Digital Design & Computer Arch.

Lecture 1: Introduction and Basics

Prof. Onur Mutlu

ETH Zürich

Spring 2021

25 February 2021

Brief Self Introduction

- **Onur Mutlu**
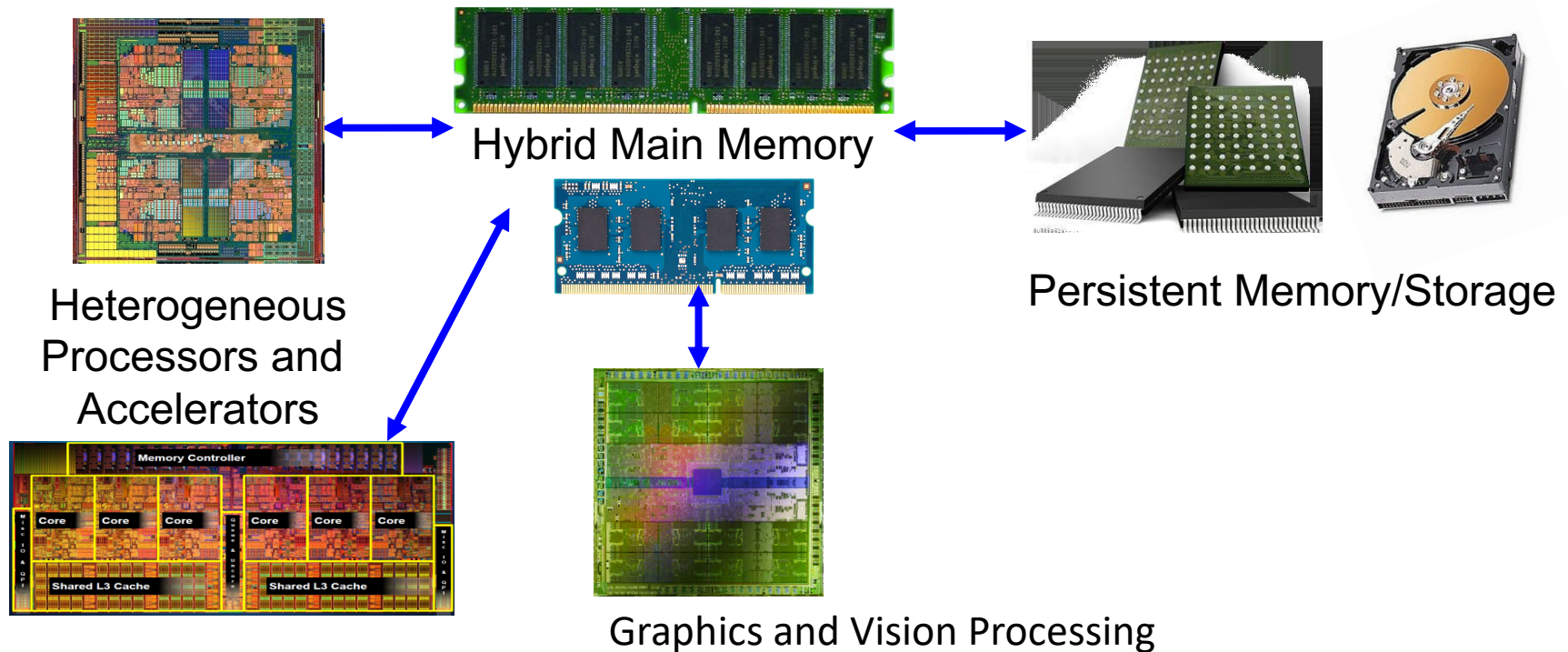 - Full Professor @ ETH Zurich ITET (INFK), since September 2015
  - Strecker Professor @ Carnegie Mellon University ECE/CS, 2009-2016, 2016-…
  - PhD from UT-Austin, worked at Google, VMware, Microsoft Research, Intel, AMD
  - https://people.inf.ethz.ch/omutlu/
  - omutlu@gmail.com (Best way to reach me)
  - https://people.inf.ethz.ch/omutlu/projects.htm

- **Research and Teaching in:**
  - Computer architecture, computer systems, hardware security, bioinformatics
  - Memory and storage systems
  - Hardware security, safety, predictability
  - Fault tolerance, robust systems
  - Hardware/software cooperation
  - Architectures for bioinformatics, health, medicine, intelligent decision making
  - …

Current Research Mission

Computer architecture, HW/SW, systems, bioinformatics, security

Hybrid Main Memory

Heterogeneous Processors and Accelerators

Persistent Memory/Storage

Graphics and Vision Processing

Build fundamentally better architectures

Four Key Current Directions

- Fundamentally Secure/Reliable/Safe Architectures

- Fundamentally Energy-Efficient Architectures
  - Memory-centric (Data-centric) Architectures

- Fundamentally Low-Latency and Predictable Architectures

- Architectures for AI/ML, Genomics, Medicine, Health

Onur Mutlu's SAFARI Research Group

Computer architecture, HW/SW, systems, bioinformatics, security, memory

https://safari.ethz.ch/safari-newsletter-april-2020/

38+ Researchers
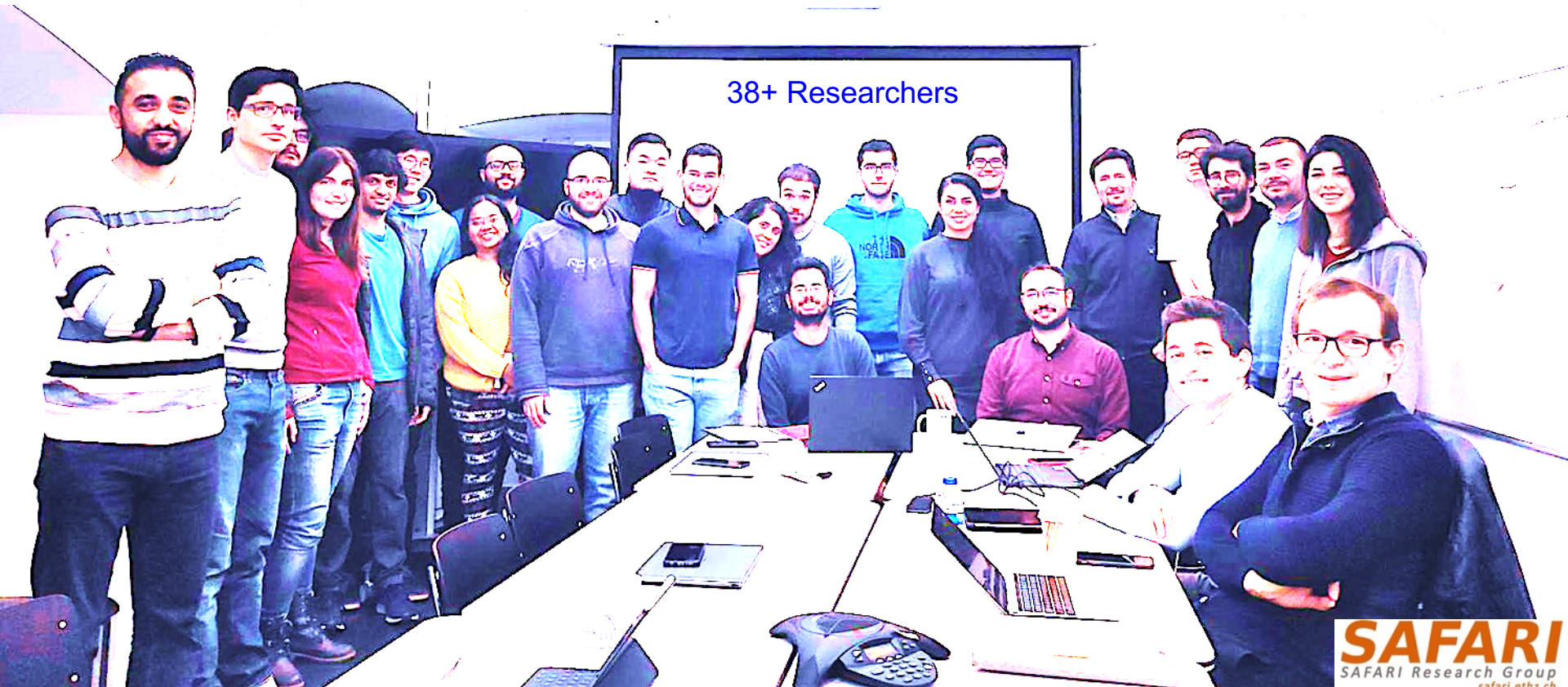
Think BIG, Aim HIGH!

SAFARI

https://safari.ethz.ch

SAFARI Newsletter January 2021 Edition

- https://safari.ethz.ch/safari-newsletter-january-2021/

Principle: Teaching and Research

...

Teaching drives Research

Research drives Teaching

...

Principle: Insight and Ideas

Focus on Insight

Encourage New Ideas

Focus on

learning and scholarship

Principle: Environment of Freedom

Create an environment that values

free exploration, openness, collaboration, hard work, creativity

Principle: Learning and Scholarship

The quality of your work defines your impact

Research & Teaching: Some Overview Talks

https://www.youtube.com/onurmutlulectures

- ## Future Computing Architectures
  - https://www.youtube.com/watch?v=kgiZlSOcGFM&list=PL5Q2soXY2Zi8D_5MGV6EnXEJHnV2YFBJl&index=1

- ## Enabling In-Memory Computation
  - https://www.youtube.com/watch?v=njX_14584Jw&list=PL5Q2soXY2Zi8D_5MGV6EnXEJHnV2YFBJl&index=16

- ## Accelerating Genome Analysis
  - https://www.youtube.com/watch?v=r7sn41lH-4A&list=PL5Q2soXY2Zi8D_5MGV6EnXEJHnV2YFBJl&index=41

- ## Rethinking Memory System Design
  - https://www.youtube.com/watch?v=F7xZLNMIY1E&list=PL5Q2soXY2Zi8D_5MGV6EnXEJHnV2YFBJl&index=3

- ## Intelligent Architectures for Intelligent Machines
  - https://www.youtube.com/watch?v=c6_LgzuNdkw&list=PL5Q2soXY2Zi8D_5MGV6EnXEJHnV2YFBJl&index=25

- ## The Story of RowHammer
  - https://www.youtube.com/watch?v=sgd7PHQQ1AI&list=PL5Q2soXY2Zi8D_5MGV6EnXEJHnV2YFBJl&index=39

SAFARI

An Interview on Research and Education

- **Computing Research and Education (@ ISCA 2019)**
  - https://www.youtube.com/watch?v=8ffSEKZhmvo&list=PL5Q2soXY2Zi_4oP9LdL3cc8G6NIjD2Ydz

- **Maurice Wilkes Award Speech (10 minutes)**
  - https://www.youtube.com/watch?v=tcQ3zZ3JpuA&list=PL5Q2soXY2Zi8D_5MGV6EnXEJHnV2YFBJl&index=15

More Thoughts and Suggestions

- Onur Mutlu,
  "Some Reflections (on DRAM)"
  Award Speech for ACM SIGARCH Maurice Wilkes Award, *at the* **ISCA** *Awards Ceremony*, Phoenix, AZ, USA, 25 June 2019.
  [Slides (pptx) (pdf)]
  [Video of Award Acceptance Speech (Youtube; 10 minutes) (Youku; 13 minutes)]
  [Video of Interview after Award Acceptance (Youtube; 1 hour 6 minutes) (Youku; 1 hour 6 minutes)]
  [News Article on "ACM SIGARCH Maurice Wilkes Award goes to Prof. Onur Mutlu"]

- Onur Mutlu,
  "How to Build an Impactful Research Group"
  *57th Design Automation Conference Early Career Workshop (**DAC**)*, Virtual, 19 July 2020.
  [Slides (pptx) (pdf)]

SAFARI

How to Approach This Course

"Formative Experience"

How to Approach This Course

"High investment, high return"

How to Approach This Course

Learning experience

Long-term tradeoff analysis

Critical thinking & decision making

How to Approach This Course

Concepts & Ideas

Fundamentals

Cutting-edge

Hands-on learning

What Will We Learn in This Course?

How Computers Work
(from the ground up)

Answer Continued

And Why We Care

Why Do We Have Computers?

Why Do We Do Computing?

Answer

To Solve Problems

Answer Reworded

To Gain Insight

Hamming, "Numerical Methods for Scientists and Engineers," 1962. [25]

Answer Extended

To Enable
a Better Life & Future

SAFARI

How Does a Computer Solve Problems?

Answer

Orchestrating Electrons

In today's dominant technologies

SAFARI

How Do Problems Get Solved by Electrons?

The Transformation Hierarchy

Computer Architecture (expanded view)

Computer Architecture (narrow view)

Problem
Algorithm
Program/Language
System Software
SW/HW Interface
Micro-architecture
Logic
Devices
Electrons

Levels of Transformation

"The purpose of computing is [to gain] insight" (*Richard Hamming*)
We gain and generate insight by solving problems
How do we ensure problems are solved by electrons?

Algorithm

Step-by-step procedure that is **guaranteed to terminate** where **each step is precisely stated** and **can be carried out by a computer**

- **Finiteness**
- **Definiteness**
- **Effective computability**

Many algorithms for the same problem

Problem
Algorithm
Program/Language
Runtime System (VM, OS, MM)
ISA (Architecture)
Microarchitecture
Logic
Devices
Electrons

ISA (Instruction Set Architecture)

Interface/contract between SW and HW.

What the programmer assumes hardware will satisfy.

Microarchitecture
An implementation of the ISA

Digital logic circuits
Building blocks of micro-arch (e.g., gates)

Computer Architecture

- is the science and art of designing computing platforms (hardware, interface, system SW, and programming model)

- to achieve a set of design goals
  - E.g., highest performance on earth on workloads X, Y, Z
  - E.g., longest battery life at a form factor that fits in your pocket with cost < $$$ CHF
  - E.g., best average performance across all known workloads at the best performance/cost ratio
  - …

  - Designing a supercomputer is different from designing a smartphone → But, many fundamental principles are similar

SAFARI

Different Platforms, Different Goals

SAFARI

Source: http://www.sia-online.org (semiconductor industry association)

Different Platforms, Different Goals

SAFARI Source: https://iq.intel.com/5-awesome-uses-for-drone-technology/

Different Platforms, Different Goals

Source: https://taxistartup.com/wp-content/uploads/2015/03/UK-Self-Driving-Cars.jpg

Different Platforms, Different Goals

Source: http://sm.pcmag.com/pcmag_uk/photo/g/google-self-driving-car-the-guts/google-self-driving-car-the-guts_dwx8.jpg

Different Platforms, Different Goals

Source: http://datacentervoice.com/wp-content/uploads/2015/10/data-center.jpg

Different Platforms, Different Goals

Source: https://fossbytes.com/wp-content/uploads/2015/06/Supercomputer-TIANHE2-china.jpg

Different Platforms, Different Goals

Figure 3. TPU Printed Circuit Board. It can be inserted in the slot for an SATA disk in a server, but the card uses PCIe Gen3 x16.

Figure 4. Systolic data flow of the Matrix Multiply Unit. Software has the illusion that each 256B input is read at once, and they instantly update one location of each of 256 accumulator RAMs.

Jouppi et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit", ISCA 2017.

Different Platforms, Different Goals

- ML accelerator: 260 mm^2, 6 billion transistors, 600 GFLOPS GPU, 12 ARM 2.2 GHz CPUs.
- Two redundant chips for better safety.

Axiom

To achieve the highest energy efficiency and performance:

we must take the expanded view
of computer architecture

Problem
Algorithm
Program/Language
System Software
SW/HW Interface
Micro-architecture
Logic
Devices
Electrons
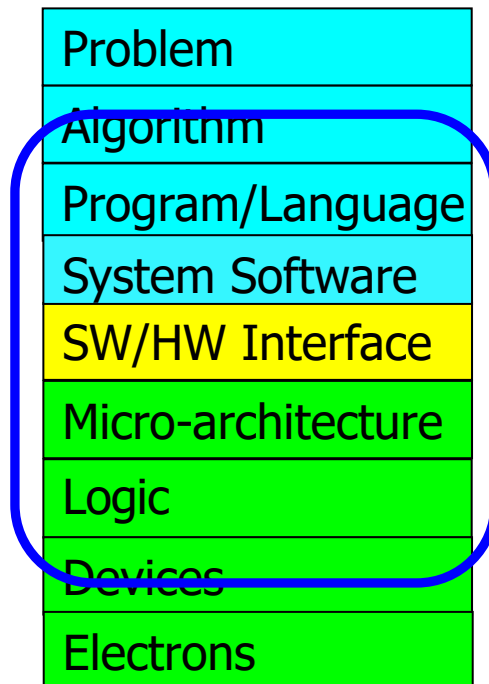
Co-design across the hierarchy:
Algorithms to devices

Specialize as much as possible
within the design goals

What is Computer Architecture?

- The science and art of designing, selecting, and interconnecting hardware components and designing the hardware/software interface to create a computing system that meets functional, performance, energy consumption, cost, and other specific goals.

Why Study Computer Architecture?

- **Enable better systems**: make computers faster, cheaper, smaller, more reliable, …
  - By exploiting advances and changes in underlying technology/circuits

- **Enable new applications**
  - Life-like 3D visualization 20 years ago? Virtual reality?
  - Self-driving cars?
  - Personalized genomics? Personalized medicine?

- **Enable better solutions** to problems
  - Software innovation is built on trends and changes in computer architecture
    - > 50% performance improvement per year has enabled this innovation

- **Understand why computers work the way they do**

Computer Architecture Today (I)

- Today is a very exciting time to study computer architecture

- Industry is in a large paradigm shift (to novel architectures) – many different potential system designs possible

- Many difficult problems *motivating* and *caused by* the shift
  - Huge hunger for data and new data-intensive applications
  - Power/energy/thermal constraints
  - Complexity of design
  - Difficulties in technology scaling
  - Memory bottleneck
  - Reliability problems
  - Programmability problems
  - Security and privacy issues

- No clear, definitive answers to these problems

Computer Architecture Today (II)

- Computing landscape is very different from 10-20 years ago

- Applications and technology both demand novel architectures

Heterogeneous
Processors and
Accelerators

Hybrid Main Memory

Persistent Memory/Storage

General Purpose GPUs

Every component and its interfaces, as well as entire system designs are being re-examined

Axiom

To achieve the highest energy efficiency and performance:

we must take the expanded view
of computer architecture

Problem
Algorithm
Program/Language
System Software
SW/HW Interface
Micro-architecture
Logic
Devices
Electrons
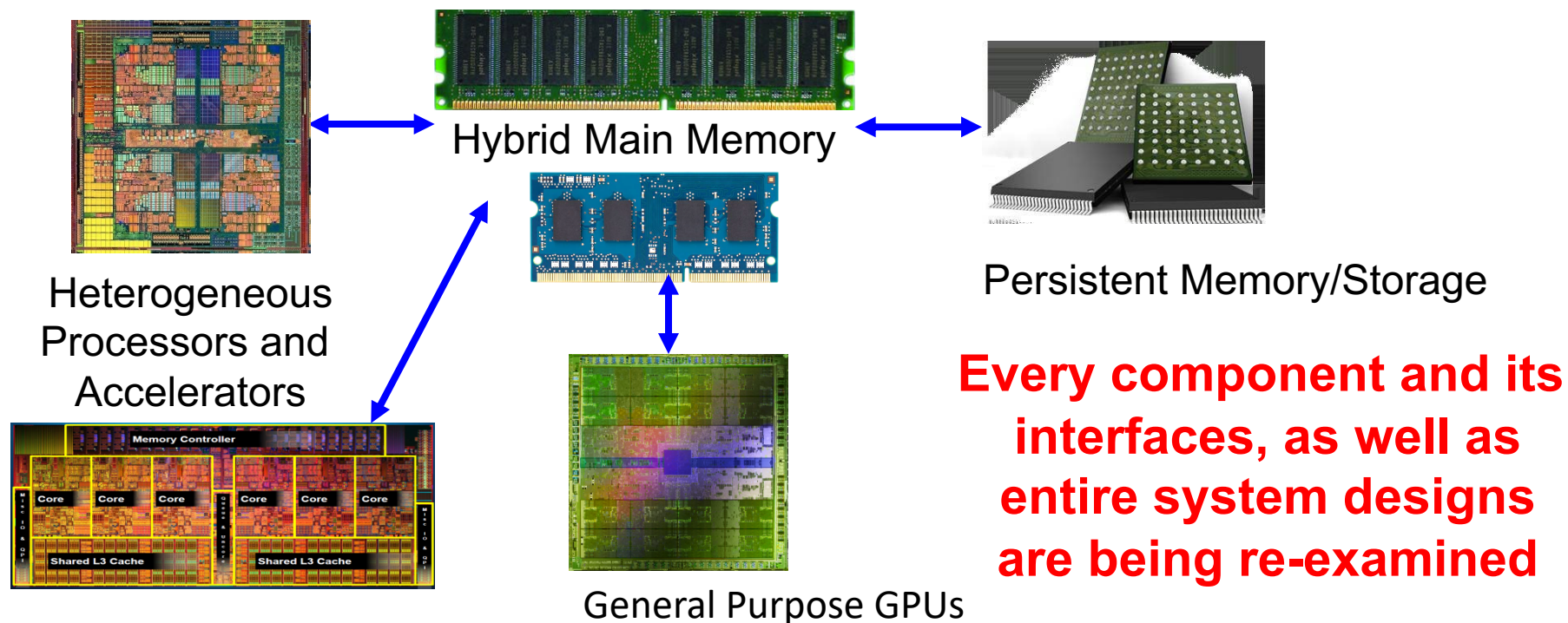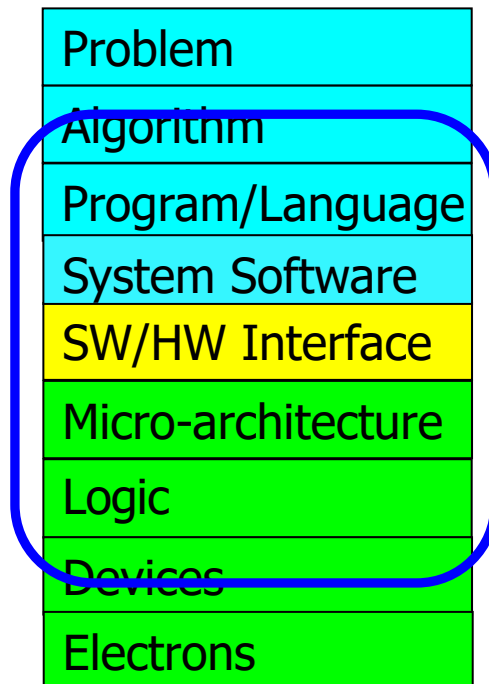
Co-design across the hierarchy:
Algorithms to devices

Specialize as much as possible
within the design goals

Historical: Opportunities at the Bottom

There's Plenty of Room at the Bottom

From Wikipedia, the free encyclopedia

"There's Plenty of Room at the Bottom: An Invitation to Enter a New Field of Physics" was a lecture given by physicist Richard Feynman at the annual American Physical Society meeting at Caltech on December 29, 1959.[1] Feynman considered the possibility of direct manipulation of individual atoms as a more powerful form of synthetic chemistry than those used at the time. Although versions of the talk were reprinted in a few popular magazines, it went largely unnoticed and did not inspire the conceptual beginnings of the field. Beginning in the 1980s, nanotechnology advocates cited it to establish the scientific credibility of their work.

