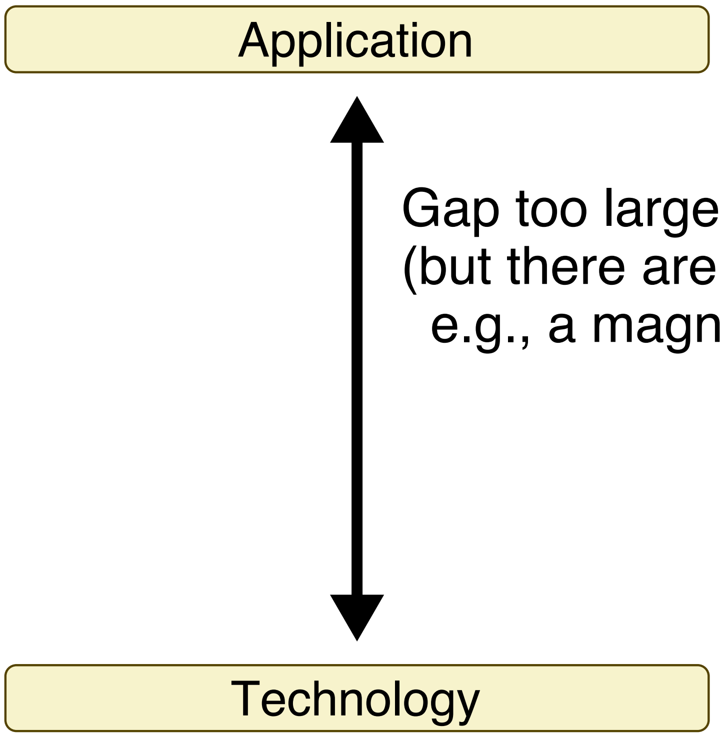
ENGRI 1210
Recent Trends in Computer Engineering

Christopher Batten

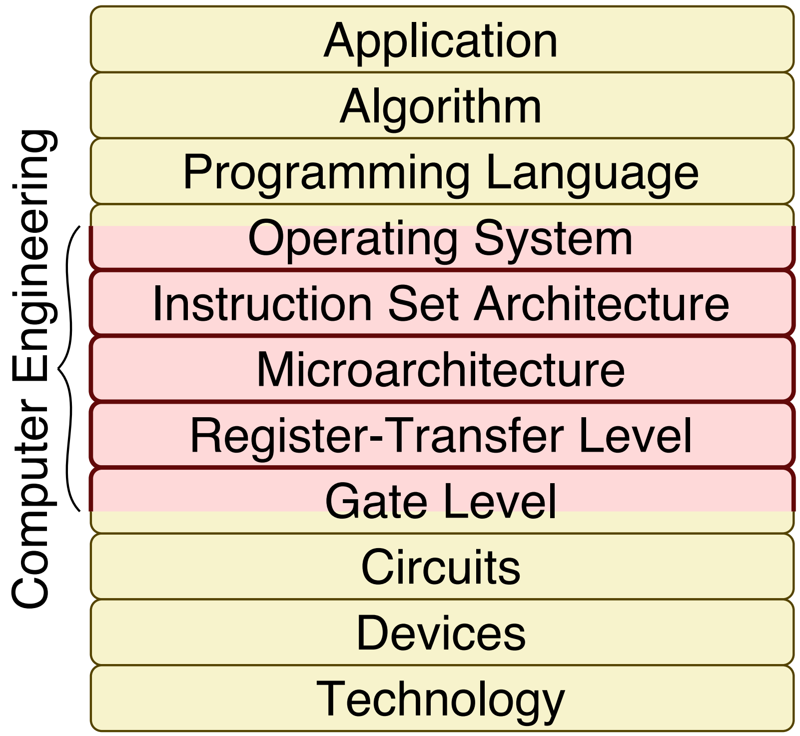
School of Electrical and Computer Engineering
Cornell University

The Computer Systems Stack



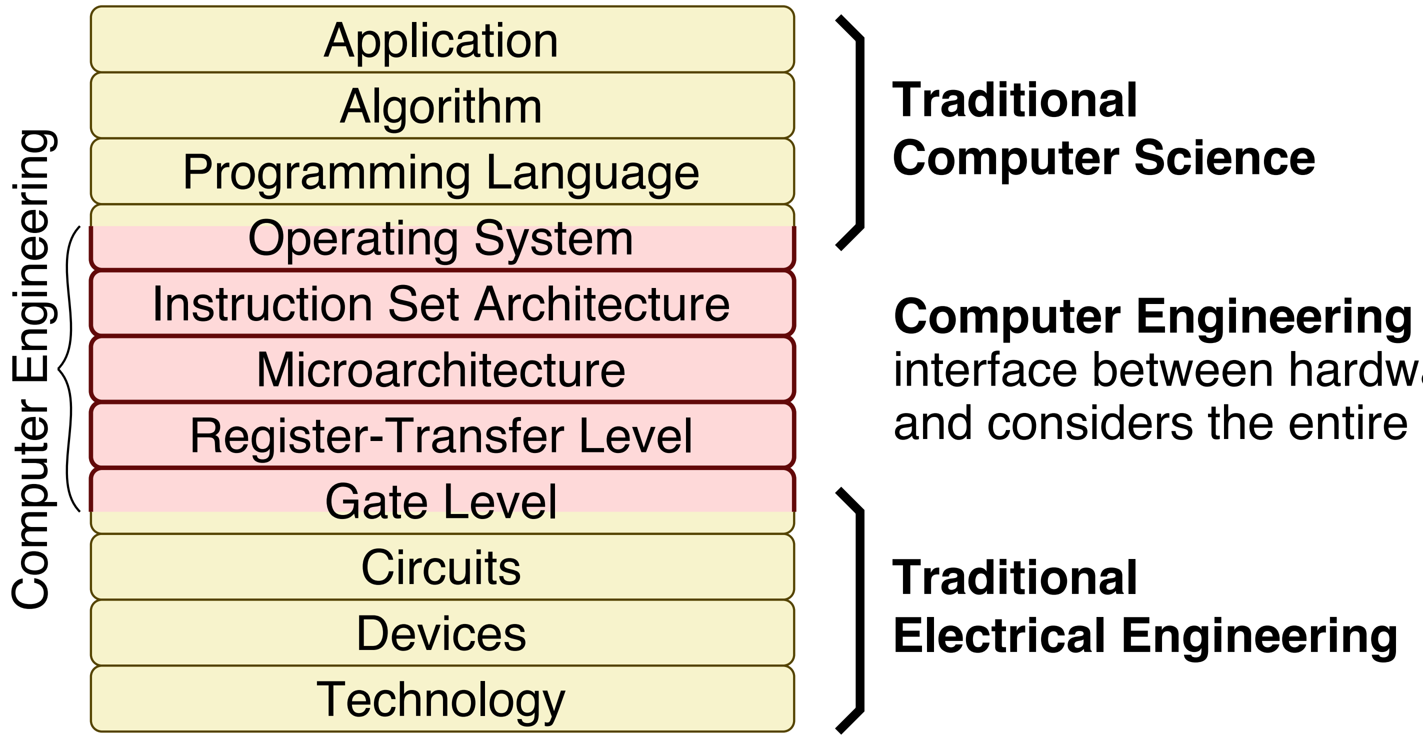
Gap too large to bridge in one step
(but there are exceptions,
e.g., a magnetic compass)

