EECS 598 VLSI for Wireless Communication and Machine Learning

Introduction

Prof. Hun-Seok Kim
hunseok@umich.edu

Practical Information

- Instructor
  - Prof. Hun-Seok Kim
  - 2406 EECS, hunseok@umich.edu
  - Office Hours: Mon 11am – noon, Tue/Thur 5 - 5:30pm @ 2406 EECS, or by appointment
- GSI
  - Yufan Yue (funkyyue@umich.edu)
  - Office hours: Wed/Thur 10am – 11am @ 2420 EECS
- Course organization
  - Tue/Thur 3:30 - 5pm, 1010 Dow
  - Lectures: 4/5 of the semester
  - Student presentations: 1/5 of the semester
- Course material
  - Textbook not required
  - Lecture slides and published papers
- Webpages
  - Canvas (https://umich.instructure.com)

Class Organization

- EECS 598 special topics
  - On advanced topics not covered in core courses
  - Mostly based on publications, not textbooks
    - Good to have: VLSI Digital Signal Processing Systems, by K. Parhi
  - Combination of lecture and independent study
- No midterm
- 4 bi-weekly homeworks
  - Typically involves Matlab (or any other your preferred language) and/or Verilog coding
- Project (groups of 4 students)
- Final exam: take-home mini project

Goals and Expectations

- Prereq: (EECS 351 and (EECS312 or EECS 370))
  - Familiarity with Matlab/Python and RTL/Verilog (won't be covered in lectures)
- Course workload is mostly tailored by yourself
  - HW won't be graded thoroughly
    - No unique solution for RTL and Matlab programming
  - Define your own term project scope
- Successful projects can lead to conference / journal publications
- Learn from your classmates (be sure to attend your classmates' presentations and provide good reviews)
- Bridge the gap between VLSI and signal processing, wireless comm, ML
  - VLSI students who had only learned microprocessor / accelerator designs → get to know interesting signal processing applications and algorithms
  - SP students who could only draw block diagrams → make real, functional systems in hardware

Course Goals

- Research
  - Initiate new research projects
  - Help advance your own research
  - Discover common interests for collaborations
- Develop the breadth and enlarge your comfort zone
  - VLSI ←→Signal Processing / Wireless Comm. ←→ Machine Learning
- Develop 'system' expertise

Lectures

- Emphasis of this term's lectures
  - Some lecture slides (including this introduction) are based on Prof. Zhengya Zhang's (U of M) EECS 598 and Naresh Shanbhag's (UIUC) ECE 560
  - Another reference is Prof. Keshab Parhi's textbook: VLSI Digital Signal Processing Systems: Design and Implementation
  - Digital signal processing for
    - Wireless communications
    - Signal processing
    - Machine learning
- Main topics
  - Digital circuit overview
  - Signal processing kernels
  - Systolic architecture
  - Wireless communication modem design
  - Deep neural network algorithms and architectures
  - Hardware architecture-friendly algorithms for signal processing

Grading Policy

- Bi-weekly homework: 30% (4 assignments)
- Take home exam: 30%
- Presentation: 10%
- Presentation reviews: 5%
- Project: 25%

- HW late penalty -25% per day
- This is an advanced graduate class: Grade should not be the sole motivation of your work
- Won't curve the grade

Homework and Exam Policies

- You can discuss HW with classmates but it should be done individually. Do NOT share your HW codes.
- Homework will NOT be thoroughly graded and returned. It is mostly for your own exercise. Homework solutions will not be posted (multiple possible solutions). Come to office hours or post questions on Canvas if you have any.
- Your peer presentation reviews will be graded and forwarded to presenters without reviewer information. You can select two presentations (out of four) per lecture to review.
- Final exam is a mini take-home project that involves HDL simulation and/or Matlab programming. You are NOT allowed to discuss it with classmates. Details will be posted near the end of the semester.