Historical: Opportunities at the Bottom (II)

There's Plenty of Room at the Bottom

From Wikipedia, the free encyclopedia

Feynman considered some ramifications of a general ability to manipulate matter on an atomic scale. He was particularly interested in the possibilities of denser computer circuitry, and microscopes that could see things much smaller than is possible with scanning electron microscopes. These ideas were later realized by the use of the scanning tunneling microscope, the atomic force microscope and other examples of scanning probe microscopy and storage systems such as Millipede, created by researchers at IBM.

Feynman also suggested that it should be possible, in principle, to make nanoscale machines that "arrange the atoms the way we want", and do chemical synthesis by mechanical manipulation.

He also presented the possibility of "swallowing the doctor", an idea that he credited in the essay to his friend and graduate student Albert Hibbs. This concept involved building a tiny, swallowable surgical robot.

Historical: Opportunities at the Top

There's plenty of room at the Top: What will drive computer performance after Moore's law?

Charles E. Leiserson[1], Neil C. Thompson[1,2,*], Joel S. Emer[1,3], Bradley C. Kuszmaul[1,†], Butler W. Lampson[1,4], …

+ See all authors and affiliations

Much of the improvement in computer performance comes from decades of miniaturization of computer components, a trend that was foreseen by the Nobel Prize–winning physicist Richard Feynman in his 1959 address, "There's Plenty of Room at the Bottom," to the American Physical Society. In 1975, Intel founder Gordon Moore predicted the regularity of this miniaturization trend, now called Moore's law, which, until recently, doubled the number of transistors on computer chips every 2 years.

Unfortunately, semiconductor miniaturization is running out of steam as a viable way to grow computer performance—there isn't much more room at the "Bottom." If growth in computing power stalls, practically all industries will face challenges to their productivity. Nevertheless, opportunities for growth in computing performance will still be available, especially at the "Top" of the computing-technology stack: software, algorithms, and hardware architecture.

Axiom, Revisited

There is plenty of room both at the top and at the bottom

but **much more so**

when you

communicate well between and optimize across

the top and the bottom

Hence the Expanded View

Computer Architecture (expanded view)
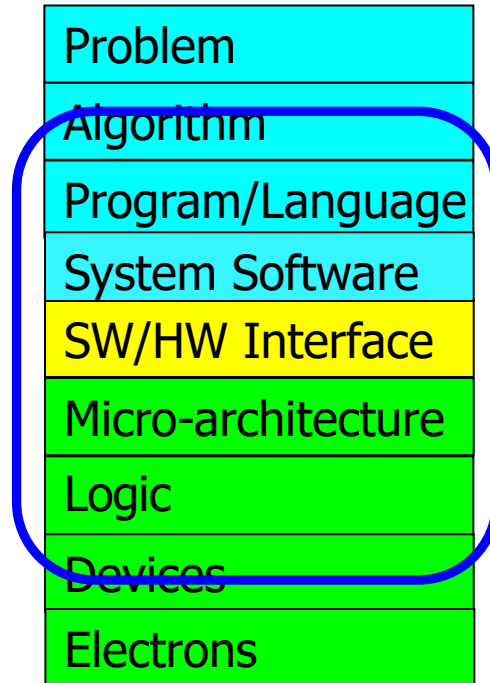
Problem
Algorithm
Program/Language
System Software
SW/HW Interface
Micro-architecture
Logic
Devices
Electrons

Many Interesting Things Are Happening Today in Computer Architecture

Many Interesting Things Are Happening Today in Computer Architecture

Performance
and
Energy Efficiency

Intel Optane Persistent Memory (2019)

- Non-volatile main memory
- Based on 3D-XPoint Technology

SAFARI https://www.storagereview.com/intel_optane_dc_persistent_memory_module_pmm

PCM as Main Memory: Idea in 2009

- Benjamin C. Lee, Engin Ipek, Onur Mutlu, and Doug Burger,
  "Architecting Phase Change Memory as a Scalable DRAM Alternative"
  Proceedings of the 36th International Symposium on Computer Architecture (**ISCA**), pages 2-13, Austin, TX, June 2009. Slides (pdf)

Architecting Phase Change Memory as a Scalable DRAM Alternative

Benjamin C. Lee†   Engin Ipek†   Onur Mutlu‡   Doug Burger†

†Computer Architecture Group
Microsoft Research
Redmond, WA
{blee, ipek, dburger}@microsoft.com

‡Computer Architecture Laboratory
Carnegie Mellon University
Pittsburgh, PA
onur@cmu.edu

SAFARI

PCM as Main Memory: Idea in 2009

- Benjamin C. Lee, Ping Zhou, Jun Yang, Youtao Zhang, Bo Zhao, Engin Ipek, Onur Mutlu, and Doug Burger,
  "Phase Change Technology and the Future of Main Memory"
  IEEE Micro, Special Issue: Micro's Top Picks from 2009 Computer Architecture Conferences (**MICRO TOP PICKS**), Vol. 30, No. 1, pages 60-70, January/February 2010.

PHASE-CHANGE TECHNOLOGY AND THE FUTURE OF MAIN MEMORY

Cerebras's Wafer Scale Engine (2019)
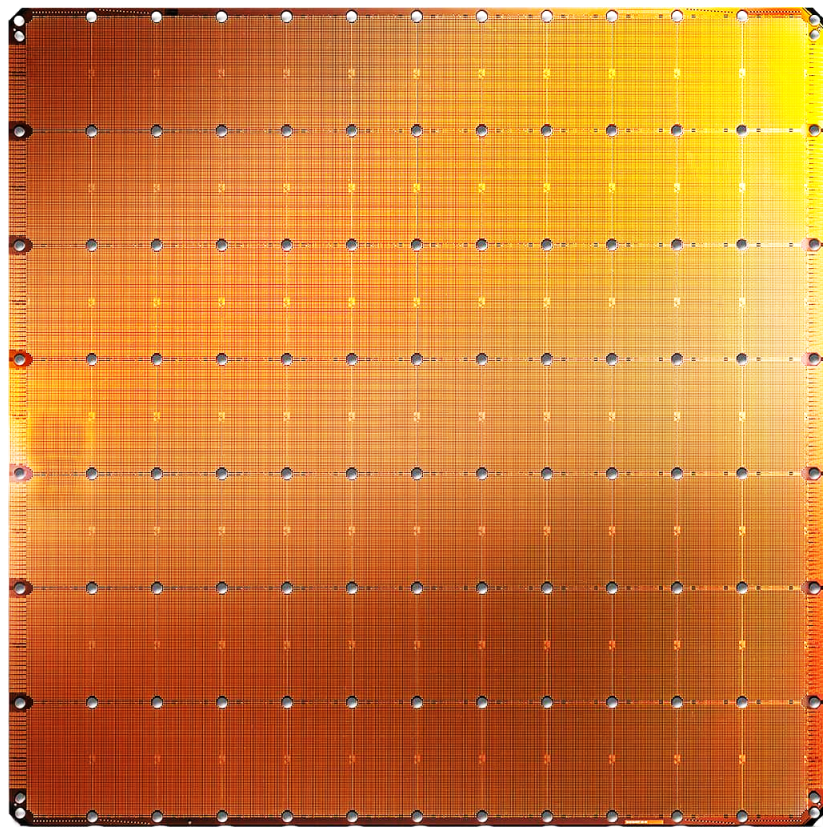
- The largest ML accelerator chip

- 400,000 cores

Cerebras WSE
1.2 Trillion transistors
46,225 mm^2

Largest GPU
21.1 Billion transistors
815 mm^2
NVIDIA TITAN V

UPMEM Processing-in-DRAM Engine (2019)

- **Processing in DRAM Engine**

- Includes **standard DIMM modules**, with a **large number of DPU processors** combined with DRAM chips.

- Replaces **standard** DIMMs
  - DDR4 R-DIMM modules
    - 8GB+128 DPUs (16 PIM chips)
    - Standard 2x-nm DRAM process
  - **Large amounts of** compute & memory bandwidth
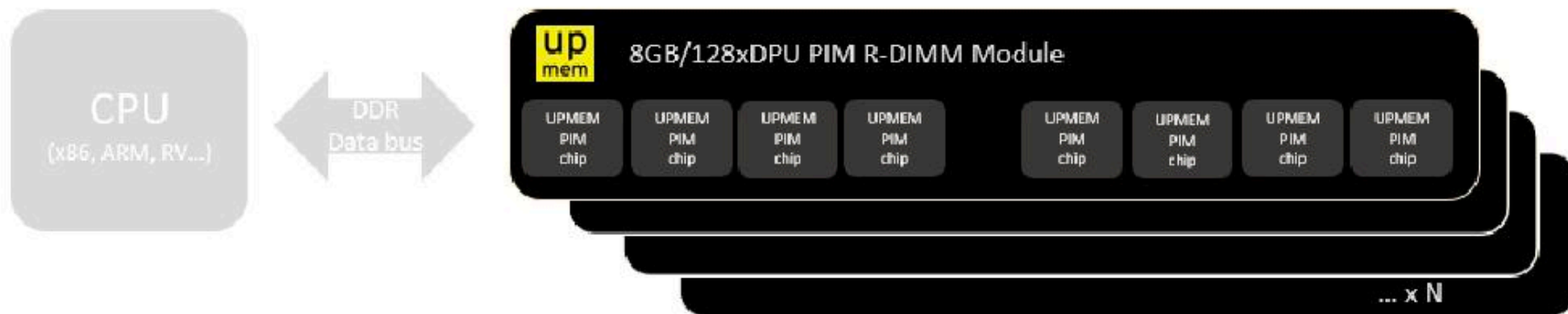
Samsung Function-in-Memory DRAM (2021)

Samsung Newsroom

Samsung Develops Industry's First High Bandwidth Memory with AI Processing Power

Korea on February 17, 2021

Audio   Share

The new architecture will deliver over twice the system performance and reduce energy consumption by more than 70%

Samsung Electronics, the world leader in advanced memory technology, today announced that it has developed the industry's first High Bandwidth Memory (HBM) integrated with artificial intelligence (AI) processing power — the HBM-PIM. The new processing-in-memory (PIM) architecture brings powerful AI computing capabilities inside high-performance memory, to accelerate large-scale processing in data centers, high performance computing (HPC) systems and AI-enabled mobile applications.

Kwangil Park, senior vice president of Memory Product Planning at Samsung Electronics stated, "Our groundbreaking HBM-PIM is the industry's first programmable PIM solution tailored for diverse AI-driven workloads such as HPC, training and inference. We plan to build upon this breakthrough by further collaborating with AI solution providers for even more advanced PIM-powered applications."

Samsung Function-in-Memory DRAM (2021)

- **FIMDRAM based on HBM2**

SID1 Core-die (HBM2)

SID0 Core-die (FIMDRAM)

Buffer-die
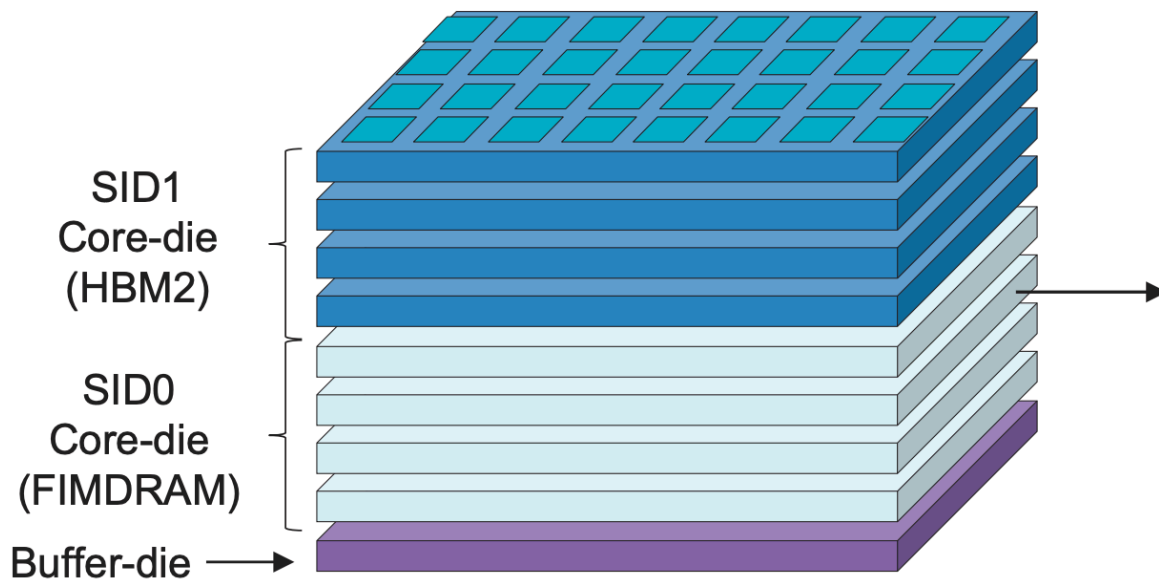
[3D Chip Structure of HBM with FIMDRAM]

Chip Specification

128DQ / 8CH / 16 banks / BL4

32 PCU blocks (1 FIM block/2 banks)

1.2 TFLOPS (4H)

**FP16 ADD /
Multiply (MUL) /
Multiply-Accumulate (MAC) /
Multiply-and- Add (MAD)**

Young-Cheon Kwon[1], Suk Han Lee[1], Jaehoon Lee[1], Sang-Hyuk Kwon[1],
Je Min Ryu[1], Jong-Pil Son[1], Seongil O[1], Hak-Soo Yu[1], Haesuk Lee[1],
Soo Young Kim[1], Youngmin Cho[1], Jin Guk Kim[1], Jongyoon Choi[1],
Hyun-Sung Shin[1], Jin Kim[1], BengSeng Phuah[1], HyoungMin Kim[1],
Myeong Jun Song[1], Ahn Choi[1], Daeho Kim[1], SooYoung Kim[1], Eun-Bong Kim[1],
David Wang[2], Shinhaeng Kang[1], Yuhwan Ro[3], Seungwoo Seo[3], JoonHo Song[3],
Jaeyoun Youn[1], Kyomin Sohn[1], Nam Sung Kim[1]

[1]Samsung Electronics, Hwaseong, Korea
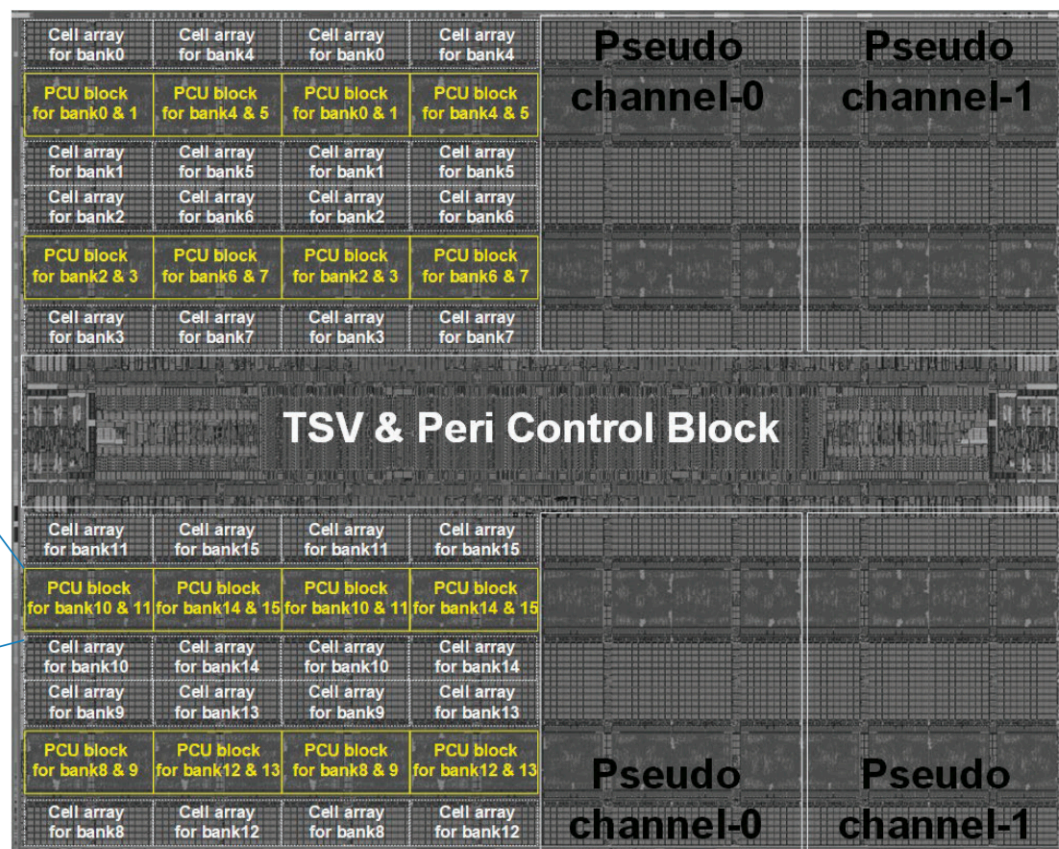[2]Samsung Electronics, San Jose, CA
[3]Samsung Electronics, Suwon, Korea

Programmable Computing Unit

- **Configuration of PCU block**
  - Interface unit to control data flow
  - Execution unit to perform operations
  - Register group
    - 32 entries of CRF for instruction memory
    - 16 GRF for weight and accumulation
    - 16 SRF to store constants for MAC operations
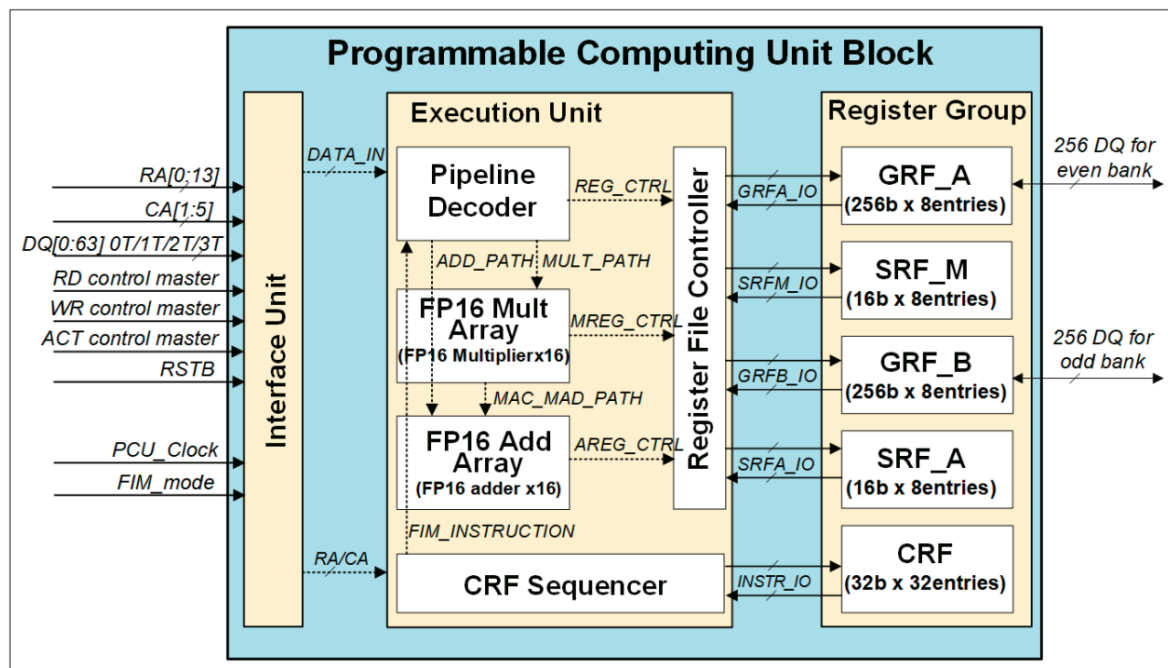
[Block diagram of PCU in FIMDRAM]

Samsung Function-in-Memory DRAM (2021)

[Available instruction list for FIM operation]

Type	CMD	Description
Floating Point	ADD	FP16 addition
	MUL	FP16 multiplication
	MAC	FP16 multiply-accumulate
	MAD	FP16 multiply and add
Data Path	MOVE	Load or store data
	FILL	Copy data from bank to GRFs
Control Path	NOP	Do nothing
	JUMP	Jump instruction
	EXIT	Exit instruction

Young-Cheon Kwon[1], Suk Han Lee[1], Jaehoon Lee[1], Sang-Hyuk Kwon[1],
Je Min Ryu[1], Jong-Pil Son[1], Seongil O[1], Hak-Soo Yu[1], Haesuk Lee[1],
Soo Young Kim[1], Youngmin Cho[1], Jin Guk Kim[1], Jongyoon Choi[1],
Hyun-Sung Shin[1], Jin Kim[1], BengSeng Phuah[1], HyoungMin Kim[1],
Myeong Jun Song[1], Ahn Choi[1], Daeho Kim[1], SooYoung Kim[1], Eun-Bong Kim[1],
David Wang[2], Shinhaeng Kang[1], Yuhwan Ro[3], Seungwoo Seo[3], JoonHo Song[3],
Jaeyoun Youn[1], Kyomin Sohn[1], Nam Sung Kim[1]

[1]Samsung Electronics, Hwaseong, Korea
[2]Samsung Electronics, San Jose, CA
[3]Samsung Electronics, Suwon, Korea

Chip Implementation

- **Mixed design methodology to implement FIMDRAM**
  - Full-custom + Digital RTL
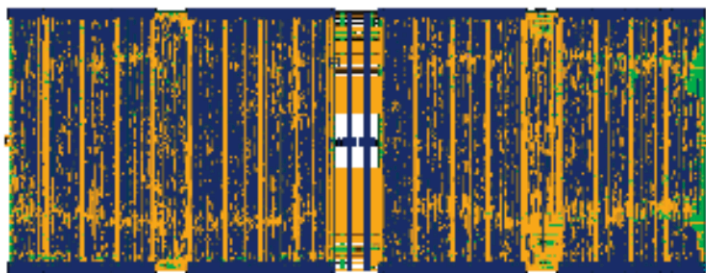
[Digital RTL design for PCU block]

Young-Cheon Kwon[1], Suk Han Lee[1], Jaehoon Lee[1], Sang-Hyuk Kwon[1], Je Min Ryu[1], Jong-Pil Son[1], Seongil O[1], Hak-Soo Yu[1], Haesuk Lee[1], Soo Young Kim[1], Youngmin Cho[1], Jin Guk Kim[1], Jongyoon Choi[1], Hyun-Sung Shin[1], Jin Kim[1], BengSeng Phuah[1], HyoungMin Kim[1], Myeong Jun Song[1], Ahn Choi[1], Daeho Kim[1], SooYoung Kim[1], Eun-Bong Kim[1], David Wang[2], Shinhaeng Kang[1], Yuhwan Ro[3], Seungwoo Seo[1], JoonHo Song[3], Jaeyoun Youn[1], Kyomin Sohn[1], Nam Sung Kim[1]

[1]Samsung Electronics, Hwaseong, Korea
[2]Samsung Electronics, San Jose, CA
[3]Samsung Electronics, Suwon, Korea

63

Specialized Processing in Memory (2015)

- Junwhan Ahn, Sungpack Hong, Sungjoo Yoo, Onur Mutlu, and Kiyoung Choi,
  "A Scalable Processing-in-Memory Accelerator for Parallel Graph Processing"
  Proceedings of the [42nd International Symposium on Computer Architecture](#) (**ISCA**), Portland, OR, June 2015.
  [[Slides (pdf)](#)] [[Lightning Session Slides (pdf)](#)]

A Scalable Processing-in-Memory Accelerator for Parallel Graph Processing

Junwhan Ahn   Sungpack Hong[§]   Sungjoo Yoo   Onur Mutlu[†]   Kiyoung Choi

junwhan@snu.ac.kr, sungpack.hong@oracle.com, sungjoo.yoo@gmail.com, onur@cmu.edu, kchoi@snu.ac.kr

Seoul National University   [§]Oracle Labs   [†]Carnegie Mellon University

Simple Processing in Memory (2015)

- Junwhan Ahn, Sungjoo Yoo, Onur Mutlu, and Kiyoung Choi,
  "PIM-Enabled Instructions: A Low-Overhead, Locality-Aware Processing-in-Memory Architecture"
  Proceedings of the 42nd International Symposium on Computer Architecture (**ISCA**), Portland, OR, June 2015.
  [Slides (pdf)] [Lightning Session Slides (pdf)]

PIM-Enabled Instructions: A Low-Overhead, Locality-Aware Processing-in-Memory Architecture

Junwhan Ahn   Sungjoo Yoo   Onur Mutlu[†]   Kiyoung Choi
junwhan@snu.ac.kr, sungjoo.yoo@gmail.com, onur@cmu.edu, kchoi@snu.ac.kr
Seoul National University   [†]Carnegie Mellon University

SAFARI

Processing in Memory on Mobile Devices

- Amirali Boroumand, Saugata Ghose, Youngsok Kim, Rachata Ausavarungnirun, Eric Shiu, Rahul Thakur, Daehyun Kim, Aki Kuusela, Allan Knies, Parthasarathy Ranganathan, and Onur Mutlu,
"Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks"
Proceedings of the 23rd International Conference on Architectural Support for Programming Languages and Operating Systems (**ASPLOS**), Williamsburg, VA, USA, March 2018.

Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks

Amirali Boroumand[1]   Saugata Ghose[1]   Youngsok Kim[2]
Rachata Ausavarungnirun[1]   Eric Shiu[3]   Rahul Thakur[3]   Daehyun Kim[4,3]
Aki Kuusela[3]   Allan Knies[3]   Parthasarathy Ranganathan[3]   Onur Mutlu[5,1]

In-DRAM Processing (2013)

- Vivek Seshadri et al., "**Ambit: In-Memory Accelerator for Bulk Bitwise Operations Using Commodity DRAM Technology**," MICRO 2017.

Ambit: In-Memory Accelerator for Bulk Bitwise Operations Using Commodity DRAM Technology

Vivek Seshadri[1,5]   Donghyuk Lee[2,5]   Thomas Mullins[3,5]   Hasan Hassan[4]   Amirali Boroumand[5]
Jeremie Kim[4,5]   Michael A. Kozuch[3]   Onur Mutlu[4,5]   Phillip B. Gibbons[5]   Todd C. Mowry[5]

[1]Microsoft Research India   [2]NVIDIA Research   [3]Intel   [4]ETH Zürich   [5]Carnegie Mellon University

In-DRAM Bulk Bitwise Execution (2017)

- Vivek Seshadri and Onur Mutlu,
  "In-DRAM Bulk Bitwise Execution Engine"
  Invited Book Chapter in Advances in Computers, to appear
  in 2020.
  [Preliminary arXiv version]

In-DRAM Bulk Bitwise Execution Engine

Vivek Seshadri   Onur Mutlu
Microsoft Research India   ETH Zürich
visesha@microsoft.com   onur.mutlu@inf.ethz.ch

Coming Up Next Month @ ASPLOS 2021...

SIMDRAM: A Framework for Bit-Serial SIMD Processing Using DRAM
Extended Abstract

Nastaran Hajinazar[◇]   *Geraldo F. Oliveira[◇]   Sven Gregorio[◇]   João Dinis Ferreira[◇]   Nika Mansouri Ghiasi[◇]

Minesh Patel[◇]   Mohammed Alser[◇]   Saugata Ghose[⊙]   Juan Gómez-Luna[◇]   Onur Mutlu[◇]

[◇]*ETH Zürich*   [*]*Simon Fraser University*   [⊙]*University of Illinois at Urbana–Champaign*

Coming Up Next Week @ HPCA 2021…

- Christina Giannoula, Nandita Vijaykumar, Nikela Papadopoulou, Vasileios Karakostas, Ivan Fernandez, Juan Gómez-Luna, Lois Orosa, Nectarios Koziris, Georgios Goumas, and Onur Mutlu,
  "SynCron: Efficient Synchronization Support for Near-Data-Processing Architectures"
  Proceedings of the 27th International Symposium on High-Performance Computer Architecture (**HPCA**), Virtual, February-March 2021.

SynCron: Efficient Synchronization Support for Near-Data-Processing Architectures

Christina Giannoula[†‡]   Nandita Vijaykumar[*‡]   Nikela Papadopoulou[†]   Vasileios Karakostas[†]   Ivan Fernandez[§‡]

Juan Gómez-Luna[‡]   Lois Orosa[‡]   Nectarios Koziris[†]   Georgios Goumas[†]   Onur Mutlu[‡]

[†]*National Technical University of Athens*   [‡]*ETH Zürich*   [*]*University of Toronto*   [§]*University of Malaga*

SAFARI

A Modern Primer on Processing in Memory

Onur Mutlu[a,b], Saugata Ghose[b,c], Juan Gómez-Luna[a], Rachata Ausavarungnirun[d]

SAFARI Research Group

[a]ETH Zürich
[b]Carnegie Mellon University
[c]University of Illinois at Urbana-Champaign
[d]King Mongkut's University of Technology North Bangkok

Onur Mutlu, Saugata Ghose, Juan Gomez-Luna, and Rachata Ausavarungnirun,
"A Modern Primer on Processing in Memory"
*Invited Book Chapter in **Emerging Computing: From Devices to Systems - Looking Beyond Moore and Von Neumann**, Springer, to be published in 2021.*

A Tutorial on PIM

- Onur Mutlu,
  "Memory-Centric Computing Systems"
  Invited Tutorial at *66th International Electron Devices Meeting (**IEDM**)*, Virtual, 12 December 2020.
  [Slides (pptx) (pdf)]
  [Executive Summary Slides (pptx) (pdf)]
  [Tutorial Video (1 hour 51 minutes)]
  [Executive Summary Video (2 minutes)]
  [Abstract and Bio]
  [Related Keynote Paper from VLSI-DAT 2020]
  [Related Review Paper on Processing in Memory]

  https://www.youtube.com/watch?v=H3sEaINPBOE

https://www.youtube.com/onurmutlulectures

73

Detailed Lectures on PIM (I)

- Computer Architecture, Fall 2020, Lecture 6
  - **Computation in Memory** (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=oGcZAGwfEUE&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN&index=12

- Computer Architecture, Fall 2020, Lecture 7
  - **Near-Data Processing** (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=j2GIigqn1Qw&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN&index=13

- Computer Architecture, Fall 2020, Lecture 11a
  - **Memory Controllers** (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=TeG773OgiMQ&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN&index=20

- Computer Architecture, Fall 2020, Lecture 12d
  - **Real Processing-in-DRAM with UPMEM** (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=Sscy1Wrr22A&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN&index=25

https://www.youtube.com/onurmutlulectures

Detailed Lectures on PIM (II)

- Computer Architecture, Fall 2020, Lecture 15
  - **Emerging Memory Technologies** (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=AlE1rD9G_YU&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN&index=28

- Computer Architecture, Fall 2020, Lecture 16a
  - **Opportunities & Challenges of Emerging Memory Technologies** (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=pmLszWGmMGQ&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN&index=29

- Computer Architecture, Fall 2020, Guest Lecture
  - **In-Memory Computing: Memory Devices & Applications** (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=wNmqQHiEZNk&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN&index=41

TESLA Full Self-Driving Computer (2019)

- ML accelerator: 260 mm^2, 6 billion transistors, 600 GFLOPS GPU, 12 ARM 2.2 GHz CPUs.
- Two redundant chips for better safety.
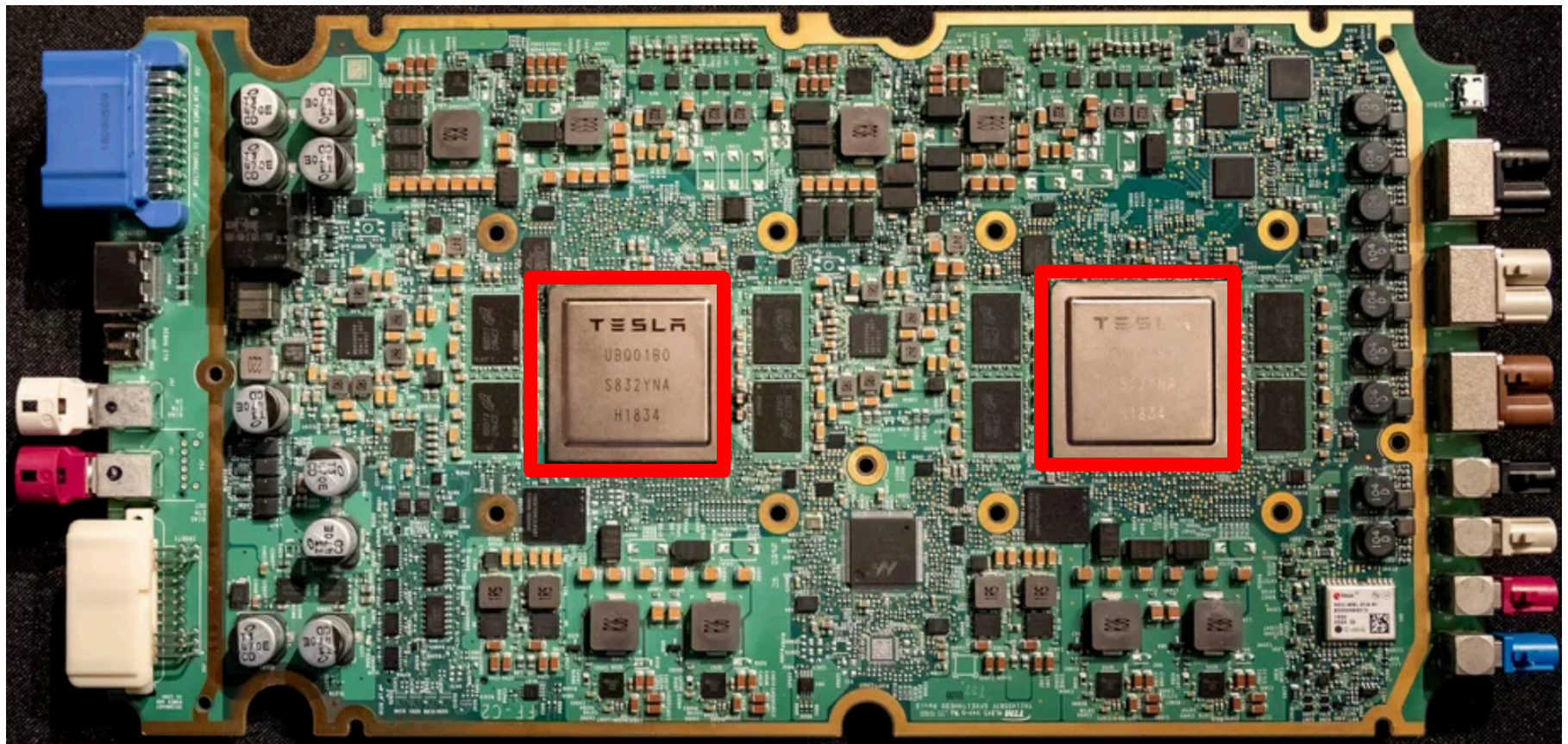
Google TPU Generation I (~2016)
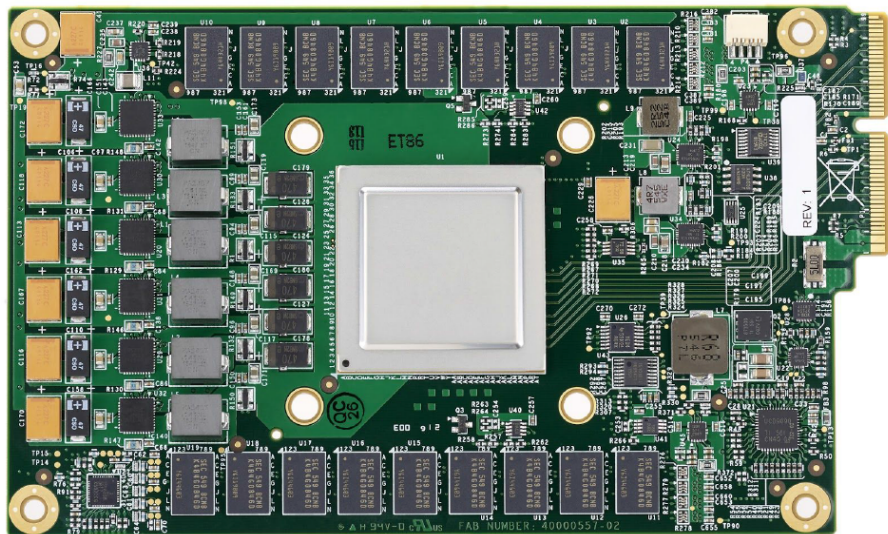
Figure 3. TPU Printed Circuit Board. It can be inserted in the slot for an SATA disk in a server, but the card uses PCIe Gen3 x16.
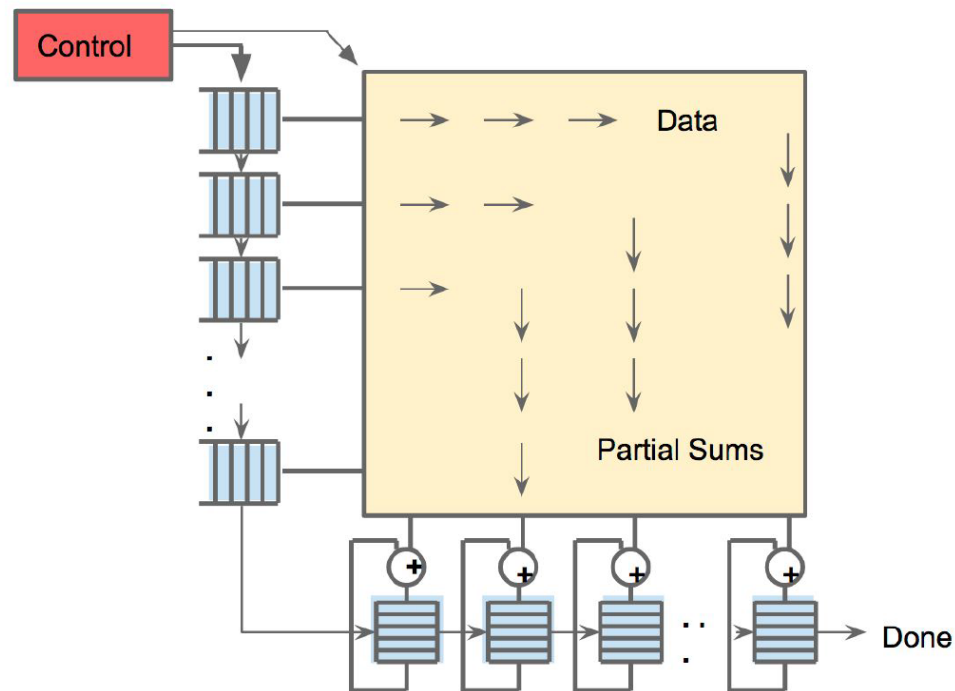
Figure 4. Systolic data flow of the Matrix Multiply Unit. Software has the illusion that each 256B input is read at once, and they instantly update one location of each of 256 accumulator RAMs.

Jouppi et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit", ISCA 2017.

Google TPU Generation II (2017)

https://www.nextplatform.com/2017/05/17/first-depth-look-googles-new-second-generation-tpu/
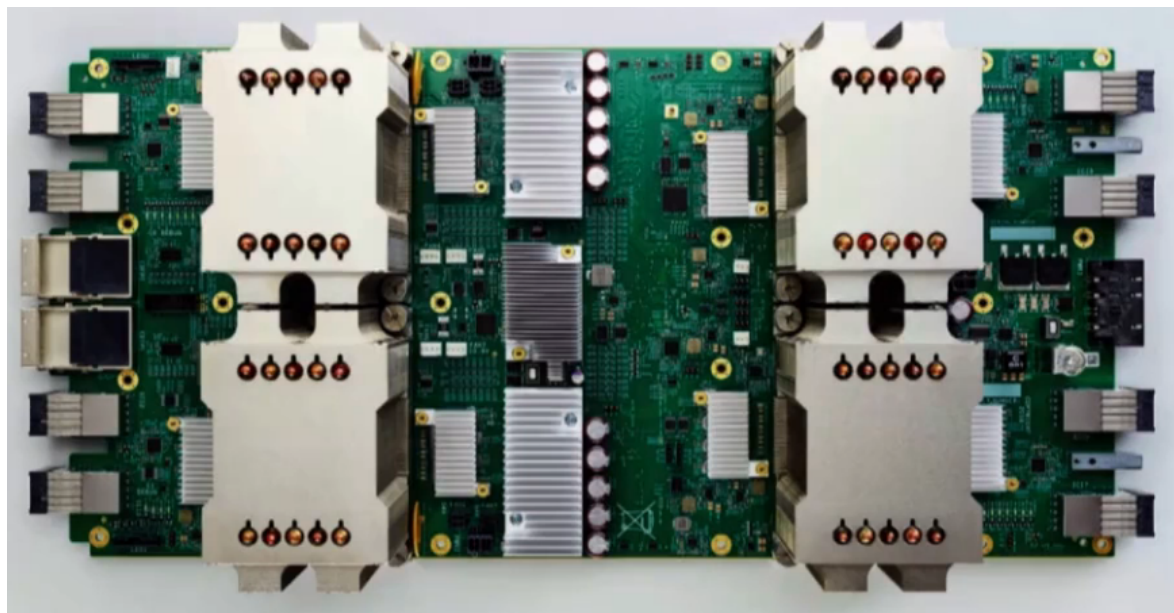
4 TPU chips
vs 1 chip in TPU1

High Bandwidth Memory
vs DDR3

Floating point operations
vs FP16

45 TFLOPS per chip
vs 23 TOPS

Designed for training
and inference
vs only inference

An Example Modern Systolic Array: TPU (II)

As reading a large SRAM uses much more power than arithmetic, the matrix unit uses systolic execution to save energy by reducing reads and writes of the Unified Buffer [Kun80][Ram91][Ovt15b]. Figure 4 shows that data flows in from the left, and the weights are loaded from the top. A given 256-element multiply-accumulate operation moves through the matrix as a diagonal wavefront. 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 256 accumulators. From a correctness perspective, software is unaware of the systolic nature of the matrix unit, but for performance, it does worry about the latency of the unit.
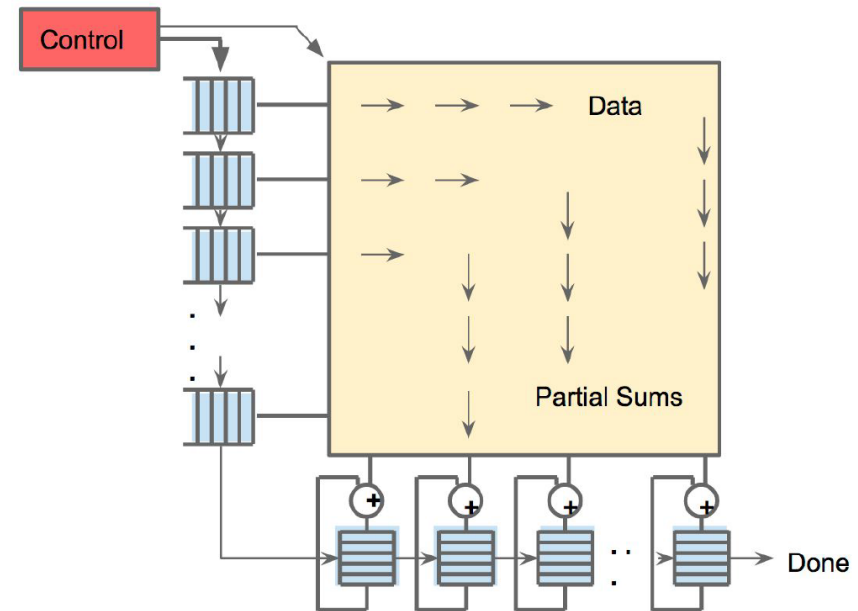
Jouppi et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit", ISCA 2017.

An Example Modern Systolic Array: TPU (III)
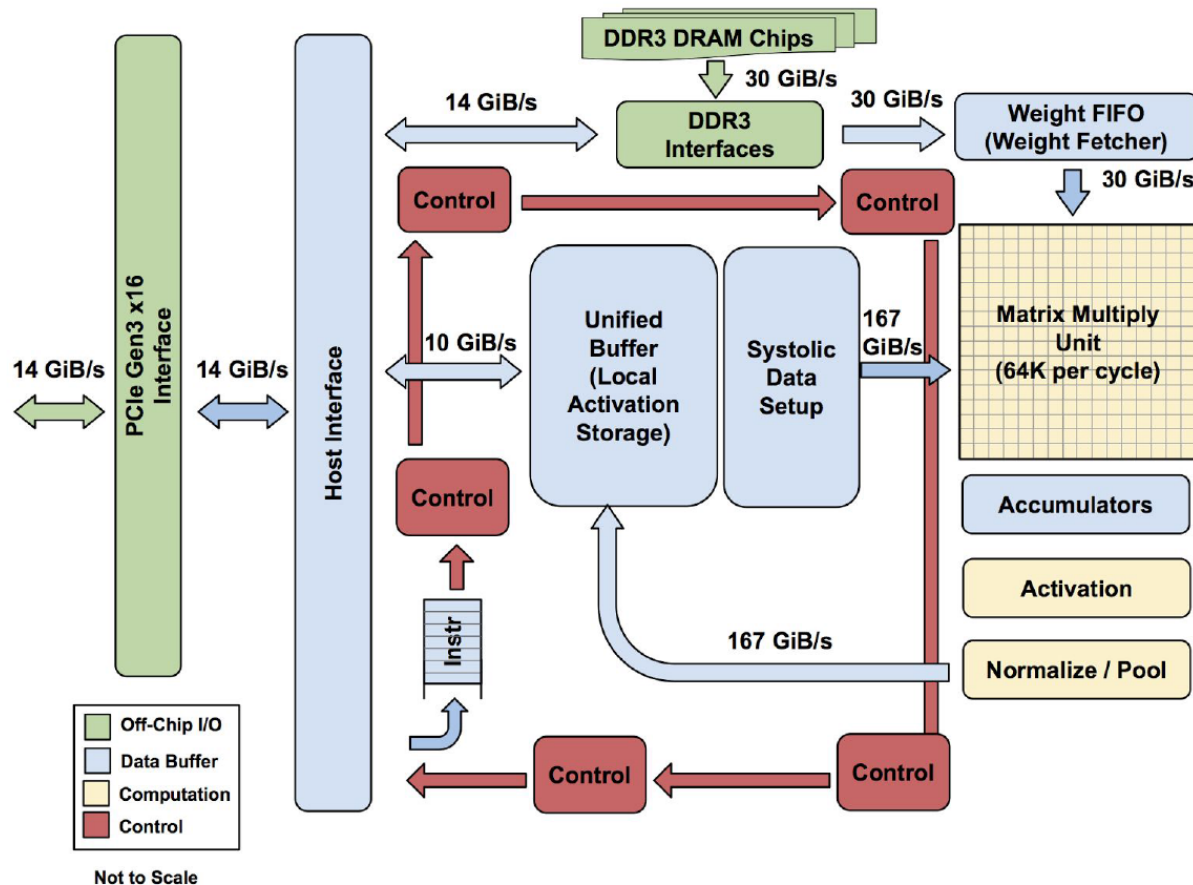
Figure 1. TPU Block Diagram. The main computation part is the yellow Matrix Multiply unit in the upper right hand corner. Its inputs are the blue Weight FIFO and the blue Unified Buffer (UB) and its output is the blue Accumulators (Acc). The yellow Activation Unit performs the nonlinear functions on the Acc, which go to the UB.