The Computer Systems Stack



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

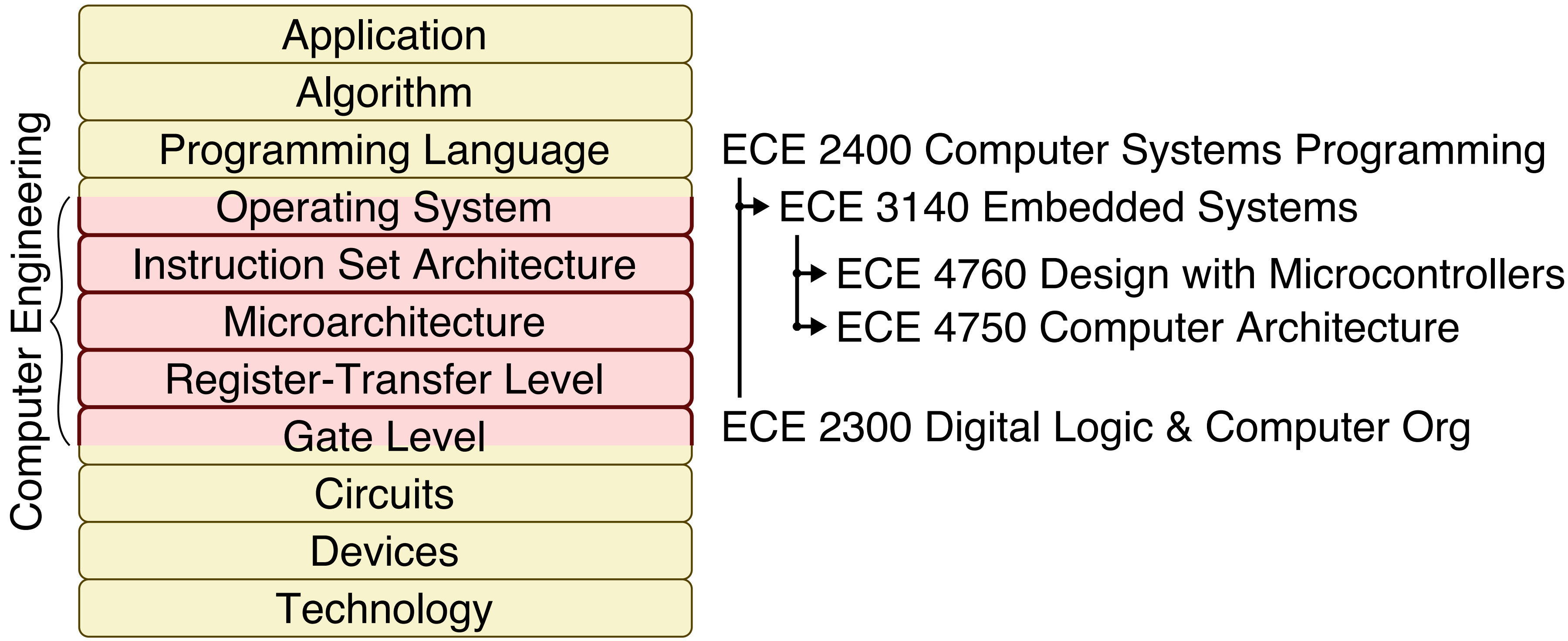
Electrical Engr vs. Comp Sci vs. Comp Engr



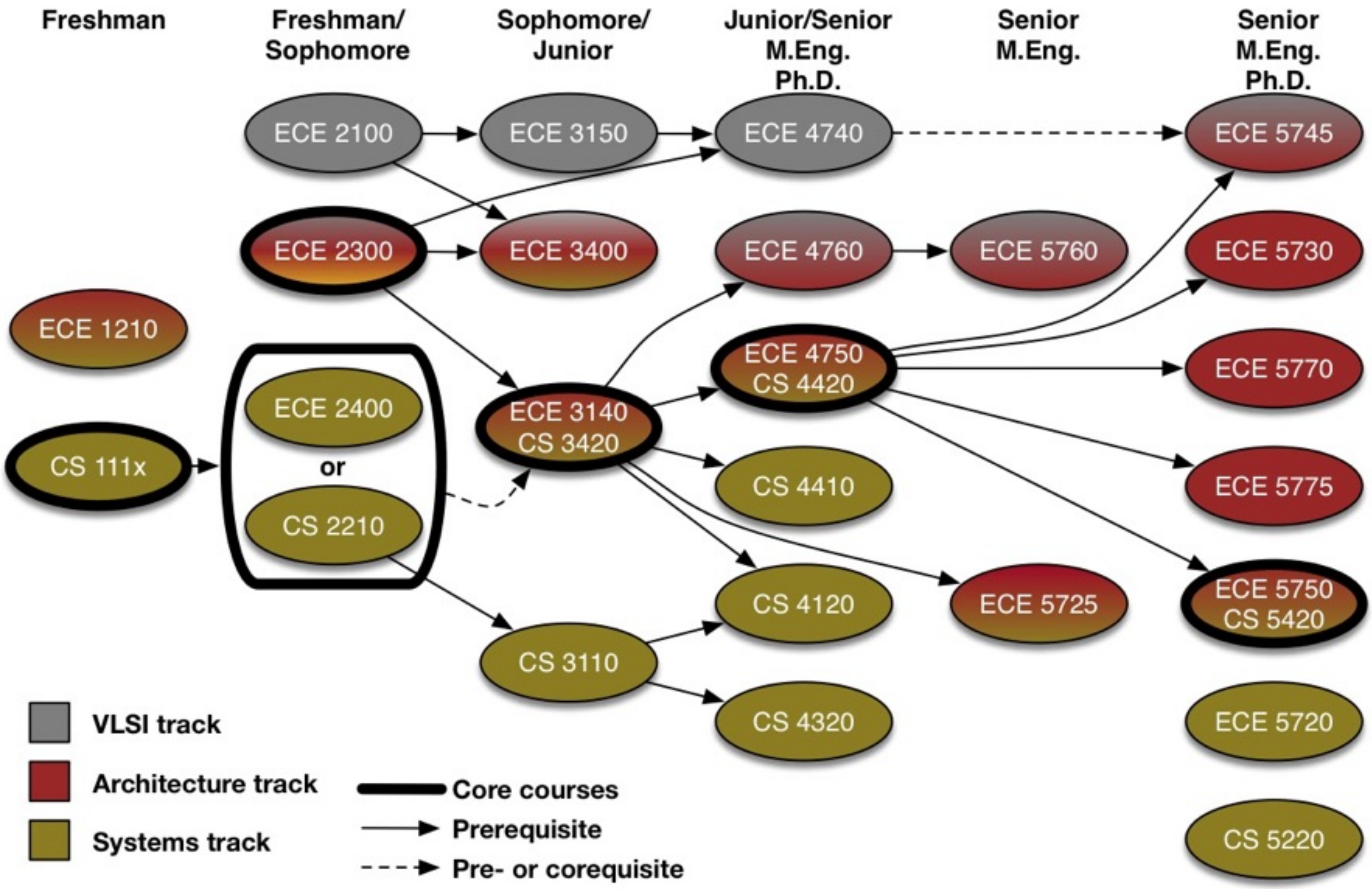
Computer Engineering is at the interface between hardware and software and considers the entire system

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



Cornell Computer Engineering Curriculum



- Application
- Algorithm
- PL
- OS
- ISA
- µArch
- RTL
- Gates
- Circuits
- Devices
- Technology

Agenda

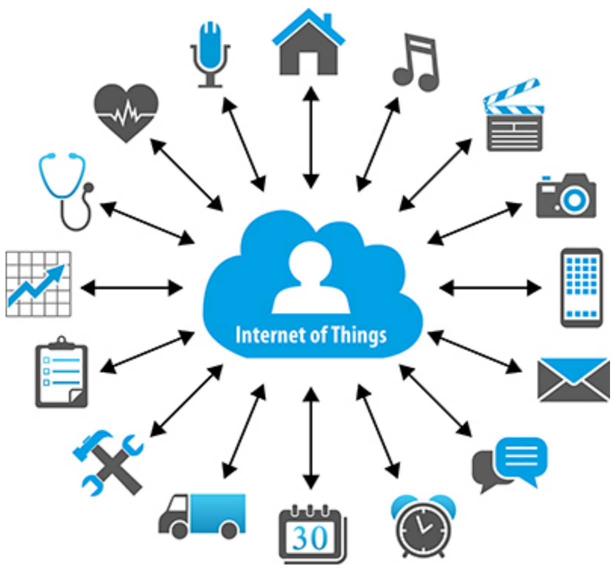
The Computer Systems Stack

Trends in Computer Engineering

Hardware Acceleration for Deep Learning

Three Key Trends in Computer Engineering

Trend #1: Growing Diversity in Applications and Systems



- Application
- Algorithm
- PL
- OS
- ISA
- µArch
- RTL
- Gates
- Circuits
- Devices
- Technology