Project Requirements

- Groups of 4
- Find your topic and group early!
- Topic can be anything related to VLSI for wireless communication, signal processing, or/and machine learning
  - Examples in the next slide
- If you have an ongoing unfinished research work that is related to this course, it is ok to make it as your term project
  - In this case, your project needs to be 'completed' as a standalone course project
- Define your own project scope
  - HDL implementation, synthesis only without APR, or preliminary APR, …
  - Discuss with the instructor and GSI to define the right scope
  - Find the topic and scope by reading relevant papers and their references

Examples of Previous Project Topic

- MIMO-OFDM-Based Inverse-Free MMSE Detector With Gram- Schmidt QR Decomposition
- SCNN: An Accelerator for Compressed Sparse Convolutional Neural Networks
- Accelerator for Deconvolution in Generative Adversial Networks (GAN)
- Reconfigurable SISO FEC Decoder for Polar, LDPC and Turbo
- Low-Delay and Low-Cost Sigma-Delta Adaptive Controller for Active Noise Control
- SVD Systolic Array Implementation with CORDIC
- Bundle Adjustment for simultaneous localization and mapping (SLAM)
- 8-bit Floating-Point Systolic-Array based Transformer Accelerator with Local-Attention-Reuse
- Engine WiFi baseband modem design for secure distance ranging
- Accelerator for heterogeneous transform domain neural networks
- A Flexible-Mapping CNN Accelerator with Fast Winograd Algorithm

Project Requirements

- Pick a topic by surveying literature and read at least 5 relevant papers
  - Discover unknowns and check references
  - Teach the topic to the class
  - Cover both the fundamentals and the cutting-edge developments
- Presentation
  - Will be scheduled near the end of the semester
  - 20 min per group
  - Slides should be uploaded and distributed before presentation
  - Presentation will be before you complete the project. It is for a review of intermediate progress and results.
  - Includes literature summary, introduction/motivation, technical novelty/challenges, design details, and status
- Attend presentations (mandatory)
  - Learn new areas
  - Ask questions to make sure you understand the topic and help your classmates improve
  - Pick two presentations per lecture session to review
  - Ask questions and provide well thought-out reviews
  - Reviews will be sent to presenters (without reviewer identity)
- Project report
  - Includes introduction, motivation, technical novelty/challenges, design details, and evaluation results.
  - Double-column conference paper format is recommended
  - No strict length limitations: $5 - 8$ pages per group recommended (excluding bibliography and appendix)
  - Take the reviews in consideration

Tools for HW and Project

- Matlab (any other preferred language) for signal processing algorithm evaluation and modeling
  - No tutorial will be provided
  - Optimized for vector / matrix processing
  - Minimize using 'for' loops …
- Verilog HDL (Hardware Description Language)
  - Tutorial material will be posted but no lectures on the tutorial
  - Logic synthesis and simulation on CAEN machines
- Project design scope can vary
  - HDL implementation + synthesis + (optional) APR
  - Optional tools (available upon request)
    - PDK for APR
    - Cadence support for transistor level design
    - FPGA programming / demonstration

References for literature search

- No textbook required
- Lecture slides will be distributed
  - Slides will only cover summary points
- Read well-cited papers from leading journals and conference proceedings
  - Access them from IEEE Xplore http://ieeexplore.ieee.org (accessible by on-campus networks or UM VPN)
  - Try Google scholar (https://scholar.google.com/)
  - Check references or references of references in papers
- **Use well-cited reference papers in good journal / conferences** to deeply understand the lecture material and to find project topics

References for literature search

- Relevant journals
  - Journal of Solid-State Circuits (JSSC)
  - Journal on Selected Areas in Communications (JSAC)
  - Transactions on Very Large Scale Integration Systems (TVLSI)
  - Transactions on Circuits and Systems (TCAS)
  - Transactions on Circuits and Systems for Video Technology (TCASVT)
  - Transactions on Signal Processing (TSP)
  - Transactions on Wireless Communication (Twireless)
  - Journal of Signal Processing Systems (JSPS)
  - Transactions on Pattern Analysis and Machine Intelligence (TPAMI)
  - …
- Relevant conference proceedings
  - International Solid-State Circuits Conference (ISSCC)
  - Symposium on VLSI Circuits (VLSI)
  - International Symposium on Computer Architecture (ISCA)
  - International Conference on Mobile Computing and Networking (Mobicom)
  - International Symposium on Circuits and Systems (ISCAS)
  - International Conference on Acoustics, Speech, and Signal Processing (ICASSP)
  - International Symposium on Microarchitecture (Micro)
  - Conference of the ACM Special Interest Group on Data Communication (SIGCOMM)
  - International Symposium on High-Performance Computer Architecture (HPCA)
  - International Conference on Machine Learning (ICML)
  - Conference on Computer Vision and Pattern Recognition (CVPR)
  - International Conference on Learning Representations (ICLR)
  - …

Course Schedule (tentative)

