronics"

978-1-5090-4940-0/18/$31.00

1.3 Future Mobility- Enhanced Society Enabled by Semiconductor Technology

Yukihiro Kato, Senior Executive Director

DENSO, Aichi, Japan

1. Introduction

The automotive industry is in the midst of a once-in-a-century transformation. The industry began adopting semiconductors in the 1960s, starting with rectifier diodes for alternators. Then, the spread of electronically controlled engines that supported emission control in the 1970s triggered wide-scale installation of semiconductor devices (Figure 1.3.1). The automotive semiconductor market has grown rapidly since the 1980s [1]. Further growth is expected as electrification, autonomous vehicles, and connected vehicles develop (Figure 1.3.2). In the past, when semiconductors were introduced into automobiles, the purpose was to improve performance by replacing conventional mechanical controls with electronic ones; it is important to note that no changes were made to the major components. But, for the projected disruptive transformation (already underway), dramatic changes to all major automobile components and related concepts are expected. Electrification (replacing the internal combustion engine (ICE) with an electric motor), automated driving (replacing the human driver with artificial intelligence (AI)), and connected vehicles (replacing independent automobiles with ones that are incorporated into society) will play important roles in this transformation as the automobile evolves in a major way.

To achieve this evolution, semiconductor technology must also evolve. Semiconductor technology has progressed steadily according to Moore's Law since 1965: Digital integrated-circuit density has increased, and high-speed processing with low power consumption is now available. In addition, technology has improved the device characteristics of analog circuits in terms of their high-frequency performance, enabling the implementation of various wired and wireless systems. Figure 1.3.3 shows the requirements for automotive semiconductors. The automotive industry has typically adopted semiconductors that are several generations older than those in the latest consumer products, because reliability and a proven record have been the industry's first priority in the market. This trend, however, will surely change, as state-of-the-art technology is required to satisfy future demands for the automobile. Unfortunately, little attention has been paid to power semiconductor device design, as the need for such devices has been limited to the railroad and power generation industries. However, in recent years, the automotive industry has emerged as a new market for power semiconductor devices. New power semiconductor technology is being developed, and the proliferation of these devices in automotive power control has begun. As well, these new semiconductor technologies will be deployed in future automobiles as automated driving becomes popular. The automobile will come to be regarded not just as property, simply as utilitarian or to emphasize status, but rather as a mobility tool that brings the joy of driving, riding and meeting people. That will also open the door on a new world in the social system.

2. Automotive Industry Trend

In the past, the automobile has provided a means of transport; however in the future, it will change to provide passengers on-the-move with better utilization of time and space. In this new world, cities will be designed in combination around the evolving functions of automobiles. Industries there must offer both convenience and security, such as one-time reservation of smooth transportation, and also provide solutions to issues for sustaining new generation of the society. Utilizing clean energy such as electricity and hydrogen, the automobile of the future will evolve to make various decisions, based not only on information from various embedded sensors, but also on cloud information (such as a car or pedestrian in blind spots, traffic conditions, or weather forecasts). At the same time, the advanced automobile will be able to detect health-and-alertness condition of the driver and passengers, bringing us to a state where people and automobiles can coexist in a harmonious society. To employ external data requires a communications network having extreme integrity, with a secure system that can receive up-to-the-second authentic information. In building such a system having high reliability, the concept of functional safety is important; international standards for which are currently under way. Figure 1.3.4 depicts an example of a corresponding future urban transport network.

With the arrival of electrification, automated driving and connected vehicles, coordination and optimization of transport networks employing multi-modal transportation, valet parking, and collaboration with supply chains will become the norm. These functions will be centered on a cyber mobility center that can make forecasts based on data analysis, and provide updates and visualizations of the latest information. This level of automated transportation system will operate highly efficiently by merging information and mobility, allowing environmentally friendly transportation of goods and people at low cost. Traffic accidents can be reduced by using advanced driver-assistance systems (ADAS) and automated driving technology. Mobility will become ever-easier for older people, and none will suffer in traffic jams under a highly-controlled traffic system. As information will be shared seamlessly from homes to moving vehicles, a medical environment that brings confidence and safety can be also expected. But, there are many technological challenges to be overcome in preparing this bright future for coming generations!

3. Electrification

Electrification of automobile powertrains will progress further in the near future. Although ICEs will be used for power generation and as an auxiliary power source, electrification will be indispensable to achieving the efficient use of renewable energy: Although ICEs will still be used in power generation, electric motors will form the main powertrain to improve energy efficiency throughout society. As the cruising range of electric vehicles is improved, downsizing and weight reduction of components are key factors for development. Figure 1.3.5 shows the configuration of a typical electrified automobile. In such a vehicle, the fuel tank is replaced by batteries, the ICE by an electric motor, and the engine control unit by a power control unit (PCU). These occupies fairly large space in a car to radiate excess heat. Thus, one of the biggest development challenges is to increase the efficiency and decrease the size of the PCU. Figure 1.3.6 shows power devices adopted for various inverters. Surprisingly, an automotive motor actually requires about as much power as that for a single train (100-to-150kW)! However, the space allowed for installing a PCU is less than one cubic meter. Due to the proximity of the powertrain and cabin space in an automobile, there is less space for heat radiation, so efficiency must be enhanced. To improve efficiency, the development of high-speed switching devices has been proposed. The railway industry has improved efficiency by replacing gate turn-off thyristor power devices with insulated-gate-bipolar-transistor (IGBT) devices. IGBTs have been optimized for use in vehicles [2]. In automobiles, the IGBT is mounted on a double-sided heat radiating package [3] using water-cooled engine technology, achieving a PCU with the world's highest efficiency. The most promising candidate for further efficiency improvement is to replace the IGBT device with an SiC MOSFET [4]. In railway applications, the adoption of SiC MOSFETs, characterized by high efficiency (achieved by their MOS structure) and by high-temperature operation (because of their wide bandgap), is expanding in recent years. As well, power devices for PCUs require a high breakdown voltage and low on-resistance. These are determined by the characteristics of the semiconductor material and the device structure. SiC is a wide bandgap semiconductor material with a breakdown voltage resistance of 3MV/cm, which is one or more orders-of-magnitude higher than for pure Si. Thus, SiC-MOSFET devices can be thinner than IGBT devices. As well, the drift layer where electrons pass through is thin, and as a result, both a low on-resistance and a high breakdown can be achieved simultaneously. As the saturated drift velocity of SiC (2km/s) is twice that of Si, high-speed switching is possible. As a result, further efficiency improvements can be achieved. As previously indicated, by developing power devices with structures that utilize the material properties of SiC, PCU cooling can be changed from water cooling to air cooling. If air cooling is adopted, the piping for the cooling water can be eliminated, and drastic downsizing and weight reductions can be achieved. In addition, the high-speed switching performance of the SiC MOSFET is effective in downsizing future wireless charging systems, enabling wireless charging on the road. One of the greatest challenges in SiC-MOSFET development is to reduce the defect density on the surface of these devices. In our search for improvement, we have adopted the repeated a-face (RAF) method to prototype a highly reliable low-defect SiC single-crystal mass-producible 6-inch wafer (Fig. 1.3.7) [5]. As well, GaN is expected to be a candidate for power device materials after SiC. However, GaN requires further improvements in terms of efficiency and cost for automotive usage.

4. Automated Driving

In recent years, there has been much technological research and development directed to automated driving. Some automakers have already reported the results of actual driving tests on public roads and highways, and of unmanned driving tests in dedicated areas. In automated driving, multiple targets must be recognized simultaneously with high precision as shown in Fig. 1.3.8. Sensors for automated driving include cameras (image sensors) and millimeter-wave radar that can detect targets under adverse conditions such as in total darkness or in fog. Ultrasonic sensors for short-range detection and high-precision LiDAR mapping are also employed [6]. As the costs of such sensors fall, multiplicity of sensors have been installed in automobiles. The main components for imaging are an image sensor and an image processor, which also includes a microcomputer that performs calculations to output signals for warnings, driver assistance, and automated driving. The technology required for image processing must allow the recognition of objects quickly and with low power. Generally speaking, image-processor recognition accuracy has a trade-off relationship with power consumption; thus, reducing processor power consumption while maintaining performance will improve competitiveness. Heat generated by semiconductors in automotive use raises a major issue, since an automobile has very space limited. Facilities for cooling hot semiconductors raises production costs.

In addition to the power problem resulting from motive control, there is yet another associated with supporting the required intelligent environment. Although automated recognition through machine learning, based on deep neural networks (DNNs) and artificial intelligence, is making progress, obtaining realistic power consumption for automotive environment with DNNs based on a general-purpose GPU is still a challenge. Currently individual object recognition performed using conventional signal processing may be more power efficient. There are two methods to reduce power consumption: One is to increase the efficiency of the DNN itself by improving the software, and the other is to adopt DNN-dedicated scalable hardware for the required latency [7]. In machine learning, performance is determined by the quality and quantity of the training data. In automotive usage, all learning is implemented before being shipped. For automotive usage that does not require learning after being shipped, use of a dedicated IP is desirable to provide cost reduction and low power consumption. Potentially, a DNN IP can achieve the same level of recognition as a human driver, as portrayed in Fig. 1.3.9. It can quickly perform the necessary processes, such as determination of targets' motion directions and spaces that are irrelevant to driving, that are challenging for conventional signal processors. Generally, it is important to have installed DNNs equipped with knowledge about all possible situations that can be expected in a real driving scene and not to be faced with the unexpected after its shipment.

Millimeter-wave radar enables highly precise target detection, ranging from far to near distances. For these detectors, angle resolution has been vastly improved by increasing the number of channels (antennas) [8], and by adopting an electronic scanning method based on phase control (Fig. 1.3.10). High-frequency at first, 24GHz and 76GHz devices employed GaAs HEMTs. At present, more SiGe bipolar devices have been used with production volume growing as millimeter-wave radar becomes popular. In the future, it is expected that more CMOS devices will be adopted to support the increased number of digital elements, along with the growing complexity of control [9]. As CMOS scaling progresses, with f_{max} reported to exceed 300GHz even in 55nm technology, more and more such devices will be integrated in millimeter-wave radars. Adoption of more CMOS devices will enable the integration of more-complex chips as shown in Fig.1.3.11, which will spur further downsizing and cost reductions. As well, higher levels of integration through multiple antennas and phased-array technology are being discussed for 5th generation (5G) cellular communications, providing further progress in high-frequency technology of relevance to automotive sensor electronics.

In applications where route calculations are performed based on information from image processors and millimeter-wave radar data, the data flow processor (DFP,) shows great promise [10]: For automated driving, both CPUs and GPUs are not ideal: while CPUs are well suited to efficiently perform complex processing, they are relatively slow; while GPUs are much faster at performing large-scale parallel processing operations involving large numbers of identical repetitive operations, they are not suitable for performing complex tasks. On the other hand, the DFP that we have developed is a processor that will be well suited for high speed processing that matches the pace of human natural reflexes. Unlike CPUs and GPUs, it will possess unique characteristics, being able to handle multiple calculations, and to perform them using flexible parallel processing. As depicted in Fig. 1.3.12, the DFP can flexibly create multiple processes, with power consumption less than one-tenth of that of a GPU: In the DFP, by changing the border per-calculation-content and the length of calculation, efficient processing is possible without wasting processing time. Using a DFP enables responsive judgment much faster than with a GPU alone. As illustrated in Figure 1.3.13, for route calculation, a GPU will calculate all possible directions, while in the DFP, more efficient calculation is enabled by dropping the path which is unnecessary. Overall, power consumption continues to be of great concern: in this respect, it is important to note that, currently, for fast response-signal processing, most power is consumed by the image processor. In general, further technical advancement is required to achieve low power consumption for all automotive usage.

To build a high-precision surround-monitoring system that performs route calculations based on sensing information (as described above), the following steps are essential: thorough field testing, quantitative evaluation through simulation, and quantitative evaluation through test-course driving. As well, reliability requirements for automated driving cannot be supported by considerations of component lifetime and reduction of individual failure alone; rather, the safety of the entire system (defined as functional safety) is paramount! Thus, both the reliability defined by the Automotive Electronics Council (AEC) Quality Criteria 100 (Q100), and support for functional safety defined by the Automotive Safety Integrity Level (ASIL) are indispensable for automotive semiconductor industry. Furthermore, in terms of manufacturing technology, establishment of testing and implementation technologies for DNNs and millimeter-wave radar is essential. Overall, the device market for automated driving including sensors and signal processors is expected to expand greatly, along with further market growth expected for use in infrastructure and individual homes.

5. Connected Vehicles

Connected vehicles are expected to lead to the creation of new businesses targeted at improved convenience, such as cooperation with home security and home delivery services (Fig. 1.3.14) [11]. To facilitate these connectivity applications, a Plug & Play function similar to a smartphone must be enabled in the automobile. At present, because an application is connected to its basic function through a dedicated interface (I/F) too little flexibility is allowed. Adopting service-oriented architecture (SOA) and a common I/F will enable Plug & Play and the use of a variety of services.

There is a growing tendency to reduce latency as much as possible through the use of 5G communications to obtain the speed necessary for automotive usage [12]. The latency requirement level does not depend on a communication partner such as a car or the cloud, but rather depends on the purpose of the communication. To achieve an advanced level of control, it is necessary to make the latency much shorter than that in the previous system. Various standards for automotive communications (called V2X (vehicle-to-everything)) are under discussion at the IEEE [13]. Unlike cellular phones, the lifetime of an automotive wireless-communication product may exceed 10 years after launch. Thus, extra care must be taken in considering the flow of standardization and the timing of product commercialization.

As the concept of "best-effort" communication, and the requirements for confidence and safety are contradictory, communication security must be considered from the device level [14-to-16]. Currently, automotive communication using the cellular-phone network is secured by having a common key for each SIM card and exclusive chip. In the near future, SOCs (security operation centers) will monitor communications to identify hacking and assist recovery, thereby continuously improving security. Secured communication for exclusive use by the automotive sector will be ensured by a wireless chip with built-in public key using a PKI (public key infrastructure). Ultimately, operation centers will collect surveillance data obtained by the in-vehicle devices, improving security continuously as well. As connected vehicles become more common, the transfer of huge amounts of information between the driver/passengers and the automobile will become the norm, increasing the importance of human-machine interfaces (HMI): While, in the past, information was conveyed to the driver through video, audio, or even vibration, in the future,

 978-1-5090-4940-0/18/$31.00

optimal methods that integrate information from multiple sources according to its importance, must be adopted. However, projected modern infotainment and the aforementioned vehicle control information flow are in conflict: a trade-off will be necessary! Potentially, adapting virtual technology developed by the IT industry, to the automotive situation will enable the merging of both open content and vehicle control in a secure and robust manner. Equipment such as head-up displays and danger-sensing alarm seats [17] can convey information to a driver efficiently without interfering with the driving task; such interfaces will be essential even in the era of automated driving. HMIs are not simply one-way (providing information only to the driver), but rather two-way (providing adequate services by monitoring the driver): the HMI can provide information on important conditions, namely the driver's state, environmental state, and vehicle state, to support appropriate driving and comfort assistance by feedback to the driver. That is to say, all information is managed by the HMI-ECU Manager to support the driver (see Fig. 1.3.15). To achieve this, a large number of sensors are needed to monitor the driver's state, including facial expression, gestures, body temperature, voice, and brain waves, as well as the driving environment. In addition, even with a driver, for some interactions, no human intervention is necessary (such as for automated parking and emergency automatic braking).

Overall, connected vehicles are independent systems as well as information sources and actuators that must be supported by large-scale cloud computing. Calculations that require immediacy are performed by the vehicle's own system. Events that need to interface with the infrastructure, first interact with an intermediate network such as an edge device. Connected vehicles use the cloud for connection and form part of the flow of information. As in mobile communications in general, the amount of data required in connected vehicles covers a wide range, from low volume and low data rates (such as for beacons) to large volumes and high data rates (such as to provide 3D images). While, radio-communication technology has advanced greatly, in the future with extended use of semiconductor technology, it must provide high reliability with confidence and safety. Correspondingly, communications technology based on millimeter waves [18] with frequencies higher than those presently used will be necessary for large-volume communication.

6. Future Mobility-Enhanced Society

Future societal issues of Global warming, air pollution, the aging society, and cyber security cannot be solved by using automobile technology alone. Rather, it is essential to build a system that comprehensively grasps the flow of information and goods, while simultaneously sharing high-level information on safety. In all of this, electrification, automated driving, and connected vehicles will play a key role. In the future, in support of this system, cyber mobility centers will make forecasts based on data analysis of available broad-ranging data, providing information updates and visualizations. Semiconductor technology is expected to achieve lower-power consumption, higher reliability, and improved performance; moreover, enhanced software technology is also necessary. Overall, global industrial cooperation beyond the borders of semiconductors, components, and systems is essential.

References:

[1] Strategy Analytics, Automotive Electronics System Demand Forecast 2017.
[2] S. Miura et al., "Development of Power Devices for Power Cards," DENSO Technical Review, vol. 16, pp. 38-45, 2011.
[3] N. Hirano et al., "Structural Development of Double-sided Cooling Power Modules," DENSO Technical Review, vol. 16, pp. 30-37, 2011.
[4] H. Kono et al., "1.2KV Vertical Power SiC MOSFET with High Temperature Characteristics," Toshiba Review, Vol. 65, No. 1, pp. 23-26, 2010.
[5] D. Nakamura et al., "Ultrahigh-Quality Silicon Carbide Single Crystals," *Nature* 430, 1009-1012, 2004.
[6] K. Matsugatani, "Sensing Technologies for Realizing Automated Driving," *DENSO Technical Review*, vol. 21, pp. 13-21, 2016.
[7] T. Funazaki et al., "Deep Neural Networks Accelerator for Pixel Labeling," DS-01, Meeting on Image Recognition and Understanding (MIRU), 2016.
[8] M. Ogawa et al., "Development Trend of Automotive Millimeter-Wave Radar," IEICE technical report, vol. 113, No. 203, pp. 17-20, 2013.
[9] Y. Watanabe, "Semiconductor Technology and Its Challenges for Millimeter-Wave Applications," IEICE technical report, Vol. 114, No. 111, pp. 51-54, 2014.
[10] DENSO news releases Aug.8, 2017,"DENSO Establishes A New Company Designing Key Components Enabling Automated Driving"
https//www.denso.com/global/en/news-releases/2017/20170808-g01/
[11] N. Lu et al., "Connected Vehicles: Solutions and Challenges," IEEE Internet of Things Journal, Vol. 1, No. 4, pp. 289-299, 2014.
[12] S. Onoe, "Evolution of 5G Mobile Technology Toward 2020 and Beyond," ISSCC, pp. 23-28, Feb. 2016.
[13] Marc Klaassen, "Wireless Transceivers for Car2Car & Car2X systems and Applications", Forum F4 Presentation: "mm-Wave Advances for Active Safety and Communication Systems", ISSCC 2014.
[14] IEEE 802.11p, IEEE Standard for Information technology Telecommunications and Information exchange between systems Local and metropolitan area networks Specific requirements, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications.
[15] S. Checkoway et al. (2011, Aug.) *Comprehensive Experimental Analyses of Automotive Attack Surfaces* [Online]. Available:
http://www.autosec.org/publications.html
[16] Ed Markey (2015, Feb.) *Tracking & Hacking: Security & Privacy Gaps Put American Drivers at Risk* Available:
http://www.markey.senate.gov/imo/media/doc/2015-02-06_MarkeyReport-
[17] A. M. Corley, "The Danger-Sensing Driver's Seat," *IEEE Spectrum*, Vol. 47, No. 7, pp. 12-13, 2010.
[18] L. Reger, "The Road Ahead for Securely-Connected Cars," *ISSCC,* pp. 29-33, Feb. 2016.

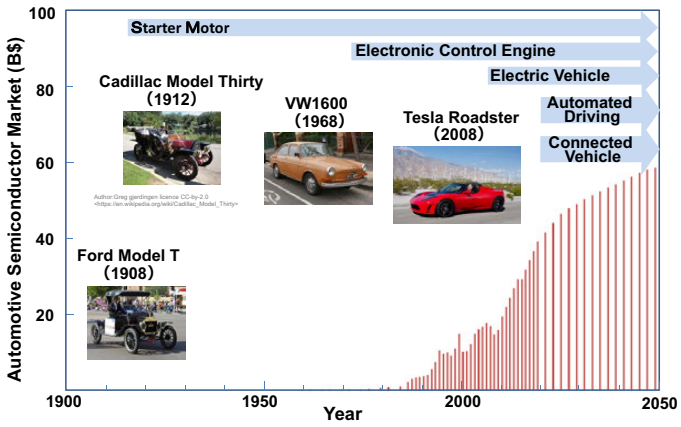


Figure 1.3.1: Automotive Semiconductor Market.

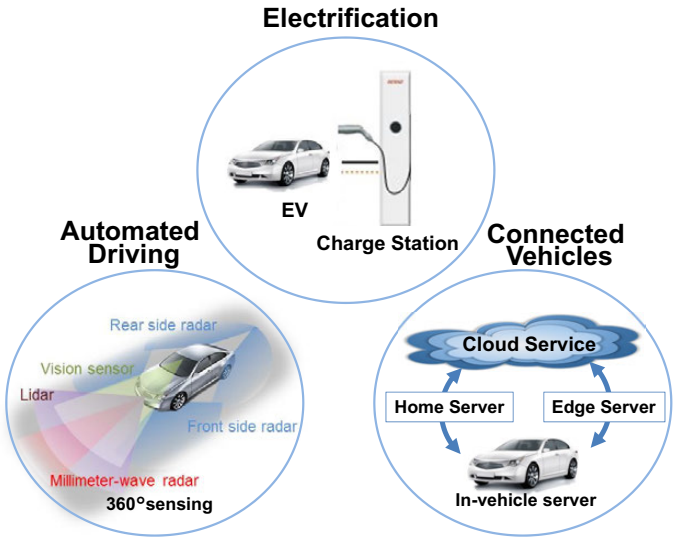


Figure 1.3.2: Core Technologies for Future Mobility.

Quality	Ambient Temperature	-40~150°C
	Humidity	<95%
	Lifetime	20years
	Defect Ratio	<1ppm
Sample Delivery	Several Years before Vehicle Production	
Delivery Period	>10 years	

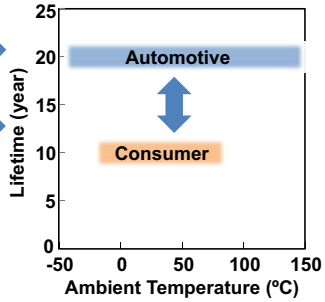


Figure 1.3.3: Automotive Semiconductor Requirements.

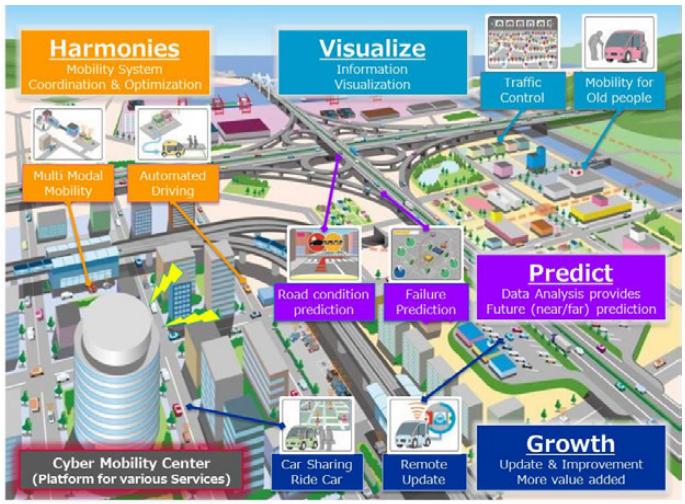


Figure 1.3.4: Future Urban Transport Network.

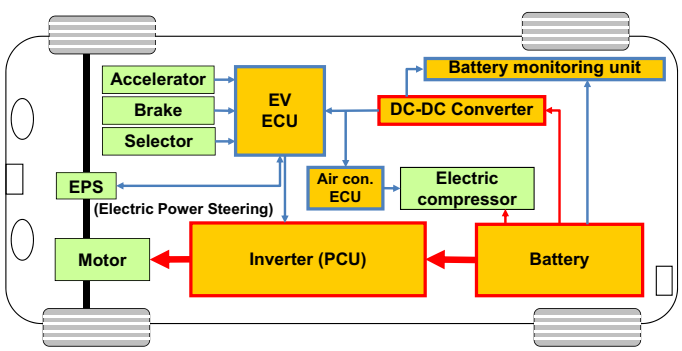


Figure 1.3.5: Electric-Vehicle System Configuration.

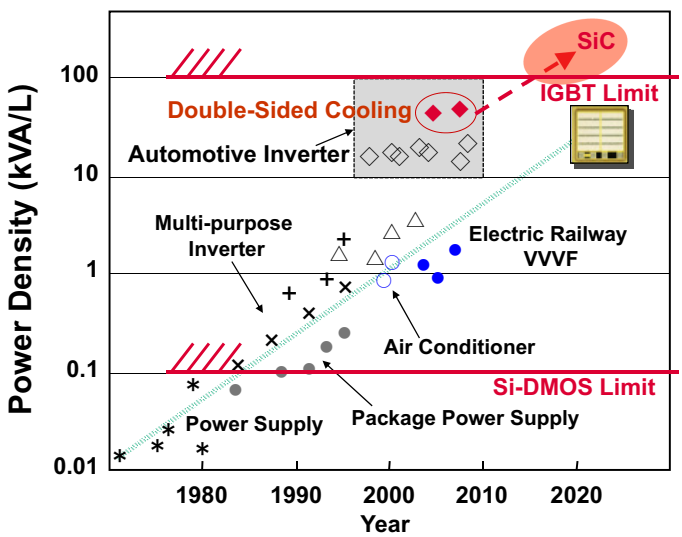


Figure 1.3.6: Power Devices for Inverter.

 978-1-5090-4940-0/18/$31.00

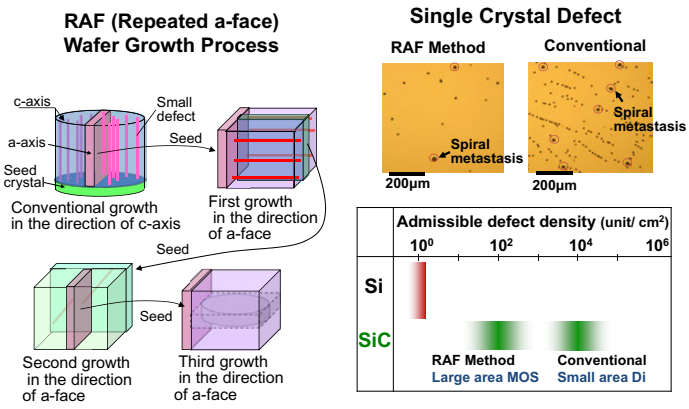


Figure 1.3.7: High-Quality SiC Wafer.

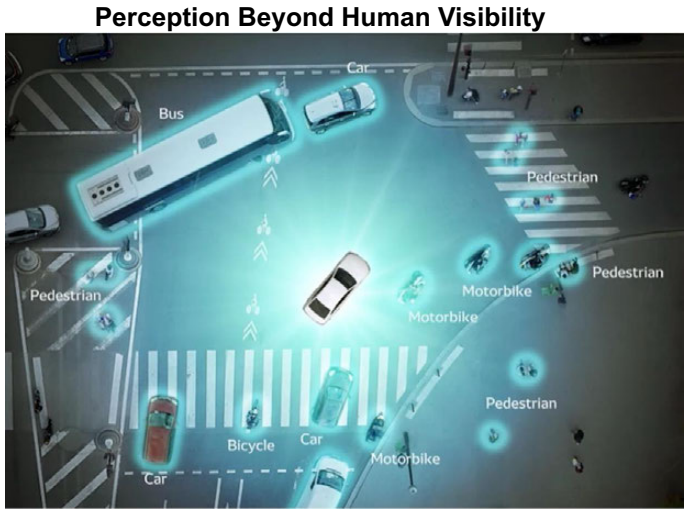


Figure 1.3.8: Sensing for Automated Driving.

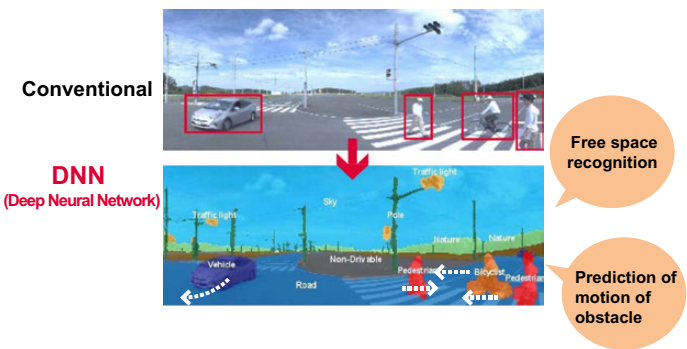


Figure 1.3.9: DNN-IP Processing Result.

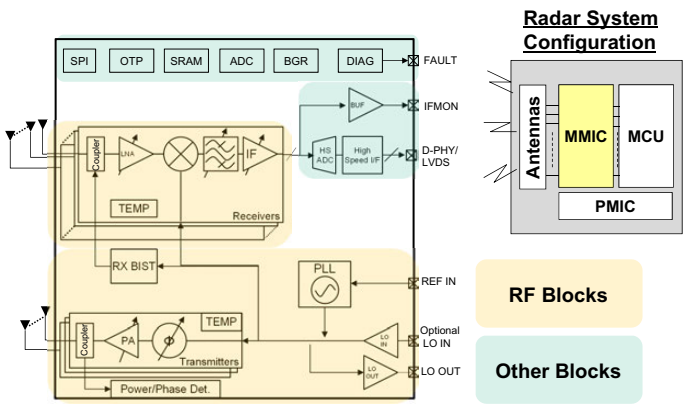


Figure 1.3.10: Block Diagram of A CMOS Radar MMIC.

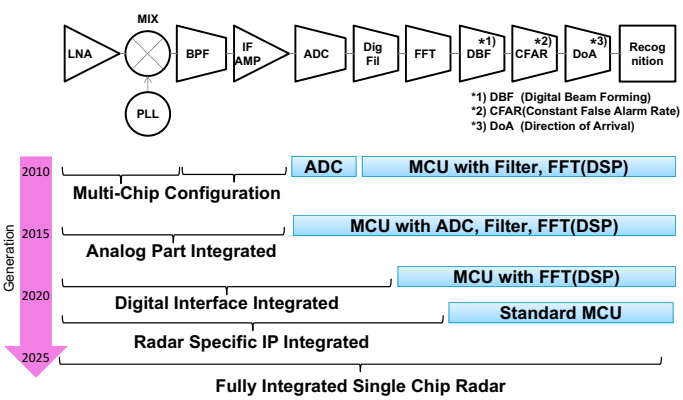


Figure 1.3.11: Typical Radar Receiver System.

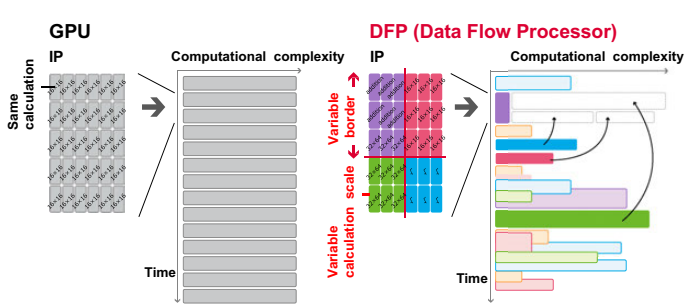
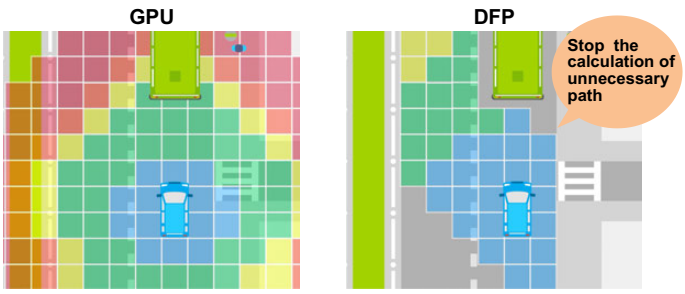


Figure 1.3.12: Difference between GPU and DFP.



Search for a new route, when a front bus stopped.

Figure 1.3.13: Characteristic of DFP Processing.

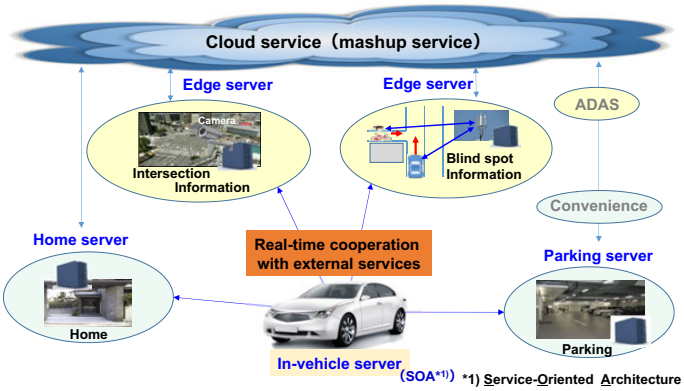


Figure 1.3.14: Vision of Connected Service.

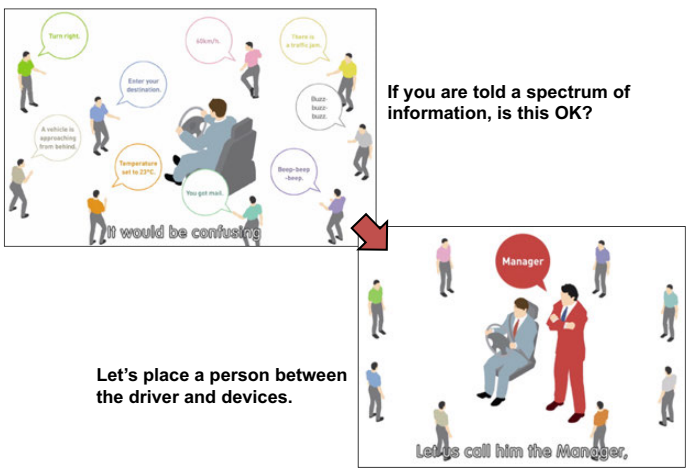


Figure 1.3.15: HMI-ECU Manager.

1.4 50 years of Computer Architecture: From the Mainframe CPU to the Domain-Specific TPU and the Open RISC-V Instruction Set

David Patterson, Google and UC Berkeley

1. Our Story Begins in the Early 1960s

IBM had four incompatible computer lines. Each had its own unique *instruction-set architecture* (*ISA*); I/O system; system software (assemblers, compilers, libraries); and market niches (business, scientific, real time). IBM engineers bet that they could invent a single ISA that would work for customers of all four lines. Moreover, the same program would run correctly on any implementation of that ISA, though at different speed and cost. That vision required a new way to build computers that would be *binary compatible* from the cheapest 8-bit model to the fastest 64-bit version.

While datapaths were straightforward, the hardest part of computer design then and now is control. Maurice Wilkes, a computing pioneer, proposed that control be built from Read Only Memory (ROM) [1]. At the time, Random Access Memory (RAM) was much less expensive than logic, and ROM was much less expensive than RAM. Wilkes dubbed it *microprogramming*, in that designing control resembled programming at a low, detailed level, and he called each word of the control ROM a *microinstruction*.

IBM engineers embraced microprogramming to deliver their grand goal. Each model of the new computer family would have its own ISA interpreter written in microinstructions customized to that model. The more hardware to control, the wider the microinstruction, and wider datapaths generally needed fewer microinstructions for the ISA interpreter since they computed more quickly. The cheapest model used 4000 microinstructions 50 bits wide and the fastest one used 2800 microinstructions 87 bits wide.

On April 7, 1964, IBM announced the most important technical event in the company's history:

> *"System/360 represents a sharp departure from concepts of the past in designing and building computers. …*
>
> *System/360 is a single system spanning the performance range of virtually all current IBM computers. …*
>
> *System/360 purchase prices range from $133,000 to $5,500,000."*

In today's dollars, the range was $1,054,000 to $43,570,000.

IBM bet the company that binary compatibility would work, and it won that bet. Correspondingly, *mainframe computers* dominated the high end of the information technology (IT) field for decades. IBM still sells a descendant of the System/360 more than 50 years later, making it the longest lasting ISA.

2. Complex-Instruction-Set Computers (CISC)

Built from small-scale and medium-scale integrated circuits, *minicomputers* were next on the IT scene. Logic, RAM, and ROM were all made from the same transistor in the 1970s. Moore's Law meant that the control store could be much larger, yet still affordable and fast. That increase led to larger microprograms, which supported larger ISAs with many sophisticated instructions. But, as microprograms grew, they were more likely to have bugs. To make microprograms easier to repair, control store became RAM as it was about the same speed as semiconductor ROM.

A shining example of minicomputers and complex ISAs was the VAX-11/780, which Digital Equipment Corporation (DEC) announced October 25, 1977. It had 5120 microinstructions that were 96 bits wide. The VAX line of minicomputers was a workhorse of the IT industry for the next decade.

3. The Dominant Microprocessor ISA

Moore's Law suggested that microprocessors would eventually compete with minicomputers in performance but at much lower cost. Just as IBM still sells an offshoot of the System/360, Intel founder Gordon Moore believed that the ISA that succeeded the Intel 8080 8-bit microprocessor would last the lifetime of the company.

In 1975, a year after Intel completed the 8080, Moore started a skunk works in Oregon to invent an ISA worthy of Intel's future. He hired many new PhDs in computer science to deliver that goal.

Their ISA was a complete break from the 8080: It was a stack computer with no general-purpose registers; instruction lengths could be any number of bits; addresses were 32-bit capabilities; and they wrote a custom operating system for it in the esoteric programming language Ada. With considerable fanfare, in 1981 Intel announced the iAPX 432 [2]:

"The vacuum tube, the transistor, the microprocessor—at least once in a generation an electronic device arises to shock and strain designers' understanding. The latest such device is the iAPX 432 micromainframe processor, a processor as different from the current crop of microprocessors (and indeed, mainframes) as those devices are from the early electromechanical analog computers of the 1940s."

Alas, its large microprogram could not fit into a single chip, so the iAPX 432 was expensive, slow, and late. When the Oregon engineers told Moore in 1977 that their project would take much longer than he wanted, Intel launched an emergency project to develop a 16-bit successor to the 8080. The small team was given in only 52 weeks to invent an ISA, and design the chip. Given the abrupt schedule, the team designed the ISA in only 3 weeks of elapsed time with a total effort of 10 person weeks. With the goal of being assembly language compatible with the 8080, the ISA widened the 8080 accumulators and added a 20-bit segmented address space. Intel released the 8086 in 1978 with neither high hope or hype.

Around the same time, IBM started a group in Florida to develop a consumer product that would compete with the Apple I computer. They seriously considered using the Motorola 68000 microprocessor, whose ISA resembled the IBM System/360, but the 68000 was late. IBM went instead with a cost-reduced variant of the Intel 8086. IBM announced the resulting Personal Computer on August 12, 1981:

"This is the computer for just about everyone who has ever wanted a personal system at the office, on the university campus or at home .... We believe its performance, reliability, and ease of use, make it the most advanced, affordable personal computer in the marketplace."

IBM projected sales of 250,000 but sold 100,000,000 PCs, turning the Intel 8086 into an "overnight" success. When binary compatibility for PC software was factored in, the IBM PC also gave the 8086 a very bright future.

Gordon Moore correctly predicted that Intel's next ISA would dominate its future, but it was the emergency substitute 8086, not the anointed iAPX 432, that fulfilled the promise. (Intel discontinued that ISA in 1986.) Subsequently in 1985, Intel extended the 8086 ISA address size to 32 bits with the 80386, enabling the 80x86 ISA to dominate microprocessors for the next 15 years.

4. Reduced-Instruction-Set Computers (RISC)

Returning to larger computers, DEC engineers found 20% of VAX instructions were responsible for 60% of the microcode, but only accounted for 0.2% of execution time [3]. Contemporaneously, John Cocke and his group at IBM Research ported an experimental compiler to IBM System/370 that only used simple register-register and load/store instructions. These programs ran much faster than those generated from existing compilers that utilized the full ISA. Such results called into question the big microcoded interpreters in their larger control stores and the complex ISAs they enabled.

It was realized that an alternative was having an ISA so simple that it did not require a microcoded interpreter. Another way to think of it was that such instructions were as simple as microinstructions, but not as wide. The next insight was to convert the fast RAM of control store into an instruction cache of user-visible instructions. The contents of fast instruction memory could change to hold what the executing application needs rather than always containing an ISA interpreter.

