ENGRI 1210
Recent Trends in Computer Engineering
Christopher Batten

School of Electrical and Computer Engineering
Cornell University

(stay tuned for two exciting announcements at end!)
Gap too large to bridge in one step (but there are exceptions, e.g., a magnetic compass)
In its broadest definition, computer engineering is the development of the abstraction/implementation layers that allow us to execute information processing applications efficiently using available manufacturing technologies.
In its broadest definition, computer engineering is the development of the abstraction/implementation layers that allow us to execute information processing applications efficiently using available manufacturing technologies.
Cornell Computer Engineering Curriculum

- Application
- Algorithm
- Programming Language
- Operating System
- Compiler
- Instruction Set Architecture
- Microarchitecture
- Register-Transfer Level
- Gate Level
- Circuits
- Devices
- Technology

ECE 2400 Computer Systems Programming
ECE 3140 Embedded Systems
ECE 4760 Design with Microcontrollers
ECE 4750 Computer Architecture
ECE 2300 Digital Logic & Computer Org
Cornell Computer Engineering Curriculum

Freshman
- ECE 1210
- CS 111x

Freshman/Sophomore
- ECE 2100
- ECE 2300
- ECE 2400
- CS 2210

Sophomore/Junior
- ECE 3150
- ECE 3400
- ECE 3140
- CS 3420

Junior/Senior M.Eng. Ph.D.
- ECE 4740
- ECE 4760
- ECE 4750
- CS 4420

Senior M.Eng.
- ECE 5745
- ECE 5760
- ECE 4750
- CS 4420

Senior M.Eng. Ph.D.
- ECE 5730
- ECE 5770
- ECE 5775
- ECE 5725

Core courses
- VLSI track
- Architecture track
- Systems track

Prerequisite
- Pre- or corequisite
Agenda

The Computer Systems Stack

Trends in Computer Engineering

Hardware Acceleration for Deep Learning
Three Key Trends in Computer Engineering

Trend #1: Growing Diversity in Applications and Systems

Students entering the field of computer engineering have a unique opportunity to shape the future of computing and how it will impact society

Trend #2: Software/Arch Interface Changing Radically

Trend #3: Technology/Arch Interface Changing Radically
Bell’s Law

Roughly every decade a new, smaller, lower priced computer class forms based on a new programming platform resulting in entire new industries.
**M3: Michigan Micro Mote**

![Diagram and cross-sectional view of proposed 1.0 mm sensing platform.](image)

**Fig. 2.** Diagram and cross-sectional view of proposed 1.0 mm sensing platform.

Means to protect the battery during the time period between system assembly and deployment, wireless synchronization and batch programming, and flexible design to enable use in multiple application domains. A key challenge to implementing such features in a very small form factor is the additional power consumption incurred. To achieve long lifetime in a mm-scale system, power consumption of each component must be aggressively reduced to stay within a stringent power budget of 10 nW.

To this end, we propose a 1.0 mm general purpose reconfigurable sensor node platform with a heterogeneous stackable multi-layer structure. The key components implemented to realize this form-factor includes ultra-low power CMOS (Integrated Circuit), a 228 pW standby power optical wakeup receiver, ultra-low power power management unit (PMU) and brown-out detector (BOD).

**II. SYSTEM OVERVIEW**

The 1.0 mm sensing platform is designed with stacked integrated circuit (IC) dies fabricated in three different technologies. Fig. 2 shows the dimension of each die and the wirebonding scheme for electrical connectivity of the sensor system. To enforce 1.0 mm volume, each layer measures less than 2.21 mm and the length of each layer has to be reduced by 140 μm compared to the lower layer to provide enough clearance for bond-wires. The height of each IC layer is thinned to 50 μm, while the custom thin-film Li battery is 150 μm thick. The system’s die-stacked structure with wire-bonding provides maximum functionality (or silicon area) per unit volume and also enables easy expansion of the system with additional layers. End users can create a sensor system for new applications by designing an application-specific layer in a preferred technology, which complies with the system power and energy budget, and providing an identical inter-layer communication interface.

**Fig. 2 shows the system block diagram.** The various components in the system are categorized as CPU, memory, power management, timer, and sensors.

