Memory Systems and Memory-Centric Computing

Lecture 1: Trends, Challenges, Opportunities

Onur Mutlu

omutlu@gmail.com

https://people.inf.ethz.ch/omutlu

15 July 2024

HiPEAC ACACES Summer School 2024

SAFARI **ETH**zürich

Brief Self Introduction

- ## Onur Mutlu
  - Full Professor @ ETH Zurich ITET (INFK), since Sept 2015
  - Strecker Professor @ Carnegie Mellon University ECE (CS), 2009-2016, 2016-…
  - Started the Comp Arch Research Group @ Microsoft Research, 2006-2009
  - Worked @ Google, VMware, Microsoft Research, Intel, AMD
  - PhD in Computer Engineering from University of Texas at Austin in 2006
  - BS in Computer Engineering & Psychology from University of Michigan in 2000
  - https://people.inf.ethz.ch/omutlu/   omutlu@gmail.com

- ## Research and Teaching in:
  - **Computer architecture, systems, hardware security, bioinformatics**
  - Memory and storage systems
  - Robust & dependable hardware systems: security, safety, predictability, reliability
  - Hardware/software cooperation
  - New computing paradigms; architectures with emerging technologies/devices
  - Architectures for bioinformatics, genomics, health, medicine, AI/ML
  - …

SAFARI Research Group

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

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

40+ Researchers

SAFARI
SAFARI Research Group
safari.ethz.ch

Think BIG, Aim HIGH!

SAFARI   https://safari.ethz.ch

SAFARI Newsletter January 2021 Edition

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

SAFARI Newsletter December 2021 Edition

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

SAFARI Newsletter June 2023 Edition

- https://safari.ethz.ch/safari-newsletter-june-2023/

SAFARI Newsletter July 2024 Edition

- **https://safari.ethz.ch/safari-newsletter-july-2024/**

SAFARI Introduction & Research

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

Think BIG, Aim HIGH!

SAFARI   https://www.youtube.com/watch?v=mV2OuB2djEs

SAFARI PhD and Post-Doc Alumni

- https://safari.ethz.ch/safari-alumni/

- Hasan Hassan (Rivos), EDAA Outstanding Dissertation Award 2023; S&P 2020 Best Paper Award, 2020 Pwnie Award, IEEE Micro TP HM 2020
- Christina Giannoula (Univ. of Toronto), NTUA Best Dissertation Award 2023
- Minesh Patel (Rutgers, Asst. Prof.), DSN Carter Award Best Thesis 2022; ETH Medal 2023; MICRO'20 & DSN'20 Best Paper Awards; ISCA HoF 2021
- Damla Senol Cali (Bionano Genomics), SRC TECHCON 2019 Best Student Presentation Award; RECOMB-Seq 2018 Best Poster Award
- Nastaran Hajinazar (Intel)
- Gagandeep Singh (AMD/Xilinx), FPL 2020 Best Paper Award Finalist
- Amirali Boroumand (Stanford Univ → Google), SRC TECHCON 2018 Best Presentation Award
- Jeremie Kim (Apple), EDAA Outstanding Dissertation Award 2020; IEEE Micro Top Picks 2019; ISCA/MICRO HoF 2021
- Nandita Vijaykumar (Univ. of Toronto, Assistant Professor), ISCA Hall of Fame 2021
- Kevin Hsieh (Microsoft Research, Senior Researcher)
- Justin Meza (Facebook), HiPEAC 2015 Best Student Presentation Award; ICCD 2012 Best Paper Award
- Mohammed Alser (ETH Zurich), IEEE Turkey Best PhD Thesis Award 2018
- Yixin Luo (Google), HPCA 2015 Best Paper Session
- Kevin Chang (Facebook), SRC TECHCON 2016 Best Student Presentation Award
- Rachata Ausavarungnirun (KMUNTB, Assistant Professor), NOCS 2015 and NOCS 2012 Best Paper Award Finalist
- Gennady Pekhimenko (Univ. of Toronto, Assistant Professor), ISCA Hall of Fame 2021; ASPLOS 2015 SRC Winner
- Vivek Seshadri (Microsoft Research)
- Donghyuk Lee (NVIDIA Research, Senior Researcher), HPCA Hall of Fame 2018
- Yoongu Kim (Software Robotics → Google), IFIP JCL Award'24, TCAD'19 Top Pick Award; IEEE Micro Top Picks'10; HPCA'10 Best Paper Session
- Lavanya Subramanian (Intel Labs → Facebook)

- Samira Khan (Univ. of Virginia, Assistant Professor), HPCA 2014 Best Paper Session
- Saugata Ghose (Univ. of Illinois, Assistant Professor), DFRWS-EU 2017 Best Paper Award
- Jawad Haj-Yahya (Huawei Research Zurich, Principal Researcher)
- Lois Orosa (Galicia Supercomputing Center, Director)
- Jisung Park (POSTECH, Assistant Professor)
- Gagandeep Singh (AMD/Xilinx, Researcher)
- Juan Gomez-Luna (NVIDIA, Researcher), ISPASS 2023 Best Paper Session

SAFARI

An Interview on Computing Futures

Interview with Onur Mutlu @ ISCA 2019 on computing research & education (after Maurice Wilkes Award)

6,749 views • Oct 19, 2019

👍 195   👎 0   ↱ SHARE   ≡+ SAVE   ...

Onur Mutlu Lectures
19.1K subscribers

ANALYTICS   EDIT VIDEO

https://www.youtube.com/watch?v=8ffSEKZhmvo

Suggestions on Research, Education, PhD

SAFARI https://www.youtube.com/watch?v=83tlorht7Mc&list=PL5Q2soXY2Zi8D_5MGV6EnXEJHnV2YFBJl&index=54

Current Mission

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

Hybrid Main Memory

Heterogeneous
Processors and
Accelerators

Persistent Memory/Storage

Graphics and Vision Processing
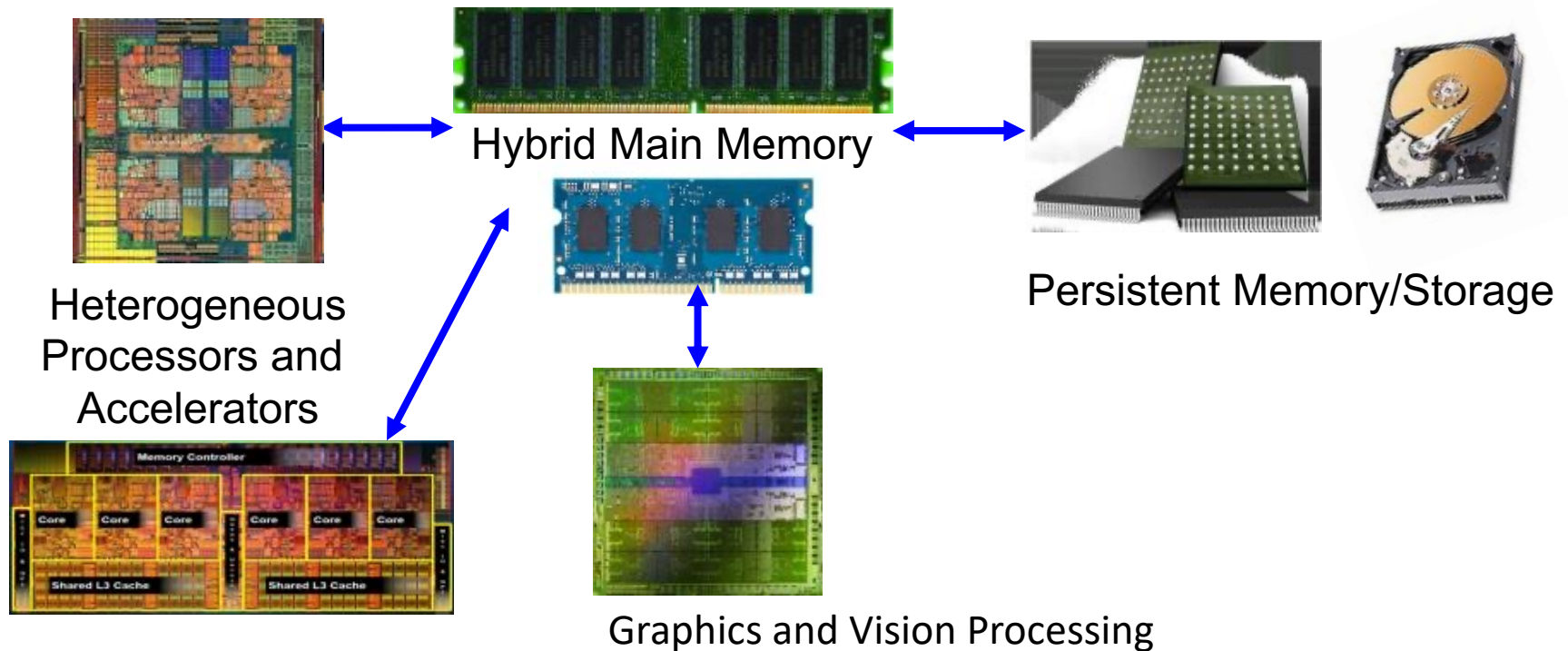
Build fundamentally better computers

Current Research Mission & Major Topics

Build fundamentally better computers

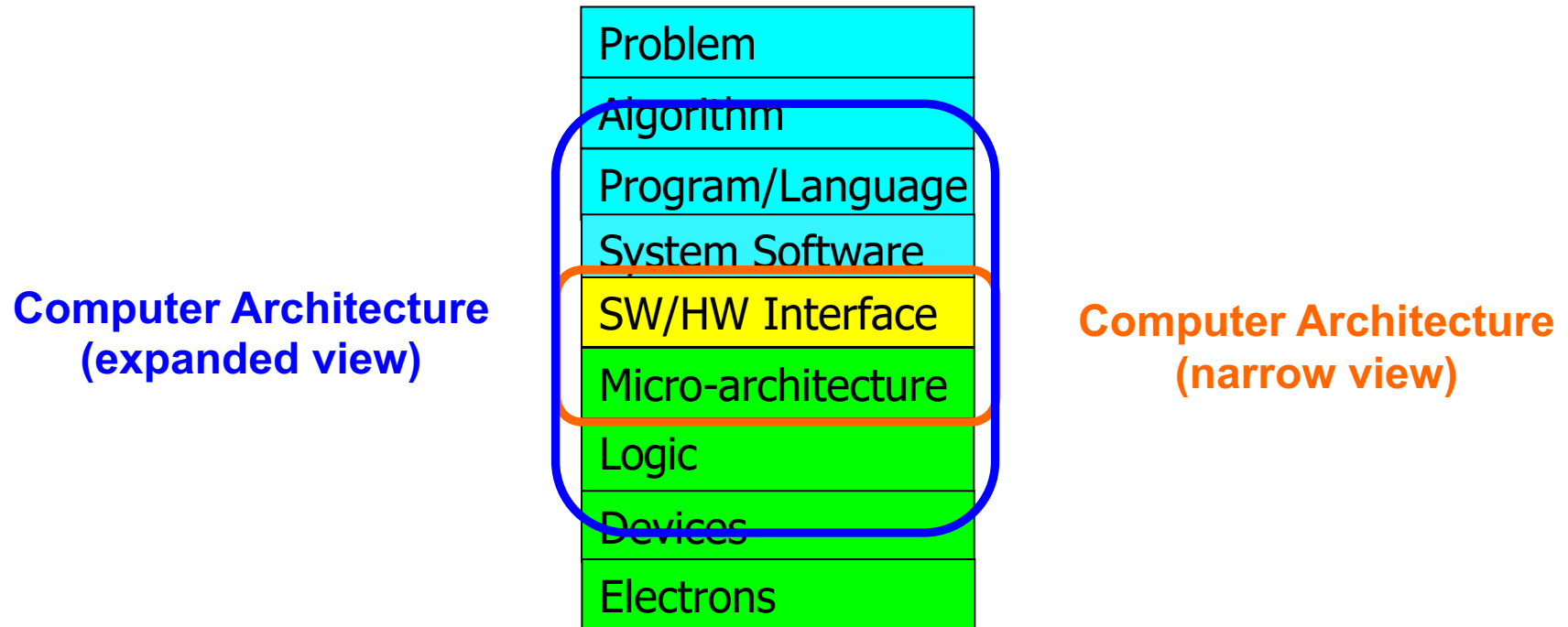| Problem |
|---|
| Algorithm |
| Program/Language |
| System Software |
| SW/HW Interface |
| Micro-architecture |
| Logic |
| Devices |
| Electrons |
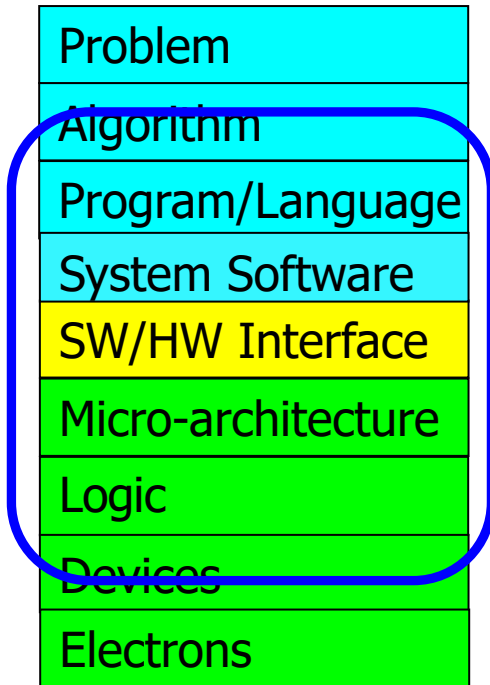
Broad research spanning apps, systems, logic with architecture at the center

- **Data-centric arch. for low energy & high perf.**
  - Proc. in Mem/DRAM, NVM, unified mem/storage

- **Low-latency & predictable architectures**
  - Low-latency, low-energy yet low-cost memory
  - QoS-aware and predictable memory systems

- **Fundamentally secure/reliable/safe arch.**
  - Tolerating all bit flips; patchable HW; secure mem

- **Architectures for ML/AI/Genomics/Health/Med**
  - Algorithm/arch./logic co-design; full heterogeneity

- **Data-driven and data-aware architectures**
  - ML/AI-driven architectural controllers and design
  - Expressive memory and expressive systems

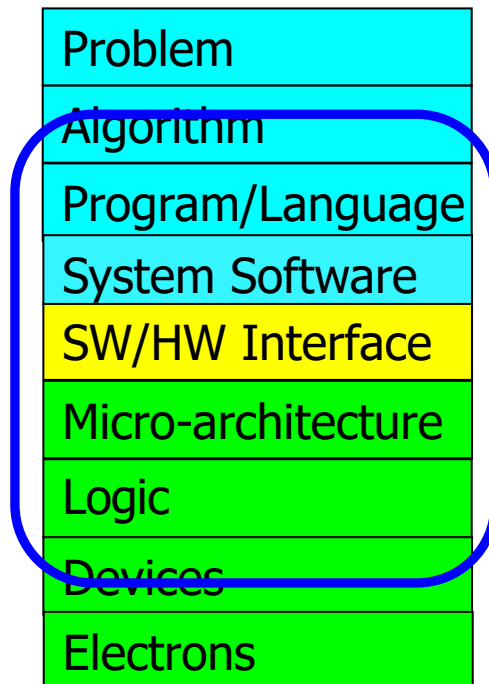*SAFARI*

Five Key Current Directions

- Fundamentally Robust (Secure/Reliable/Safe) Architectures

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

- Fundamentally Low-Latency and Predictable Architectures

- Fundamentally Intelligent and Evolving Architectures
    - ML/AI-Assisted (Data-driven) and Data-aware Architectures

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

SAFARI

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

Axiom

To achieve the highest efficiency, performance, robustness:

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

Principle: Teaching and Research

...

Teaching drives Research

Research drives Teaching

...

Open Source Tools: SAFARI GitHub

SAFARI Research Group at ETH Zurich and Carnegie Mellon University

Site for source code and tools distribution from SAFARI Research Group at ETH Zurich and Carnegie Mellon University.

👥 **440** followers ◯ ETH Zurich and Carnegie Mellon U... 🔗 https://safari.ethz.ch/ ✉ omutlu@gmail.com

🏠 **Overview**   📖 Repositories **98**   🗂 Projects   📦 Packages   👤 People **13**

📖 ramulator (Public)

A Fast and Extensible DRAM Simulator, with built-in support for modeling many different DRAM technologies including DDRx, LPDDRx, GDDRx, WIOx, HBMx, and various academic proposals. Described in the...

● C++   ☆ 532   ⑂ 206

📖 prim-benchmarks (Public)

PrIM (Processing-In-Memory benchmarks) is the first benchmark suite for a real-world processing-in-memory (PIM) architecture. PrIM is developed to evaluate, analyze, and characterize the first publ...

● C   ☆ 126   ⑂ 47

📖 MQSim (Public)

MQSim is a fast and accurate simulator modeling the performance of modern multi-queue (MQ) SSDs as well as traditional SATA based SSDs. MQSim faithfully models new high-bandwidth protocol implement...

● C++   ☆ 268   ⑂ 143

📖 rowhammer (Public)

Source code for testing the Row Hammer error mechanism in DRAM devices. Described in the ISCA 2014 paper by Kim et al. at http://users.ece.cmu.edu/~omutlu/pub/dram-row-hammer_isca14.pdf.

● C   ☆ 211   ⑂ 42

📖 SoftMC (Public)

SoftMC is an experimental FPGA-based memory controller design that can be used to develop tests for DDR3 SODIMMs using a C++ based API. The design, the interface, and its capabilities and limitatio...

● Verilog   ☆ 120   ⑂ 27

📖 Pythia (Public)

A customizable hardware prefetching framework using online reinforcement learning as described in the MICRO 2021 paper by Bera et al. (https://arxiv.org/pdf/2109.12021.pdf).

● C++   ☆ 109   ⑂ 34

https://github.com/CMU-SAFARI/

18

Referenced Papers, Talks, Artifacts

- All are available at

    https://people.inf.ethz.ch/omutlu/projects.htm

    https://www.youtube.com/onurmutlulectures

    https://github.com/CMU-SAFARI/

SAFARI

Quick Course Overview

What Will You Learn in This Course?

- **Memory Systems and Memory-Centric Computing**
  - July 15-19, 2024

- Topic 1: Memory Trends, Challenges, Opportunities, Basics
- Topic 2: Memory-Centric Computing
- Topic 3: Memory Robustness: RowHammer, RowPress & Beyond
- Topic 4: Machine Learning Driven Memory Systems
- Topic 5 (another course): Architectures for Genomics and ML
- Topic 6 (unlikely): Non-Volatile Memories and Storage
- Topic 7 (unlikely): Memory Latency, Predictability & QoS

- Major Overview Reading:
  - Mutlu et al., "A Modern Primer on Processing in Memory," Book Chapter on Emerging Computing and Devices, 2022.

Course Website & Some Study Materials

- **https://safari.ethz.ch/memory_systems/ACACES2024/**

- "A Modern Primer on Processing in Memory" (Emerging Computing, 2022)
https://arxiv.org/abs/2012.03112

- "Fundamentally Understanding and Solving RowHammer" (ASP-DAC, 2023)
https://arxiv.org/abs/2211.07613

- "Intelligent Architectures for Intelligent Computing Systems" (DATE, 2021)
https://arxiv.org/abs/2012.12381

- "Accelerating Neural Network Inference With Processing-in-DRAM: From the Edge to the Cloud" (IEEE Micro, 2022) https://arxiv.org/abs/2209.08938

- "Accelerating Genome Analysis via Algorithm-Architecture Co-Design" (DAC, 2023)
https://arxiv.org/abs/2305.00492

- "Memory-Centric Computing" (DAC, 2023) https://arxiv.org/abs/2305.20000

- "RowHammer: A Retrospective" (TCAD, 2019) https://arxiv.org/abs/1904.09724

- "Accelerating Genome Analysis: A Primer on an Ongoing Journey" (IEEE Micro, 2020) https://arxiv.org/abs/2008.00961

Course Information

- My Contact Information
  - Onur Mutlu
  - omutlu@gmail.com
  - https://people.inf.ethz.ch/omutlu
  - +41-79-572-1444 (my cell phone)
  - Find me during breaks and/or email any time.

- Website for Course Slides, Papers, Updates
  - **https://safari.ethz.ch/memory_systems/ACACES2024/**

- For the curious – ACACES 2013 & 2018 courses:
  - https://people.inf.ethz.ch/omutlu/acaces2013-memory.html
  - https://people.inf.ethz.ch/omutlu/acaces2018.html

SAFARI

This Course

- Will cover many problems and potential solutions related to the design of memory systems & memory-centric computers

- The design of memory systems poses many
  - Difficult research and engineering problems
  - Important fundamental problems
  - Industry-relevant problems
  - **Problems whose solutions can revolutionize the world**

- Many creative and insightful solutions are needed to solve these problems

- Goal: Acquire the basics to develop such solutions (by covering fundamentals and cutting-edge research)

How To Make the Best Out of This Course

- Be alert during lectures – they will be fast paced

- Do the readings (and explore even more)
  - I will provide many references

- Go back and reinforce fundamentals (as needed)
  - I will provide pointers to basic computer architecture materials (lecture videos, slides, readings, exams, …)

- Remember "Chance favors the prepared mind." (Pasteur)

Unfortunately, No Time For:

- Memory Latency

- Memory Interference and QoS, Predictable Performance
  - QoS-aware Memory Systems

- Emerging Memory Technologies and Hybrid Memories

- Interconnects

- Caching, Prefetching, Memory Hierarchy Design

- You can find many materials on these at my online lectures
  - https://people.inf.ethz.ch/omutlu/teaching.html

Links for Basic Materials

- Digital Design & Computer Architecture Course (Spring 2023):
  - **https://safari.ethz.ch/digitaltechnik/spring2023/**
  - **https://www.youtube.com/onurmutlulectures**
  - https://www.youtube.com/playlist?list=PL5Q2soXY2Zi-EImKxYYY1SZuGiOAOBKaf

Links for More Advanced Materials

- Computer Architecture Course (Fall 2021):
  - **https://safari.ethz.ch/architecture/fall2021/**
  - **https://www.youtube.com/onurmutlulectures**
  - https://www.youtube.com/playlist?list=PL5Q2soXY2Zi-Mnk1PxjEIG32HAGILkTOF

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 2023

PIM Review and Open Problems (II)

A Workload and Programming Ease Driven Perspective of Processing-in-Memory

Saugata Ghose[†]   Amirali Boroumand[†]   Jeremie S. Kim[†§]   Juan Gómez-Luna[§]   Onur Mutlu[§†]

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

Saugata Ghose, Amirali Boroumand, Jeremie S. Kim, Juan Gomez-Luna, and Onur Mutlu,
"Processing-in-Memory: A Workload-Driven Perspective"
Invited Article in IBM Journal of Research & Development, Special Issue on Hardware for Artificial Intelligence, to appear in November 2019.
[Preliminary arXiv version]

Future Computing Platforms
Challenges and Opportunities

Why Do We Do Computing?

Answer

To Solve Problems

Answer Reworded

To Gain Insight

SAFARI   Hamming, "Numerical Methods for Scientists and Engineers," 1962.

Answer Extended

To Enable
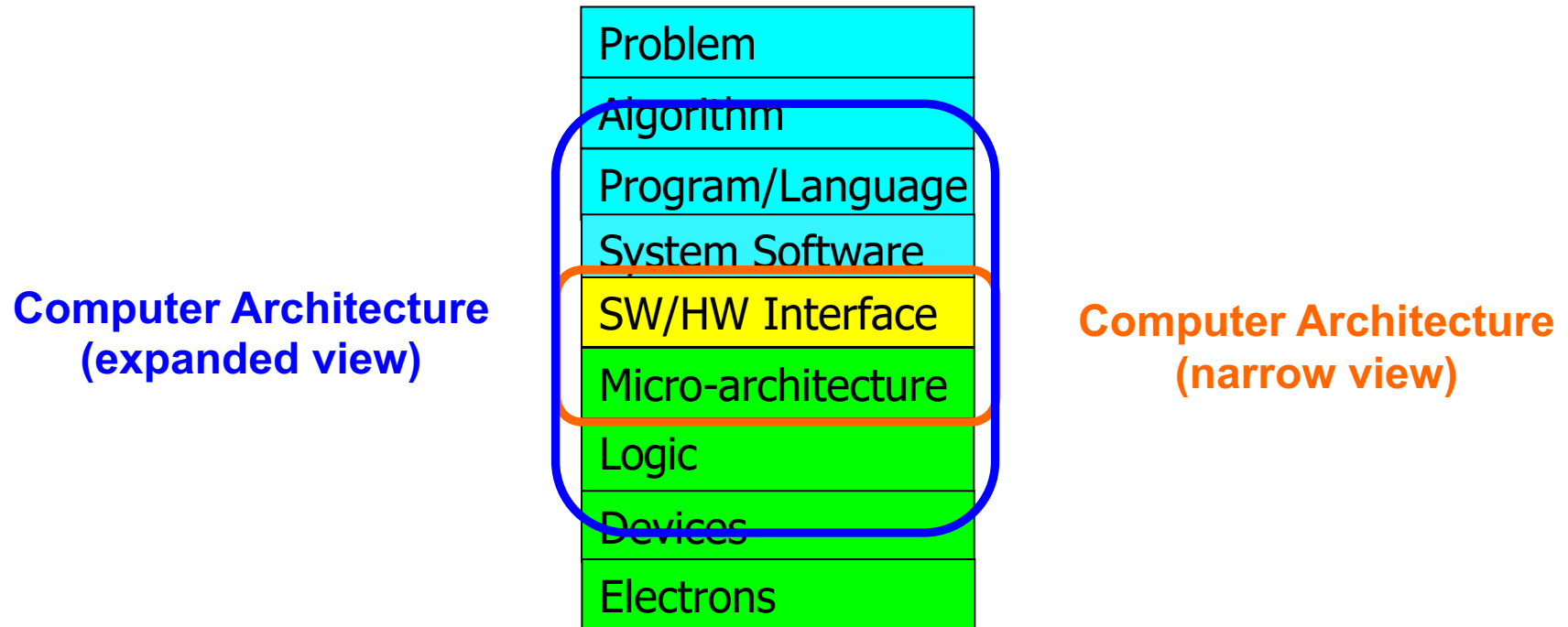# a Better Life & Future

SAFARI

How Does a Computer Solve Problems?