Lecture	Date	Topic	HW due
1	8/27 Tu	Introduction	
2	8/29 Th	Digital Integrated Circuits Review	
3	9/3 Tu	Digital Integrated Circuits Review	
4	9/5 Th	Fixed point, Sampling, Quantization	
5	9/10 Tu	Fixed point, Sampling, Quantization	
6	9/12 Th	DFG, pipelining	
7	9/17 Tu	DFG, pipelining	
8	9/19 Th	Parallel processing, strength reduction, fast convolution	
9	9/24 Tu	Parallel processing, strength reduction, fast convolution	HW1
10	9/26 Th	Systolic array	
11	10/1 Tu	CORDIC, matrix operations	Project Proposal
12	10/3 Th	CORDIC, matrix operations	
13	10/8 Tu	----- Kim out -----	
14	10/10 Th	Signal transformation	HW2
	10/15 Tu	----- Fall study break ----	

Course Schedule (tentative)

15	10/17 Th	Signal transformation	
16	10/22 Tu	Machine learning, deep neural network	
17	10/24 Th	Machine learning, deep neural network	HW3
18	10/29 Tu	Machine learning, deep neural network	
19	10/31 Th	Wireless communication systems and signal processing	
20	11/5 Tu	Wireless communication systems and signal processing	
21	11/7 Th	Source and channel coding	HW4
22	11/12 Tu	Students Presentations	
23	11/14 Th	Students Presentations	
24	11/19 Tu	Students Presentations	
25	11/21 Th	Work on the project (no lectures)	
26	11/26 Tu	Work on the project (no lectures)	
	11/28 Th	---- Thanksgiving	
27	12/3 Tu	Work on the project (no lectures)	

Project due: Dec 6th Friday midnight
Final exam due: Dec 13th Friday midnight

Where do DSPs go?

- Communications and networking
  - Modulation/demodulation
  - Encoding/decoding
  - Encryption/decryption
  - Detection, estimation
  - …
- Data processing, multimedia and consumer
  - Natural language / speech processing
  - Image, video, graphics processing
  - Perception and recognition deep learning
- Applications
  - Mobile phones, augmented / virtual reality
  - IoT devices, wearable devices
  - Smart TV, smart audio boxes
  - Robots, drones
  - Industrial applications
  - …

What is special about VLSI for DSP?

- From a sales' point of view
  - Very large volume, performance critical, and extremely cost sensitive and power limited
- From application designer's point of view
  - Well-defined algorithm (fixed for a standard)
  - Real-time throughput requirement (new samples need to be processed as received)
- Implementation of DSP
  - Microprocessors / Microcontrollers: low performance, high overhead
  - Multi-core mobile application processors: flexibility, quick turn-around, per-unit cost is still high
  - FPGA: same as above, but custom architecture enables a higher performance and efficiency
  - Application specific integrated circuit (ASIC): less flexibility, longer turn-around, fixed NRE cost (which is high), per-unit cost is low, custom architecture enables the highest performance and efficiency
- DSP VLSI design metrics
  - **Functional features**
  - **Silicon area, speed (performance), power/energy**

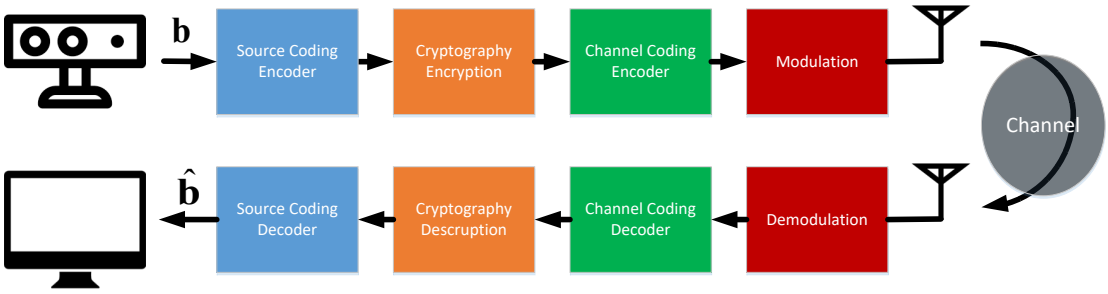
What is special about VLSI ASIC for DSP?