Many (Other) AI/ML Chips

- Alibaba
- Amazon
- Facebook
- Google
- Huawei
- Intel
- Microsoft
- NVIDIA
- Tesla
- Many Others and Many Startups…

- **Many More to Come…**

SAFARI

Many (Other) AI/ML Chips

AI Chip Landscape

https://basicmi.github.io/AI-Chip/

Many Interesting Things Are Happening Today in Computer Architecture

Many Interesting Things Are Happening Today in Computer Architecture

Reliability
and
Security

The Story of RowHammer

- One can predictably induce bit flips in commodity DRAM chips
  - >80% of the tested DRAM chips are vulnerable

- First example of how a simple hardware failure mechanism can create a widespread system security vulnerability

WIRED   Forget Software—Now Hackers Are Exploiting Physics

| BUSINESS | CULTURE | DESIGN | GEAR | SCIENCE |

ANDY GREENBERG   SECURITY   08.31.16   7:00 AM

SHARE

FORGET SOFTWARE—NOW HACKERS ARE EXPLOITING PHYSICS

SHARE
18276

TWEET

Modern DRAM is Prone to Disturbance Errors
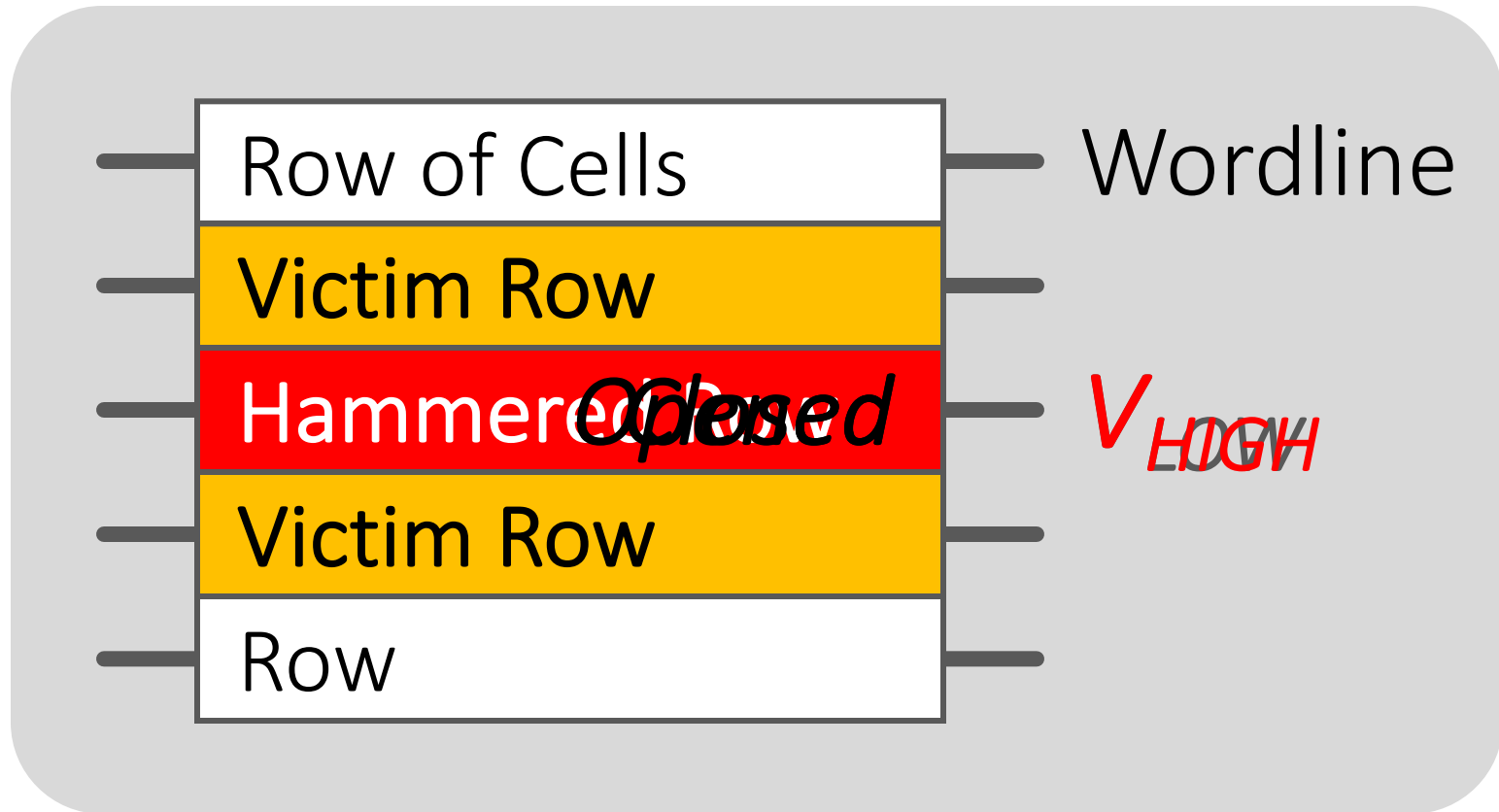
Repeatedly reading a row enough times (before memory gets refreshed) induces disturbance errors in **adjacent rows** in most real DRAM chips you can buy today

Flipping Bits in Memory Without Accessing Them: An Experimental Study of DRAM Disturbance Errors, (Kim et al., ISCA 2014)

Most DRAM Modules Are Vulnerable

A company   **B** company   **C** company
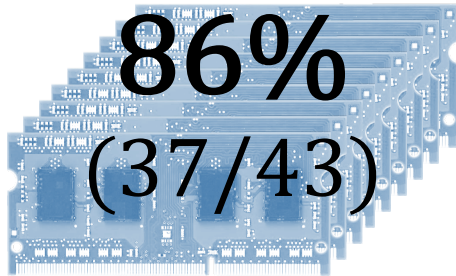
86% (37/43)   **83%** (45/54)   **88%** (28/32)

Up to

1.0×10^7 errors

Up to

2.7×10^6 errors

Up to

3.3×10^5 errors

Flipping Bits in Memory Without Accessing Them: An Experimental Study of DRAM Disturbance Errors, (Kim et al., ISCA 2014)

88
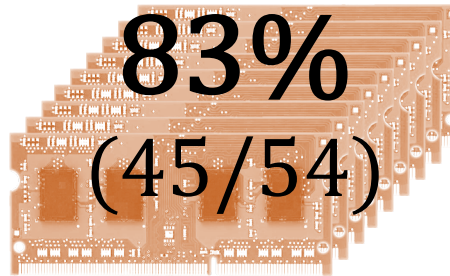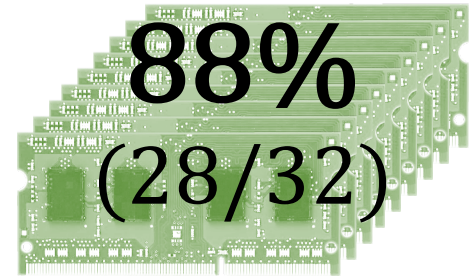
One Can Take Over an Otherwise-Secure System

Flipping Bits in Memory Without Accessing Them: An Experimental Study of DRAM Disturbance Errors

Abstract. Memory isolation is a key property of a reliable and secure computing system — an access to one memory address should not have unintended side effects on data stored in other addresses. However, as DRAM process technology

Project Zero

Flipping Bits in Memory Without Accessing Them: An Experimental Study of DRAM Disturbance Errors (Kim et al., ISCA 2014)

News and updates from the Project Zero team at Google

Exploiting the DRAM rowhammer bug to gain kernel privileges (Seaborn+, 2015)

Monday, March 9, 2015

Exploiting the DRAM rowhammer bug to gain kernel privileges

Security: RowHammer (2014)

Rowhammer

It's like breaking into an apartment by repeatedly slamming a neighbor's door until the vibrations open the door you were after

More Security Implications (I)

"We can gain unrestricted access to systems of website visitors."

Rowhammer.js: A Remote Software-Induced Fault Attack in JavaScript (DIMVA'16)

Source: https://lab.dsst.io/32c3-slides/7197.html

More Security Implications (II)
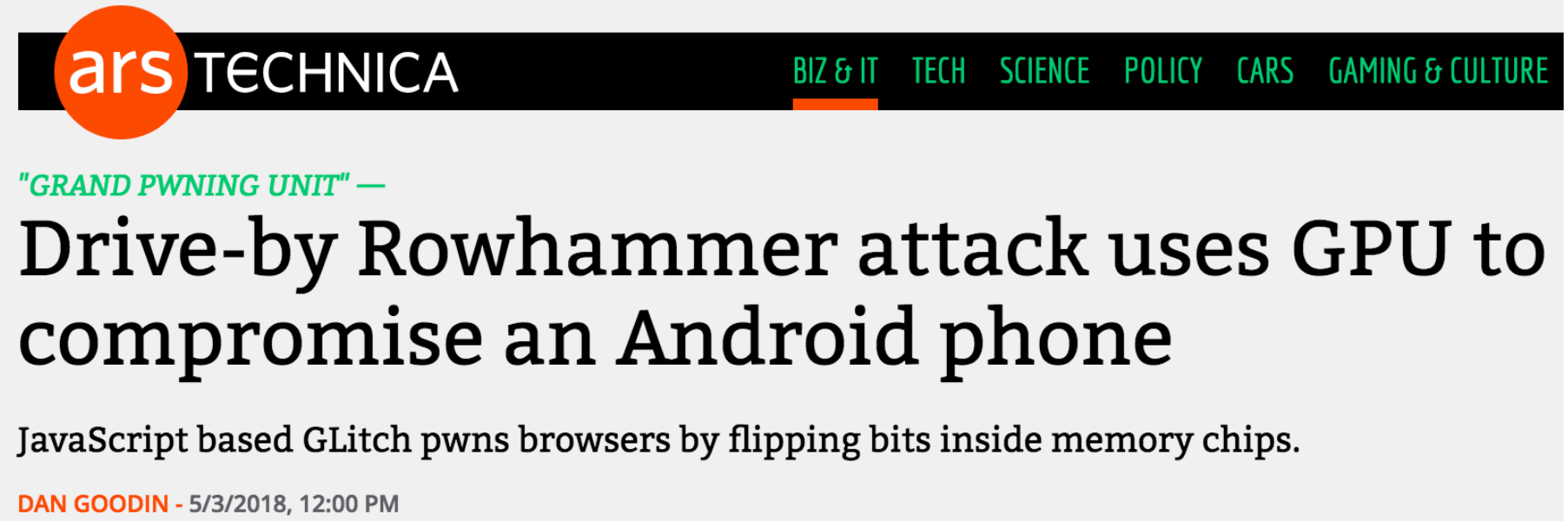
"Can gain control of a smart phone deterministically"

Hammer And Root

Millions of Androids

Drammer: Deterministic Rowhammer Attacks on Mobile Platforms, CCS'16

92

More Security Implications (III)

- Using an integrated GPU in a mobile system to remotely escalate privilege via the WebGL interface

ars TECHNICA   BIZ & IT   TECH   SCIENCE   POLICY   CARS   GAMING & CULTURE

"GRAND PWNING UNIT" —

Drive-by Rowhammer attack uses GPU to compromise an Android phone

JavaScript based GLitch pwns browsers by flipping bits inside memory chips.

DAN GOODIN - 5/3/2018, 12:00 PM

Grand Pwning Unit: Accelerating Microarchitectural Attacks with the GPU

Pietro Frigo
Vrije Universiteit
Amsterdam
p.frigo@vu.nl

Cristiano Giuffrida
Vrije Universiteit
Amsterdam
giuffrida@cs.vu.nl

Herbert Bos
Vrije Universiteit
Amsterdam
herbertb@cs.vu.nl

Kaveh Razavi
Vrije Universiteit
Amsterdam
kaveh@cs.vu.nl

More Security Implications (IV)

- Rowhammer over RDMA (I)

THROWHAMMER —

Packets over a LAN are all it takes to trigger serious Rowhammer bit flips

The bar for exploiting potentially serious DDR weakness keeps getting lower.

DAN GOODIN - 5/10/2018, 5:26 PM

Throwhammer: Rowhammer Attacks over the Network and Defenses

Andrei Tatar
VU Amsterdam

Radhesh Krishnan
VU Amsterdam

Elias Athanasopoulos
University of Cyprus

Cristiano Giuffrida
VU Amsterdam

Herbert Bos
VU Amsterdam

Kaveh Razavi
VU Amsterdam

More Security Implications (V)

- Rowhammer over RDMA (II)

Nethammer—Exploiting DRAM Rowhammer Bug Through Network Requests

Nethammer:
Inducing Rowhammer Faults through Network Requests

Moritz Lipp
Graz University of Technology

Misiker Tadesse Aga
University of Michigan

Michael Schwarz
Graz University of Technology

Daniel Gruss
Graz University of Technology

Clémentine Maurice
Univ Rennes, CNRS, IRISA

Lukas Raab
Graz University of Technology

Lukas Lamster
Graz University of Technology

More Security Implications (VI)

-
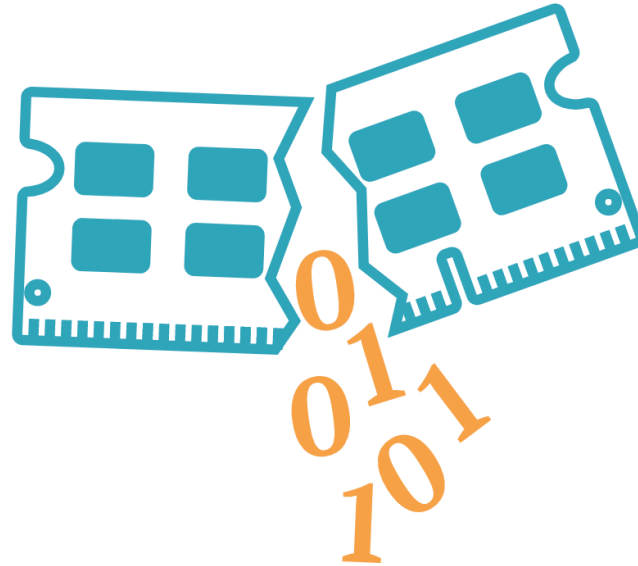
RAMBleed

RAMBleed: Reading Bits in Memory Without Accessing Them

Andrew Kwong
University of Michigan
ankwong@umich.edu

Daniel Genkin
University of Michigan
genkin@umich.edu

Daniel Gruss
Graz University of Technology
daniel.gruss@iaik.tugraz.at

Yuval Yarom
University of Adelaide and Data61
yval@cs.adelaide.edu.au

More Security Implications (VII)

- **USENIX Security 2019**

Terminal Brain Damage: Exposing the Graceless Degradation in Deep Neural Networks Under Hardware Fault Attacks

Sanghyun Hong, Pietro Frigo[†], Yiğitcan Kaya, Cristiano Giuffrida[†], Tudor Dumitraş

University of Maryland, College Park
[†]Vrije Universiteit Amsterdam

A Single Bit-flip Can Cause Terminal Brain Damage to DNNs

One specific bit-flip in a DNN's representation leads to accuracy drop over 90%

Our research found that a specific bit-flip in a DNN's bitwise representation can cause the accuracy loss up to 90%, and the DNN has 40-50% parameters, on average, that can lead to the accuracy drop over 10% when individually subjected to such single bitwise corruptions...

Read More

More Security Implications (VIII)

- **USENIX Security 2020**

DeepHammer: Depleting the Intelligence of Deep Neural Networks through Targeted Chain of Bit Flips

Fan Yao
University of Central Florida
fan.yao@ucf.edu

Adnan Siraj Rakin
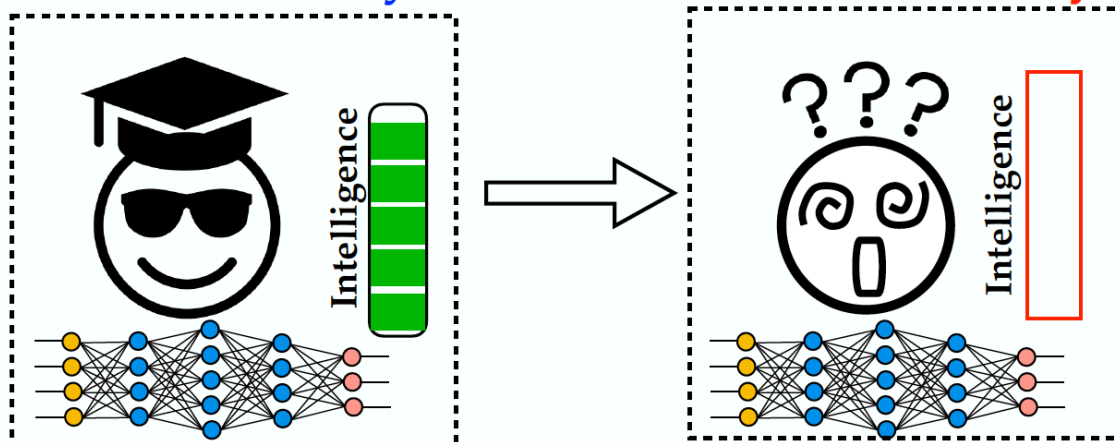*Arizona State University*
asrakin@asu.edu

Deliang Fan
dfan@asu.edu

Degrade the **inference accuracy** to the level of **Random Guess**

Example: ResNet-20 for CIFAR-10, **10** output classes

Before attack, **Accuracy: 90.2%** After attack, **Accuracy: ~10% (1/10)**

RowHammer: Seven Years Ago…

- Yoongu Kim, Ross Daly, Jeremie Kim, Chris Fallin, Ji Hye Lee, Donghyuk Lee, Chris Wilkerson, Konrad Lai, and Onur Mutlu,
"Flipping Bits in Memory Without Accessing Them: An Experimental Study of DRAM Disturbance Errors"
Proceedings of the 41st International Symposium on Computer Architecture (**ISCA**), Minneapolis, MN, June 2014.
[Slides (pptx) (pdf)] [Lightning Session Slides (pptx) (pdf)] [Source Code and Data]

Flipping Bits in Memory Without Accessing Them: An Experimental Study of DRAM Disturbance Errors

Yoongu Kim[1]   Ross Daly*   Jeremie Kim[1]   Chris Fallin*   Ji Hye Lee[1]
Donghyuk Lee[1]   Chris Wilkerson[2]   Konrad Lai   Onur Mutlu[1]

[1]Carnegie Mellon University   [2]Intel Labs

RowHammer: Now and Beyond...

- Onur Mutlu and Jeremie Kim,
  "RowHammer: A Retrospective"
  IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (**TCAD**) *Special Issue on Top Picks in Hardware and Embedded Security*, 2019.
  [Preliminary arXiv version]
  [Slides from COSADE 2019 (pptx)]
  [Slides from VLSI-SOC 2020 (pptx) (pdf)]
  [Talk Video (30 minutes)]

RowHammer: A Retrospective

Onur Mutlu[§‡]   Jeremie S. Kim[‡§]
[§]ETH Zürich   [‡]Carnegie Mellon University

RowHammer in 2020

RowHammer in 2020 (I)

- Jeremie S. Kim, Minesh Patel, A. Giray Yaglikci, Hasan Hassan, Roknoddin Azizi, Lois Orosa, and Onur Mutlu,
"Revisiting RowHammer: An Experimental Analysis of Modern Devices and Mitigation Techniques"
Proceedings of the 47th International Symposium on Computer Architecture (**ISCA**), Valencia, Spain, June 2020.
[Slides (pptx) (pdf)]
[Lightning Talk Slides (pptx) (pdf)]
[Talk Video (20 minutes)]
[Lightning Talk Video (3 minutes)]

Revisiting RowHammer: An Experimental Analysis of Modern DRAM Devices and Mitigation Techniques

Jeremie S. Kim[§†]   Minesh Patel[§]   A. Giray Yağlıkçı[§]

Hasan Hassan[§]   Roknoddin Azizi[§]   Lois Orosa[§]   Onur Mutlu[§†]

[§]*ETH Zürich*   [†]*Carnegie Mellon University*

RowHammer in 2020 (II)

- Pietro Frigo, Emanuele Vannacci, Hasan Hassan, Victor van der Veen, Onur Mutlu, Cristiano Giuffrida, Herbert Bos, and Kaveh Razavi,
  "TRRespass: Exploiting the Many Sides of Target Row Refresh"
  Proceedings of the 41st IEEE Symposium on Security and Privacy (**S&P**), San Francisco, CA, USA, May 2020.
  [Slides (pptx) (pdf)]
  [Lecture Slides (pptx) (pdf)]
  [Talk Video (17 minutes)]
  [Lecture Video (59 minutes)]
  [Source Code]
  [Web Article]
  Best paper award.
  Pwnie Award 2020 for Most Innovative Research. Pwnie Awards 2020

TRRespass: Exploiting the Many Sides of Target Row Refresh

Pietro Frigo*[†]   Emanuele Vannacci*[†]   Hasan Hassan[§]   Victor van der Veen[¶]
Onur Mutlu[§]   Cristiano Giuffrida*   Herbert Bos*   Kaveh Razavi*

*Vrije Universiteit Amsterdam   [§]ETH Zürich   [¶]Qualcomm Technologies Inc.