Answer

Orchestrating Electrons

In today's dominant technologies

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

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

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

Mohammed Alser, Zülal Bingöl, Damla Senol Cali, Jeremie Kim, Saugata Ghose, Can Alkan, Onur Mutlu **"Accelerating Genome Analysis: A Primer on an Ongoing Journey"** IEEE Micro, August 2020.

Accelerating Genome Analysis: A Primer on an Ongoing Journey
Sept.-Oct. 2020, pp. 65-75, vol. 40
DOI Bookmark: 10.1109/MM.2020.3013728

FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications
July-Aug. 2021, pp. 39-48, vol. 41
DOI Bookmark: 10.1109/MM.2021.3088396

MinION from ONT

SmidgION from ONT

An Example System in Your Pocket

Sensors

Storage

Main Memory

SoC
with lots of
compute
& caches

Main Memory

Storage
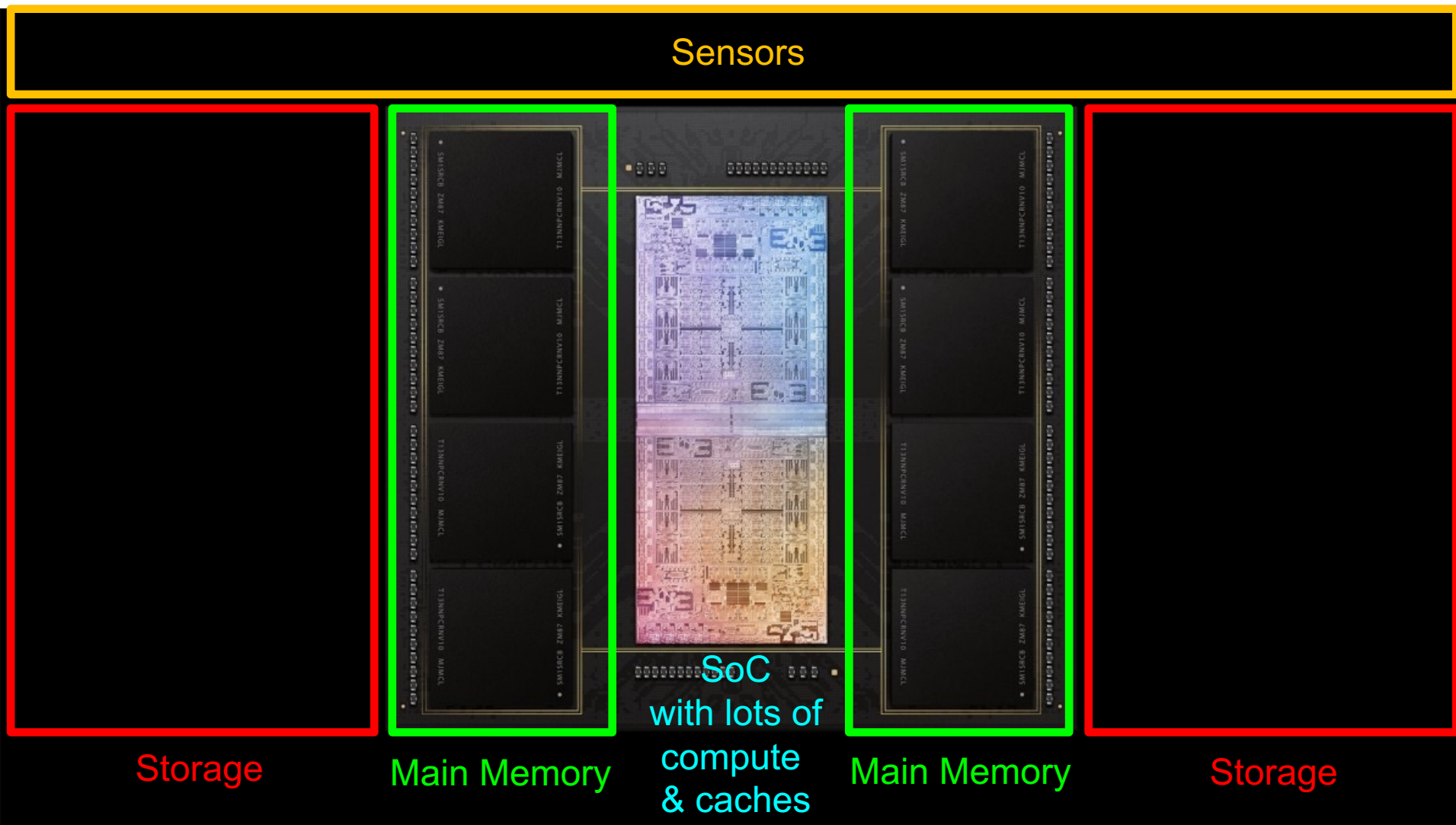
Apple M1 Ultra System (2022)

Different Platforms, Different Goals

SAFARI

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

SAFARI

Different Platforms, Different Goals

Source: https://www.itmagazine.ch/artikel/72401/Fugaku_Der_schnellste_Supercomputer_der_Welt.html

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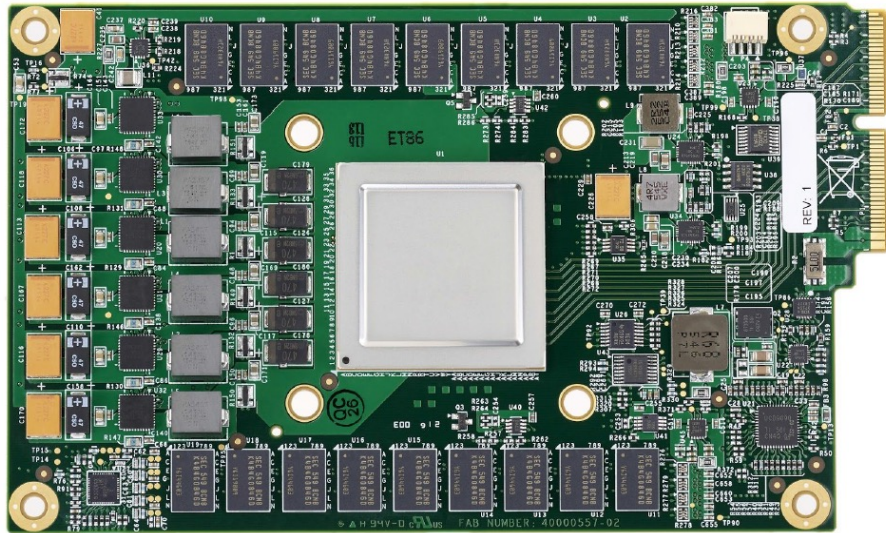
Figure 3. TPU Printed Circuit Board. It can be inserted in the slot for an SATA disk in a server, but the card uses PCIe Gen3 x16.
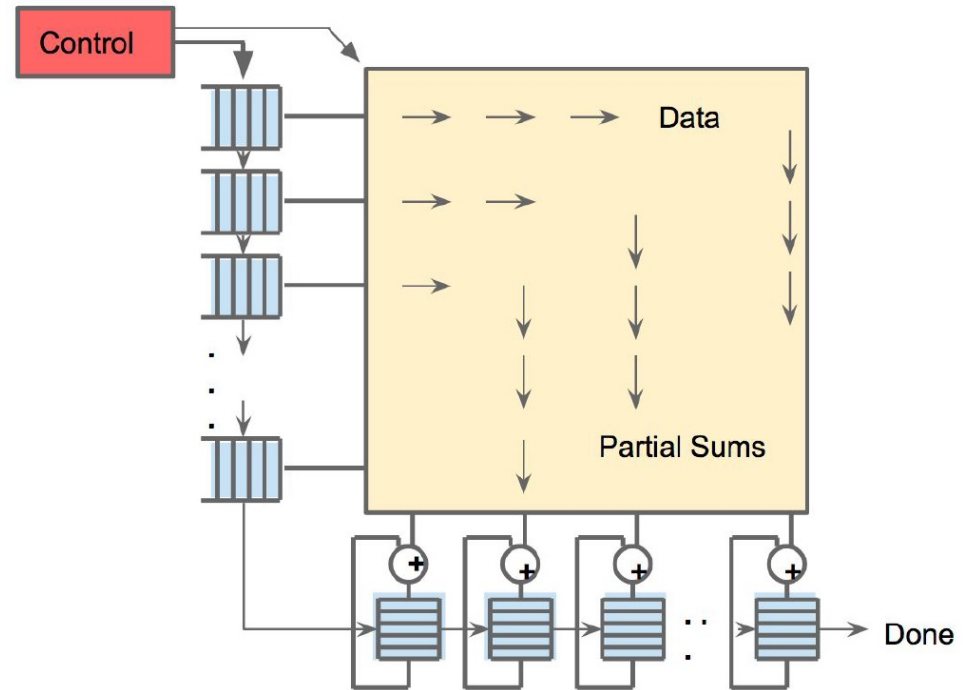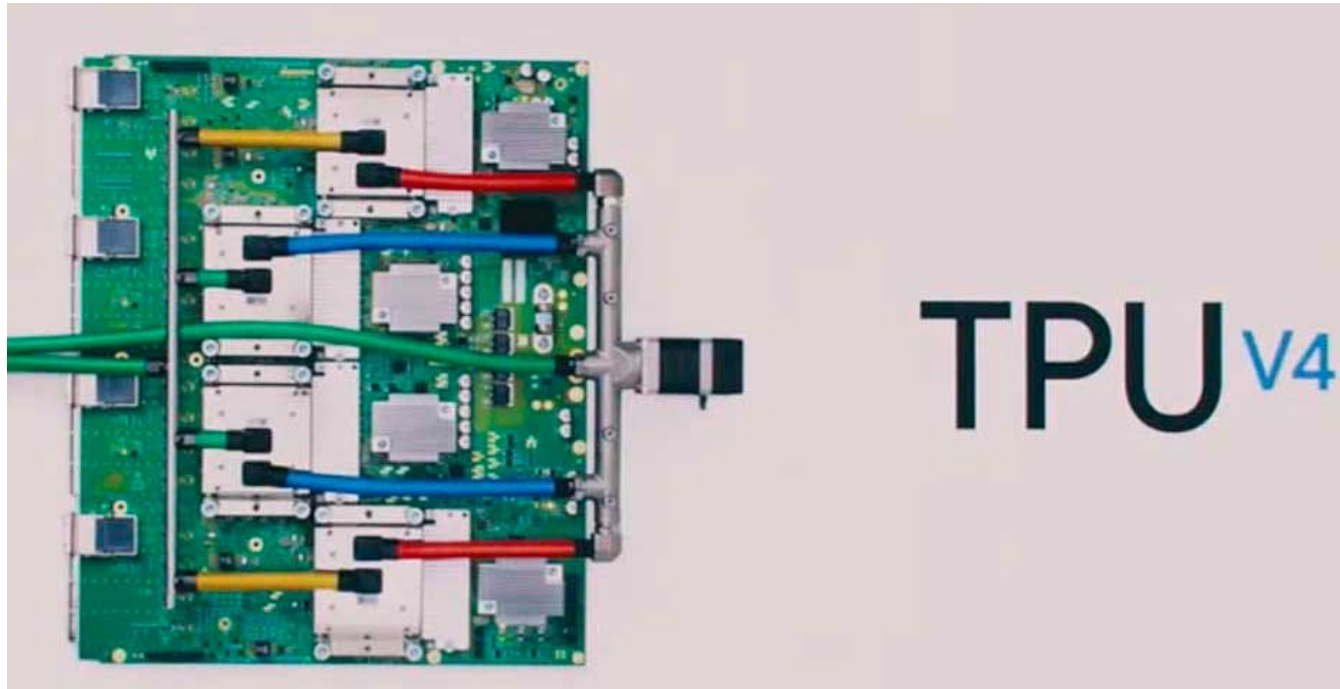

Figure 4. Systolic data flow of the Matrix Multiply Unit. Software has the illusion that each 256B input is read at once, and they instantly update one location of each of 256 accumulator RAMs.

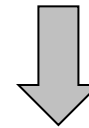Jouppi et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit", ISCA 2017.

Different Platforms, Different Goals

New ML applications (vs. TPU3):
- Computer vision
- Natural Language Processing (NLP)
- Recommender system
- Reinforcement learning that plays Go

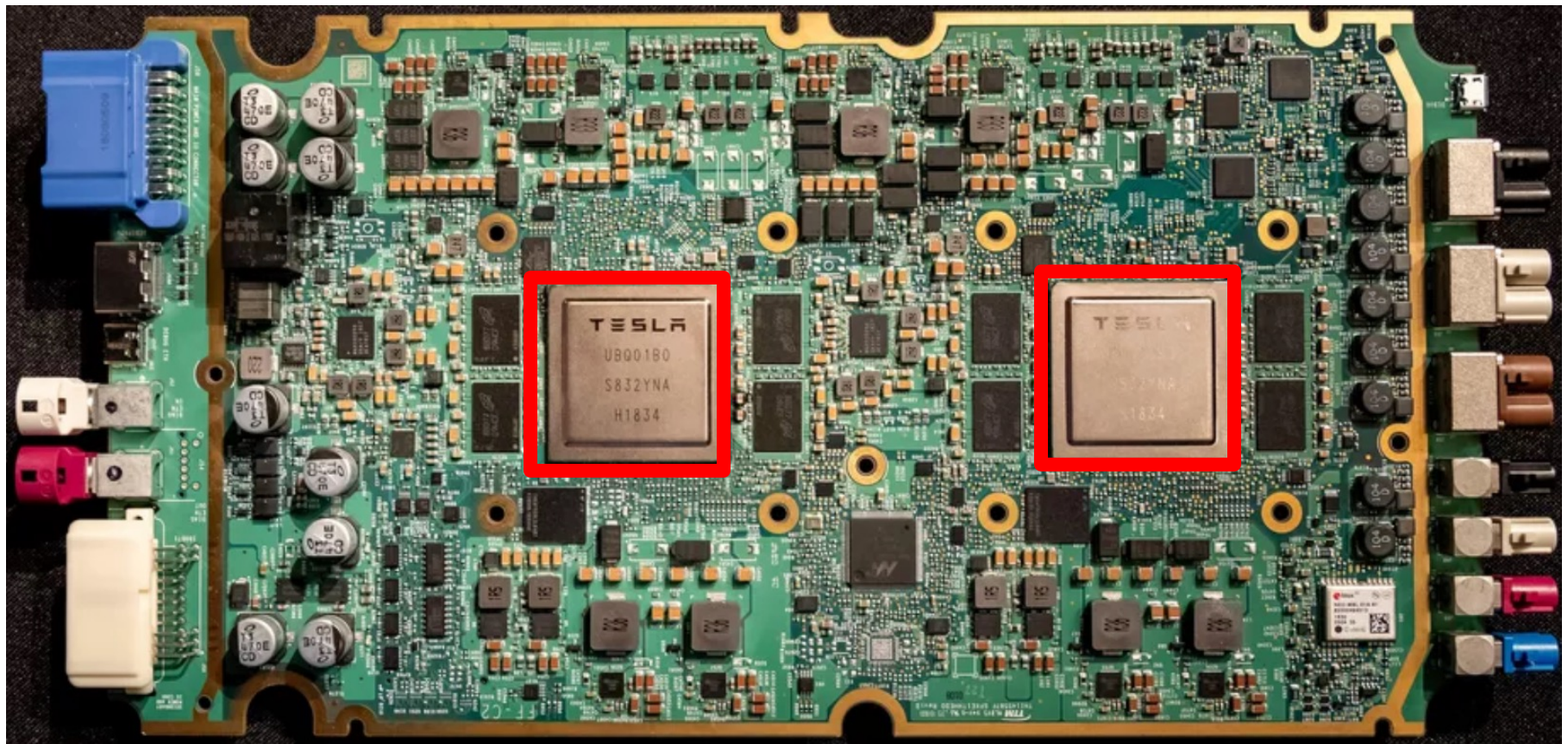**250 TFLOPS per chip in 2021** vs 90 TFLOPS in TPU3

⬇

1 ExaFLOPS per board

https://spectrum.ieee.org/tech-talk/computing/hardware/heres-how-googles-tpu-v4-ai-chip-stacked-up-in-training-tests

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.
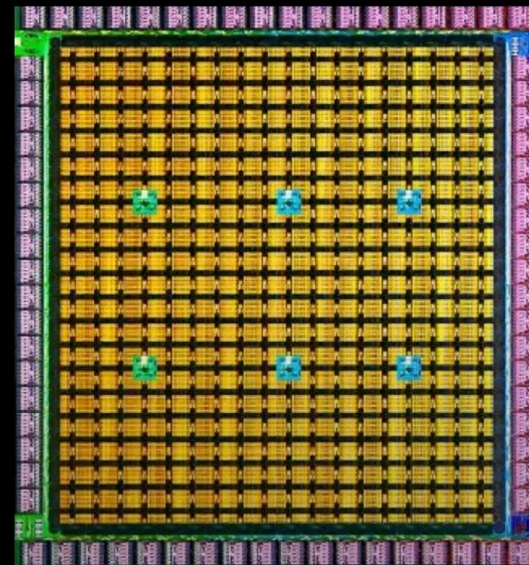
■ **Tesla Dojo Chip & System**