- Application-centric
  - New applications constantly emerging (e.g., IoT sensor signal processing, biomedical signal processing, machine learning), providing more opportunities for innovation
- Custom architecture
  - A much larger architecture exploration space (beyond Patterson and Hennessey)
  - Higher efficiency is often possible for data-driven designs (stream architecture, massively parallel architecture, neuromorphic architecture)
- New opportunities for cross-layer designs
  - Algorithm-architecture-circuit co-design
- Extends your career horizon

Communication System-on-Chips (SoCs)

- Explosive growth in communication SoCs
  - All the necessary ingredients: advances in semiconductors (substrate), computing (logic and architecture), and communication theory (application)
- Semiconductor revolution
  - Invention of transistors (1948), integrated circuits (1965)
  - Moore's Law (1968)
- Computing revolution
  - Turing machine (1937)
  - Von Neumann architecture (1945)
- Communication theory
  - Channel capacity theory (Shannon, 1948)
  - Birth of information theory

Wireless Communication System

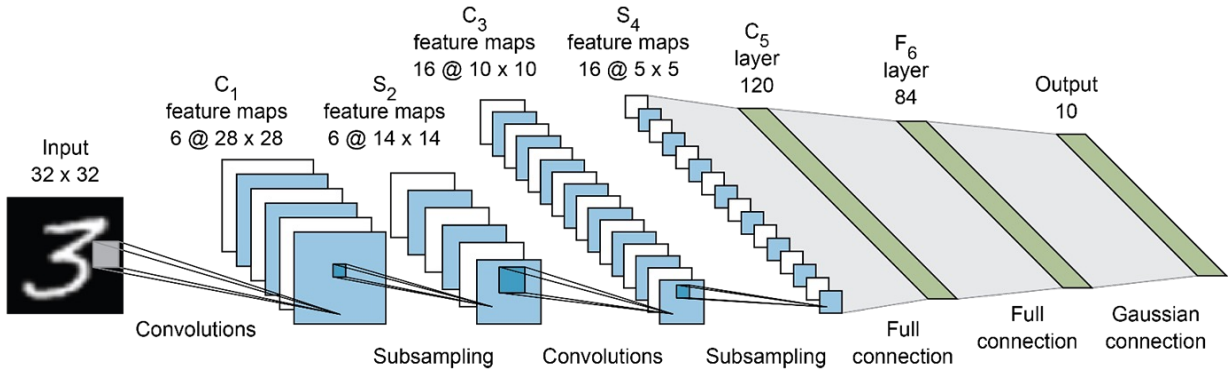


- Processing order matters!
- This course will (briefly) cover all blocks
- Goal: reliable information transfer
  - Maximize the probability $b = \hat{b}$

Wireless Communication System

- Source encoding/decoding
  - Remove redundancy in information for efficient transmission
  - Examples: H.264 / H.265 (HEVC), MPEG, JPEG
  - Digital processing. Encoding is more complex than Decoding.
- Encryption / Description
  - Encrypts / authenticate message for secure communication
  - Examples: AES, ECC
  - Digital processing. Encryption and decryption have similar complexity
- Channel encoding/decoding
  - Add redundancy efficiently to protect information against errors
  - Examples: convolution code, turbo, LDPC, polar code
  - Digital processing. Decoding is more complex than Encoding.
- Modulation/demodulation
  - Map coded bits or symbols to waveforms propagate through physical channels
  - Mixed-signal (RF, analog, digital) implementation
  - Demodulation is more complex than modulation

Deep Neural Networks



Many traditional algorithms are being replaced by deep learning based approaches

- Computer vision
- Natural language processing
- Contents creation
- Medical, finance, infrastructure, industrial applications
- Wireless communication, etc.

Deep Neural Networks

- Deep neural network is not a recently discovered algorithm and it has been unpopular for decades
- Why has the DNN/AI booming started just recently?
- Key ingredients for highly functional DNNs
  - High performance, highly efficient processors to expedite data- and computation-intensive training and inference with lower energy footprint
  - Large datasets for training
- Role of the VLSI/DNN algorithm researchers / engineers for future DNNs/AI models?
- Goal of this course is to educate such researchers / engineers