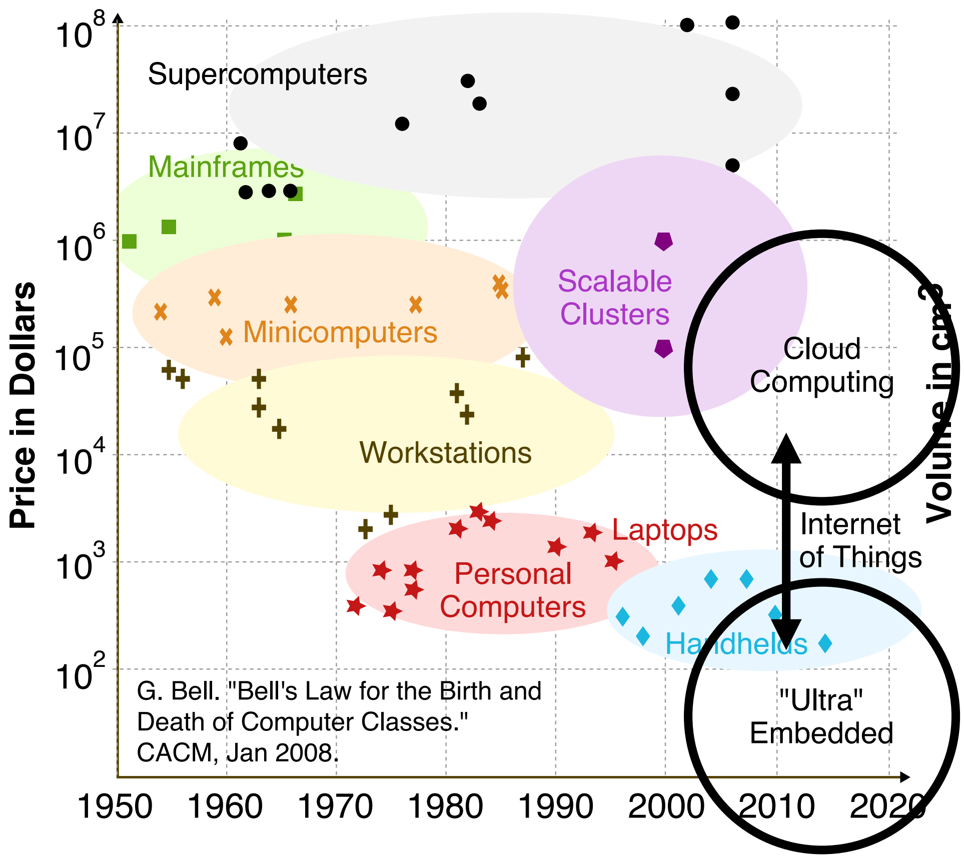
Trend #2: Software/Arch Interface Changing Radically

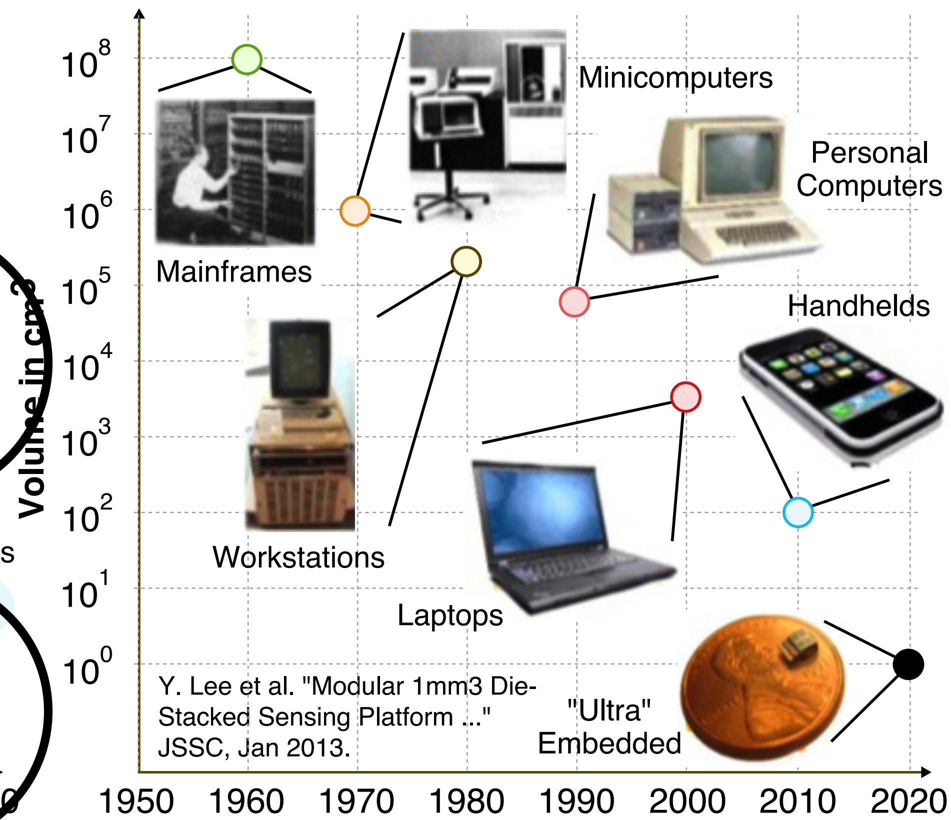
Trend #3: Technology/Arch Interface Changing Radically