This simple ISA also made it easy to have pipelined implementations, which led to faster clock rates. The advances of Moore's Law meant that in the early 1980s, a 32-bit datapath with small caches could fit in a single chip. The lack of chip crossings of an integrated design also meant faster operation. This ISA style became known as a *Reduced-Instruction-Set Computer* (*RISC*) in contrast to CISC [4].

One skepticism directed at RISC was that it would execute more instructions than a CISC, so how could it be better? The following formula resolved the dilemma:

$$\frac{Intructions}{Program} \times \frac{Clock\ cycles}{Instruction} \times \frac{Time}{Clock\ cycle} = \frac{Time}{Program}$$

CISC implementations executed fewer instructions per program, but RISC computers executed on average many fewer clock cycles per instruction (CPI). To their credit, DEC engineers published a paper 10 years later comparing a CISC to a RISC with a similar datapath. They found about a factor of three advantage for RISC [5]: the VAX executed roughly half as many instructions as the RISC, but its CPI was about six times higher.

RISC research projects at IBM, Berkeley, and Stanford paved the way for the commercial success of ARM [6], MIPS, POWER, SPARC [7], and many more. Correspondingly, Reduced Instruction Set Computers (RISC) became the dominant ISA of workstations, minicomputers, and servers from the mid 1980s to the turn of the century.

5. RISC versus CISC Today

The 80x86 eventually surpassed the RISC architectures by using hardware to translate the 80x86 ISA into internal RISC-like instructions, which allowed the 80x86 to copy any performance enhancements developed for RISC processors. (Intel calls them *micro operations* and AMD calls them *RISC operations*.) Examples include deep pipelines, fetching multiple instructions per clock cycle, and branch prediction. Given superior semiconductor processing and circuit design, the 80x86 eventually had the fastest processors, and overtook the market in small servers from RISCs in addition to dominating the PC market.

As Moore's Law increased the transistor budgets, the extra overhead in area and energy of hardware translation was affordable for PCs and servers, but it was too costly for the embedded market. For example, 100% of Android and Apple phones and tablets use RISC processors. The size of the embedded market means billions of chips are shipped each year using RISC processors from ARM (Advanced RISC Machine), Synopsis ARC (Argonaut RISC Core), Cadence Tensilica, and MIPS. Figure 1.4.1 shows that RISC shipments grew about 24% annually recently, providing a sevenfold increase since 2007.

The 80x86 CISC ISA dominated chip shipments in the PC era, but RISC architectures control the PostPC era. We demark that era with the introduction of the iPhone in 2007. Annual 80x86 shipments peaked in 2011 at 365M, but have been declining about 8% annually since; Intel and AMD made fewer 80x86 chips in 2016 than in 2007. While 80x86 dominates the cloud portion of the PostPC Era, Reddi [8] estimates that the entire installed base for the Amazon, Google, and Microsoft clouds is 10M servers. While these chips are expensive, their volume is negligible; 10M RISC chips ship every 4 hours!

6. Very-Long-Instruction-Word (VLIW) Computers

Some researchers believed that Very-Long-Instruction-Word (VLIW) computers would replace CISC and RISC as the dominant ISA style. Similar in concept to the wide microinstructions, a single wide VLIW instruction commands multiple operations. Figure 1.4.2 gives an example of an instruction that could perform two integer operations, two memory accesses, and two floating-point operations. The supporting hardware would then include two integer units, two load-store units, and two floating-point units. Instead of the 32-bit instructions of RISC architectures, VLIW instructions were more than 100 bits.

In pure VLIW, there were no interlocks between instructions. If a VLIW instruction calculated a result need by a subsequent instruction, and the latency for that operation was three clock cycles, the dependent instruction had to execute exactly three clock cycles later to get the correct results. It was up to the software to schedule the operations properly, which made the hardware easier to design but upped the demands on compilers.

Figure 1.4.2 shows three distinct latencies for different operations within one VLIW instruction, which was typically true. While such a task would be challenging for the assembly-language programmer, VLIW advocates believed that advanced compiler technology could successfully map high-level language programs to the software-scheduled hardware.

In the mid 1990s, Intel faced a decision about the fate of the 80x86 ISA. Its 32-bit address space would soon be insufficient for many applications. While one choice was to extend the old ISA to 64 bits, similar to how it transitioned from 16 to 32 bits with the 80386, Intel had business and technical reasons to change the ISA. A business reason was to eject AMD as an ISA partner, as AMD also had the right to make microprocessors that were compatible with the 80x86 ISA. The RISC architectures also posed a business opportunity to Intel, as a new ISA might capture that market. The technical reason was that the 80x86 was an antiquated ISA that had relatively few registers, was missing some useful operations, and had legacy features that were maintained only for backwards binary-compatibility.

Consequently, Intel joined forces with Hewlett Packard in 1994 to promote *Explicitly-Parallel-Instruction Computing* (*EPIC*) as their 64-bit address successors to the 80x86, and HP PA-RISC ISAs. EPIC was essentially a binary-compatible variation of VLIW. The initial goal was to ship the first chip in 1997. Intel announced the official name of the processor, *Itanium*, on October 4, 1999:

"EPIC is the old term for what is now known as the Itanium processor family architecture, co-developed by HP and Intel. This design philosophy will one day replace RISC and CISC. It is a gateway into the 64-bit future..."

There was considerable publicity about the Itanium even before the first chips were available, with dozens of companies (HP, Microsoft, Silicon Graphics, Bull, Hitachi, ...) giving up their RISC products to embrace Itanium, which was widely viewed as the inevitable winner given the backing of HP and Intel.

Since AMD was not invited to participate, it had no choice but to develop its own 64-bit ISA. The AMD64 ISA extended 80x86 to 64-bit addresses by widening all registers plus adding a few more registers to help compilers, following the precedent of the 80386.

Alas, the Itanium was an "EPIC" failure! The first version did not ship until 2001. The performance difficulties included scheduling VLIW code for unpredictable branches and for variable memory latency due to unpredictable cache misses, and for the explosion in code size that increased instruction cache misses due to wider instructions and their low utilization. As the Stanford computer scientist Donald Knuth summarized [10]:

"The Itanium approach...was supposed to be so terrific—until it turned out that the wished-for compilers were basically impossible to write."

Considering the billions invested and the extensive promotion, the chip delays and under performance caused online forums to rechristen it the *Itanic* after the infamous "unsinkable" Titanic"!

The VLIW Itanium architecture was heralded as the 64-bit address successor to the 80x86, but instead it was AMD64 that Intel was eventually forced to adopt. (Intel ended the Itanium line in May 2017.) Ironically, the emergency replacement 8086 ISA and its descendants beat both of Intel's anointed and trumpeted successors: the iAPX 432 and the Itanium.

VLIW failed for general-purpose computing, but found a home in Digital Signal Processors (DSPs), which avoid three of the weaknesses of VLIW: DSP programs are small so code size matters less; its branches are usually very predictable; and the hardware provides program-controlled memories instead of caches, which offers fixed memory latency.

7. Domain-Specific Architectures

Architects rode Moore's Law and Dennard Scaling to turn increased resources into performance. They designed sophisticated processors and memory hierarchies that exploited parallelism between instructions without the knowledge of the programmer. Architects eventually ran out of techniques for instruction-level parallelism that could be exploited efficiently. The end of

 978-1-5090-4940-0/18/$31.00

Dennard scaling and the lack of greater (efficient) instruction-level parallelism in 2004 forced the industry to switch from a single energy-hogging processor per microprocessor to multiple efficient processors or *cores* per chip.

A law that is as true today as when Gene Amdahl stated it in 1967, demonstrates the diminishing returns to increasing the number of processors: Amdahl's Law states that the sequential part of the task limits the theoretical speedup from parallelism; if ⅛ of the task is serial, the maximum speedup is 8 even if one adds 100 processors, and the rest of the task is easily parallel.

Figure 1.4.3 shows how the impact of these three laws on processor performance for the past 40 years. If the trends continue, single-program performance using standard benchmarks will not double for 20 years! At the present state-of-the-art,

- transistors are not getting much better (due to the ending of Moore's Law),
- the peak power per mm^2 of chip is increasing (due to the end of Dennard scaling), but the power budget per chip is limited (due to electromigration, mechanical and thermal limits), and
- we have already played the multicore card (limited by Amdahl's Law).

In view of these inevitable limitation, architects now widely believe that the only path left for major improvements in performance-cost-energy is *domain-specific architectures* (*DSAs*).

8. An Example DSA: the Google TPU

A trailblazing example of DSA concept is the Google *Tensor Processing Unit* (*TPU*), first deployed in 2015, which serves billions of people [11]. It runs *deep neural network* (*DNN*) inference 15-to-30 times faster with 30-to-80 times better energy efficiency than contemporary CPUs and GPUs in similar semiconductor technologies.

The block diagram of the TPU in Figure 1.4.4 shows that the host sends instructions over the PCIe bus to an instruction buffer. The internal blocks are typically connected together by 256-byte-wide paths. Starting in the upper-right corner, the *Matrix Multiply Unit* is the heart of the TPU. It contains 256×256 multiply-accumulators (*MACs*) that can perform 8-bit multiply-and-adds on integers. The 16-bit products are collected in the 4MiB of 32-bit *Accumulators* below the matrix unit. The 4MiB represents 4096, 256-element, 32-bit accumulators. The matrix unit produces one 256-element partial sum per clock cycle.

The weights for the matrix unit are staged through an on-chip Weight FIFO that reads from an off-chip 8GiB DRAM called *Weight Memory*. The weight FIFO holds up to four 256×256 tiles of 8-bit weights. The intermediate results are held in the 24MiB on-chip *Unified Buffer*, which serves as inputs to the Matrix Unit. A programmable DMA controller transfers data to-or-from CPU Host memory and the Unified Buffer. To function dependably at Google scale, the TPUs internal and external memories have error detection and correction hardware.

The philosophy of the TPU microarchitecture is to keep the large matrix unit busy. Toward that end, the instruction that reads the weights follows the *decoupled-access/execute* philosophy [12], in that it can complete after sending its address, but before the weight is fetched from Weight Memory. The matrix unit will stall if the input activation or weight data is not ready.

Since readout of a large SRAM uses much more power than arithmetic, the matrix unit uses *systolic execution* [13] to save energy by reducing reads and writes of the Unified Buffer. It relies on data from two directions arriving at cells in an array at regular intervals where they are combined. A given 256-element multiply-accumulate operation moves through the matrix as a diagonal wave front. The weights are preloaded, and take effect with the advancing wave alongside the first data of a new block. Control and data are pipelined to give the illusion that the 256 inputs are read at once, and that they instantly update one location of each of all 256 accumulators.

The TPU can be compared to two large chips: an 18-core Intel Haswell CPU (662 mm^2) and the Nvidia Kepler K80 GPU (561 mm^2). The TPU is about half the area of the Haswell or Kepler, and half the power, yet packs 25 times the MACs (65,536 8-bit vs. 2,496 32-bit), and 3.5 times the on-chip memory (28MiB vs. 8MiB) as the GPUs.

Figure 1.4.5 illustrates Roofline Performance models for six neural network applications that represent 95% of the TPUs datacenter inference workload in 2016. The Roofline Performance model [14] for high-performance computer modelling illustrates the effectiveness of the applications. This simple visual model offers insights into the causes of performance bottlenecks. The peak computation rate forms the "flat" part of the roofline, and memory bandwidth is the "slanted" part of the roofline.

The six DNN applications are generally further below their ceilings for Haswell CPU and K80 GPU than is the case for the TPU. Response time is the reason! Many of these DNN applications are parts of end-user-facing services. For example, the 99th-percentile response-time limit of one application was 7ms. Haswell and the K80 run at just 42% and 37%, respectively, of the highest throughput achievable if the response-time limit is relaxed. These bounds affect the TPU as well, but at 80%, it is operating much closer to its highest application throughput. On average, the TPU is 29 and 15 times faster than the CPU and GPU.

In datacenters, cost-performance trumps mere performance. Since Google does not disclose its costs, but power is correlated with total cost of ownership, Figure 1.4.6 plots performance/Watt, showing GPU vs. CPU (blue), TPU vs. CPU (red), and TPU vs. GPU (orange). Quantitatively, the relative performance/Watt (ignoring host power) is 83 for TPU vs. CPU, and 29 for TPU vs. GPU. Figure 1.4.6 also shows a hypothetical TPU (TPU') using the same GDDR5 memory as in the GPU. The relative performance/Watt (ignoring host power) of TPU' leaps to an amazing 196× over the Haswell CPU (green bar), and 68× over the K80 GPU (purple bar).

But, more importantly, Google's Tensor Processing Unit for deep-neural-network inference has been deployed successfully in the cloud since 2015, and is used regularly by billions of people. Seven factors explain its energy-performance advantages, given in priority order:

1. The TPU has 1 very large, 2-dimensional multiply units, while the CPU and GPU have 18 and 13 smaller, 1-dimensional multiply units. The matrix multiplies inherent to DNNs benefit from 2-dimensional hardware.
2. The TPUs DNN inference applications uses 8-bit integers rather than 32-bit floating point to improve efficiency of computation, memory bandwidth, and memory capacity.
3. The 2-dimensional organization enables systolic arrays, which reduce register accesses and energy.
4. The TPU drops features required by CPUs and GPUs that DNNs do not use, which makes the TPU cheaper, saves energy, and allows transistors to be repurposed for domain-specific optimizations.
5. The TPU has 1 program thread while the K80 has 13 and the CPU has 18. A single thread makes it easier to stay within a fixed latency limit that the neural network applications demand, as well as saving energy.
6. The TPU has enough flexibility to implement the DNNs of 2017, as well as the lesser demands of 2013.
7. The original production applications were written using TensorFlow, making them easy to port to the TPU with high-performance, rather than requiring rewriting on the very different TPU hardware.

9. RISC-V

Despite the promise of DSAs, we still need general-purpose processors to run the programs that are not domain specific, such as operating systems, user interfaces, compilers, and so on. Amazingly, 30 years after the first commercial RISC architectures hit the market, the consensus is still that RISC is the best ISA style for general-purpose processors. No one has proposed a new CISC architecture since 1985, nor a new general-purpose VLIW since 2000.

This ISA consensus plus the move to DSAs have led to a new take on ISAs, called *RISC-V* [15]. (It is pronounced "RISC five" since it is the fifth RISC architecture from UC Berkeley.) RISC-V is unconventional not only because it is a recent ISA (born this decade when most alternatives date from the 1970s or 1980s), but also because it is an *open* ISA. Unlike practically all prior ISAs, its future is free from the decisions and ultimate fate of a single corporation. It belongs instead to an open nonprofit foundation (riscv.org) with 100 members.

The goal of the RISC-V Foundation is to maintain the stability of RISC-V, evolve it slowly and carefully for technological requirements, and to try to make it as popular for processors as Linux is for operating systems. RISC-V benefits from starting 25 years later than other popular ISAs, which allowed its architects to borrow the good ideas but to not repeat the mistakes of the past [15].

Keeping with its heritage, RISC-V is a minimalist ISA; in fact, the base ISA is remarkably similar to its great-great-grandparent RISC-I [9]. One indication of complexity is the size of the documentation. The ISA manual for x86-32 is 2198 pages or 2,186,259 words [16]. The RISC-V equivalents are 236 pages and 76,702 words [17]. If one read manuals as a (terribly boring) full-time job (8 hours a day for 5 days a week), it would take a month for one pass over the x86-32 but less than a day for the RISC-V. Using this common-sense metric, RISC-V is 1/30th the complexity.

Beyond being minimalist open and recent, RISC-V is unusual since, again unlike almost all prior ISAs, it is *modular*. At the core is a base ISA that runs a full software stack (operating system, libraries, debuggers, compilers). The base is frozen and will never change, which gives compiler writers, operating system developers, and assembly language programmers a stable target. The modularity comes from optional standard extensions that hardware can include or not, depending on the needs of the application. Example optional extensions are multiply and divide, single- and double-precision floating-point arithmetic, atomic instructions, compact instruction encoding, and vector instructions [18]. Even including all optional extensions, a summary of the minimalist RISC-V ISA fits into only two pages [19].

To achieve the software-desirable goal of a single ISA that works from the smallest to the largest computers, it must lead to efficient designs both for edge devices and the cloud. To empower large-scale computers, RISC-V offers 64-bit, as well as 32-bit address versions. Minimalism and modularity enable small and low energy implementations of RISC-V, which helps embedded applications. While some argue that ISA complexity doesn't matter for high-end processors, it does matter for low-cost applications, which the lack of success of the 80x86 illustrates. A universal ISA by definition must work well everywhere.

A modern ISA must also reserve opcode space to support the closest possible coupling between general-purpose cores and domain-specific accelerators. In the 1970s and 1980s, when Moore's Law was in full force, there was little thought of saving opcode space for future accelerators. Architects instead valued larger address and immediate fields to reduce the number of instructions executed per program. RISC-V enables DSAs simply by reserving opcodes for custom accelerators.

An open ISA also enables sharing of implementations of RISC-V either for free, or for profit by any organization. Competition, a free market, and open implementations may lower costs and increase innovation, similar to the benefits of open-source software. Open designs also reduce the odds of unwanted malicious secrets being hidden in a processor.

10. Conclusion

For at least the past decade, computer architecture researchers have been publishing innovations based on simulations using limited benchmarks claiming improvements for general-purpose processors *of 10 percent or less,* while the DSAs like the TPU reports gains for a domain-specific architecture deployed in real hardware running genuine production applications *of more than a factor of 10* [20]. Order-of-magnitude differences between commercial products are rare in computer architecture, which explains architects' enthusiasm for DSAs. DSAs need ISA support, which RISC-V offers.

The goal of RISC-V is to be effective for all computing devices, from the smallest to the fastest; to be modular to allow tailoring of implementations to the needs of the application; to be long-lived by having a non-profit foundation as its owner that evolves it slowly in response to technology change; and to preserve opcode space to support DSAs.

Ironically, the end of Dennard scaling and Moore's Law may rejuvenate computer architecture research and development, as advances must now come from innovation visible in ISAs. DSAs and RISC-V may play leading roles in this renaissance!

References:

[1] Wilkes, Maurice, William Renwick, David Wheeler, "The Design of the Control Unit of An Electronic Digital Computer", *Proceedings of the IEE-Part B: Radio and Electronic Engineering* 105:20, pp.121-128, 1958.

[2] Hemenway, Jack, Robert Grappel, "Understand the Newest Processor to Avoid Future Shock", *Electronic Design News*, pp. 129-36, April 29, 1981.

[3] Emer, Joel, Douglas Clark, "A Characterization of Processor Performance in the VAX-11/780", *Proc. International Symposium on Computer Architecture*, pp, 301-310, 1984.

[4] Patterson, David, David Ditzel, "The Case for the Reduced Instruction Set Computer", *ACM SIGARCH Computer Architecture News* 8, no. 6, pp. 25-33, 1980.

[5] Bhandarkar, Dileep, Douglas Clark, "Performance from Architecture: Comparing a RISC and a CISC with Similar Hardware Organization", In *Proc. Architectural Support for Prog. Languages and Operating Systems Symposium*, pp. 310-319, 1991.

[6] Furber, Steve. "Microprocessors: The Engines of the Digital Age." *Proc. Royal Society Series A* 473:2199. The Royal Society, 2017.

[7] "Chip Hall of Fame: Sun Microsystems SPARC Processor", https://spectrum.ieee.org/tech-history/silicon-revolution/chip-hall-of-fame-sun-microsystems-sparc-processor, IEEE Spectrum June 2017.

[8] Reddi, Vijay, "A Decade of Mobile Computing", *Computer Architecture Today*, July 21, 2017.

[9] Patterson, David, "How Close is RISC-V to RISC-I?", *ASPIRE Blog*, June 19, 2017

[10] Knuth, Donald E., Andrew Binstock. Interview with Donald Knuth. *InformIT,* 04-01, 2010

[11] Jouppi, Norman P., Cliff Young, Nishant Patil, David Patterson, *et al*, "In-Datacenter Performance Analysis of a Tensor Processing Unit", *Proc. International Symposium on Computer Architecture*, 2017.

[12] Smith, James, "Decoupled Access/Execute Computer Architectures", *Proc. International Symposium on Computer Architecture*, 1982

[13] Kung, H.T., Charles Leiserson. "Algorithms for VLSI processor Arrays", *Introduction to VLSI systems*, 1980.

[14] Williams, Samuel, Andrew Waterman, David Patterson, "Roofline: An Insightful Visual Performance Model for Multicore Architectures." *Communications of the ACM* 52, no. 4, pp. 65-76, 2009

[15] Patterson, David, "Reduced Instruction Set Computers Then and Now", *IEEE Computer*, December 2017.

[16] Baumann, Andrew, "Hardware is the New Software", *Proc. 16th Workshop on Hot Topics in Operating Systems*, pp. 132–137, 2017.

[17] Waterman, Andrew, Krste Asanovic, editors, "*The RISC-V Instruction Set Manual, Volume I: User-Level ISA*, Version 2.2", https://riscv.org/specifications/. "*The RISC-V Instruction Set Manual Volume II: Privileged Architecture* Version 1.10", https://riscv.org/specifications/privileged-isa/ May 2017.

[18] Patterson, David, Andrew Waterman, "SIMD Considered Harmful", *Computer Architecture Today*, September 18, 2017

[19] Patterson, David and Andrew Waterman, *The RISC-V Reader: An Open Architecture Atlas"*, Strawberry Canyon, 1st edition, 200 pages, www.riscvbook.com, 2017.

[20] Hennessy, John L., David A. Patterson, "*Computer Architecture: A Quantitative Approach"*, 6th edition, Elsevier, 2018.

 978-1-5090-4940-0/18/$31.00

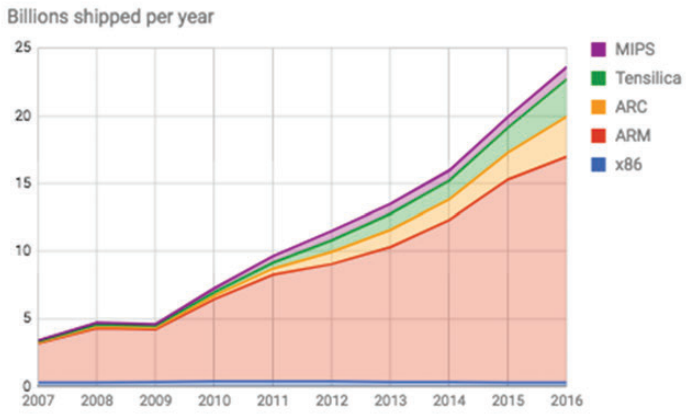


Figure 1.4.1: Billions of chips shipped 2007 to 2016; 99% of 2016 chips are RISC [9].

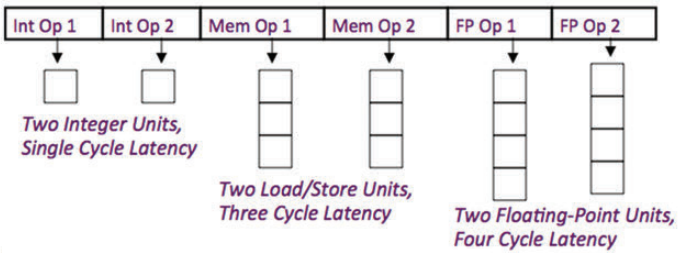


Figure 1.4 2: Hypothetical VLIW Instruction.

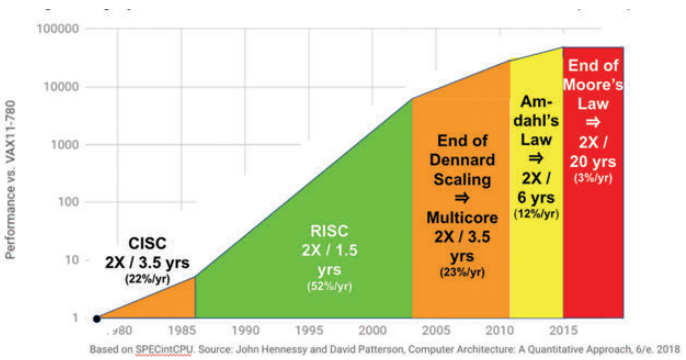


Figure 1.4.3: Average performance gain for a single program over time versus VAX 11/780 using SPECintCPU [20].

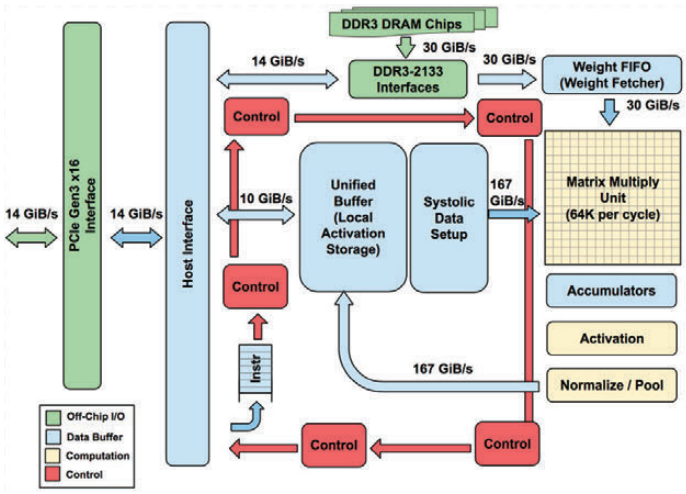


Figure 1.4.4: Block diagram of the TPU [11].

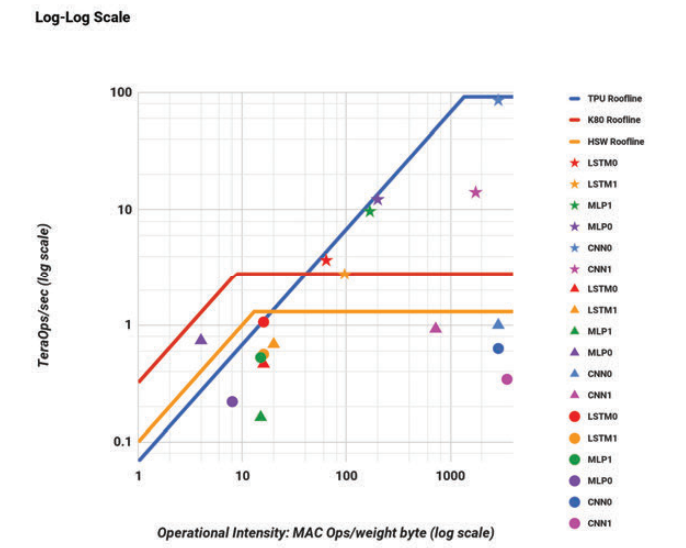


Figure 1.4.5: Roofline Performance Model of the CPU, GPU, and TPU [11].

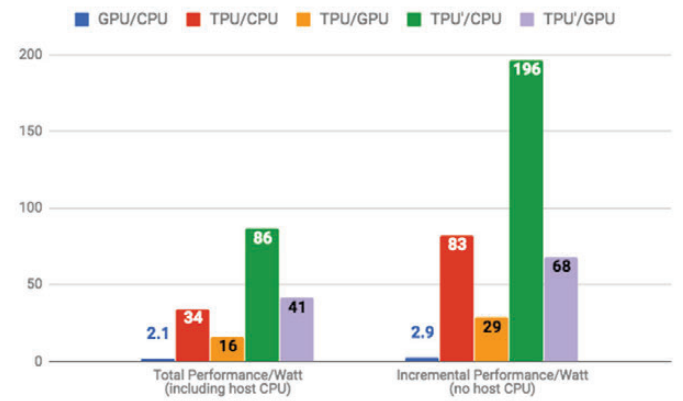


Figure 1.4.6: Relative Performance per Watt of a CPU, GPU, the original TPU, and a revised TPU' [11].

Session 2 Overview: *Processors*

DIGITAL ARCHITECTURES AND SYSTEMS SUBCOMMITTEE

Session Chair:
Thomas Burd
AMD, Santa Clara, CA

Associate Chair:
Muhammad Khellah
Intel, Hillsboro, OR

Subcommittee Chair: *Byeong-Gyu Nam,* *Chungnam National University, Korea*

Continued growth in cloud-to-edge applications is driving innovations in digital processors. The first two papers of this session cover next-generation server-class processors. This is followed by an energy-efficient 14nm graphics processor. An SoC configurable with 1-4 chips on an MCM to service multiple markets is described next. The last three papers demonstrate the first implementation of the datagram transport layer security (DTLS) protocol in hardware, an MSP430-compatible microcontroller with dual-mode enabling minimum-power and minimum-energy, and a net-zero-energy (NZE) smart mote SiP for IoT applications.

1:30 PM

2.1 SkyLake-SP: A 14nm 28-Core Xeon® Processor

S. M. Tam, Intel, Santa Clara, CA

In Paper 2.1, Intel describes SkyLake-SP Xeon®, a 28-core server-class CPU in a 14nm tri-gate process featuring a MESH on-die interconnect fabric, on-die IVRs, and 6 DDR4 channels capable of 2666MT/s per channel.

2:00 PM

2.2 IBM z14™: 14nm Microprocessor for the Next-Generation Mainframe

C. Berry, IBM Systems, Poughkeepsie, NY

In Paper 2.2, IBM presents a 14nm FinFET z14, with 50% more L2 cache, 2× larger L3 caches, 25% more cores, enhanced branch prediction, and cryptography, running 200MHz faster than the previous generation under the same power envelope.

 978-1-5090-4940-0/18/$31.00

2:30 PM

2.3 An Energy-Efficient Graphics Processor Featuring Fine-Grain DVFS with Integrated Voltage Regulators, Execution-Unit Turbo, and Retentive Sleep in 14nm Tri-Gate CMOS

P. Meinerzhagen, Intel, Hillsboro, OR

In Paper 2.3, Intel describes a 14nm graphics processor featuring fine-grain DVFS with execution-unit turbo and retentive sleep, enabling up to 32% energy reduction at iso-performance.

3:15 PM

2.4 "Zeppelin": An SoC for Multichip Architectures

N. Beck, AMD, Boxborough, MA

In Paper 2.4, AMD describes "Zeppelin", a chiplet SoC manufactured using 14nm FinFET technology with eight x86 "Zen" cores per chip. Individual chiplets are connected with AMD's coherent Infinity Fabric in a range of products, from a single-die package up to a 4-die multi-chip module (MCM) with 2-socket capability.

3:45 PM

2.5 An Energy-Efficient Reconfigurable DTLS Cryptographic Engine for End-to-End Security in IoT Applications

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

In Paper 2.5, MIT presents the first implementation of the datagram transport layer security (DTLS) protocol in a 65nm 4mm^2 test chip, resulting in a 10× reduction in code size and a 438× improvement in energy-efficiency over a software solution.

4:15 PM

2.6 A 595pW 14pJ/Cycle Microcontroller with Dual-Mode Standard Cells and Self-Startup for Battery-Indifferent Distributed Sensing

S. Jain, National University of Singapore, Singapore

In Paper 2.6, the National University of Singapore presents an MSP430 compatible microcontroller enabling minimum-power (595pW) and minimum-energy mode (14-33pJ/cycle) in a 9.5mm^2 0.18µm chip with cold start up from a 0.54mm^2 solar cell at 55lux.

4:45 PM

2.7 A cm-Scale Self-Powered Intelligent and Secure IoT Edge Mote Featuring an Ultra-Low-Power SoC in 14nm Tri-Gate CMOS

D. Kurian, Intel, Bangalore, India

In Paper 2.7, Intel demonstrates a complete cm-scale self-powered and secure IoT edge mote in 14nm, with 0.2mW (idle), 25mw (peak) power consumption integrated with an x86 core, CNN and crypto engines, sub-mW wake-up radio, and a 512KB memory, operable from 200kHz to 950MHz.

2.1 SkyLake-SP: A 14nm 28-Core Xeon® Processor

Simon M. Tam, Harry Muljono, Min Huang, Sitaraman Iyer, Kalapi Royneogi, Nagmohan Satti, Rizwan Qureshi, Wei Chen, Tom Wang, Hubert Hsieh, Sujal Vora, Eddie Wang

Intel, Santa Clara, CA

SkyLake-SP (Scalable Performance), code name SKX, is the next generation Xeon® server processor fabricated on the Intel® 14nm tri-gate CMOS technology with 11-metal layers [1,2]. The SKX processor family has three core-count configurations. Each SKX core is accompanied by 1MB of dedicated L2 (2nd level cache) and 1.375MB of non-exclusive L3 (3rd level cache). At its maximum configuration of 28 cores, the SKX processor supports 6 DDR4 channels (2666MT/s), 3×20-lanes UPI processor-to-processor links (10.4GT/s) and x48+4 PCIE links (8GT/s). SKX supports per-core power-performance optimization enabled by on-die integrated voltage regulators (FIVR) [3, 4]. A new 2-dimensional synchronous on-die MESH fabric interconnects all the on-die components. Fig. 2.1.1 shows the overall architecture of the SKX processor.

Comparing to previous generation Xeon® server processors, SKX's higher maximum core-count, increased frequency, and improved IPC provided the generational performance improvements across all relevant server benchmarks. Aggressive core dynamic capacitance (Cdyn) reduction and frequency push were exercised to achieve the power and frequency targets. To facilitate high-volume production and silicon debug, SKX incorporated extensive DFT (design for test) features with high SCAN coverage and analog debug capabilities for critical analog circuits.

SKX deploys a highly configurable floorplan architecture to accommodate multiple products/sockets requirements. Key integration components are: (1) CORE-TILE and (2) on-die-fabric (MESH). The CORE-TILE unit integrates the core, LLC, and the core-to-MESH agent into one readily array-able modular object. The MESH is a 2D synchronous fabric architected to scale with core-tile count at higher data bandwidth and reduced latency relative to its RING predecessor [3] at lower frequencies. The 28-core SKX floorplan started from a 5-by-6 CORE-TILE array. Two CORE-TILEs, one each from the left and right columns were replaced with MEMORY CONTROLLER modules. At the top of the die, the "NORTHCAP" contains the IO agents, serial-IP ports, the clock generator unit (CGU), global power management and the fuse unit. Adding global chassis structures, the on-die-voltage-regulators (FIVR) and the DDR-IO complete the full-chip assembly. Fig. 2.1.2 shows the evolution of the 28-core floorplan.

Figure 2.1.3 shows the 9 primary VCC domains of the 28-core SKX processor physically partitioned into 35 VCC planes. Multiple FIVRs in conjunction with 5 mother board (MB) voltage regulators (MBVRs) serve these VCC partitions. FIVR vs. MBVR assignment was chosen based on die area, current compliance, VCC noise specifications and the power delivery efficiency attributed to MB and package IR loss. FIVR and a dedicated all-digital PLL for each core enables per-core voltage-frequency optimization to achieve the lowest "average" core power. Core-droop mitigation based on specific workload was added to help mitigate the core voltage droop. For the 28-core SKX, two FIVRs supply the un-CORE VCC (VCCCLM) serving the LLC and non-CORE components in the CORE-TILE. The entire un-CORE belongs to one clock domain and is served by a single ADPLL. LC filtered VCCs with the L realized on the package are also available for critical analog circuits (e.g. clock circuits in the high-speed IOs). Clock distribution schemes similar to that in [2] are adopted for the SKX processor. A new block to detect and enable throttling the peak current of FIVR input supply was added in SKX. This helped reduce the amount of VR caps needed to support current surge on that supply.

The MESH forms the high-bandwidth on-die 2D global synchronous fabric interconnecting the AGENTs with the COREs and CACHEs (Fig. 2.1.4). The MESH eliminated the RING-to-RING-bridge logic from the previous RING architecture to provide a low latency 2D cross-point interconnecting every CORE-TILE with its 4 neighbors. It adopts a simple data routing algorithm comprising data being first routed in the vertical direction and then routed horizontally to the destination. This dataflow control, however, makes the vertical MESH latency the most critical factor in determining the overall MESH performance. At low VCCCLM for low un-CORE power, it is difficult to achieve single-cycle vertical MESH latency above ~2GHz due to the vertical TILE-to-TILE RC delay. The solution is to move the critical section of the V-MESH from the VCCCLM domain to a higher fixed voltage VCCIO supply. This technique enabled single-cycle vertical TILE to TILE latency without needing to raise the un-core VCC thus reducing the overall power.

The SKX has a performance enhanced and customized core. Specifically, the SKX core has two AVX (advanced vector extensions) processing units with AVX-512, a larger L2 cache (1MB, 1024 sets by 16 ways) achieving 2× the floating-point operation performance and higher cache bandwidth vs. its predecessor [3]. Additionally, each CORE-TILE has a 2048 sets by 11 ways L3, and a 2048 sets by 12-ways snoop filter (SF) to support the size of the L2. The SKX caches incorporated column and/or row repairs to achieve high manufacturability. Significant FUSE resources were allocated to provide a statistically robust cache repair capability to achieve low VCC in a 28-CORE-TILEs SKX embedded with 28MB L2 and 38.5MB L3. Silicon data showed the minimum VCC exceeding the product requirements. Fig. 2.1.4 shows the CORE-TILE containing the CORE, AVX, L2, LLC, and the CACHE-HOME AGENT (CHA) that interface the cache to the MESH.

The chip features 128 lanes of high-speed IOs including 48 PCIE, 4 DMI, 16 on-package PCIE lanes running at 2.5/5.0/8.0GT/s and 60 UPI lanes running at speeds up to 10.4GT/s. The TX architecture and circuits are based on the design reported in [5]. The RX architecture is an evolution of [5] and includes support for the PCIE Separate Refclk Independent SSC (SRIS) ECN. To achieve the higher speed of 10.4GT/s with minimal power impact relative to prior generations and to enable faster layout convergence, the RX front-end was re-architected to eliminate the variable gain amplifier (VGA) functionality from the CTLE. Instead, a front-end attenuator was introduced before the CTLE. The CTLE circuit topology was carefully constructed to completely avoid the need for a precision resistor. PVT variation was addressed by limiting the number of stages in the CTLE to 2 and by making key performance parameters ratios of device parameters. The topology of the CTLE is shown in Fig. 2.1.5. Simulations show higher than 10.8dB of AC peaking at Nyquist rate of 5.2GHz. The overall transceiver occupies 20% less area and 17% less power, (simulated) compared to the prior generation [3].

SKX has 6-channel DDR4 interfaces each capable of supporting 2-DIMM per channel and speed up to 2666MT/s achieving 128GB/s total memory bandwidth. The 6-channel interface is split into two independent and identical physical sections (3 channels each) residing to the left and right edges of the die. Data bytes are located on north and south sub-sections of the channel layout, while Command, Control, Clock signals and PVT compensation circuitry are located toward the middle section. This floorplan facilitates package routing escapes and pin-out order matching between the CPU and DIMM card resulting in shortened package and board routing length to improve signal integrity. Fig. 2.1.6 shows the SKX DDR4 receiver (RX) Architecture.

The SKX server processor fabricated on a 14nm CMOS process is fully functional and is performing to its specifications. Fig. 2.1.7 shows the die photograph of the 28-core SkyLake-SP processor.

References:

[1] S. Natarajan, et al., "A 14nm Logic Technology Featuring 2nd-Generation Finfet, Air-Gapped Interconnects, Self-Aligned Double Patterning And A 0.0588 μm² SRAM Cell Size," *IEDM*, pp. 3.7.1-3.7.3, 2014.

[2] E. Fayneh, et al., "14nm 6th-Generation Core Processor SoC with Low Power Consumption and Improved Performance," *ISSCC*, pp.72-73, 2016.

[3] B. Bowhill, et al., "The Xeon® Processor E5-2600 v3: A 22nm 18-Core Product Family," *ISSCC*, pp. 78-79, 2015.

[4] A. Nalamalpu, "Design Optimization of Computing Systems - from the Transistor to the Data Center," *ISSCC*, 2017.

[5] F. Spagna, et al., "A 78mW 11.8Gb/s Serial Link Transceiver with Adaptive RX Equalization and Baud-Rate CDR in 32nm CMOS," *ISSCC*, pp. 366-377, 2010.

 978-1-5090-4940-0/18/$31.00

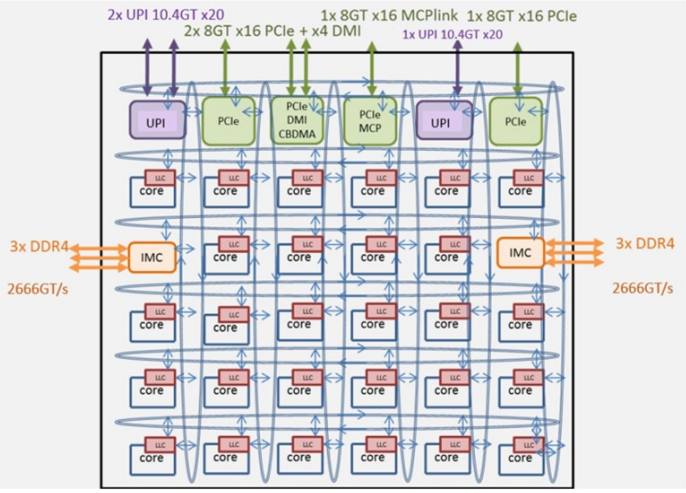


Figure 2.1.1: SKX processor architecture.