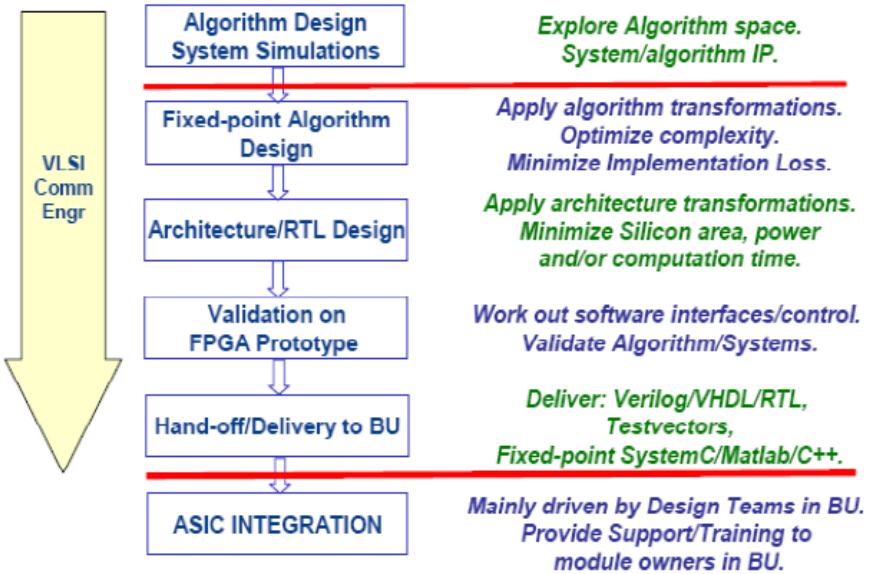
Recent Trends

- Advances in communication and signal processing algorithms
  - Emerging technologies: deep learning, 5G/6G, ultra-low power IoT
  - Newer standards, higher system performance, better energy efficiency
  - Data driven, deep learning-based approaches replacing hand-crafted algorithms
- Advances in semiconductor technology
  - Technology scaling, more opportunities to implement more complex functions at lower cost
  - Data-driven deep learning became possible due to semiconductor (GPU) technology advancement
- Design challenges
  - Mismatch between algorithm specification and IC design constraints
  - Higher complexity, lower power, more difficult to implement
  - Memory wall, interconnect bottleneck, power consumption, reliability problems

VLSI DSP Design

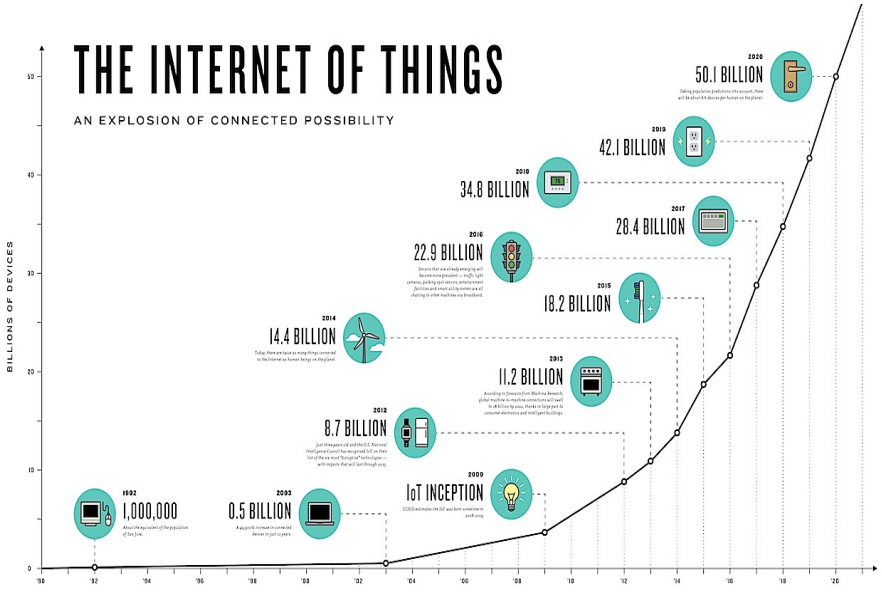
- Systematic techniques for the design of VLSI architectures for DSP
  - Understand the link between DSP algorithms, architectures and circuits
  - Systematic mapping of algorithms to VLSI architectures to achieve high performance and low power
  - A system-level, vertically-integrated design