RowHammer in 2020 (III)

- Lucian Cojocar, Jeremie Kim, Minesh Patel, Lillian Tsai, Stefan Saroiu, Alec Wolman, and Onur Mutlu,
  "Are We Susceptible to Rowhammer? An End-to-End Methodology for Cloud Providers"
  Proceedings of the *41st IEEE Symposium on Security and Privacy* (**S&P**), San Francisco, CA, USA, May 2020.
  [Slides (pptx) (pdf)]
  [Talk Video (17 minutes)]

Are We Susceptible to Rowhammer? An End-to-End Methodology for Cloud Providers

Lucian Cojocar, Jeremie Kim[§†], Minesh Patel[§], Lillian Tsai[‡],
Stefan Saroiu, Alec Wolman, and Onur Mutlu[§†]
Microsoft Research, [§]ETH Zürich, [†]CMU, [‡]MIT

Coming Up Next Week @ HPCA 2021…
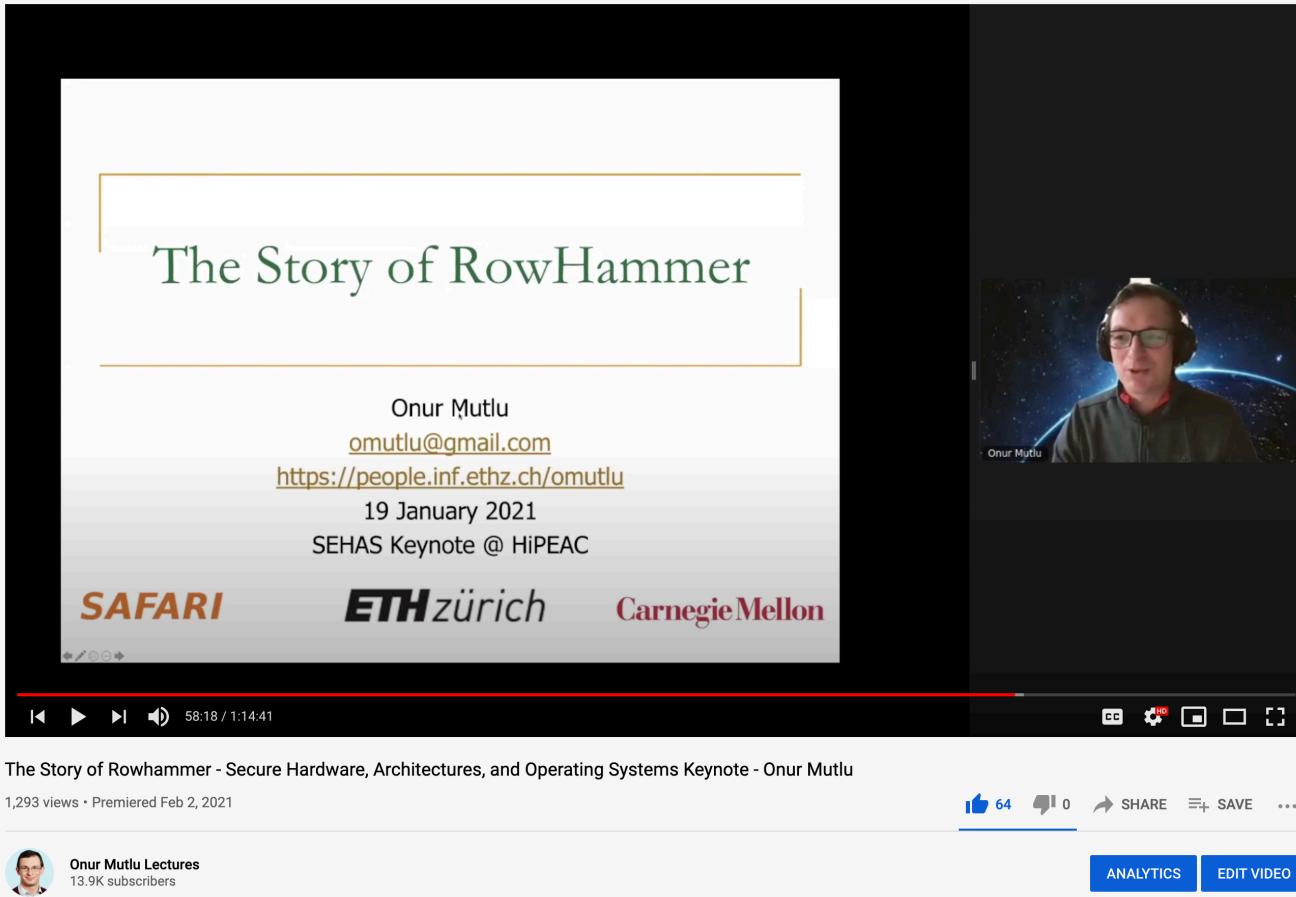
- A. Giray Yaglikci, Minesh Patel, Jeremie S. Kim, Roknoddin Azizi, Ataberk Olgun, Lois Orosa, Hasan Hassan, Jisung Park, Konstantinos Kanellopoulos, Taha Shahroodi, Saugata Ghose, and Onur Mutlu, **"BlockHammer: Preventing RowHammer at Low Cost by Blacklisting Rapidly-Accessed DRAM Rows"** *Proceedings of the 27th International Symposium on High-Performance Computer Architecture* (**HPCA**), Virtual, February-March 2021.

BlockHammer: Preventing RowHammer at Low Cost by Blacklisting Rapidly-Accessed DRAM Rows

A. Giray Yağlıkçı[1]   Minesh Patel[1]   Jeremie S. Kim[1]   Roknoddin Azizi[1]   Ataberk Olgun[1]   Lois Orosa[1]

Hasan Hassan[1]   Jisung Park[1]   Konstantinos Kanellopoulos[1]   Taha Shahroodi[1]   Saugata Ghose[2]   Onur Mutlu[1]

[1]*ETH Zürich*   [2]*University of Illinois at Urbana–Champaign*

SAFARI

The Story of RowHammer Lecture ...

- Onur Mutlu,
  "The Story of RowHammer"
  Keynote Talk at *Secure Hardware, Architectures, and Operating Systems Workshop* (**SeHAS**), *held with HiPEAC 2021 Conference*, Virtual, 19 January 2021.
  [Slides (pptx) (pdf)]
  [Talk Video (1 hr 15 minutes, with Q&A)]

The Story of Rowhammer - Secure Hardware, Architectures, and Operating Systems Keynote - Onur Mutlu

1,293 views • Premiered Feb 2, 2021

y

Detailed Lectures on RowHammer

- Computer Architecture, Fall 2020, Lecture 4b
  - RowHammer (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=KDy632z23UE&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN&index=8

- Computer Architecture, Fall 2020, Lecture 5a
  - RowHammer in 2020: TRRespass (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=pwRw7QqK_qA&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN&index=9

- Computer Architecture, Fall 2020, Lecture 5b
  - RowHammer in 2020: Revisiting RowHammer (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=gR7XR-Eepcg&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN&index=10

- Computer Architecture, Fall 2020, Lecture 5c
  - Secure and Reliable Memory (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=HvswnsfG3oQ&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN&index=11

Rowhammer

Security: Meltdown and Spectre (2018)

Source: J. Masters, Redhat, FOSDEM 2018 keynote talk.
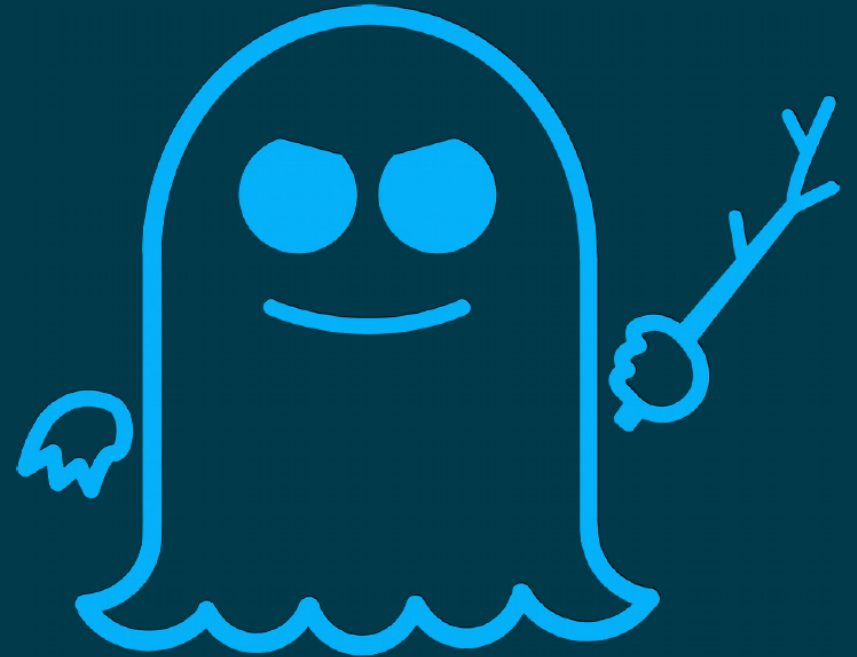
Meltdown and Spectre

- Someone can steal secret data from the system even though
  - your program and data are perfectly correct and
  - your hardware behaves according to the specification and
  - there are no software vulnerabilities/bugs

- Why?
  - Speculative execution leaves traces of secret data in the processor's cache (internal storage)
    - It brings data that is not supposed to be brought/accessed if there was no speculative execution
  - A malicious program can inspect the contents of the cache to "infer" secret data that it is not supposed to access
  - A malicious program can actually force another program to speculatively execute code that leaves traces of secret data

Project Zero

News and updates from the Project Zero team at Google

Wednesday, January 3, 2018

Reading privileged memory with a side-channel

Posted by Jann Horn, Project Zero

We have discovered that CPU data cache timing can be abused to efficiently leak information out of mis-speculated execution, leading to (at worst) arbitrary virtual memory read vulnerabilities across local security boundaries in various contexts.

Source: https://googleprojectzero.blogspot.ch/2018/01/reading-privileged-memory-with-side.html

Many Interesting Things Are Happening Today in Computer Architecture

Many Interesting Things Are Happening Today in Computer Architecture

More Demanding Workloads

Increasingly Demanding Applications

Dream

and, they will come

As applications push boundaries, computing platforms will become increasingly strained.

New Genome Sequencing Technologies

Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions

Damla Senol Cali ✉, Jeremie S Kim, Saugata Ghose, Can Alkan, Onur Mutlu

Oxford Nanopore MinION

Data → performance & energy bottleneck

Why Do We Care? An Example

200 Oxford Nanopore sequencers have left UK for China, to support rapid, near-sample coronavirus sequencing for outbreak surveillance

Fri 31st January 2020

Following extensive support of, and collaboration with, public health professionals in China, Oxford Nanopore has shipped an additional 200 MinION sequencers and related consumables to China. These will be used to support the ongoing surveillance of the current coronavirus outbreak, adding to a large number of the devices already installed in the country.

Each MinION sequencer is approximately the size of a stapler, and can provide rapid sequence information about the coronavirus.

700Kg of Oxford Nanopore sequencers and consumables are on their way for use by Chinese scientists in understanding the current coronavirus outbreak.

SAFARI

116

Population-Scale Microbiome Profiling

City-Scale Microbiome Profiling
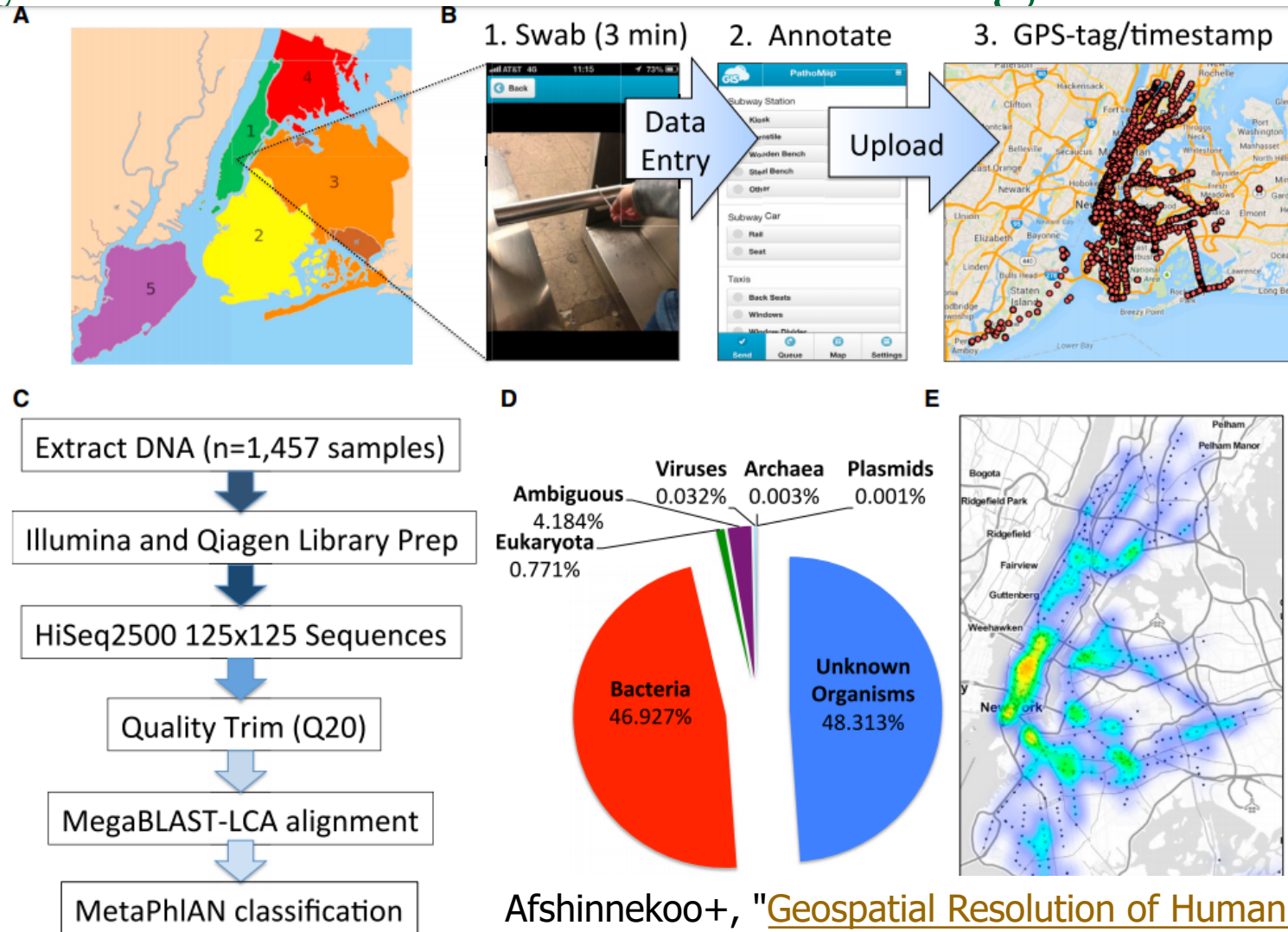
Figure 1. The Metagenome of New York City

(A) The five boroughs of NYC include (1) Manhattan (green)
(B) The collection from the 466 subway stations of NYC across the 24 subway lines involved three main steps: (1) collection with Copan Elution swabs, (2) data entry into the database, and (3) uploading of the data. An image is shown of the current collection database, taken from http://pathomap.giscloud.com.
(C) Workflow for sample DNA extraction, library preparation, sequencing, quality trimming of the FASTQ files, and alignment with MegaBLAST and MetaPhlAn to discern taxa present

Afshinnekoo+, "Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics", Cell Systems, 2015

Example: Rapid Surveillance of Ebola Outbreak

Figure 1: Deployment of the portable genome surveillance system in Guinea.

Quick+, "Real-time, portable genome sequencing for Ebola surveillance", *Nature*, 2016

High-Throughput Genome Sequencers

Illumina MiSeq

Pacific Biosciences Sequel II

Oxford Nanopore PromethION

Oxford Nanopore MinION

Oxford Nanopore SmidgION

Illumina NovaSeq 6000

Pacific Biosciences RS II

… and more! All produce data with different properties.

The Genomic Era

Cost per Raw Megabase of DNA Sequence
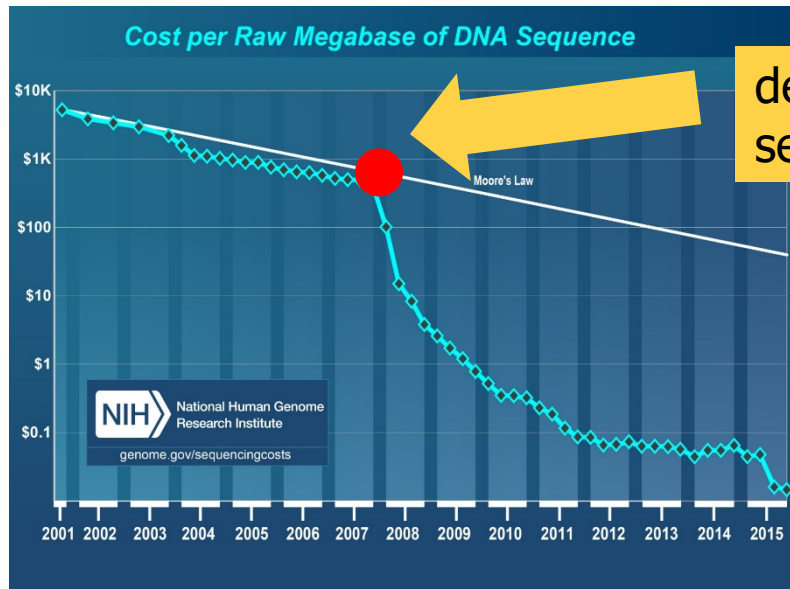
development of high-throughput sequencing (HTS) technologies

Number of Genomes Sequenced
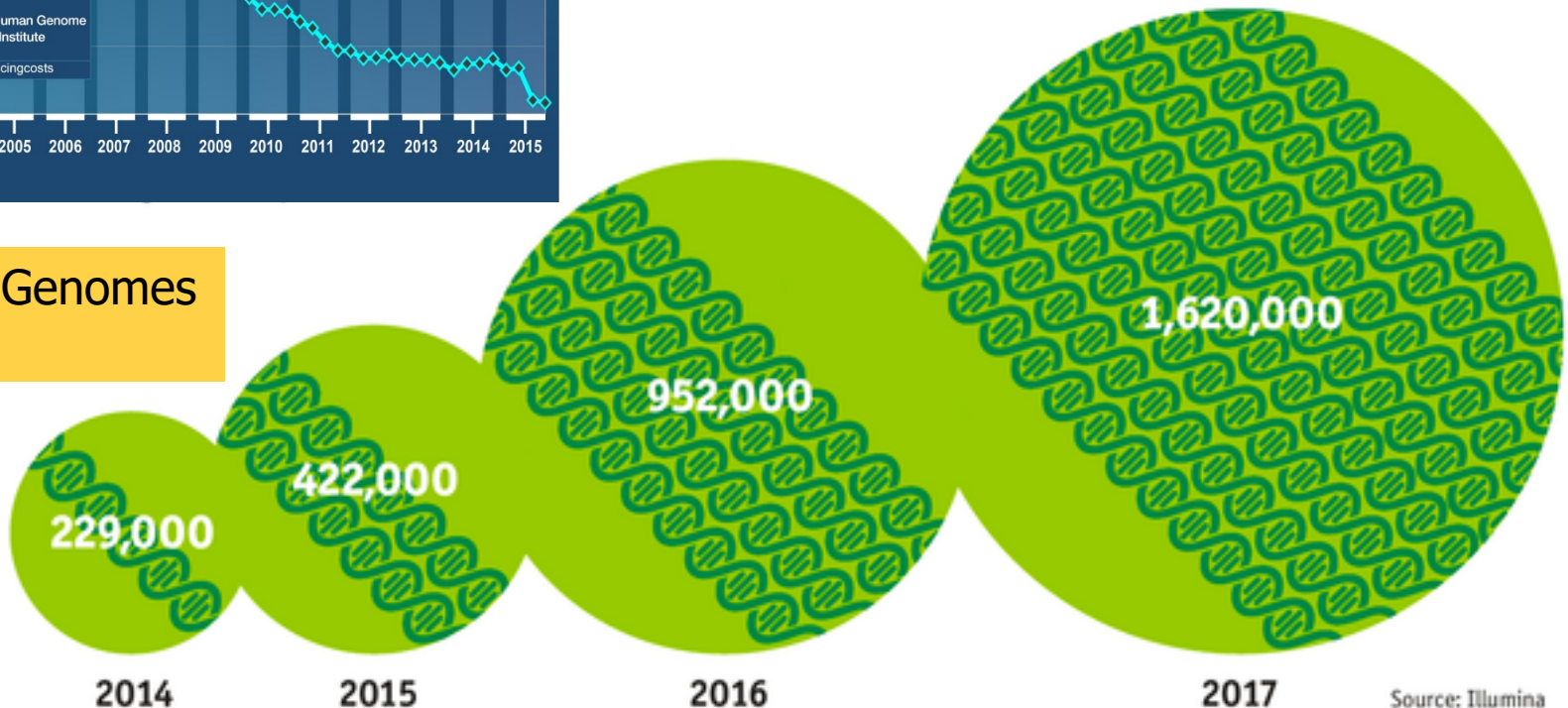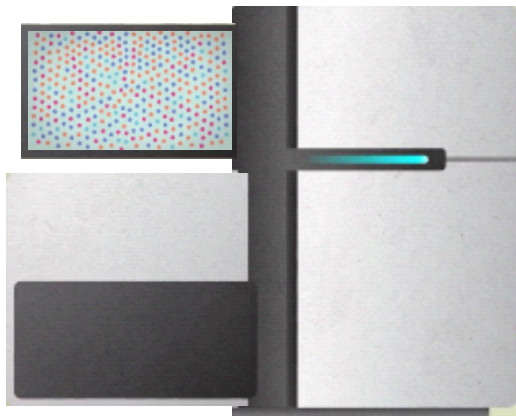
229,000 — 2014
422,000 — 2015
952,000 — 2016
1,620,000 — 2017

Source: Illumina
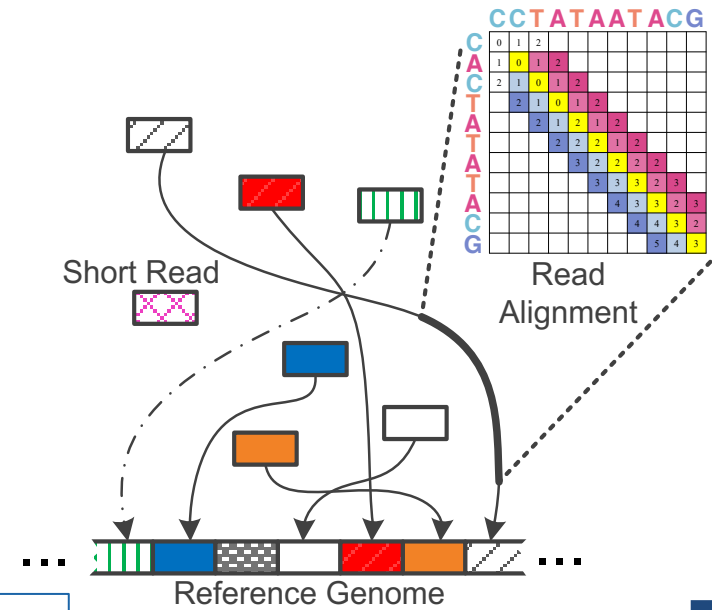
1 Sequencing

2 Read Mapping

Genome Analysis

Data → performance & energy bottleneck

3 Variant Calling

4 Scientific Discovery

GateKeeper: FPGA-Based Alignment Filtering

- Mohammed Alser, Hasan Hassan, Hongyi Xin, Oguz Ergin, Onur Mutlu, and Can Alkan
  "GateKeeper: A New Hardware Architecture for Accelerating Pre-Alignment in DNA Short Read Mapping"
  Bioinformatics, [published online, May 31], 2017.
  [Source Code]
  [Online link at Bioinformatics Journal]

GateKeeper: a new hardware architecture for accelerating pre-alignment in DNA short read mapping

Mohammed Alser ✉, Hasan Hassan, Hongyi Xin, Oğuz Ergin, Onur Mutlu ✉, Can Alkan ✉

In-Memory DNA Sequence Analysis

- Jeremie S. Kim, Damla Senol Cali, Hongyi Xin, Donghyuk Lee, Saugata Ghose, Mohammed Alser, Hasan Hassan, Oguz Ergin, Can Alkan, and Onur Mutlu,
"GRIM-Filter: Fast Seed Location Filtering in DNA Read Mapping Using Processing-in-Memory Technologies"
BMC Genomics, 2018.
Proceedings of the 16th Asia Pacific Bioinformatics Conference (**APBC**), Yokohama, Japan, January 2018.
arxiv.org Version (pdf)

GRIM-Filter: Fast seed location filtering in DNA read mapping using processing-in-memory technologies

Jeremie S. Kim[1,6*], Damla Senol Cali[1], Hongyi Xin[2], Donghyuk Lee[3], Saugata Ghose[1], Mohammed Alser[4], Hasan Hassan[6], Oguz Ergin[5], Can Alkan[4*] and Onur Mutlu[6,1*]

Shouji (障子) [Alser+, Bioinformatics 2019]

Mohammed Alser, Hasan Hassan, Akash Kumar, Onur Mutlu, and Can Alkan,
"Shouji: A Fast and Efficient Pre-Alignment Filter for Sequence Alignment"
Bioinformatics, [published online, March 28], 2019.
[Source Code]
[Online link at Bioinformatics Journal]

Sequence alignment

Shouji: a fast and efficient pre-alignment filter for sequence alignment

Mohammed Alser[1,2,3,*], Hasan Hassan[1], Akash Kumar[2], Onur Mutlu[1,3,*] and Can Alkan[3,*]

[1]Computer Science Department, ETH Zürich, Zürich 8092, Switzerland, [2]Chair for Processor Design, Center For Advancing Electronics Dresden, Institute of Computer Engineering, Technische Universität Dresden, 01062 Dresden, Germany and [3]Computer Engineering Department, Bilkent University, 06800 Ankara, Turkey

*To whom correspondence should be addressed.