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

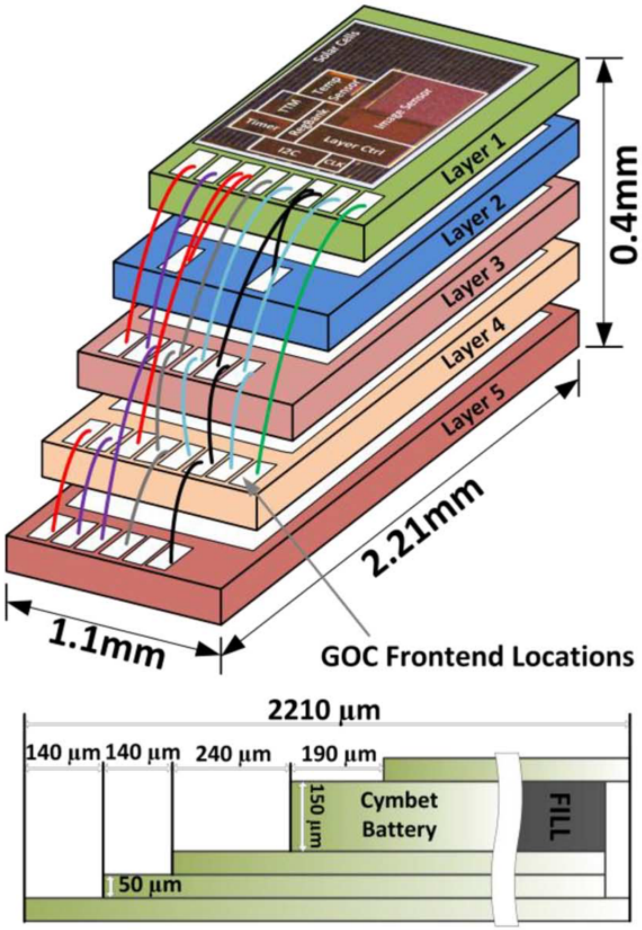
Bell's Law

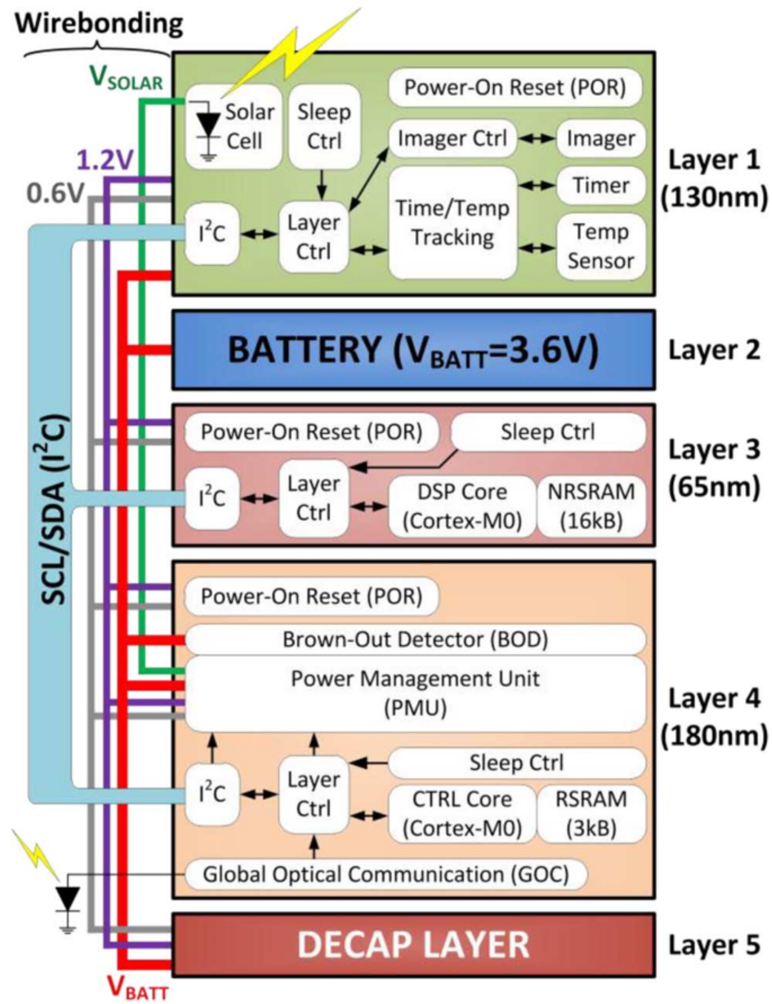
Roughly every decade a new, smaller, lower priced computer class forms based on a new programming platform resulting in entire new industries





M3: Michigan Micro Mote

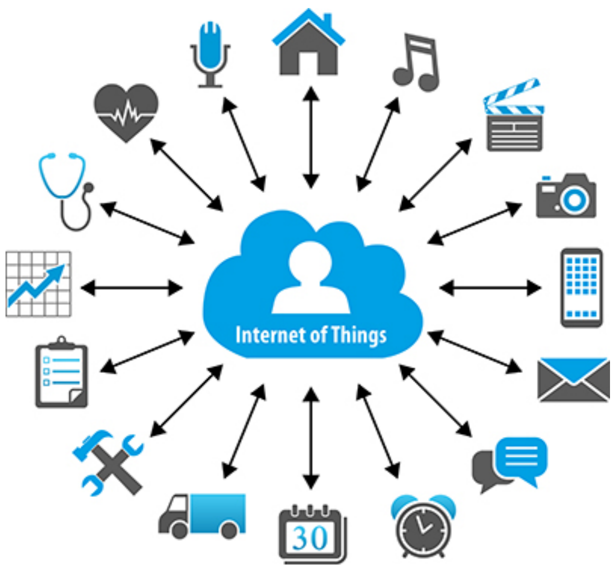




Adapted from Y. Lee et al., JSSC, 2013.

Three Key Trends in Computer Engineering

Trend #1: Growing Diversity in Applications and Systems



- Application
- Algorithm
- PL
- OS
- ISA
- µArch
- RTL
- Gates
- Circuits
- Devices
- Technology

Trend #2: Software/Arch Interface Changing Radically

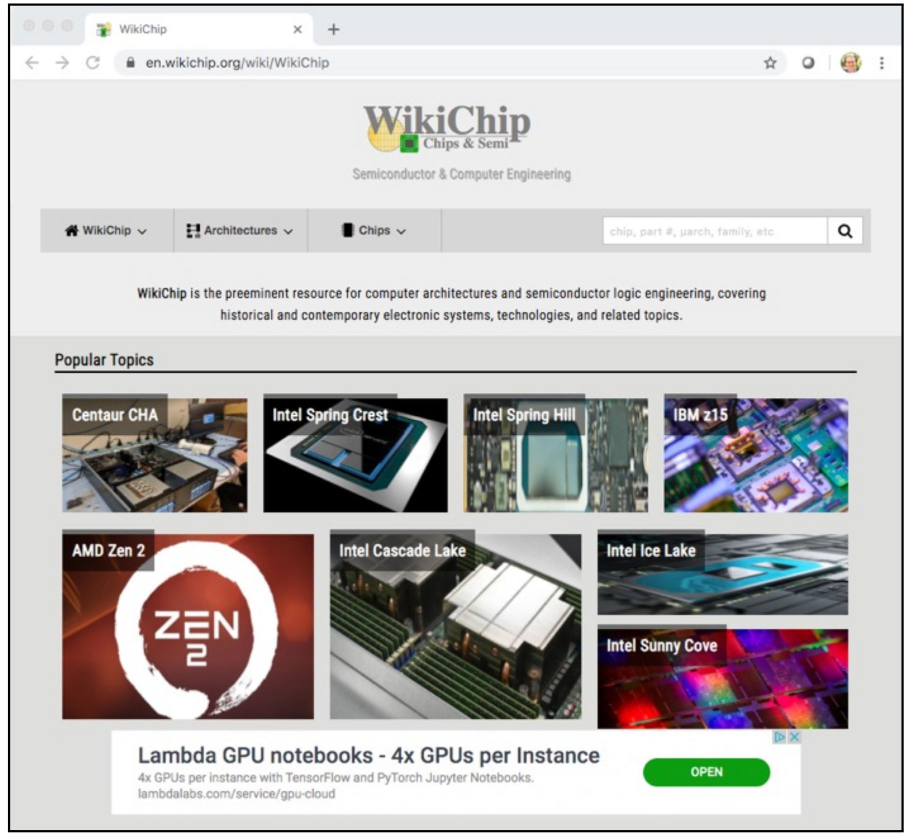
Trend #3: Technology/Arch Interface Changing Radically

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

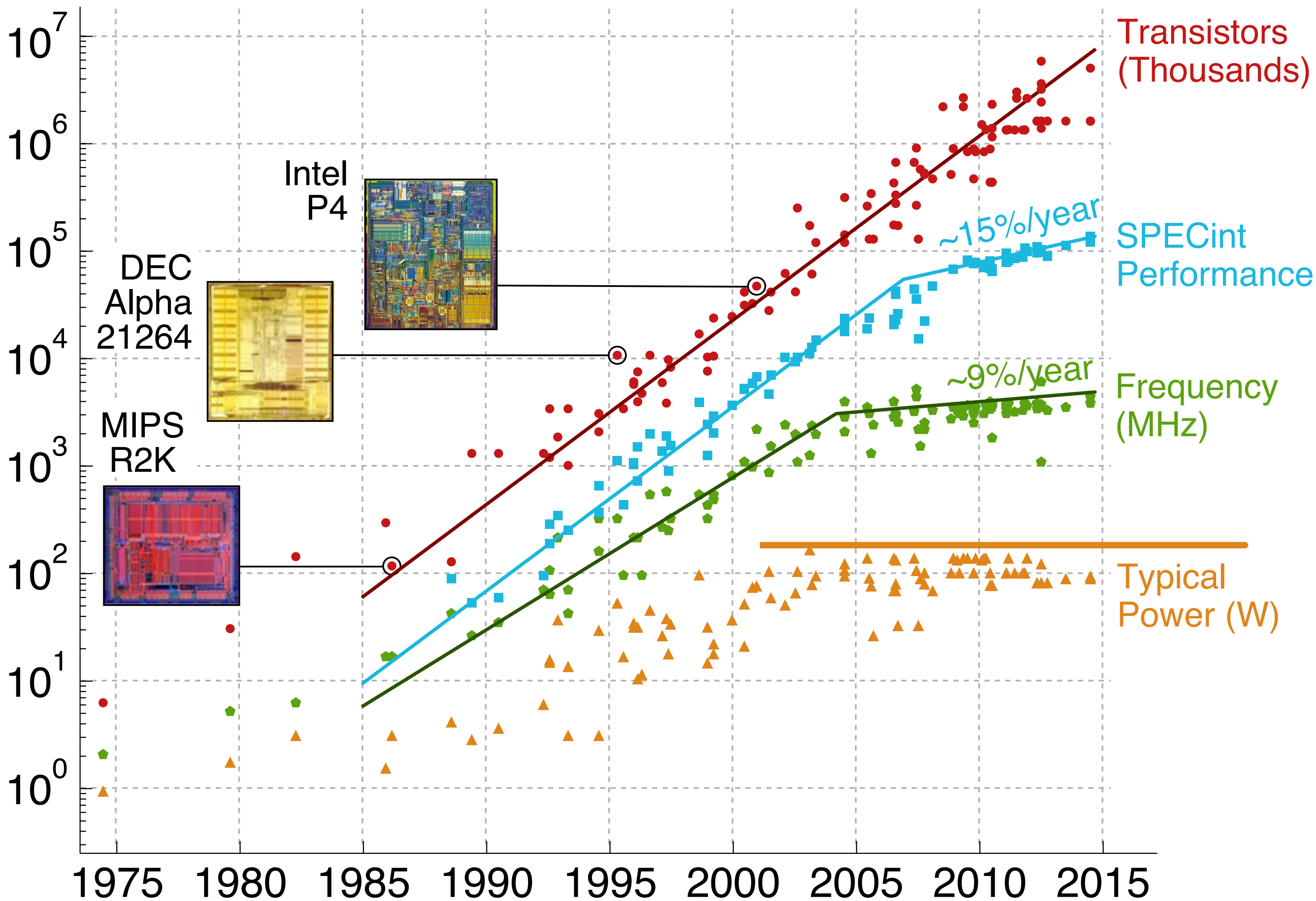
Activity: Specifications of Modern Processors