Vertical Design Methodology



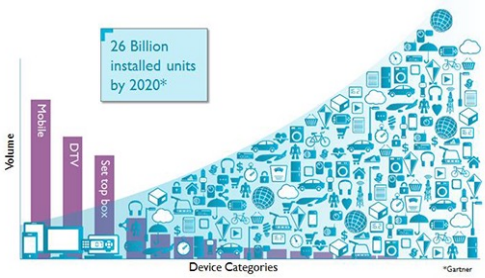
Courtesy: D. Shaver, Texas Instruments

Internet-of-Things (IoT) Opportunity



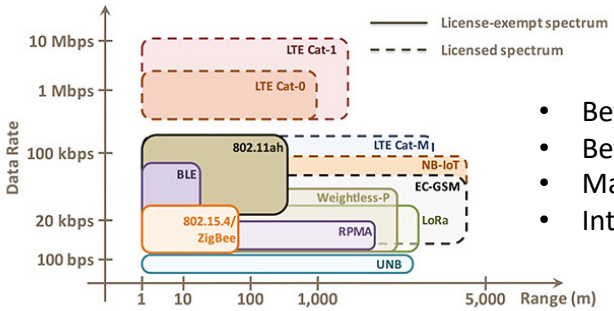
Source: https://www.ncta.com/platform/industry-news/infographic-the-growth-of-the-internet-of-things/ (based on Cisco data)

IoT Challenges: Communication



- Billions of Heterogeneous IoT Nodes
- Extreme variation of use cases

[Source: www.arm.com]



- Better energy efficiency
- Better adaptability & scalability
- Massively concurrent connection
- Interoperability, deep coverage

Source: LoRa alliance, iotbusinessnews.com

IoT Challenges: Information/Signal Processing

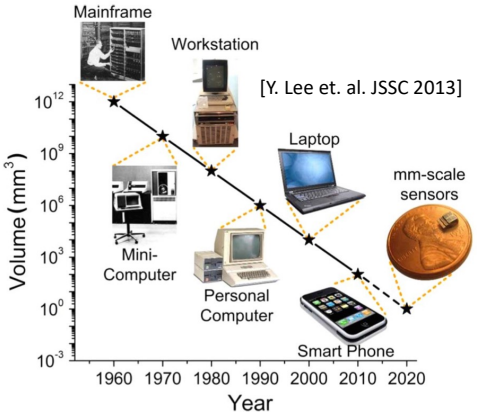
Distributed Information / Sensing

Centralized Processing

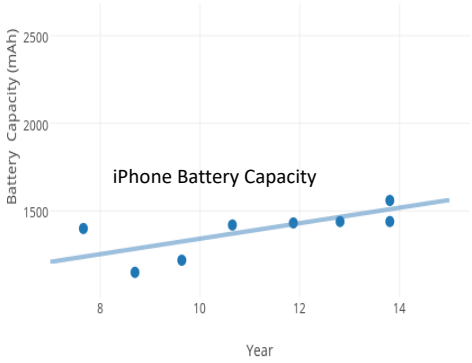
- Location mismatch between data and processing
  - when computing resource centric (cloud centric) architecture is adopted
- Communication overhead dominates overall power and latency
- Lack of local intelligence / data processing limits IoT usability
- Need to bring computing and intelligence closer to the data source and user

IoT Challenges: Energy Problem in the IoT Era

Exponential System Volume Scaling



Linear Battery Capacity Increase



- To Enable Ubiquitous IoT
  - Semi-perpetual or energy autonomous operation is desired
  - Improving energy efficiency is a major challenge

Deep Learning Energy Crisis

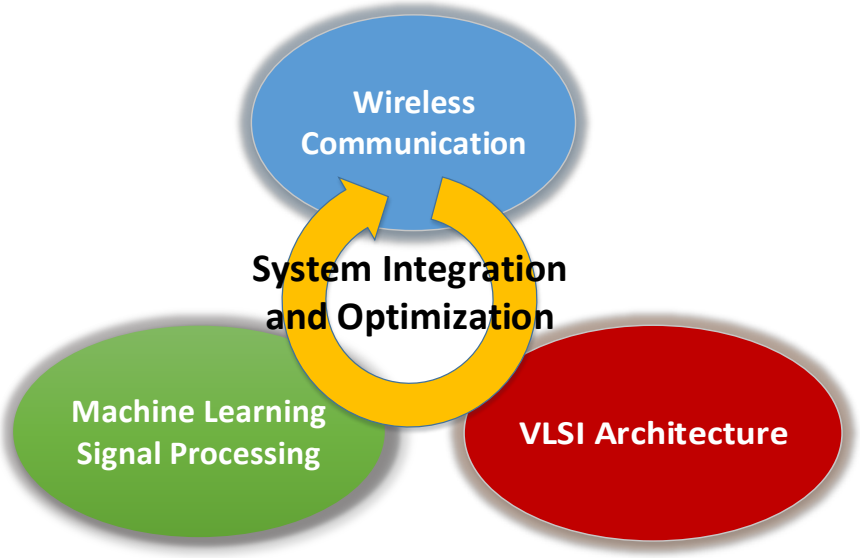
- As of 2024, Data centers consume about 3 percent of the world's electricity
  - This can be doubled by 2030