D1 Chip

362 TFLOPs BF16/CFP8
22.6 TFLOPs FP32

10TBps/dir. On-Chip Bandwidth
4TBps/edge. Off-Chip Bandwidth

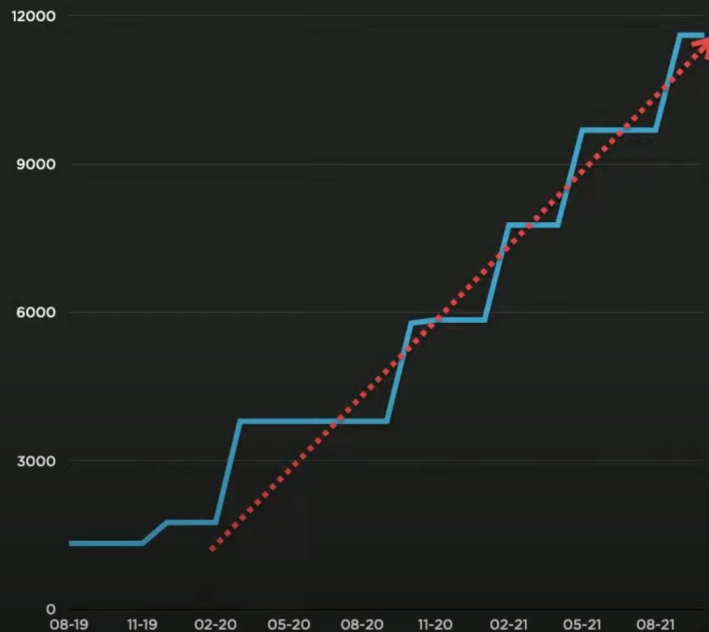400W TDP

645mm²
7nm Technology

50 Billion
Transistors

11+ Miles
Of Wires

https://www.youtube.com/watch?v=j0z4FweCy4M&t=6340s

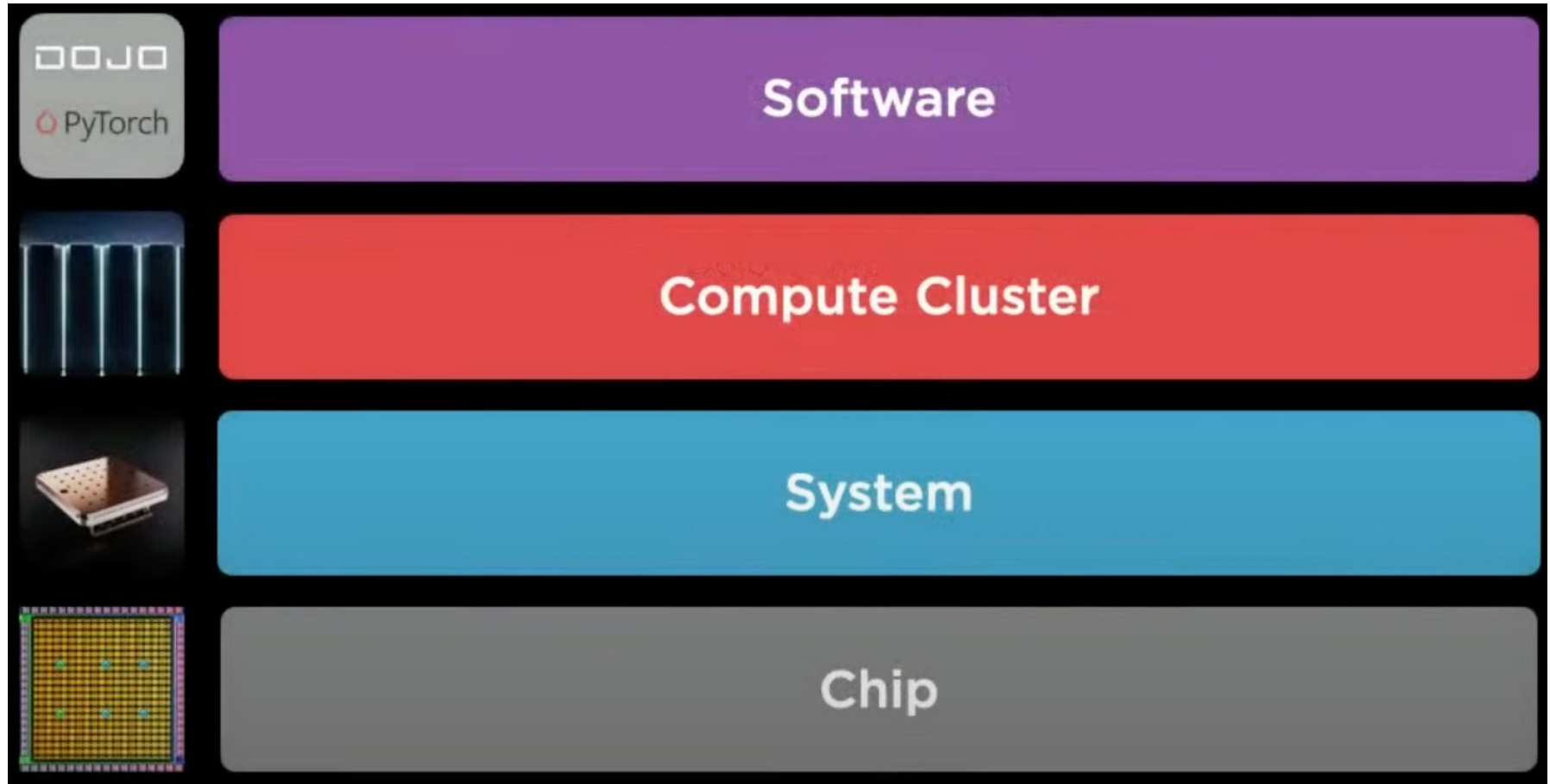# Different Platforms, Different Goals

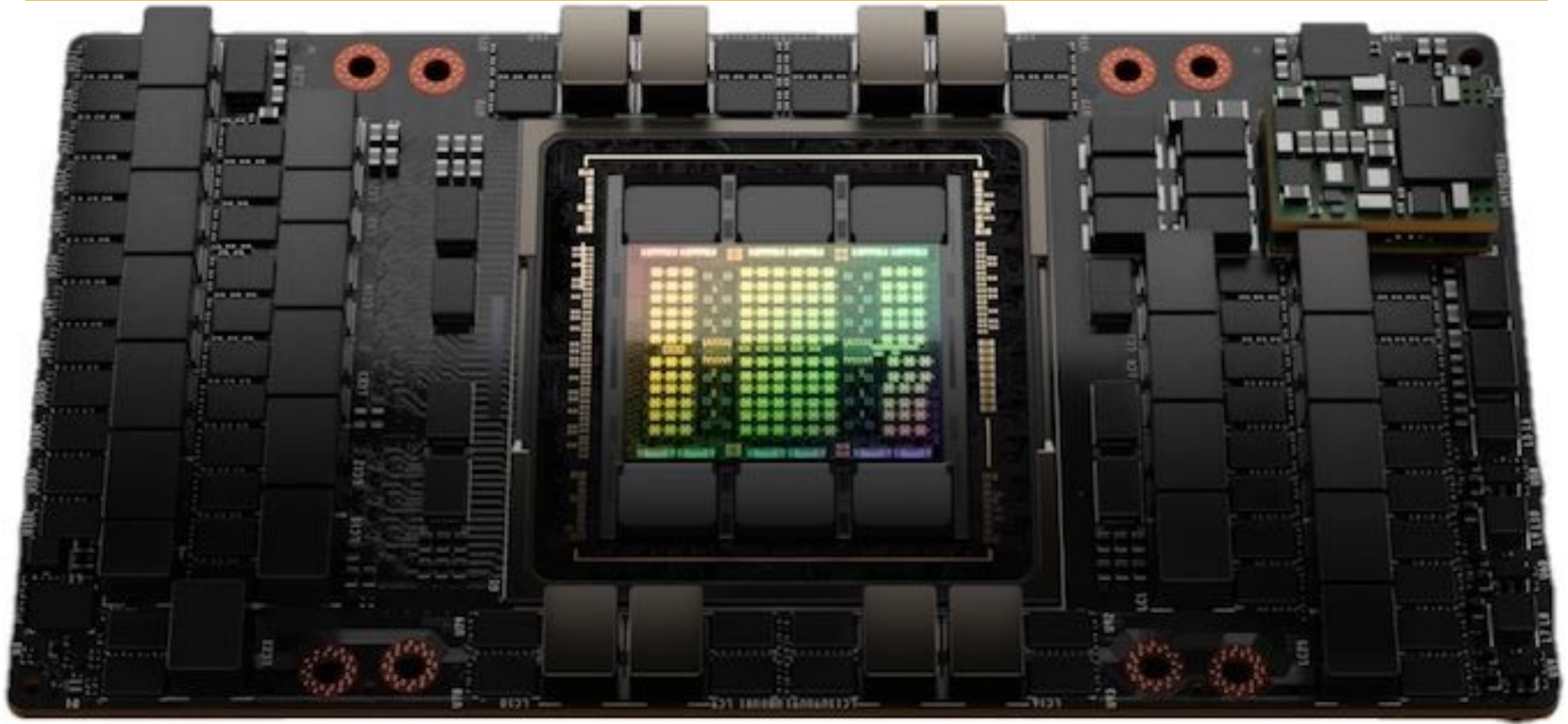- ## Tesla Dojo Chip & System

Different Platforms, Different Goals

- Tesla Dojo Chip & System

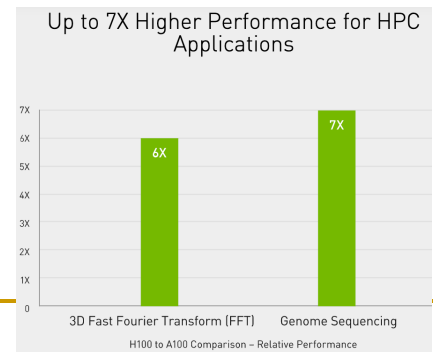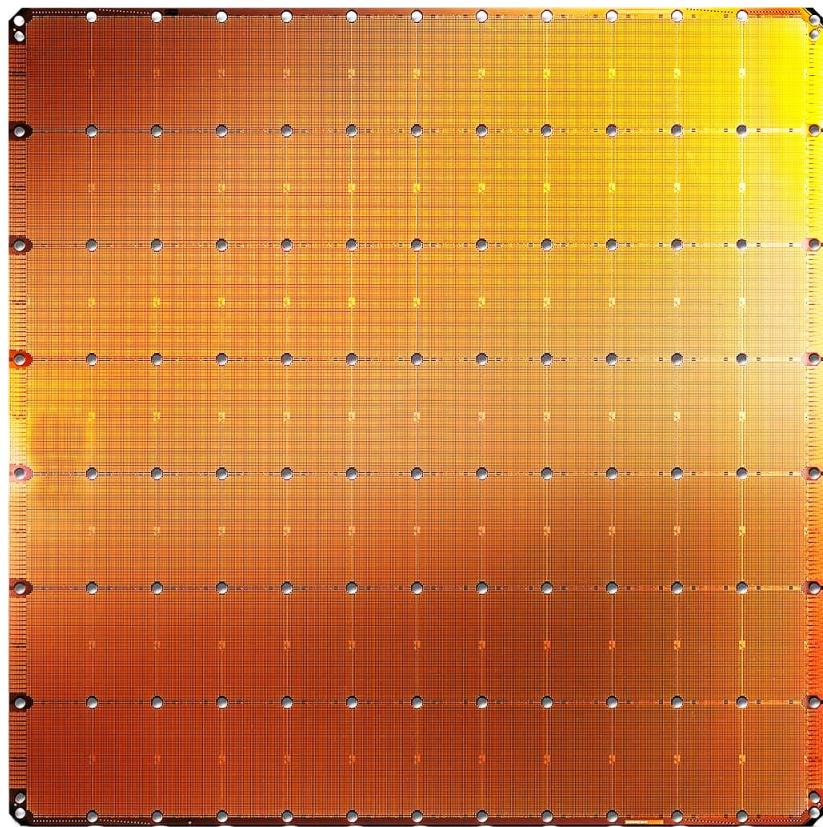# Different Platforms, Different Goals

NVIDIA is claiming a **7x improvement** in dynamic programming algorithm (**DPX instructions**) performance on a single H100 versus naïve execution on an A100.

Up to 7X Higher Performance for HPC Applications

3D Fast Fourier Transform (FFT) — 6X
Genome Sequencing — 7X

H100 to A100 Comparison – Relative Performance

https://www.nvidia.com/en-us/data-center/h100/

Cerebras's Wafer Scale Engine (2019)

- **The largest ML accelerator chip**

- **400,000 cores**

Cerebras WSE
1.2 Trillion transistors
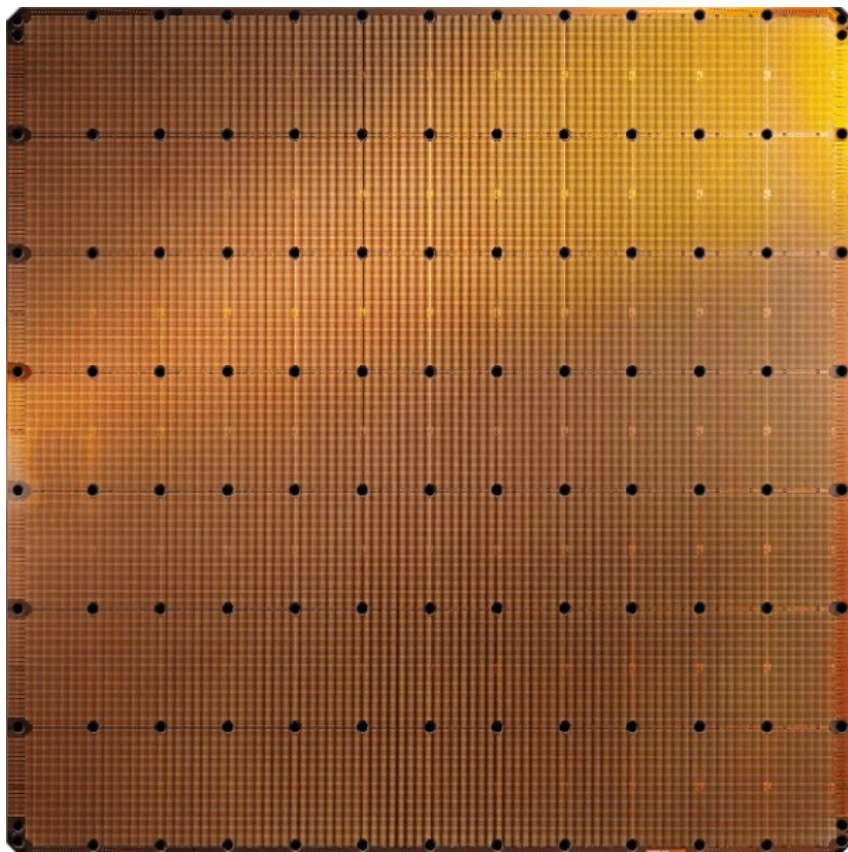46,225 mm^2

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

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

Cerebras's Wafer Scale Engine-2 (2021)

- **The largest ML accelerator chip (2021)**

- **850,000 cores**

Cerebras WSE-2
2.6 Trillion transistors
46,225 mm^2

Largest GPU
54.2 Billion transistors
826 mm^2

NVIDIA Ampere GA100

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/

Many (Other) (AI/ML) Chips

- Alibaba
- Amazon
- Facebook
- Google
- Huawei
- Intel
- Microsoft
- NVIDIA
- Tesla
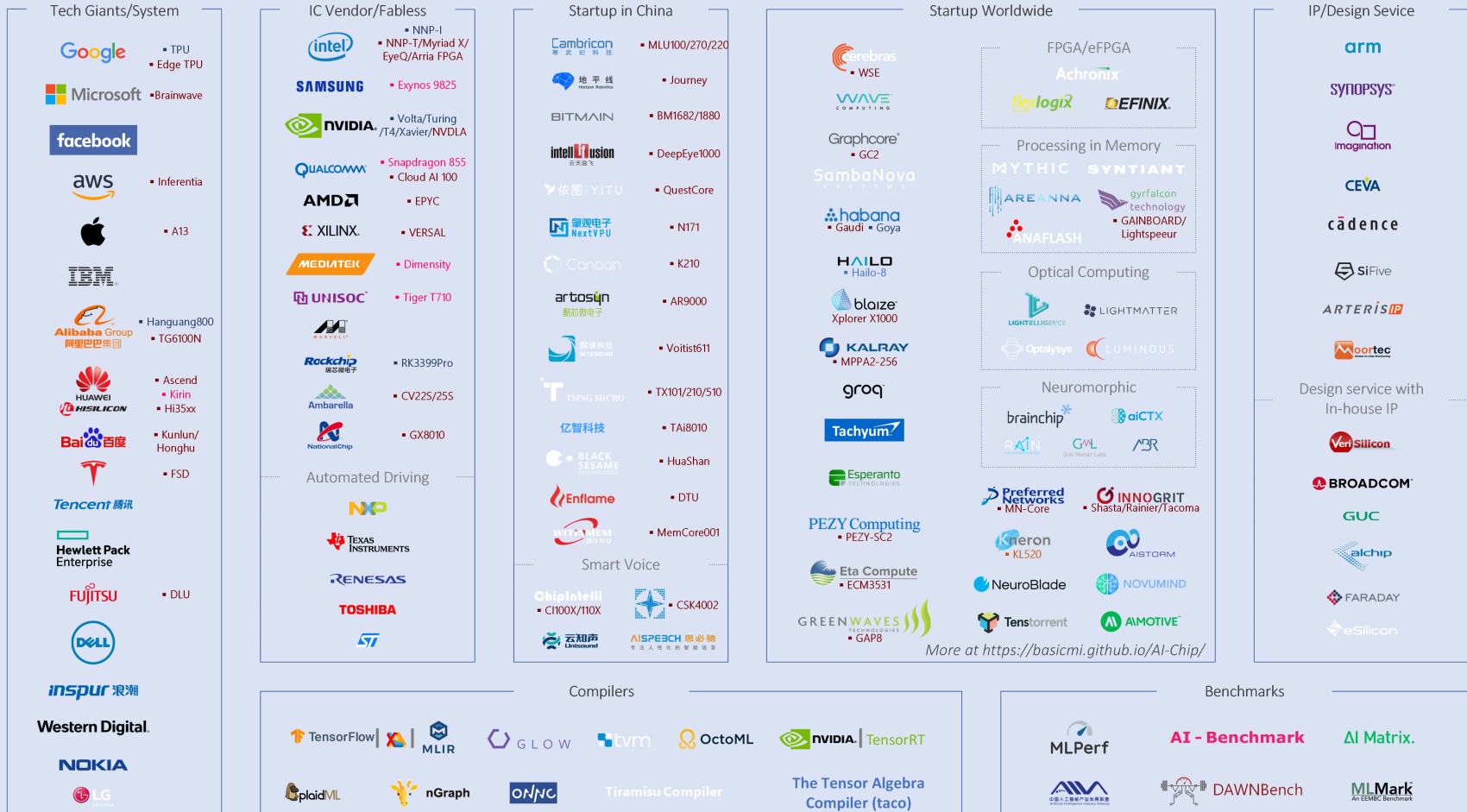- Many Others and Many Startups are Building Their Own Chips…

- **Many More to Come…**

Many (Other) AI/ML Chips (2019)

AI Chip Landscape

V0.7 Dec., 2019   S.T.

- MLPerf results available   - AI-Benchmark results available

Tech Giants/System
- Google — TPU, Edge TPU
- Microsoft — Brainwave
- facebook
- aws — Inferentia
- Apple — A13
- IBM
- Alibaba Group — Hanguang800, TG6100N
- HUAWEI HiSilicon — Ascend, Kirin, Hi35xx
- Baidu — Kunlun/Honghu
- Tesla — FSD
- Tencent
- Hewlett Packard Enterprise
- FUJITSU — DLU
- DELL
- inspur
- Western Digital
- NOKIA
- LG

IC Vendor/Fabless
- intel — NNP-I, NNP-T/Myriad X/ EyeQ/Arria FPGA
- SAMSUNG — Exynos 9825
- NVIDIA — Volta/Turing /T4/Xavier/NVDLA
- QUALCOMM — Snapdragon 855, Cloud AI 100
- AMD — EPYC
- XILINX — VERSAL
- MEDIATEK — Dimensity
- UNISOC — Tiger T710
- MARVELL
- Rockchip — RK3399Pro
- Ambarella — CV22S/25S
- NationalChip — GX8010

Automated Driving
- NXP
- TEXAS INSTRUMENTS
- RENESAS
- TOSHIBA
- ST

Startup in China
- Cambricon — MLU100/270/220
- Horizon Robotics — Journey
- BITMAIN — BM1682/1880
- intellifusion — DeepEye1000
- YITU — QuestCore
- NextVPU — N171
- Canaan — K210
- artosyn — AR9000
- INTENGINE — Voitist611
- TSING MICRO — TX101/210/510
- TAi8010
- BLACK SESAME — HuaShan
- Enflame — DTU
- WITMEM — MemCore001

Smart Voice
- Chipintelli — CI100X/110X
- CSK4002
- Unisound — GAP8

Startup Worldwide
- Cerebras — WSE
- WAVE COMPUTING
- Graphcore — GC2
- SambaNova
- habana — Gaudi, Goya
- HAILO — Hailo-8
- blaize — Xplorer X1000
- KALRAY — MPPA2-256
- groq
- Tachyum
- Esperanto
- PEZY Computing — PEZY-SC2
- Eta Compute — ECM3531
- GREENWAVES — GAP8

FPGA/eFPGA
- Achronix
- flexlogix
- EFINIX

Processing in Memory
- MYTHIC
- SYNTIANT
- AREANNA
- gyrfalcon technology — GAINBOARD/ Lightspeeur
- ANAFLASH

Optical Computing
- LIGHTELLIGENCE
- LIGHTMATTER
- Optalysys
- LUMINOUS

Neuromorphic
- brainchip
- aiCTX
- RAIN
- GML GrAI Matter Labs
- ABR
- Preferred Networks — MN-Core
- INNOGRIT — Shasta/Rainier/Tacoma
- Kneron — KL520
- AISTORM
- NeuroBlade
- NOVUMIND
- Tenstorrent
- AIMOTIVE

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

IP/Design Sevice
- arm
- SYNOPSYS
- imagination
- CEVA
- cadence
- SiFive
- ARTERIS IP
- Moortec

Design service with In-house IP
- VeriSilicon
- BROADCOM
- GUC
- alchip
- FARADAY
- eSilicon

Compilers
- TensorFlow
- MLIR
- GLOW
- tvm
- OctoML
- NVIDIA TensorRT
- plaidML
- nGraph
- ONNC
- Tiramisu Compiler
- The Tensor Algebra Compiler (taco)

Benchmarks
- MLPerf
- AI - Benchmark
- AI Matrix.
- AIIA
- DAWNBench
- MLMark An EEMBC Benchmark

All information contained within this infographic is gathered from the internet and periodically updated, no guarantee is given that the information provided is correct, complete, and up-to-date.

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
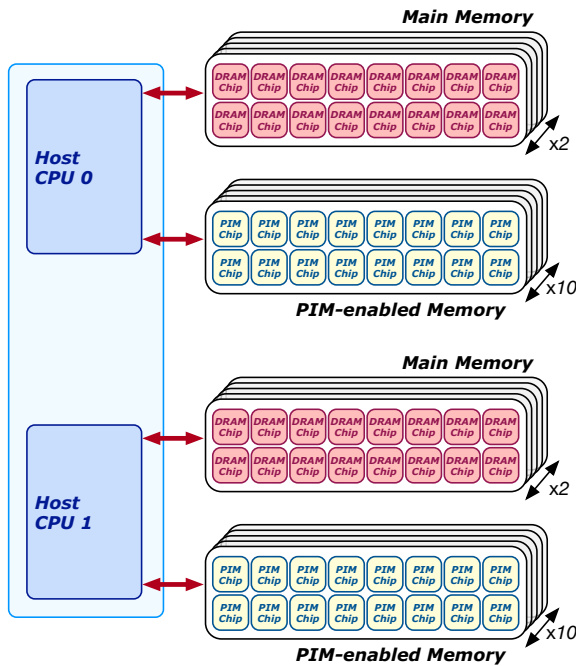 - **Large amounts of** compute & memory bandwidth
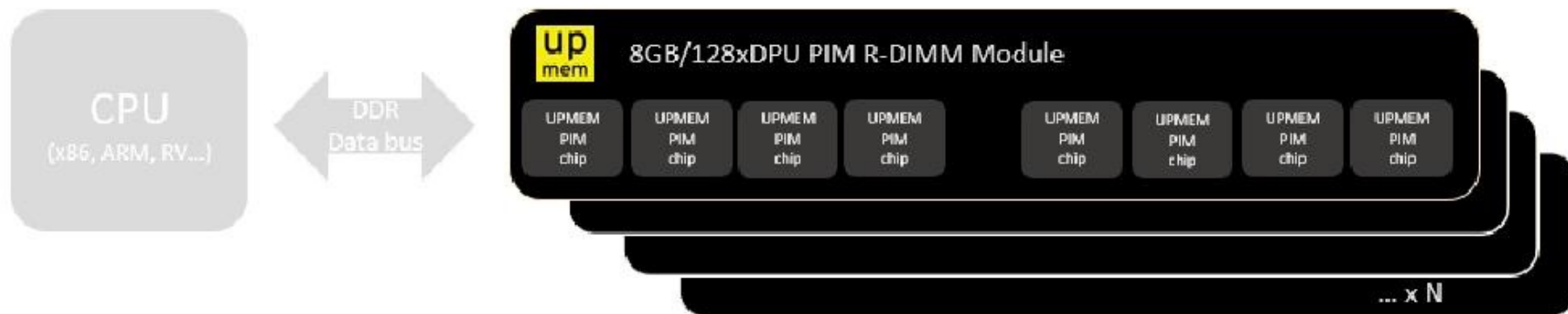
UPMEM Memory Modules

- E19: 8 chips DIMM (1 rank). DPUs @ 267 MHz
- P21: 16 chips DIMM (2 ranks). DPUs @ 350 MHz

SAFARI
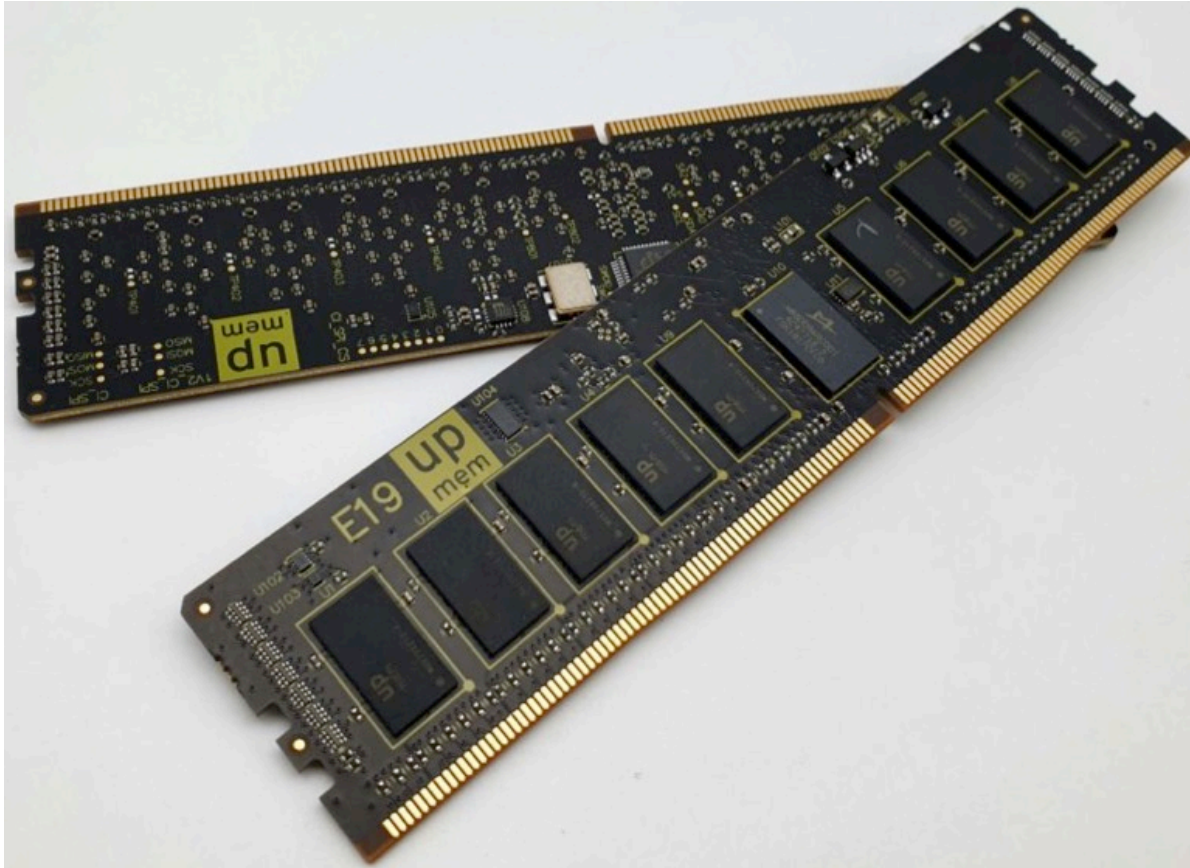
2,560-DPU Processing-in-Memory System

Benchmarking a New Paradigm: An Experimental Analysis of a Real Processing-in-Memory Architecture

JUAN GÓMEZ-LUNA, ETH Zürich, Switzerland
IZZAT EL HAJJ, American University of Beirut, Lebanon
IVAN FERNANDEZ, ETH Zürich, Switzerland and University of Malaga, Spain
CHRISTINA GIANNOULA, ETH Zürich, Switzerland and NTUA, Greece
GERALDO F. OLIVEIRA, ETH Zürich, Switzerland
ONUR MUTLU, ETH Zürich, Switzerland

Many modern workloads, such as neural networks, databases, and graph processing, are fundamentally memory-bound. For such workloads, the data movement between main memory and CPU cores imposes a significant overhead in terms of both latency and energy. A major reason is that this communication happens through a narrow bus with high latency and limited bandwidth, and the low data reuse in memory-bound workloads is insufficient to amortize the cost of main memory access. Fundamentally addressing this *data movement bottleneck* requires a paradigm where the memory system assumes an active role in computing by integrating processing capabilities. This paradigm is known as *processing-in-memory* (*PIM*).

Recent research explores different forms of PIM architectures, motivated by the emergence of new 3D-stacked memory technologies that integrate memory with a logic layer where processing elements can be easily placed. Past works evaluate these architectures in simulation or, at best, with simplified hardware prototypes. In contrast, the UPMEM company has designed and manufactured the first publicly-available real-world PIM architecture. The UPMEM PIM architecture combines traditional DRAM memory arrays with general-purpose in-order cores, called *DRAM Processing Units* (*DPUs*), integrated in the same chip.

This paper provides the first comprehensive analysis of the first publicly-available real-world PIM architecture. We make two key contributions. First, we conduct an experimental characterization of the UPMEM-based PIM system using microbenchmarks to assess various architecture limits such as compute throughput and memory bandwidth, yielding new insights. Second, we present *PrIM* (*Processing-In-Memory benchmarks*), a benchmark suite of 16 workloads from different application domains (e.g., dense/sparse linear algebra, databases, data analytics, graph processing, neural networks, bioinformatics, image processing), which we identify as memory-bound. We evaluate the performance and scaling characteristics of PrIM benchmarks on the UPMEM PIM architecture, and compare their performance and energy consumption to their state-of-the-art CPU and GPU counterparts. Our extensive evaluation conducted on two real UPMEM-based PIM systems with 640 and 2,556 DPUs provides new insights about suitability of different workloads to the PIM system, programming recommendations for software designers, and suggestions and hints for hardware and architecture designers of future PIM systems.

https://arxiv.org/pdf/2105.03814.pdf

Experimental Analysis of the UPMEM PIM Engine

Benchmarking a New Paradigm: An Experimental Analysis of a Real Processing-in-Memory Architecture

JUAN GÓMEZ-LUNA, ETH Zürich, Switzerland
IZZAT EL HAJJ, American University of Beirut, Lebanon
IVAN FERNANDEZ, ETH Zürich, Switzerland and University of Malaga, Spain
CHRISTINA GIANNOULA, ETH Zürich, Switzerland and NTUA, Greece
GERALDO F. OLIVEIRA, ETH Zürich, Switzerland
ONUR MUTLU, ETH Zürich, Switzerland

Many modern workloads, such as neural networks, databases, and graph processing, are fundamentally memory-bound. For such workloads, the data movement between main memory and CPU cores imposes a significant overhead in terms of both latency and energy. A major reason is that this communication happens through a narrow bus with high latency and limited bandwidth, and the low data reuse in memory-bound workloads is insufficient to amortize the cost of main memory access. Fundamentally addressing this *data movement bottleneck* requires a paradigm where the memory system assumes an active role in computing by integrating processing capabilities. This paradigm is known as *processing-in-memory* (*PIM*).