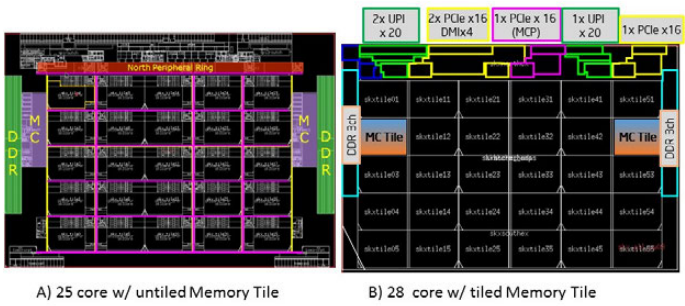


Figure 2.1.2: SKX core-tile-based floorplan evolution.

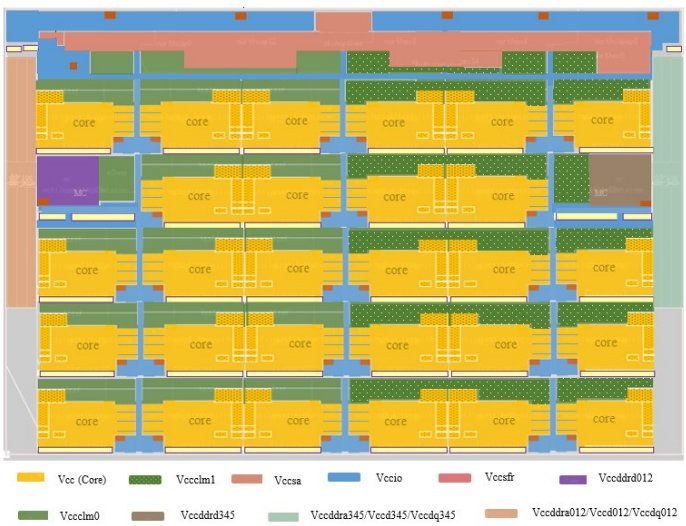


Figure 2.1.3: VCC domains of a 28-core SKX processor.

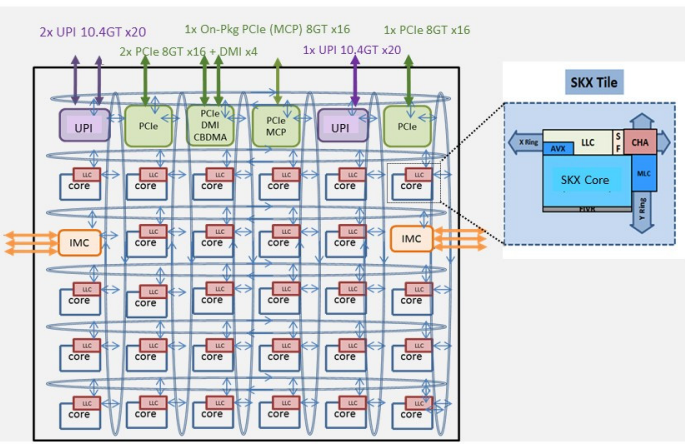


Figure 2.1.4: MESH architecture connecting cores and caches.

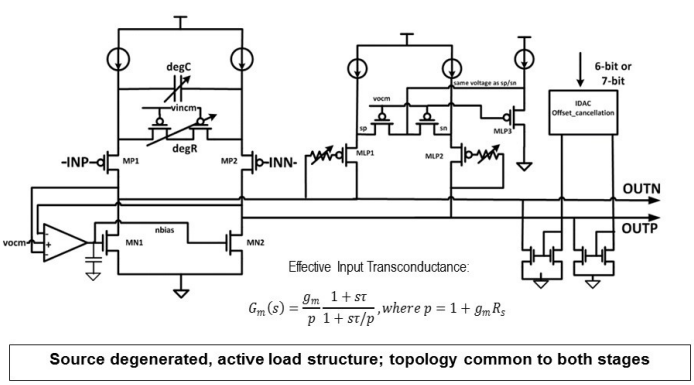


Figure 2.1.5: SKX SERDES CTLE circuit topology.

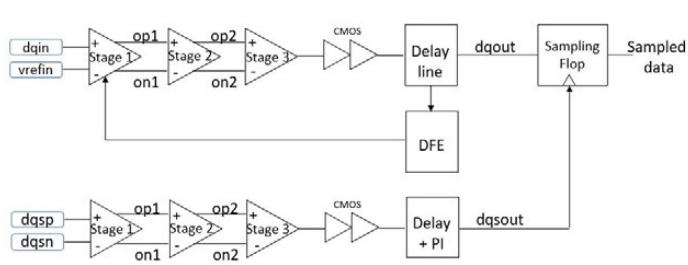


Figure 2.1.6: SKX DDR4 receiver (RX) architecture.

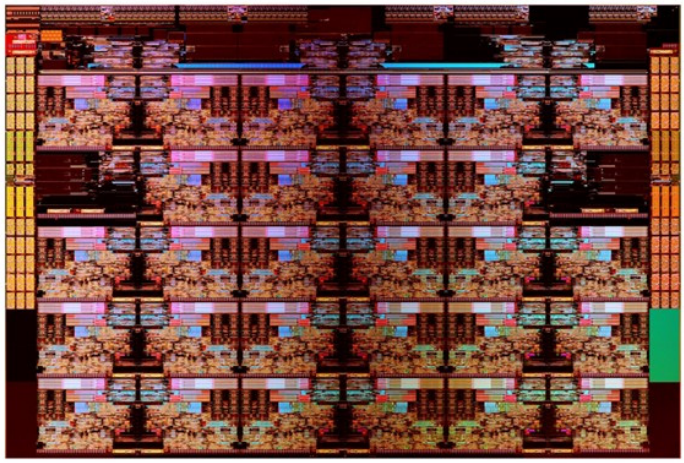


Figure 2.1.7: SKX die photograph.

978-1-5090-4940-0/18/$31.00

2.2 IBM z14™: 14nm Microprocessor for the Next-Generation Mainframe

Christopher Berry[1], James Warnock[2], John Isakson[3], John Badar[3], Brian Bell[4], Frank Malgioglio[1], Guenter Mayer[5], Dina Hamid[1], Jesse Surprise[1], David Wolpert[1], Ofer Geva[6], Bill Huott[1], Leon Sigal[2], Sean Carey[1], Richard Rizzolo[1], Ricardo Nigaglioni[3], Mark Cichanowski[3], Dureseti Chidambarrao[7], Christian Jacobi[1], Anthony Saporito[1], Arthur O'neill[1], Robert Sonnelitter[1], Christian Zoellin[1], Michael Wood[1], Jose Neves[1]

[1]IBM Systems, Poughkeepsie, NY; [2]IBM Systems, Yorktown Heights, NY
[3]IBM Systems, Austin, TX; [4]IBM Systems, Rochester, MN
[5]IBM Systems, Boeblingen, Germany; [6]IBM Systems, Tel Aviv, Israel
[7]IBM Systems, Hopewell Junction, NY

The IBM Z microprocessor in the z14 system has been redesigned to improve performance, system capacity, and security [1] over the previous z13 system [2]. The system contains up to 24 central processor (CP) and 4 system controller (SC) chips. Each CP, shown in die photo A (Fig. 2.2.7), operates at 5.2GHz and is comprised of 10 cores, 2 PCIe Gen3 interfaces, an IO bus controller (GX), 128MB of L3 embedded DRAM (eDRAM) cache, X-BUS interfaces connecting to 2 other CP chips and one SC chip, and a redundant array of independent memory (RAIM) interface. Each core on the CP chip has 4MB of eDRAM L2 Data cache and 2MB of eDRAM L2 Instruction cache, with 128KB SRAM Instruction and 128KB SRAM Data L1 caches. Each SC, shown in die photo B (Fig. 2.2.7), operates at 2.6GHz and has 672MB of L4 eDRAM cache, X-BUS interfaces connecting to CP chips in the drawer and A-BUS interfaces connecting SCs on the other drawers. Both chips are 696mm^2 and are designed in Global Foundries 14nm high performance (14HP) SOI FinFET technology with 17 layers of copper interconnect [3]. The CP contains 6.1B transistors, while the SC contains 9.7B transistors. The total IO bandwidth of the CP and SC are 2.9Tb/s and 5.5Tb/s, respectively.

The maximum system has 4 processor drawers, each with 6 CP chips and 1 SC chip, for a total of 240 physical cores. The maximum number of user configurable cores are 170, with 26 cores available for system-assist processors (SAP), and the others left unusable for power, yield or sparing purposes. The maximum system memory available is 32TB. The system topology was changed slightly in z14 to improve performance by reducing latency of cache access and coherency requests. In Fig. 2.2.1, each drawer is arranged into 2 clusters of 3 fully interconnected CP chips. Each of the chips in each cluster connects directly to the SC. The SC on each drawer connects to the SC on each of the other 3 drawers.

Reliability and availability being part of the foundation for the IBM Z system, the L3 cache error-correction code (ECC) was improved significantly in z14. The previous bit-based single error-correction double-error detection (SEC/DED) code was replaced with symbol based Reed-Solomon ECC code. This allows the L3 to correct any failures within a single symbol and detect any failures within two symbols, allowing complete failure of an eDRAM in a data access to be correctable. To achieve this the data must be striped across enough eDRAMs to ensure each ECC word only has one symbol accessed from any given eDRAM instance. Additionally, to minimize data storage overhead associated with the more robust symbol ECC, the size of individually correctable chunks of data doubled, going from 64b to 128b. The implications of this combined with doubling the cache size, from a physical design perspective, was a 4× increase in the connectivity between the individual 1152 eDRAM instances. The most significant contributors, as shown in Fig. 2.2.2, are the local eDRAM interconnect as well as the global interconnect (yellow, purple, and green), which are each twice as wide connected to twice as many eDRAMs. From a performance perspective, there were latency and L3 availability implications. Both were mitigated with a combination of careful floorplanning and performance-aware logic and circuit implementation.

As with all technology transitions new challenges need to be overcome. One of those challenges were new stress and distortion effects created by combining wide-scale usage of eDRAM, made of high-density deep-trench (DT) capacitors, and FinFET technology. Higher DT density gradients across the die could lead to distortions that worsen focus and gate-height control. New analysis was needed to assess the impact of the DT density gradient. Based on the analysis, multiple adjustments and iterations to the content, kerf and floorplan were required to reduce the simulated distortions. One of the more significant adjustments associated with this was in the SC chip. The earliest design point contained an SRAM L4 directory to reduce latency. The distortion analysis found that the DT density gradient was too large and a design change to an eDRAM L4 directory was required to reduce the gradient and distortions sufficiently. As shown in Fig. 2.2.3, the final DT density gradient (and hence distortion) after many iterations of the floor plan change was much smaller with the eDRAM directory than with the SRAM directory. Simulations indicated distortions would be small enough with a decrease of over 3× relative to the early design point and the chip data with the final design confirmed that distortions were contained.

Power reduction was an important theme in designing the z14. For each new technology, it is increasingly difficult to balance performance and power to achieve yield targets for such a large chip. For z14 many foundational aspects of our design and assumptions associated with them were revisited, with a new emphasis on power savings. One area where we found significant power savings, of nearly 10W, was in our local clock block (LCB) design. A high level before and after schematic for the design is in Fig. 2.2.4. In the original LCB, the functional and scan-clock generation portions of the design are intertwined and optimized for total FET width and clock delay. Revisiting the analysis for different operational modes, it was clear the design was not power optimized for normal machine operation. New gates were added to the design, increasing total FET width, to gate off the scan clocks when in functional mode. The added gate width and associated leakage power was compensated for with a reduction in the active power for the overall power reduction in functional mode.

The cryptographic assists for the Advanced Encryption Standard (AES) in the z14 core co-processor (COP) unit are executed on a significantly re-designed pipeline that computes two AES rounds in three cycles [1]. This provides a latency reduction of 25% compared to the previous AES pipeline. The COP unit has two pipelines allowing for 6 blocks of 16B to be processed in parallel. The longer pipeline reduced routing pressure on the feedback path and refactoring of the linear steps of the algorithm together with a twisted BDD S-box enabled timing closure under the challenging cycle time target. A new datapath in the COP unit can supply encryption results to the GHASH hardware directly and further speeds up the AES Galois-Counter Mode (GCM). These improvements result in a 6-7× throughput enhancement compared to z13 for the AES modes used in transmission and storage of data. For secure hashing, the z14 COP unit has added fixed-function hardware for the SHA-3 encryption standard that computes a round every two cycles.

The logical directory in the core was optimized for power, area and performance. In previous generations, the L1 and L2 caches were virtual-address indexed and absolute-address tagged which causes large amounts of directory and translation lookaside buffer (TLB) data to be accessed for each load and store. In z14, the L1 cache directory was updated to a logical-indexed, logical-tagged structure, enabling an L1 hit to be determined without a TLB access. Integration of the 2nd-level TLB and L2 directory pipeline with the L1 caches have enabled the L2 access to be very fast, less than 1.6ns for the 4MB L2 data cache. A block diagram of the updated structure can be found in Fig. 2.2.5.

Despite the accelerating complexity of circuit function and technology enablement, the CP chip continues to improve upon fundamental performance capabilities over the previous generations. These diverse improvements are distilled into the achievable performance per volt in Fig. 2.2.6, plotting the hardware characterization at the minimum functional voltage (Vmin) needed to achieve the design's desired 5.4GHz chip frequency guardband. This minimum voltage is shown plotted across a ±5% normalized process delay range alongside similar results from the previous 45nm, 32nm, and 22nm processor generations. As shown, the voltage-frequency design point achieved by the CP extends the trend of efficiency improvements despite both technology and design challenges, including the elimination of the highest device threshold in the move to 14nm, the increased frequency guardband, addition of two cores, and increased L2 and L3 cache sizes.

Acknowledgements:
The authors would like to thank the entire IBM Z team, the IBM Enterprise Systems Product Engineering team, the IBM Research team, and the IBM EDA team for all their hard work and contributions to the success of this project.

References:
[1] C. Jacobi, et al., "The Next Generation IBM Z Systems Processor," *Hot Chips*. 2017.
[2] J. Warnock, et al., "22nm Next-Generation IBM System z Microprocessor," *ISSCC,* pp. 70-71, 2015.
[3] C-H. Lin, et al., "High Performance 14nm SOI FinFET CMOS Technology with 0.0174μm2 embedded DRAM and 15 Levels of Cu Metallization", *IEDM*, pp. 3.8.1-3.8.3., 2014.

 978-1-5090-4940-0/18/$31.00

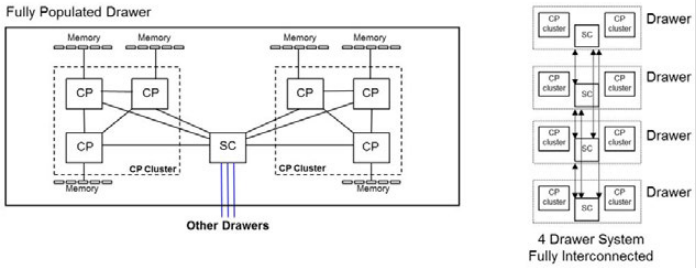


Figure 2.2.1: Drawer topology on left with system topology on right.

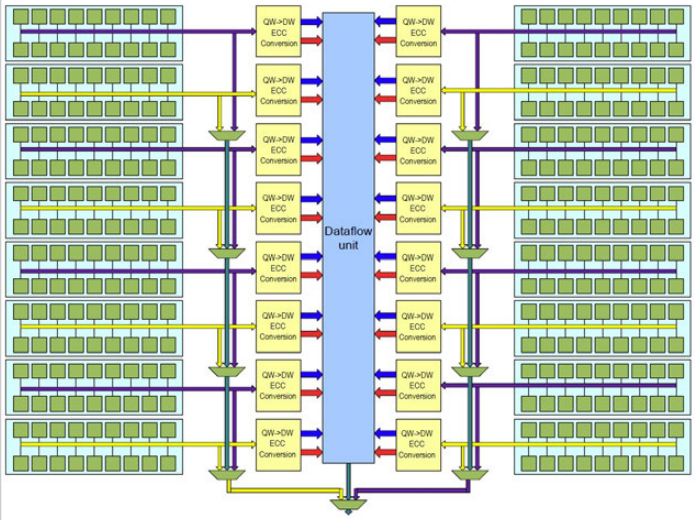


Figure 2.2.2: Dataflow of one fourth of the L3 Cache. Green boxes represent individual eDRAMs.

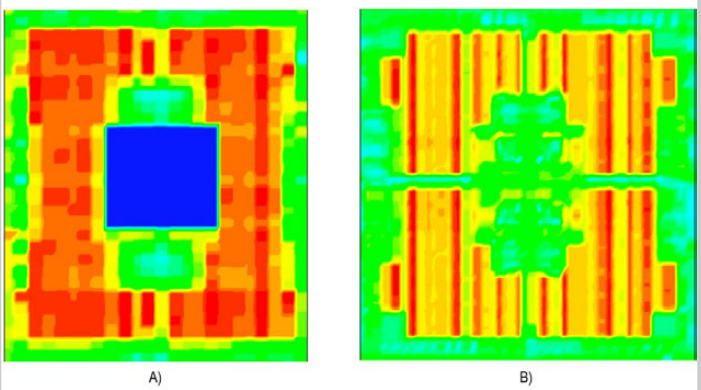


Figure 2.2.3: DT density on SC chip before (A) and after (B) floorplan and design improvements.

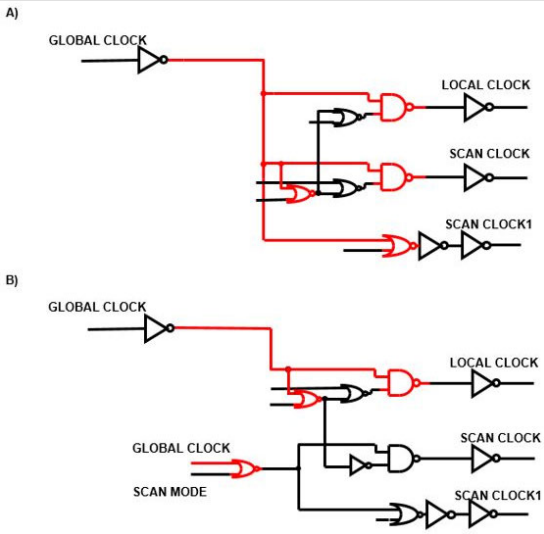


Figure 2.2.4: Local clock block before (A) and after (B) the redesign to reduce power in functional mode.

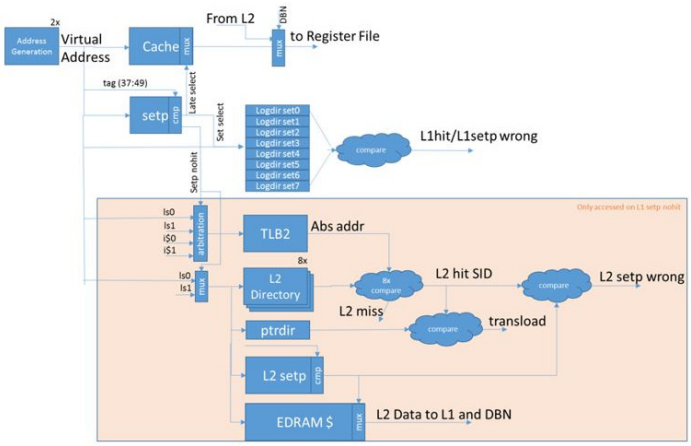


Figure 2.2.5: Diagram of updated L1 and L2 directories and TLB structures.

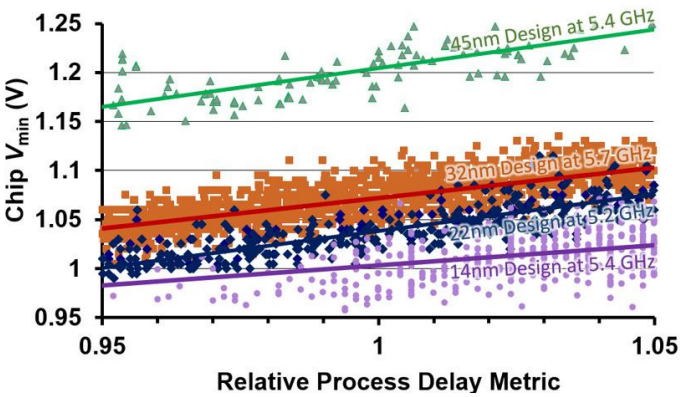


Figure 2.2.6: Process delay metric plotted against chip V_{min} across 4 different technology generations.

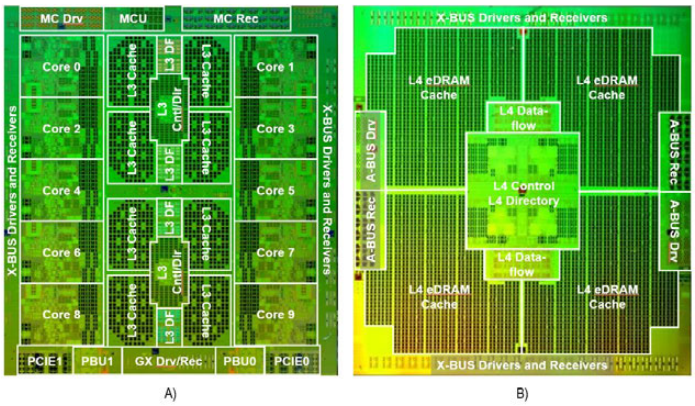


Figure 2.2.7: CP chip on left (A); SC chip on right (B).

978-1-5090-4940-0/18/$31.00

2.3 An Energy-Efficient Graphics Processor Featuring Fine-Grain DVFS with Integrated Voltage Regulators, Execution-Unit Turbo, and Retentive Sleep in 14nm Tri-Gate CMOS

Pascal Meinerzhagen[1], Carlos Tokunaga[1], Andres Malavasi[1], Vaibhav Vaidya[1], Ashwin Mendon[1], Deepak Mathaikutty[1], Jaydeep Kulkarni[1], Charles Augustine[1], Minki Cho[1], Stephen Kim[1], George Matthew[1], Rinkle Jain[1], Joseph Ryan[1], Chung-Ching Peng[1], Somnath Paul[1], Sriram Vangal[1], Brando Perez Esparza[1], Luis Cuellar[1], Michael Woodman[1], Bala Iyer[1], Subramaniam Maiyuran[2], Gautham Chinya[1], Chris Zou[1], Yuyun Liao[1], Krishnan Ravichandran[1], Hong Wang[1], Muhammad Khellah[1], James Tschanz[1], Vivek De[1]

[1]Intel, Hillsboro, OR; [2]Intel, Folsom, CA

Graphics workloads are highly dynamic in nature, using multi-threaded SIMD execution units (EUs), fixed-function units, samplers, and media accelerators to provide ever-increasing amounts of graphics performance. These workloads are often limited by power and thermal constraints, requiring dynamic voltage/frequency scaling (DVFS) of the graphics processor (GPU). This coarse-grain DVFS, driven by a power-management IC (PMIC) setting a shared rail voltage (V_{IN}), incurs performance loss while waiting for PLL re-lock and slow-rail voltage transitions. In addition, it does not allow a performance-critical unit (e.g. an EU) to use on demand a higher V/F (e.g. for EU turbo) without an energy penalty for the rest of the GPU.

In this paper, we present a GPU in 14nm tri-gate CMOS featuring: (1) fine-grain DVFS where the key power/performance critical blocks for 3D graphics processing, i.e., the EUs, are powered by a digitally controlled integrated voltage regulator (IVR) for fast on-die voltage (V_{OUT}) control, (2) retentive sleep, and (3) reliability-aware wake-up, all implemented using distributed power gates (PG). The GPU responds quickly to instantaneous workload demands by boosting performance of only the bandwidth-critical units while reducing V/F of the non-critical units. In particular, we demonstrate EU turbo as an example of fine-grain DVFS in a GPU for improving performance and energy efficiency of EU-dominant workloads. During EU-intensive periods, EUs are set to a high voltage V_{HIGH} (V_{IN}=V_{HIGH}) and operate at a readily available 2× clock frequency, while other blocks operate at a low voltage V_{LOW} using IVRs (V_{OUT}=V_{LOW}) and the 1× clock. The IVR supports a wide range of load currents and operating voltages, including near-threshold voltage (NTV). The embedded graphics register file (GRF) and ROM arrays use voltage boosting from V_{IN} as an assist technique to achieve NTV operation. The IVRs are dynamically configured to clamp V_{OUT} to the retention limit (V_{RET}) during short stall periods. Digitally controlled and programmable underdrive of selected PGs is utilized to aid IVR regulation, retention clamping, and reliability-aware wake-up.

The system prototype (Figs. 2.3.1 and 2.3.7) features a Gen9LP GPU design [1] containing 3 sub-slices (SS) of 6 EUs each. Two of these SS modules (SSM) feature the IVR, retention clamp, turbo support, and array V_{MIN} reduction techniques, while the remaining one is left unmodified as a reference. A system agent (SA) serves as an interface to feed the GPU with high-bandwidth data for 3D graphics workloads, and to handle GPU configuration, boot, and power management. To enable accurate power/performance measurement for key graphics workloads, the SA includes a 4MB paging cache and a controller that communicates to a host PC through an FPGA and PCI interface. The testchip also includes multiple JTAG scan chains for configuration, debug and observability of the different implemented techniques. EU turbo operation at 2× clock (Fig. 2.3.2) is enabled by EU logic modifications as well as a glitch-less clock divider/mux and synchronization logic to enable data transfer to/from the 1× clock domain. A power/turbo controller sequences the V_{OUT} and clock frequency changes to efficiently boost EU performance. Voltage level shifters at EU boundaries enable a seamless interface to other GPU blocks and also allow the IVR to compensate for process/temperature-induced EU-EU V_{MIN} variations. Read and write word-lines in the 32KB GRF and ROM are boosted using the IVR input voltage (V_{IN}) for EU V_{MIN} improvement.

A digitally controlled hybrid DLDO/SCVR [2] IVR offers high conversion efficiency over a wide V_{OUT} range. Fully distributed designs of both DLDO and SCVR (Fig. 2.3.3) are implemented to reduce demands for low-resistance metal resources in the thick upper layers, while maintaining low IR drop. The DLDO is designed for low V_{OUT} ripple and to meet the PG transistor self-heating and EM reliability constraints across a large range of load currents and V_{OUT} values, especially under low-load and large dropout conditions [3]. The DLDO uses an under-drive voltage (V_{UD}) for the two-way stacked primary PG (PPG) to limit PG current density. V_{UD} is generated by an R-2R DAC in the central IVR controller. The parallel secondary PGs (SPGs) are turned on by the DLDO controller only if the PPGs are fully utilized at high load or in bypass mode. The 1,420 PPG + SPG units are distributed in a checker-board pattern across the EU. A central spine of PG drivers controls the PGs at half-row granularity. The PG drivers form a distributed shift register (SR), receiving increment/decrement instructions from a central IVR controller. In order to minimize the DLDO control power overhead, the SR clock is gated at the fine granularity of PG drivers, as opposed to at only 4 clock sections [4]. The DLDO uses a linear fine-grain control loop for steady-state conditions, and a non-linear coarse-grain control for fast droop mitigation. Fast droop mitigation is achieved by sending an asynchronous preset signal to all flip-flops in the SR, which quickly turns on all PPGs, and optionally all SPGs as well. The DLDO can be dynamically configured as a retention clamp, where it sets V_{UD} such that V_{OUT} does not fall below the retention voltage (V_{RET}). The R-2R DAC is also used to generate a programmable V_{UD} ramp for self-heating and EM compliant EU wake-up.

The SCVR consists of 6 distributed power tiles (Fig. 2.3.3). 40% of on-die high-density MIM caps on top of the EUs are used to implement the fly capacitors. It operates across a 0.3-to-0.7V V_{OUT} range, and automatically transitions between 3:2, 2:1, and 3:1 V_{IN}/V_{OUT} ratios using the same two fly caps. The SCVR uses a fast digital controller for regulation and droop mitigation, and a slow controller for mode transitions, ripple management, and efficiency tracking. IVR reference voltage generator and voltage comparator circuits are shared between the DLDO and SCVR.

The measured current efficiency of the DLDO, when running an active workload, ranges from 93% to 95% for 0.785-to-1.11V V_{OUT} with 1.15V V_{IN}, at 25°C. The measured average power efficiency of the DLDO is only 5.5% below ideal over this operating range (Fig. 2.3.4). Under light load conditions, when EUs are idle, the measured DLDO power efficiency is 13.6% below ideal. The SCVR provides a 0.44-to-0.57V V_{OUT} range with a peak (average) power efficiency of 72% (69.8%) in the 2:1 mode and 58% (56%) in the 3:2 mode. With the DLDO configured in the open loop as retention clamp, V_{UD} values around V_{IN}-V_{TH} successfully clamp V_{OUT} close to V_{RET}≈0.5V, thus reducing EU leakage current by 62.5% for V_{IN}=1.15V, or 25% for V_{IN}=0.8V, during short stall periods.

Figure 2.3.5 demonstrates DLDO-enabled EU turbo with bypass-mode voltage boost (Fig. 2.3.6 oscilloscope capture) to improve GPU performance or reduce energy. The baseline GPU uses a PMIC for slow, coarse-grain DVFS from 0.51V/50MHz to 1.2V/400MHz for varying compute demands. Baseline measurements are taken from a modified SSM with EU turbo turned off. For a measured workload with 53.6% EU utilization, EU turbo offers performance improvements of up to 40%, with an average 36.6% improvement across the entire DVFS range. However, without independent IVRs for all major GPU blocks on the shared V_{IN} rail, the GPU energy/performance for EU turbo degrades significantly since V_{IN} is raised to support the 2× clock for the EUs in bypass mode. Therefore, all major blocks in the GPU must have independent IVRs to fully exploit the benefits of EU turbo. For 100% EU utilization and independent IVRs for all major GPU blocks, EU turbo enables up to 19% (average 17%) energy reduction at iso-performance. For lower EU utilizations, such as 80% or lower, EU turbo energy consumption can be worse than baseline, especially for high performance levels. This energy penalty is reduced by using a dual-rail system where only the EUs are IVR-enabled, while all other GPU blocks are powered by a separate PMIC-controlled external rail at V_{LOW}, in which case EU turbo enables up to 32% (average 29%) energy reduction at iso-performance for 100% EU utilization.

The oscilloscope captures of Fig. 2.3.6 demonstrate R-2R DAC-based reliability-aware EU wake-up enabled by a digitally controlled PPG V_{UD} ramp, as well as DLDO regulation of V_{OUT} to 0.7V, along with transition to bypass mode (V_{IN}=0.97V) triggered by an EU turbo request. In addition, wide-range hybrid DLDO/SCVR IVRs can be used to enable energy-efficient DVFS at the SoC level [3], where the GPU and CPU share the same V_{IN} rail (Fig. 2.3.6).

Acknowledgements:
The authors thank the many members of the Intel Labs circuit research, microarchitecture research, and silicon/system prototyping teams that contributed to this work, as well as the Intel Visual and Parallel Computing Group. This research was, in part, funded by the U.S. Government (DARPA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References:

[1] "The Compute Architecture of Intel Processor Graphics Gen 9," available online at: https://software.intel.com/sites/default/files/managed/c5/9a/The-Compute-Architecture-of-Intel-Processor-Graphics-Gen9-v1d0.pdf
[2] S. Kim, et al., "Enabling Wide Autonomous DVFS in a 22nm Graphics Execution Core Using a Digitally Controlled Hybrid LDO/Switched-Capacitor VR with Fast Droop Mitigation," *ISSCC*, pp. 154-155, 2015.
[3] R. Muthukaruppan, et al., "A Digitally Controlled Linear Regulator for Per-Core Wide-Range DVFS of Atom Cores in 14nm Tri-Gate CMOS Featuring Non-Linear Control, Adaptive Gain and Code Roaming," *ESSCIRC*, pp. 275-278, 2017.
[4] 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. 98-99, 2015.

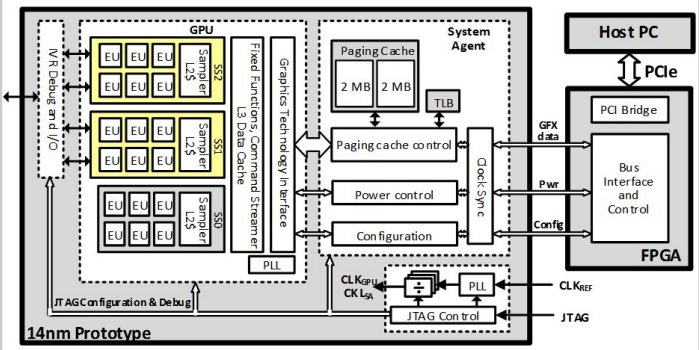


Figure 2.3.1: 14nm graphics processor system block diagram with FPGA for SoC emulation and host PC interface.

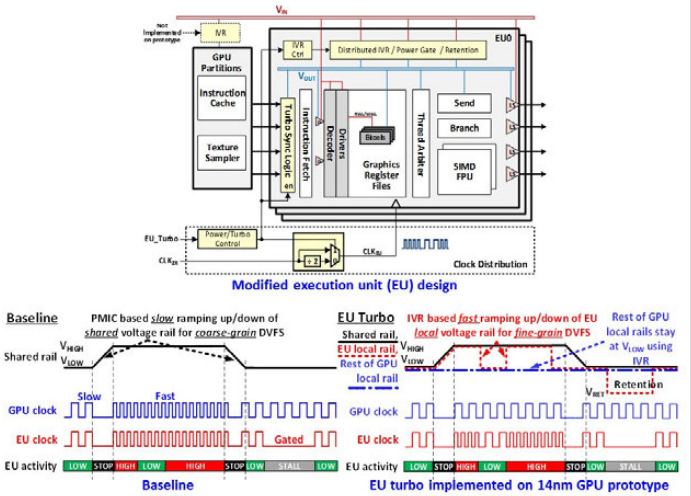
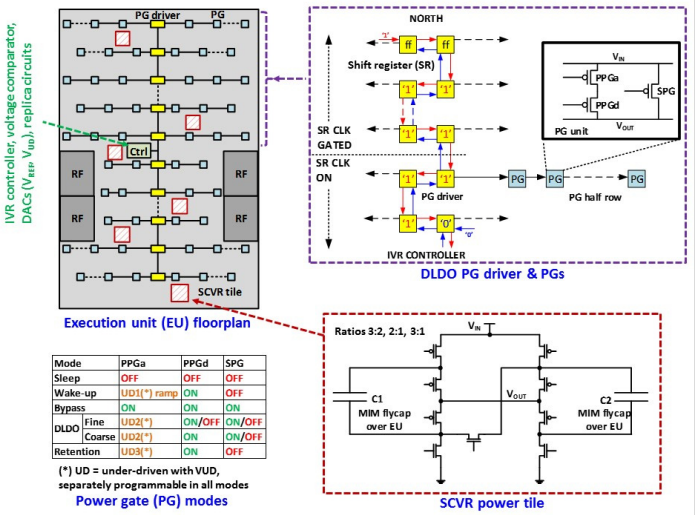


Figure 2.3.2: Modified execution unit (EU) design with IVRs (DLDO, SCVR), retention, GRF boost, and usage model of IVRs for fast, fine-grain, intra-GPU DVFS.



Mode	PPGa	PPGd	SPG
Sleep	OFF	OFF	OFF
Wake-up	UD1(*) ramp	ON	OFF
Bypass	ON	ON	ON
DLDO Fine	UD2(*)	ON/OFF	ON/OFF
DLDO Coarse	UD2(*)	ON	ON/OFF
Retention	UD3(*)	ON	OFF

(*) UD = under-driven with VUD, separately programmable in all modes

Power gate (PG) modes

Figure 2.3.3: Floorplan of distributed IVRs in the modified EU, and distributed DLDO and SCVR design details.

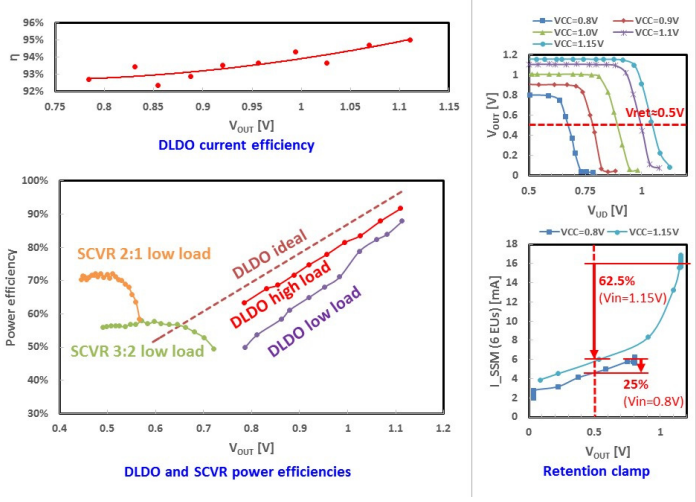
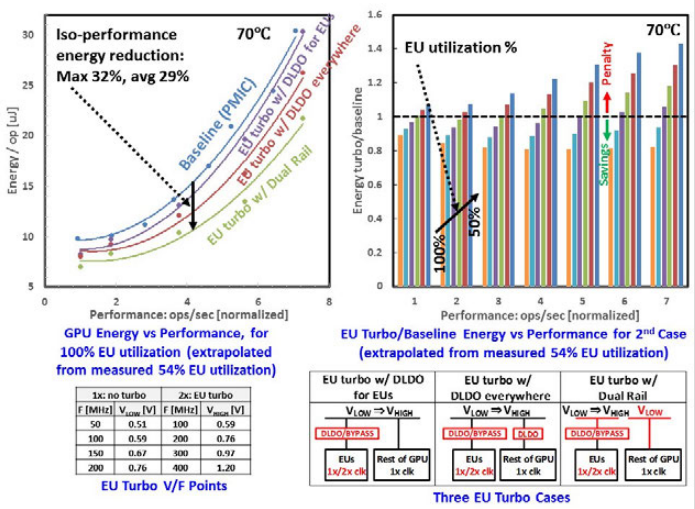


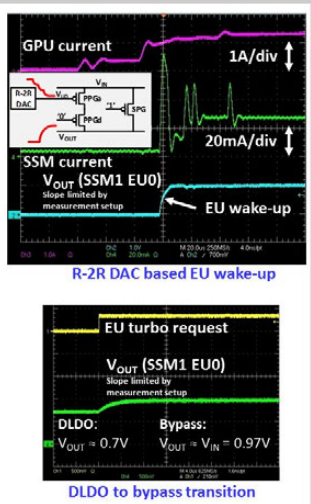
Figure 2.3.4: Measured DLDO and SCVR current and power efficiencies, and up to 62.5% power reduction enabled by retention clamp (open-loop DLDO with primary PG under-drive).



1x: no turbo		2x: EU turbo	
F [MHz]	V_LOW [V]	F [MHz]	V_HIGH [V]
50	0.51	100	0.59
100	0.59	200	0.76
150	0.67	300	0.97
200	0.76	400	1.20

EU Turbo V/F Points

Figure 2.3.5: Measurement results for fine-grain, intra-GPU DVFS: EU turbo enables up to 32% energy reduction at iso-performance.



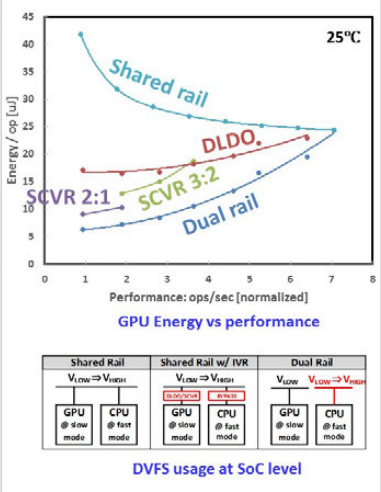
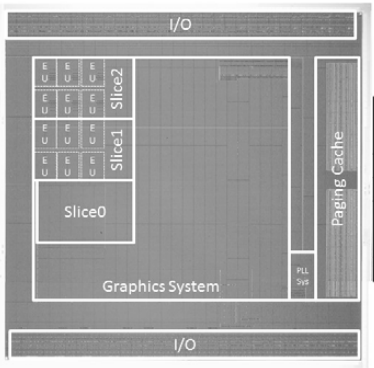


Figure 2.3.6: Oscilloscope captures showing R-2R DAC-based EU wake-up for PG transistor self-heating and EM compliance, DLDO operation, transition to bypass, and measurement results for SoC level use of IVRs (DLDO, SCVR).



Technology	14nm, 10-metal layer, tri-gate high-K/MG CMOS
Test-chip die area	8.0 x 8.0 mm^2
Transistor count	1.542 Billion
GPU V/F	0.51V, 50Mhz – 1.20V, 400MHz (turbo)
Package	FCBGA1595l

Figure 2.3.7: Die micrograph and design details.