**CPU**

The sensor system operation sequence is managed by a control microprocessor, which requires low computational performance and hence can be optimized for low power operation. However, some sensors, such as an imager, require high performance for digital signal processing (DSP) operations. For this reason, two ARM® Cortex-M0 processors are located in separate layers with different functionality as follows:

1) The DSP CPU efficiently handles data streaming from the imager (or other sensors), thus is built in 65 nm CMOS (Layer 3) with a large 16 kB non-retentive SRAM (NRSRAM). In such an advanced technology node, the DSP CPU runs faster than the control CPU (fabricated in 180 nm) and accommodates the larger memory capacity that is required for complex DSP operation. However, due to the high leakage current in this process, the SRAM has to be power-gated in standby mode and is non-retentive.

2) The CTRL CPU manages the system using an always-on 3kB retentive SRAM (RSRAM) to maintain the operating program, and is built in low leakage 180 nm CMOS (Layer 4).

Adapted from Y. Lee et al., JSSC, 2013.
Three Key Trends in Computer Engineering

Trend #1: Growing Diversity in Applications and Systems

Students entering the field of computer engineering have a unique opportunity to shape the future of computing and how it will impact society.

Trend #2: Software/Arch Interface Changing Radically

Trend #3: Technology/Arch Interface Changing Radically
Activity: Specifications of Modern Processors

http://tiny.cc/engri1210-2

1. Breakout into groups of 3 students
2. Browse WikiChip
3. Find a few processors
4. Enter year, frequency, core count, power in Google form
Trends in High-Performance Processors

Data collected by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, C. Batten
Parallelization & Specialization Are Now Critical

Data collected by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, C. Batten
Celerity System-on-Chip

UCSD, Washington, Cornell, Michigan w/ DARPA CRAFT Program

- 5 × 5mm in TSMC 16 nm FFC
- 385 million transistors
- 511 RISC-V cores
  - 5 Linux-capable Rocket cores
  - 496-core tiled manycore
  - 10-core low-voltage array
- 1 BNN accelerator
- 1 synthesizable PLL
- 1 synthesizable LDO Vreg
- 3 clock domains
- 672-pin flip chip BGA package
- 9-months from PDK access to tape-out
Three Key Trends in Computer Engineering

Trend #1: Growing Diversity in Applications and Systems

Students entering the field of computer engineering have a unique opportunity to shape the future of computing and how it will impact society.
Technology Scaling is Slowing


System Performance

Vacuum Tube
Discrete Transistor
Integrated Bipolar
Integrated CMOS

7nm, ~50B Transistors

New Technologies
- Vertical MOSFETs
- Graphene
- Carbon Nanotubes
- Nanorelays
- Quantum Computing
- Molecular Computing
- Memristers
- Phase-Change Mem
- Spintronics
- 3D Integration
- Nanophotonics

Adapted from D. Brooks Keynote at NSF XPS Workshop, May 2015.
Three Key Trends in Computer Engineering

Trend #1: Growing Diversity in Applications and Systems

Students entering the field of computer engineering have a unique opportunity to shape the future of computing and how it will impact society.
Agenda

The Computer Systems Stack

Trends in Computer Engineering

Hardware Acceleration for Deep Learning
Image Recognition
Training vs. Inference

Training

many images

Inference

few images

Model

forward "starfish"

backward error

labels

"dog"

Model

forward "dog"

labels

"dog"
ImageNet Large-Scale Visual Recognition Challenge

Software: Deep Neural Network

Hardware: Graphics Processing Units

Top 5 Error Rate

- '10: 28%
- '11: 26%
- '12: 16%
- '13: 14%
- '14: 12%
- '15: 7%
- '16: ~100%
- '17: Human Error Rate

Entries Using GPUs

- '10: 0%
- '11: 0%
- '12: 14%
- '13: 74%
- '14: 89%
- '15: ~100%
- '16: 3.6%
- '17: 3%
ML Hardware Acceleration in the Cloud

**NVIDIA DGX-1**
- Graphics processor specialized just for machine learning
- Available as part of a complete system with both the software and hardware designed by NVIDIA

**Google TPU**
- Custom chip specifically designed to accelerate Google's TensorFlow C++ library
- Tightly integrated into Google's data centers
- 15–30× faster than contemporary CPU and GPUs