Recent research explores different forms of PIM architectures, motivated by the emergence of new 3D-stacked memory technologies that integrate memory with a logic layer where processing elements can be easily placed. Past works evaluate these architectures in simulation or, at best, with simplified hardware prototypes. In contrast, the UPMEM company has designed and manufactured the first publicly-available real-world PIM architecture. The UPMEM PIM architecture combines traditional DRAM memory arrays with general-purpose in-order cores, called *DRAM Processing Units* (*DPUs*), integrated in the same chip.

This paper provides the first comprehensive analysis of the first publicly-available real-world PIM architecture. We make two key contributions. First, we conduct an experimental characterization of the UPMEM-based PIM system using microbenchmarks to assess various architecture limits such as compute throughput and memory bandwidth, yielding new insights. Second, we present *PrIM* (*Processing-In-Memory benchmarks*), a benchmark suite of 16 workloads from different application domains (e.g., dense/sparse linear algebra, databases, data analytics, graph processing, neural networks, bioinformatics, image processing), which we identify as memory-bound. We evaluate the performance and scaling characteristics of PrIM benchmarks on the UPMEM PIM architecture, and compare their performance and energy consumption to their state-of-the-art CPU and GPU counterparts. Our extensive evaluation conducted on two real UPMEM-based PIM systems with 640 and 2,556 DPUs provides new insights about suitability of different workloads to the PIM system, programming recommendations for software designers, and suggestions and hints for hardware and architecture designers of future PIM systems.

https://arxiv.org/pdf/2105.03814.pdf

Understanding a Modern PIM Architecture

SAFARI Live Seminar: Understanding a Modern Processing-in-Memory Architecture

2,579 views • Streamed live on Jul 12, 2021
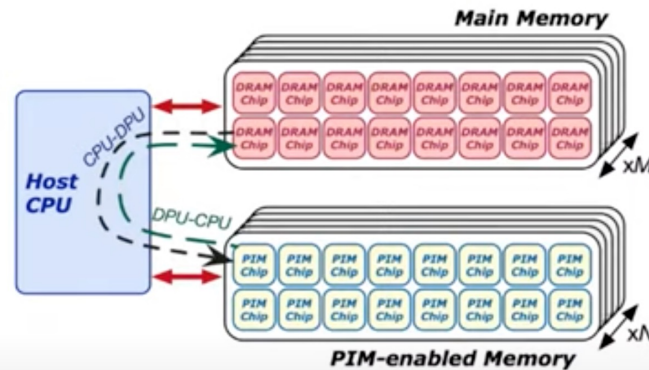
👍 93   👎 0   ↗ SHARE   ≡+ SAVE   ...

Onur Mutlu Lectures
18.7K subscribers

SUBSCRIBED

More on Analysis of the UPMEM PIM Engine

SAFARI Live Seminar: Understanding a Modern Processing-in-Memory Architecture

1,868 views • Streamed live on Jul 12, 2021

Onur Mutlu Lectures
17.6K subscribers

Talk Title: Understanding a Modern Processing-in-Memory Architecture: Benchmarking and Experimental Characterization
Dr. Juan Gómez-Luna, SAFARI Research Group, D-ITET, ETH Zurich

https://www.youtube.com/watch?v=D8Hjy2iU9l4&list=PL5Q2soXY2Zi_tOTAYm--dYByNPL7JhwR9

More on Analysis of the UPMEM PIM Engine

Understanding a Modern Processing-in-Memory Arch: Benchmarking & Experimental Characterization; 21m

3,482 views • Premiered Jul 25, 2021

👍 38   👎 0   ➤ SHARE   ➕ SAVE   •••

Onur Mutlu Lectures
17.9K subscribers

ANALYTICS   EDIT VIDEO

https://www.youtube.com/watch?v=Pp9jSU2b9oM&list=PL5Q2soXY2Zi8_VVChACnON4sfh2bJ5IrD&index=159

FPGA-based Processing Near Memory

- Gagandeep Singh, Mohammed Alser, Damla Senol Cali, Dionysios Diamantopoulos, Juan Gómez-Luna, Henk Corporaal, and Onur Mutlu,
  "FPGA-based Near-Memory Acceleration of Modern Data-Intensive Applications"
  IEEE Micro (**IEEE MICRO**), 2021.

FPGA-based Near-Memory Acceleration of Modern Data-Intensive Applications

Gagandeep Singh◇   **Mohammed Alser**◇   **Damla Senol Cali**⋈

Dionysios Diamantopoulos▽   **Juan Gómez-Luna**◇

Henk Corporaal⋆   **Onur Mutlu**◇⋈

◇*ETH Zürich*   ⋈*Carnegie Mellon University*
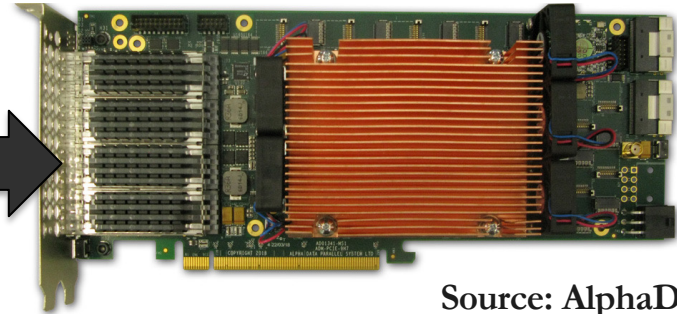⋆*Eindhoven University of Technology*   ▽*IBM Research Europe*

Near-Memory Acceleration using FPGAs

Source: IBM

Source: AlphaData

IBM POWER9 CPU   **HBM-based FPGA board**

OCAPI

Near-HBM FPGA-based accelerator

Two communication technologies: **CAPI2** and **OCAPI**
Two memory technologies: **DDR4** and **HBM**
Two workloads: **Weather Modeling** and **Genome Analysis**

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 →
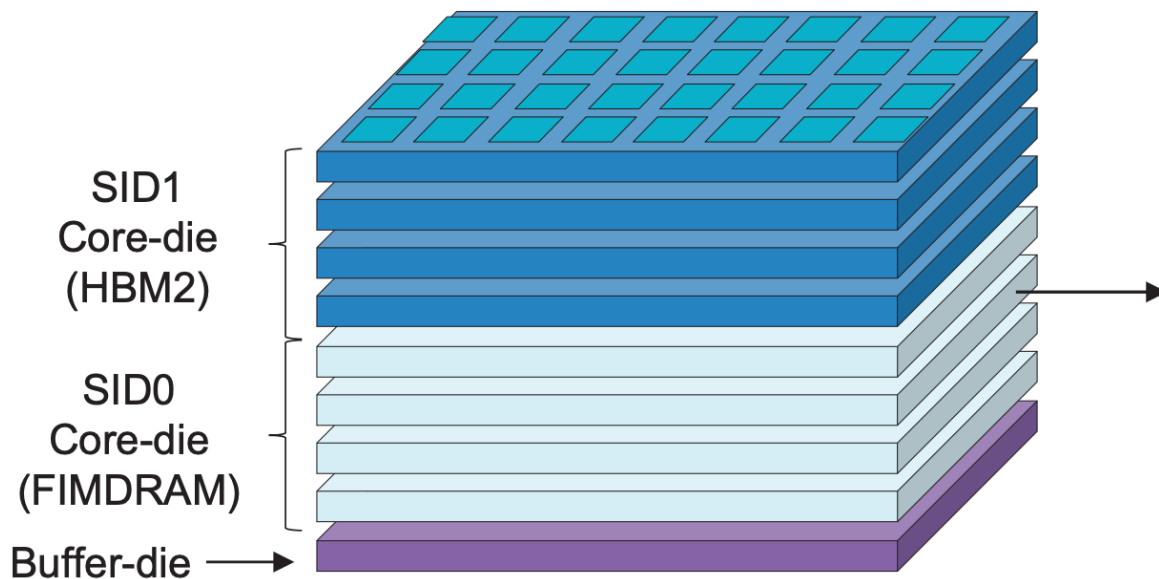
[3D Chip Structure of HBM with FIMDRAM]

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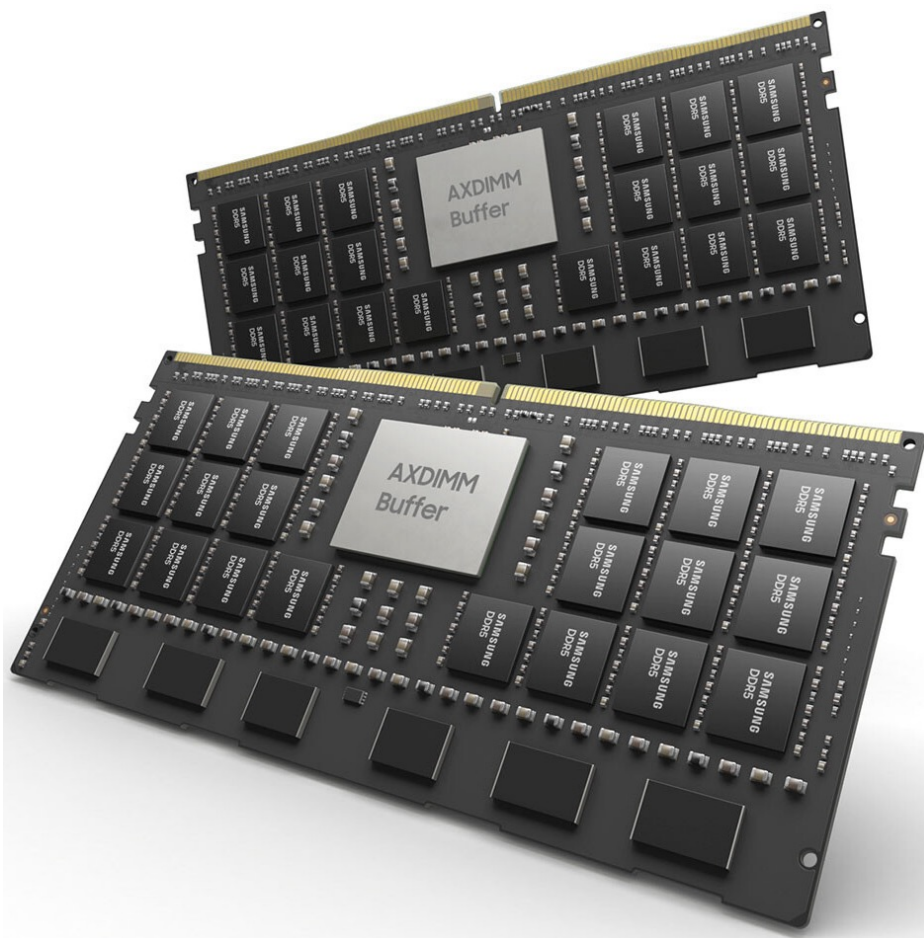
[1]Samsung Electronics, Hwaseong, Korea
[2]Samsung Electronics, San Jose, CA
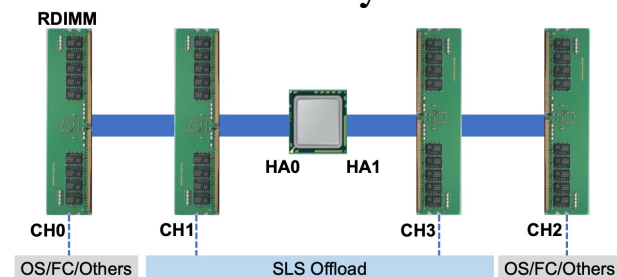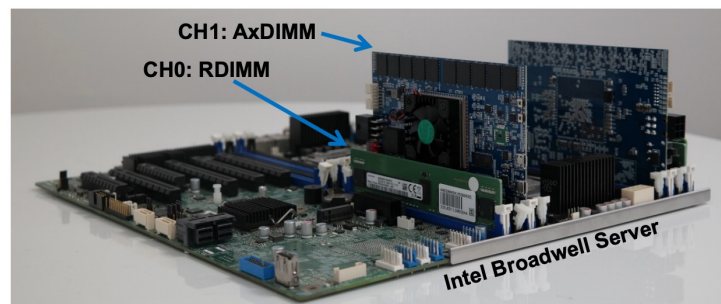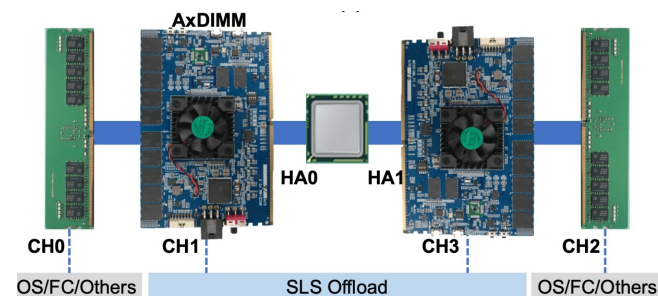[3]Samsung Electronics, Suwon, Korea

Samsung AxDIMM (2021)

- ## DDRx-PIM
  - DLRM recommendation system

Baseline System

AxDIMM System

CH1: AxDIMM
CH0: RDIMM

Intel Broadwell Server

Ke et al. "Near-Memory Processing in Action: Accelerating Personalized Recommendation with AxDIMM", IEEE Micro (2021)

SK Hynix Accelerator-in-Memory (2022)

INSIGHT   SK hynix STORY   PRESS CENTER   MULTIMEDIA

SK hynix Develops PIM, Next-Generation AI Accelerator

February 16, 2022

Seoul, February 16, 2022

SK hynix (or "the Company", www.skhynix.com) announced on February 16 that it has developed PIM*, a next-generation memory chip with computing capabilities.

PIM(Processing In Memory): A next-generation technology that provides a solution for data congestion issues for AI and big data by adding computational functions to semiconductor memory

It has been generally accepted that memory chips store data and CPU or GPU, like human brain, process data. SK hynix, following its challenge to such notion and efforts to pursue innovation in the next-generation smart memory, has found a breakthrough solution with the development of the latest technology.

SK hynix plans to showcase its PIM development at the world's most prestigious semiconductor conference, 2022 ISSCC*, in San Francisco at the end of this month. The company expects continued efforts for innovation of this technology to bring the memory-centric computing, in which semiconductor memory plays a central role, a step closer to the reality in devices such as smartphones.

ISSCC: The International Solid-State Circuits Conference will be held virtually from Feb. 20 to Feb. 24 this year with a theme of "Intelligent Silicon for a Sustainable World"

For the first product that adopts the PIM technology, SK hynix has developed a sample of GDDR6-AiM (Accelerator* in memory). The GDDR6-AiM adds computational functions to GDDR6* memory chips, which process data at 16Gbps. A combination of GDDR6-AiM with CPU or GPU instead of a typical DRAM makes certain computation speed 16 times faster. GDDR6-AiM is widely expected to be adopted for machine learning, high-performance computing, and big data computation and storage.
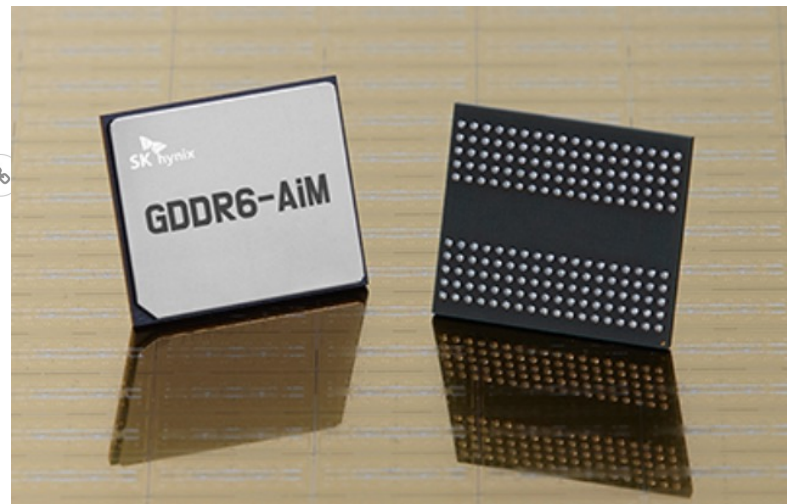
11.1 A 1ynm 1.25V 8Gb, 16Gb/s/pin GDDR6-based Accelerator-in-Memory supporting 1TFLOPS MAC Operation and Various Activation Functions for Deep-Learning Applications

Seongju Lee, SK hynix, Icheon, Korea

In Paper 11.1, SK Hynix describes an 1ynm, GDDR6-based accelerator-in-memory with a command set for deep-learning operation. The 8Gb design achieves a peak throughput of 1TFLOPS with 1GHz MAC operations and supports major activation functions to improve accuracy.

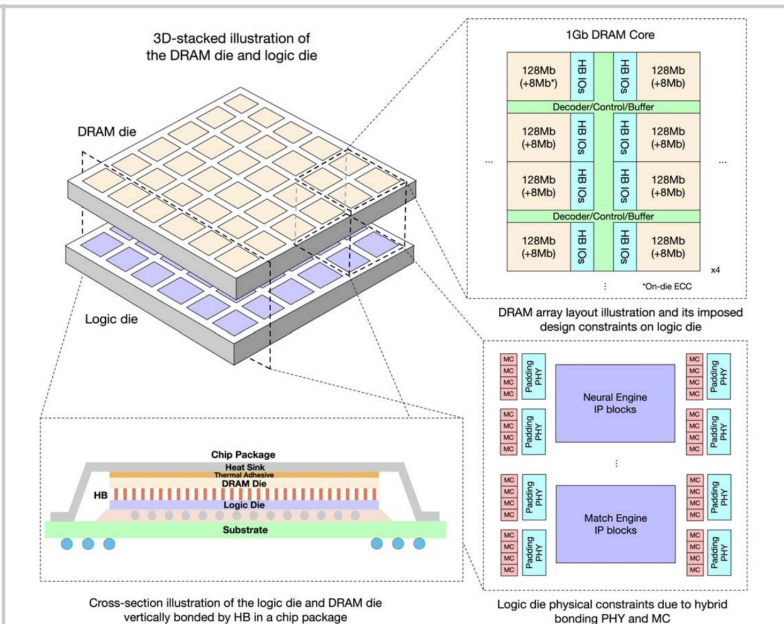# AliBaba PIM Recommendation System (2022)

Figure 29.1.2: Illustration of 3D-stacked chip, cross-illustration of package, DRAM array layout and design blocks on logic die.
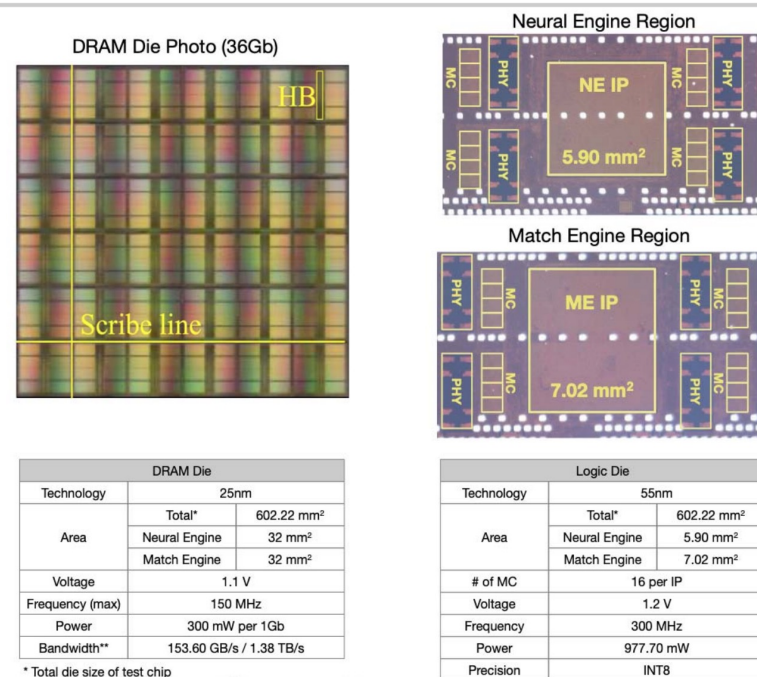

Figure 29.1.7: Die micrographs of DRAM die, NE and ME. Detailed specifications of DRAM die and logic die.

29.1   184QPS/W 64Mb/mm² 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System

Dimin Niu[1], Shuangchen Li[1], Yuhao Wang[1], Wei Han[1], Zhe Zhang[2], Yijin Guan[2], Tianchan Guan[3], Fei Sun[1], Fei Xue[1], Lide Duan[1], Yuanwei Fang[1], Hongzhong Zheng[1], Xiping Jiang[4], Song Wang[4], Fengguo Zuo[4], Yubing Wang[4], Bing Yu[4], Qiwei Ren[4], Yuan Xie[1]

74

Computational CXL-Memory Solution for Accelerating Memory-Intensive Applications

Joonseop Sim [ID], Soohong Ahn [ID], Taeyoung Ahn [ID],
Seungyong Lee [ID], Myunghyun Rhee, Jooyoung Kim [ID],
Kwangsik Shin, Donguk Moon [ID],
Euiseok Kim, and Kyoung Park [ID]

Abstract—CXL interface is the up-to-date technology that enables effective memory expansion by providing a memory-sharing protocol in configuring heterogeneous devices. However, its limited physical bandwidth can be a significant bottleneck for emerging data-intensive applications. In this work, we propose a novel CXL-based memory disaggregation architecture with a real-world prototype demonstration, which overcomes the bandwidth limitation of the CXL interface using near-data processing. The experimental results demonstrate that our design achieves up to 1.9× better performance/power efficiency than the existing CPU system.