978-1-5090-4940-0/18/$31.00

2.4 "Zeppelin": An SoC for Multichip Architectures

Noah Beck[1], Sean White[1], Milam Paraschou[2], Samuel Naffziger[2]

[1]AMD, Boxborough, MA
[2]AMD, Fort Collins, CO

Codenamed "Zeppelin", AMD's next-generation System-on-a-Chip (SoC) was designed for use in multiple products and packages in multiple markets, including server, mainstream PC desktop, and high-end desktop. Utilizing GLOBALFOUNDRIES' 14nm LPP FinFET process technology, the "Zeppelin" SoC has over 4.8B transistors. It contains high-performance AMD x86 cores codenamed "Zen" [1][2], caches, memory controllers, PCIe®, SATA, and other IO controllers, and integrated x86 southbridge chipset capabilities. All these functions are connected on the SoC and between multichip packages and multi-socket systems by AMD Infinity Fabric.

The "Zeppelin" SoC was architected with leadership server capabilities as the top priority, but retained features in the single "Zen"-based SoC to support other complementary markets as well:

- Server market: 4-chip SP3 package with 8 DDR4 channels and 128 PCIe® Gen3 lanes, scalable to 2-socket systems with coherent interconnect.
- Client market: Single-chip AM4 package with 2 DDR4 channels and 24 PCIe® Gen3 lanes, platform-compatible with other AMD SoCs.
- High-end desktop market: 2-chip sTR4 package with 4 DDR4 channels and 64 PCIe® Gen3 lanes.

Within the Infinity Fabric (IF), the Scalable Data Fabric (SDF) plane is designed to provide the coherent data transport between cores using the cache-coherent master (CCM), memory using the unified memory controller (UMC), and IO using the IO master/slave (IOMS) (Fig. 2.4.2). The chip-to-chip communication method in the SDF is key to this multichip package approach, enabled by the Coherent AMD Socket Extender (CAKE) component of the SDF. CAKE is designed to take requests/responses from the local chip's SDF, and encode them into flits at 128b per clock cycle. CAKE is bidirectional, also decoding flits each cycle. The flits are suitable for transmission over any serializer/deserializer (SerDes) interface to another chip within the system. To eliminate clock-domain-crossing latency, the clock of all the SDF components, including the CAKE component in "Zeppelin", run at the system DRAM's MEMCLK frequency.

Two different SerDes types are used with CAKE in "Zeppelin": one for Infinity Fabric on-package (IFOP) traffic and one for Infinity Fabric inter-socket (IFIS) traffic. The IFOP SerDes is designed for minimum power across short in-package trace lengths, while the IFIS SerDes is needed for communication across longer socket-to-socket trace lengths.

A custom IFOP SerDes design was created achieving a power efficiency of 2pJ/b. Key elements in achieving the power target:

- 32b of low-swing, single-ended data with differential clock consuming ~50% the power of an equivalent differential driver.
- Zero power driver state during logic-0 transmission due to transmit/receive impedance termination to ground while the driver pullup is disabled, also applied during link idle.
- Data-bit inversion encoding optimizes bit patterns transmitted to take advantage of the low-power logic-0 state (Fig. 2.4.1) saving 10% average power per bit.

To support the high reliability required by server systems, CRC is transmitted along with every cycle of data. The IFOP SerDes transmission bitrate is 4 transfers per CAKE clock. The bandwidth of an IFOP link is overprovisioned by about a factor of two relative to DDR4 channel bandwidth for mixed read/write traffic to provide robust multi-chip performance scaling.

The SerDes used for IFIS also supports PCIe® and SATA protocols. To align the package pinout with standard PCIe® device lane counts, the IFIS SerDes transmits and receives 16 differential data lanes at roughly 11pJ/b. The IFIS SerDes interface runs at 8 transfers per CAKE clock. Due to the 16b data width and in-band CRC overhead, the bandwidth of an IFIS link has 8/9 of the bandwidth of an IFOP link.

Optimization of SerDes placement in the SoC floorplan required careful consideration of the pinout and routing challenges of the 2-socket-capable 4-chip SP3 package (Fig. 2.4.2). The package pinout is dominated by eight DDR4 channels (4 per side) and eight 16-lane high-speed SerDes links (4 each top and bottom). To support routing the DDR4 channels in the package, the chips on the left side were rotated 180° compared to the chips on the right. To support common platform configurations, the SerDes providing IFIS links to a second socket were routed to the top of the package with each chip providing one such link. The remaining SerDes links, supporting PCIe®/SATA, were routed to the bottom of the package.

While only one 16-lane SerDes per chip needs to support IF traffic at a time, muxing the IFIS capability to SerDes placed at opposing corners of the chip allowed half of the DDR4 routes to coexist on the same two package signal routing layers as the IFIS/PCIe®/SATA SerDes routes (Fig. 2.4.3). A second PCIe® controller on the chip enabled rotated chips to support PCIe® to the bottom of the package, and also enabled support for the I/O-heavy single-socket SP3 option – up to 128 lanes of PCIe®.

Three IFOP SerDes from each chip were required to support full connectivity in the 4-chip package. However, four SerDes were ultimately placed in order to keep all IFOP SerDes package routes restricted to two package layers, and to allow those two package layers to also provide half of the DDR4 package routes. Two IFOP SerDes are placed on the side opposite DDR4 and used by all chips. The DDR4 PHY itself is placed in between two additional IFOP SerDes. One of these two SerDes is unused and clock gated on each chip. The resulting IO and core complex locations are overlaid on the die image in Fig. 2.4.4. The total SoC die area is 213mm^2. Having met the server 4-chip package routing challenges, a 2-chip sTR4 package (Fig. 2.4.5) was created with half of the high-speed IO pins, and a single-chip AM4 package was created to drop into previously existing AM4 platforms.

The IFOP SerDes and digital logic, such as CAKE, that support the 4-chip architecture of SP3, add area which adds to cost; the total silicon area in SP3 is 852mm^2. Creating a monolithic 32-core die without the multichip support would only save about 10% of the area, resulting in a 777mm^2 die [3]. AMD projects that the large die would cost ~40% more to manufacture and test than four small chips. Adding to the cost benefits, the multichip design provides ~20% higher full 32-core yield than would the single-chip version. To make only the top-of-stack 32-core parts, the cost for the large die jumps to 70% more than the cost of the 4 small chips. A very high-yielding multichip assembly process is required, or the improved silicon yields are lost at the package level. AMD internal data has demonstrated success at achieving assembly yields that have a negligible impact on overall cost. In order to ensure that chips with similar maximum frequency capabilities can be matched to each other for assembly into the same package, on-die frequency sensors containing representative critical-path logic are consulted before chips are selected for assembly into packages [4].

The SP3 package delivers power with ±25mV accuracy to all cores, as seen in Fig. 2.4.6, which shows the measured variation in core voltage across the SP3 package, chips and cores running a maximum power pattern at 2.5GHz. On-die per-core low-drop-out (LDO) voltage regulators reduce voltage to faster cores to save power. Idle cores are power-gated for maximum power savings.

The design choices made in the "Zeppelin" SoC definition enabled a wide variety of products for both legacy and new platforms, ranging from single-chip up to 8-chip in the largest 2-socket configuration. The careful balance between compute cores, memory and IO capability on the base design, with the high-bandwidth fabric provides scalable performance across these products (Fig. 2.4.7). As Moore's Law slows in its ability to deliver more transistors per area, multichip architectures such as "Zeppelin" are necessary to provide continued increases in functionality that can be delivered to a single package.

References:

[1] T. Singh, et al., "The Next-Generation High-Performance x86 Core: Zen," *ISSCC*, pp. 52-53, 2017.
[2] M. Clark, "A New x86 Core Architecture for the Next Generation of Computing," *Hot Chips*, 2016.
[3] K. Lepak, et al., "The Next Generation AMD Enterprise Server Product Architecture," *Hot Chips*, 2017.
[4] S. Sundaram, et al., "Bristol Ridge: A 28-nm x86 Performance-Enhanced Microprocessor Through System Power Management," *IEEE JSSC*, vol. 52, no. 1, pp. 89-97, 2017.

 978-1-5090-4940-0/18/$31.00

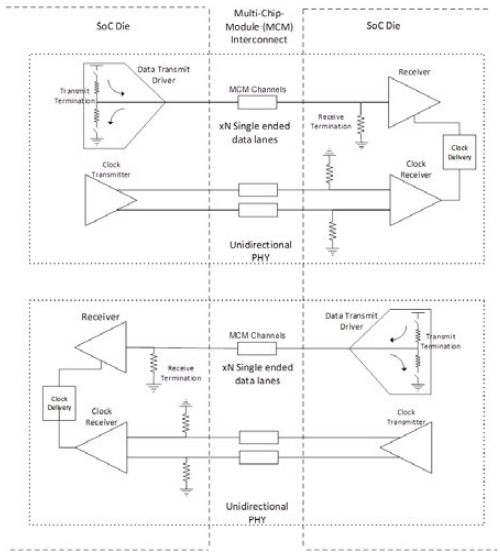


Figure 2.4.1: Infinity Fabric on-package SerDes link circuit diagram.

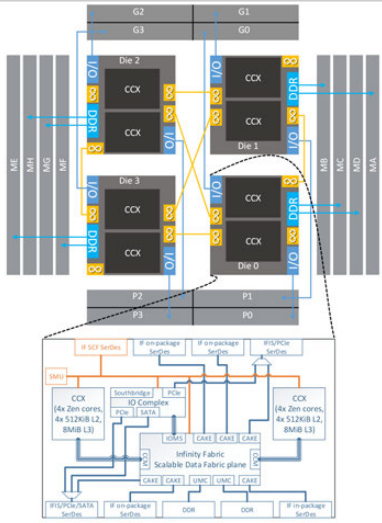


Figure 2.4.2: SP3 package pinout with high-speed connectivity and die architectural detail.

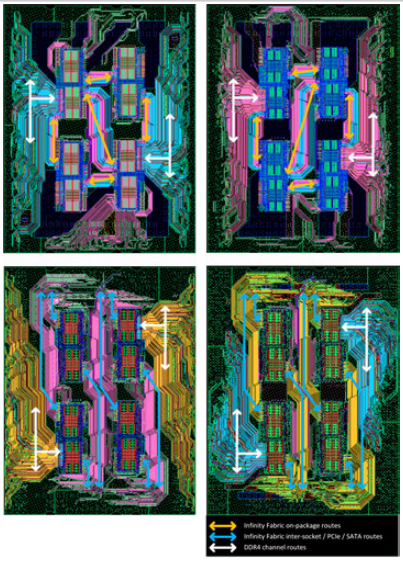


Figure 2.4.3: Four signal routing layers of the SP3 package.

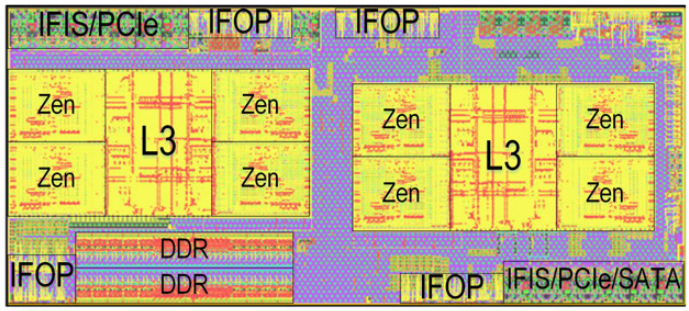


Figure 2.4.4: Die image.

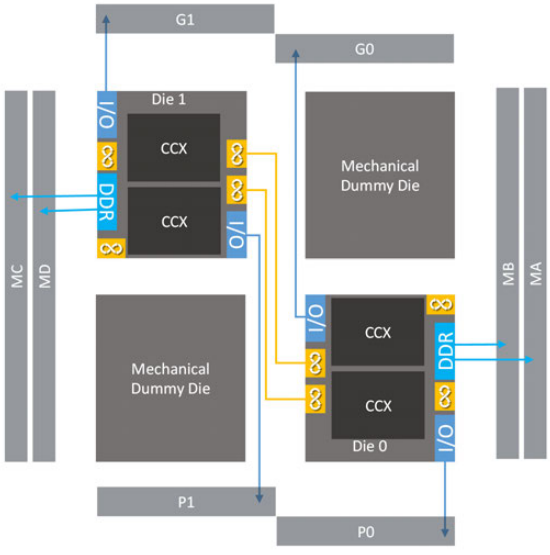


Figure 2.4.5: sTR4 high-end desktop package.

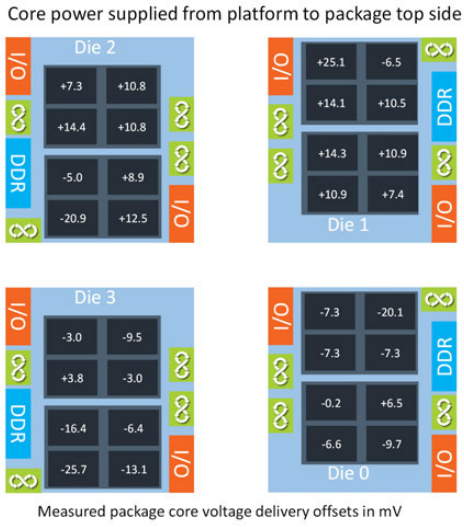


Figure 2.4.6: Measured variation in core voltage across SP3 package core locations.

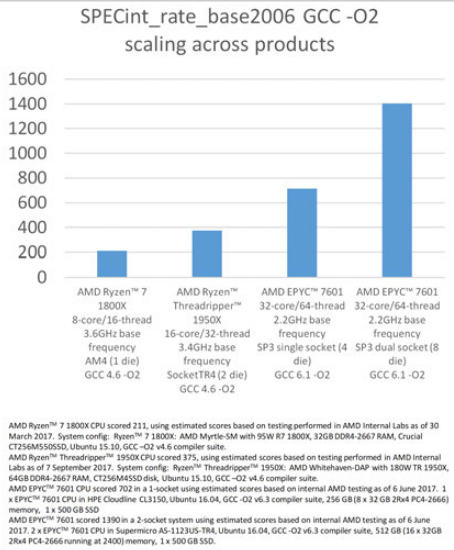


Figure 2.4.7: SPECint_rate_base2006 GCC -O2 scaling.

978-1-5090-4940-0/18/$31.00 ©2018 IEEE

2.5 An Energy-Efficient Reconfigurable DTLS Cryptographic Engine for End-to-End Security in IoT Applications

Utsav Banerjee, Chiraag Juvekar, Andrew Wright, Arvind, Anantha P. Chandrakasan

Massachusetts Institute of Technology, Cambridge, MA

End-to-end security protocols, like Datagram Transport Layer Security (DTLS) [1], enable the establishment of mutually authenticated confidential channels between edge nodes and the cloud, even in the presence of untrusted and potentially malicious network infrastructure. While this makes DTLS an ideal solution for IoT, the associated computational cost makes software-only implementations prohibitively expensive for resource-constrained embedded devices. We address this challenge through three key contributions: reconfigurable cryptographic accelerators enable two orders of magnitude energy savings, a dedicated DTLS engine offloads control flow to hardware reducing program code and memory usage by ~10×, and an on-chip RISC-V core exercises the flexibility of the cryptographic accelerators to demonstrate security applications beyond DTLS.

Figure 2.5.1 summarizes the two major phases of the DTLS protocol: handshake and application data. The handshake phase consists of four steps. In the first step, the client (edge node) and the server agree upon protocol parameters such as the cryptographic algorithms to be used. Next, a Diffie-Hellman key exchange is performed to establish a shared secret over the untrusted channel. Following this, the client and the server authenticate each other through digital certificate verification. Finally, the two parties verify the integrity of the information exchanged in the above steps, to prevent man-in-the-middle attacks. At this point, a mutually authenticated confidential channel has been established between the client and the server, which can then be used in the second phase to exchange encrypted application data. We accelerate both phases of the DTLS protocol in hardware.

Figure 2.5.2 shows the system block diagram, the DTLS modes we support, and details of the computations required to implement these modes. Our system consists of a 3-stage RISC-V processor [2] supporting the RV32I instruction set, with a 16KB instruction cache and a 64KB data memory. An SD card is used as the backing store for larger programs. A memory-mapped DTLS engine (DE), comprised of a protocol controller, a dedicated 2KB RAM, and AES-128 GCM, SHA-256 and prime curve elliptic curve cryptography (ECC) primitives, accelerates the DTLS protocol. Sleep mode is implemented on the RISC-V, to save power, by gating its clock when cryptographic tasks are delegated to the DE. The DE uses a dedicated hardware interrupt to wake the processor on completion of these tasks. The DE is clocked by a software-controlled divider to decouple the processor operating frequency from the long critical paths in the ECC accelerator. In addition to full verification of the server certificate in step three of the handshake phase, the DE also supports caching of server certificate information to speed up future handshakes. This cached mode reduces an ECDSA-Verify operation to gain 1.56× savings in handshake energy.

Even in the cached mode, ECC computations, such as ECDHE and ECDSA, account for over 99% of the DTLS handshake energy. Fig. 2.5.3 describes an energy-efficient ECC accelerator that reduces this overhead. A pre-computation-based comb algorithm [3] is used for elliptic-curve scalar multiplication (ECSM), and a 4KB cache can store pre-computed comb data for up to six points, including generator points and public keys, thus reducing ECSM energy by 2.5× compared to a baseline implementation. A 256b wide interleaved reduction-based modular multiplier is implemented to support all Weierstrass and Montgomery curves over prime fields up to 256b, with higher bits of the data-path gated when working with smaller primes. The use of interleaved reduction allows us to handle arbitrary primes without any special structure, enabling support for NIST, SEC and ANSI curves. Resource-constrained ECC implementations [4,5] typically use projective coordinates to avoid modular inversion in the ECSM inner loop, at the cost of extra multiplications and a final expensive Fermat inversion. This work implements a dedicated 31k-gate modular inverter, allowing the use of affine coordinates, which saves 1.93× in ECSM energy by trading off the extra multiplications for cheaper Euclid inversions. Furthermore, a zero-less signed digit representation [3] of the scalar k is used to prevent simple power analysis side-channel attacks on the ECSM.

Figure 2.5.4 shows the detailed architecture of the DTLS engine and a comparison of resource utilization in three scenarios: DTLS fully implemented as RISC-V software (SW), the cryptographic kernels accelerated in hardware and only the DTLS controller implemented in software (SW+HW), and DTLS fully implemented in hardware (HW). Three blocks in the DE, that result in resource utilization improvement, are highlighted in Fig. 2.5.4. The use of cryptographic accelerators alone results in over 2 orders of magnitude improvement in run time and energy efficiency (SW vs. SW+HW). Similarly, the elimination of ECC code reduces instruction cache thrashing by 70×. Next, the DTLS controller implements a micro-coded state machine for packet framing, computation of the session transcript, parsing and validation of X.509 digital certificates and HMAC-DRBG-based pseudo-random number generation. This reduces code size by ~60KB and instruction cache misses by 60× (SW+HW vs. HW). Finally, the DTLS RAM implements a micro stack for storing temporary variables computed during the DTLS handshake. This 1.25KB micro stack results in 13KB reduction in data memory usage on the RISC-V (SW+HW vs. HW).

Figure 2.5.5 demonstrates the reconfigurability of the DE. The ECC primitive in the DE can accelerate all prime curves up to 256b. ECSM energies and run times as a function of prime bitwidth, using the secp160r1, secp192r1, secp224r1 and secp256r1 curves, are shown in Fig. 2.5.5. Security applications beyond DTLS can be implemented on the RISC-V, using the cryptographic accelerators in standalone mode. We illustrate this flexibility using three benchmark applications: (a) ECMQV, an alternative to ECDHE+ECDSA-based authenticated key exchange, (b) Schnorr Prover, an interactive zero-knowledge prover of identity, and (c) Merkle Hashing, used to ensure data integrity in peer-to-peer network protocols. The reduction in resource utilization for all three applications is shown in Fig. 2.5.5. The ECC-based applications experience over 200× increase in energy efficiency, while Merkle hashing sees 6× energy savings.

Figure 2.5.6 compares this work with embedded systems that integrate multiple cryptographic accelerators. This work implements a flexible ECC accelerator which supports arbitrary primes up to 256b, in contrast with [4] and [5] which only support fixed 192 and 255b curves respectively. [6] only supports binary-field modular multiplication in hardware. Our ECC accelerator is 458× and 9× more energy-efficient than [4] and [5], respectively, at comparable security levels. In addition to the resource savings enabled by the individual cryptographic accelerators, offloading DTLS control flow to the DE realizes a further 3× reduction in energy and 5× reduction in run time.

The chip was fabricated in a 65nm LP CMOS process, occupies 2.4mm^2 active area, and supports voltage scaling from 1.2V down to 0.8V. All measurements for the RISC-V and the DE are reported at 16MHz and 0.8V. The RISC-V processor occupies 34k NAND Gate Equivalents (GE), and achieves 0.96DMIPS/MHz, consuming 40.36µW/MHz. The DTLS engine occupies 149k GE, and uses 6.75KB of SRAM. The DE consumes 44.08µJ per DTLS handshake, and 0.89nJ per byte of application data. Therefore, through the design of reconfigurable energy-efficient cryptographic accelerators and a dedicated protocol controller, this work makes DTLS a practical solution for implementing end-to-end security on resource-constrained IoT devices.

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

References:
[1] E. Rescorla, et al., "Datagram Transport Layer Security Version 1.2," *IETF RFC*, vol. 6347, 2012.
[2] A. Waterman, et al., "The RISC-V Instruction Set Manual, Volume I: User-Level ISA, Version 2.0," *EECS Department, University of California, Berkeley*, Tech. Rep. UCB/EECS-2014-54, 2014.
[3] M. Hedabou, et al., "Countermeasures for Preventing Comb Method Against SCA Attacks," *ISPEC LNCS*, vol. 3439, pp. 85-96, 2005.
[4] M. Hutter, et al., "A Cryptographic Processor for Low-Resource Devices: Canning ECDSA and AES Like Sardines," *WISTP LNCS*, vol. 6633, pp. 144-159, 2011.
[5] M. Hutter, et al., "NaCl's crypto_box in Hardware," *Int. Workshop on CHES*, pp. 81-101, 2015.
[6] Y. Zhang, et al., "Recryptor: A Reconfigurable In-Memory Cryptographic Cortex-M0 Processor for IoT," *IEEE Symp. VLSI Circuits*, pp. C264-C265, 2017.

 978-1-5090-4940-0/18/$31.00

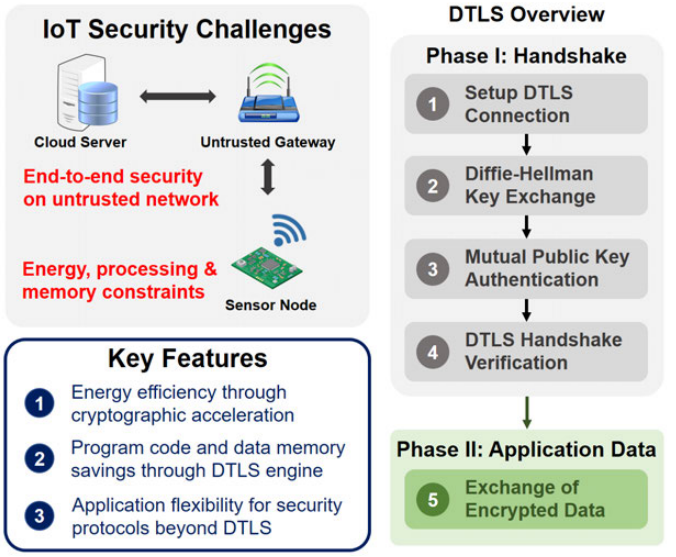


Figure 2.5.1: End-to-end security for IoT - challenges and proposed solutions.

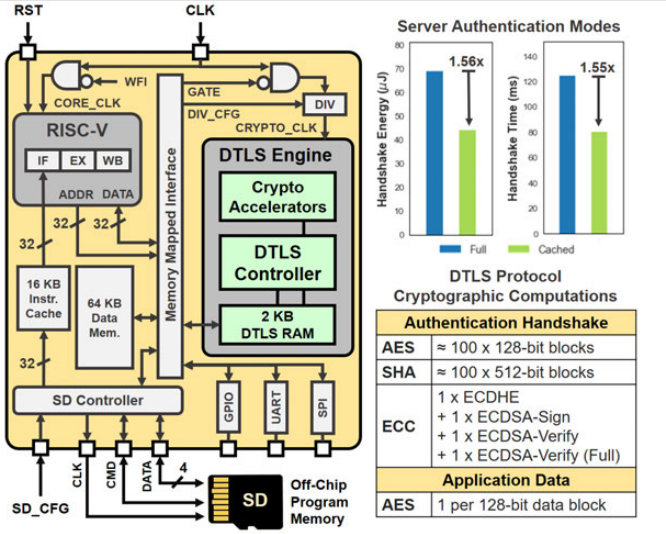


Figure 2.5.2: System diagram with supported server authentication modes and DTLS computation details.

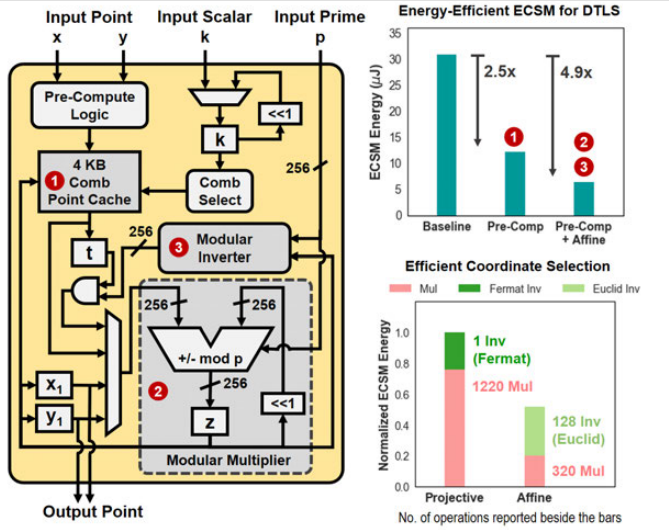


Figure 2.5.3: Proposed prime-field Elliptic Curve Cryptography accelerator and energy savings.

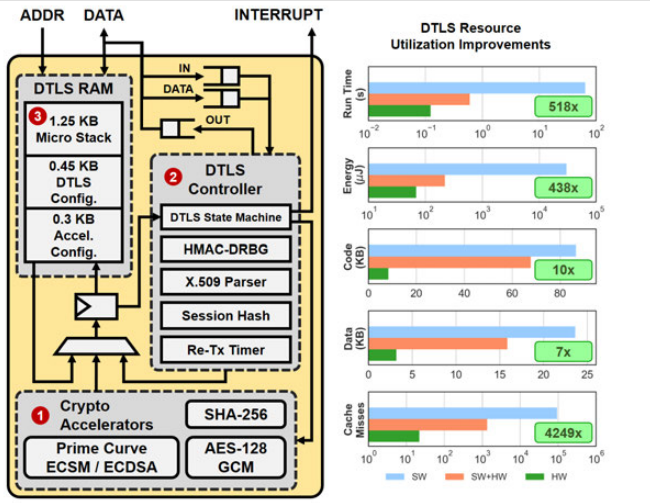


Figure 2.5.4: Architecture of the DTLS engine with resource utilization improvements over software-only and software-plus-hardware implementations.

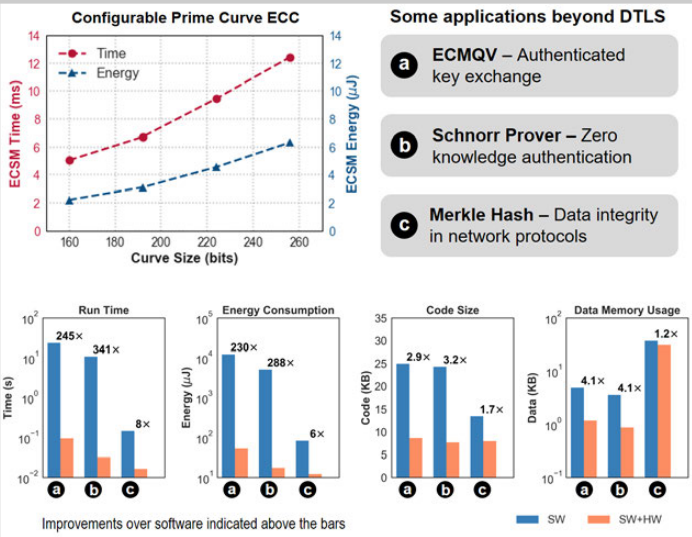
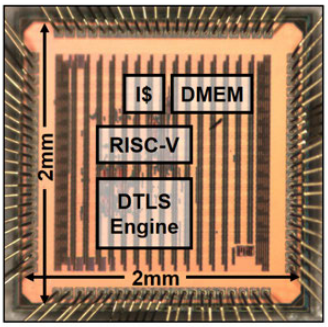


Figure 2.5.5: Flexibility of the DTLS engine with reconfigurable ECC and application benchmarks.

Specifications	WISTP'11 [4] [a]	CHES'15 [5] [a]	VLSIC'17 [6]	This work
Technology	350 nm	130 nm	40 nm	65 nm
Supply voltage (V)	3.3	1.2	0.7	0.8
Frequency (MHz)	0.847	1	28.8	16
App. Processor	–	–	ARM Cortex M0	RISC-V RV32I
Cryptographic Accelerator				
Logic gates	12.8k	32.6k	–	149k
SRAM	0.25 KB	0.28 KB	8 KB	6.75 KB
Hardware ECSM support	Only NIST P-192	Only Curve25519	– [b]	**All prime curves up to 256 bits**
Base point ECSM energy (μJ)	1423.6 (192 bit)	–	– [b]	**3.11** (192 bit)
	–	56.8 (255 bit)		**6.34** (256 bit)
AES energy (nJ)	8558.04	521.01 [c]	7.05	6.21
SHA energy (nJ)	6876.3 [d]	–	48.7 [d]	24.3 [d]
DTLS in hardware	No	No	No	**Yes**
DTLS energy	–	–	–	**44.08 μJ** (Handshake) **0.89 nJ/B** (App. Data)

[a] Post-synthesis data reported for [4] and [5]
[b] [6] implements only modular multiplication in binary fields in hardware
[c] [5] implements Salsa20 instead of AES for encryption
[d] [4] implements SHA-1; [6] implements SHA-3; This work implements SHA-2

Figure 2.5.6: Comparison with integrated cryptographic accelerators for embedded systems.



Chip Specifications	
Technology	65 nm LP CMOS
Supply voltage	0.8 – 1.2 V
Package	64-pin QFN
Die size	2 mm x 2 mm
Core size	1.54 mm x 1.54 mm
DTLS Engine	
Logic gates	149k (NAND2 equiv.)
SRAM	6.75 KB
Max. frequency	16 MHz at 0.8 V 20 MHz at 1.2 V
DTLS energy	44.08 µJ (Handshake) 0.89 nJ/B (App. Data)
RISC-V Processor	
Logic gates	34k (NAND2 equiv.)
SRAM	16 KB Instr. Cache 64 KB Data Mem.
Max. frequency	20 MHz at 0.8 V 78 MHz at 1.2 V
Dhrystone	0.96 DMIPS/MHz
Dhrystone Energy	40.36 µW/MHz at 0.8 V

Figure 2.5.7: Chip micrograph and performance summary.

978-1-5090-4940-0/18/$31.00

2.6 A 595pW 14pJ/Cycle Microcontroller with Dual-Mode Standard Cells and Self-Startup for Battery-Indifferent Distributed Sensing

Longyang Lin, Saurabh Jain, Massimo Alioto

National University of Singapore, Singapore, Singapore

Battery-indifferent sensor nodes require continuous operation in spite of the intermittently available battery energy, and hence require low peak-power operation to fit the fluctuating power made available by the harvester when the battery is out of energy (Fig. 2.6.1). Such harvested power can be very limited (e.g., nW and below) in aggressively miniaturized systems in the millimeter scale, and is typically well below the leakage consumption of the circuit being powered. Recently, purely harvested continuous operation with an on-chip harvester with sub-leakage sub-nW minimum power has been demonstrated for battery-less operation [1], at the cost of drastically lower performance (i.e., clock frequency in the Hz range) and larger energy. On the other hand, conventional miniaturized sensor nodes pursue minimum energy per operation to maximize the battery lifetime [2-6], but are not able to operate in the sub-leakage regime, and are hence unsuitable for purely harvested operation. Hence, existing solutions cannot interchangeably operate in purely harvested and battery-powered mode, as their design targeting minimum power (energy) severely degrades performance and energy efficiency (peak power consumption).

In this work, a dual-mode architecture comprising a microcontroller and a power management module is presented, which can operate both in normal (NM) and leakage suppression mode (LSM). NM and LSM modes, respectively, allow minimum-energy and minimum-power sub-leakage configurations, as required by battery-powered and purely harvested operation. Standard cells are configured as conventional CMOS gates in NM mode, whereas they are configured as dynamic leakage-suppression (DLS) logic [1] in LSM mode, so that their current is pushed below leakage. A self-startup scheme is introduced to enable cold start at reduced harvested power, overcoming the need for a large harvested power peak in [1] at start-up.

The proposed dual-mode architecture in Fig. 2.6.2 consists of a microcontroller (MCU) and a power management sub-system (PM). The microcontroller is an MSP430-compatible core, has 1KB latch-based instruction and data memory (each having 4 separately power-gated sections), 128B synthesized boot ROM, an on-chip clock generator, and a GPIO interface for communication with sensors and other peripherals. The MCU is designed with a dual-mode standard cell library (Fig. 2.6.3), which is configured in either NM or LSM mode depending on the *mode* signal generated by the configuration block, based on the battery condition (Fig. 2.6.2). The power controller includes a ripple self-startup circuit, a DC-DC converter, and a power-mode configuration block (the latter two are off-chip to allow ample testing flexibility in the various configurations). The latter block generates the signals to turn on/off the self-startup circuit, the DC-DC converter, and the power gating signals. When the battery is available, the system operates in normal mode with high energy efficiency and performance. When the battery is out of energy and under limited harvested power (e.g., solar cell at dim light), the system is configured in LSM mode to operate at sub-leakage power.

Figure 2.6.3 shows the operation of the proposed dual-mode standard cells, where four extra transistors (M1, M2, M5, M6) are added to the conventional CMOS gate (inverter gate, in this example). When *mode* = 0 (i.e., NM mode), M1 (M5) is turned on and boosted by ΔV to pull up node *n1* to V_{DD} (pull down *n2* to ground), which disables the feedback paths from transistors M2 and M6, and allows conventional CMOS gate operation. Voltage boosting by ΔV=0.4V in NM mode is sufficient to compensate the threshold voltage drop of M1 and M5, and is delivered by the DC-DC converter (powered by the battery in this mode). When *mode* = 1 (i.e., LSM mode), M1 and M5 are off and the inverter operates as DLS logic [1], assuring minimum (sub-leakage) power thanks to the reverse gate biasing of M1-M6 (i.e., super-cutoff). In a 35-stage ring oscillator (Fig. 2.6.3), LSM mode shows 750× power reduction at 0.4V compared to NM mode, while the latter exhibits 3,800× speed-up and 5× energy reduction. Based on this principle, a dual-mode standard cell library is designed.

In Fig. 2.6.4, measurements on the dual-mode microcontroller designed with the above library and an automated design flow show that the minimum-power point in LSM mode occurs at 0.45V and is 595pW, which is 198× lower than the minimum power point in normal mode. When running a moving average program, the minimum energy point in NM mode is 33pJ/cycle at 0.45V with fully enabled memory banks (14pJ/cycle when using 512B memory, with other banks being power gated), which is 8.2× (19.4×) lower energy than in LSM mode. At the same minimum energy point, the MCU runs at 19kHz, which is 7,755× faster than in LSM mode. From Fig. 2.6.4, the dual-mode reconfiguration breaks the tradeoff between minimum-power and minimum-energy encountered in conventional single-mode systems. Compared to a conventional CMOS design, LSM mode reduces power down to the sub-nW range like DLS logic [1]. Conversely, NM mode avoids the drastic speed (7,755×) and energy (8.2×) degradation of DLS.

Although operation in LSM mode reduces the current drawn by the MCU to the nA range once bootstrapped, the DC current absorbed when the harvester voltage is progressively raised is much larger, as was observed in [1] for DLS logic. For example, in Fig. 2.6.4, the current (power) at V_{DD}=0.2V in LSM mode is 17.1× (7.6×) larger than the value at the minimum-power point V_{DD}=0.45V, because transistors in DLS cells are less negatively gate biased at lower V_{DD}, and hence draw an exponentially larger current than at the minimum-power point [1]. This issue was addressed in [1] by requiring the harvester power to be significantly raised at power-up, and then allowed to be smaller during in-field operation. However, this limits the ability of the system to boot up again after a harvesting power outage, as it will not boot until a large harvested power becomes available again. To fundamentally solve this issue, a ripple power gating self-startup mechanism (Fig. 2.6.5) is introduced to allow cold start with limited harvested power. Instead of powering up the entire microcontroller all at once, the latter is partitioned in smaller power domains that are sequentially powered by the ripple self-startup block in Fig. 2.6.2, which progressively turns on the relevant header sleep transistors. The gate count in each power domain is small enough to keep its power-up peak current lower than the minimum targeted harvested power, as set by the application (e.g., low illuminance in a solar cell). The ripple power gating stage in Fig. 2.6.5 contains a hysteresis voltage detector that turns on the corresponding sleep transistor after a delay (tunable for testing purposes), when the harvester voltage reaches the 250mV trigger level during self-startup. Fig. 2.6.5 shows the measured waveform of signals progressively activating the power domains during self-startup, as powered by an on-chip 0.54mm^2 solar cell at 55lux illuminance (as typical of twilight). Without the proposed self-startup, more than 380lux would be needed for a safe start-up (bright office lighting).

Compared to prior art (Fig. 2.6.6), the proposed dual-mode architecture improves the minimum energy by 3.2×, and speed by five orders of magnitude compared with [1], while achieving an energy that is comparable to [2-6]. In sub-leakage operation, the dual-mode architecture offers more than 1,000× improvement in minimum power compared with [2-6], allowing the system to fully function at 55lux light intensity with a 0.54mm^2 on-chip solar cell.

Acknowledgements:
The authors thank Gopalakrishnan Ponnusamy for his kind help with testing, and acknowledge the kind support by TSMC and the Singaporean Ministry of Education grant MOE2014-T2-2-158.

References:
[1] W. Lim, et al., "Batteryless Sub-nW Cortex-M0+ Processor with Dynamic Leakage-Suppression Logic," *ISSCC,* pp. 146-147, 2015.
[2] 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.
[3] D. Bol, et al., "A 25MHz 7μW/MHz Ultra-Low-Voltage Microcontroller SoC in 65nm LP/GP CMOS for Low-Carbon Wireless Sensor Nodes," *ISSCC,* pp. 490-491, 2012.
[4] S. Paul, et al., "A Sub-cm^3 Energy-Harvesting Stacked Wireless Sensor Node Featuring a Near-Threshold Voltage IA-32 Microcontroller in 14-nm Tri-Gate CMOS for Always-ON Always-Sensing Applications," *IEEE JSSC*, vol. 52, no. 4, pp. 961-971, 2017.
[5] Y. Lee, et al., "A Modular 1mm^3 Die-Stacked Sensing Platform with Optical Communication and Multi-Modal Energy Harvesting," *ISSCC,* pp. 402-403, 2012.
[6] H. Reyserhove, et al., "A Differential Transmission Gate Design Flow for Minimum Energy Sub-10-pJ/Cycle ARM Cortex-M0 MCUs," *IEEE JSSC*, vol. 52, no. 7, pp. 1904-1914, 2017.

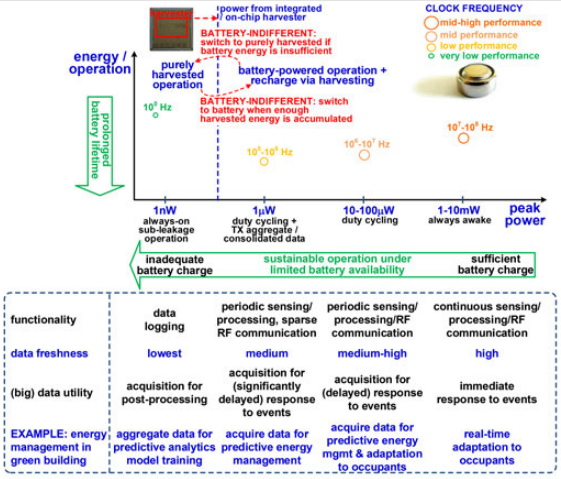


Figure 2.6.1: Battery-indifferent operation of sensor nodes with an integrated harvester need to achieve sub-nW power (minimum power) when the battery is out of energy, while reducing energy when battery-powered (minimum energy).