SneakySnake [Alser+, Bioinformatics 2020]

Mohammed Alser, Taha Shahroodi, Juan-Gomez Luna, Can Alkan, and Onur Mutlu,
"SneakySnake: A Fast and Accurate Universal Genome Pre-Alignment Filter for CPUs, GPUs, and FPGAs"
Bioinformatics, to appear in 2020.
[Source Code]
[Online link at Bioinformatics Journal]

Subject Section

SneakySnake: A Fast and Accurate Universal Genome Pre-Alignment Filter for CPUs, GPUs, and FPGAs

Mohammed Alser [1,2,*], Taha Shahroodi [1], Juan Gómez-Luna [1,2],
Can Alkan [4,*], and Onur Mutlu [1,2,3,4,*]

[1] Department of Computer Science, ETH Zurich, Zurich 8006, Switzerland
[2] Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich 8006, Switzerland
[3] Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh 15213, PA, USA
[4] Department of Computer Engineering, Bilkent University, Ankara 06800, Turkey

GenASM Framework [MICRO 2020]

- Damla Senol Cali, Gurpreet S. Kalsi, Zulal Bingol, Can Firtina, Lavanya Subramanian, Jeremie S. Kim, Rachata Ausavarungnirun, Mohammed Alser, Juan Gomez-Luna, Amirali Boroumand, Anant Nori, Allison Scibisz, Sreenivas Subramoney, Can Alkan, Saugata Ghose, and Onur Mutlu,
"GenASM: A High-Performance, Low-Power Approximate String Matching Acceleration Framework for Genome Sequence Analysis"
Proceedings of the 53rd International Symposium on Microarchitecture (**MICRO**), Virtual, October 2020.
[Lighting Talk Video (1.5 minutes)]
[Lightning Talk Slides (pptx) (pdf)]
[Talk Video (18 minutes)]
[Slides (pptx) (pdf)]

GenASM: A High-Performance, Low-Power Approximate String Matching Acceleration Framework for Genome Sequence Analysis

Damla Senol Cali[†⋈]   Gurpreet S. Kalsi[⋈]   Zülal Bingöl[▽]   Can Firtina[◇]   Lavanya Subramanian[‡]   Jeremie S. Kim[◇†]
Rachata Ausavarungnirun[⊙]   Mohammed Alser[◇]   Juan Gomez-Luna[◇]   Amirali Boroumand[†]   Anant Nori[⋈]
Allison Scibisz[†]   Sreenivas Subramoney[⋈]   Can Alkan[▽]   Saugata Ghose[⋆†]   Onur Mutlu[◇†▽]

[†]*Carnegie Mellon University*   [⋈]*Processor Architecture Research Lab, Intel Labs*   [▽]*Bilkent University*   [◇]*ETH Zürich*
[‡]*Facebook*   [⊙]*King Mongkut's University of Technology North Bangkok*   [⋆]*University of Illinois at Urbana–Champaign*
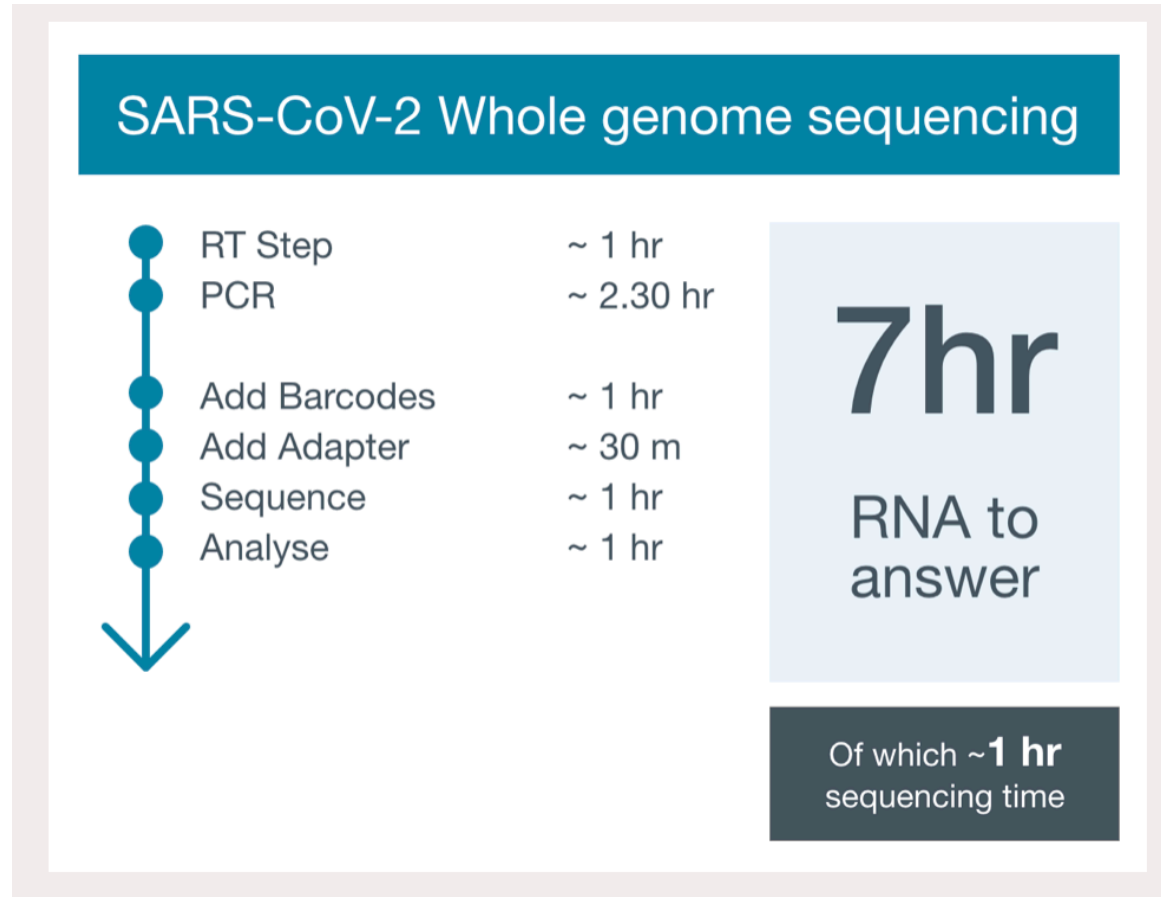
Future of Genome Sequencing & Analysis

MinION from ONT
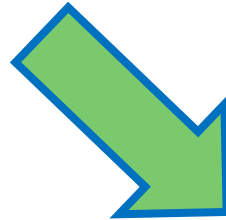
SmidgION from ONT

COVID-19 Nanopore Sequencing (I)

From ONT (https://nanoporetech.com/covid-19/overview)

COVID-19 Nanopore Sequencing (II)

How are scientists using nanopore sequencing to research COVID-19?

From ONT (https://nanoporetech.com/covid-19/overview)

Accelerating Genome Analysis: Overview

- Mohammed Alser, Zulal Bingol, Damla Senol Cali, Jeremie Kim, Saugata Ghose, Can Alkan, and Onur Mutlu,
  "Accelerating Genome Analysis: A Primer on an Ongoing Journey"
  IEEE Micro (**IEEE MICRO**), Vol. 40, No. 5, pages 65-75, September/October 2020.
  [Slides (pptx)(pdf)]
  [Talk Video (1 hour 2 minutes)]

Accelerating Genome Analysis: A Primer on an Ongoing Journey

Mohammed Alser
ETH Zürich

Zülal Bingöl
Bilkent University

Damla Senol Cali
Carnegie Mellon University

Jeremie Kim
ETH Zurich and Carnegie Mellon University

Saugata Ghose
University of Illinois at Urbana–Champaign and Carnegie Mellon University

Can Alkan
Bilkent University

Onur Mutlu
ETH Zurich, Carnegie Mellon University, and Bilkent University

More on Fast Genome Analysis …

- Onur Mutlu,
  "Accelerating Genome Analysis: A Primer on an Ongoing Journey"
  Invited Lecture at Technion, Virtual, 26 January 2021.
  [Slides (pptx) (pdf)]
  [Talk Video (1 hour 37 minutes, including Q&A)]
  [Related Invited Paper (at IEEE Micro, 2020)]

Detailed Lectures on Genome Analysis

- Computer Architecture, Fall 2020, Lecture 3a
  - **Introduction to Genome Sequence Analysis** (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=CrRb32v7SJc&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN&index=5

- Computer Architecture, Fall 2020, Lecture 8
  - **Intelligent Genome Analysis** (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=ygmQpdDTL7o&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN&index=14

- Computer Architecture, Fall 2020, Lecture 9a
  - **GenASM: Approx. String Matching Accelerator** (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=XoLpzmN-Pas&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN&index=15

- Accelerating Genomics Project Course, Fall 2020, Lecture 1
  - **Accelerating Genomics** (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=rgjl8ZyLsAg&list=PL5Q2soXY2Zi9E2bBVAgCqLgwiDRQDTyId

Data Overwhelms Modern Machines

In-memory Databases
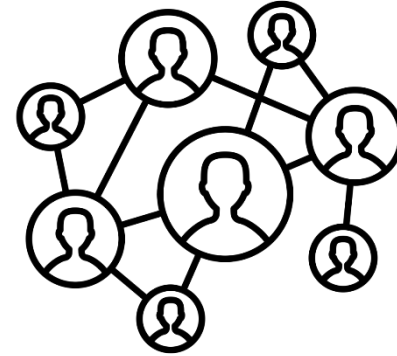
Graph/Tree Processing

Data → performance & energy bottleneck

In-Memory Data Analytics
[Clapp+ (**Intel**), IISWC'15; Awan+, BDCloud'15]

Datacenter Workloads
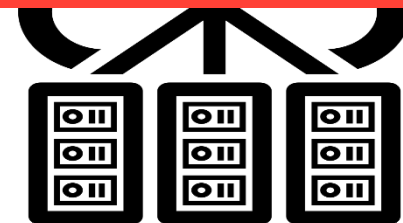[Kanev+ (**Google**), ISCA'15]

SAFARI

Data Overwhelms Modern Machines

Chrome

TensorFlow Mobile

Data → performance & energy bottleneck

Video Playback

Google's **video codec**

Video Capture

Google's **video codec**

Data Movement Overwhelms Modern Machines

- Amirali Boroumand, Saugata Ghose, Youngsok Kim, Rachata Ausavarungnirun, Eric Shiu, Rahul Thakur, Daehyun Kim, Aki Kuusela, Allan Knies, Parthasarathy Ranganathan, and Onur Mutlu, **"Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks"** *Proceedings of the 23rd International Conference on Architectural Support for Programming Languages and Operating Systems* (**ASPLOS**), Williamsburg, VA, USA, March 2018.

62.7% of the total system energy is spent on **data movement**

Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks

Amirali Boroumand[1]   Saugata Ghose[1]   Youngsok Kim[2]

Rachata Ausavarungnirun[1]   Eric Shiu[3]   Rahul Thakur[3]   Daehyun Kim[4,3]

Aki Kuusela[3]   Allan Knies[3]   Parthasarathy Ranganathan[3]   Onur Mutlu[5,1]

Data Movement vs. Computation Energy

Communication Dominates Arithmetic

Dally, HiPEAC 2015

20mm

64-bit DP 20pJ

26 pJ   256 pJ   16 nJ   DRAM Rd/Wr

256-bit buses

500 pJ   Efficient off-chip link

256-bit access 8 kB SRAM

50 pJ

A memory access consumes ~100-1000X the energy of a complex addition

SAFARI
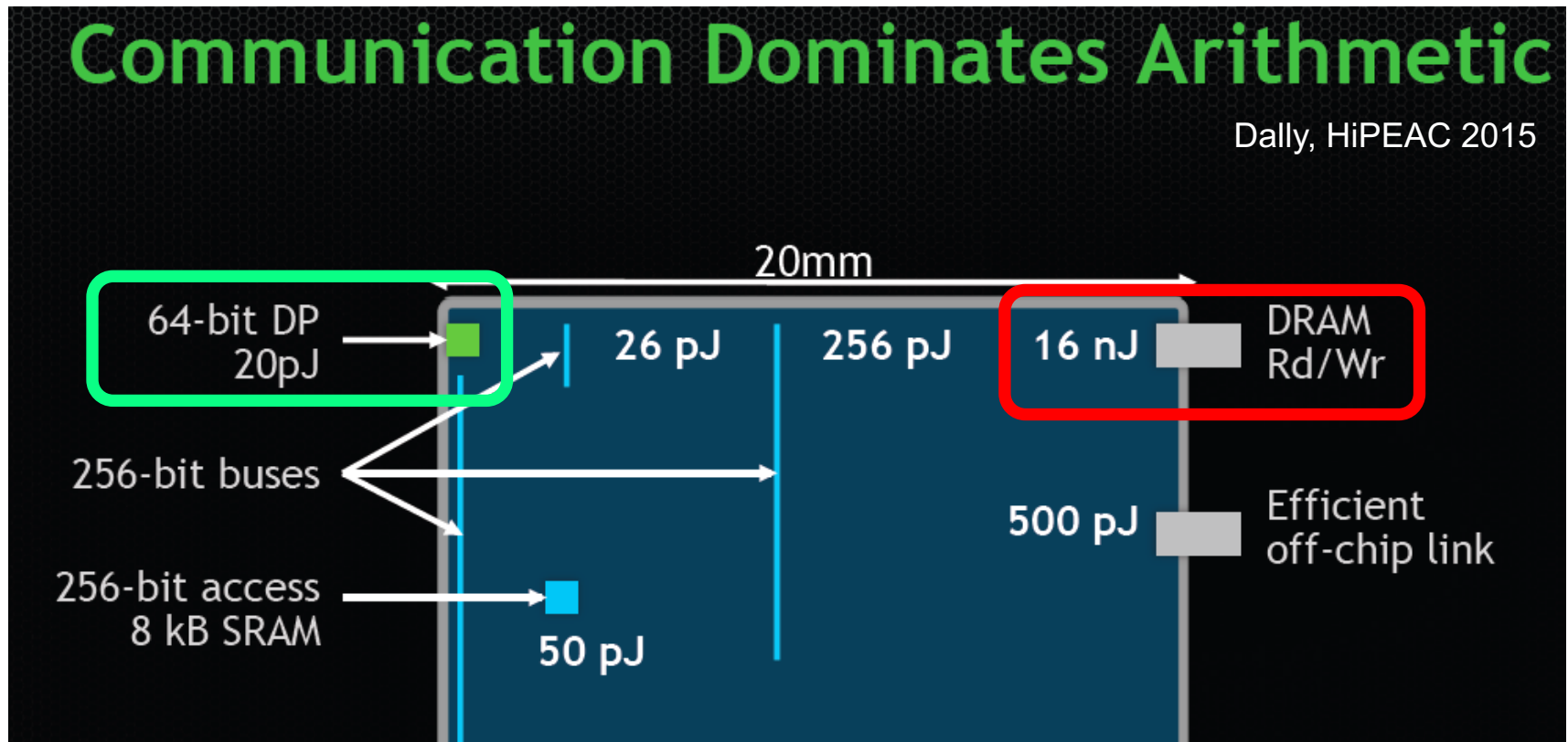
Many Interesting Things Are Happening Today in Computer Architecture

Many Novel Concepts Investigated Today

- **New Computing Paradigms (Rethinking the Full Stack)**
  - Processing in Memory, Processing Near Data
  - Neuromorphic Computing
  - Fundamentally Secure and Dependable Computers

- **New Accelerators (Algorithm-Hardware Co-Designs)**
  - Artificial Intelligence & Machine Learning
  - Graph Analytics
  - Genome Analysis

- **New Memories and Storage Systems**
  - Non-Volatile Main Memory
  - Processing in Memory, Intelligent Memory

Increasingly Demanding Applications

Dream

and, they will come

As applications push boundaries, computing platforms will become increasingly strained.

SAFARI

Increasingly Diverging/Complex Tradeoffs

Communication Dominates Arithmetic
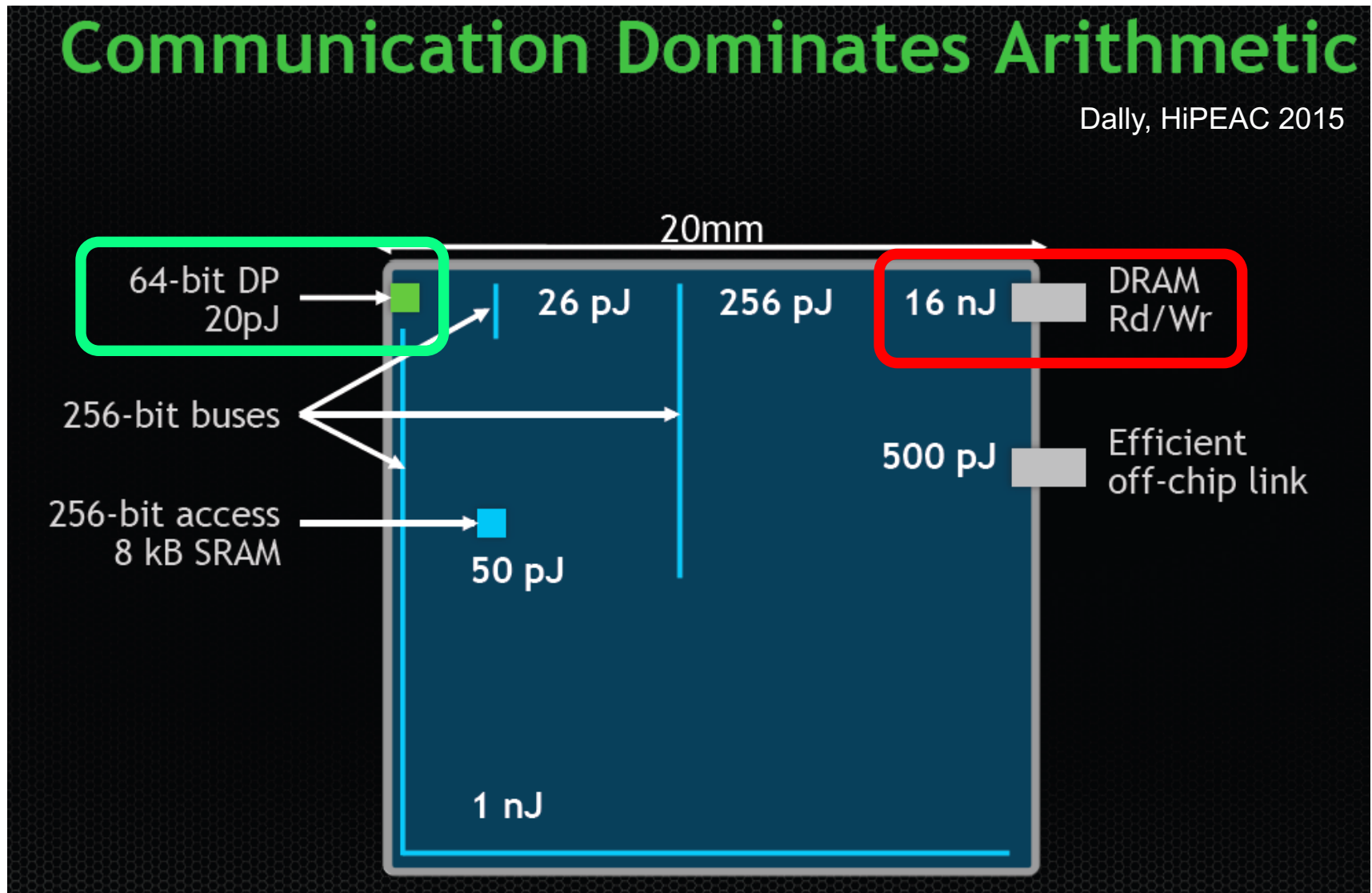
Dally, HiPEAC 2015

20mm

64-bit DP
20pJ

256-bit buses

256-bit access
8 kB SRAM

26 pJ

256 pJ

16 nJ — DRAM Rd/Wr

500 pJ — Efficient off-chip link

50 pJ

1 nJ

Data Movement vs. Computation Energy

Communication Dominates Arithmetic

Dally, HiPEAC 2015

20mm

64-bit DP 20pJ

26 pJ 256 pJ 16 nJ DRAM Rd/Wr

256-bit buses

500 pJ Efficient off-chip link

256-bit access 8 kB SRAM

50 pJ

A memory access consumes ~100-1000X the energy of a complex addition

SAFARI
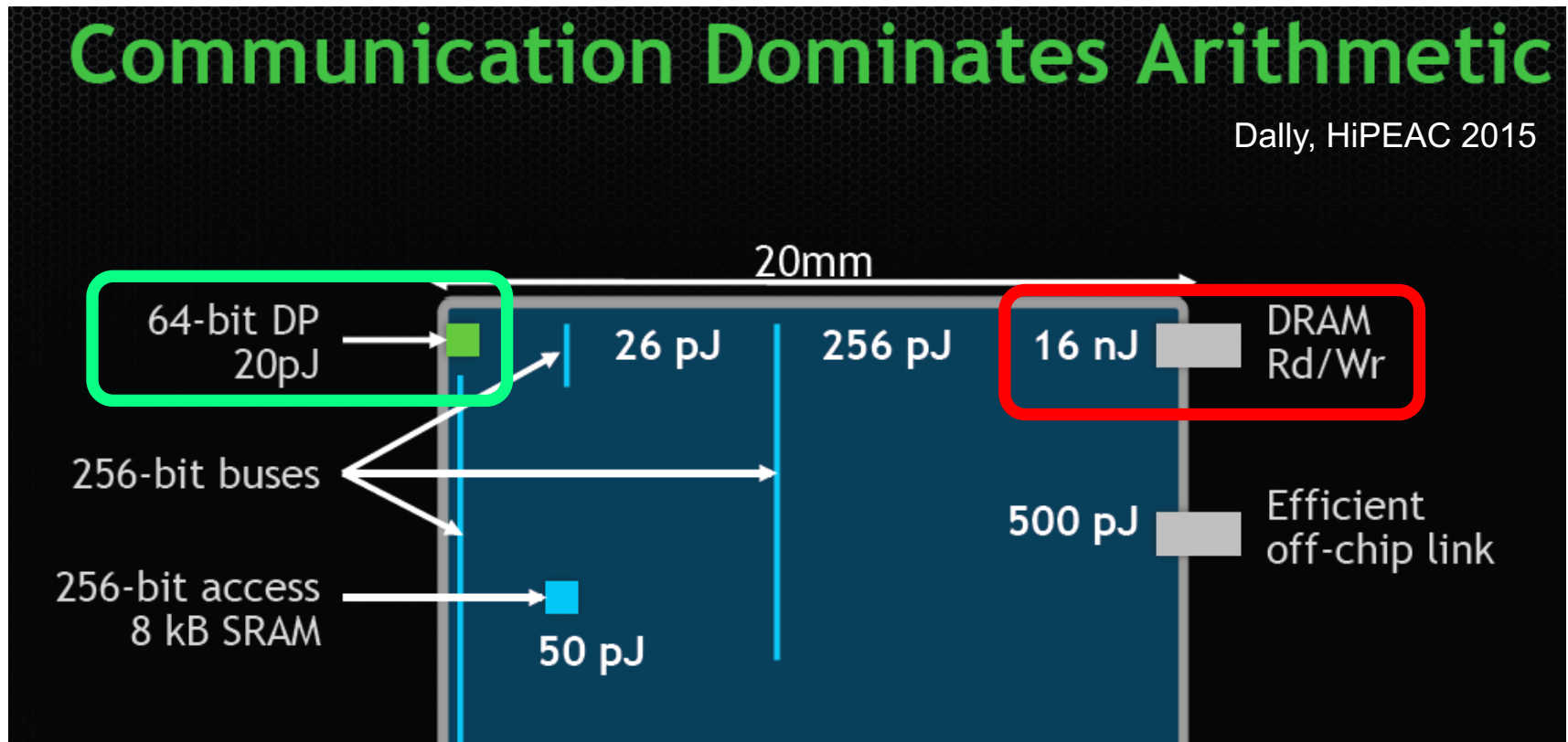
Increasingly Complex Systems

Past systems
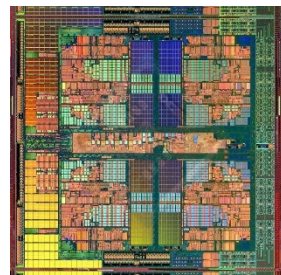
Microprocessor   Main Memory   Storage (SSD/HDD)