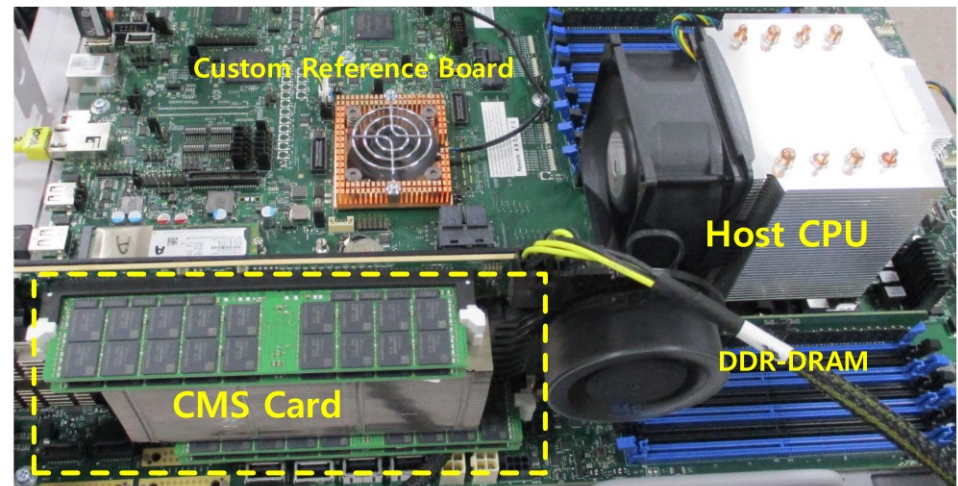**Index Terms**—Compute express link (CXL), near-data-processing (NDP)

Fig. 6. FPGA prototype of proposed CMS card.
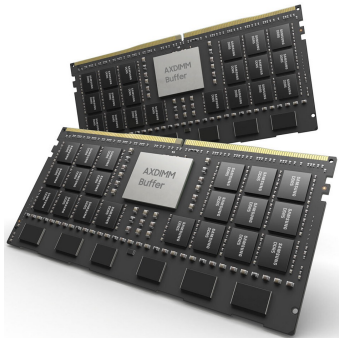
Samsung CXL Processing Near Memory (2023)

Samsung Processing in Memory Technology at Hot Chips 2023

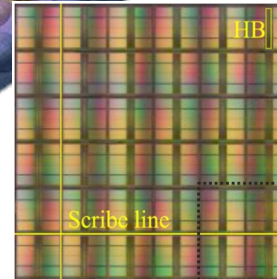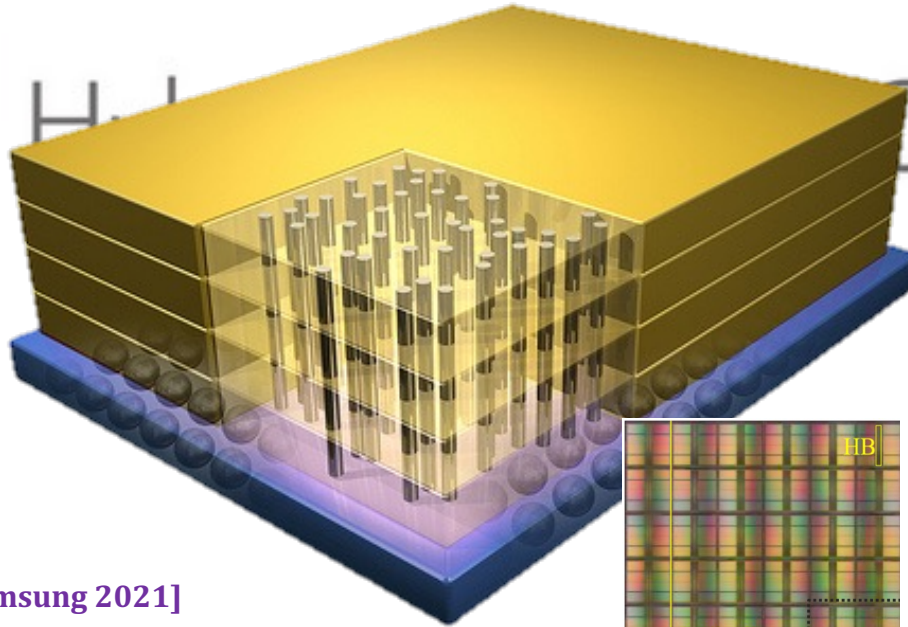By **Patrick Kennedy** - August 28, 2023

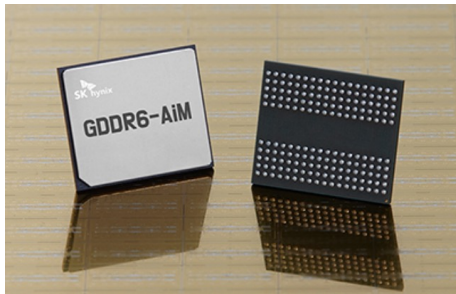Samsung PIM PNM For Transformer Based AI HC35_Page_24
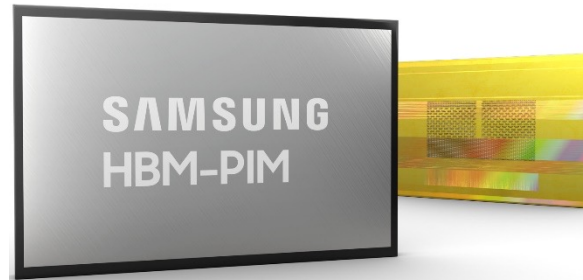
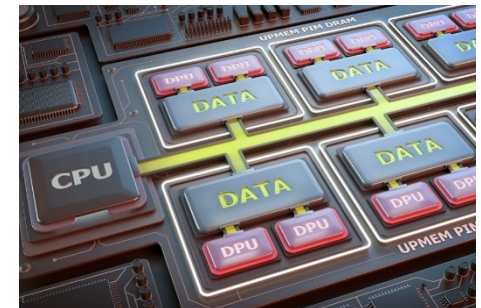# Processing-in-Memory Landscape (2022)

[Samsung 2021]

[Alibaba 2022]

[SK Hynix 2022]

[Samsung 2021]

[UPMEM 2019]

SAFARI

And, many other experimental chips and startups

Future of Genome Sequencing & Analysis

Mohammed Alser, Zülal Bingöl, Damla Senol Cali, Jeremie Kim, Saugata Ghose, Can Alkan, Onur Mutlu
"Accelerating Genome Analysis: A Primer on an Ongoing Journey" IEEE Micro, August 2020.

Accelerating Genome Analysis: A Primer on an Ongoing Journey
Sept.-Oct. 2020, pp. 65-75, vol. 40
DOI Bookmark: 10.1109/MM.2020.3013728

FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications
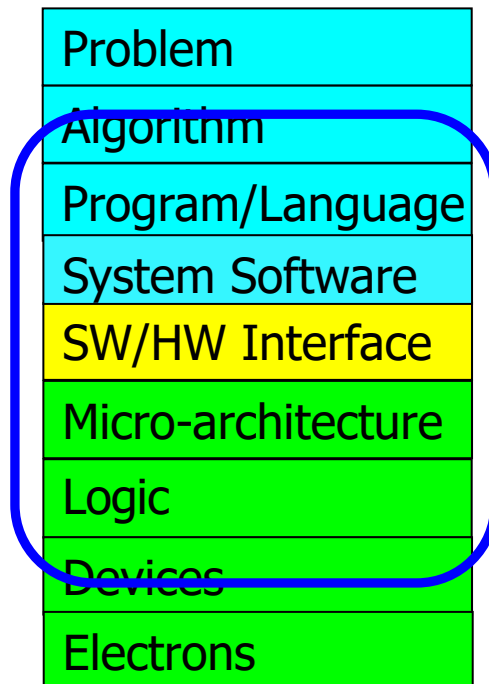July-Aug. 2021, pp. 39-48, vol. 41
DOI Bookmark: 10.1109/MM.2021.3088396

MinION from ONT

SmidgION from ONT

Axiom

To achieve the highest efficiency, performance, robustness:

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 Kind of a Future
Do We Want?

How Reliable/Secure/Safe is This Bridge?

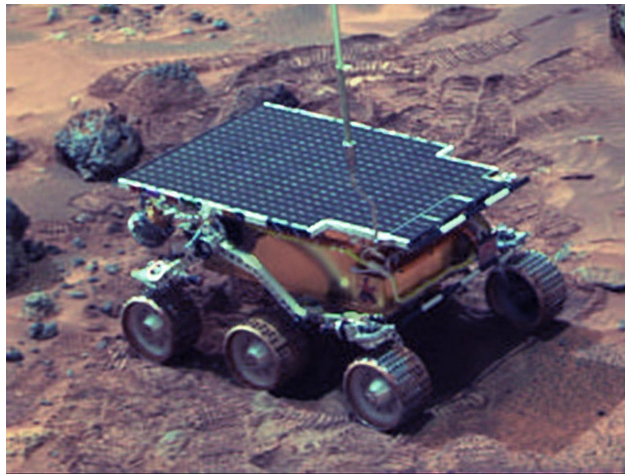# Collapse of the "Galloping Gertie"

Another View

How Secure Are These People?

Security is about preventing unforeseen consequences

How Safe & Secure Is **This** Platform?

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

How Robust Are **These** Platforms Really?

SAFARI

https://www.kennedyspacecenter.com/explore-attractions/nasa-now
https://www.cnet.com/pictures/nasas-wildest-rides-extreme-vehicles-for-earth-and-beyond/7/

Robust
(Reliable, Secure, Safe)

Do We Want This?

Or This?

SAFARI Source: V. Milutinovic

Sustainable

and

Energy Efficient

Many Difficult Problems: Climate

SAFARI   Source: https://farm9.staticflickr.com/8571/16376102935_8628150df8_o.jpg
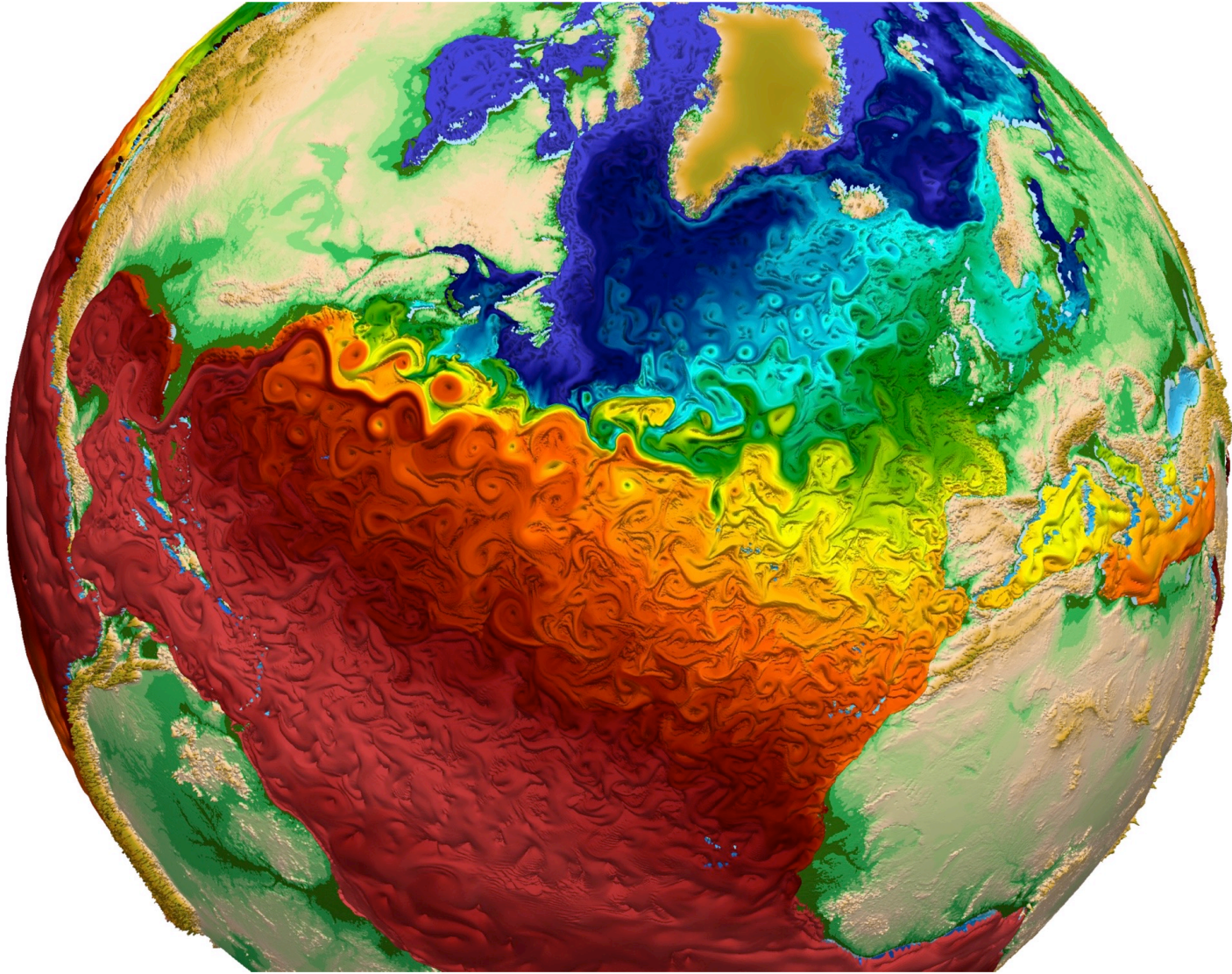
Many Difficult Problems: Congestion

Source: https://blogs-images.forbes.com/jimgorzelany/files/2015/10/China-G4-backup-this-oct-reuters.jpg

Many Difficult Problems: Intelligence

SAFARI

Many Difficult Problems: Public Health

Source: https://blog.wego.com/7-crowded-places-and-events-that-you-will-love/

Many Difficult Problems: Genome Analysis

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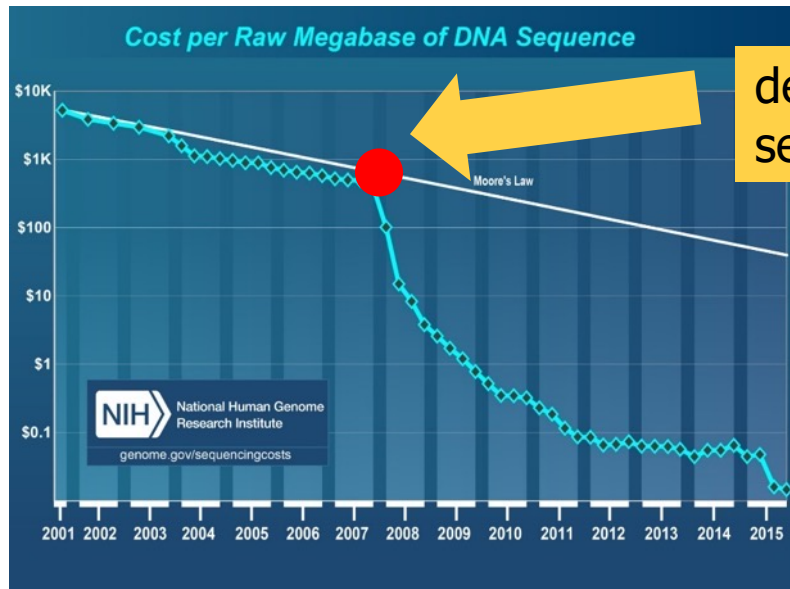
development of high-throughput sequencing (HTS) technologies

Number of Genomes Sequenced

229,000 — 2014
422,000 — 2015
952,000 — 2016
1,620,000 — 2017

Source: Illumina
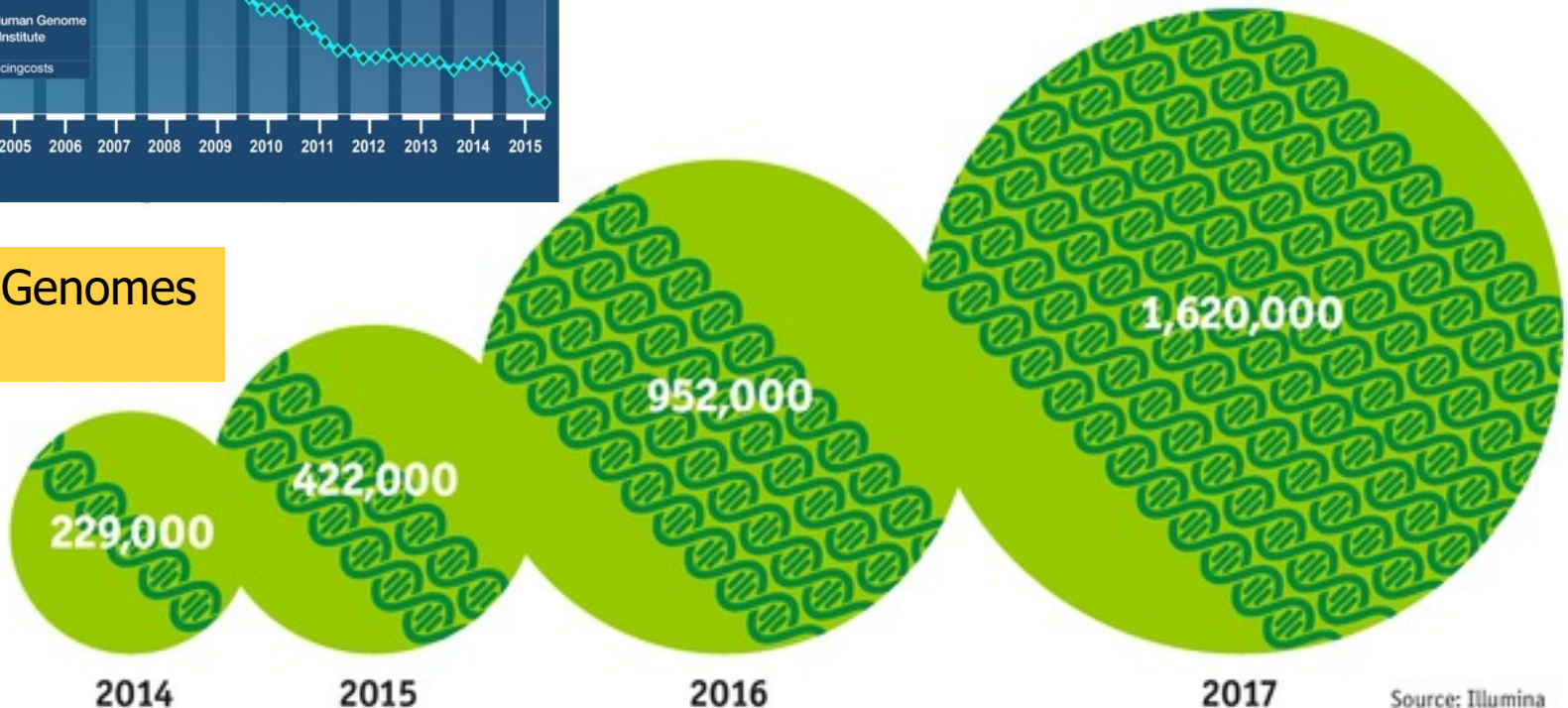
We Need Faster & Scalable Genome Analysis
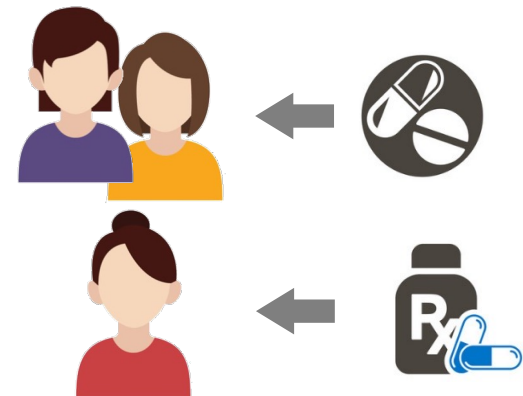

Understanding **genetic variations, species, evolution, …**

Predicting the presence and relative abundance of **microbes** in a sample

Rapid surveillance of **disease outbreaks**

Developing **personalized medicine**

And, many, many other applications …

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

Senol Cali+, "**Nanopore Sequencing Technology and Tools for Genome Assembly: Computational Analysis of the Current State, Bottlenecks and Future Directions**," Briefings in Bioinformatics, 2018.
[Open arxiv.org version]

Accelerating Genome Analysis [DAC 2023]

- Onur Mutlu and Can Firtina,
  "Accelerating Genome Analysis via Algorithm-Architecture Co-Design"
  Invited Special Session Paper in Proceedings of the 60th Design Automation Conference (**DAC**), San Francisco, CA, USA, July 2023.
  [arXiv version]

Accelerating Genome Analysis via Algorithm-Architecture Co-Design

Onur Mutlu   Can Firtina
ETH Zürich

Accelerating Genome Analysis [IEEE MICRO 2020]

- 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

Beginner Reading on Genome Analysis

Mohammed Alser, Joel Lindegger, Can Firtina, Nour Almadhoun, Haiyu Mao, Gagandeep Singh, Juan Gomez-Luna, Onur Mutlu
"[From Molecules to Genomic Variations to Scientific Discovery: Intelligent Algorithms and Architectures for Intelligent Genome Analysis](#)"
Computational and Structural Biotechnology Journal, 2022
[[Source code](#)]

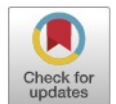Review

From molecules to genomic variations: Accelerating genome analysis via intelligent algorithms and architectures

Mohammed Alser *, Joel Lindegger, Can Firtina, Nour Almadhoun, Haiyu Mao, Gagandeep Singh, Juan Gomez-Luna, Onur Mutlu *

ETH Zurich, Gloriastrasse 35, 8092 Zürich, Switzerland

Future of Genome Sequencing & Analysis

Mohammed Alser, Zülal Bingöl, Damla Senol Cali, Jeremie Kim, Saugata Ghose, Can Alkan, Onur Mutlu
"Accelerating Genome Analysis: A Primer on an Ongoing Journey" IEEE Micro, August 2020.

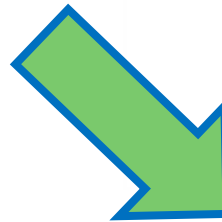**Accelerating Genome Analysis: A Primer on an Ongoing Journey**
Sept.-Oct. 2020, pp. 65-75, vol. 40
DOI Bookmark: 10.1109/MM.2020.3013728

FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications
July-Aug. 2021, pp. 39-48, vol. 41
DOI Bookmark: 10.1109/MM.2021.3088396
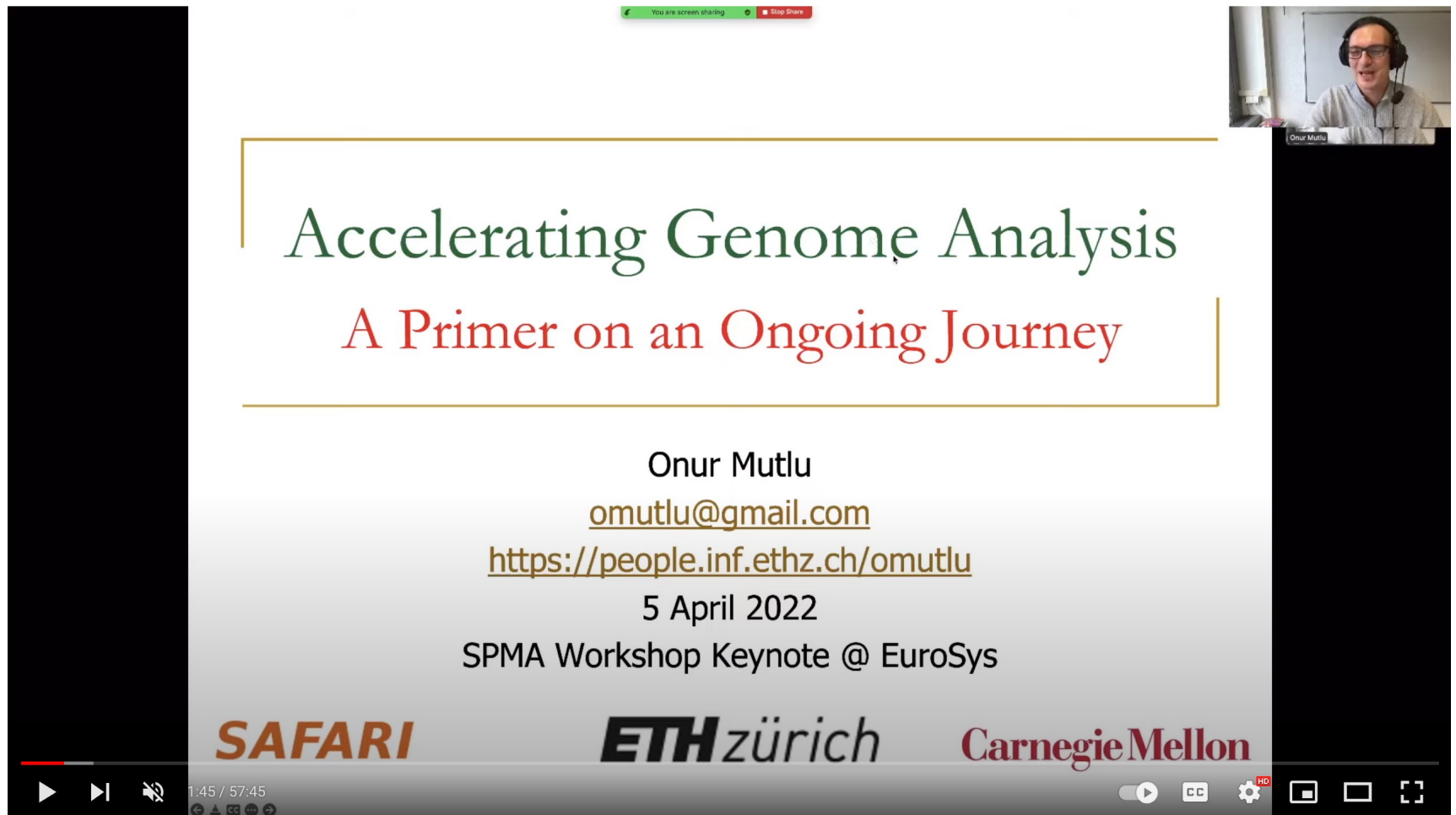
MinION from ONT

SmidgION from ONT

More on Fast & Efficient Genome Analysis …

https://www.youtube.com/watch?v=NCagwf0ivT0

Genomics Course (Fall 2022)

- **Fall 2022 Edition:**
  - https://safari.ethz.ch/projects_and_seminars/fall2022/doku.php?id=bioinformatics

- **Spring 2022 Edition:**
  - https://safari.ethz.ch/projects_and_seminars/spring2022/doku.php?id=bioinformatics

- **Youtube Livestream (Fall 2022):**
  - https://www.youtube.com/watch?v=nA41964-9r8&list=PL5Q2soXY2Zi8tFlQvdxOdizD_EhVAMVQV
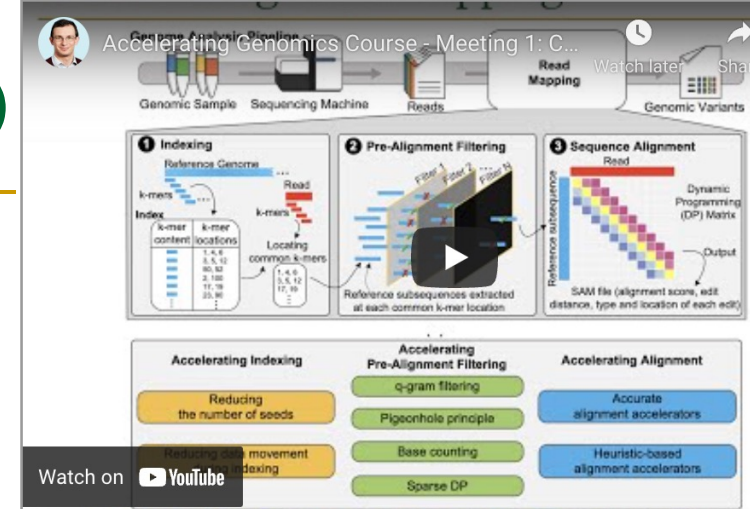
- **Youtube Livestream (Spring 2022):**
  - https://www.youtube.com/watch?v=DEL_5A_Y3TI&list=PL5Q2soXY2Zi8NrPDgOR1yRU_Cxxjw-u18