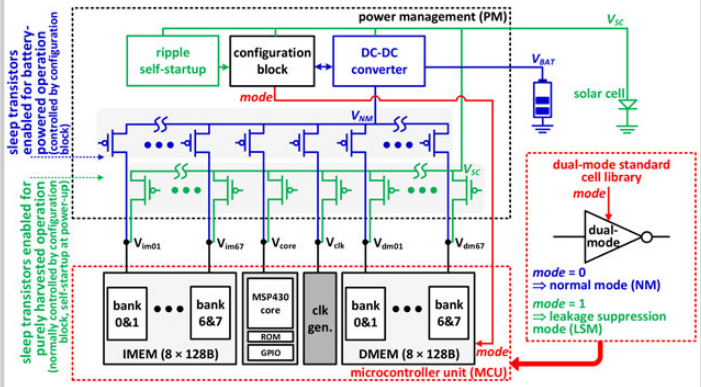


Figure 2.6.2: The proposed dual-mode system architecture comprises microcontroller (MCU) and power management (PM), and can operate in normal (minimum energy) or leakage suppression mode (minimum power).

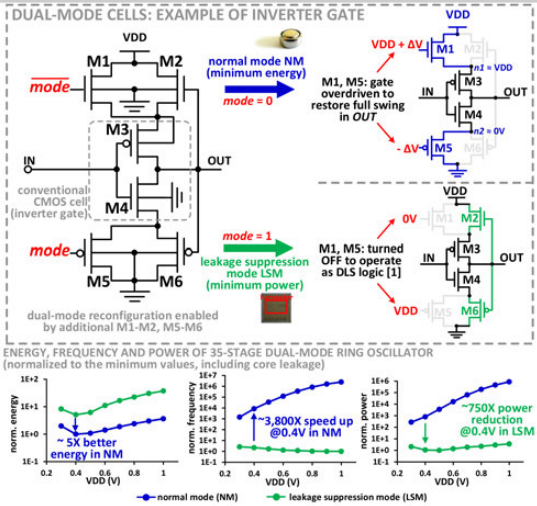


Figure 2.6.3: Dual-mode inverter and its operation in normal and leakage suppression mode (top), and measured energy/frequency/power of a dual-mode ring oscillator (bottom).

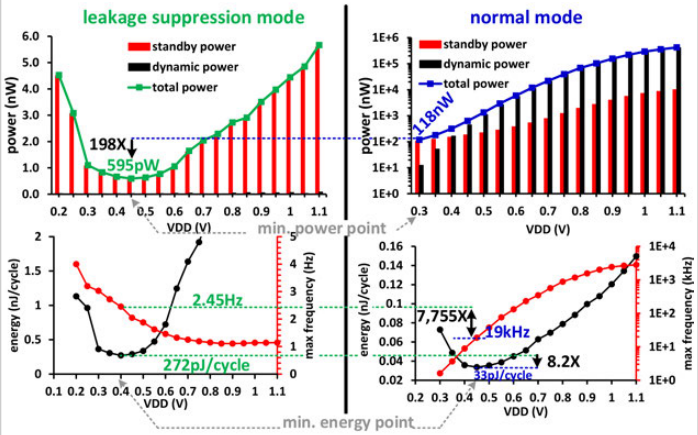


Figure 2.6.4: Measured power (top) and energy (bottom) of the microcontroller system in LSM (left) and NM (right), when running a program computing the moving average of an input acquired through the GPIO (entire 2KB memory is active, T=25°C).

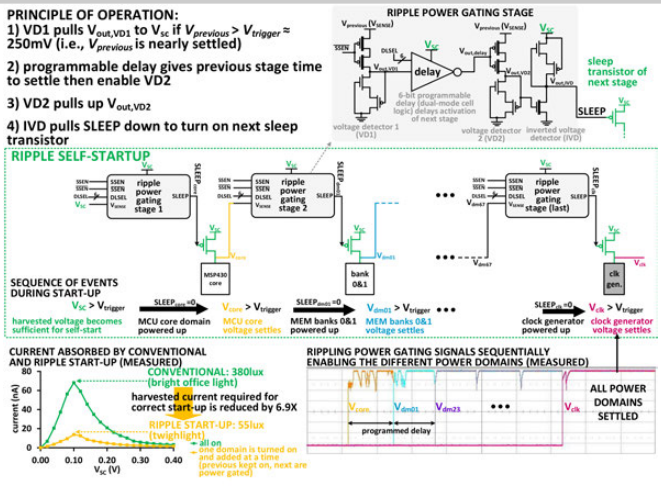


Figure 2.6.5: Block diagram of power management with ripple self-startup (mid), consisting of cascaded power gating stages (top) with sequential activation of power domains. Measured current and sequence of enable signals activating the ripple self-startup stages are shown on the bottom.

	technology	processor	area	normalized area (F^2)[1]	memory	VDD range	frequency range	minimum power	minimum energy
this work	180nm	MSP430 compatible	5.33mm²	1.65×10^8	2KB latch-based	0.2V - 1.1V	NM: 1.6kHz - 2.8MHz LSM: 1-4Hz	595pW (LSM) @0.45V, 2Hz	14-33pJ/cycle @0.45V (NM, depending on active MEM banks)
[1] ISSCC'15	180nm	ARM Cortex M0+	2.04mm²	6.3×10^7	256B latch-based	0.16V - 1.15V	2Hz - 15Hz	295pW @0.55V, 2Hz	44.7pJ/inst @0.55V
[2] ISSCC'15	65nm	ARM Cortex M0+	1.28mm²	3.03×10^8	8KB-ULV, 16KB	0.25V - 1.2V	29kHz - 66MHz	1µW @0.25V, 25kHz[2]	11.7pJ/cycle @0.39V
[3] ISSCC'12	65nm	MSP430 compatible	0.42mm²	9.9×10^7	18KB	0.32V - 1.2V	8MHz - 71MHz	20µW @0.32V, 8MHz[2]	7pJ/cycle @0.375V
[4] JSSC'17	14nm	x86 IA	0.79mm²	4.03×10^8	96KB	0.308V - 1.0V	0.5MHz - 297MHz	21µW @0.308V, 0.5MHz	17.18pJ/cycle@0.37V
[5] ISSCC'12	180nm	ARM Cortex M0	1.7mm²	5.3×10^7	3KB	0.6V	160kHz - 330kHz	3.3µW @0.6V, 1kHz	17.2pJ/inst @0.26V
[6] JSSC'17	40nm	ARM Cortex M0	0.39mm²	2.4×10^8	64KB	0.2V - 0.5V	0.8MHz - 50MHz	143.4µW @0.2V, 0.8MHz	43.22pJ/cycle@0.37V

[1] F = technology minimum feature size
[2] estimated from the reported power plot

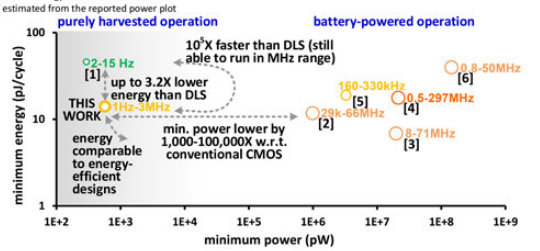


Figure 2.6.6: Performance summary and comparison.

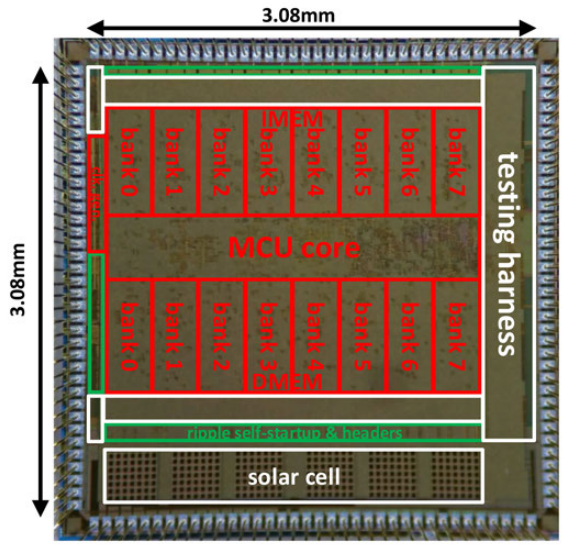


Figure 2.6.7: Die micrograph.

978-1-5090-4940-0/18/$31.00

2.7 A cm-Scale Self-Powered Intelligent and Secure IoT Edge Mote Featuring an Ultra-Low-Power SoC in 14nm Tri-Gate CMOS

Tanay Karnik[1], Dileep Kurian[2], Paolo Aseron[1], Richard Dorrance[1], Erkan Alpman[1], Angela Nicoara[1], Roman Popov[1], Leonid Azarenkov[1], Mikhail Moiseev[1], Li Zhao[1], Santosh Ghosh[1], Rafael Misoczki[1], Ankit Gupta[2], Akhila M[2], Sriram Muthukumar[2], Saurabh Bhandari[2], Yada Satish[2], Kartik Jain[2], Robert Flory[1], Chanitnan Kanthapanit[1], Eduardo Quijano[3], Bradley Jackson[1], Hao Luo[4], Suhwan Kim[1], Vaibhav Vaidya[1], Adel Elsherbini[5], Renzhi Liu[1], Farhana Sheikh[1], Omesh Tickoo[1], Ilya Klotchkov[1], Manoj Sastry[1], Sheldon Sun[1], Mukesh Bhartiya[2], Anuradha Srinivasan[1], Yatin Hoskote[6], Hong Wang[4], Vivek De[1]

[1]Intel, Hillsboro, OR; [2]Intel, Bangalore, India; [3]Intel, Guadalajara, Mexico
[4]Intel, Santa Clara, CA; [5]Intel, Chandler, AZ; [6]ARM, Austin, TX

Energy efficiency, performance and security of compact, self-powered, smart, secure and connected motes at the edge of IoT are critical for realizing intelligent, robust and sustainable end-to-end cyberphysical systems that deliver compelling new capabilities based on big data analytics. Integrated ultra-low-power compute [1] and wireless connectivity [2], neural-network-based inference accelerators [3], energy-efficient and compact multi-sensing front end and crypto engines, high density and low leakage embedded memory, and fine-grain power management are essential ingredients of the mote SoC.

We present a complete "edge-gateway-cloud" IoT system prototype for an example application that highlights the key advanced capabilities of the cm-scale, self-powered, intelligent and secure mote hardware platform at the edge, featuring a 2.5mm×2.5mm, 12M transistors ultra-low-power SoC in 14nm tri-gate CMOS (Figs. 2.7.1 and 2.7.7). The mote SoC integrates (Figs. 2.7.1 and 2.7.2): (i) a sub-mW 307K-transistor x86 host application processor operating across a wide voltage/frequency (V/F) range including near-threshold-voltage (NTV) for maximum energy efficiency; (ii) an efficient 600K-transistor convolutional neural network (CNN) accelerator, as part of a CNN-based 1.04M-transistor visual recognition and classification engine; (iii) a 70K-transistor lightweight crypto engine for secure boot and protected transport/storage of sensor/image data; (iv) a digitally reconfigurable adaptive multi-sensing analog front end (AFE); (v) a sub-mW always-ON wake-up radio receiver with a 278K-transistor digital baseband for remote SoC activation/sleep control from the gateway; (vi) 512KB high-density and low-leakage shared embedded memory; and (vii) an embedded power management unit (PMU) for smart and fine-grain power control of the SoC subsystems. The cm-scale mote hardware platform also contains solar cells, a photovoltaic (PV) harvester and a master PMU controller, a rechargeable solid-state electrolyte battery, voltage regulators, storage flash, a BLE radio, a CMOS camera, and multiple environmental sensors.

In the example application, these motes are deployed with adhesive trap papers across large agricultural fields. They wake up periodically during daytime to capture image and environmental sensor data. Each mote locally identifies and counts the number of moths captured, and transmits the information securely to a gateway drone via BLE for cloud analytics that drive optimal pesticide application schedules. Sustainable energy-autonomous operation during the lifetime of the mote is enabled by solar energy harvesting.

The mote SoC (Fig. 2.7.2) has 4 independent supply voltage domains: 0.55V for logic, ROM and register file (RF) arrays; 0.8V for a dense SRAM array; 1.05V for radio; and, 1.7V for IO transceivers. Except for the always-ON subsystem, comprising the wake-up radio, PMU and AON IO, all other subsystems – compute and crypto, memory, visual and peripheral – can be power gated individually by the PMU to minimize energy consumption in active, idle and sleep states. The entire SoC is turned OFF in the "sleep" state at night to conserve energy. The platform RTC timer and ambient light sensor are ON at night. The always-on subsystem is continuously ON during daytime. A wake interrupt can trigger various "active" states for sensor or image data capture, or image recognition and classification, or BLE transmission. The wake is qualified with a battery charge threshold. The unified tightly coupled memory (TCM) consists of 8KB ROM and 64KB RF arrays with single cycle access for fast boot and improved performance. Another shared 64KB RF array is used as temporary local data memory, and the 384KB shared SRAM array is used mainly for storing and processing image data. All arrays are gated OFF by default, and turned ON at first access to minimize leakage power. Individual SRAM banks support data-retention mode.

The image processing and blob detection engine in the visual subsystem (Fig. 2.7.3) first performs image segmentation to remove all background pixels using color-based thresholding, binarization and clustering. A histogram-based adaptive thresholding scheme is implemented where a grayscale conversion is done followed by image thresholding to segment the image. A mean-shift segmentation is used to identify possible moth candidates. The hue-based segmentation scheme lowers the number of blobs generated, thus reducing the amount of evaluations in the recognition engine. The detected blobs are fed into a pre-trained CNN classification accelerator for moth recognition. The CNN topology (Fig. 2.7.3) supports 8b resolution.

A lightweight crypto engine (Fig. 2.7.3) accelerates secure boot based on a quantum-compute-resistant public-key signature algorithm. It consists of a Keccak-400 Hash computation block with built-in lightweight DMA logic. A compact 64b authenticated encryption (PRINCE + CCM) engine with an all-digital TRNG preserves confidentiality and integrity of sensor/image data during local storage & transportation. An adaptive multi-sensing AFE (Fig. 2.7.3) with a 12b 1MS/s SAR ADC and digitally controlled reconfiguration enables near-simultaneous real-time capture of 5 environmental sensors' data.

The sub-mW always-ON Wake-up Radio (WuR) (Fig. 2.7.4) achieves a sensitivity of -72dBm @ 10^{-3} BER without high-Q external components. The mixer-first receiver's noise figure (NF) is dominated by the baseband, but improved by a switched-capacitor multiplier, where kTC noise is optimized for baseband power vs. NF trade-off. The three fully differential single-stage amplifiers with real output poles are used for baseband amplification and filtering. An additional buffer drives a low-power 6b SAR ADC to digitize the incoming data. The LO signal is generated with a resistively starved ring oscillator frequency-locked to an external 32kHz RTC. Demodulation of the On-Off Keying (OOK) WuR packet is implemented in the digital domain using adaptive threshold detection to enable flexible and efficient integration in a scaled process node. Acquisition and synchronization of the WuR packet is achieved by correlating and then folding the preamble back upon itself. The OOK demodulator calculates the ideal OOK demodulation threshold based on the received power of the WuR preamble. All of the building blocks for the 240×240 preamble correlator, preamble folding, and adaptive threshold calculation can be shared, resulting in an extremely compact and low latency design.

Mote SoC wake-up via the always-ON wake-up radio is demonstrated (Fig. 2.7.5). Measurements show that the SoC operates across a wide V/F range of 0.4V/200KHz to 0.85V/950MHz with 80µW-to-17mW power consumption. NTV operation for minimum energy consumption of 6.2pJ/cycle is achieved at 0.5V/100MHz. A sufficiently low BER is achieved for -71dBm WuR receiver input power. The total platform power during an image capture and processing is 16-24mW, with 1-7mW from the camera during capture, 14mW and 1-3mW in the SoC gated IO and gated logic blocks, respectively, and 100-200µW in the embedded SRAM.

We achieved successful operation of the system prototype including image capture, CNN-based visual recognition and classification and crypto for moths in an adhesive trap paper (Fig. 2.7.6). Effective utilization of the embedded shared memory enables 160ms image capture with segmentation and on-the-fly blob detection, 70ms blob extraction and CNN configuration, and 5ms classification per blob. Peak memory utilization is 380KB during CNN operation. The lightweight crypto accelerators enable 56ms secure boot, as well as low latency hash and encryption.

References:

[1] S. Paul, et al., "A Sub-Cm^3 Energy-Harvesting Stacked Wireless Sensor Node Featuring a Near-Threshold Voltage IA-32 Microcontroller in 14-Nm Tri-Gate CMOS For Always-On Always-Sensing Applications," *IEEE JSSC*, vol. 52, no. 4, pp. 961-971, 2017.

[2] N. Roberts, et al., "A 236nw -56.5dbm-Sensitivity Bluetooth Low-Energy Wakeup Receiver with Energy Harvesting in 65nm CMOS," *ISSCC,* pp. 450-451, 2016.

[3] 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.

 978-1-5090-4940-0/18/$31.00

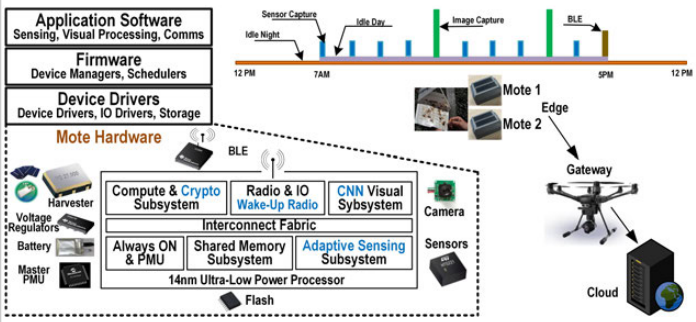


Figure 2.7.1: System architecture depicting software layers, details of the hardware layer with its subsystems and all peripheral components. Power consumption across a day cycle is included, as well as a picture of the actual moth trap with adhesive paper, final cm-scale motes communicating to the drone gateway and cloud.

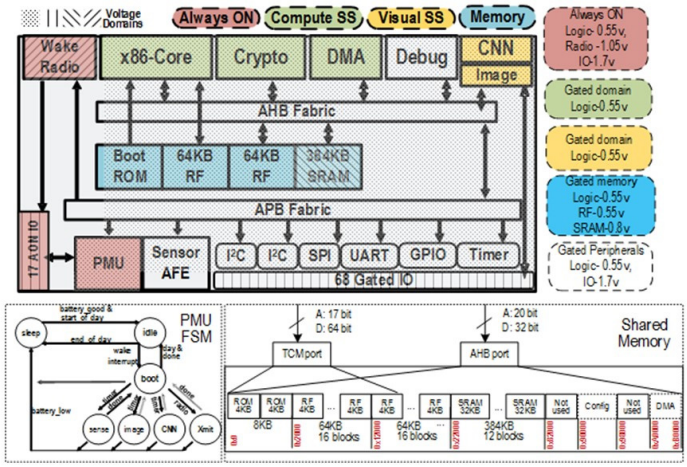


Figure 2.7.2: Microarchitecture details of the SoC with clearly identified voltage and power domains. Simplified PMU state transition diagram explains the implemented FSM. Shared memory map shows 1-cycle TCM access, 2-cycle AHB access and power-managed SRAM.

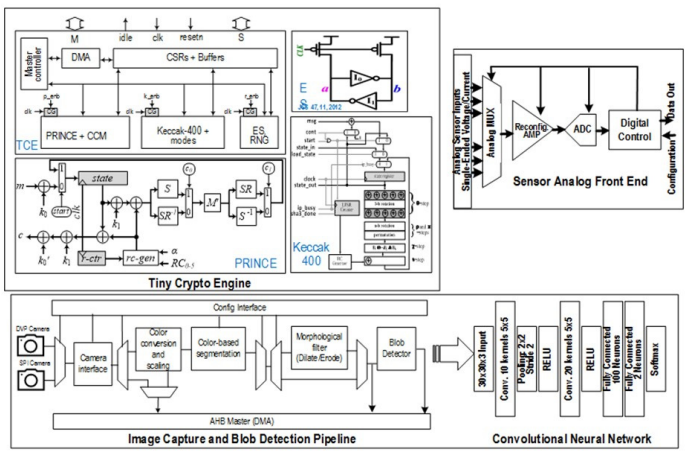


Figure 2.7.3: Internal details of the crypto engine and sensor analog front end. Image capture, blob detection and moth identification pipeline depicts the details of implemented CNN.

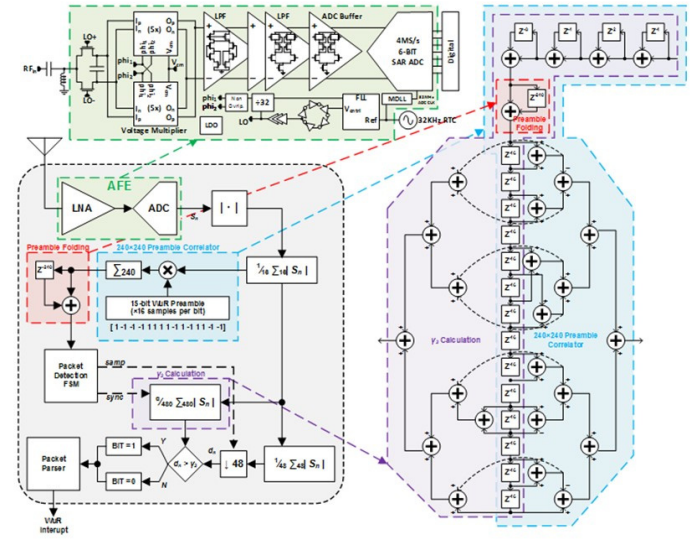


Figure 2.7.4: Ultra-low-power wakeup radio analog front end and detailed architecture of the radio baseband.

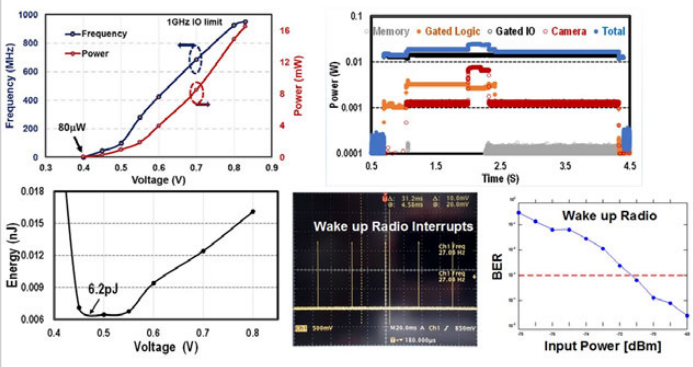


Figure 2.7.5: Power and frequency vs. voltage plot indicates wide dynamic range. Oscilloscope power was captured on 4 voltage rails. AON rails are very low and not shown. Energy versus voltage shows the 6.2pJ minimum energy point. Finally, repeated off-the-air wake-up radio interrupt captures and BER vs. input power of the radio is included.

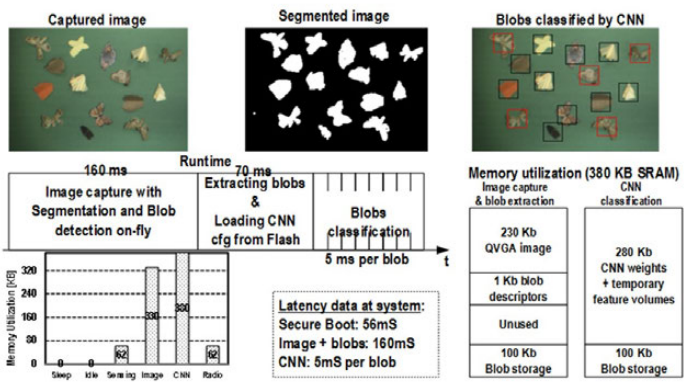
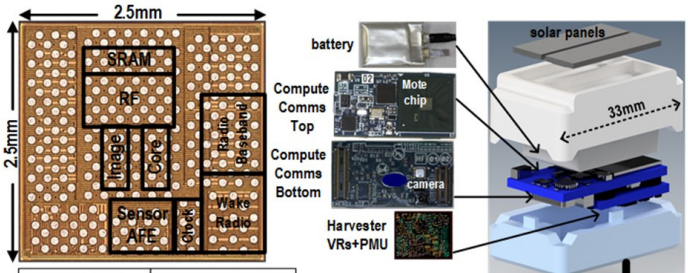


Figure 2.7.6: Sample image, blob segmentation and classification pictures demonstrating the functional mote. Runtime, system latency and memory utilization numbers are included to show the visual recognition with a modest amount of shared memory.



Technology	14nm Tri-Gate CMOS
Logic Transistors	8 million
Memory Transistors	3.6 million
Voltage	0.4 – 1.0V
Frequency Range	200KHz – 950MHz
Min System Power	0.2mW idle, 25mW peak

Block	Logic Gate Count
AFE Wrapper	24K
Image	440K
Wake Radio BB	278K
CNN	600K
Core	307K
Crypto	70K
PMU-Clock	50K
Rest	290K

Figure 2.7.7: Die shot shows individual major blocks. The actual tiny boards are shown to fit inside a 3D drawing of the form factor. It then points to the actual final 3D form factor mote. Technology and implementation details of the SoC are included.

978-1-5090-4940-0/18/$31.00

Session 3 Overview: *Analog Techniques*

ANALOG SUBCOMMITTEE

Session Chair:
Youngcheol Chae
Yonsei University, Seoul, Korea

Associate Chair:
Mahdi Kashmiri
Robert Bosch, Palo Alto, CA

Subcommittee Chair: ***Kofi Makinwa,*** *Delft University of Technology, Delft, The Netherlands*

Analog techniques continue to defy simple categories. This session illustrates the diversity and vigor of modern analog circuitry. Entries span the range of amplifiers, Class-D audio, references, programmable filters and oscillators. New frontiers of precision, power, and performance are established. The first paper describes a low-noise voltage buffer with 0.6pA input current and 0.6μV offset. The second and third papers describe sub-0.5V operation of a crystal oscillator and an RC oscillator with Allan deviation floor down to 250ppb. The next three papers expand the performance of Class-D audio amplifiers in terms of power, THD+N, and quiescent current. The last paper describes a programmable FIR filter operating at 3.25GS/s.

1:30 PM

3.1 A Quiet Digitally Assisted Auto-Zero-Stabilized Voltage Buffer with 0.6pA Input Current and 0.6μV Offset.

T. Rooijers, Delft University of Technology, Delft, The Netherlands

In Paper 3.1, Delft University of Technology presents an auto-zeroed stabilized voltage buffer with 0.6pA input current and 0.6μV offset. A digitally assisted offset-reduction scheme reduces its excess low-frequency (LF) noise while achieving a voltage noise of 29nV/√Hz.

2:00 PM

3.2 A Regulation-Free Sub-0.5V 16/24MHz Crystal Oscillator for Energy-Harvesting BLE Radios with 14.2nJ Startup Energy and 31.8μW Steady-State Power

K-M. Lei, University of Macau, Macau, China

In Paper 3.2, the University of Macau presents a sub-0.5V 16/24MHz crystal oscillator for energy-harvesting BLE radios with only 14.2nJ startup energy and 31.8μW steady-state power.

 978-1-5090-4940-0/18/$31.00

2:30 PM

3.3 A CMOS Dual-RC Frequency Reference with ±250ppm Inaccuracy from -45°C to 85°C

Ç. Gürleyük, Delft University of Technology, Delft, The Netherlands

In Paper 3.3, Delft University of Technology presents an RC-based frequency reference that achieves an inaccuracy of ±250ppm from -45°C to 85°C and an Allan Deviation floor of 250ppb.

3:15 PM

3.4 A 2×20W 0.0013% THD+N Class-D Audio Amplifier with Consistent Performance up to Maximum Power Level

E. Cope, Qualcomm, Tempe, AZ

In Paper 3.4, Qualcomm presents a 2×20W Class-D amplifier with a peak THD+N of 0.0013% and a 0.006% THD+N at its full power level. The feedback path only needs to process the error signal between the reference and output and thus the performance at full power level is enhanced by coefficient adjustment, lowering the loop order, and freezing the modulation index.

3:45 PM

3.5 A 0.0004% (-108dB) THD+N, 112dB-SNR, 3.15W Fully Differential Class-D Audio Amplifier with G_m Noise Cancellation and Negative Output-Common-Mode Injection Techniques

W-C. Wang, MediaTek, Hsinchu, Taiwan

In Paper 3.5, MediaTek presents a 3.15W Class-D amplifier achieving 0.0004% (-108dB) THD+N and 112dB SNR (A-weighted). Such high linearity is achieved with negative output-common-mode injection and G_m noise-cancellation techniques.

4:15 PM

3.6 A 0.96mA Quiescent Current, 0.0032% THD+N, 1.45W Class-D Audio Amplifier with Area-Efficient PWM-Residual-Aliasing Reduction

S-H. Chien, National Cheng Kung University, Tainan, Taiwan

In Paper 3.6, National Cheng Kong University presents a Class-D audio amplifier that proposes a PWM-residual-aliasing reduction technique, providing about 33% quiescent current reduction.

4:45 PM

3.7 A Low-Power 3.25GS/s 4th-Order Programmable Analog FIR Filter Using Split-CDAC Coefficient Multipliers for Wideband Analog Signal Processing

S. Park, Virginia Tech, Blacksburg, VA

In Paper 3.7, Virginia Tech presents a 3.25GS/sec 4th-order programmable FIR filter for wideband analog signal processing in 32nm SOI CMOS technology. Split CDACs are used to generate the programmable coefficient multipliers providing high linearity up to the Nyquist rate.

3.1 A Quiet Digitally Assisted Auto-Zero-Stabilized Voltage Buffer with 0.6pA Input Current and 0.6µV Offset

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 with both low offset and low input current. Chopper amplifiers can achieve low offset, but the switching of their input chopper gives rise to significant input current (40 to 110pA) [2-4]. Auto-zero (AZ) amplifiers require less input switching, but exhibit more voltage noise. However, ping-pong amplifiers continuously swap two auto-zeroed input stages, leading to more switching [5,7]. In this work, an AZ stabilized topology is proposed, in which a single amplifier is always present in the signal path. Only one input switch is required, resulting in an input current of 0.6pA (max), a 66× improvement on the state-of-the art [4]. Furthermore, a digitally assisted offset-reduction scheme reduces its low-frequency (LF) noise to the theoretical √5× limit. It also achieves a state-of-the-art maximum offset of 0.6µV.

The AZ stabilized amplifier is shown in Fig. 3.1.1. It consists of a 3-stage unity-gain buffer (BUF), whose offset (V_{os1}) and 1/*f* noise are cancelled by an AZ stabilization loop, which consists of an integrator (INT) and two OTAs: AZ1 and AZ3. During phase ϕ_2, AZ1 is auto-zeroed. Its input is shorted and the resulting output current is integrated on capacitors $C_{int21\text{-}int22}$ (10 pF each). AZ2 then converts the result into a current that cancels V_{os1}. The corresponding correction voltage V_{EF} is stored on the input capacitors $C_{21\text{-}22}$ (5pF each) of AZ2. During phase ϕ_1, AZ1 senses the buffer's offset, which, due to feedback, appears between its inputs. The output current of AZ1 is then integrated on capacitors $C_{int31\text{-}int32}$ (10pF each), thus generating, via AZ3, an appropriate cancellation current. The corresponding correction voltage V_{GH} is stored on the input capacitors $C_{31\text{-}32}$ (5pF each) of AZ3. To achieve µV-level offset, each stabilization loop should have >120dB gain. This is achieved by a single gain-boosted integrator amplifier (INT), which is multiplexed between the two loops.

The main drawback of auto-zeroing is increased LF noise due to the fold-back of wideband thermal noise. In the chosen topology, the LF voltage-noise spectral density is theoretically limited to $\sqrt{5}\cdot e_n$, where e_n is the thermal noise density of AZ1 and BUF (Fig. 3.1.2). This can be understood as follows. During one AZ cycle, AZ1 first auto-zeros itself before auto-zeroing BUF, thus contributing $\sqrt{2}\cdot e_n$ to the total LF noise. Adding the contribution of BUF, this leads to a total of $\sqrt{3}\cdot e_n$. However, after AZ1 auto-zeros itself, its sampled noise is stored by $C_{21\text{-}22}$ and so reaches the output in both AZ phases. This means that its contribution to the total LF noise is correlated, leading to a minimum LF noise density of $\sqrt{5}\cdot e_n$.

To reduce noise folding, the ratio between the AZ loop bandwidth (BW_{AZ}) and the auto-zeroing frequency f_{AZ} should be minimized. Reducing the former is preferable, since increasing f_{AZ} increases the rate of switching spikes, and hence, input current. This requires either large integration capacitors or a large ratio between the transconductances of AZ1 & BUF and AZ2 & AZ3, respectively. Simulations show that a ratio of about 500× is required to reach the √5× limit (Fig. 3.1.2). However, this limits the offset correction range of each loop to ~500µV. In order to handle the expected mV-range, a digitally-assisted coarse/fine AZ scheme is proposed.

The digitally-assisted local (through AZ2) and overall AZ loops (through AZ3) are shown in Fig. 3.1.3. A coarse (5-bit + Polarity-bit) current DAC (IDAC) is used in parallel with each fine analog loop. The IDAC state is controlled by a SAR, which is driven by a comparator that samples the integrator output. At startup, the coarse loop minimizes the integrator swing; BW_{AZ} is temporarily increased for fast settling after each bit trial. The SAR bits are then fixed, BW_{AZ} is decreased (via the end of conversion (EOC) bit) and the analog loop turned on to cancel the remaining offset and drift. With an IDAC LSB corresponding to 100µV of offset, this greatly relaxes the job of the analog loop.

Output spikes can occur at the transitions between the two operating phases of the AZ scheme. These are mainly due to the charge injection of the input switches of AZ1 ($SW_{4\text{-}6}$) and the finite time needed for the integrator's output to settle to the different voltages (V_{EF}, V_{GH}) required to cancel the offsets of AZ1 and BUF, respectively. The digitally-assisted AZ loop greatly relaxes the swing and settling time of the integrator, minimizing its contribution to the output spikes. To minimize spikes due to on-chip crosstalk, the AZ clock is applied to the chip as a differential current via low-impedance inputs. The transitions of the resulting differential voltage are then re-synchronized by on-chip current-steering SR-latches.

Most of the buffer's input current is due to the charge injection and clock feedthrough of SW_6. Being a transmission gate, this depends on the mismatch between its PMOS and NMOS switches, which, in turn, depends on the input voltage, and so cannot be completely cancelled. Capacitors $C_{11\text{-}12}$ (1 pF) minimize the common-mode transient (due to leakage) at the input of AZ1 during phase ϕ_2, since this also causes output spikes. However, due to the buffer's residual offset V_{os_res}, these capacitors are also a source of input current. During ϕ_1, C_{11} samples $V_{in} + V_{os\_res}$, while C_{12} samples Vin. The associated charge is then shared during ϕ_2, which means that C_{11} must be charged from $V_{in} + V_{os\_res}/2$ back to V_{in} during the next ϕ_1 phase. This results in an average input current $I_{in} = C_{11}V_{os\_res}f_{AZ}/2$. For V_{os_res} = 0.6µV, I_{in} = 5fA, which is quite negligible.

The buffer's extremely low input current is evaluated by reading out the voltage across an on-chip hold capacitor C_H (36pF). This can be set to an external voltage via a low-leakage sampling switch $SW_{1,2}$. To minimize the leakage of the sampling switch, extra hold capacitors C_T (3 pF) and, via SW_3, C_B (36 pF) ensure that the channel and body-diodes of SW_2 are operated at zero reverse bias. The same technique is applied to AZ1's input switches $SW_{4\text{-}6}$, as their leakage will otherwise discharge $C_{11\text{-}12}$ and cause CM transients. The S&H switches $SW_{1\text{-}3}$ are simultaneously closed to sample the external voltage and opened to start the hold phase. The voltage drift across C_H is then an accurate measure of the buffer's input current, avoiding the need for low-leakage bootstrapped ESD diodes.

The digitally-assisted AZ stabilized voltage buffer was realized in a 0.18µm CMOS process (Fig. 3.1.7). It has an active area of 0.55mm², 0.12mm² of which is taken by the S&H circuit and draws 210µA from a 1.8V supply. With a 1V input and f_{AZ} = 15kHz, measurements show that its input current is below 0.6pA (15 samples), and that its offset does not exceed 0.6µV (Fig. 3.1.4). In Fig. 3.1.5, the buffer's voltage noise spectral density is shown. Over BW_{AZ} a LF noise density of 29nV/√Hz is achieved, which equals the √5× noise limit. No tones at f_{AZ} can be seen, demonstrating the effectiveness of the spike reduction techniques. The voltage drift across C_H is also shown (typical sample, 1V input). With AZ off, there is negligible leakage, illustrating the effectiveness of the low-leakage techniques. With AZ on, the variation in input current over the buffer's input range (0.1 to 1.3 V) indicates that it is indeed mainly due to the charge injection of SW_6. In Fig. 3.1.6 the performance of the auto-zeroed voltage buffer is summarized and compared with the state-of-the-art. It achieves 66× less input current (0.6pA), as well as state-of-the-art offset (0.6µV) and competitive LF voltage noise (29nV/√Hz).

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] 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.

[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 20nV/√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.

 978-1-5090-4940-0/18/$31.00

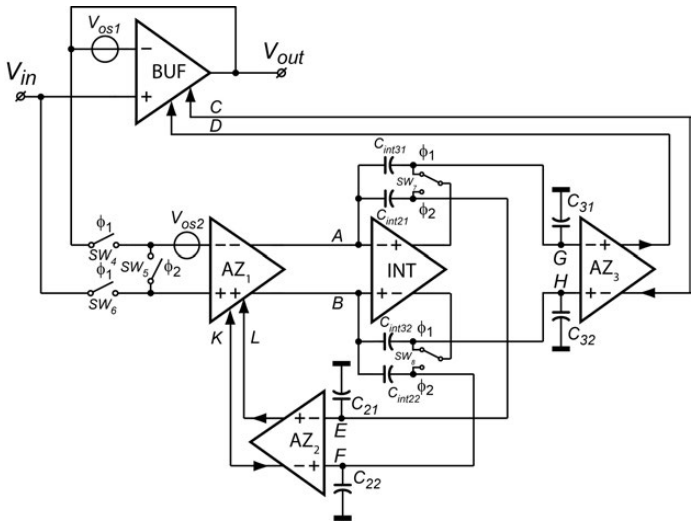


Figure 3.1.1: Block diagram of an auto-zero stabilized voltage buffer.

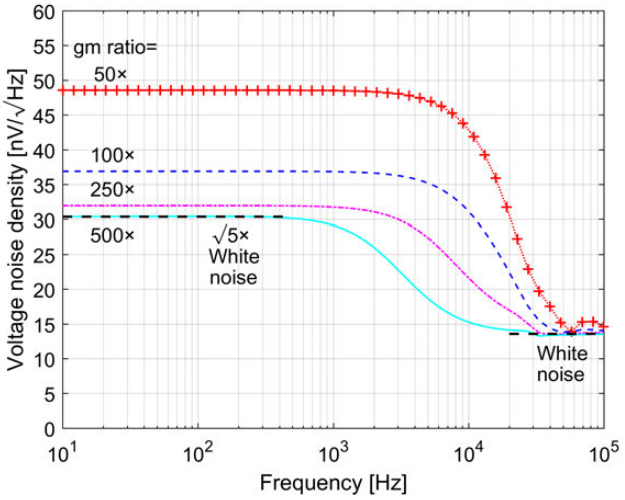


Figure 3.1.2: Simulated voltage noise density for different transconductance (g_m) ratios of AZ1 & BUF and AZ2 & AZ3, respectively (50×, 100×, 250× and 500×).

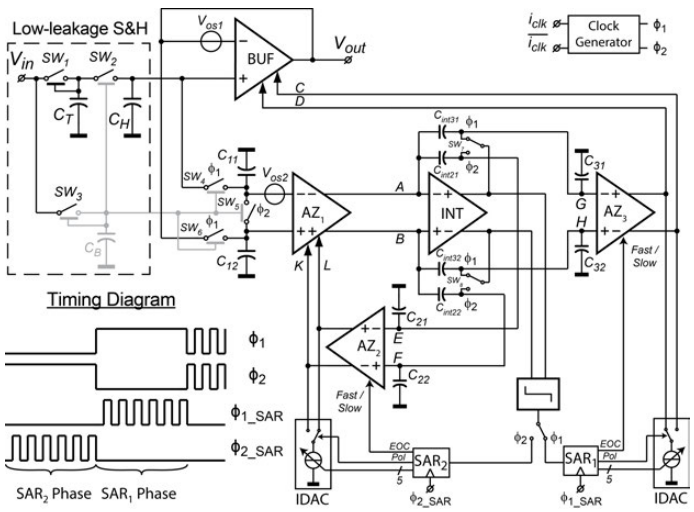


Figure 3.1.3: Block and timing diagram of the digitally-assisted auto-zero stabilized voltage buffer.