Increasingly Complex Systems

FPGAs

Modern systems

Hybrid Main Memory

Persistent Memory/Storage

Heterogeneous
Processors and
Accelerators

Memory Controller

Core   Core   Core   Core   Core   Core

Shared L3 Cache   Shared L3 Cache

(General Purpose) GPUs
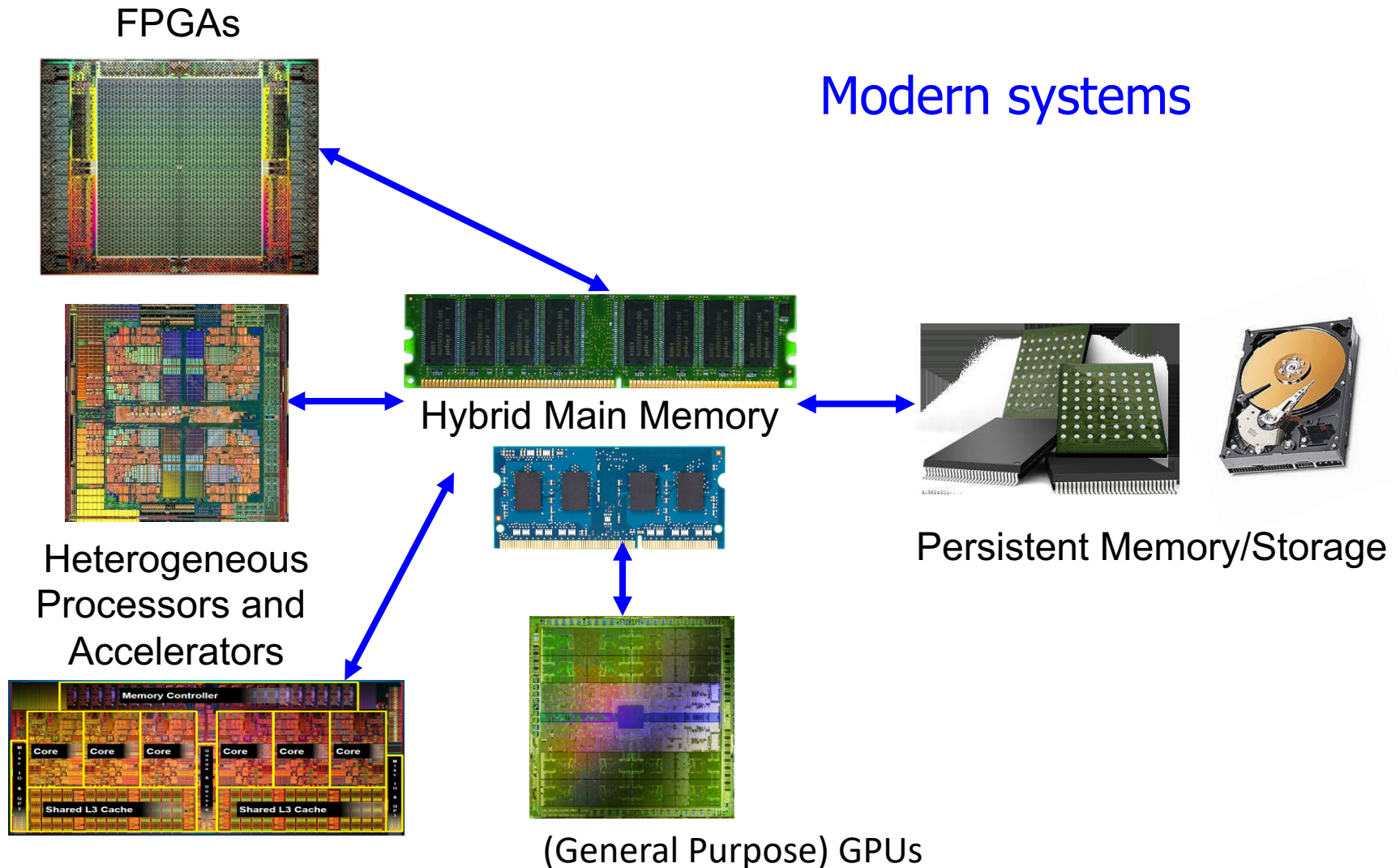
Computer Architecture Today

- Computing landscape is very different from 10-20 years ago

- Applications and technology both demand novel architectures

Heterogeneous Processors and Accelerators

Hybrid Main Memory

Persistent Memory/Storage

General Purpose GPUs

Every component and its interfaces, as well as entire system designs are being re-examined
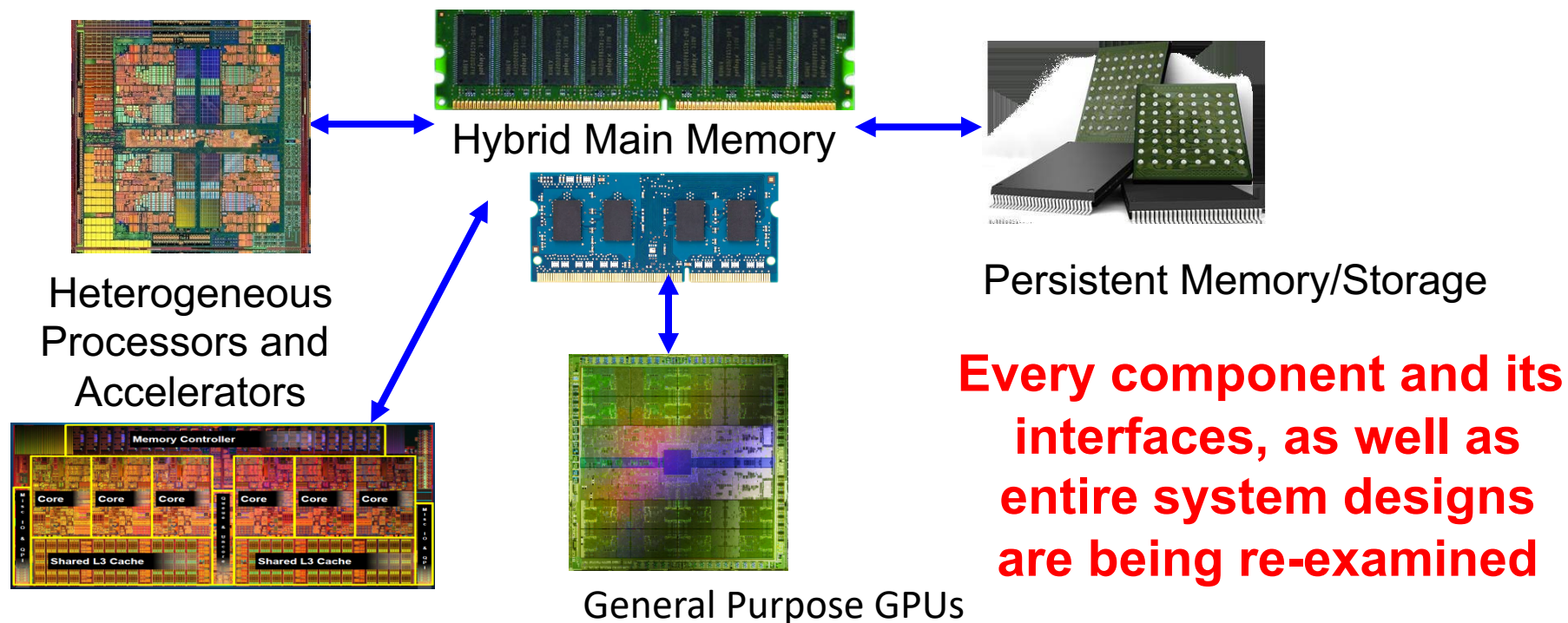
Computer Architecture Today (II)

- You can revolutionize the way computers are built, if you understand both the hardware and the software (and change each accordingly)

- You can invent new paradigms for computation, communication, and storage

- Recommended book: Thomas Kuhn, "The Structure of Scientific Revolutions" (1962)
  - Pre-paradigm science: no clear consensus in the field
  - Normal science: dominant theory used to explain/improve things (business as usual); exceptions considered anomalies
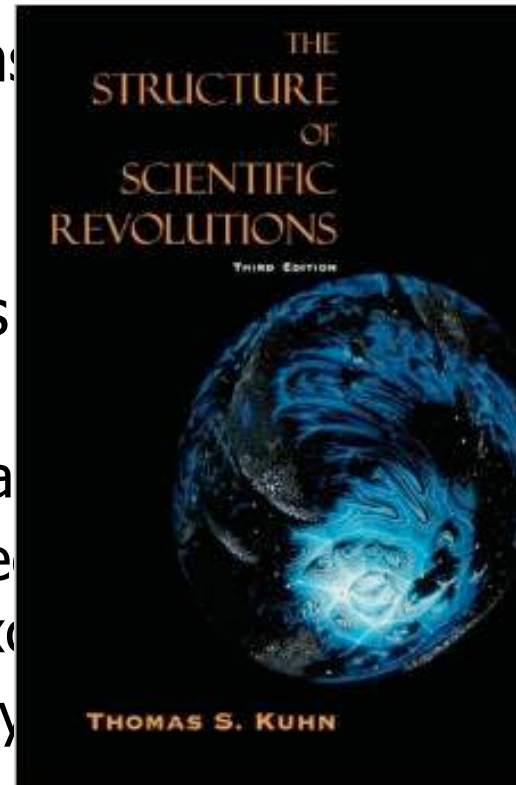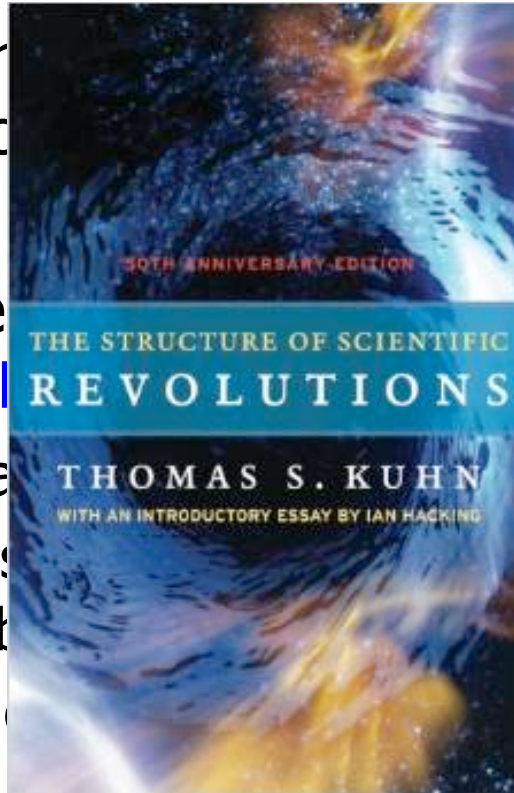 - Revolutionary science: underlying assumptions re-examined

Computer Architecture Today (II)

- You can revolutionize the way computers are built, if you understand both the hardware and the software (and change each accordingly)

- You can in_____ms communic_____

- Recomme_____as _____ure of Scientific _____)
  - Pre-para_____ea_____ield
  - Normal _____ne_____improve things (b_____exc_____anomalies
  - Revoluti_____rly_____examined

Takeaways

- **It is an exciting time to be understanding and designing computing architectures**

- Many challenging and exciting problems in platform design
  - That no one has tackled (or thought about) before
  - That can have huge impact on the world's future

- Driven by huge hunger for data (Big Data), new applications (ML/AI, graph analytics, genomics), ever-greater realism, …
  - We can easily collect more data than we can analyze/understand

- Driven by significant difficulties in keeping up with that hunger at the technology layer
  - Five walls: Energy, reliability, complexity, security, scalability

Let's Start with Some Puzzles

a.k.a. Computer Architecture resembles Building Architecture

What Is This?

What About This?

What Do the Following Have in Common?

Gare do Oriente, Lisbon

Milwaukee Art Museum

154

Athens Olympic Stadium

City of Arts and Sciences, Valencia

Florida Polytechnic University (I)

Source: http://www.architectmagazine.com/design/buildings/florida-polytechnic-university-designed-by-santiago-calatrava_o

Oculus, New York City

What do All Those Have in Common with Bahnhof Stadelhofen?

Answer: All Designed by a Famous Architect

- ETH Alumnus, PhD Civil Engineering

- "The train station has several of the features that became signatures of his work; straight lines and right angles are rare."

Santiago Calatrava Valls (born 28 July 1951) is a Spanish architect, structural engineer, sculptor and painter, particularly known for his bridges supported by single leaning pylons, and his railway stations, stadiums, and museums, whose sculptural forms often resemble living organisms.[1] His best-known works include the Milwaukee Art Museum, the Turning Torso tower in Malmo, Sweden, the Margaret Hunt Hill Bridge in Dallas, Texas, and the Museum of Tomorrow in Rio de Janeiro,

Your First Comp. Architecture Assignment

- Go and find the closest Calatrava building to this classroom
    - For the ones who like a challenge, find the furthest building that was designed by Calatrava to his classroom ☺

- Appreciate the beauty & out-of-the-box and creative thinking
- Think about tradeoffs in the design
    - Strengths, weaknesses, goals of design
- Derive principles on your own for good design and innovation

- Due date: **Any time during this course**
    - Later during the course is better
    - Apply what you have learned in this course
    - Think out-of-the-box

But First, Today's First Assignment

Find The Differences of This and That

This

164

That

Many Tradeoffs Between Two Designs

- You can list them after you complete the first assignment…

Aside: Evaluation Criteria for the Designs

- Functionality (Does it meet the specification?)
- Reliability
- Space requirement
- Cost
- Expandability
- Comfort level of users
- Happiness level of users
- Aesthetics
- Security

- …

- How to evaluate goodness of design is always a critical question → "Performance" evaluation and metrics

A Key Question

- How was Calavatra able to design especially his key buildings?
- Can have many guesses
  - (Ultra) hard work, perseverance, dedication (over decades)
  - Experience
  - Creativity, Out-of-the-box thinking
  - A good understanding of past designs
  - Good judgment and intuition
  - Strong skill combination (math, architecture, art, engineering, …)
  - Funding ($$$$), luck, initiative, entrepreneurialism
  - Strong understanding of and commitment to fundamentals
  - Principled design
  - …

- (You will be exposed to and hopefully develop/enhance many of these skills in this course)

Principled Design

- "To me, there are two overriding principles to be found in nature which are most appropriate for building:
    - one is the optimal use of material,
    - the other the capacity of organisms to change shape, to grow, and to move."
    - *Santiago Calatrava*

- "Calatrava's constructions are inspired by natural forms like plants, bird wings, and the human body."

Gare do Oriente, Lisbon, Revisited

A Principled Design

Zoomorphic architecture

From Wikipedia, the free encyclopedia

Zoomorphic architecture is the practice of using animal forms as the inspirational basis and blueprint for architectural design. "While animal forms have always played a role adding some of the deepest layers of meaning in architecture, it is now becoming evident that a new strand of biomorphism is emerging where the meaning derives not from any specific representation but from a more general allusion to biological processes."[1]

Some well-known examples of Zoomorphic architecture can be found in the TWA Flight Center building in New York City, by Eero Saarinen, or the Milwaukee Art Museum by Santiago Calatrava, both inspired by the form of a bird's wings.[3]

What Does This Remind You Of?

Source: https://www.dezeen.com/2016/08/29/santiago-calatrava-oculus-world-trade-center-transportation-hub-new-york-photographs-hufton-crow/

The Architect's Answer

Design   [edit]

Calatrava said that the Oculus resembles a bird being released from a child's hand. The roof was originally designed to mechanically open to increase light and ventilation to the enclosed space. Herbert Muschamp, architecture critic of *The New York Times*, compared the design to the Bethesda Terrace and Fountain in Central Park, and wrote in 2004:

Source: https://en.wikipedia.org/wiki/World_Trade_Center_station_(PATH)

Strengths and Praise

"Santiago Calatrava's design for the World Trade Center PATH station should satisfy those who believe that buildings planned for ground zero must aspire to a spiritual dimension. Over the years, many people have discerned a metaphysical element in Mr. Calatrava's work. I hope New Yorkers will detect its presence, too. With deep appreciation, I congratulate the Port Authority for commissioning Mr. Calatrava, the great Spanish architect and engineer, to design a building with the power to shape the future of New York. It is a pleasure to report, for once, that public officials are not overstating the case when they describe a design as breathtaking.[43]"

Design Constraints and Criticism

However, Calatrava's original soaring spike design was scaled back because of security issues. The *New York Times* observed in 2005:

> In the name of security, Santiago Calatrava's bird has grown a beak. Its ribs have doubled in number and its wings have lost their interstices of glass.... [T]he main transit hall, between Church and Greenwich Streets, will almost certainly lose some of its delicate quality, while gaining structural expressiveness. It may now evoke a slender stegosaurus more than it does a bird.[45]

Stegosaurus

From Wikipedia, the free encyclopedia

For the pachycephalosaurid of a similar name, see Stegoceras.

Stegosaurus (/ˌstɛgəˈsɔːrəs/[1]) is a genus of armored dinosaur. Fossils of this genus date to the Late Jurassic period, where they are found in Kimmeridgian to early Tithonian aged strata, between 155 and 150 million years ago, in the western United States and Portugal. Several
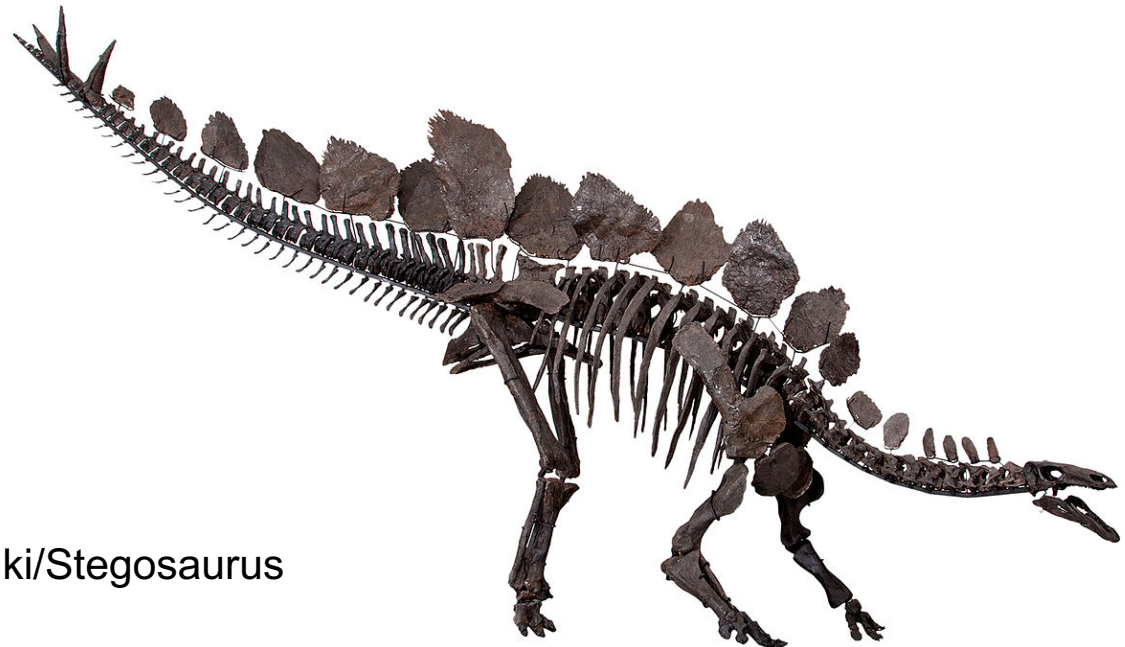
Source: https://en.wikipedia.org/wiki/Stegosaurus

176

Design Constraints: Noone is Immune

However, Calatrava's original soaring spike design was scaled back because of security issues. The *New York Times* observed in 2005:

> In the name of security, Santiago Calatrava's bird has grown a beak. Its ribs have doubled in number and its wings have lost their interstices of glass.... [T]he main transit hall, between Church and Greenwich Streets, will almost certainly lose some of its delicate quality, while gaining structural expressiveness. It may now evoke a slender stegosaurus more than it does a bird.[45]

The design was further modified in 2008 to eliminate the opening and closing roof mechanism because of budget and space constraints.[46]

The Transportation Hub has been dubbed "the world's most expensive transportation hub" for its massive cost for reconstruction—$3.74 billion dollars.[48][58] By contrast, the proposed two-mile PATH extension

Digital Design & Computer Arch.

Lecture 1: Introduction and Basics

Prof. Onur Mutlu

ETH Zürich

Spring 2021

25 February 2021

We Will Likely Not Cover the Following Slides in Lecture 1

The Lecture Was Slightly Different When I Was at CMU

What Is This?

Answer: Masterpiece of A Famous Architect

Fallingwater

From Wikipedia, the free encyclopedia

Fallingwater or **Kaufmann Residence** is a house designed by architect Frank Lloyd Wright in 1935 in rural southwestern Pennsylvania, 43 miles (69 km) southeast of Pittsburgh.[4] The home was built partly over a waterfall on Bear Run in the Mill Run section of Stewart Township, Fayette County, Pennsylvania, in the Laurel Highlands of the Allegheny Mountains.

Time cited it after its completion as Wright's "most beautiful job";[5] it is listed among *Smithsonian*'s Life List of 28 places "to visit before you die."[6] It was designated a National Historic Landmark in 1966.[3] In 1991, members of the American Institute of Architects named the house the "best all-time work of American architecture" and in 2007, it was ranked twenty-ninth on the list of America's Favorite Architecture according to the AIA.
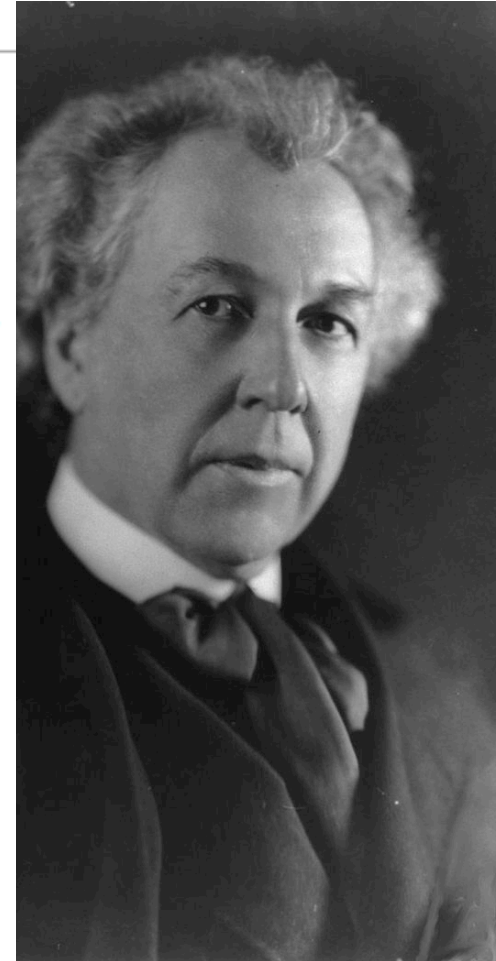
Find The Differences of This and That

This

That

A Key Question

- **How was Wright able to design his masterpiece?**
- **Can have many guesses**
  - ❑ (Ultra) hard work, perseverance, dedication (over decades)
  - ❑ Experience
  - ❑ Creativity, Out-of-the-box thinking
  - ❑ A good understanding of past designs
  - ❑ Good judgment and intuition
  - ❑ Strong skill combination (math, architecture, art, engineering, …)
  - ❑ Funding ($$$$), luck, initiative, entrepreneurialism
  - ❑ Strong understanding of and commitment to fundamentals
  - ❑ Principled design
  - ❑ …

- **(You will be exposed to and hopefully develop/enhance many of these skills in this course)**

A Quote from The Architect Himself

- "architecture […] based upon principle, and not upon precedent"

A Principled Design

Organic architecture

From Wikipedia, the free encyclopedia

Organic architecture is a philosophy of architecture which promotes harmony between human habitation and the natural world through design approaches so sympathetic and well integrated with its site, that buildings, furnishings, and surroundings become part of a unified, interrelated composition.