- ChatGPT4 training requires 50 Giga Watt-hours energy
  - 25,000 A100 GPUs for 100 days
  - $100M total energy cost
  - This does not include inference energy to use the model after training

Source: https://www.forbes.com/sites/arielcohen/2024/05/23/ai-is-pushing-the-world-towards-an-energy-crisis/

- **Reducing energy foot-print is a major challenge for future deep learning applications**

How to address IoT challenges: cross-optimization



Related Research

- **Computer vision**
  - Real-time 6D-vision (3D-depth + 3D-motion) accelerator

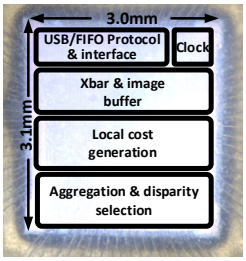
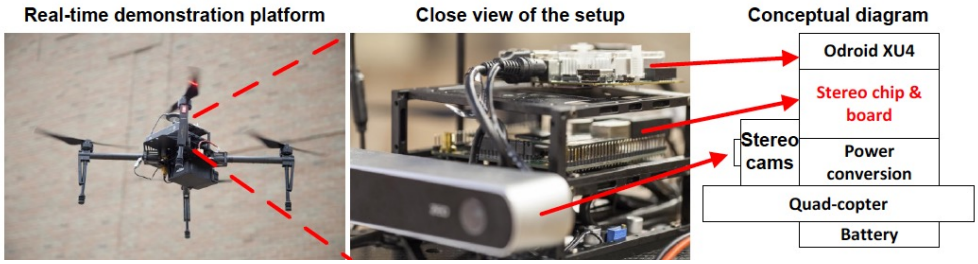
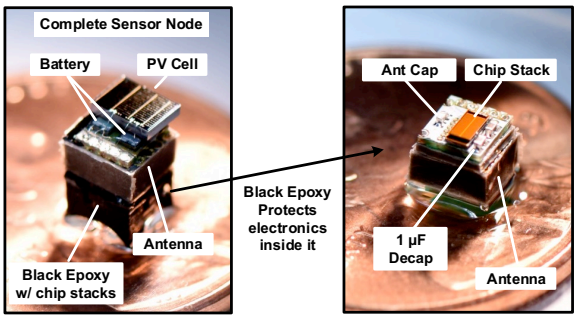


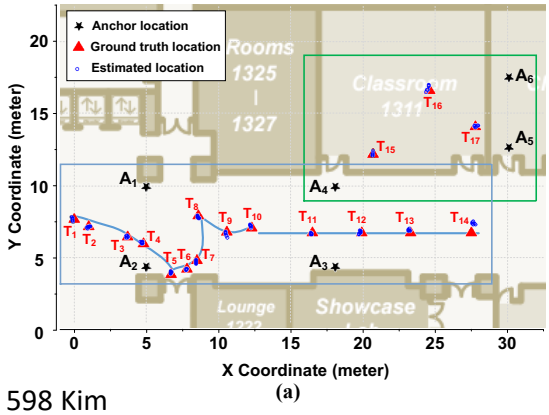
Fig. 9 Die photo



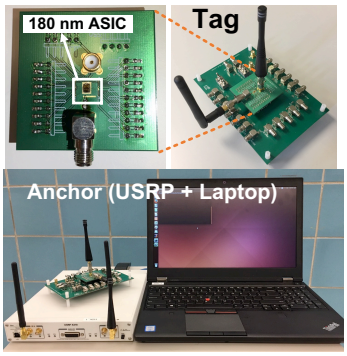
Related Research

- **Wireless communication for mm-scale Michigan Micro-Mote (M³) IoT platforms**
- **RF-Echo: Decimeter-accurate indoor non-line-of-sight localization system**



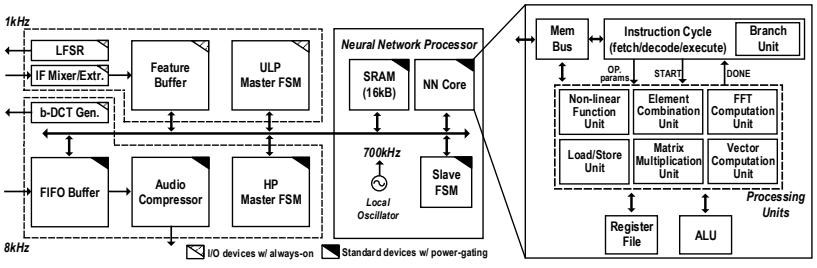


(a)