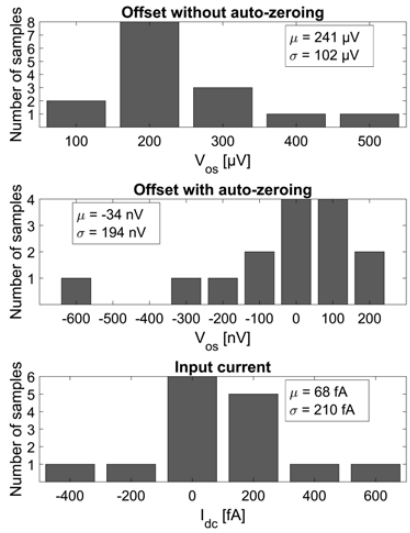


Figure 3.1.4: Histograms (15 samples) of the measured offset (without and with AZ) and input current (f_{AZ} = 15 kHz, V_{in} = 1V).

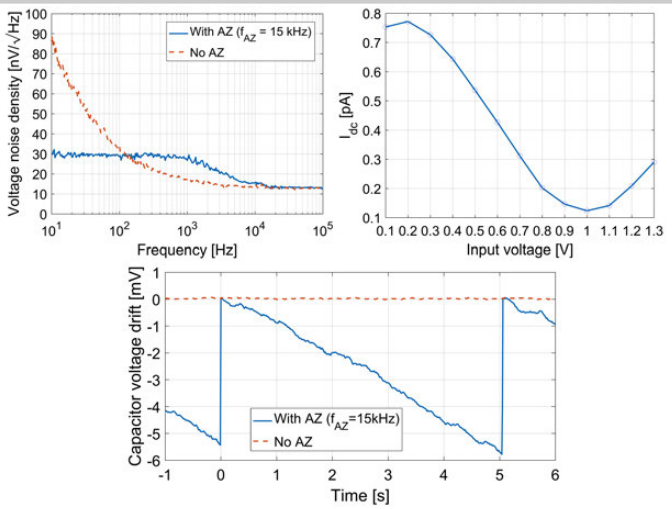


Figure 3.1.5: Measured voltage noise density: with and without AZ (Top left). The input current (I_{dc}) vs the input voltage for a typical sample (Top right). The capacitor voltage drift of a typical sample with and without AZ (Bottom).

	This work	[2]	[3]	[4]	[5]	[6]	[7]
Dynamic technique(s)	**Auto-zeroing**	Chopping	Chopping	Chopping and Auto-zeroing	Chopping and Auto-zeroing	Auto-zeroing	Chopping and Auto-zeroing
Input current (Max)	**0.6 pA**	110 pA	72 pA	40 pA	-	50 pA	-
Offset (Max)	**0.6 µV**	1 µV	0.78 µV	3 µV	2.8 µV	5 µV	4 µV
Voltage noise (nV/√Hz)	**29**	10.5	5.9	20	38 (AZ) 27 (CH&AZ)	75	140 (AZ) 28 (CH&AZ)
NEF	**7.4**	4.8	8.7*	21.8*	43.5*	-	-
GBW (MHz)	**1.45**	1.8	4	2.5	0.8	1	-
PSRR (dB)	**125**	120	142	-	138	130	128
Frequency (kHz)	**15**	30	200	15 / 7.5	28 / 14	4	11 / 7.33
Supply current	**210 µA**	143 µA	1.47 mA	800 µA	1.7 mA	750 µA	480 µA
Supply voltage	**1.8 V**	5 V	2.5 - 5.5 V	5 V	2.7 - 5.5 V	2.7 V	3.3 - 5.5 V
Die area (mm^2)	**1.4**	1.8	1.26	0.67	2.5	-	1.48

* Estimated value [2]

Figure 3.1.6: Performance summary and comparison with previous works.

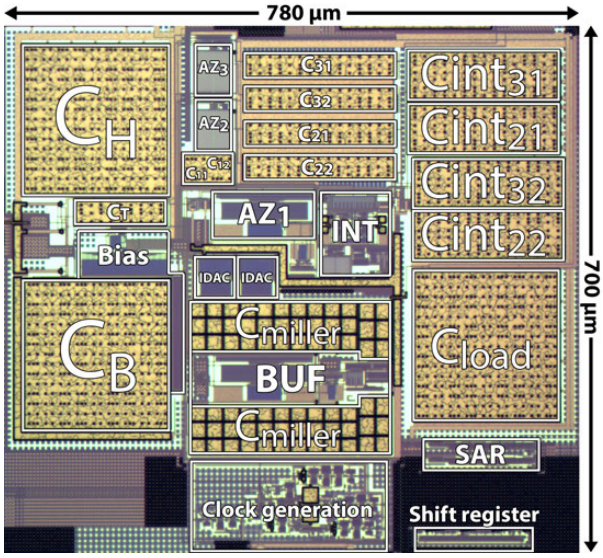


Figure 3.1.7: Die micrograph.

978-1-5090-4940-0/18/$31.00

3.2 A Regulation-Free Sub-0.5V 16/24MHz Crystal Oscillator for Energy-Harvesting BLE Radios with 14.2nJ Startup Energy and 31.8µW Steady-State Power

Ka-Meng Lei[1], Pui-In Mak[1], Man-Kay Law[1], Rui P. Martins[1,2]

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

This paper reports a regulation-free sub-0.5V crystal oscillator (XO) for Bluetooth Low-Energy (BLE) radios [1] that can be self-powered by harvesting the ambient energies, avoiding the loss and cost of the interim power converters. An ultra-low-voltage *Dual-Mode g_m Scheme* assisted by *Scalable Self-reference Chirp Injection (SSCI)* is proposed (Fig. 3.2.1) for the XO to surmount the operating challenges under an inconstant sub-0.5V V_{DD} (e.g. thermoelectric [2], ~80mV/K dependence) in both startup and steady states. Compatibility with different crystals (16/24MHz) is achieved, together with lower startup energy (14.2nJ) and steady-state power (31.8µW) than the recent art [3-5].

When an energy-limited BLE radio is duty-cycled to save the average power, both the startup energy (E_S) and time (t_S) of its XO are crucial to reduce the response latency and redundant power [3]. Herein, we present two circuit techniques that aid t_S reduction without momentarily raising the startup power, culminating in a lower E_S and relaxed power-source design.

Dual-Mode g_m Scheme: An XO with a 1-stage g_m (A_{XO-1}) is commonly used to optimize the phase noise (PN) [3-5]. As shown in Fig. 3.2.2 (upper-left), its impedance between the I/O (Z_{amp-1}) is given by

$$Z_{amp-1} = -\frac{g_m}{4\omega_0^2 C_L^2} + \frac{1}{j\omega_0 C_L} \ ,$$

where C_L is the crystal's load capacitance and $\omega_0 = 2\pi f_0$, where f_0 is the oscillation frequency. Since Z_{amp} is shunted by the crystal's stray capacitance (C_S), the negative resistance (R_N) of the overall impedance looking from the crystal core (Z_C) is

$$R_N \equiv -Re(Z_c) = -\frac{Re(Z_{amp})}{[\omega_0 C_S Re(Z_{amp})]^2 + [1-\omega_0 C_S Im(Z_{amp})]^2} \ .$$

A large R_N favors t_S reduction [4]. For A_{XO-1} where $Im(Z_{amp-1})$ is negative (capacitive), the maximum R_N is $C_L/[2\omega_0 C_S(C_L+C_S)]$ for $g_m = 4\omega_0 C_L(1+C_L/C_S)$, which is the upper limit if raising only g_m [4,5]. For instance, R_N is limited to 1.2kΩ with a C_S of 2pF, even an oversized g_m of 14.5mS could be applied, under typical f_0=24MHz and C_L=6pF.

To surmount the aforesaid R_N limit, a positive Im(Z_{amp}) is conceived to counteract the effect of C_S. Herein we propose a 3-stage g_m (A_{XO-3}) with designated capacitive loads (Z_{o1-2}) to mimic an inductor during the startup (Fig. 3.2.2, upper-right). Although a multistage g_m has been attempted in [6] to save the steady-state power, its inductive feature has not been revealed for t_S reduction. When Z_{amp-3} behaves inductively, Im(Z_{amp-3})>0 can be achieved over 13 to 46MHz as shown in the locus plot (Fig. 3.2.2, lower-left). For instance, given a small g_m of 2.3mS, R_N of A_{XO-3} is 2.4kΩ after paralleling with a C_S of 2pF (Fig. 3.2.2, lower-right), which is ~9× higher against that of A_{XO-1} with the same g_m. It is clear that A_{XO-3} is inferior to A_{XO-1} in terms of PN. Thus, when the crystal has gained sufficient energy during the startup, A_{XO-3} will be turned off (with an external control signal), leaving A_{XO-1} to sustain the oscillation.

Scalable Self-reference Chirp Injection (SSCI): Signal injection close to f_0 of the crystal is proven efficient and robust for t_s reduction [4]. The XO in [3] exhibits a slashed t_S (<400µs) by dithered-signal injection to the crystal, but entails trimming to handle the PVT variations on the injection oscillator. While chirp-modulated-signal injection [4] averts calibration, the related RC-sweeping unit is area hungry (~90% of the area) due to its large time constant. Also, PVT variations of the ring oscillator (RO) and RC elements hinder its utility. To address those pitfalls, we introduce SSCI to generate a chirping signal accelerating the startup of the XO (Fig. 3.2.3). Unlike the RC-based chirping [4], here a 5-stage uncalibrated RO is incorporated with a finite state machine (FSM) to digitally scale the injection time (t_{CI}), which is decided by the number of exciting cycles at each cap-bank value C_{OSC}. This scalability covers 10 to 38MHz for robustness, and renders the XO compatible with different crystals (i.e., L_M and optimum t_{CI} are package-dependent [4]). To maximize the injection energy (i.e., 50% duty cycle), the chirp-modulated signal is a div-by-2 output from the RO. It serves as the exciting signal for the crystal via an output driver, and trigger signal for the FSM. The FSM counts the pulses, and sequentially raises C_{OSC} by sending the control f_{ctrl} to RO. The RO is powered down by the FSM automatically after the injection.

For sub-0.5V operation, subthreshold common-source (CS) amplifiers with resistive loads are applied for both A_{XO-1} and A_{XO-3}. Unlike the current-source load [3,5], the resistive load aids to uphold a moderate g_m even when V_{DD}<0.35V, under a small bias current (simulated at I_{dc}=100µA). For instance, the simulated g_m of A_{XO-1} is 1.3mS at V_{DD}=0.3V and –40°C, being 4× higher than the current-source load (assume an identical g_m with V_{DD}=0.35V at 20°C). The resistive load has a tradeoff of f_0 variation, but is manageable for the BLE standard: <±50ppm.

A_{XO-3} is an ac-coupled 3-stage CS amplifier (Fig. 3.2.4) assisted by a constant-g_m bias circuit. The latter secures A_{XO-3} to be inductive and a stable R_N for robust-and-fast startup against PVT. Only the micro-current (<5µA) digital and constant-g_m bias circuits entail a 0.7V that can be generated by an on-chip charge pump as in [1]. As the constant-g_m bias circuit is off after startup, the power and noise overheads are negligible. Monte-Carlo-simulated R_N (mean) of A_{XO-3} is >9.1× higher than that of A_{XO-1}, and the boosting factor is immune to C_s from 1 to 3pF. For the RO of the SSCI, it is realized by a source-degenerated CS stage to reduce its frequency deviation against PVT, while enabling sub-0.5V operation.

The XO was fabricated in 65nm CMOS with on-chip C_L of 6pF. f_0 is flexible between 16 and 24MHz. Tested with a 24MHz crystal, t_S is 530µs if only the A_{XO-3} technique is enabled during the startup (Fig. 3.2.5). With both A_{XO-3} and SSCI enabled, t_S is further shortened to 400µs (3.3× reduction) and E_S is 14.2nJ (2.8× reduction), for a 90% oscillation amplitude [4]. Note that t_S faces nonlinear reduction with respect to the achieved R_N-boosting factor of 9.6 (A_{XO-3}'s R_N over A_{XO-1}'s R_N); since g_m of M_{1-3} (Fig. 3.2.4, right) deviate from their small-signal values when the oscillation swing is developing, resulting in an aggravated R_N. Also, the XO entails an overhead time to enter the steady-state after switching to A_{XO-1} (i.e., 240µs for the case of A_{XO-3} + SSCI). The A_{XO-3}-to-A_{XO-1} switching time can tolerate ±50% uncertainty for t_S variation <10%. The XO takes ~300µs to settle for a ±20ppm f_0 accuracy [5]. The steady-state power is 31.8µW at 0.35V and the PN is -134dBc/Hz at 1kHz offset adequate for the BLE standard.

For robustness, the XO upholds a steady-state output swing >80% of V_{DD} at 0.3 to 0.5V, and t_S variation is <25% from its mean (400µs). Only the RO of the SSCI fails to start when V_{DD} is down to 0.25V, but A_{XO-3} is still in place to aid t_S reduction. Over -40 to 90°C, t_S variation is <7.5%, being 4.7× less than [3]. The frequency deviation ($\Delta f_0/f_0$) is ≤19.7 and ≤14.1ppm, respectively, over such V_{DD} and temperature ranges.

This XO succeeds in conforming to the BLE standard with adequate margin for aging (Fig. 3.2.6). The achieved E_S and area efficiency are >2.6× and >3.1× better than [3-5], respectively. The experimental setup and chip micrograph are depicted in Fig. 3.2.7-left, where the contribution of E_S and t_s reduction by each technique is appended in Fig. 3.2.7-right. By combining two startup techniques (A_{XO-3} + SSCI), the startup energy E_S and time t_S are reduced by 2.8× (40 → 14.2nJ) and 3.3× (1.3 → 0.4ms), respectively. Consistent results are measured for different crystals (16/24MHz), demonstrating a regulation-free sub-0.5V BLE-compliant XO.

Acknowledgements:
The authors thank the Macau Science and Technology Development Fund (FDCT) - SKL Fund and University of Macau - MYRG2017-00223-AMSV for financial support.

References:
[1] W.-H. Yu, et al., "A 0.18V 382µW Bluetooth Low-Energy (BLE) Receiver with 1.33nW Sleep Power for Energy-Harvesting Applications in 28nm CMOS," *ISSCC*, pp. 414-415, Feb. 2017.
[2] Micropelt MPG-D655. Datasheet:
http://micropelt.com/downloads/datasheet_mpg_ d655.pdf.
[3] D. Griffith, et al., "A 24MHz Crystal Oscillator with Robust Fast Start-Up Using Dithered Injection," *ISSCC*, pp. 104-105, Feb. 2016.
[4] 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.
[5] 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.
[6] S. Iguchi, et al., "93% Power Reduction by Automatic Self-Power Gating (ASPG) and Multistage Inverter for Negative Resistance (MINR) in 0.7V, 9.2µW, 39MHz Crystal Oscillator," *IEEE Symp. VLSI Circuits,* pp. 142-143, June 2013.

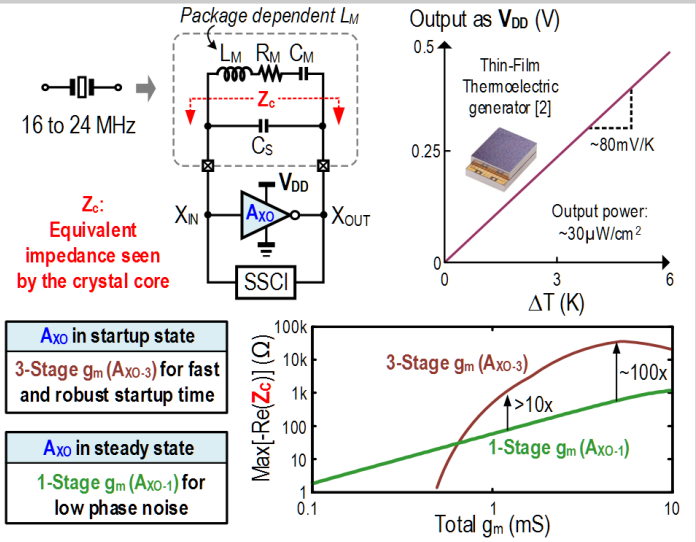


Figure 3.2.1: A 16/24MHz BLE-compliant crystal oscillator to operate from an unregulated sub-0.5V energy source, e.g. thermoelectric [2].

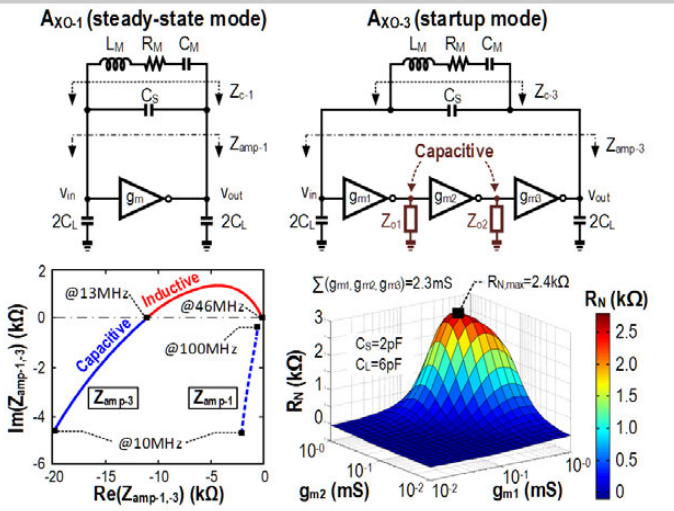
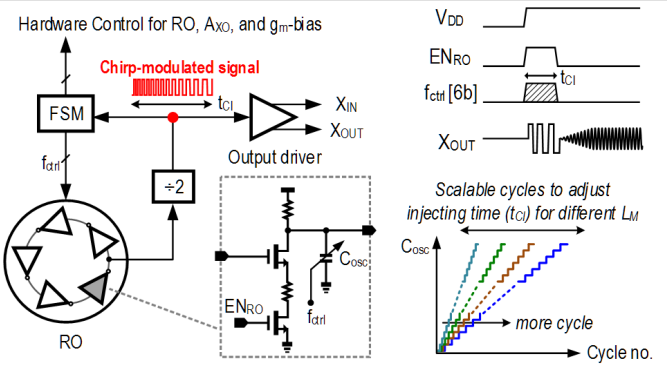


Figure 3.2.2: (Top) Schematics of crystal+A_{XO-1} for steady state and crystal+A_{XO-3} for startup. (Bottom) Simulated locus plot of $Z_{amp-1,-3}$, and A_{XO-3} shows a much higher $R_{N,max}$ at 24MHz.



Signal-injection techniques to kick-start the XO

	t_S reduction effectiveness	E_S reduction effectiveness	Trimming on auxiliary oscillator	Precision of auxiliary oscillator
Constant frequency	✓✓✓	✓✓✓	Required	Very Critical
Dithering	✓✓	✓✓	Required	Critical
Chirping	✓	✓	**Not Required**	**Relax**

Figure 3.2.3: SSCI generates a chirping signal to kick-start the XO using an untrimmed RO with relaxed precision. The FSM provides feasibility to scale t_{CI}, accommodating different crystal packages (i.e., L_M).

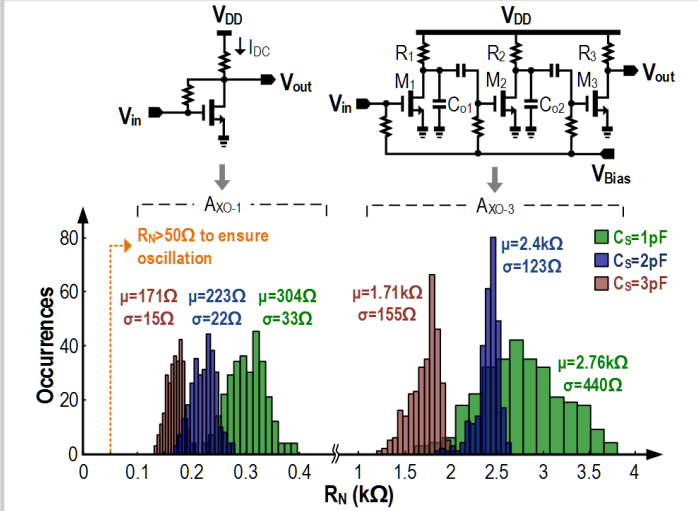


Figure 3.2.4: ULV 1-stage (left) and 3-stage (right) g_m implementation, and their Monte-Carlo-simulated R_N (N=300, V_{DD}=0.35V). For the 3-stage g_m, $C_{o1,2}$ renders Z_{amp-3} inductive.

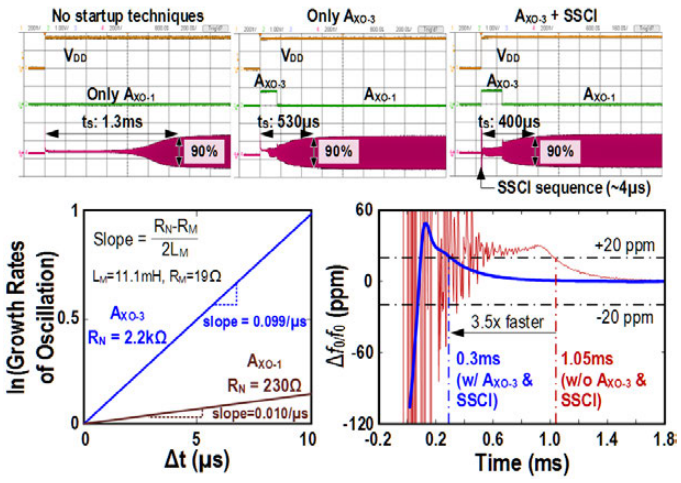


Figure 3.2.5: (Top) Measured startup times with and without proposed techniques. (Bottom, left) Estimated R_N from the exponential growth of X_{OUT}'s amplitude before the transistors enter triode region. (Bottom, right) Transient f_0 profiles of the XO (V_{DD}=0.35V, T=20°C).

		This work		JSSC'16 [4]	ISSCC'16 [3]		ISSCC'17 [5]
Applications		BLE		Bluetooth	BLE		BLE
Fast-startup techniques		ULV inductive 3-stage g_m + SSCI		Chirp injection + g_m-boosting	Dithered injection		Dynamic load + g_m-boosting
Steady-state techniques		ULV 1-stage g_m + resistive load		1-stage inverter	1-stage g_m + current-source load		
CMOS process (nm)		65		180	65		90
Active area (mm²)		**0.023**		0.12	0.08		0.072
Supply voltage, V_{DD} (V)		**0.35***		1.5	1.68		1.0
Temperature, T_{Range} (°C)		-40 to 90		-30 to 125	-40 to 90		-40 to 90
Load capacitance, C_L (pF)		6		6 (off-chip)	6	9	10
Frequency, f_0 (MHz)		**16**	**24**	39.25	24	24	24
Startup energy, E_S (nJ)		**15.8**	**14.2**	349	--	--	36.7
Startup time, t_S (μs)		460	400	158	64	435	200†
$\Delta t_S/t_S$ over T_{range}		9.8%	7.5%	7%	±35%	±20%	26.6%
XO inaccuracy $\Delta f_0/f_0$ (ppm)	versus T_{Range} (@ 0.35V)	< -14.7 / +7.2 # (-40 to 90°C)		±5.5 (-30 to 125°C)	N/A		N/A
	versus V_{DD} (@ 20°C)	< -13.0 / +4.9 # (0.3 to 0.5V)		±0.6 (1.2 to 1.8V)	N/A		N/A
Steady-state power (μW)		**31.6**	**31.8**	181	393	693	95

* Digital & constant-g_m bias circuits are at 0.7V (current budget: 5μA) to be generated by an on-chip charge pump as [1]. † Amplitude >90% and $\Delta f_0/f_0$ <±20ppm. # worse values from 16/24MHz crystals, the BLE spec. is <±50ppm.

Figure 3.2.6: Performance summary and comparison with the recent art [3-5].

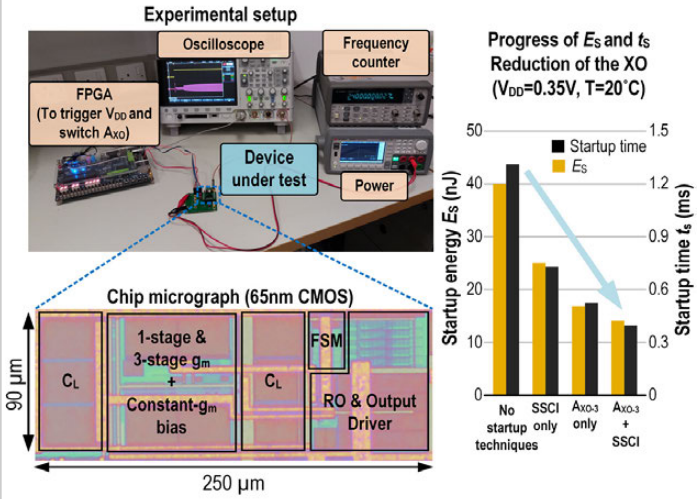


Figure 3.2.7: (Left) Experimental setup and die micrograph in 65nm CMOS. (Right) By A_{XO-3} + SSCI techniques, the startup energy E_S and time t_S are reduced by 2.8× and 3.3×, respectively.

978-1-5090-4940-0/18/$31.00

3.3 A CMOS Dual-RC Frequency Reference with ±250ppm Inaccuracy from -45°C to 85°C

Çağrı Gürleyük, Lorenzo Pedalà, Fabio Sebastiano, Kofi A. A. Makinwa

Delft University of Technology, Delft, The Netherlands

To comply with wired communication standards such as USB, SATA and PCI/PCI-E, systems-on-chip require frequency references with better than 300ppm accuracy. LC-based references achieve 100ppm accuracy [1], but suffer from high power consumption (~20mW). Thermal diffusivity (TD) references require less power (~2mW), at the expense of less accuracy (1000ppm) [2]. RC-based references offer the lowest power consumption, but their accuracy is typically limited to ~0.1% [3]. In RC relaxation oscillators, comparator offset and delay are the major sources of inaccuracy [4,5]. References based on frequency-locked loops (FLLs) circumvent these by locking an oscillator's frequency to the time-constant of an RC filter, but their accuracy is then limited by the nonlinear temperature dependency of on-chip resistors [3,6].

This paper describes a 7MHz RC-based frequency reference that solves this problem by accurately combining the complementary temperature dependencies of two integrated resistors in the digital domain. Measurements on 12 samples show that it achieves an inaccuracy of ±250ppm from -45°C to 85°C and an Allan Deviation floor of 250ppb. These results represent a 3.5× reduction in inaccuracy and a 16× reduction in long-term drift compared to state-of-the-art CMOS RC-based frequency references.

The proposed frequency reference (Fig. 3.3.1) consists of a frequency-locked loop (FLL), in which the frequency of a digitally controlled oscillator (DCO) is locked to a composite phase shift derived from two Wien-Bridge (WB) filters. Due to the finite temperature coefficients (TCs) of their resistors, the phase shift of each WB will be a function of temperature. Phase shifts with complementary TCs can then be generated by realizing WBs from different resistor types, e.g. silicided p-poly (TC = 0.36%/°C) and unsilicided n-poly (TC = -0.17%/°C). By digitizing and appropriately combining the complementary phase shifts of the two such WBs, ϕ_{sp} and ϕ_n, a temperature *independent* phase-shift, ϕ_e, can be realized.

As shown in Fig. 3.3.1, the phase shift of each WB is digitized by a Phase Domain ΔΣ-Modulator (PDΔΣM). The resulting bitstream output is then decimated by a CIC (cascaded integrator-comb) filter. The nonlinear temperature dependence of the resulting ϕ_{sp} and ϕ_n is first corrected by fixed polynomials [$p_{sp}(.)$ and $p_n(.)$], and then combined to generate ϕ_e. The gain provided by the FLL's digital integrator drives ϕ_e to zero, thus making the DCO's output frequency f_{DCO} (=7MHz) temperature independent as well.

Figure 3.3.2 shows the block diagram of a WB and its 2nd-order PDΔΣM [8]. Both are driven at $f_{DRV} = f_{DCO}/16$, which is also used to generate the modulator's sampling clock f_s and its phase references, ϕ_0 and ϕ_1. A chopper demodulator detects the difference between the WB's phase shift, ϕ_{WB}, and the phase reference selected by the bitstream, BS. The resulting DC signal is driven to zero by the loop filter's gain, ensuring that the average value of the selected phase references is in quadrature with ϕ_{WB}, and therefore, that BS is a digital representation of ϕ_{WB}. As in [8], the 1st integrator is based on a 2-stage opamp, while the 2nd integrator employs a g_m-C OTA, which uses R_{ff} to realize the modulator's feed-forward coefficient. Figure 3.3.2 also shows the measured output spectrum of the PDΔΣM that digitizes WB_{sp} when it is driven by a fixed 7MHz clock and the CIC filter output. The decimation factor of the CIC filter involves a trade-off between modulator resolution (smaller bandwidth) and suppression of DCO drift and noise (wider bandwidth). A decimation factor of 1024 places the filter's first notch at ~425Hz, which ensures sufficient suppression of quantization noise (highlighted area in Fig. 3.3.2). After decimation, the PDΔΣM+CIC combination achieves a phase resolution of ~0.025m° (rms), which translates into negligible DCO jitter: <0.5ps (rms). A digital gain following the integrator sets the dominant pole of the entire FLL at ~50Hz.

Figure 3.3.3 shows the circuit diagram of the DCO. It consists of a 9-stage current-starved ring oscillator, which is driven by a 5b coarse current-steering DAC, and a 13b fine current-output R-2R DAC. The coarse DAC covers a ±50% range around the 7MHz nominal output frequency, while the fine DAC covers a ±7.5% range with a 120Hz LSB. The FLL loop primarily controls the fine DAC using linear feedback, but the coarse DAC can be updated when the digital integrator is close to saturation. To ensure feedback stability, the fine DAC must be monotonic, and so a segmented architecture based on a 5b unary DAC and an 8b R-2R ladder was used. Its reference is generated from the supply voltage via a resistive divider, and then applied to the DAC via a buffer and a gain-boosted current mirror (g_m and M_1). An RC lowpass filter (R_{lpf}-C_{lpf}), with a cut-off much higher than the FLL pole, suppresses the DAC's wide-band noise, this reducing the DCO's jitter. This coarse-fine architecture results in a fine LSB small enough (18ppm) to keep the DCO's quantization noise well below the expected accuracy, while achieving a large enough range to handle process variations.

The prototype (Fig. 3.3.7) was fabricated in a TSMC 0.18µm CMOS process. 12 samples in ceramic DIL packages were characterized. The two WB and PDΔΣM channels occupy 1.24mm² and draw 180µA from a 1.8V supply. The DCO occupies 0.35mm² and draws 250µA from a separate 1.8V supply. Digital circuitry is implemented in an external FPGA for flexibility.

The phase vs. temperature characteristic of the two WBs was initially determined with the help of a fixed 7MHz reference frequency. As in [8], after correcting for the inherent nonlinearity of the PDΔΣM (same for all samples & resistor types), each sample was trimmed at two temperatures (-35°C, 75°C), and then, for each resistor type, the remaining systematic error is corrected by a fixed 4th-order polynomial (the same for all samples). The resulting polynomials and calibration coefficients were then loaded in the FPGA, the FLL closed and a second temperature sweep done to characterize its output frequency over temperature. Figure 3.3.4 shows the frequency output of the 12 samples and the residual frequency error over temperature and supply voltage. The frequency error is less than ±250ppm over the temperature range from -45°C to 85°C, corresponding to a TC of 3.85ppm/°C (box method). Over a 1.7-to-2V supply range, the worst-case peak-to-peak frequency error is 548ppm, corresponding to a worst-case supply sensitivity of 0.18%/V. Figure 3.3.5 shows the period jitter of the DCO in open-loop and closed-loop modes. The open-loop and closed-loop period jitter is $22ps_{rms}$ and $23ps_{rms}$, respectively, showing that short-term jitter is mainly dominated by the DCO. The Allan Deviation (Fig. 3.3.5) is greatly improved by closed-loop operation, reaching a 250ppb floor beyond 3s measurement time.

Figure 3.3.6 summarizes the performance of the frequency reference and compares it to state-of-the-art RC oscillators with low TC and long-term stability. The proposed frequency reference achieves the lowest inaccuracy over multiple samples and the lowest long-term drift. This demonstrates that CMOS RC frequency references can achieve enough accuracy at a low power consumption to enable wired communication standards on systems-on-chip

Acknowledgements:
The authors would like to thank Infineon Technologies for financial support.

References:
[1] M. S. McCorquodale, et al., "A 0.5-to-480MHz Self-Referenced CMOS Clock Generator with 90ppm Total Frequency Error and Spread-Spectrum Capability," *ISSCC*, pp. 350-351, 2008.
[2] S. M. Kashmiri, et al., "A Scaled Thermal-Diffusivity-Based 16 MHz Frequency Reference in 0.16 µm CMOS," *IEEE JSSC*, vol. 47, no. 7, pp. 1535-1545, July 2012.
[3] J. Lee, et al., "A 1.4V 10.5MHz Swing-Boosted Differential Relaxation Oscillator with 162.1dBc/Hz FOM and $9.86ps_{rms}$ Period Jitter in 0.18µm CMOS," *ISSCC*, pp. 106-107, 2016.
[4] S. Jeong, et al., "A 5.8 nW CMOS Wake-Up Timer for Ultra-Low-Power Wireless Applications," *IEEE JSSC*, vol. 50, no. 8, pp. 1754-1763, Aug. 2015.
[5] T. Jang, et al. "A 4.7nW 13.8ppm/°C Self-Biased Wakeup Timer Using a Switched-Resistor Scheme," *ISSCC*, pp. 102-103, 2016.
[6] M. Choi, et al., "A 99nW 70.4kHz Resistive Frequency Locking On-Chip Oscillator with 27.4ppm/°C Temperature Stability," *IEEE Symp. VLSI Circuits*, pp. C238-C239, 2015.
[7] D. Griffith, et al., "A 190nW 33kHz RC Oscillator with ±0.21% Temperature Stability and 4ppm Long-Term Stability," *ISSCC*, pp. 300-301, 2014.
[8] S. Pan, et al., "A Resistor-Based Temperature Sensor with a $0.13pJ{\cdot}K^2$ Resolution FOM," *ISSCC*, pp. 158-159, 2017.

 978-1-5090-4940-0/18/$31.00

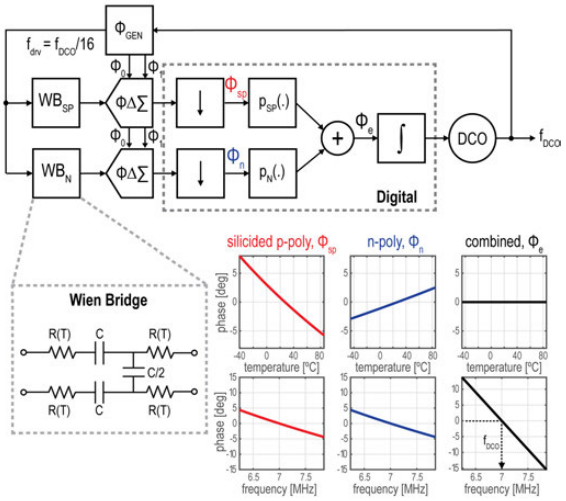


Figure 3.3.1: Block diagram of the proposed dual-RC reference (top), showing also the Wien-Bridge (WB) filter (bottom left) and the simulated phase response of the two WBs and their combination (bottom right).

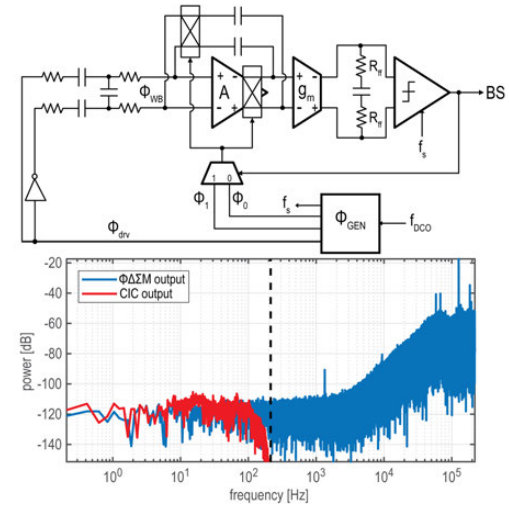


Figure 3.3.2: Circuit diagram of a Phase Domain ΔΣ-Modulator (PDSDM) (top) and PSD of the measured bitstream of the silicided p-poly bridge, using an external 7MHz clock (bottom).

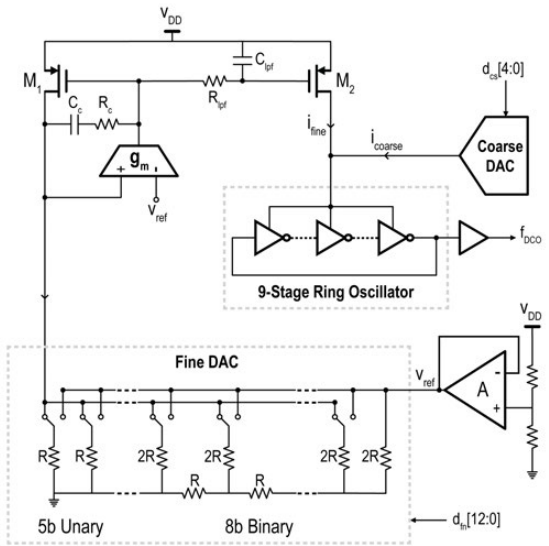


Figure 3.3.3: Circuit diagram of the Digitally Controlled Oscillator (DCO).

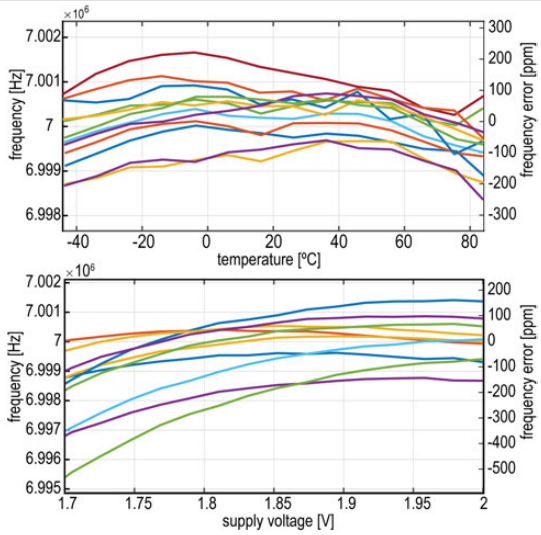


Figure 3.3.4: Measured frequency and frequency error over temperature (top) and supply voltage (bottom).

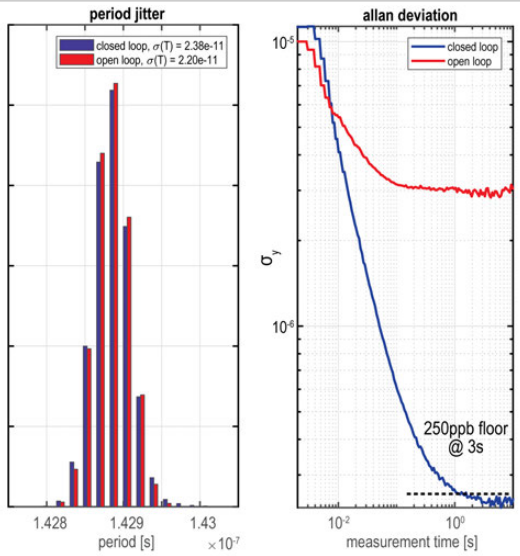


Figure 3.3.5: Measured period jitter (left) and Allan Deviation (right).

	This Work	Zhang VLSI2017	Jang [5] ISSCC2016	Hsiao VLSI2012	Choi [6] VLSI2016	Griffith [7] ISSCC2014	Savanth ISSCC2017
Process [nm]	180	180	180	60	180	65	65
Frequency [Hz]	8e6	24e6	3e3	3.2768e4	7.04e4	3.2768e4	1.3e6
TC [ppm/°C]	3.85	3.2[1]	13.8	16.67	27.4	38.18	96
T Range [°C]	-45 to 85	-40 to 150	-25 to 85	-20 to 100	-40 to 80	-20 to 90	0 to 150
Voltage [%/V]	0.18	0.03	0.49	0.125	0.5	0.09	0.49
V Range [V]	1.7 to 2.0	1.8 to 5.0	0.85 to 1.4	3.2 to 1.6	1.2 to 3.0	1.15 to 1.45	0.9 to 1.9
# of Samples	12	1	1	4	1	5	2
Allan Deviation Floor [ppm]	0.25	-	63	-	7	4	-
Power	750μW	200μW	4nW	4.48μW	99.4nW	0.19μW	0.92μW

[1] Utilizes a thin-film resistor

Figure 3.3.6: Performance summary and comparison table with previous work.