`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
5. Come back into main zoom room

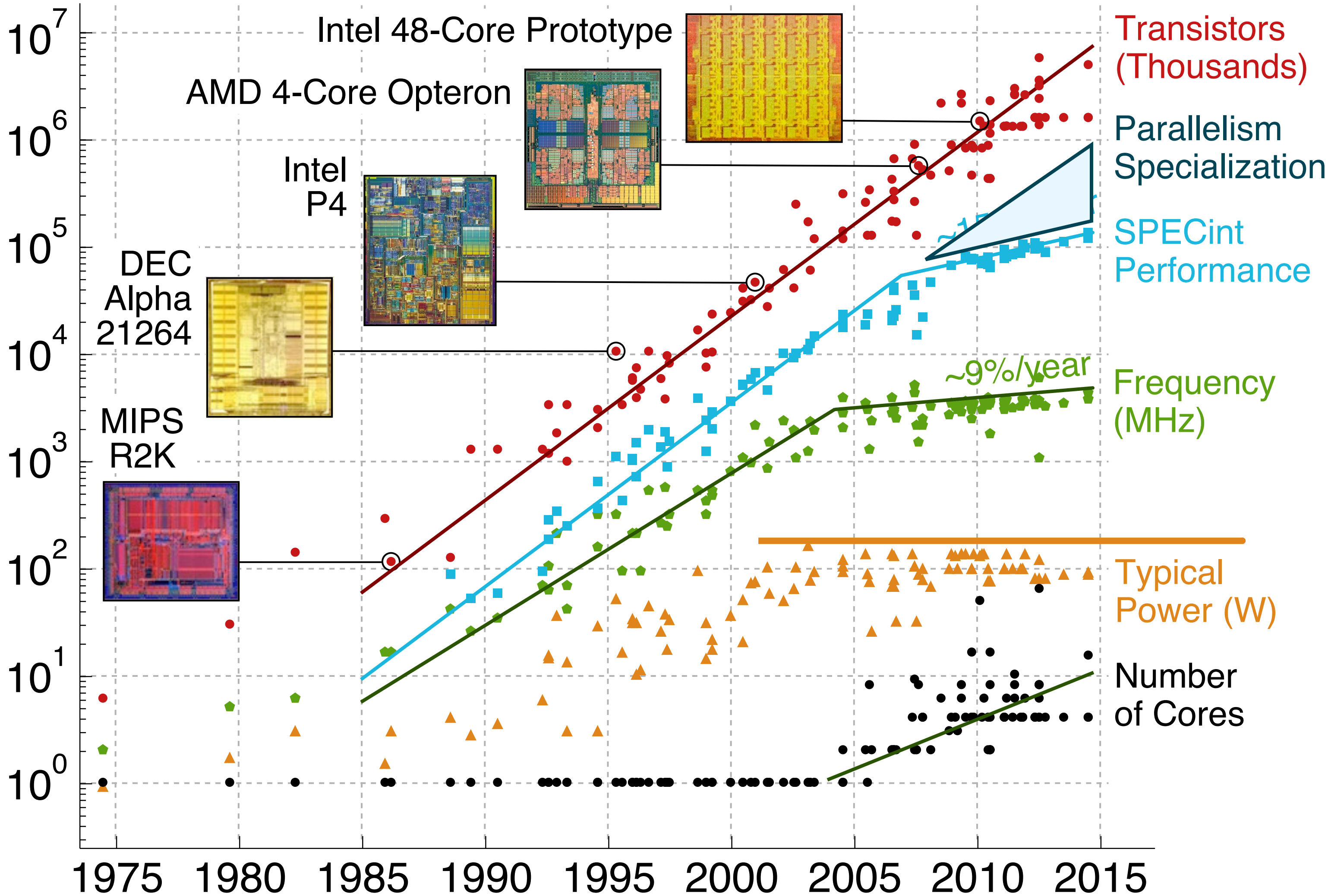


Trends in High-Performance Processors



Data collected by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, C. Batten

Parallelization & Specialization Are Now Critical

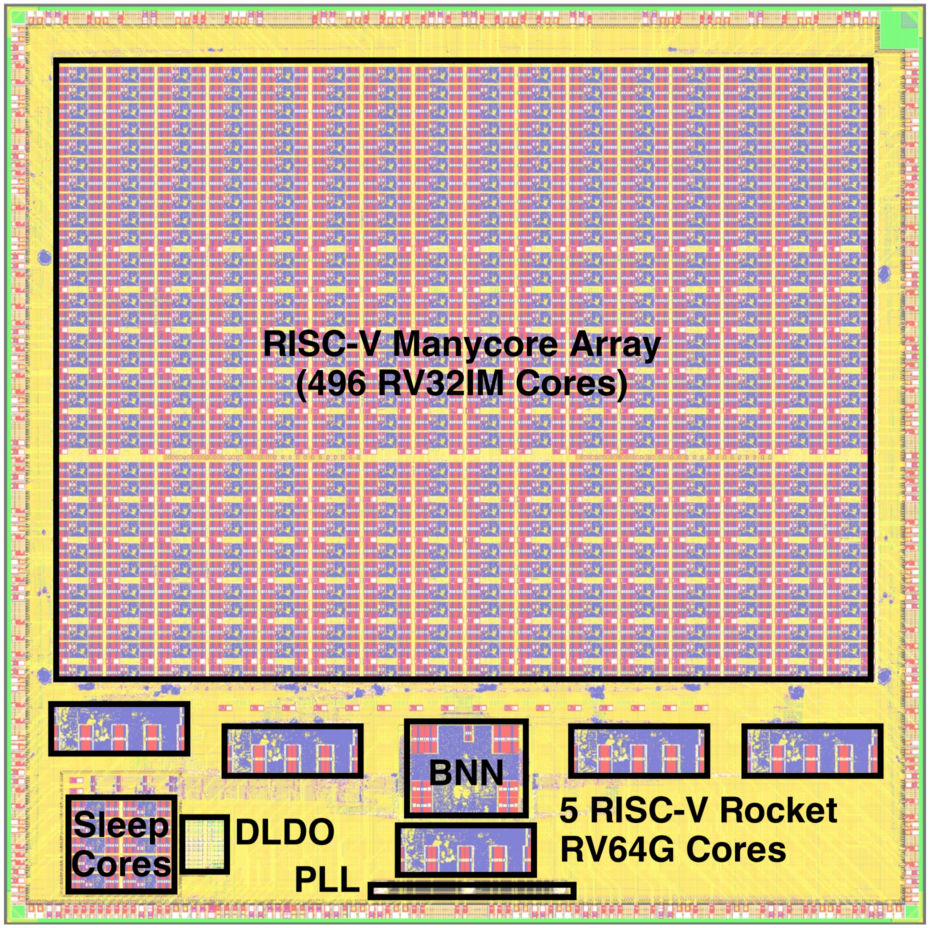


Data collected by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, C. Batten