- Project course
  - Taken by Bachelor's/Master's students
  - Genomics lectures
  - Hands-on research exploration
  - Many research readings

https://www.youtube.com/onurmutlulectures

Spring 2022 Meetings/Schedule

Week	Date	Livestream	Meeting	Learning Materials
W1	11.3 Fri.	YouTube Live	M1: P&S Accelerating Genomics Course Introduction & Project Proposals (PDF) (PPT)	Required Materials Recommended Materials
W2	18.3 Fri.	YouTube Live	M2: Introduction to Sequencing (PDF) (PPT)	
W3	25.3 Fri.	YouTube Premiere	M3: Read Mapping (PDF) (PPT)	
W4	01.04 Fri.	YouTube Premiere	M4: GateKeeper (PDF) (PPT)	
W5	08.04 Fri.	YouTube Premiere	M5: MAGNET & Shouji (PDF) (PPT)	
W6	15.4 Fri.	YouTube Premiere	M6: SneakySnake (PDF) (PPT)	
W7	29.4 Fri.	YouTube Premiere	M7: GenStore (PDF) (PPT)	
W8	06.05 Fri.	YouTube Premiere	M8: GRIM-Filter (PDF) (PPT)	
W9	13.05 Fri.	YouTube Premiere	M9: Genome Assembly (PDF) (PPT)	
W10	20.05 Fri.	YouTube Live	M10: Genomic Data Sharing Under Differential Privacy (PDF) (PPT)	
W11	10.06 Fri.	YouTube Premiere	M11: Accelerating Genome Sequence Analysis (PDF) (PPT)	

BIO-Arch Workshop at RECOMB 2023

- **April 14, 2023**

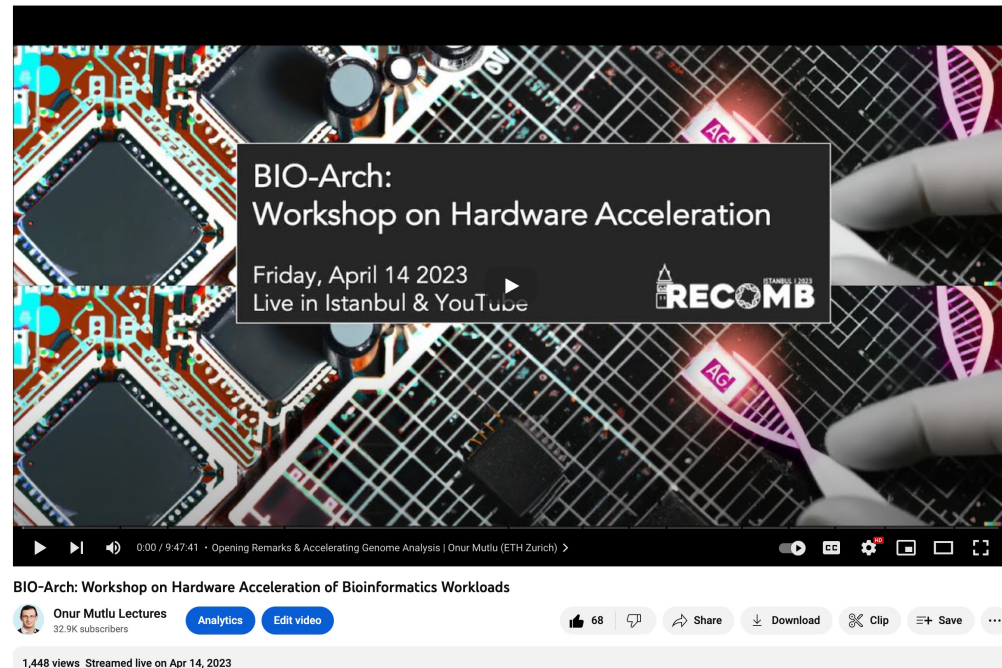## BIO-Arch: Workshop on Hardware Acceleration of Bioinformatics Workloads

About

BIO-Arch is a new forum for presenting and discussing new ideas in accelerating bioinformatics workloads with the co-design of hardware & software and the use of new computer architectures. Our goal is to discuss new system designs tailored for bioinformatics. BIO-Arch aims to bring together researchers in the bioinformatics, computational biology, and computer architecture communities to strengthen the progress in accelerating bioinformatics analysis (e.g., genome analysis) with efficient system designs that include hardware acceleration and software systems tailored for new hardware technologies.

Venue

BIO-Arch will be held in The Social Facilities of İstanbul Technical University on **April 14**. Detailed information about how to arrive at the venue location with various transportation options can be found on the RECOMB website.

Our panel discussion will be held in conjunction with the main RECOMB conference. The panel discussion will be held in Marriott Şişli on **April 17 at 17:00**. You can find

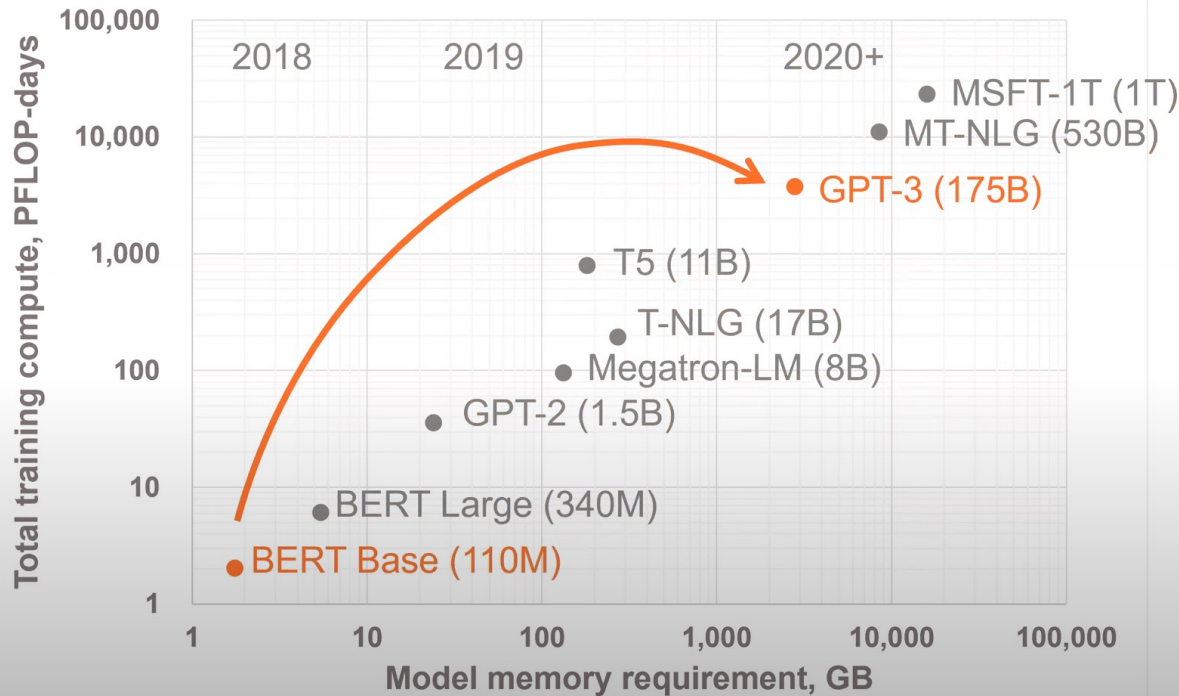BIO-Arch: Workshop on Hardware Acceleration of Bioinformatics Workloads

Onur Mutlu Lectures
32.9K subscribers

1,448 views Streamed live on Apr 14, 2023

https://www.youtube.com/watch?v=2rCsb4-nLmg

https://safari.ethz.ch/recomb23-arch-workshop/

104

Huge Demand for Performance & Efficiency

Exponential Growth of Neural Networks

Memory and compute requirements

1800x more compute

In just **2 years**
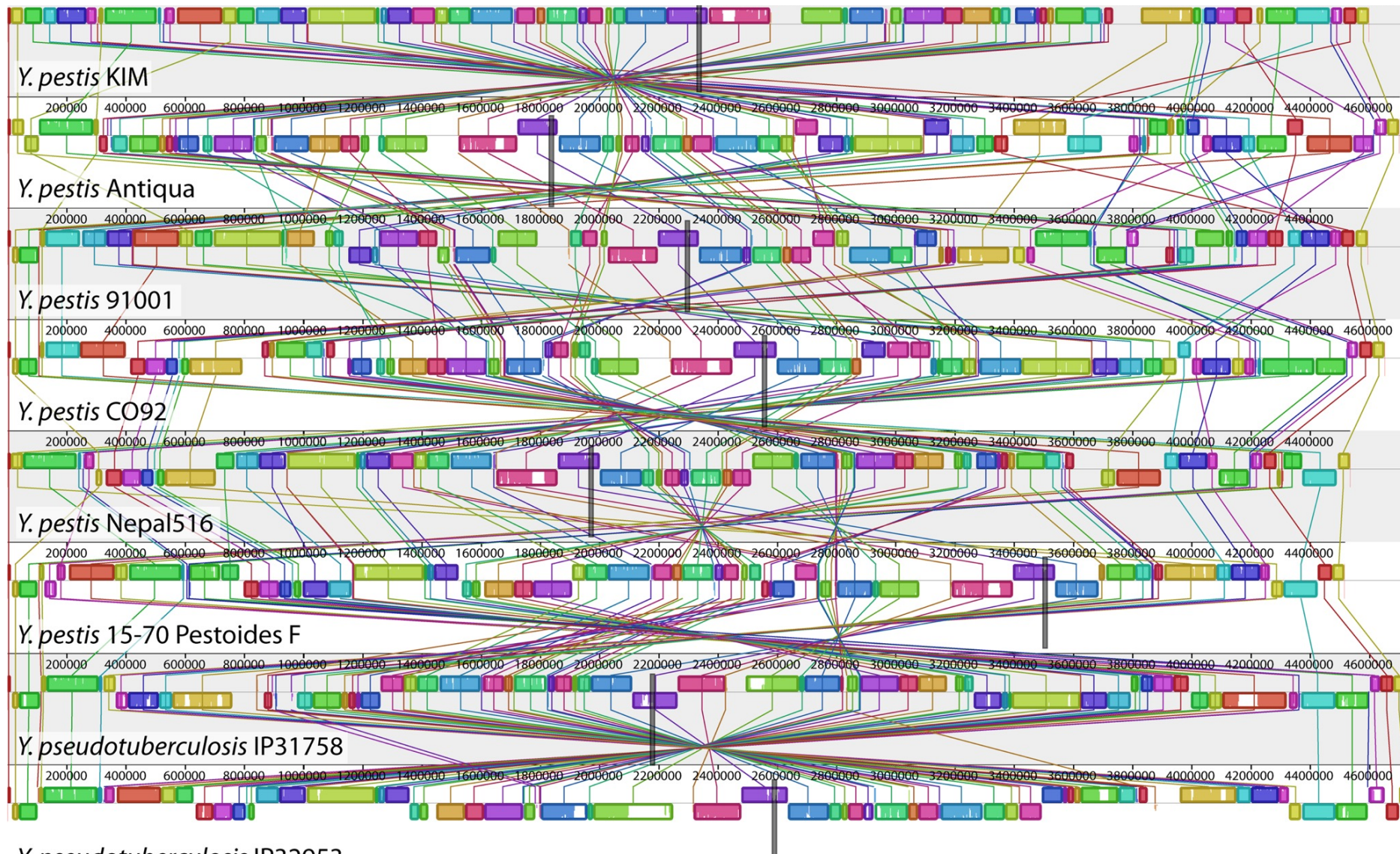
Tomorrow, **multi-trillion** parameter models

Source: https://youtu.be/Bh13Idwcb0Q?t=283

High Performance

(to solve

the **toughest** & **all** problems)

Personalization: Medicine

Comparative Genomics & Medicine

108

Personalized Medical 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

Senol Cali+, "**Nanopore Sequencing Technology and Tools for Genome Assembly: Computational Analysis of the Current State, Bottlenecks and Future Directions**," Briefings in Bioinformatics, 2018.
[Preliminary arxiv.org version]

Personalized Robotics

SAFARI

Source: https://www.asme.org/topics-resources/content/five-personal-robots-coming-home-soon

Challenge and Opportunity for Future

Personalized and Private

(in every aspect of life:
health, medicine,
spaces, devices, robotics, ...)

What Limits Us
in Computing Today?

This Course is About …

- Questioning what limits us in designing the best computing architectures for the future

- Providing directions for fundamentally better designs

- Advocating principled approaches

Increasingly Demanding Applications

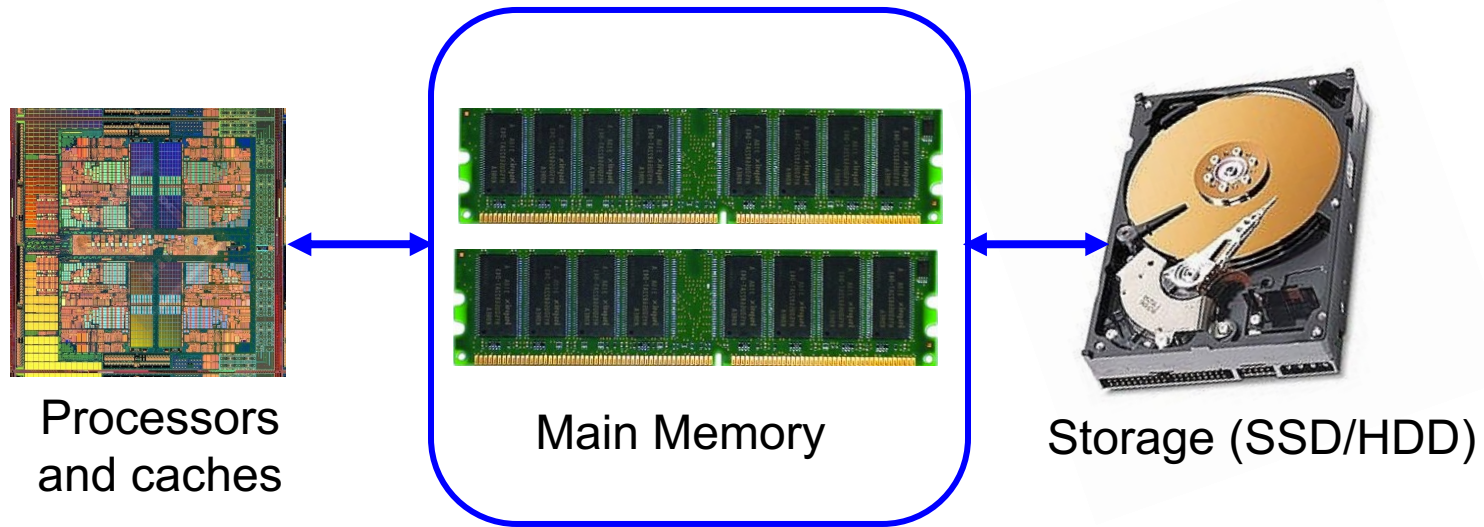# Dream...

and, they will come

As applications push boundaries, computing platforms become increasingly strained

SAFARI

Key Realization

Realization

Modern Systems are Bottlenecked by

Data Storage and Movement

Focus is on Data Storage Systems (Memory)

Processors and caches   Main Memory   Storage (SSD/HDD)

- **Main memory is a critical component of all computing systems**: server, mobile, embedded, desktop, sensor

- **Main memory system must scale** (in *size*, *technology*, *efficiency*, *cost*, and *management algorithms*) to maintain performance growth and technology scaling benefits

Focus is on Data Storage Systems (Memory)

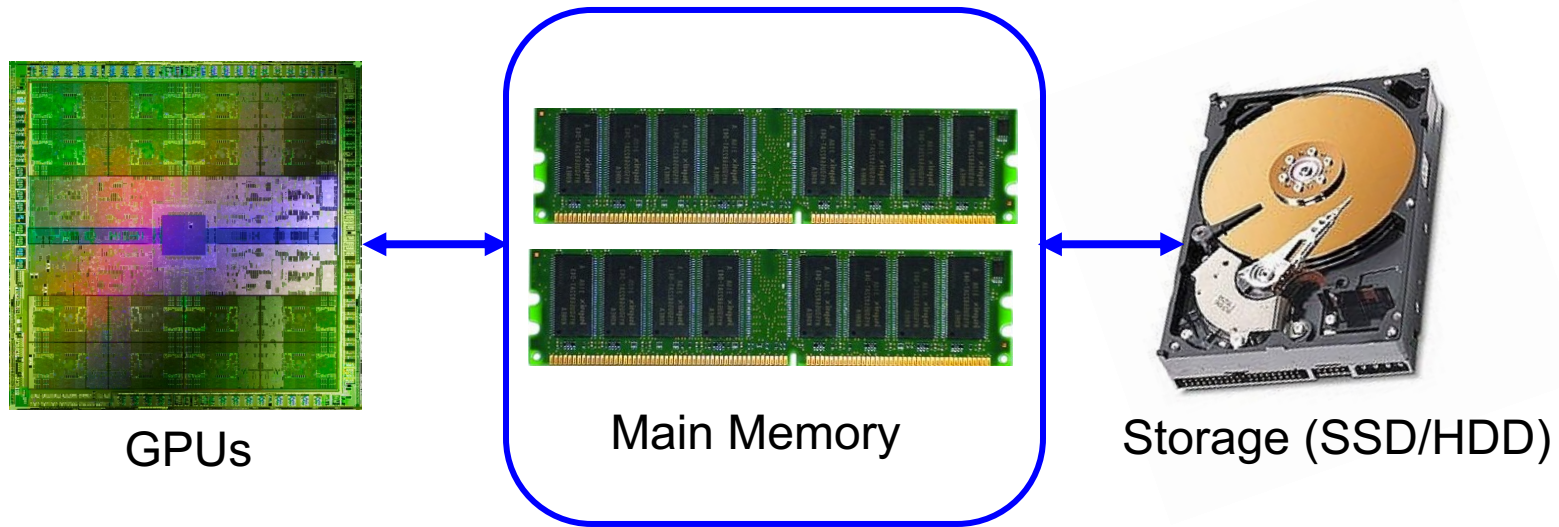FPGAs   Main Memory   Storage (SSD/HDD)

- **Main memory is a critical component of all computing systems**: server, mobile, embedded, desktop, sensor

- **Main memory system must scale** (in *size*, *technology*, *efficiency*, *cost*, and *management algorithms*) to maintain performance growth and technology scaling benefits

Focus is on Data Storage Systems (Memory)

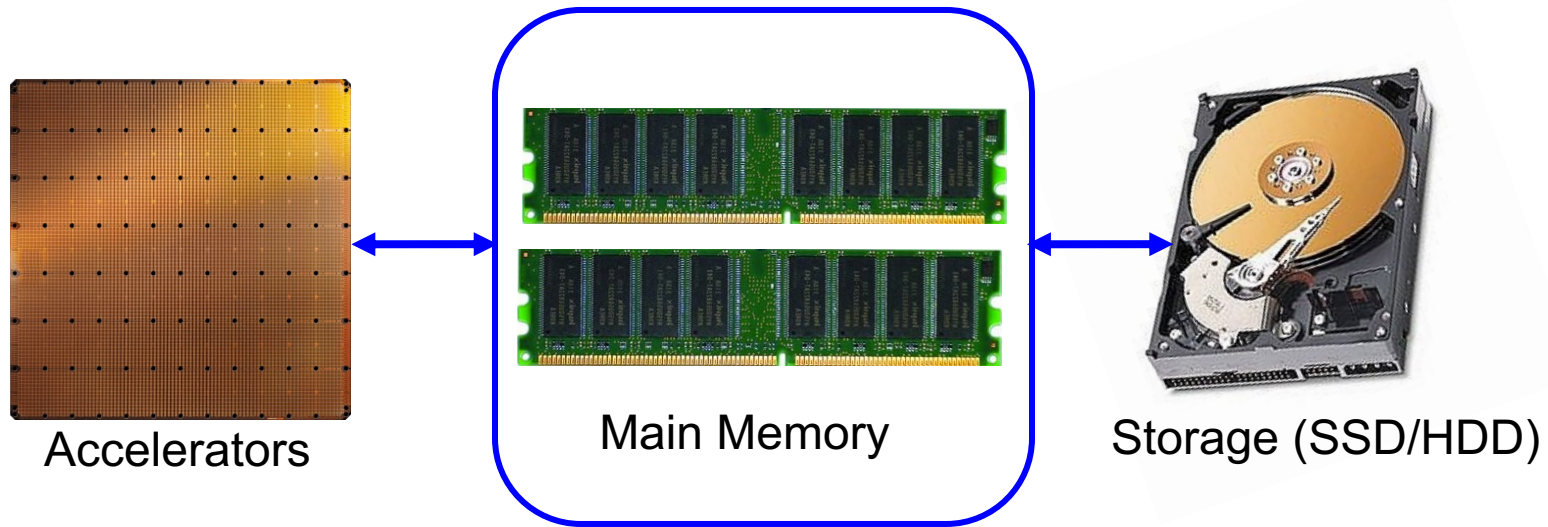GPUs   Main Memory   Storage (SSD/HDD)

- **Main memory is a critical component of all computing systems**: server, mobile, embedded, desktop, sensor

- **Main memory system must scale** (in *size*, *technology*, *efficiency*, *cost*, and *management algorithms*) to maintain performance growth and technology scaling benefits

Focus is on Data Storage Systems (Memory)

Accelerators   Main Memory   Storage (SSD/HDD)

- **Main memory is a critical component of all computing systems**: server, mobile, embedded, desktop, sensor

- **Main memory system must scale** (in *size*, *technology*, *efficiency*, *cost*, and *management algorithms*) to maintain performance growth and technology scaling benefits
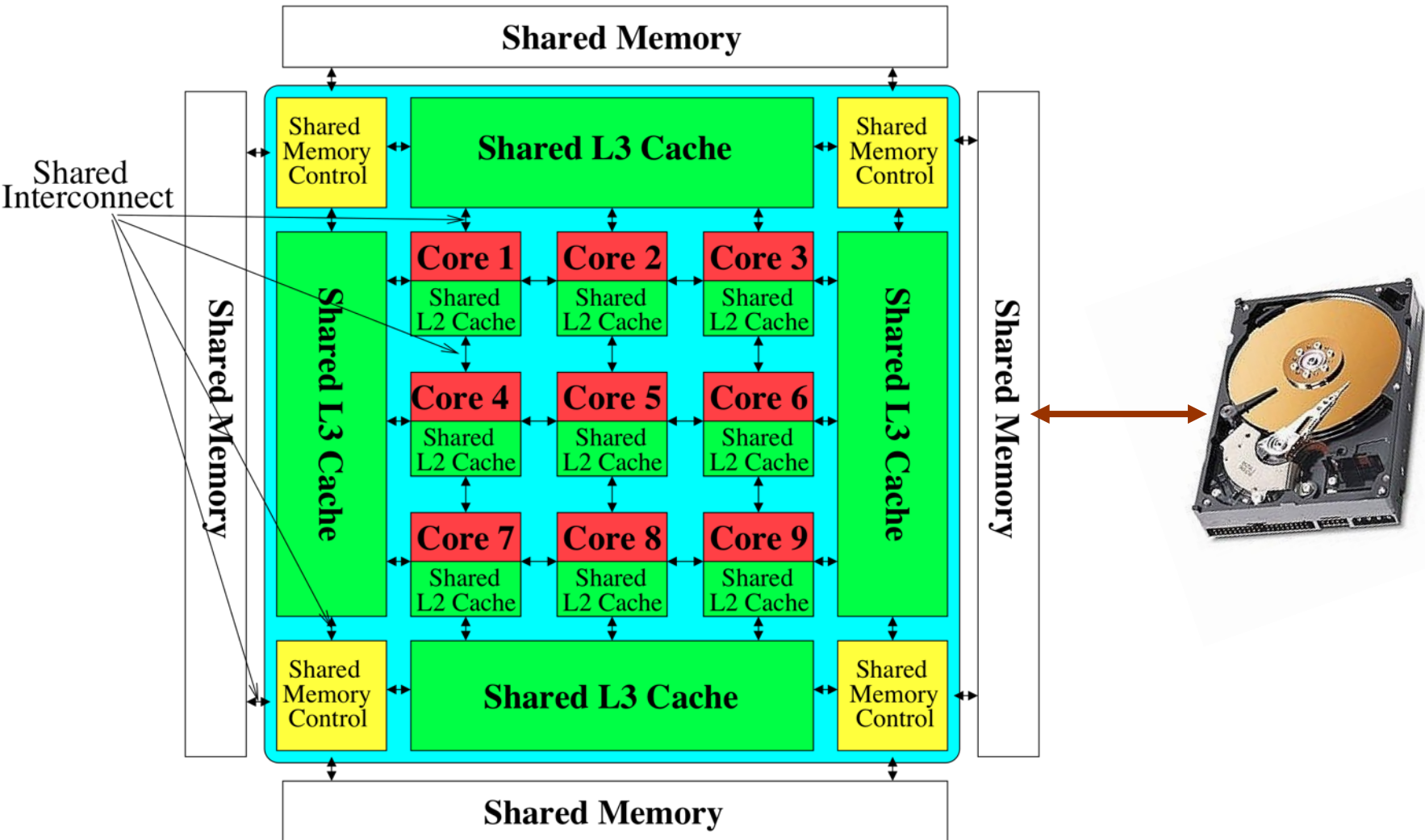
Memory & Storage