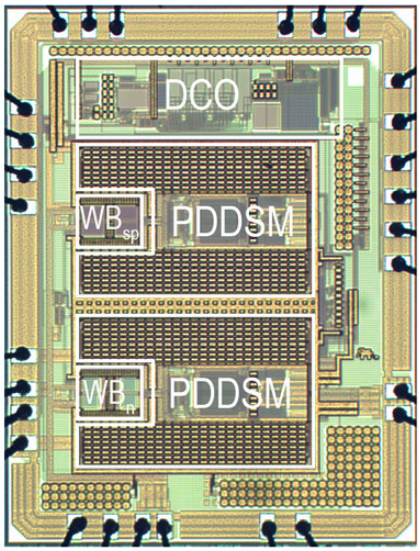


Figure 3.3.7: Die Micrograph.

978-1-5090-4940-0/18/$31.00

3.4 A 2×20W 0.0013% THD+N Class-D Audio Amplifier with Consistent Performance up to Maximum Power Level

Eric Cope, Julian Aschieri, Tony Lai, Franklin Zhao, Walter Grandfield, Michael Clifford, Pete Rathfelder, Qiyuan Liu, Siddartha Kavilipati, Aaron Vandergriff, Gerald Miaille

Qualcomm, Tempe, AZ

Conventional Class-D amplifiers, although more power efficient than Class-AB amplifiers, typically do not deliver the same audio quality. The non-ideal switching behavior of the output power stages can degrade the linearity, noise and power-supply-rejection-ratio (PSRR) performance of Class-D amplifiers if employed in open-loop configurations [1]. Closed-loop Class-D amplifiers shape the non-idealities of the power amplifiers (PAs) and provide improved performance [2]. Conventional analog feedback amplifiers (AFAs) sense the PA output and feed it back to compare with the audio input signal in the analog domain. Compared with AFAs, digital feedback amplifiers (DFAs) have emerged with benefits of improved control of the loop filter and pulse-width modulation (PWM) in the digital domain. However, the DFA architecture usually demands a high-performance analog-to-digital converter (ADC) to digitize the PA output in the feedback path; This ADC's non-idealities may become the bottleneck of the system [3]. In this paper, a 2-channel Class-D digital error-feedback amplifier (DEFA) with a peak THD+N of 0.0013% is presented. A series of proposed techniques enable the DEFA to maintain its performance up to the maximum power level available.

Figure 3.4.1 shows the block diagram of the proposed Class-D DEFA together with AFA and DFA. The first beneficial feature of DEFA is its capability for a direct digital PWM controller. In AFA, the audio input needs to go through the analog loop filter, the analog PWM and the PA before being delivered to the speaker. However, in DEFA, the digital input D_{IN} directly controls the digital PWM. Thus, the DEFA architecture can be made compatible with multiple modulation schemes including AD, BD and 4-phase modulation as in DFA [3]. The dead time of the PA can also be made dynamically adjustable based on the modulation levels, which offers the possibility of optimizing between the PA's linearity and its power efficiency. Besides, the spread- and shift-spectrum of the PWM carrier is made simple with only digital reconfiguration, which is clearly another benefit.

In addition to the direct PWM control, the DEFA features a unique error-processing scheme in its feedback path. There are basically two noise-shaping schemes functioning in the proposed DEFA architecture. The primary noise shaping (PNS) and the digital-to-analog converter (DAC) provide a high-quality input reference A_{IN} in the feedback path. The path consisting of the analog loop filter, and the FLASH quantizer constitutes the secondary noise shaping (SNS). The error signal A_{ERROR} is generated based on the difference between the input reference A_{IN} and the PA output A_{OUT}. Thus, the loop filter of the SNS loop only needs to process the error signal rather than the PA output signal A_{OUT} [4]. This feature enables the DEFA system to maintain its stability and performance up to its maximum power level, addressing the clipping issue of the existing Class-D amplifier solutions.

Figure 3.4.2 shows the proposed PNS DAC, which provides a high-fidelity reference for the DEFA. A 135dB (flat) signal to noise-and-distortion ratio (SNDR) primary noise shaping is achieved using a 5th-order $\Delta\Sigma$ modulator (DSM) with a 4b quantizer running at 6.5MHz. The pulse-density-modulation (PDM) data from the DSM is then converted to PWM data to control the DAC. The main reason to employ a PWM DAC is for its immunity to inter-symbol interferences (ISI) which is a known issue in conventional PDM DACs. The encoding from PDM to PWM is done using a rotating digital ramp together with sixteen comparators as illustrated in Fig. 3.4.2. In this way, the PWM data get effectively "Barrel Shifted", thus providing time-domain mismatch shaping in each DAC element. For instance, for a PNS DSM output of N, the pulse width for each PWM segment is then N/16 T_{QTZ} and rotates with a time shift of 1/16 T_{QTZ} between each two successive segments. The PNS DSM is optimized to reduce code changes and thus minimize jitter-induced error. Simulation results show that for 10ps RMS jitter, the signal to jitter-induced-noise ratio (SJNR) of the PNS DAC is 126.8dB.

Figure 3.4.3 shows the implementation of the SNS loop for the proposed DEFA. To deliver 20W to 8Ω and 4Ω loads with 90% efficiency, an on-chip NMOS/NMOS PA with bootstrap control is employed. The supply rails for the driver of the bottom NMOS are GND and AVDD, while the supply rails for the driver of the top NMOS switch are boosted to the domain of PVDD and PVDD+AVDD. Level shifting inside the non-overlapping generation block is required and must be carefully designed. The DEFA also offers the option to send the gate drive of the PA switches off-chip to drive external PA switches using higher supply voltages. The PA output is fed back before the LC filter through an external resistor network. The analog loop filter is implemented based on a feedforward topology cascading of five active-RC integrators. The output of each integrator has its own overload detection and reset circuitry. The 5b FLASH ADC runs at 26MHz and quantizes the output of the loop filter. The digital PWM can be configured to support AD/BD modulation with PWM frequency choices of 400/800/1600KHz.

In this design, a series of protection techniques is proposed to address the common performance degradation issue of Class-D amplifiers at full power levels. Firstly, the proposed DEFA processes only the error signal, which is the difference between the PA output and the audio input reference. Thus, superior stability benefits come directly as a result of the amplifier architecture choice. Secondly, the digital PWM duty-cycle is digitally adjustable based on the measured PA supply from an on-chip PSUADC (Fig. 3.4.3). This ensures that the SNS loop only processes errors added from the PA non-idealities and not PA supply variations. Thirdly, the proposed DEFA has overload detection schemes both in digital and analog domains, based on modulation index and integrator output swing, respectively. Once an overload condition is detected, the coefficients of the PNS DSM will be adjusted to ensure its stability. At the same time, the 3rd and 4th integrators in the analog loop filter will be reset and thus improve the loop stability. The measurement results of the amplifier output's transient waveforms and spectrums are shown in Fig. 3.4.4, emphasizing the effect of the overload performance protection of PNS and SNS. When a heavy overloading condition takes place, if a low-end switching regulator is employed, the PA supply may droop and fail to climb up immediately after the overload condition ends. This will hurt the linearity performance of the amplifier, lead to poor audio quality and potential instabilities of the system. In this design, a PNS DAC freezer is designed to limit the modulation index to 80% if this severe overload condition is detected. The freezer keeps the amplifier with tolerable performance while protecting the switching regulator, whose effect on the PA output transient waveform is also shown in Fig. 3.4.4.

The DEFA Chip was fabricated in a 0.18μm BCD process. Each of the two-channels of DEFAs can deliver 20W to both 8Ω and 4Ω loads at 90% efficiency when powered with a 20V supply. Figure 3.4.5 shows the measured THD+N performance of the DEFA versus the output power levels. For a 6kHz input using BD modulation and 400kHz PWM frequency, a peak THD+N of 0.0013% is achieved when delivering 11W to an 8Ω load. The proposed techniques on the performance protection of the DEFA enable a <0.006% THD+N performance at 20W power level under all load conditions. The measured dynamic range of the DEFA is 115.5dB (A-weighted) and the measured noise floor modulation is within 2dB. Figure 3.4.6 compares the performance of the proposed DEFA with the state-of-the-art designs. The proposed DEFA achieves a top THD+N performance under both full-scale power delivery and across audio frequencies. The die micrograph of the chip is shown in Fig. 3.4.7 with an area of 4.3mm^2.

Reference:

[1] J.-M. Liu, et al., "A 100 W 5.1-Channel Digital Class-D Audio Amplifier with Single-Chip Design," *IEEE JSSC*, vol. 47, no. 6, pp. 1344-1354, Nov. 2014.
[2] Texas Instruments TAS5720x data sheet, http://www.ti.com/lit/ds/symlink/tas5720m.pdf, Feb. 2016.
[3] D. Schinkel, et al., "A 5×80W 0.004% THD+N Automotive Multiphase Class-D Audio Amplifier with Integrated Low-Latency $\Delta\Sigma$ ADCs for Digitized feedback after the Output Filter," *ISSCC*, pp. 86-87, Feb. 2017.
[4] X. Jiang, et al., "Integrated Class-D Audio Amplifier with 95% Efficiency and 105dB SNR," *IEEE JSSC*, vol. 49, no. 11, pp. 2387-2396, Nov. 2014.

 978-1-5090-4940-0/18/$31.00

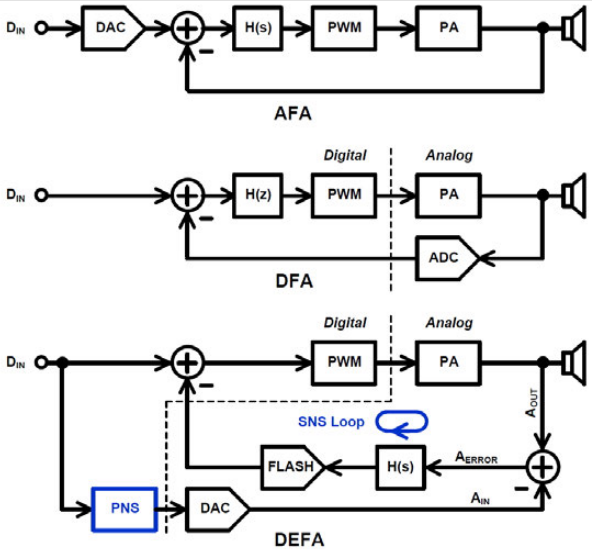


Figure 3.4.1: Block diagram of different topologies of Class-D amplifiers.

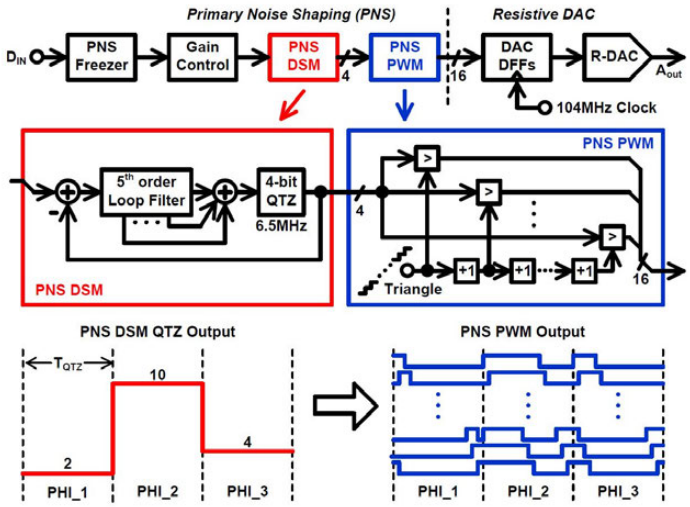


Figure 3.4.2: Block diagram of PNS DAC with PWM coding.

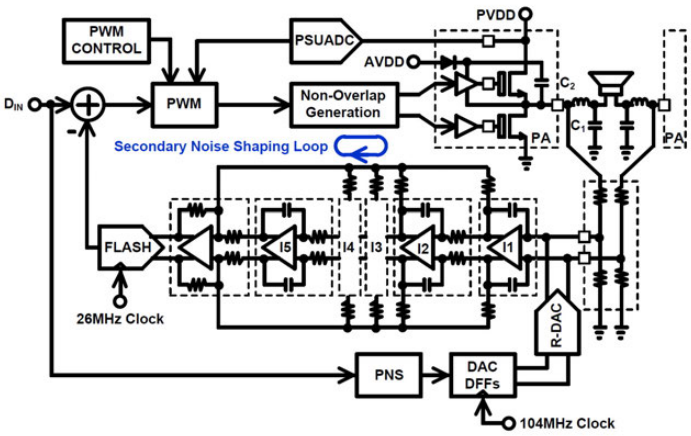


Figure 3.4.3: Block diagram of SNS loop.

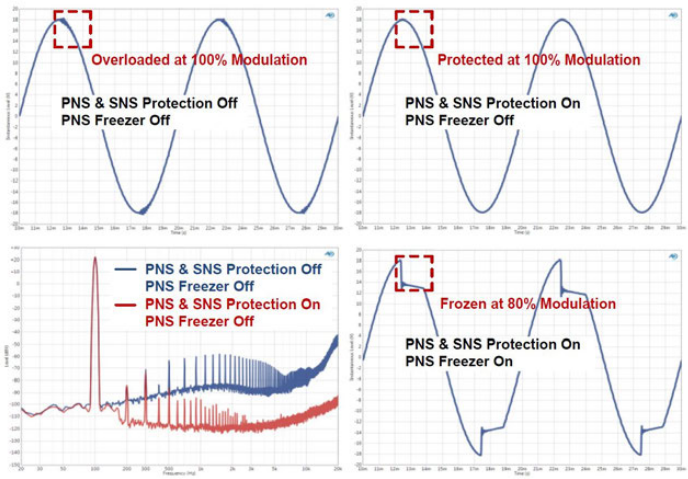


Figure 3.4.4: PNS and SNS performance protection together with PNS Freezer to maintain consistent performance of DEFA to the maximum power level.

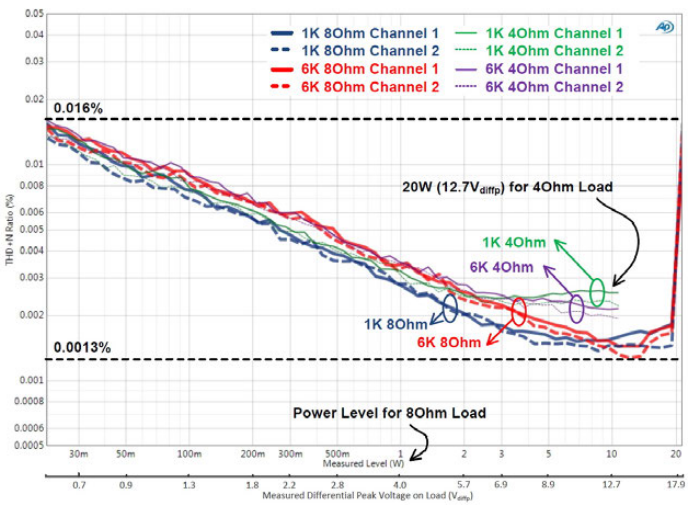


Figure 3.4.5: Measured THD+N versus output power level.

	This Work	[1]	[2]	[3]
Process	0.18 µm BCD	0.35 µm CMOS	N/A	0.14 µm BCD SOI
PVDD Supply	8 - 20 V	18 V	4.5 - 24 V	6 - 25 V
Maximum Output Power	20 W	13 W	40 W	80 W
Peak Efficiency	90%	88%	90%	N/A
THD+N @ 1 KHz 0 dBFS	0.006%	0.7%	10.0%	7.0%
Peak THD+N @ 1 KHz	0.0013%	0.075%	0.012%	0.004%
THD+N @ 1 KHz 1W	0.0029%	0.085%	0.014%	0.004%
Dynamic Range (A-weighted)	115.5 dB	84 dB	N/A	115 dB
Integrated Noise Level	20 µV	N/A	50 µV	19 µV
SNR (A-weighted)	116 dB	N/A	102 dB	N/A
PSRR (20 - 20 KHz)	80 - 50 dB	N/A	87 - 50 dB	88 (100Hz) - 60 dB
Noise Floor Modulation	<2 dB	N/A	N/A	N/A
Channel Isolation (20 - 20 KHz)	96 dB	36 dB	N/A	N/A
Quiescent Current	20.52 mA	9.4 mA	18.5 mA	N/A

Figure 3.4.6: Measured performance and comparison with the state of the art.

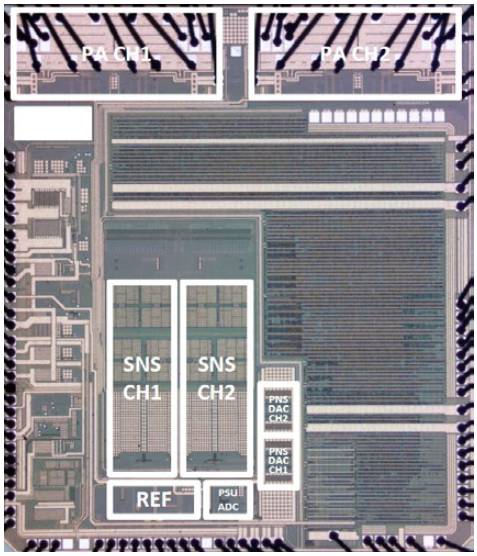


Figure 3.4.7: Die micrograph of the DEFA chip.

978-1-5090-4940-0/18/$31.00

3.5 A 0.0004% (-108dB) THD+N, 112dB-SNR, 3.15W Fully Differential Class-D Audio Amplifier with G_m Noise Cancellation and Negative Output-Common-Mode Injection Techniques

Wen-Chieh Wang, Yu-Hsin Lin, MediaTek, Hsinchu, Taiwan

Flicker (1/f) noise is a main design issue when realizing a low-noise and high-SNR audio amplifier. Utilizing large device sizes and chopper-stabilization techniques are commonly adopted approaches to mitigate 1/f noise. However, the choppers are seldom used in pulse-width-modulation (PWM) Class-D audio amplifiers (CDAs) because the conventional chopping method applied in the opamps along the signal path of CDAs inevitably results in extra aliasing and deteriorates the linearity. Therefore, a CDA with high SNR and low total harmonic distortions (THDs) has become a challenging design to ensure high-fidelity audio applications.

In this work, the Gm-noise-cancellation (GmNC) technique is introduced in the CDA to suppress the 1/f noise for the SNR improvement, further improving the linearity of the CDA. Moreover, the negative output common-mode injection (NOCMI) technique is introduced to improve the linearity by reducing the voltage fluctuation of the input common-mode (CM) of the first opamp of the CDA. This CDA achieves 0.0004% (–108dB) THD+N and 112dB SNR (A-weighted), and is capable of delivering a maximum output power of 3.15W into a 4Ω load under 5.5V battery voltage.

Figure 3.5.1 shows a conventional closed-loop PWM CDA in which the opamp of the first integrator is chopped in order to reduce its 1/f noise contributed to the CDA. The virtual ground nodes v_{XN} and v_{XP} exhibit voltage fluctuations resulting from the finite opamp gain (A_{OPAMP}) when the rail-to-rail CDA outputs v_{OP} and v_{ON} feedback to v_{XN} and v_{XP}, respectively. The magnitude of the voltage fluctuations at v_{XP} and v_{XN} is inversely proportional to A_{OPAMP}. The first transconductance of the opamp Gm1 is chopped by the chopper clock ϕ_{CK} at the frequency of f_{CHP} to overcome the 1/f noise, while the operation of the chopper causes abrupt voltage changes at the node v_{YN} (v_{YP}) where the parasitic C_{P1} is charged or discharged at every rising and falling edge of ϕ_{CK}. This results in a periodic current pulse Δi_X at the frequency of $2\times f_{CHP}$, where the magnitude of Δi_X is proportional to C_{P1} and inversely proportional to A_{OPAMP}. For the BD (3-level) switching PWM CDA with an input signal at f_{IN} and a switching frequency at f_S, the out-of-band harmonic distortions (HDs) at $(M\times 2\times f_S \pm N\times f_{IN})$ are folded back in-band by the periodic Δi_X, and the folded-back aliasing components are at $[(M\times 2\times f_S \pm N\times f_{IN}) \pm K\times 2\times f_{CHP}]$, where M, N and K are integers. Furthermore, turn-on resistance of the chopper switch in the CDA loop contributes an extra pole that degrades the bandwidth of the opamp; hence, the linearity of the CDA is further compromised.

The proposed GmNC technique to mitigate the 1/f noise of the CDA is also illustrated in Fig. 3.5.1. The impairments of the OP1 including noise and finite gain error are equivalently lumped as v_{N1} at the virtual ground inputs of OP1, v_{XN} and v_{XP}. Measuring the v_{N1}, the GmNC outputs a current i_{GmNC} to v_{XN} and v_{XP} to cancel the i_{N1} which is generated by v_{N1}. The GmNC is designed to reduce about 95% of the 1/f noise of OP1, and the GmNC is inversely proportional to R_{NC}, which is matched to the equivalent resistance R_{eq} at the virtual ground inputs of OP1. The mismatch between R_{NC} and R_{eq} can be kept less than 1% as they are of the same type. Because the GmNC is outside the main loop of the CDA, the extra 1/f noise from it can be simply removed by the conventional chopper technique, where the chopping spikes can be filtered by the lowpass filters (Q-LPFs) that are designed with a low 3dB corner to avoid the folded-back aliasing without sacrificing the bandwidth of the CDA. The proposed methodology is different from [1], which adopts negative conductance only to deal with the finite gain error of the opamp to linearize its front-end receiver, and is also different from [2], which uses a negative-*R* assistant to deal with the finite gain error and noise of the opamp to improve linearity and noise of its continuous-time delta-sigma modulator. Both negative conductance and negative-*R* are realized in different approaches, and contribute extra 1/f noise that is required to be solved. The key idea of the proposed GmNC technique is to cancel the equivalent impairments v_{N1} of the OP1, where the GmNC is outside the loop of the CDA, so that the chopper technique and Q-LPFs can be applied to solve the 1/f noise of the GmNC and chopping spikes, respectively. The proposed GmNC not only solves the 1/f noise, but also offers a benefit of improving the linearity of the CDA as the finite gain error of OP1 is resolved concurrently.

For the BD switching PWM CDA, the input virtual ground nodes of the first opamp exhibit large CM voltage fluctuations that deteriorate the linearity of the CDA owing to degradation of the gain of the first opamp. The authors of [3] proposed a method to reduce the CM voltage fluctuations to improve the CDA linearity. However, the switch pairs cannot operate at the same time due to the extra inverter logic delay in the feedback controls, which results in linearity degradation since the signal information of the CDA is transformed into the time domain by PWM modulation and any delay of the feedback control signals contributes HDs.

Figure 3.5.2 shows the proposed NOCMI technique for the BD switching PWM CDA to solve the CM voltage fluctuations at the virtual ground nodes v_{XP} and v_{XN}, especially when $(v_{OP}, v_{ON}) = (V_{BAT}, V_{BAT})$ and (ground, ground). To simplify explanations, V_{BAT}, ground and $\frac{1}{2}V_{BAT}$ are denoted by "+1", "–1" and "0", respectively. When (v_{OP}, v_{ON}) = ("+1", "+1"), the NOCMI injects a "–1" to both v_{XP} and v_{XN} through R_{CMO} to make equivalent CM of v_{XP} and v_{XN} return to "0". On the contrary, the NOCMI injects a "+1" to both v_{XP} and v_{XN} through R_{CMO} as (v_{OP}, v_{ON}) = ("–1", "–1"). Furthermore, when (v_{OP}, v_{ON}) = ("+1", "–1") or ("–1", "+1"), the NOCMI injects a "0" to both v_{XP} and v_{XN} through R_{CMO}. Consequently, the proposed NOCMI keeps the CM of v_{XP} and v_{XN} from changing for the application of the BD switching PWM CDA. It is worth to mention that the NOCMI injects only a CM signal v_{NOCMI} into both v_{XP} and v_{XN} by using negative feedback, and any feedback delay of the NOCMI does not contribute differential-mode (DM) HDs to the CDA. Moreover, the PWM CMFB in [4] is utilized to ensure the CDA outputs v_{OP} and v_{ON} are with low CM HDs; hence, the HDs of v_{NOCMI} are also kept low. R_{CMO} and R_{CMI} are incorporated with chopper switches to perform dynamic element matching for the prevention of CM-to-DM HDs, and the Q-LPF is then further adopted to filter out chopping spikes.

Figure 3.5.3 depicts the main schematic diagram of a fully differential PWM CDA that adopts the proposed GmNC and NOCMI techniques. The designed unity-gain-bandwidth (f_u) and the frequency of v_{SAW} (f_S) of this CDA are 200kHz and 650kHz, respectively. The GmNC is implemented at v_{XP} and v_{XN} to remove the noise and finite gain error of OP1. Meanwhile, OP4, R_5, R_{CMI} and R_{CMO} realize the NOCMI to reduce the CM voltage fluctuations at v_{XP} and v_{XN} to enhance linearity. R_5=0.45R_{CMI} is designed to ensure v_{NOCMI} is within the output dynamic range of OP4, and the input CM range of OP1 is sufficient to cover the residual voltage fluctuations at v_{XP} and v_{XN}. The frequency of ϕ_{CK} is at $f_S/4$. The rise and fall times of the ϕ_{CK} are around 0.2ns; the corner frequency of Q-LPF is designed at around 20MHz to provide sufficient suppression of chopping spikes. In addition, the PWM CMFB is implemented by OP3, R_6 and C_6 to improve the PSRR and reduce CM HDs of this CDA.

This CDA is fabricated in a 0.153μm CMOS process, and packaged in a QFN. It is measured with a 4Ω resistor in series with a 33μH inductor to mimic the load of a loudspeaker. The comparisons of the FFT spectrum of the proposed CDA, which enables and disables the proposed NOCMI and GmNC techniques, are shown in Fig. 3.5.4. It can be seen that the HDs are greatly improved by the NOCMI technique, and the 1/f noise is significantly improved when the GmNC is further enabled; the HDs are also reduced as expected. It demonstrates a THD+N of –108dB (0.0004%) at the output power (P_{OUT}) of 2W under V_{BAT}=5.5V. Figure 3.5.5 illustrates the comparisons of the FFT spectrum of the CDA using (NOCMI+GmNC) to that of the CDA utilizing a conventional chopper in the OP1. Though both methods are effective at reducing 1/f noise, the CDA using the conventional chopper suffers from higher HDs due to the aliasing of out-of-band HDs. The measured peak SNR (A-weighted) of the CDA is 112dB with a measured output noise of 7.5μVrms. The PSRR at 217Hz is about 118dB. Operated at V_{BAT}=5.5V, the CDA delivers maximum P_{OUT} of 3.15W to a 4Ω load at 1% THD+N, where the efficiency is 89% with a quiescent current of 1.6mA. Figure 3.5.6 compares the performance of the presented fully differential CDA using (NOCMI+GmNC) techniques to that of the state-of-the-art CDAs. Figure 3.5.7 shows the chip micrograph, and the active area is 2.28mm^2.

Acknowledgements:
The authors thank Y.-Y. Lin for technical discussion and Mediatek/ADCT for supports.

References:

[1] D.H. Mahrof, et al., "Cancellation of OpAmp Virtual Ground Imperfections by a Negative Conductance Applied to Improve RF Receiver Linearity," *IEEE JSSC*, vol. 49, no. 5, pp. 1112-1124, May 2014.
[2] M. Jang, et al., "A 55μW 93.1dB-DR 20kHz-BW Single-bit CT ΔΣ Modulator with Negative R-Assisted Integrator Achieving 178.7dB FoM in 65nm CMOS," *IEEE Symp. VLSI Circuits*, pp. C40-41, June 2017.
[3] S. Kwon, et al., "A 0.028% THD+N, 91% Power-Efficiency, 3-Level PWM Class-D Amplifier with a True Differential Front-End," *ISSCC*, pp. 96-98, Feb. 2012.
[4] W. Wang et al., "A 118dB-PSRR 0.00067% (-103.5dB) THD+N and 3.1W Fully Differential Class-D Audio Amplifier with PWM Common-Mode Control," *ISSCC*, pp. 90-91, Feb. 2016.

 978-1-5090-4940-0/18/$31.00

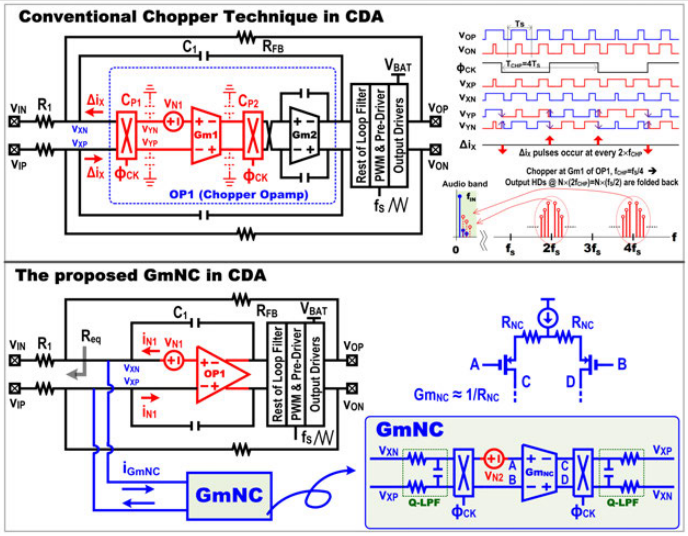
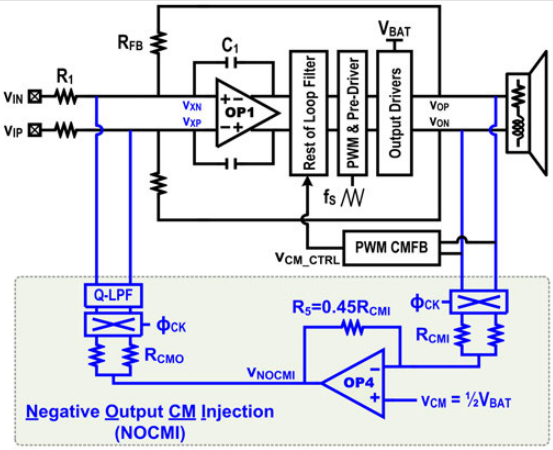


Figure 3.5.1: The CDA with the conventional chopper technique, and the CDA with the proposed GmNC technique.



BD (3-level) switching PWM CDA with NOCMI					
V_{OP}	V_{ON}	CM of V_{XP} & V_{XN}	V_{OP}	V_{ON}	CM of V_{XP} & V_{XN}
V_{BAT}	0	½V_{BAT}	0	0	~ ½V_{BAT}
0	V_{BAT}	½V_{BAT}	V_{BAT}	V_{BAT}	~ ½V_{BAT}

Figure 3.5.2: The proposed NOCMI technique in a CDA.

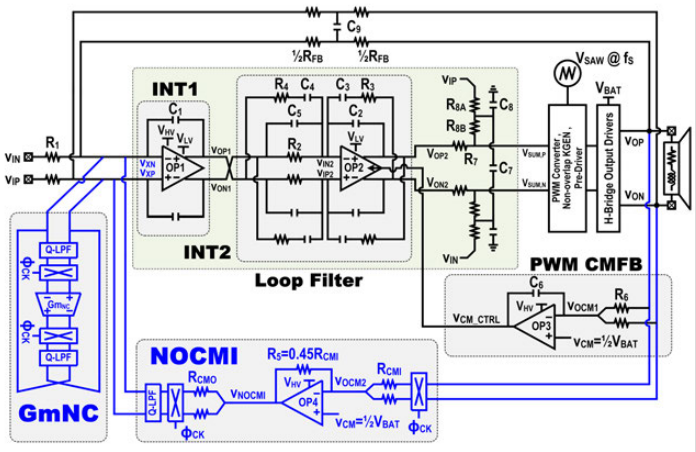


Figure 3.5.3: Schematic diagram of the proposed fully differential PWM CDA with GmNC and NOCMI techniques.

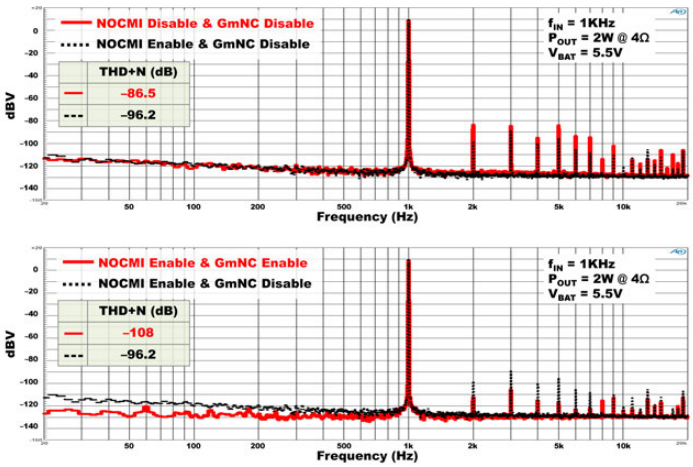


Figure 3.5.4: FFT comparisons of the CDA that enables/disables the proposed NOCMI and GmNC techniques, respectively.

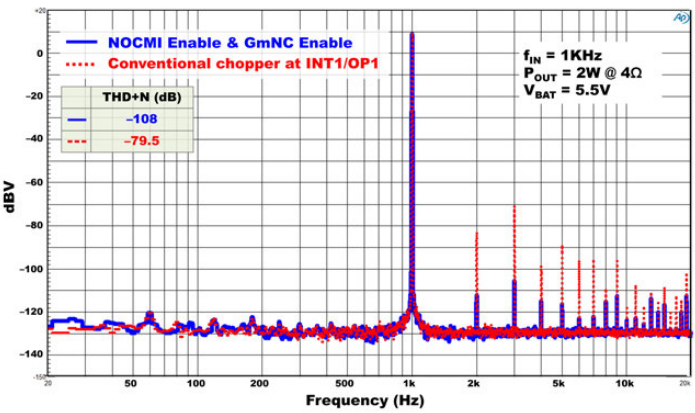


Figure 3.5.5: FFT comparisons of the CDA between conventional chopper technique and (NOCMI+GmNC) technique.

	Unit	Condition	This Work	W. Wang ISSCC 2016	M. Teplechuk ISSCC 2011	A. Nagari ISSCC 2012	S. Kwon ISSCC 2012	M. Kinyua JSSC 2015	X. Jiang JSSC 2014
Supply	V		3 ~ 5.5	3 ~ 5.5	2.5 ~ 5.5	2.5 ~ 6.6	2.7 ~ 5.2	2.5 ~ 5.5	2.5 ~ 5.5
THD+N	%	1KHz	0.0004 (4Ω)	0.00067 (4Ω)	0.00149 (4Ω)	0.025 (8Ω)	0.028 (8Ω)	0.0031 (8Ω)	0.004 (8Ω)
	dB		-108	-103.5	-96.5	-72.0	-71.1	-90.2	-88.0
Peak SNR	dB	A-weighted	112	108	103	104	97.5	105	105
PSRR	dB	217Hz	118	118	96	93	72.2	92	96
$P_{OUT,MAX}$	W	THD+N=1%	3.15 (4Ω)	3.1 (4Ω)	3.1 (4Ω)	2.5 (8Ω)	0.725 (8Ω)	1.5 (8Ω)	1.75 (8Ω)
Efficiency (η)	%	$P_{OUT,MAX}$ (4Ω)	89	89.5	90	--	--	--	--
		$P_{OUT,MAX}$ (8Ω)	91.5	92.4	93	91	90.9	85.2	95
Minimum Load	Ω		4	4	4	8	8	8	8
I_Q	mA		1.6	1.45	4	2	2	2.8	--
Switching Freq	kHz		650	650	1000	446	420	2133	722
Process			0.153 µm CMOS	0.153 µm CMOS	0.25 µm CMOS	0.13 µm CMOS	0.18 µm CMOS	55 nm CMOS	0.18 µm CMOS
Chip Area	mm^2		2.28	1.85	1.44	2.19	1.14	1.17	1.95
Package			QFN	QFN	WLCSP	WLCSP	WLCSP	QFPN	Wire-Bond
Topology			PWM	PWM	UPWM	PWM	PWM	DPWM	PWM
Loop Filter Order			3	3	2	--	2	4	4
FOM1			1391	921	156	18	16	98	--
FOM2			110461	73178	985	81	7	391	--

$$FOM1 = \frac{\eta}{I_{Q(A)} \times THD+N|_{1kHz} \times 10^5} \qquad FOM2 = \frac{\eta}{I_{Q(mA)} \times THD+N|_{1kHz} \times PSRR|_{217Hz} \times 10^5}$$

Figure 3.5.6: Performance summary and comparison with previous work.

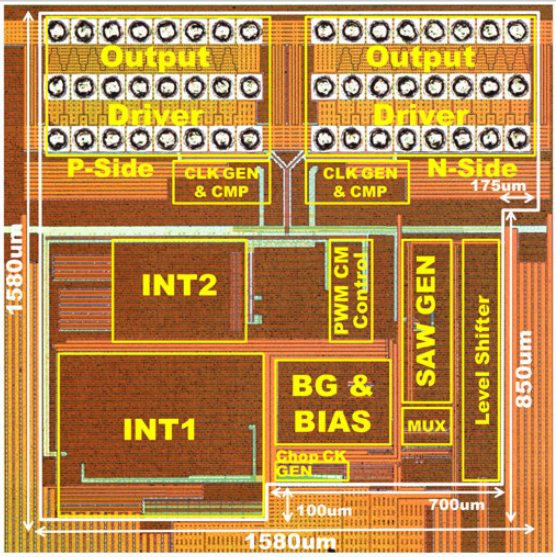


Figure 3.5.7: Die micrograph (area=2.28mm²).

978-1-5090-4940-0/18/$31.00

3.6 A 0.96mA Quiescent Current, 0.0032% THD+N, 1.45W Class-D Audio Amplifier with Area-Efficient PWM-Residual-Aliasing Reduction

Shih-Hsiung Chien, Yi-Wen Chen, Tai-Haur Kuo

National Cheng Kung University, Tainan, Taiwan

Low quiescent current (I_Q) is critical for Class-D audio amplifiers in mobile devices to extend battery usage time [1], since typical audio signals have a high crest factor of 10 to 20dB. In addition, low distortion is also important for audio fidelity. Distortion sources in closed-loop Class-D amplifiers can be classified into two types. One is attributed to the nonlinearities of PWM modulators and power stages, while the other is due to the aliasing of fed-back PWM high-frequency residuals, the latter of which comprises phase-error and duty-cycle-error distortions [2]. Figure 3.6.1 shows 2nd-order closed-loop amplifiers and existing techniques for enhancing an amplifier's linearity. Increasing the loop filter order to obtain a higher in-band loop gain by using more integrators [3] or the single-amplifier-biquad [4] suppresses all aforementioned distortions except for the phase-error distortion, which can be suppressed by adding a phase-error-free PWM modulator [2]. However, these techniques increase I_Q and/or die area due to the additional active circuits and/or several resistors and capacitors. Since phase-error distortion, as well as duty-cycle-error distortion, is caused by the fed-back PWM high-frequency residuals aliasing with the reference triangular wave V_{TRI}, a uniform PWM [5] with a sample-and-hold circuit implemented before the PWM modulation reduces the PWM residuals via an equivalent notch filtering. However, loop stability is affected by the notch filtering unless the PWM switching frequency f_{SW} is increased, but doing so increases power consumption [4]. Though the technique in [1] uses a feed-forward path with a replicated loop filter to eliminate the PWM residuals without affecting loop stability, the replicated loop filter increases both I_Q and area.

Figure 3.6.2 shows the proposed low-power area-efficient PWM-residual-aliasing reduction technique applied to a conventional 2nd-order Class-D amplifier. The amplifier's input v_{IN} contains an audio-band signal that is much slower than f_{SW}; hence, for clarity, the time-domain waveforms of the amplifier for DC input are plotted at the bottom of Fig. 3.6.2. The v_{IN} is converted into current i_{IN} (i.e. $i_{IN+} - i_{IN-}$) via R_{IN}. The loop filter output v_{INT2} is PWM-modulated, then amplified by the power stage, and finally converted into feedback current i_{FB} (i.e. $i_{FB+} - i_{FB-}$) via R_{FB} to complete the feedback mechanism. The i_{FB} mainly comprises the input audio-band signal and the PWM high-frequency components. For a conventional amplifier, since the audio-band signal of i_{FB} is similar to i_{IN}, most of the audio-band signal is cancelled without flowing into the loop filter. However, the PWM high-frequency components of i_{FB} directly flow into the loop filter as i_{C1} (i.e. $i_{C1+} - i_{C1-}$), which leads to large PWM residuals at v_{INT1} and v_{INT2}. The PWM residuals at v_{INT2} are intermodulated with V_{TRI}, resulting in fold-back in-band aliasing distortion [5].