**Microsoft Catapult**
- Custom FPGA board for accelerating Bing search and machine learning
- Accelerators developed with/by app developers
- Tightly integrated into Microsoft data center’s and cloud computing platforms
ML Hardware Acceleration at the Edge

**Amazon Echo**
- Developing AI chips so Echo line can do more on-board processing
- Reduces need for round-trip to cloud
- Co-design the algorithms and the underlying hardware

**Facebook Oculus**
- Starting to design custom chips for Oculus VR headsets
- Significant performance demands under strict power requirements

**Movidius Myriad 2**

---

**Trends in Computer Engineering**
- Hardware Xcel for Deep Learning

---

**The Computer Systems Stack**
ML Acceleration Can Incorporate All Three Trends

- **ISAAC**: Convolutional neural network accelerator which uses in-situ analog arithmetic in crossbars of emerging resistive memory devices
- **Captures all three trends**
  - New applications and systems in ultra-low-power TinyML
  - New software/architecture interface for accelerator
  - New technology/architecture interface with non-traditional devices

---

Adapted from A. Shafiee et al., ISCA, 2016.
Top-five software companies are all making chips

- Facebook: w/ Intel, in-house AI chips?
- Amazon: Echo, Oculus, networking chips
- Microsoft: Hiring for AI chips?
- Google: TPU, Pixel, convergence?
- Apple: SoCs for phones, wireless chips

Chip startup ecosystem for machine learning is thriving!
- Graphcore
- Nervana
- Cerebras
- Wave Computing
- Horizon Robotics
- Cambricon
- DeePhi
- Esperanto
- SambaNova
- Eyeriss
- Tenstorrent
- Mythic
- ThinkForce
- Groq
- Lightmatter
Take-Away Points

- We are entering an **exciting new era of computer engineering**
  - Growing diversity in applications & systems
  - Radical rethinking of software/architecture interface
  - Radical rethinking of technology/architecture interface

- This era offers tremendous challenges and opportunities, which makes it a **wonderful time to study and contribute to the field of computer engineering**
ECE 2400 Computer Systems Programming

- **Part 1: Procedural Programming**
  - introduction to C, variables, expressions, functions, conditional & iteration statements, recursion, static types, pointers, arrays, dynamic allocation

- **Part 2: Basic Algorithms and Data Structures**
  - lists, vectors, complexity analysis, insertion sort, selection sort, merge sort, quick sort, hybrid sorts, stacks, queues, sets, maps

- **Part 3: Multi-Paradigm Programming**
  - transition to C++, namespaces, flexible function prototypes, references, exceptions, new/delete, *object oriented programming* (C++ classes and inheritance for dynamic polymorphism), *generic programming* (C++ templates for static polymorphism), *functional programming* (C++ functors and lambdas), *concurrent programming* (C++ threads and atomics)

- **Part 4: More Algorithms and Data Structures**
  - trees (binary trees, binary search trees), tables (lookup tables, hash tables), graphs (DFS, BFS, shortest path first, minimum spanning trees)
ECE 2400 Computer Systems Programming

- **PA1–3: Fundamentals**
  - PA1: Math functions
  - PA2: List and Vector Data Structures
  - PA3: Sorting Algorithms

- **PA4–5: Handwriting Recognition System**
  - PA5: Linear vs. Binary Searching
  - PA5: Trees vs. Tables

- **Every programming assignment involves**
  - C/C++ “agile” programming
  - State-of-the-art tools for build systems, version control, continuous integration, code coverage
  - Performance measurement
  - Short technical report
The Computer Systems Stack

Trends in Computer Engineering

Hardware Xcel for Deep Learning

Application-Level Software

System-Level Software

Engri 1210

Recent Trends and Applications in Computer Engineering
Do I have to wait to really build a chip?
C2S2: Cornell Custom Silicon Systems Project Team

Three-year student-led project team to tapeout a custom chip in SkyWater 130nm to implement a proof-of-concept system for a campus partner.

User’s Project
~10 mm²
(2.92mm x 3.52mm)
The C2S2 project team is unique across the country!

Email cbatten@cornell.edu for more information.