A well-known example of organic architecture is Fallingwater, the residence Frank Lloyd Wright designed for the Kaufmann family in rural Pennsylvania. Wright had many choices to locate a home on this large site, but chose to place the home directly over the waterfall and creek creating a close, yet noisy dialog with the rushing water and the steep site. The horizontal striations of stone masonry with daring cantilevers of colored beige concrete blend with native rock outcroppings and the wooded environment.

A Key Question

- How was Wright able to design his masterpiece?
- Can have many guesses
  - (Ultra) hard work, perseverance, dedication (over decades)
  - Experience
  - Creativity, Out-of-the-box thinking
  - A good understanding of past designs
  - Good judgment and intuition
  - Strong skill combination (math, architecture, art, engineering, …)
  - Funding ($$$$), luck, initiative, entrepreneurialism
  - Strong understanding of and commitment to fundamentals
  - Principled design
  - …

- (You will be exposed to and hopefully develop/enhance many of these skills in this course)

Takeaways

- It all starts from the basic building blocks and design principles

- And, knowledge of how to use & apply them

- Underlying technology might change (e.g., steel vs. wood)
    - but methods of taking advantage of technology bear resemblance
    - methods used for design depend on the principles employed
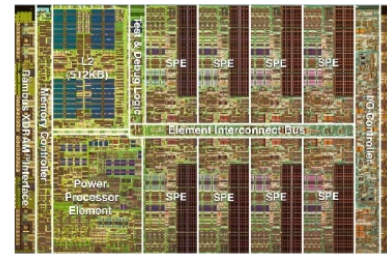
The Same Applies to Processor Chips

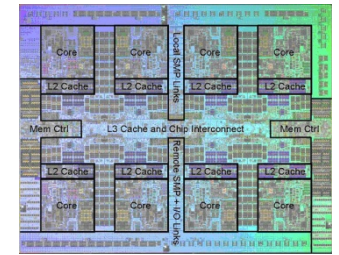- There are basic building blocks and and design principles
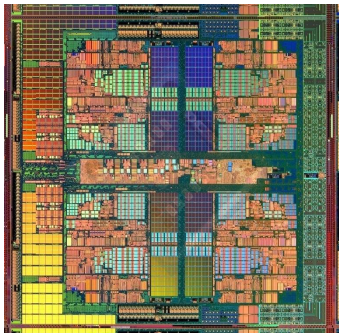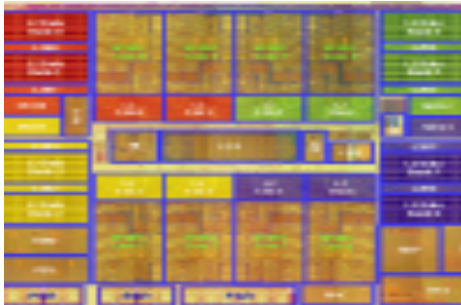

AMD Barcelona
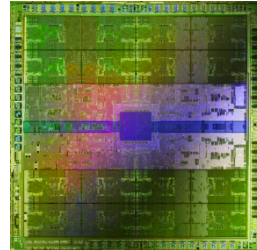4 cores
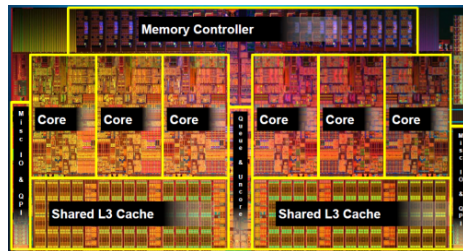
Intel Core i7
8 cores

IBM Cell BE
8+1 cores

IBM POWER7
8 cores

Sun Niagara II
8 cores

Nvidia Fermi
448 "cores"
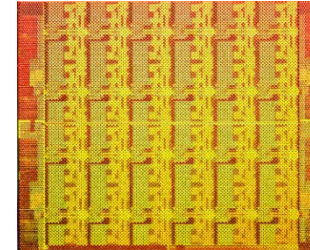
Intel SCC
48 cores, networked
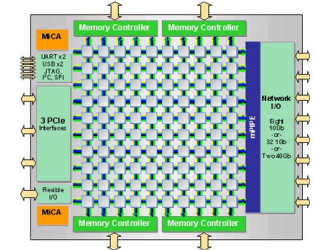
Tilera TILE Gx
100 cores, networked

The Same Applies to Computing Systems

- There are basic building blocks and and design principles

source: http://www.sia-online.org (semiconductor industry association)

The Same Applies to Computing Systems

- There are basic building blocks and and design principles

Different Platforms, Different Goals

Source: https://iq.intel.com/5-awesome-uses-for-drone-technology/

Different Platforms, Different Goals

Different Platforms, Different Goals

Different Platforms, Different Goals

Jack Dongarra
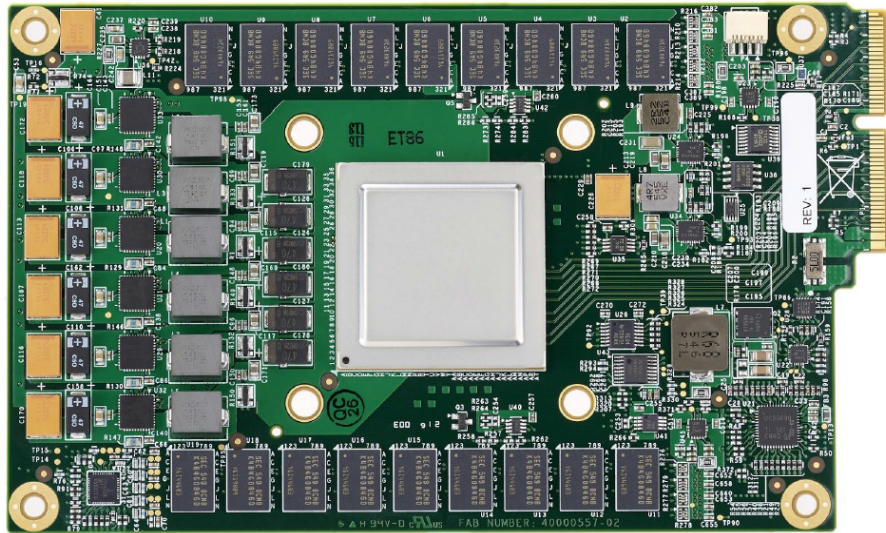
SAFARI

Source: https://fossbytes.com/wp-content/uploads/2015/06/Supercomputer-TIANHE2-china.jpg

Different Platforms, Different Goals

Figure 3. TPU Printed Circuit Board. It can be inserted in the slot for an SATA disk in a server, but the card uses PCIe Gen3 x16.

Figure 4. Systolic data flow of the Matrix Multiply Unit. Software has the illusion that each 256B input is read at once, and they instantly update one location of each of 256 accumulator RAMs.

Jouppi et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit", ISCA 2017.

Different Platforms, Different Goals

- ML accelerator: 260 mm^2, 6 billion transistors, 600 GFLOPS GPU, 12 ARM 2.2 GHz CPUs.
- Two redundant chips for better safety.

Different Platforms, Different Goals
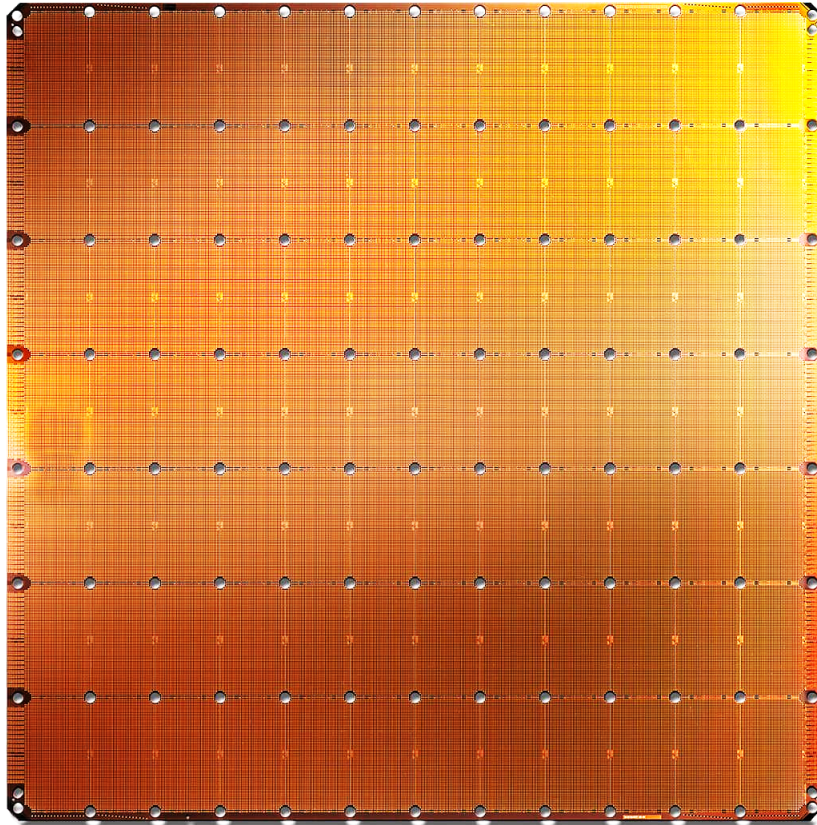
- **The largest ML accelerator chip**

- **400,000 cores**

Cerebras WSE
1.2 Trillion transistors
46,225 mm²

Largest GPU
21.1 Billion transistors
815 mm²

NVIDIA TITAN V

https://www.anandtech.com/show/14758/hot-chips-31-live-blogs-cerebras-wafer-scale-deep-learning

https://www.cerebras.net/cerebras-wafer-scale-engine-why-we-need-big-chips-for-deep-learning/

Basic Building Blocks

- Electrons
- Transistors
- Logic Gates
- Combinational Logic Circuits
- Sequential Logic Circuits
  - Storage Elements and Memory
- ...
- Cores
- Caches
- Interconnect
- Memories
- ...

Reading Assignments for This Week

- Chapter 1 in Harris & Harris
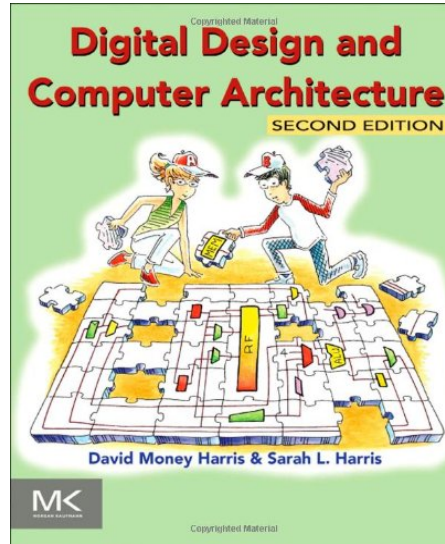
- Chapters 1-2 in Patt and Patel
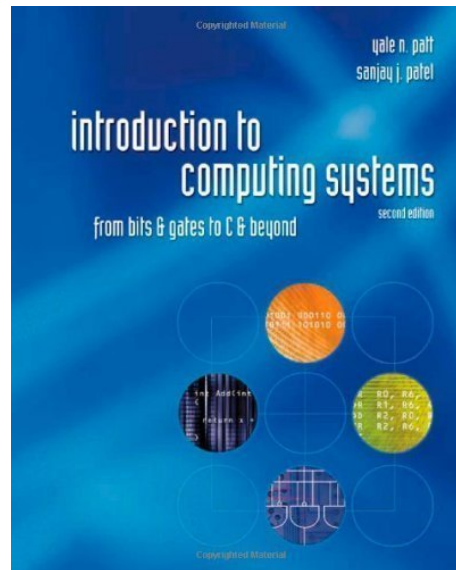
- Supplementary Lecture Slides on Binary Numbers

Major High-Level Goals of This Course

- In Digital Circuits & Computer Architecture

- Understand the basics
- Understand the principles (of design)
- Understand the precedents

- Based on such understanding:
  - learn how a modern computer works underneath
  - evaluate tradeoffs of different designs and ideas
  - implement a principled design (a simple microprocessor)
  - learn to systematically debug increasingly complex systems
  - Hopefully enable you to develop novel, out-of-the-box designs

- The focus is on basics, principles, precedents, and how to use them to create/implement good designs

Why These Goals?

- Because you are here for a Computer Science degree

- Regardless of your future direction, learning the principles of digital design & computer architecture will be useful to
  - design better hardware
  - design better software
  - design better systems
  - make better tradeoffs in design
  - understand why computers behave the way they do
  - solve problems better
  - think "in parallel"
  - think critically
  - …

Course Info and Logistics

Course Info: Instructor

- **Onur Mutlu**
  - Professor @ ETH Zurich CS, since September 2015 (started May 2016)
  - Strecker Professor @ Carnegie Mellon University ECE/CS, 2009-2016, 2016-…
  - PhD from UT-Austin, worked at Google, VMware, Microsoft Research, Intel, AMD
  - https://people.inf.ethz.ch/omutlu/
  - omutlu@gmail.com (Best way to reach me)
  - Office hours: By appointment (email me)

- **Research and Teaching in:**
  - Computer architecture, computer systems, bioinformatics, hardware security
  - Memory and storage systems
  - Hardware security
  - Fault tolerance
  - Hardware/software cooperation
  - Genome analysis and application-algorithm-hardware co-design
  - …

Course Info: Lecturer & PhD Assistants

- Head Assistant
  - Dr. Juan Gómez Luna

- Vice-Head Assistant
  - Hasan Hassan

- Lecturer
  - Dr. Frank Gurkaynak

- (Other) Key Assistants and Guest Lecturers
  - Dr. Mohammed Alser
  - Dr. Lois Orosa
  - Dr. Jawad Haj-Yahya
  - Dr. Jisung Park

Course Info: PhD Assistants

- (Other) Key Assistants and Guest Lecturers (cont.)
  - Minesh Patel
  - Giray Yaglikci
  - Can Firtina
  - Geraldo De Oliveira Junior
  - Rahul Bera
  - Konstantinos Kanellopoulos

Course Info: Student Assistants

- Roknoddin Azizibarzoki
- Tim Fischer
- Lukas Gygi
- Leo Horné
- Lara Lazier
- Artur Melo
- Chris Mnuk
- Nathan Neike
- Arpan Prasad
- Nina Richter
- João Dinis Sanches Ferreira
- Taha Shahroodi
- Roberto Starc

Course Info: Lab Assistants (I)

- **Tuesday 15-17**
  - TBD

- **Wednesday 15-17**
  - TBD

Course Info: Lab Assistants (II)

- **Friday 8-10**
  - TBD

- **Friday 10-12**
  - TBD

If You Need Help

- Post your question on Q&A Forum (soon announced)
  - **Preferred** for **technical** questions

- Write an e-mail to:
  - digitaltechnik@lists.inf.ethz.ch
  - The instructor and all assistants will receive this e-mail

- Come to office hours (CAB H 31.2)
  - Monday 1:30pm-2:30pm
  - Tuesday: 5pm-6pm
  - Wednesday: 10am-11am
    - We might need to change the room due to space limitations. In that case, we will announce it in advance

Where to Get Up-to-date Course Info?

- Website:
  - https://safari.ethz.ch/digitaltechnik/
  - Lecture slides and videos
  - Readings
  - Lab information
  - Course schedule, handouts, FAQs
  - Software
  - Plus other useful information for the course
  - Check frequently for announcements and due dates
  - This is your single point of access to all resources

- Your ETH Email
- Lecturers and Teaching Assistants

Lecture and Lab Times and Policies

- **Lectures:**
  - Thursday and Fridays, 13:15-15:00
  - HG F7 (F5 overflow)
  - Attendance is for your benefit and is therefore important
  - Some days, we will have guest lectures and exercise sessions

- **Lab sessions:**
  - See online
  - You should definitely attend the lab sessions
    - In-class evaluation (70%) and mandatory lab reports (30%)
  - Labs will start on February 28th
  - Lab information and handouts are here:
    - https://safari.ethz.ch/digitaltechnik/spring2020/doku.php?id=labs

Lab Organization

- Groups
    - Choose your preferred group in Moodle
        - https://moodle-app2.let.ethz.ch/mod/choicegroup/view.php?id=412173
        - Due 24.02.2020 at 11:59pm
    - Choose your partner
        - https://moodle-app2.let.ethz.ch/mod/feedback/view.php?id=418396
        - Due 24.02.2020 at 11:59pm
- Lab grades from previous years
    - https://moodle-app2.let.ethz.ch/mod/choice/view.php?id=412175
    - Choose among (due 26.02.2020 at 11:59pm):
        - 1) I will use my lab grades from previous years, and I won't do the labs this year
        - 2) I will use my lab grades from previous years, but I will do the labs this year
        - 3) I won't use my lab grades from previous years. I will do the labs this year

Final Exam

- **180-minute written exam**

  - Find examination rules in Course Catalogue

  - Also in the first page of previous exams
    - https://safari.ethz.ch/digitaltechnik/spring2020/doku.php?id=exams

  - Some exam questions are similar to questions in Optional HWs
    - Optional HWs are optional, but highly recommended

Demystifying Mysteries

Levels of Transformation

"The purpose of computing is [to gain] insight" (*Richard Hamming*)
We gain and generate insight by solving problems
How do we ensure problems are solved by electrons?

Algorithm

Step-by-step procedure that is **guaranteed to terminate** where **each step is precisely stated** and **can be carried out by a computer**

- **Finiteness**
- **Definiteness**
- **Effective computability**

Many algorithms for the same problem

Problem
Algorithm
Program/Language
Runtime System (VM, OS, MM)
ISA (Architecture)
Microarchitecture
Logic
Devices
Electrons

ISA (Instruction Set Architecture)

Interface/contract between SW and HW.

What the programmer assumes hardware will satisfy.

Microarchitecture
An implementation of the ISA

Digital logic circuits
Building blocks of micro-arch (e.g., gates)

Aside: A Famous Work By Hamming

- Hamming, "Error Detecting and Error Correcting Codes," Bell System Technical Journal 1950.

- Introduced the concept of Hamming distance
  - number of locations in which the corresponding symbols of two equal-length strings is different
- Developed a theory of codes used for error detection and correction

- Also see:
  - Hamming, "You and Your Research," Talk at Bell Labs, 1986.
  - http://www.cs.virginia.edu/~robins/YouAndYourResearch.html

Levels of Transformation, Revisited

- A user-centric view: computer designed for users
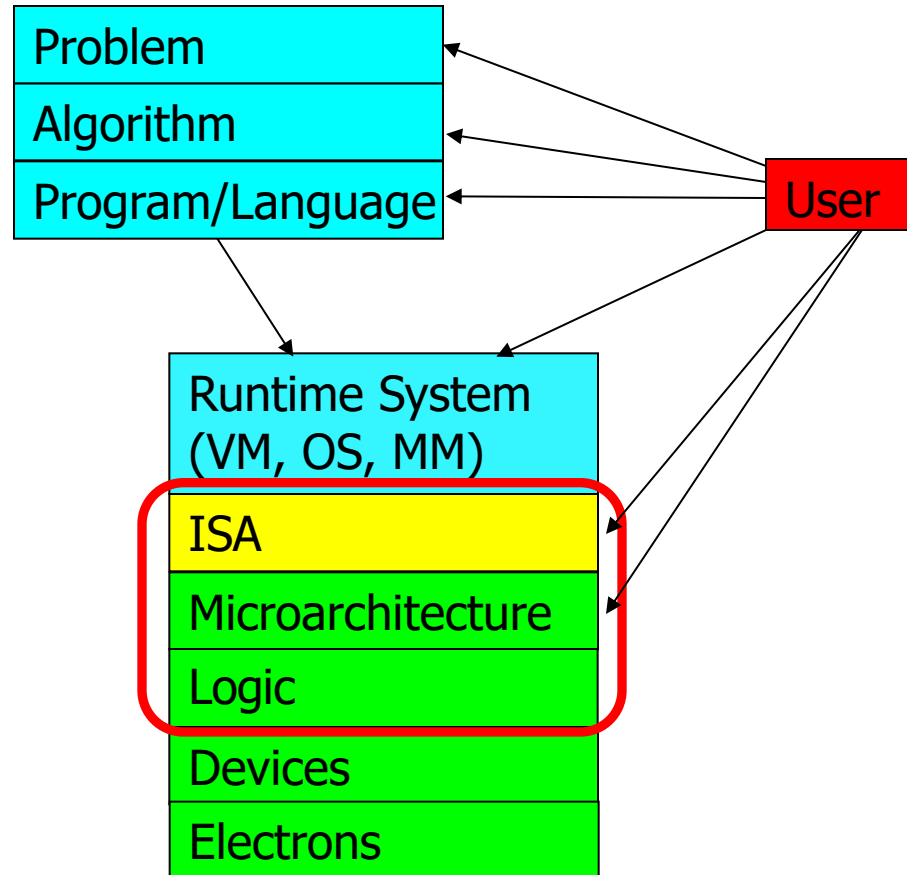
- The entire stack should be optimized for user

The Power of Abstraction

- **Levels of transformation create abstractions**
  - Abstraction: A higher level only needs to know about the interface to the lower level, not how the lower level is implemented
  - E.g., high-level language programmer does not really need to know what the ISA is and how a computer executes instructions

- **Abstraction improves productivity**
  - No need to worry about decisions made in underlying levels
  - E.g., programming in Java vs. C vs. assembly vs. binary vs. by specifying control signals of each transistor every cycle

- Then, why would you want to know what goes on underneath or above?

Crossing the Abstraction Layers

- As long as everything goes well, not knowing what happens underneath (or above) is not a problem.

- What if
    - The program you wrote is running slow?
    - The program you wrote does not run correctly?
    - The program you wrote consumes too much energy?
    - Your system just shut down and you have no idea why?
    - Someone just compromised your system and you have no idea how?

- What if
    - The hardware you designed is too hard to program?
    - The hardware you designed is too slow because it does not provide the right primitives to the software?

- What if
    - You want to design a much more efficient and higher performance system?

Crossing the Abstraction Layers

- Two goals of this course (especially the second half) are

  - to understand how a processor works underneath the software layer and how decisions made in hardware affect the software/programmer

  - to enable you to be comfortable in making design and optimization decisions that cross the boundaries of different layers and system components

Some Example "Mysteries"