Celerity System-on-Chip

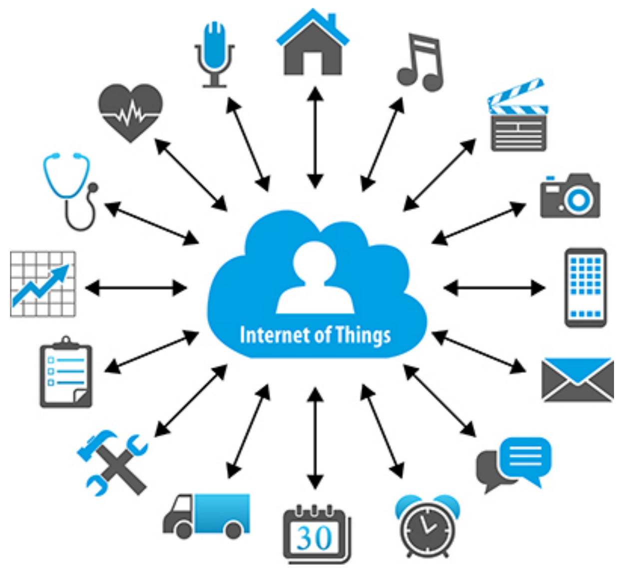
UCSD, Washington, Cornell, Michigan w/ DARPA CRAFT Program

- 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



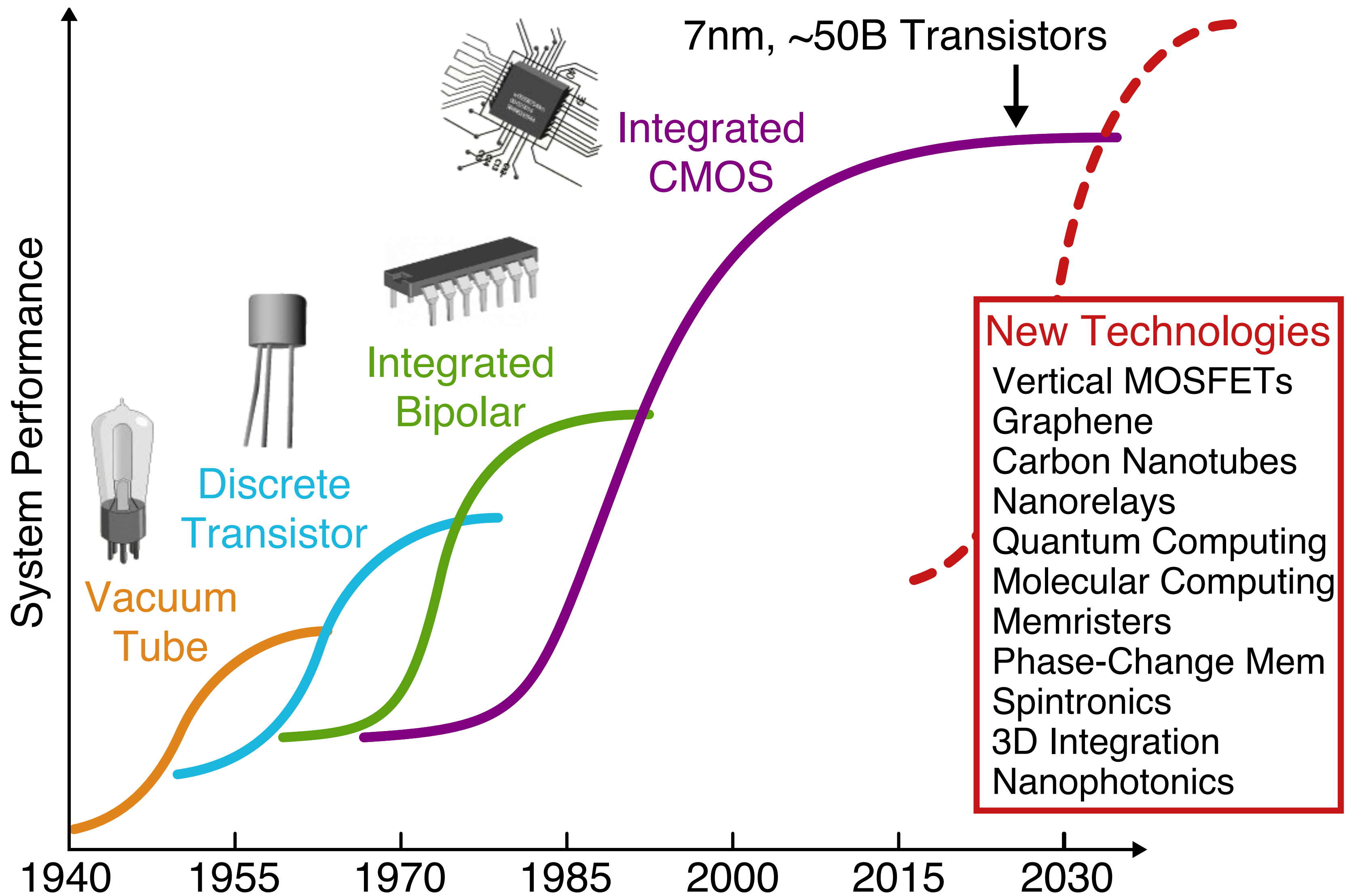
- Application
- Algorithm
- PL
- OS
- ISA
- µArch
- RTL
- Gates
- Circuits
- Devices
- Technology

Trend #2: Software/Arch Interface Changing Radically

Trend #3: Technology/Arch Interface Changing Radically

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

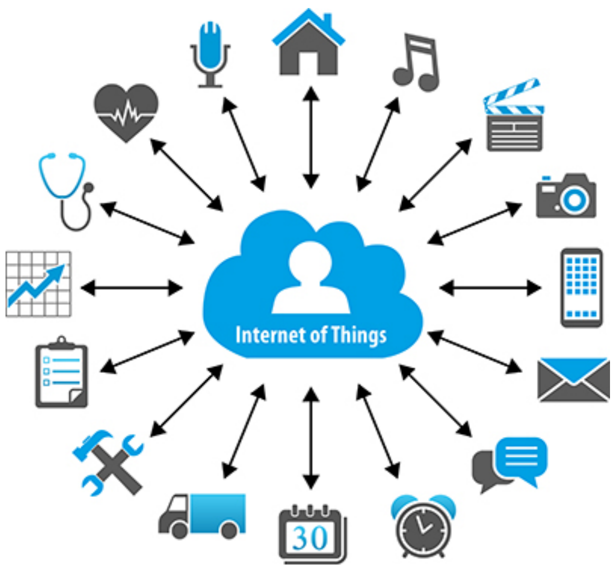
Technology Scaling is Slowing

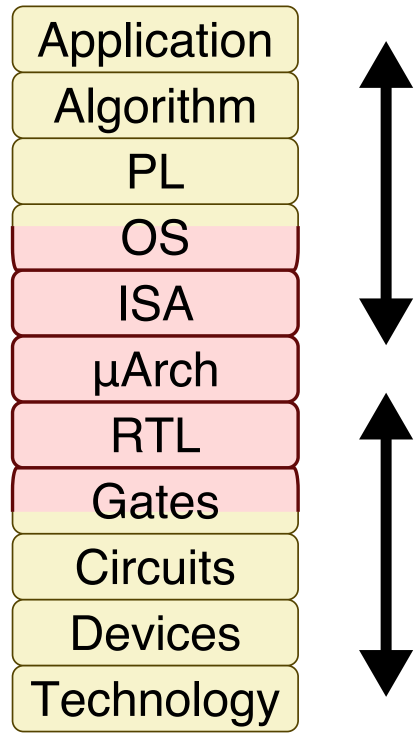


Adapted from D. Brooks Keynote at NSF XPS Workshop, May 2015.