Why Is Memory So Important? (Especially Today)

Importance of Main Memory

- Performance Perspective

- Energy Perspective

- Scaling & Robustness (Reliability/Security/Safety) Perspective

- Trends/Challenges/Opportunities in Main Memory

Perils of Processor-Centric Design

Most of the system is dedicated to storing and moving data

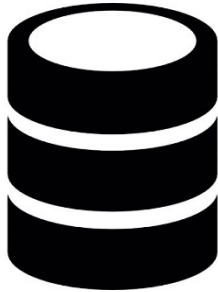**Yet, system is still bottlenecked by memory**

The Problem

Computing

is Bottlenecked by Data

Data is Key for AI, ML, Genomics, …

- Important workloads are all data intensive

- They require rapid and efficient processing of large amounts of data

- Data is increasing
  - We can generate more than we can process
  - We need to perform more sophisticated analyses on more data

Memory Is Critical for Performance (I)

In-memory Databases

[Mao+, EuroSys'12;
 Clapp+ (**Intel**), IISWC'15]
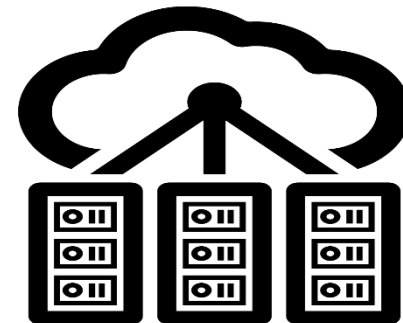
Graph/Tree Processing

[Xu+, IISWC'12; Umuroglu+, FPL'15]
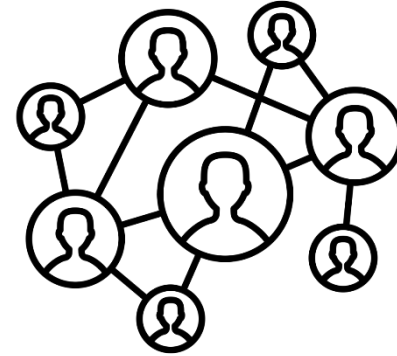
In-Memory Data Analytics

[Clapp+ (**Intel**), IISWC'15;
 Awan+, BDCloud'15]

Datacenter Workloads

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

Memory Is Critical for Performance (I)

In-memory Databases
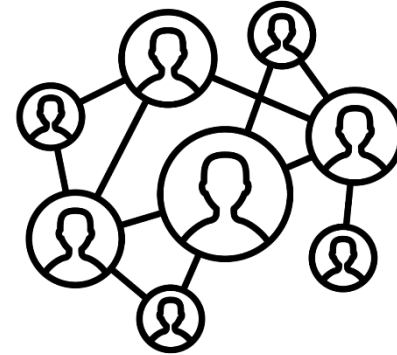
Graph/Tree Processing

Memory → bottleneck

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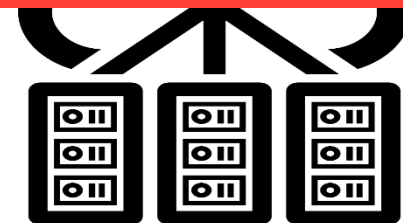
Datacenter Workloads
[Kanev+ (**Google**), ISCA'15]

Memory Is Critical for Performance (II)

Chrome

Google's web browser

TensorFlow Mobile

Google's machine learning framework

Video Playback

Google's **video codec**

Video Capture

Google's **video codec**

Memory Is Critical for Performance (II)

Chrome

TensorFlow Mobile

Memory → bottleneck

Video Playback

Google's **video codec**

Video Capture

Google's **video codec**

Genome Analysis

1 Sequencing

Billions of Short Reads
TATATATACGTACTAGTACGT
TTTAGTACGTACGT
ATACGTACTAGTACGT
ACG CCCCTACGTA
ACGTACTAGTACGT
TTAGTACGTACGT
TACGTACTAAAGTACGT
TACGTACTAGTACGT
TTTAAAACGTA
CGTACTAGTACGT
GGGAGTACGTACGT

2 Read Mapping

Short Read

Read Alignment

CCTATAATACG

Reference Genome

3 Variant Calling

reference: TTTATCGCTTCCATGACGCAG
read1:   ATCGCATCC
read2:   TATCGCATC
read3:   CATCCATGA
read4:   CGCTTCCAT
read5:   CCATGACGC
read6:   TTCCATGAC

4 Scientific Discovery

PRESCRIPTION
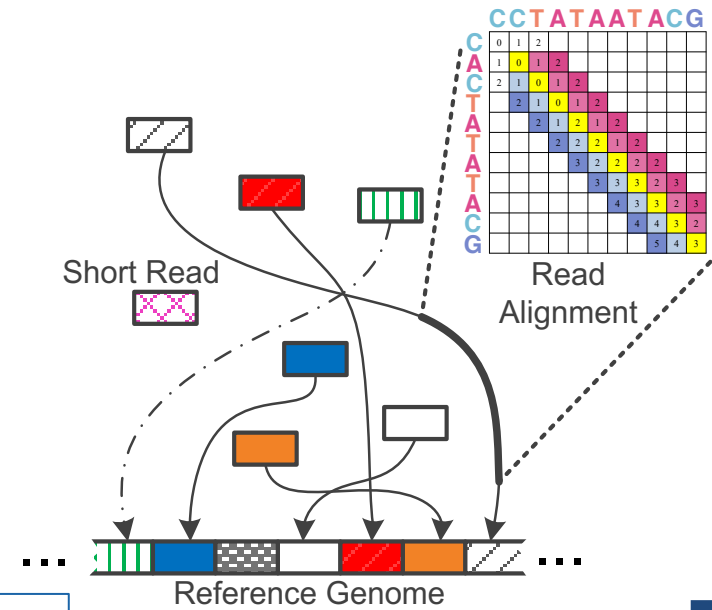
Billions of Short Reads
TATATATACGTACTAGTACGT
TTTAGTACGTACGT
ATACGTACTAGTACGT
ACG CCCCTACGTA
ACGTACTAGTACGT
TTAGTACGTACGT
TACGTACTAAAGTACGT
TACGTACTAGTACGT
TTTAAAACGTA
CGTACTAGTACGT
GGGAGTACGTACGT

CCTATAATACG

Short Read

Read Alignment

Reference Genome

Memory → bottleneck

reference: TTTATCGCTTCCATGACGCAG
read1:   ATCGCATCC
read2:   TATCGCATC
read3:   CATCCATGA
read4:   CGCTTCCAT
read5:   CCATGACGC
read6:   TTCCATGAC

PRESCRI

1 | **Sequencing**

2 | **Read Mapping**

3 | **Variant Calling**

4 | **Scientific Discovery**

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

Senol Cali+, "**Nanopore Sequencing Technology and Tools for Genome Assembly: Computational Analysis of the Current State, Bottlenecks and Future Directions**," Briefings in Bioinformatics, 2018.
[Open arxiv.org version]

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

Memory → bottleneck

State of the Main Memory System

- Recent technology, architecture, and application trends
  - lead to new requirements
  - exacerbate old requirements

- DRAM and memory controllers, as we design them today, are (and will be) unable to satisfy all requirements

- Some emerging non-volatile memory technologies (e.g., PCM) enable new opportunities: memory+storage merging

- We need to rethink the main memory system
  - to fix DRAM issues and enable emerging technologies
  - to satisfy all requirements

Major Trends Affecting Main Memory (I)

- **Need for main memory capacity, bandwidth, QoS increasing**

- **Main memory energy/power is a key system design concern**

- **DRAM technology scaling is ending**

# Major Trends Affecting Main Memory (II)

- **Need for main memory capacity, bandwidth, QoS increasing**
  - ❑ **Data-intensive applications**: increasing demand/hunger for data
  - ❑ **Multi-core**: increasing number of cores/agents
  - ❑ **Consolidation**: cloud computing, GPUs, mobile, heterogeneity

- Main memory energy/power is a key system design concern

- DRAM technology scaling is ending

SAFARI

Consequence: The Memory Capacity Gap

Core count doubling ~ every 2 years
DRAM DIMM capacity doubling ~ every 3 years

Lim et al., ISCA 2009
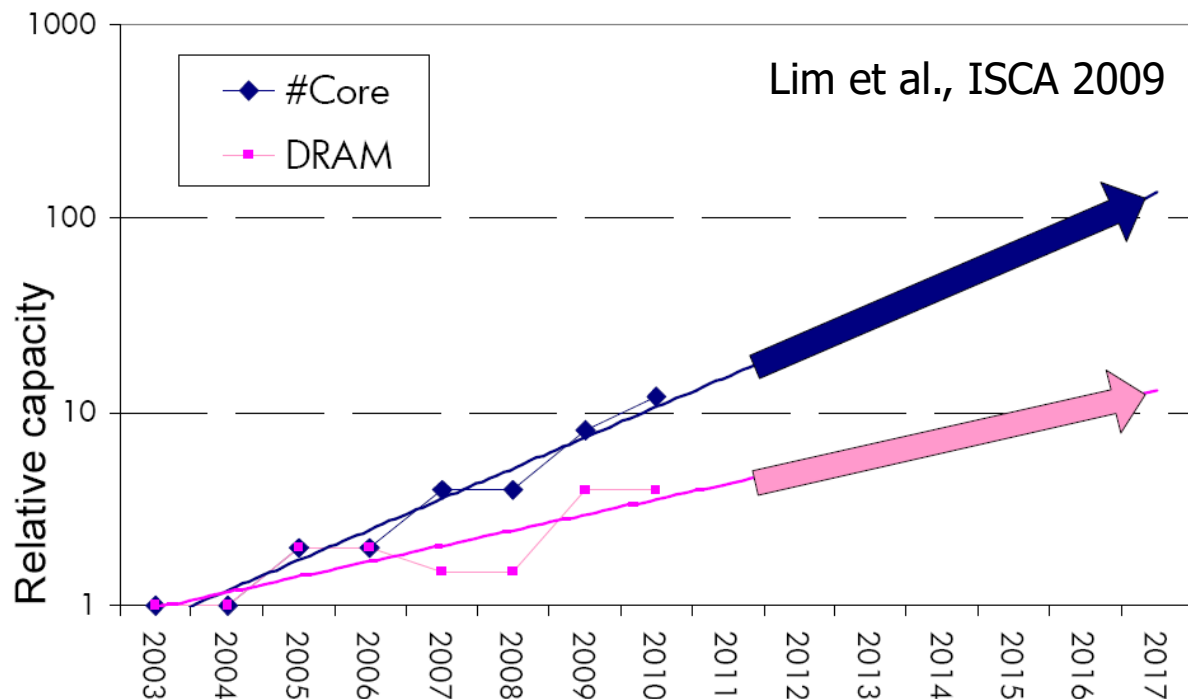
- *Memory capacity per core* expected to drop by 30% every two years
- Trends worse for *memory bandwidth per core!*

DRAM Capacity, Bandwidth & Latency
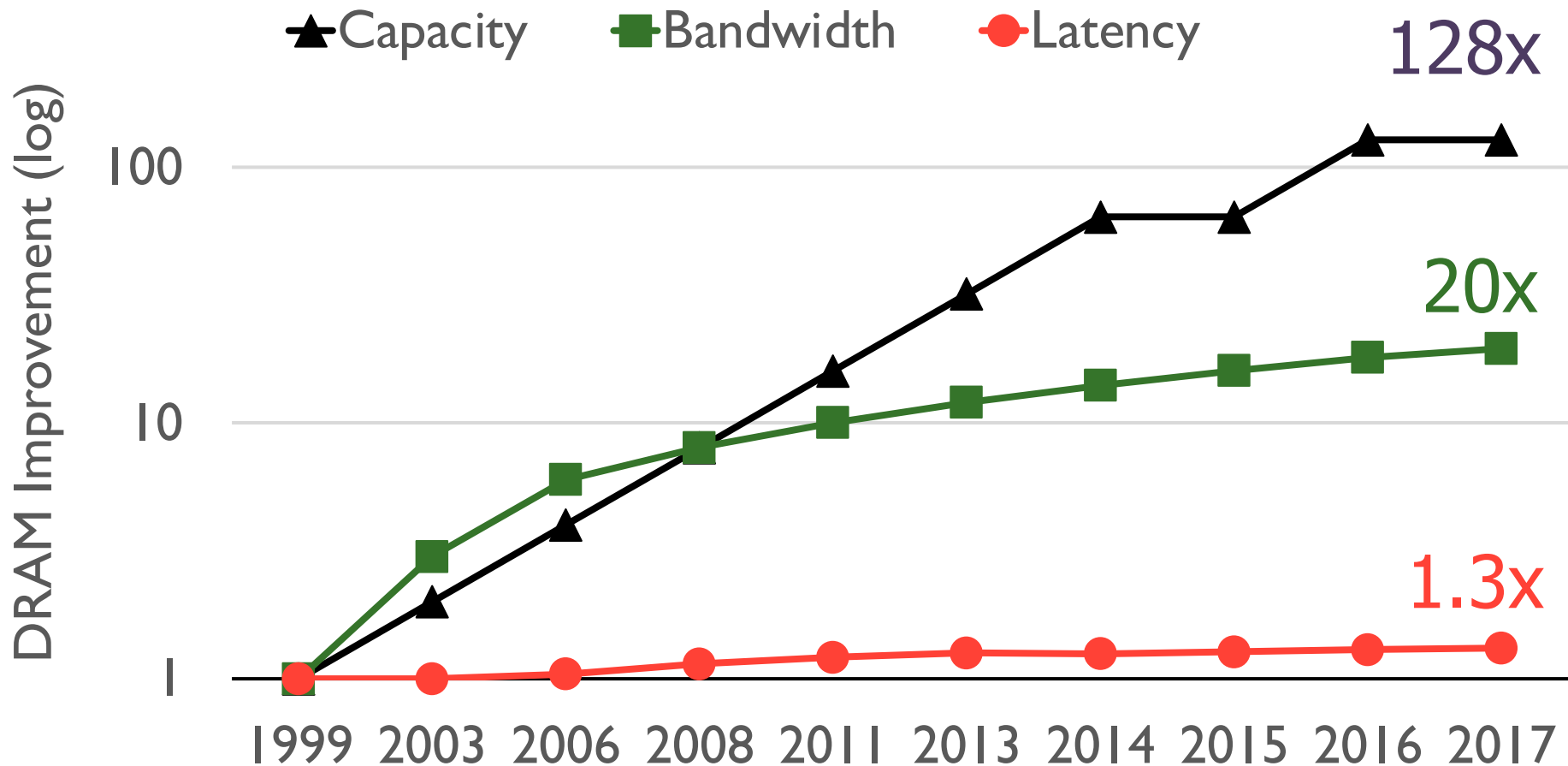
Memory Capacity Has Improved Greatly
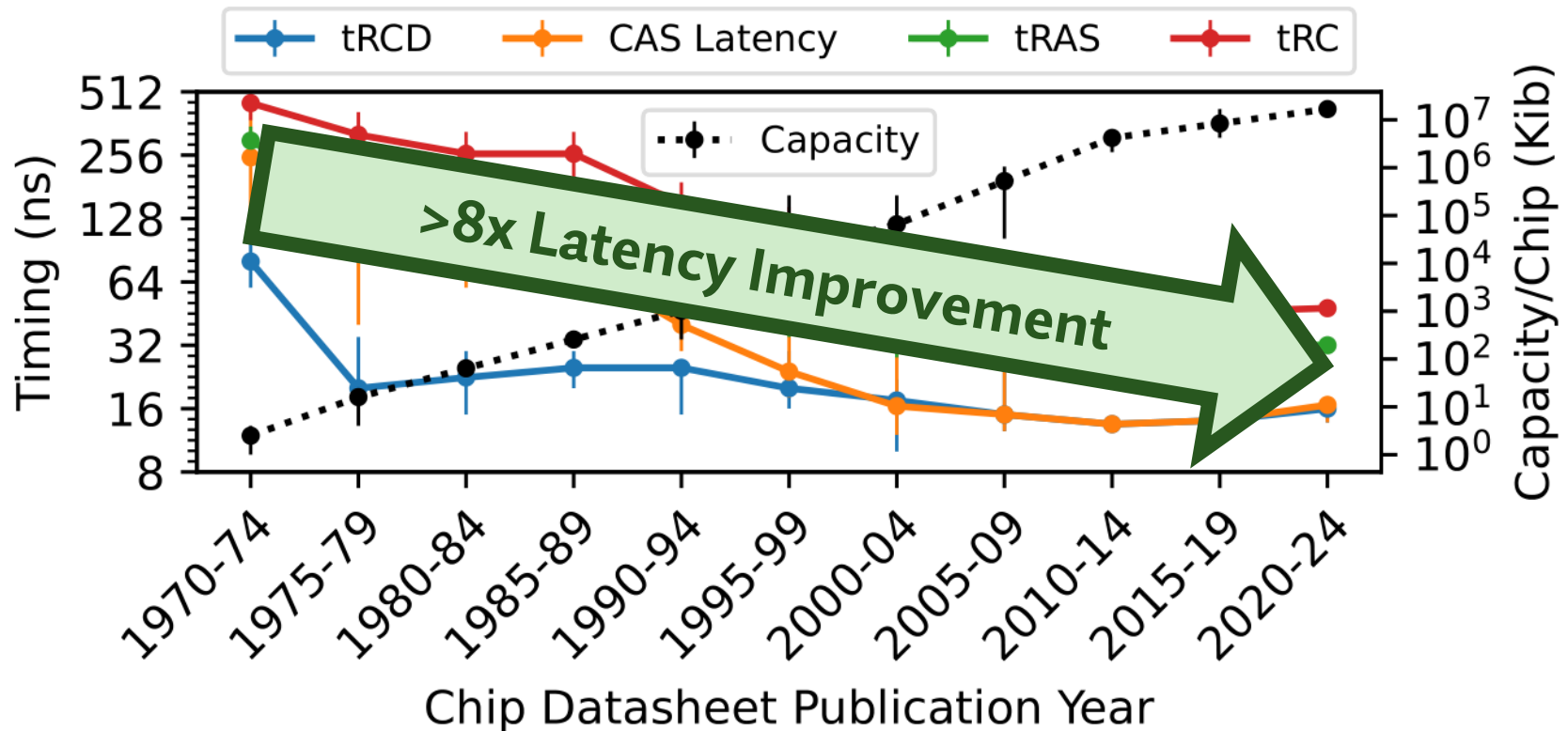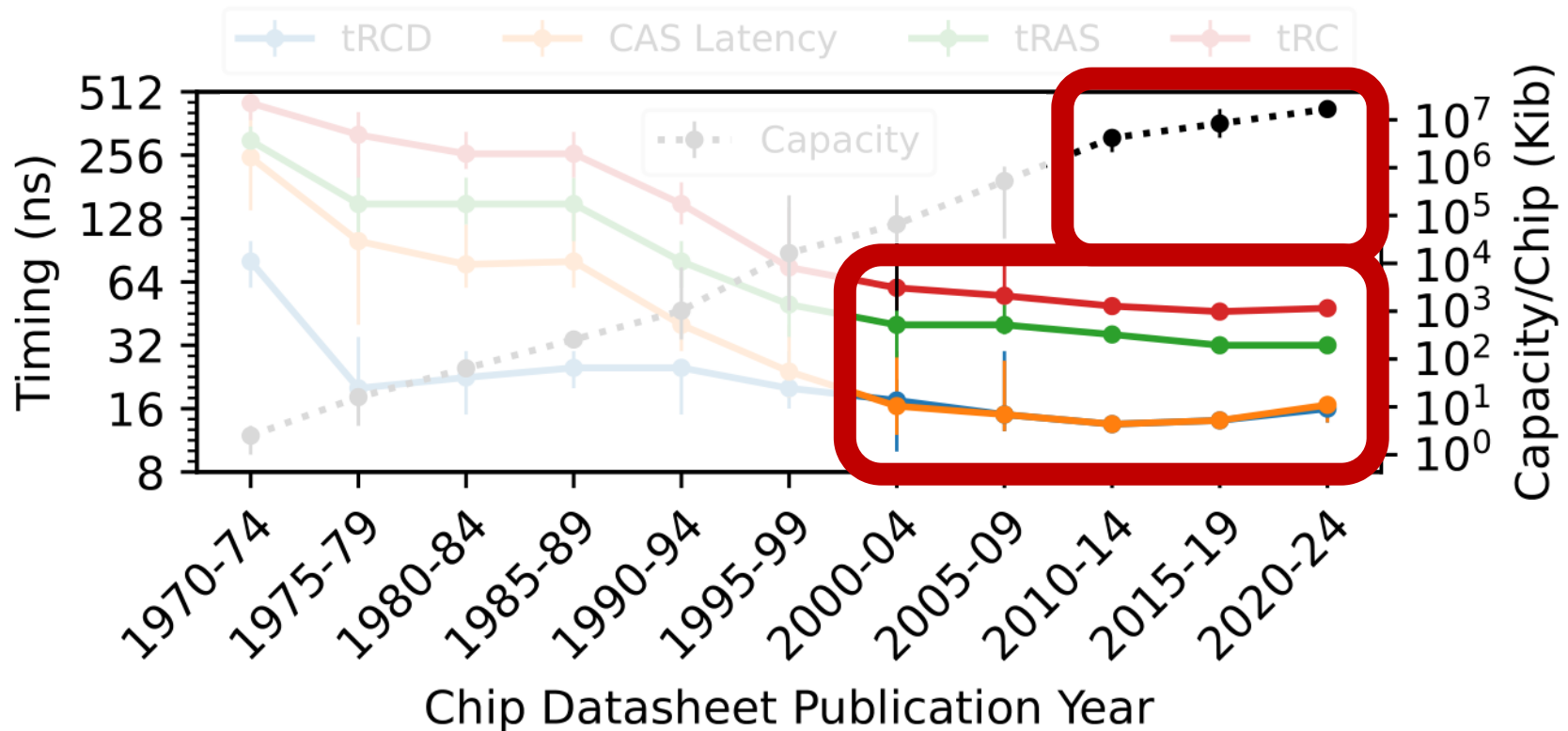
>1,000,000x Capacity Improvement

https://github.com/CMU-SAFARI/DRAM-Datasheet-Survey

https://arxiv.org/pdf/2204.10378

Patel+, "Rethinking the Producer-Consumer Relationship in Modern DRAM-Based Systems," 2024.

Memory Latency Lags Behind

https://github.com/CMU-SAFARI/DRAM-Datasheet-Survey

https://arxiv.org/pdf/2204.10378

Patel+, "Rethinking the Producer-Consumer Relationship in Modern DRAM-Based Systems," 2024.

Memory Latency Lags Behind

https://github.com/CMU-SAFARI/DRAM-Datasheet-Survey

https://arxiv.org/pdf/2204.10378

Patel+, "Rethinking the Producer-Consumer Relationship in Modern DRAM-Based Systems," 2024.

Memory Is Critical for Performance

In-memory Databases

[Mao+, EuroSys'12;
 Clapp+ (**Intel**), IISWC'15]

Graph/Tree Processing

[Xu+, IISWC'12; Umuroglu+, FPL'15]
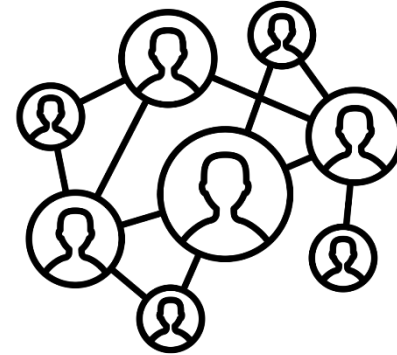
In-Memory Data Analytics

[Clapp+ (**Intel**), IISWC'15;
 Awan+, BDCloud'15]

Datacenter Workloads

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

Memory Is Critical for Performance

In-memory Databases

Graph/Tree Processing

Memory → performance bottleneck

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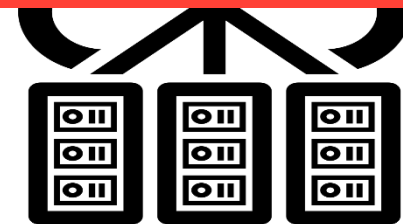
Datacenter Workloads
[Kanev+ (**Google**), ISCA'15]

Memory Is Critical for Performance

Chrome

Google's web browser

TensorFlow Mobile

Google's machine learning framework

Video Playback

Google's **video codec**

Video Capture

Google's **video codec**

Memory Is Critical for Performance

Chrome

TensorFlow Mobile

Memory → performance bottleneck

Video Playback

Google's **video codec**

Video Capture

Google's **video codec**

It's the Memory, Stupid!

- **"It's the Memory, Stupid!"** (Richard Sites, MPR, 1996)

RICHARD SITES

It's the Memory, Stupid!