The goal of the proposed technique is to generate a current pulse train i_{FF} inverse to the high-frequency components of i_{FB}, and then add it into the loop filter to cancel the PWM high-frequency components of i_{FB}. In this work, since the amplifier closed-loop gain v_{OUT}/v_{IN} and the PWM modulation gain G_{PWM} are equal, v_{IN} and v_{INT2} are similar when the amplifier input frequency f_{IN} is much lower than f_{SW}. Hence, the proposed technique is realized by directly using v_{IN} to generate the PWM pulse train, then convert it into current via R_{FF}, and finally add it to the loop filter. The capacitor C_{FF} blocks the audio-band signal of i_{FF} and passes the PWM high-frequency components. By adding i_{FF}, the resultant PWM high-frequency components of i_{C1} and the PWM residuals at v_{INT1} and v_{INT2} are greatly reduced, as shown in the bottom right of Fig. 3.6.2. Hence, the PWM-residual-aliasing distortion is decreased, which enhances linearity. The loop stability is unaffected by the proposed technique, since the original feedback loop remains unchanged.

By respectively modeling the PWM modulation inside the loop and that in the proposed technique as constant gains G_{PWM} and G_{FF}, the transfer function from v_{IN} to v_{INT2} is derived as

$$\frac{v_{INT2}(s)}{v_{IN}(s)} = \frac{1+s(R_{FF}+G_{FF}R_{IN})C_{FF}}{1+sR_{FF}C_{FF}} \times \frac{R_{FB}}{R_{IN}} \times \frac{(1+sR_2C_2)}{G_{PWM}(1+sR_2C_2)+s^2R_{FB}R_1C_1C_2}.$$

In this design, $R_{FF} = R_{FB} = 2 \cdot R_{IN}$, and $G_{PWM} = G_{FF} = 2$. The Bode plot of $v_{INT2}(s)/v_{IN}(s)$ in Fig. 3.6.3 shows that the magnitude difference between v_{INT2} and v_{IN} is negligible for f_{IN}<10kHz, while their phase difference leads to a timing delay Δt_{FF} between the two current pulse trains i_{FF} and i_{FB}. By properly designing R_{FF} and C_{FF}, the phase shift versus f_{IN} over the entire audio band is approximately linear, resulting in a near-constant Δt_{FF} due to the linear-phase property, as derived at the bottom of Fig. 3.6.3. Since this work uses BD PWM modulation, there are 2 pulses in each switching period T_{SW}. Thus, Δt_{FF} is designed as $T_{SW}/2$, so that each i_{FB} pulse is mostly cancelled by the delayed i_{FF} pulse, as shown in the waveform of Fig. 3.6.3, resulting in greatly reduced PWM residuals in i_{C1} compared with a conventional amplifier. The timing delay Δt_{FF} changes with variations in R_{FF} and C_{FF}; however, T_{SW} also changes in accordance with variations in the resistor R_{TRI} and capacitor C_{TRI} of the on-chip tri-wave generator. Hence, only the RC mismatch affects the inconsistency between Δt_{FF} and $T_{SW}/2$, and the simulated THD+N degradation due to the 3-σ RC mismatch is <0.3dB.

In circuit implementation, the two added comparators of the proposed technique are simple and identical to those in the feedback loop. The comparator offset affects the PWM transition timing of i_{FB} and i_{FF}, which may result in larger PWM residuals in i_{C1}; nevertheless, the effect on THD+N is insignificant as the input-referred offset voltage of the comparators in this work is designed as 2mV. Although additional noises, including the jitter noise injected at the PWM modulation G_{FF}, the comparator's circuit noise, and the thermal noise of R_{FF}, are introduced by the proposed technique, the input-referred transfer function from these noises to v_{IN} is $(sG_{FF}R_{IN}C_{FF})/(1+s(R_{FF}+G_{FF}R_{IN})C_{FF})$, which contains a 1st-order suppression ability over the audio band. In this design, the corner frequency of $R_{FF}C_{FF}$ is designed around 70kHz, and thus these noises are suppressed by at least 98% via the equivalent filtering.

The measured output spectrum with a 1kHz input and 215kHz f_{SW} is shown at the bottom of Fig. 3.6.3. With the proposed technique, the 3rd-harmonic distortion (HD3) is suppressed by 22.6dB, yet the output noise floor remains unchanged. The measured THD+N vs. f_{IN} in Fig. 3.6.4 shows that the proposed technique improves the THD+N by at least 15.8dB, even when f_{IN} is up to 6kHz. This chip was inputted with 30 equal-amplitude tones at around 1/3-octave bands from 23Hz to 20kHz, for which the measured HD3 suppression of the 1kHz tone is 21.2dB despite the faster-changing input. The measured THD+N vs. output power with 1kHz input in Fig. 3.6.4 shows that the minimum THD+N is improved by 16.2dB with the proposed technique enabled, resulting in only 0.0032% THD+N when delivering 0.6W to an 8Ω load.

Figure 3.6.5 shows the I_Q breakdown chart of this work. The added circuit of the proposed technique consumes only 29μA of static current to enable a 2nd-order Class-D amplifier to achieve 0.0032% THD+N while operating at only 215kHz f_{SW}. In the upper-right of Fig. 3.6.5, the simulated THD+N vs. f_{SW} for a conventional 2nd-order amplifier shows that the required f_{SW} increases to 400kHz to achieve the same THD+N. As such, with the proposed technique lowering the required f_{SW}, more than 33% quiescent current is reduced. Compared with other state of the art, this work is implemented in less-advanced cost-effective 0.5μm CMOS technology while achieving the smallest controller active area of 0.171mm^2. The added circuit of the proposed technique occupies only 0.029mm^2.

Figure 3.6.6 compares the measured performance with other state-of-the-art Class-D amplifiers that achieve <0.01% THD+N. With the proposed low-power area-efficient PWM-residual-aliasing reduction technique, the amplifier in this work, with 8Ω load, achieves a competitive THD+N while consuming the lowest quiescent current of 0.96mA, occupying the smallest active area of 0.49mm^2, and achieving a highest FOM of 306. Figure 3.6.7 shows the die micrograph.

Acknowledgements:
The authors thank the fabrication support provided by the National Chip Implementation Center, Taiwan.

References:

[1] T.-H. Kuo, et al., "A 2.4 mA Quiescent Current, 1 W Output Power Class-D Audio Amplifier with Feed-Forward PWM-Intermodulated-Distortion Reduction," *IEEE JSSC*, vol. 51, no. 6, pp. 1436-1445, June 2016.

[2] L. Guo, et al., "A 101 dB PSRR, 0.0027% THD+N and 94% Power-Efficiency Filterless Class D amplifier," *IEEE JSSC*, vol. 49, no. 11, pp. 2608–2617, Nov. 2014.

[3] X. Jiang, et al., "Integrated Class-D Audio Amplifier With 95% Efficiency and 105dB SNR," *IEEE JSSC*, vol. 49, no. 11, pp. 2387–2396, Nov. 2014.

[4] W.-C. Wang and Y.-H. Lin., "A 118dB-PSRR 0.00067%(-103.5dB) THD+N and 3.1W Fully Differential Class-D Audio Amplifier with PWM Common-Mode Control," *ISSCC*, pp. 90-91, Feb. 2016.

[5] M. Teplechuk, et al., "Filterless Integrated Class-D Audio Amplifier Achieving 0.0012% THD+N and 96dB PSRR When Supplying 1.2W," *ISSCC*, pp. 240–241, Feb. 2011.

 978-1-5090-4940-0/18/$31.00

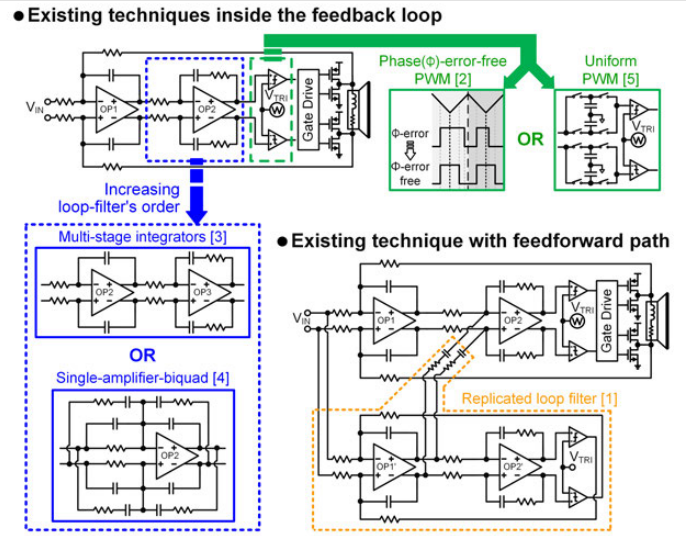


Figure 3.6.1: Conventional 2nd-order Class-D amplifiers and existing techniques for linearity enhancement.

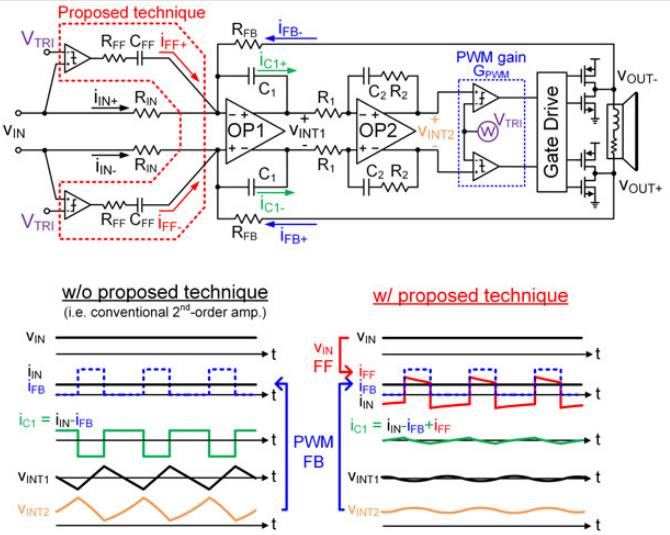


Figure 3.6.2: Proposed PWM-residual-aliasing reduction technique and the waveforms for DC input.

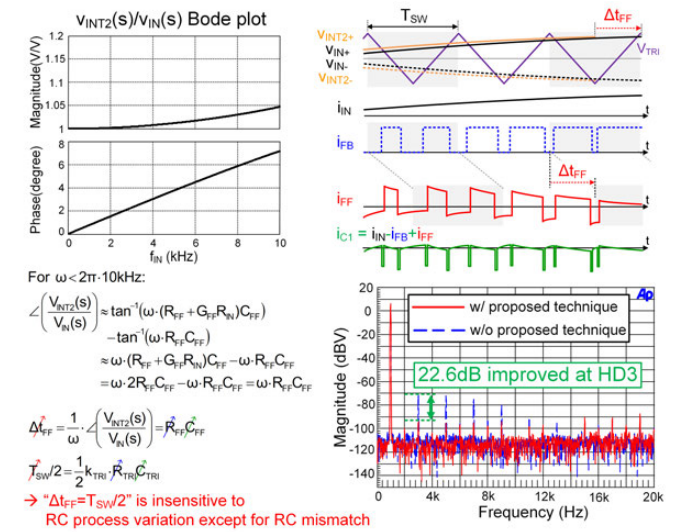


Figure 3.6.3: i_{FF} and i_{FB} waveforms with the derived phase-shifted delay Δt_{FF}, and the measured output spectrum.

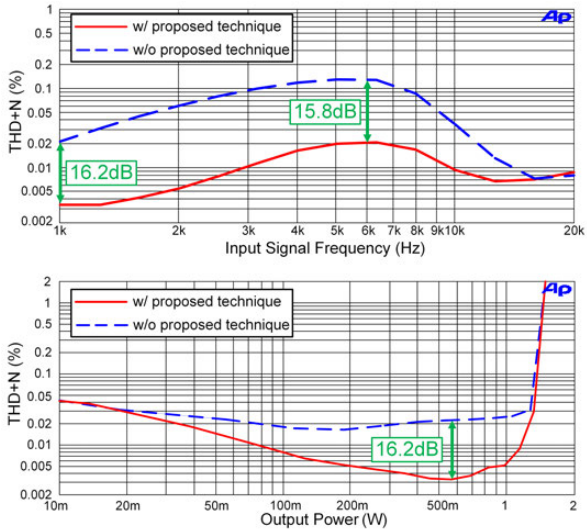


Figure 3.6.4: Measured THD+N vs. f_{IN} and THD+N vs. output power plots without and with the proposed technique.

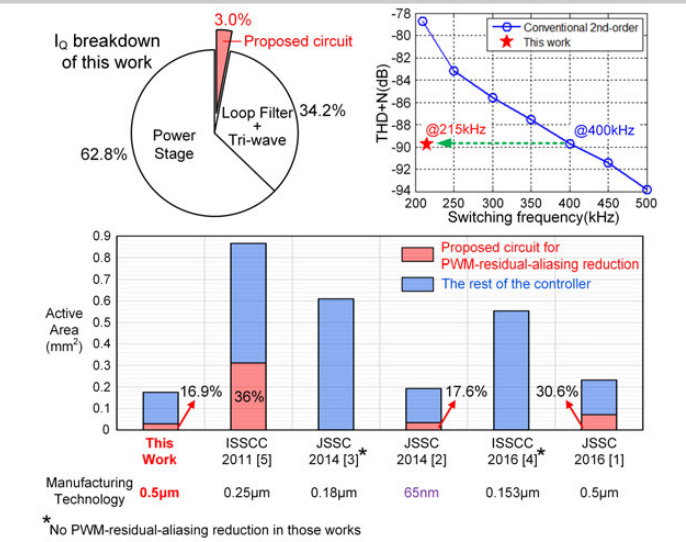


Figure 3.6.5: Quiescent current breakdown chart, the required f_{SW} comparison, and the controller area comparison.

	This Work	ISSCC 2011 [5]	JSSC 2014 [3]	JSSC 2014 [2]	ISSCC 2016 [4]	JSSC 2016 [1]
Supply Voltage (V)	2.5-5.5	2.5-5.5	2.5-5.5	1.2-4	3-5.5	2.5-5
Peak SNR (dB)	104	103	105	97	108	102
Max. P_{out}(W) @ THD+N≤1%, 8Ω load	1.45 (5V)	n/a*	1.75 (5.5V)	0.85 (4V)	n/a*	1.02 (4.2V)
Minimum THD+N (%)	0.0032	0.00122	0.004	0.0027	n/a*	0.0037
Efficiency η (%)	94	93	95	94	92.4	92
f_{sw} (kHz)	215	1000	722	320	650	200
I_Q (mA)	0.96	4	n/a*	3.1	1.45	2.4
Loop Filter's Order	2nd	2nd	4th	4th	3rd	2nd
Chip Area (mm²)	0.91	1.44	1.95	1.69	1.85	0.86
Active Area** (mm²)	0.49	1.44	1	0.6	1.23	0.57
Controller Area** (mm²)	0.171	0.866	0.61	0.193	0.672	0.232
Technology (μm)	0.5	0.25	0.18	0.065	0.153	0.5
FOM*** (@ 1kHz f_{IN})	306	190.6†	n/a*	112.3	n/a*	103.6

* n/a: not available for 8Ω measurement data

** Estimated according to the die photos in the papers

*** $FOM = \frac{\eta}{I_Q \times (THD+N)_{min,1kHz\,f_{IN}} \times 10^5}$ [5]

† Calculated by (THD+N)@ 1kHz f_{IN} for comparison

Figure 3.6.6: Measured performance summary and comparison.

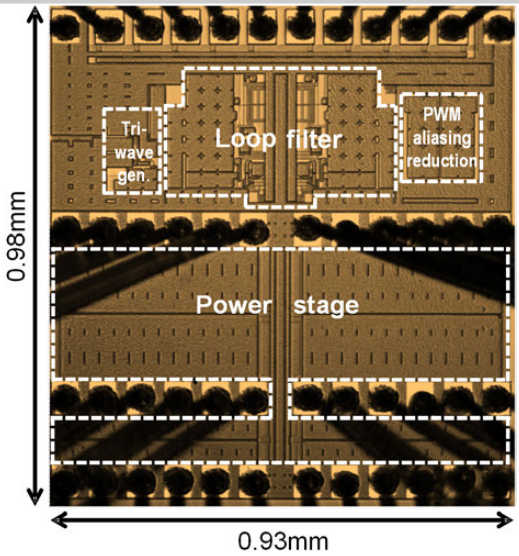


Figure 3.6.7: Die micrograph.

978-1-5090-4940-0/18/$31.00

3.7 A Low-Power 3.25GS/s 4th-Order Programmable Analog FIR Filter Using Split-CDAC Coefficient Multipliers for Wideband Analog Signal Processing

Shinwoong Park[1], Dongseok Shin[2], Kwang-Jin Koh[1], Sanjay Raman[1]

[1]Virginia Tech, Blacksburg, VA; [2]Intel, Hillsboro, OR

Discrete-time (DT) circuits provide a means to overcome the analog-circuit design challenges in deeply scaled digital CMOS technologies while benefitting from the reduced switch on-resistance and parasitic capacitance, resulting in lower dynamic power dissipation. In addition, such DT analog circuits can reduce the requirements on analog-to-digital converters that precede digital processing [1]. Recent DT domain filters achieve high-order narrowband programmable filtering with low power and high linearity even under low supply voltage [2,3]. However, DT switched capacitor circuits have not been considered for wideband analog signal processing (ASP) applications such as on-chip implementation of FIR-based beamforming [4,5]. While the AFIR filter proposed in [6] is a suitable approach for programmable wideband ASP applications, in that design only symmetric and positive coefficient sets were possible and measured performance was not shown.

The proposed DT filter in this paper is designed to improve the functionality of the AFIR filter using split-CDAC multipliers while maintaining low power and high linearity. The design is implemented in 32nm SOI CMOS technology. Figure 3.7.1 shows the block diagram of the proposed architecture. Five interleaved signal paths are employed to implement the 4-tap FIR filter. Each path is driven by the five non-overlapping clock waveforms (ϕ_1~ϕ_5) in a time-interleaved manner. For example, in path-1, the input signal is sampled and the output is reset at ϕ_1. During the next 4 clock phases (ϕ_2 to ϕ_5), the sampled signal is multiplied by the programmed 4 coefficients A, B, C, and D in a rotating sequence (rotating coefficient architecture). The design in [6] has 8 coefficient multipliers and complicated clock waveforms to implement a 4th-order FIR filter; the proposed design uses only 5 coefficient multipliers and 5 simpler clock waveforms, which helps to reduce overall size and power consumption. Input and output drivers are used for 50Ω matching for test purposes.

The coefficient multipliers are designed to be controlled individually and negative sign is available. Figure 3.7.2 depicts the bi-phase coefficient multiplier, which is implemented by adding sign select switch (D_7). This additional switch effectively provides 7b control for the coefficient values. Addition is implemented in the current domain using an NMOS gm-cell and diode-connected NMOS load for improved linearity. A current source is added to increase the load resistance, by decreasing the current through the 1/gm load, with a moderate impact on linearity. The adder is designed to have a 5GHz 3dB BW to prevent from BW loss and phase distortion up to the Nyquist rate.

The rotating coefficient operation is implemented by switching split-CDAC digital inputs, D_1-D_7, with the clock phases that are modulated by the coefficient control consisting of a total of 28 parallel digital outputs (=7b × 4 coefficients: D_{1A}, D_{1B}..., D_{7D}) from the shift register, as shown in Fig. 3.7.3. Suppose that LSB codes of the 4 coefficients are set to rotate in 1-0-1-0 order. The shift register is then programmed for a parallel output of D_{1A} - D_{1D} = '1010'; then by combining with ϕ_2 - ϕ_5 through the MUX assigned to D_1, the codes are serialized at the MUX output (D_1). In order to minimize timing skew, the ϕ_1 waveform goes through a delay (τ in Fig. 3.7.3) equivalent to that of the MUXs. The same holds true for the other switches (D_2 - D_7). The time-interleaved pattern is followed by all 5 paths that implement delay and coefficient multiplication (Fig. 3.7.3), and the 4th-order FIR filter processing is completed by summing of the outputs from each path.

The frequency response of the 4th-order AFIR filter is attained from S-parameter measurements with -10dBm input power. The effects of the input/output drivers are de-embedded, and the sinc function due to the output zero-order-hold is also compensated. In Fig. 3.7.4, CS#1 to CS#7 represent 7 example coefficient sets out of the possible type-II (even & symmetric) FIR coefficient sets to estimate the function as a programmable filter. Since each coefficient can be set independently, the coefficient sets can be also non-symmetric for other applications. To test the AFIR filter as an LPF function, every possible filter coefficient set is measured and 3dB BW and zero frequencies are evaluated, which match well with the calculations as shown in Fig. 3.7.4. It is also observed that the measured frequency responses with the selected LPF coefficient sets (CS#1 to CS#5) and BPF coefficient sets (CS#6 and CS#7) match well with the calculations. These results show that the proposed AFIR filter properly functions as both an LPF and a BPF.

Figure 3.7.5 shows the zero frequency step, 3dB BW step size and gain variation depending on the LPF coefficient sets. The worst case 3dB BW and zero frequency control step sizes are 46MHz and 35MHz, respectively, but most steps sizes are within 10MHz. The gain varies with the sum of coefficient values at low frequency. The difference between min and max is about 6dB. This gain variation can be avoided by matching the sum of coefficients to the minimum value, e.g. halving all coefficients of CS#3, then the gain becomes equal to CS#1. The accuracy of the 3dB BW, zero frequency and gain is related to the static linearity of the split-CDACs, but this does not impact the signal linearity.

The measured passband IIP3 of example coefficient sets is >11dBm. The linearity varies with the coefficient sets due to the coefficient-dependent gain variation before the adder and the adder linearity itself also depends on the amplitude from each filter tap. Input-referred noise (IRN) of each selected case is averaged over the passband. In the worst case, the LPF and BPF generate 54nV/√Hz and 28nV/√Hz of IRN, respectively. The dominant noise contribution is from the adder. Power consumption per each tap is <2mW. The variation of the power consumption is due to the variation of dynamic power of the rotating coefficient clocks. It is expected that the power consumption would increase proportionally to the number of paths, namely the FIR filter order. A summary of performance for specific coefficient sets and comparisons with recently published AFIR filters are shown in Fig. 3.7.6. The die micrograph of the fabricated chip is shown in Fig. 3.7.7.

In conclusion, a programmable discrete-time 4th-order FIR filter in 32SOI CMOS for analog signal processing is presented. With the rotating split-capacitor coefficient multiplier, frequency response is programmable with fine resolution within the FIR filter function. The advantages of the proposed design also include high linearity, small area and low power consumption. This analog FIR filter can be potentially used up to the Nyquist rate BW depending on the order of FIR filter.

Acknowledgements:
This work was funded in part by the Air Force Research Laboratory via Berrie-Hill Research Corp., prime contract # FA8650-10-D-1746/0006, and University of Dayton Research Institute, prime contract #FA8650-10-2-7028. The authors would like to thank Rohde & Schwarz for measurement equipment support.

References:

[1] M. Lehne, et al., "A 0.13-µm 1-GS/s CMOS Discrete-Time FFT Processor for Ultra-Wideband OFDM Wireless Receivers," *IEEE TMTT*, vol. 59, no. 6, pp. 1639-1650, June 2011.
[2] B. Malki, et al., "A 150 kHz–80 MHz BW Discrete-Time Analog Baseband for Software-Defined-Radio Receivers using a 5th-Order IIR LPF, Active FIR and a 10 bit 300 MS/s ADC in 28 nm CMOS," *IEEE JSSC*, vol. 51, no. 7, pp. 1593-1606, July 2016.
[3] M. Tohidian, et al., "A Fully Integrated Highly Reconfigurable Discrete-Time Superheterodyne Receiver," *ISSCC*, Feb. 2014.
[4] M. Neinhus, et al., "Finite Impulse Response-Filter-Based RF-Beamforming Network for Wideband and Ultra-Wideband Antenna Arrays," *IET Microwaves, Antennas & Propagation*, vol. 5, no. 7, pp. 844-851, May 2011.
[5] K.-J. Koh, et al., "Time-Interleaved Phased Arrays with Parallel Signal Processing in RF Modulations," *IEEE Trans. on Antennas & Propagation*, vol. 62, no. 2, pp. 677-689, Feb. 2014.
[6] S. Park, et al., "A 3.25 GS/s 4-Tap Analog FIR Filter Design with Coefficient Control Using 6-bit Split-Capacitor DAC as a Tunable Coefficient Multiplier," *IEEE DCAS*, Arlington, TX, pp. 1-4, 2016.

 978-1-5090-4940-0/18/$31.00

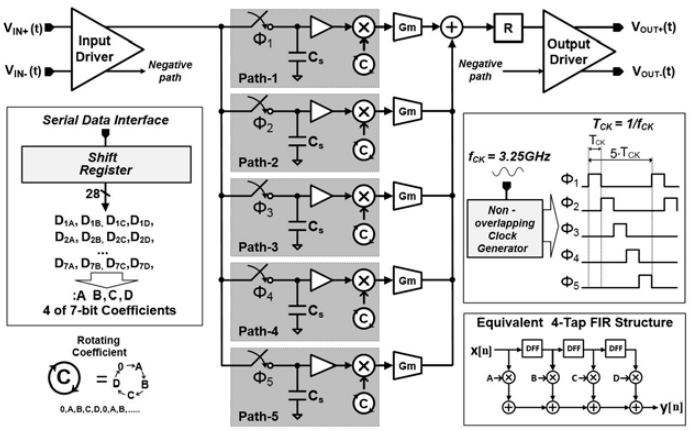


Figure 3.7.1: The proposed 4-tap analog FIR filter architecture.

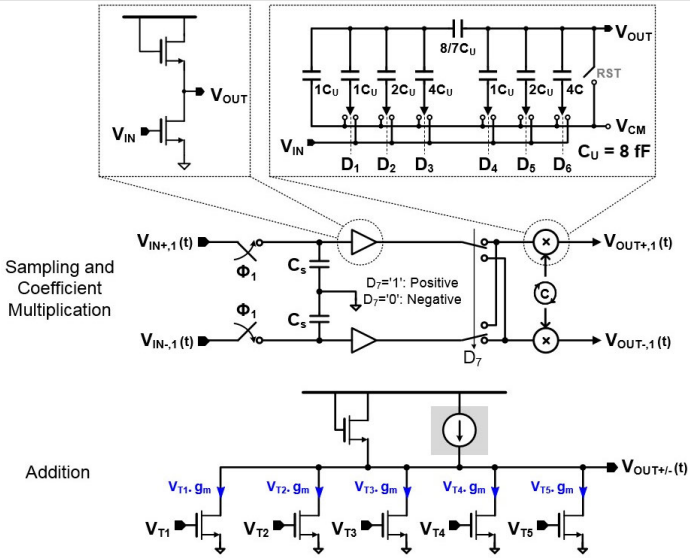


Figure 3.7.2: Fundamental building blocks of the proposed AFIR filter.

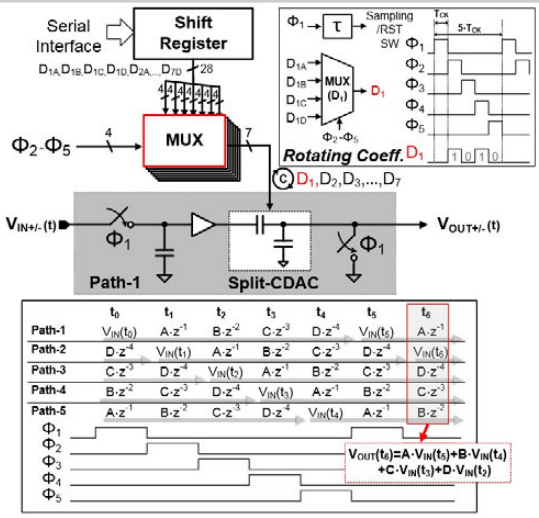


Figure 3.7.3: Implementation of sampling and coefficient rotation, and sequential delay of coefficient-multiplied signal in each path from t_0 to t_6 in the z-domain.

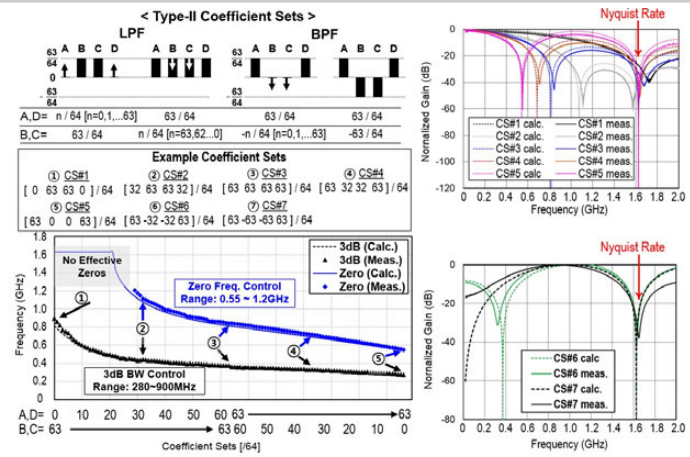


Figure 3.7.4: Description of possible type-II coefficient sets with 6b coefficients and 7 examples of coefficient sets (CS#1-to-CS#7), 3dB bandwidth/zeros in LPF coefficient sets, and simulated and measured frequency responses of LPF CSs (#1 ~ #5) and simulated and measured frequency responses of BPFs CSs (#6 and #7).

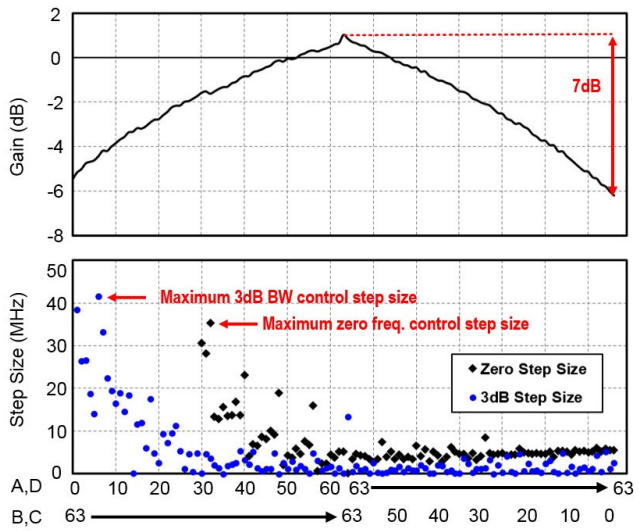


Figure 3.7.5: Measured gain, 3dB BW step size and zero frequency step size depending on the LPF coefficient sets.

	Coefficient Set (/64)	Feature	3dB BW/Center freq. (Hz)	Power* (mW)	Gain (dB)	IIP3 (dBm)**	IRN (nV/√Hz)
CS#1	0 – 32 – 32 – 0	LPF	900M	8.7	-5.5	14	35
CS#2	32 – 63 – 63 – 32	LPF	440M	8.4	-1.5	14	27
CS#3	63 – 63 – 63 – 63	LPF	380M	8.6	1.0	11	22
CS#4	63 – 32 – 32 – 63	LPF	330M	9.7	-2.0	14	32
CS#5	63 – 0 – 0 – 63	LPF	280M	10.0	-6.0	14	54
CS#6	63 – (-32) – (-32) – 63	BPF	1.07G	10.2	-3.6	16	28
CS#7	63 – (-63) – (-63) – 63	BPF	1.02G	9.5	-1.8	14	22

*Worst power consumption is 10.6mW in the extreme coefficient set case. S/H buffer and adder power is constantly about 3.9mW and clock power consumption varies with the coefficients. I/O drivers are excluded.
**From two tone tests with f_{in}=1MHz. (LPF@50MHz, BPF@center frequency). Measurement results including I/O drivers.

(a)

	ISSCC 2010 E. O'hAnnaidh, *et al.*	RWS 2011 D. Ahn, *et al.*	JSSC 2013 M.F. Huang, *et al.*	RFIC 2009 A. Yoshizawa, *et al.*	This Work
Technology	45nm	130nm	65nm	90nm	32nm
Supply (V)	1.1	1.2	1.2	1.2	0.9
Sampling Rate (GS/s)	3.2	0.36	0.48	2	3.25
# of Taps	16	12	12	6	4
Power (mW)	48	6	8.4	13.8	6.3 ~ 10.6***
Power per tap (mW)	3 (5.4*)	0.5	0.7	2.3	1.3 ~ 2
Gain	0	0	41	29	-6 ~ 1
IIP3 (dBm)	-	12	-19	-21	11~16
IRN (nV/√Hz)	SNDR = 33dB	36~149	12.3	4.11	LPF: 54, BPF: 28
Tunable BW	Y	N	Y	N	Y
Useful BW (Hz)	800M	10M	26M	250M	900MHz (LPF) 1.625GHz**
Comment	BW limited by Sinc function	Fixed BW	BW limited by Sinc function	Fixed BW	No BW limitation by Sinc function
Area(mm²)	0.15	0.23	0.52	0.19	0.78(Core: 0.10)

* Estimated power for non-decimation operation (10 → 18 signal paths)
**Potentially BW can cover up to Nyquist rate. ***Analog: 3.9mW/Digital: 2.4~6.3mW

(b)

Figure 3.7.6: Performance summary tables with (a) example coefficient sets, and (b) comparison with recently published analog FIR filters.

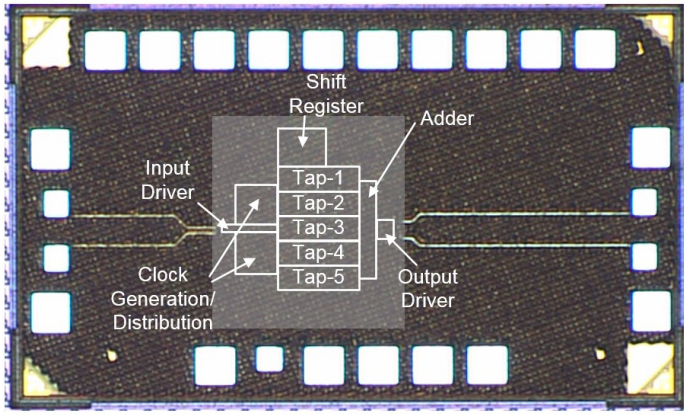


Figure 3.7.7: Die micrograph and core layout (Core area: 280×350mm²).

978-1-5090-4940-0/18/$31.00

Session 4 Overview: *mm-Wave Radios for 5G and Beyond*

WIRELESS SUBCOMMITTEE

Session Chair:
Chun-Huat Heng
National University of Singapore
Singapore

Associate Chair:
David McLaurin
Analog Devices, Raleigh, NC

Subcommittee Chair: ***Stefano Pellerano,*** *Intel, Hillsboro, OR*

Millimeter-wave beamforming and full-duplex techniques are increasingly important for 5G and next-generation radio systems. This session includes two papers describing architectures and techniques for 5G basestations, two massive MIMO papers with tileable phased-array chips scalable to hundreds of elements, an eight-element receiver supporting autonomous analog beam steering, a reconfigurable receiver for concurrent dual-band or multi-stream operation, and a full-duplex transceiver sharing a single self-interference-cancelling antenna.

INVITED PAPER

1:30 PM

4.1 Architectures and Technologies for the 5G mm-Wave Radio

T. Cameron, Analog Devices, Ottawa, Canada

The ever-increasing demand for mobile data continues to drive the expansion of mobile network capacity globally. As available spectrum becomes scarce in the traditional cellular bands, the industry looks to utilize the broad available spectrum in mm-wave bands. Whether the use case is fixed or mobile connectivity, it has been demonstrated that the challenging propagation characteristics at mm-wave frequency can be overcome through beamforming techniques.

In the Spectrum Frontiers mm-wave spectrum allocation, the FCC sets the EIRP limit for basestations at 75dBm/100MHz to enable anticipated deployments [1]. While this is the maximum EIRP limit, not all use cases demand such high power. As such a number of system architectures may be employed to achieve a power-efficient beamforming radio for a given deployment model. For smaller cell size, it may be beneficial to employ a very high level of integration and high antenna count with low power per RF chain, whereas for a larger cell it may be advantageous to employ high-power amplifiers and fewer RF chains. In the former case the gain of the antenna array is leveraged while in the latter case, power consumption may be optimized through linearization techniques.

In this presentation we will discuss two common radio architectures, one suitable for low EIRP and one for high EIRP. An efficient signal chain will be described for each case and optimized semiconductor technology choices will be discussed for each architecture. Finally we will briefly review how a fully digital beamforming approach compares to the above described analog approaches, and also review technology requirements to enable digital beamforming architectures for future mm-wave systems.

[1] "FACT SHEET: SPECTRUM FRONTIERS PROPOSAL TO IDENTIFY, OPEN UP VAST AMOUNTS OF NEW HIGH-BAND SPECTRUM FOR NEXT GENERATION (5G) WIRELESS BROADBAND", https://apps.fcc.gov/edocs_public/attachmatch/DOC-339990A1.pdf

 978-1-5090-4940-0/18/$31.00

2:00 PM

4.2 A 60GHz 144-Element Phased-Array Transceiver with 51dBm Maximum EIRP and ±60° Beam Steering for Backhaul Application

T. Sowlati, Broadcom, Irvine, CA

In Paper 4.2, Broadcom presents a 144-element phased-array transceiver using a tiled approach in 40nm CMOS for 802.11ad. It has 51dBm EIRP and supports scan angle of ±60° in azimuth and ±10° in elevation.

2:30 PM

4.3 A 23-to-30GHz Hybrid Beamforming MIMO Receiver Array with Closed-Loop Multistage Front-End Beamformers for Full-FoV Dynamic and Autonomous Unknown Signal Tracking and Blocker Rejection

M-Y. Huang, Georgia Institute of Technology, Atlanta, GA

In Paper 4.3, the Georgia Institute of Technology describes a 23-to-30GHz 8-element receiver autonomously creating spatial notches on multiple in-band blockers and performing beamforming on the desired signals over full FoV with 1-to-2µs dynamic response time. The desired signal shows -27.2dB EVM for 3Gb/s 64QAM and -33.9dB EVM for 0.8Gb/s 256QAM, after cancelling the blocker with the same modulation scheme and modulation rate.

3:15 PM

4.4 A 28GHz Bulk-CMOS Dual-Polarization Phased-Array Transceiver with 24 Channels for 5G User and Basestation Equipment

J. D. Dunworth, Qualcomm, San Diego, CA

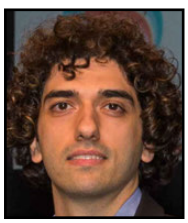
In Paper 4.4, Qualcomm proposes a 28nm, 28GHz CMOS phased-array transceiver supporting 12 elements each on 2 MIMO layers targeting 5G basestations. It attains TX P_{out} of 8dBm per element, 12% PAE including an integrated switch, and RX NF<5dB.

3:45 PM

4.5 A Reconfigurable 28/37GHz Hybrid-Beamforming MIMO Receiver with Inter-Band Carrier Aggregation and RF-Domain LMS Weight Adaptation

S. Mondal, Carnegie Mellon University, Pittsburgh, PA

In Paper 4.5, Carnegie Mellon University presents a reconfigurable multimode 28/37GHz 4-element hybrid beamforming receiver supporting either concurrent dual-band or multistream single-band operation. The receiver incorporates an LMS-like beam-steering adaptation technique, and achieves 35dB image rejection.

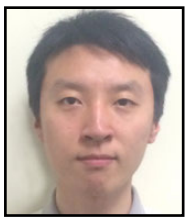


4:15 PM

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

S. Shahramian, Bell Laboratories, New Providence, NJ

In Paper 4.6, Bell Laboratories describes a scalable mm-wave transceiver with eight receive elements and sixteen transmit elements per tile in 0.18µm SiGe. The RFIC is capable of self-test and self-alignment and achieves >8dBm P_{sat} per element, <0.25W per TX or RX element, and is potentially scalable to >100 elements.



4:45 PM

4.7 A 64GHz Full-Duplex Transceiver Front-End with an On-Chip Multifeed Self-Interference-Canceling Antenna and an All-Passive Canceler Supporting 4Gb/s Modulation in One Antenna Footprint

T. Chi, Georgia Institute of Technology, Atlanta, GA

In Paper 4.7, the Georgia Institute of Technology proposes a 64GHz full-duplex transceiver sharing a single on-chip multifeed self-interference-canceling antenna. The multifeed antenna attains >35dB TX-to-RX isolation across 60 to 75GHz, and total front-end SIC is > 60dB across 63 to 65GHz.