Three Key Trends in Computer Engineering

Trend #1: Growing Diversity in Applications and Systems





Trend #2: Software/Arch Interface Changing Radically

Trend #3: Technology/Arch Interface Changing Radically

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

- Application
- Algorithm
- PL
- OS
- ISA
- µArch
- RTL
- Gates
- Circuits
- Devices
- Technology

Agenda

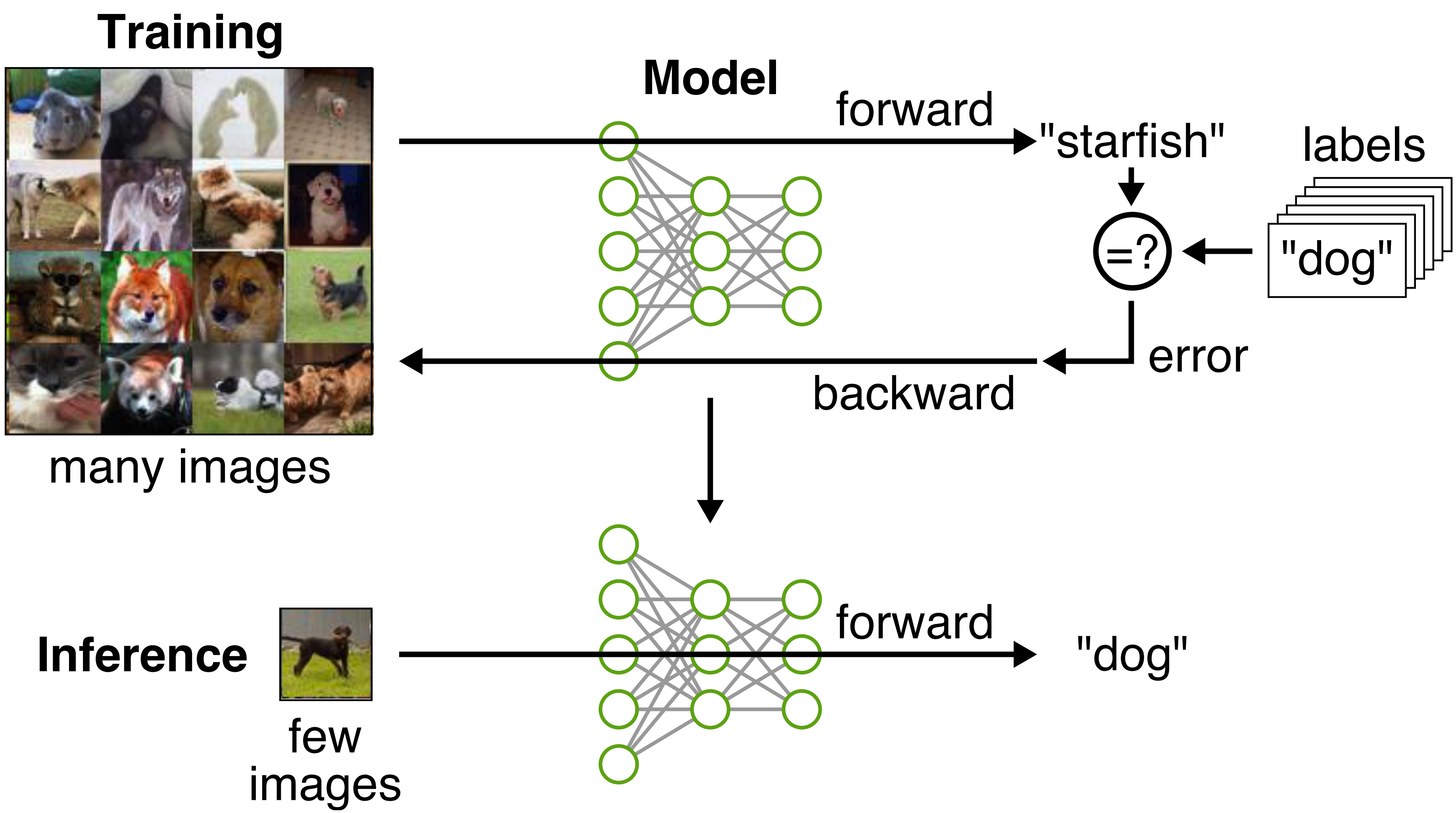
The Computer Systems Stack

Trends in Computer Engineering

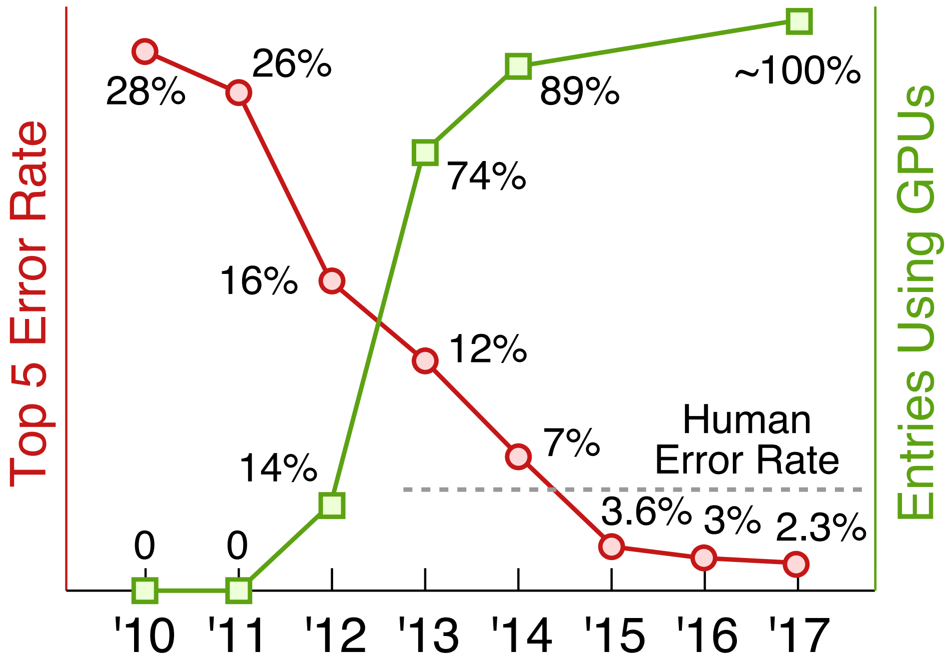
Hardware Acceleration for Deep Learning

Image Recognition

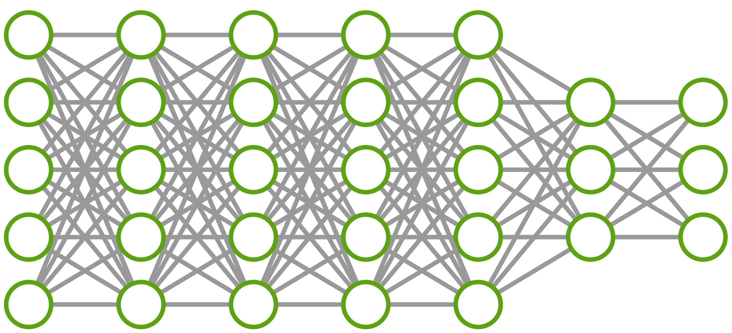
Training vs. Inference



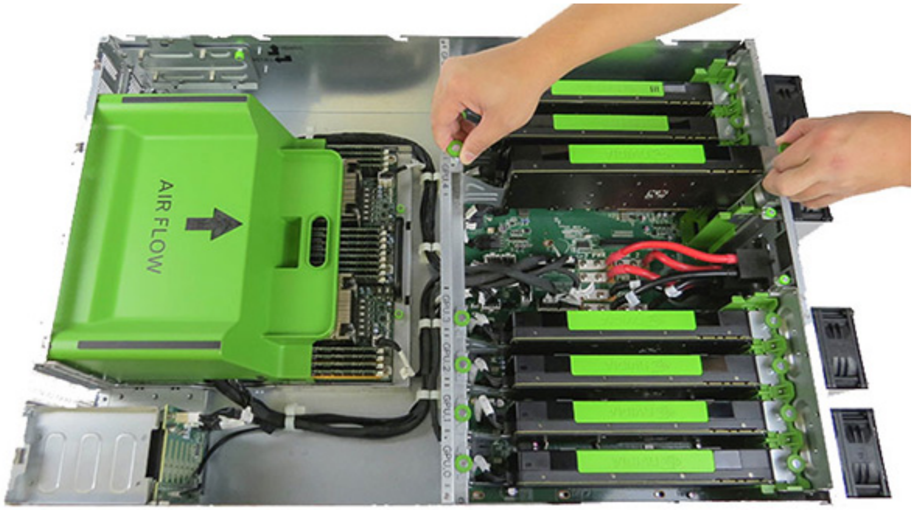
ImageNet Large-Scale Visual Recognition Challenge