M³ Audio System: 140nW Voice Activity Detection

- Serialized DCT feature extraction
- Custom-designed ultra-low power neural network processor
- 140 nW for always-on voice activity detection and acoustic event detection



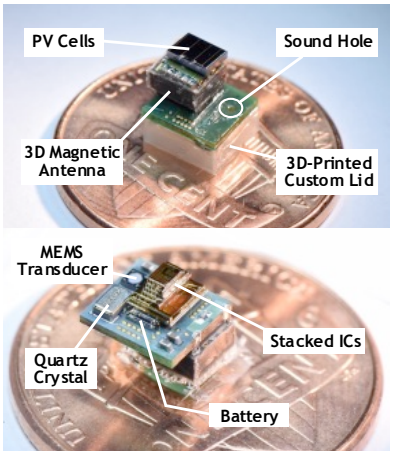
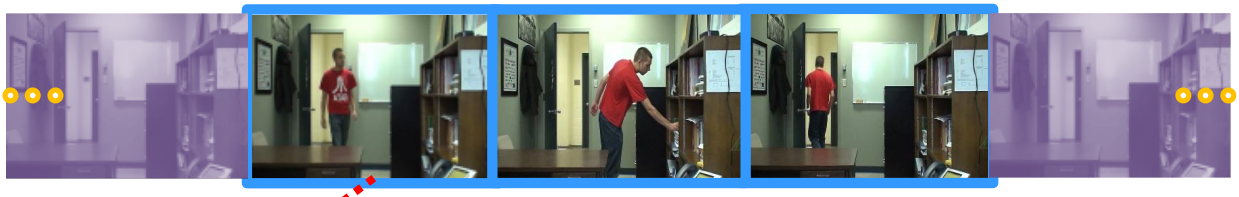
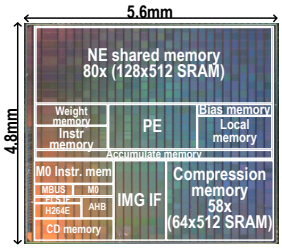


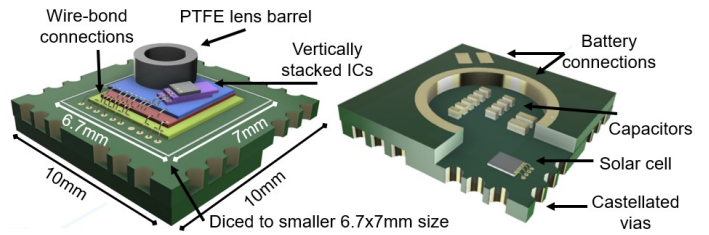
Image Signal Processor for Edge Intelligence

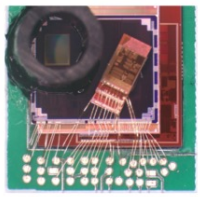


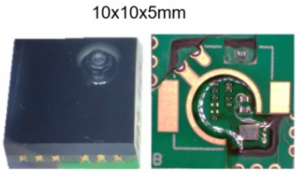


Process	40nm
Chip size	4.8mm x 5.6mm
On-chip SRAM	9Mbit
Supply voltage	0.58V~0.7V(Core) 0.29V~0.35V(Memory) 1.2V(IO)
Image size	VGA/ User-programmable subsampled VGA / 32x20
Throughput	5fps/ 0.28fps/ 0.16fps
Frequency	153kHz
Power	170μW

Image Signal Processor for Edge Intelligence

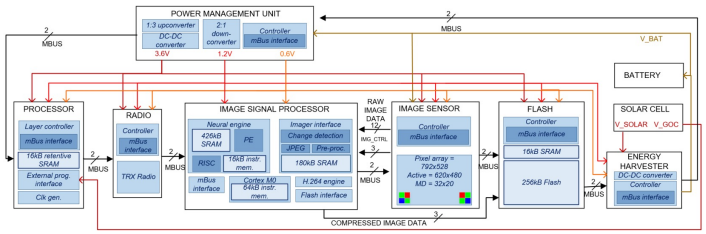


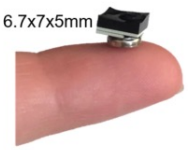




(a) Stacked IC Layers

(b) Encapsulated System





(c) Standalone System

(d) Test Setup