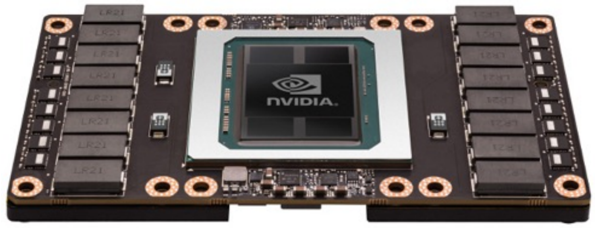
Hardware: Graphics Processing Units



Software: Deep Neural Network

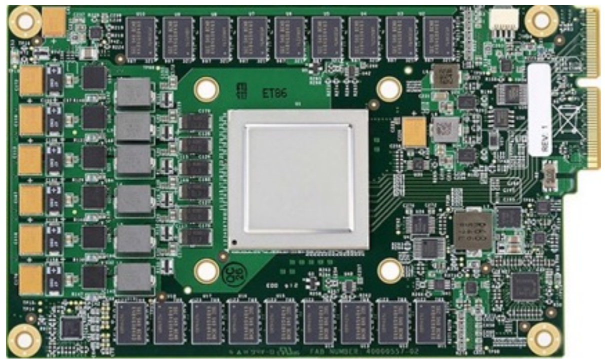


ML Hardware Acceleration in the Cloud



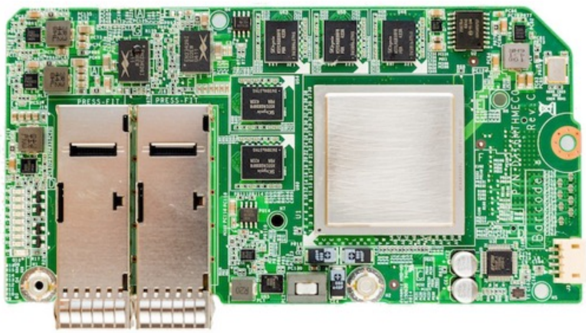
NVIDIA DGX-1

- 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

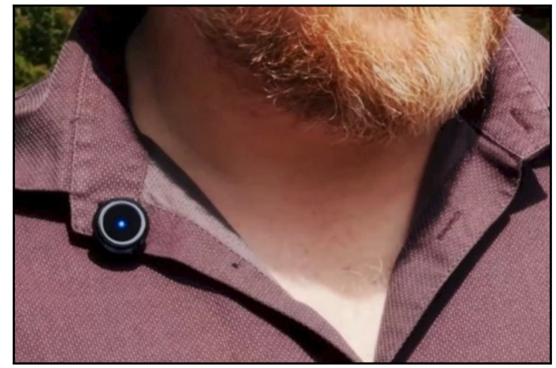
ML Hardware Acceleration at the Edge

Amazon Echo

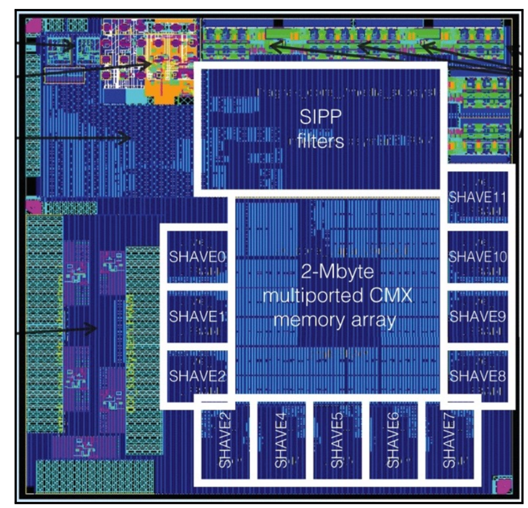
- Developing AI chips so Echo line can do more on-board processing
- Reduces need for round-trip to cloud
- Co-design the algorithms and the underlying hardware

Facebook Oculus

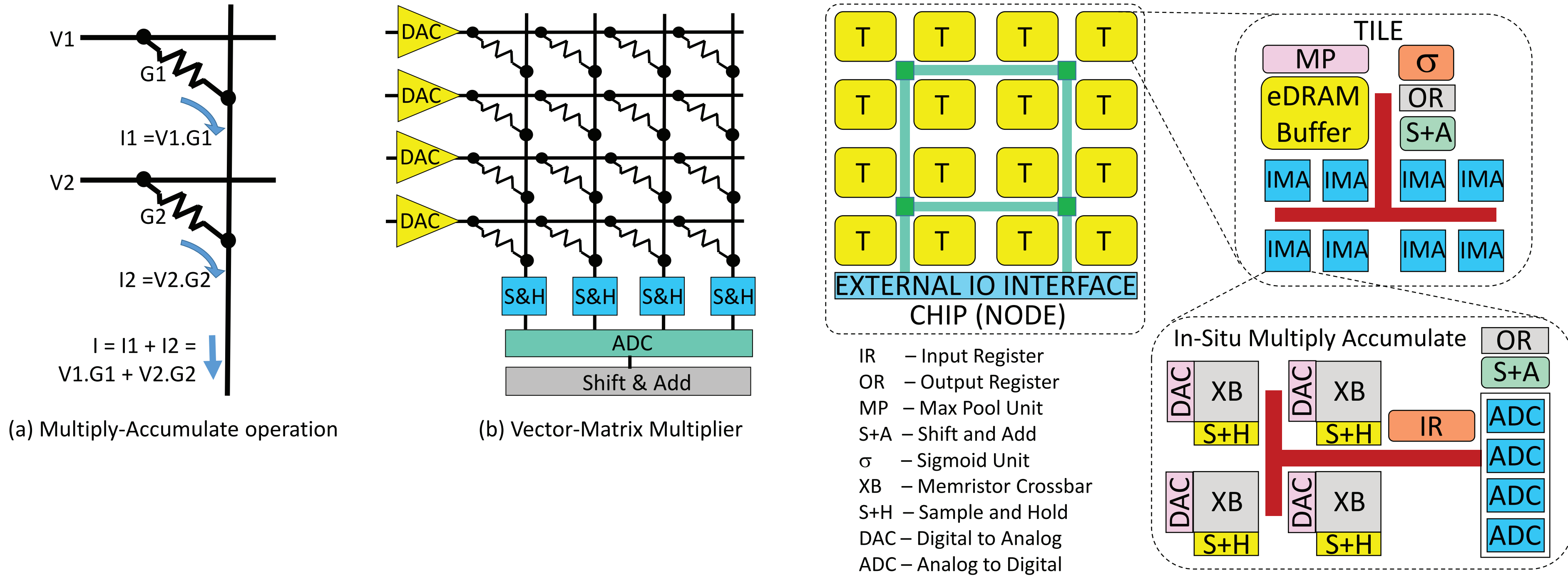
- Starting to design custom chips for Oculus VR headsets
- Significant performance demands under strict power requirements



Movidius Myriad 2



ML Acceleration Can Incorporate All Three Trends



(a) Multiply-Accumulate operation

(b) Vector-Matrix Multiplier

- 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**

Application
Algorithm
PL
OS
ISA
µArch
RTL
Gates
Circuits
Devices
Technology

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

Application-Level Software

System-Level Software