When we started the Alpha architecture design in 1988, we estimated a 25-year lifetime and a relatively modest 32% per year compounded performance improvement of implementations over that lifetime (1,000× total). We guestimated about 10× would come from CPU clock improvement, 10× from multiple instruction issue, and 10× from multiple processors.

5, 1996 ♢ MICROPROCESSOR REPORT

I expect that over the coming decade memory subsystem design will be the *only* important design issue for microprocessors.

The Performance Perspective

Data from Runahead Execution [HPCA 2003]

Mutlu+, "Runahead Execution: An Alternative to Very Large Instruction Windows for Out-of-Order Processors," HPCA 2003.
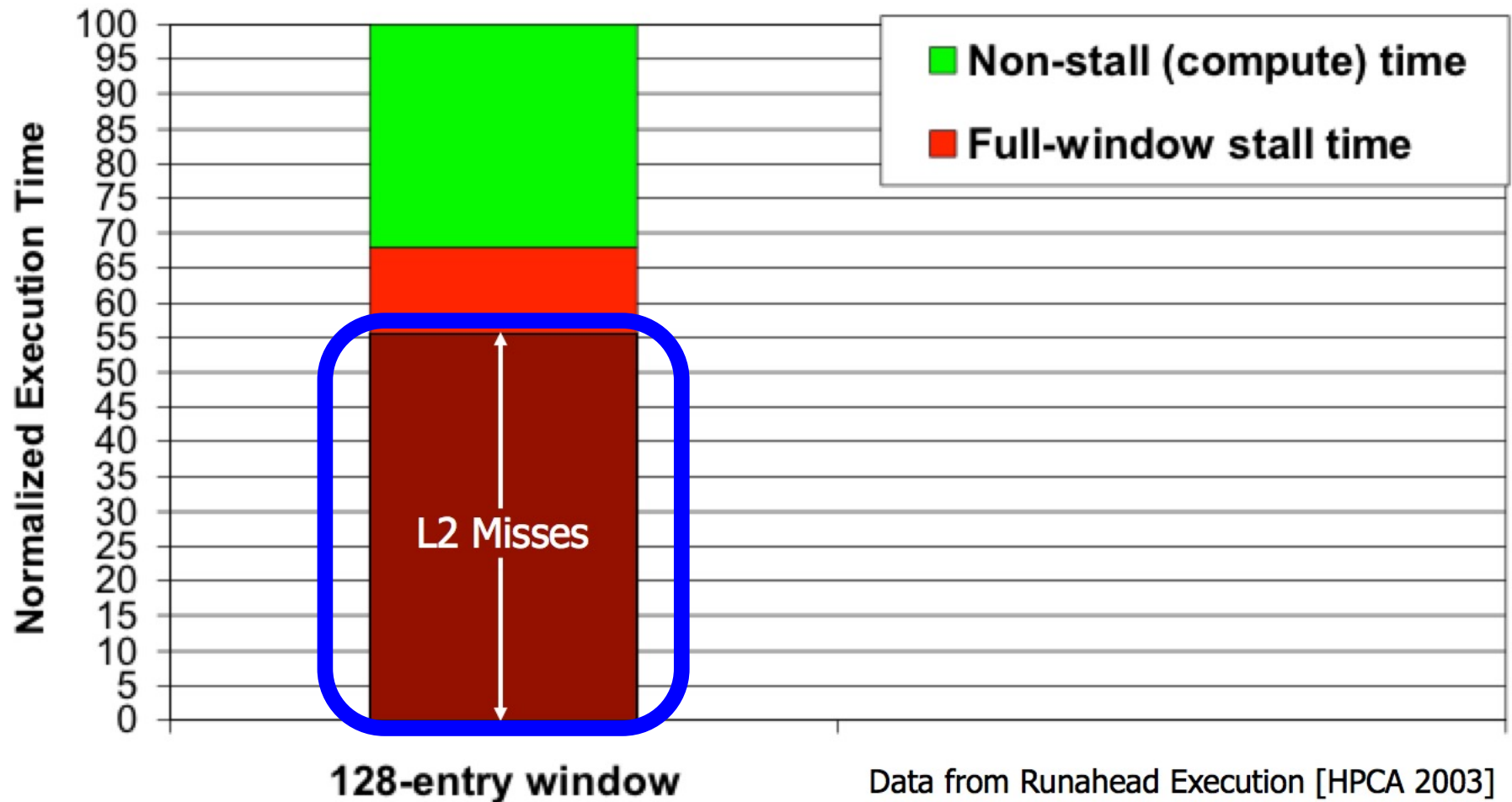
The Performance Perspective

- Onur Mutlu, Jared Stark, Chris Wilkerson, and Yale N. Patt,
  "Runahead Execution: An Alternative to Very Large Instruction Windows for Out-of-order Processors"
  Proceedings of the 9th International Symposium on High-Performance Computer Architecture (**HPCA**), pages 129-140, Anaheim, CA, February 2003. Slides (pdf)
  One of the 15 computer arch. papers of 2003 selected as Top Picks by IEEE Micro. HPCA Test of Time Award (awarded in 2021).

Runahead Execution: An Alternative to Very Large Instruction Windows for Out-of-order Processors

Onur Mutlu §   Jared Stark †   Chris Wilkerson ‡   Yale N. Patt §

§ECE Department
The University of Texas at Austin
{onur,patt}@ece.utexas.edu

†Microprocessor Research
Intel Labs
jared.w.stark@intel.com

‡Desktop Platforms Group
Intel Corporation
chris.wilkerson@intel.com

The Memory Bottleneck

- Onur Mutlu, Jared Stark, Chris Wilkerson, and Yale N. Patt,
"Runahead Execution: An Effective Alternative to Large Instruction Windows"
IEEE Micro, Special Issue: Micro's Top Picks from Microarchitecture Conferences (**MICRO TOP PICKS**), Vol. 23, No. 6, pages 20-25, November/December 2003.

RUNAHEAD EXECUTION:
AN EFFECTIVE ALTERNATIVE TO
LARGE INSTRUCTION WINDOWS

More on Runahead Execution (I)
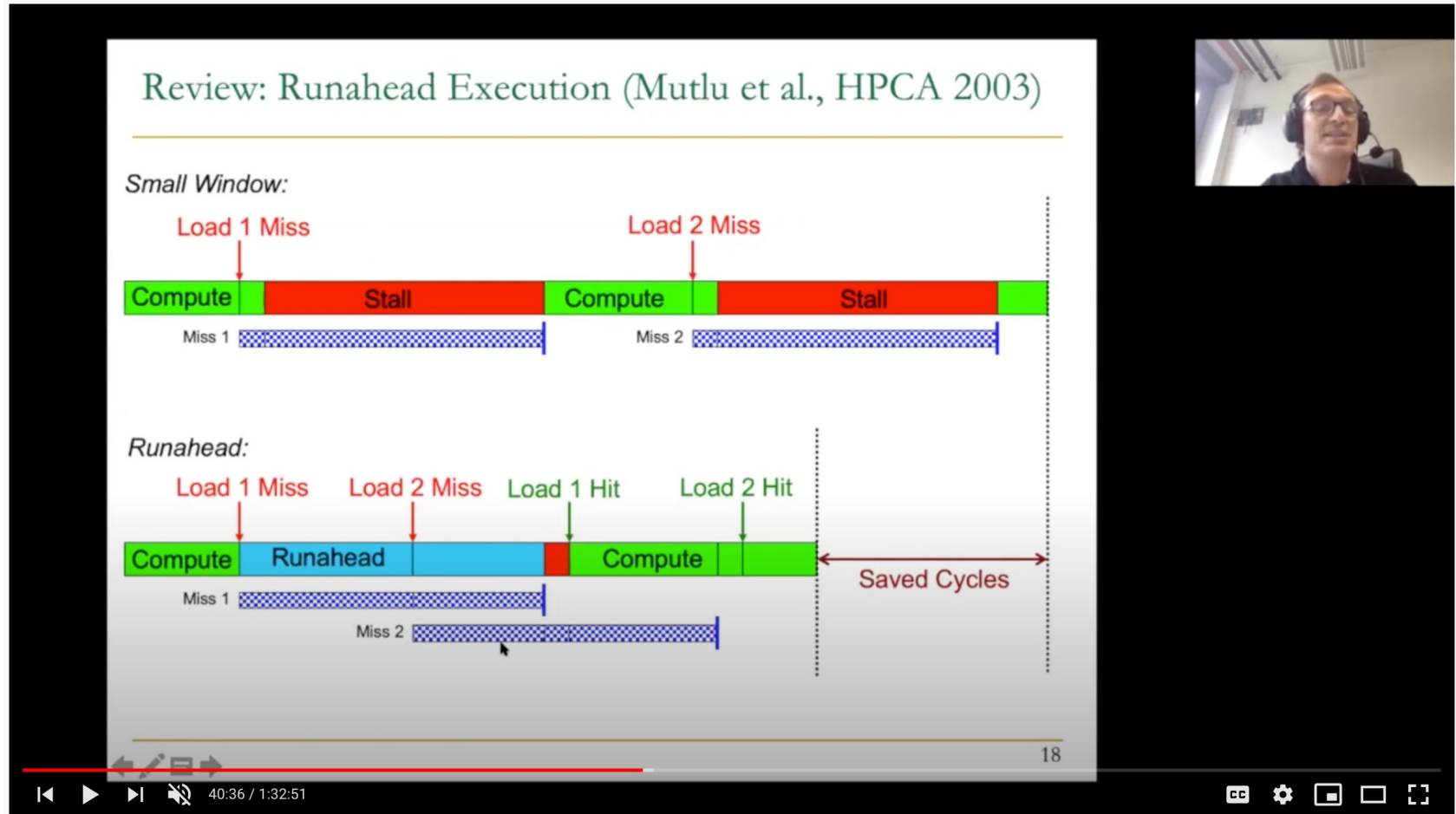
Computer Architecture - Lecture 19a: Execution-Based Prefetching (ETH Zürich, Fall 2020)

395 views • Nov 29, 2020

👍 14   👎 0   ➤ SHARE   ≡+ SAVE   ...

Onur Mutlu Lectures
16.5K subscribers

ANALYTICS   EDIT VIDEO

More on Runahead Execution (II)

The Performance Perspective (Today)

- All of Google's Data Center Workloads (2015):

Kanev+, "Profiling a Warehouse-Scale Computer," ISCA 2015.

153
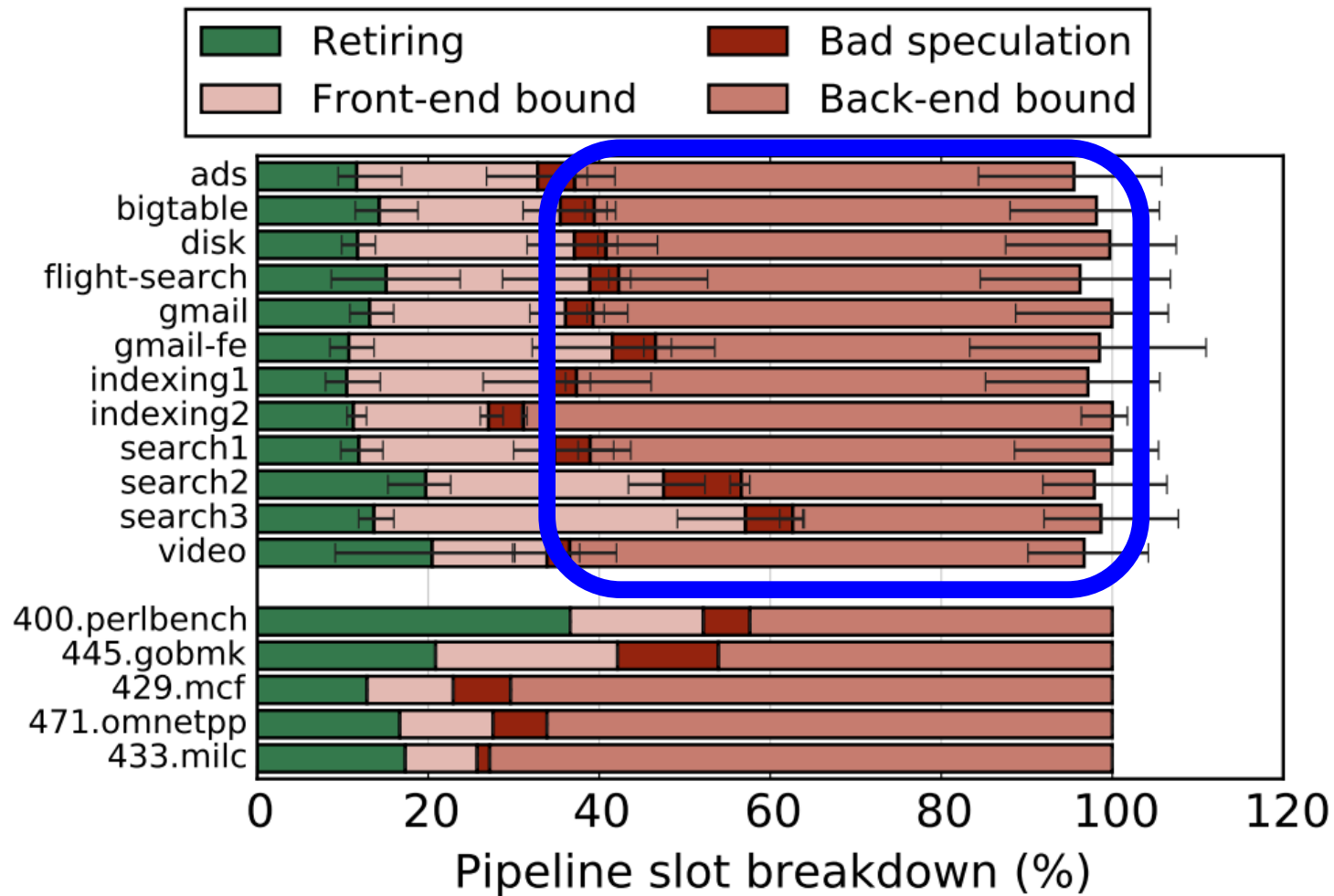
The Memory Bottleneck

- All of Google's Data Center Workloads (2015):
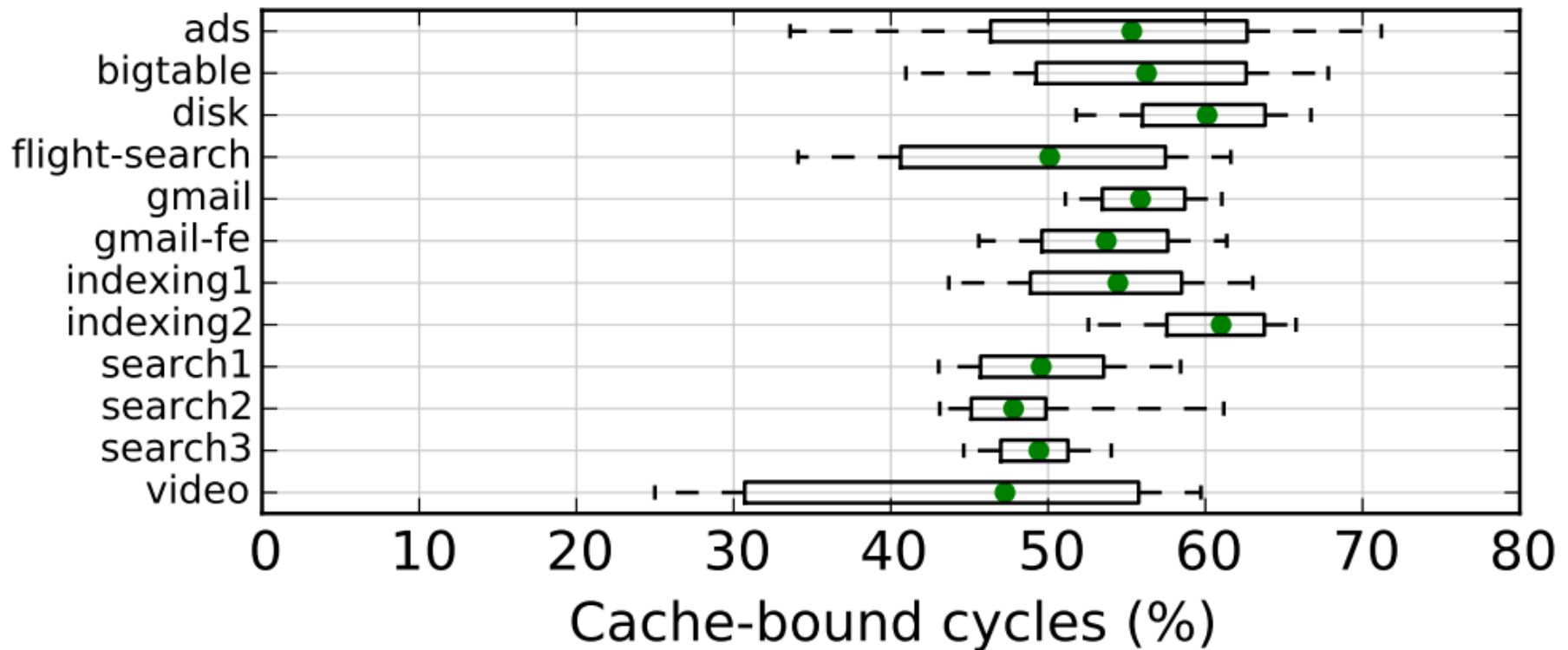
Figure 11: Half of cycles are spent stalled on caches.

Memory in a Modern System
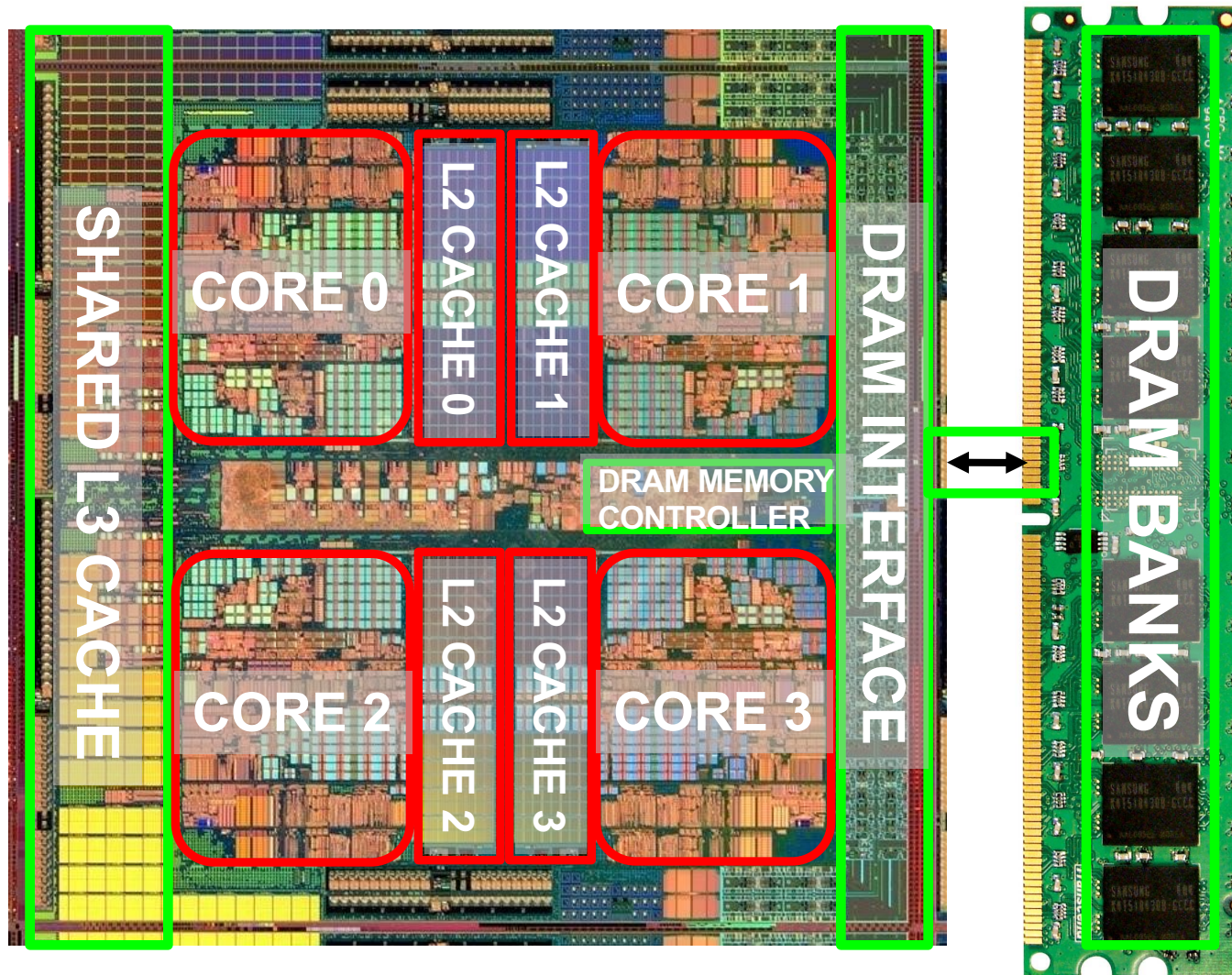
AMD Barcelona, 2006

A Large Fraction of Modern Systems is Memory

Processor chip Level 2 cache chip

Multi-chip module package
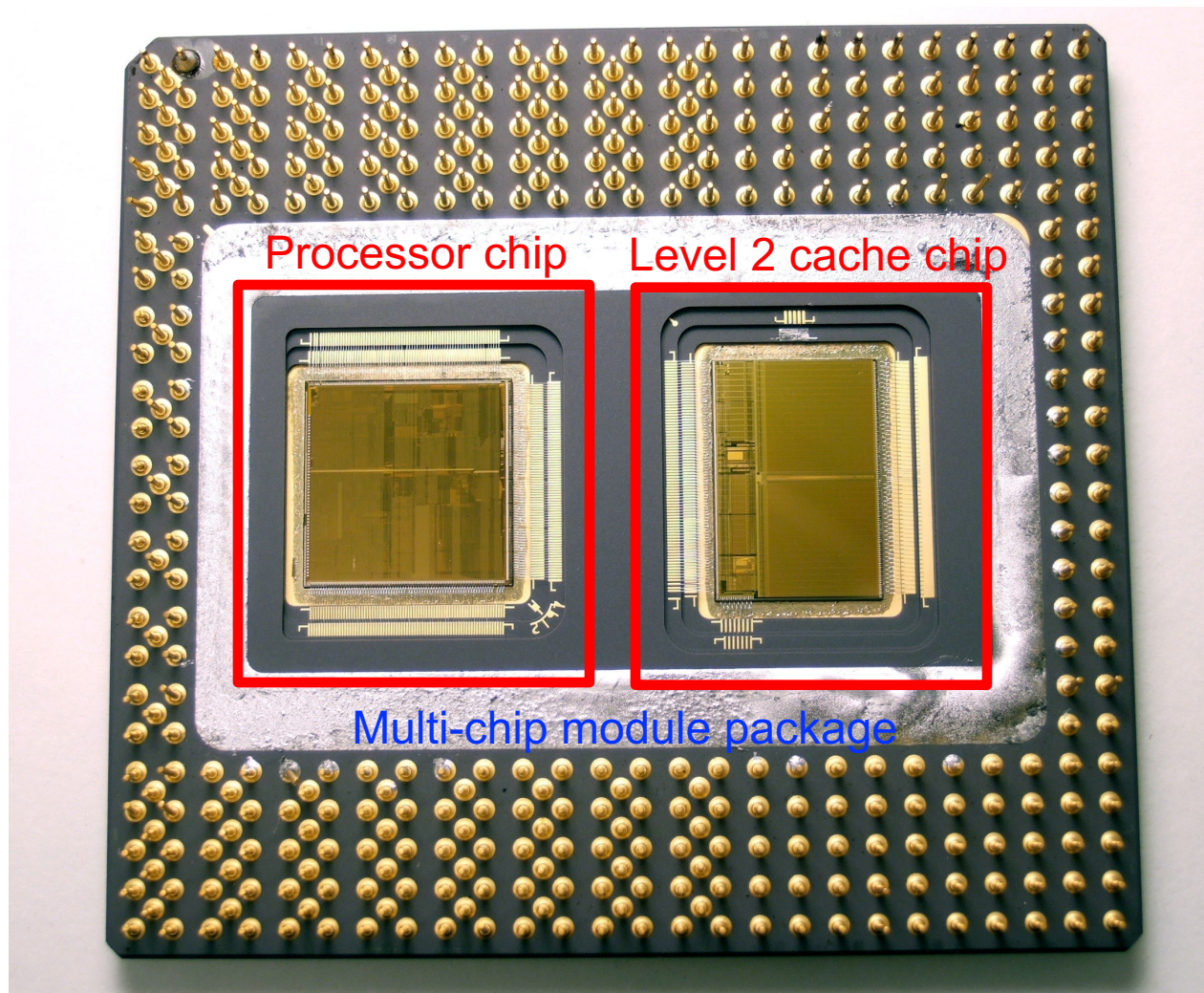
Intel Pentium Pro, 1995

A Large Fraction of Modern Systems is Memory

L2 Cache

Intel Pentium 4, 2000
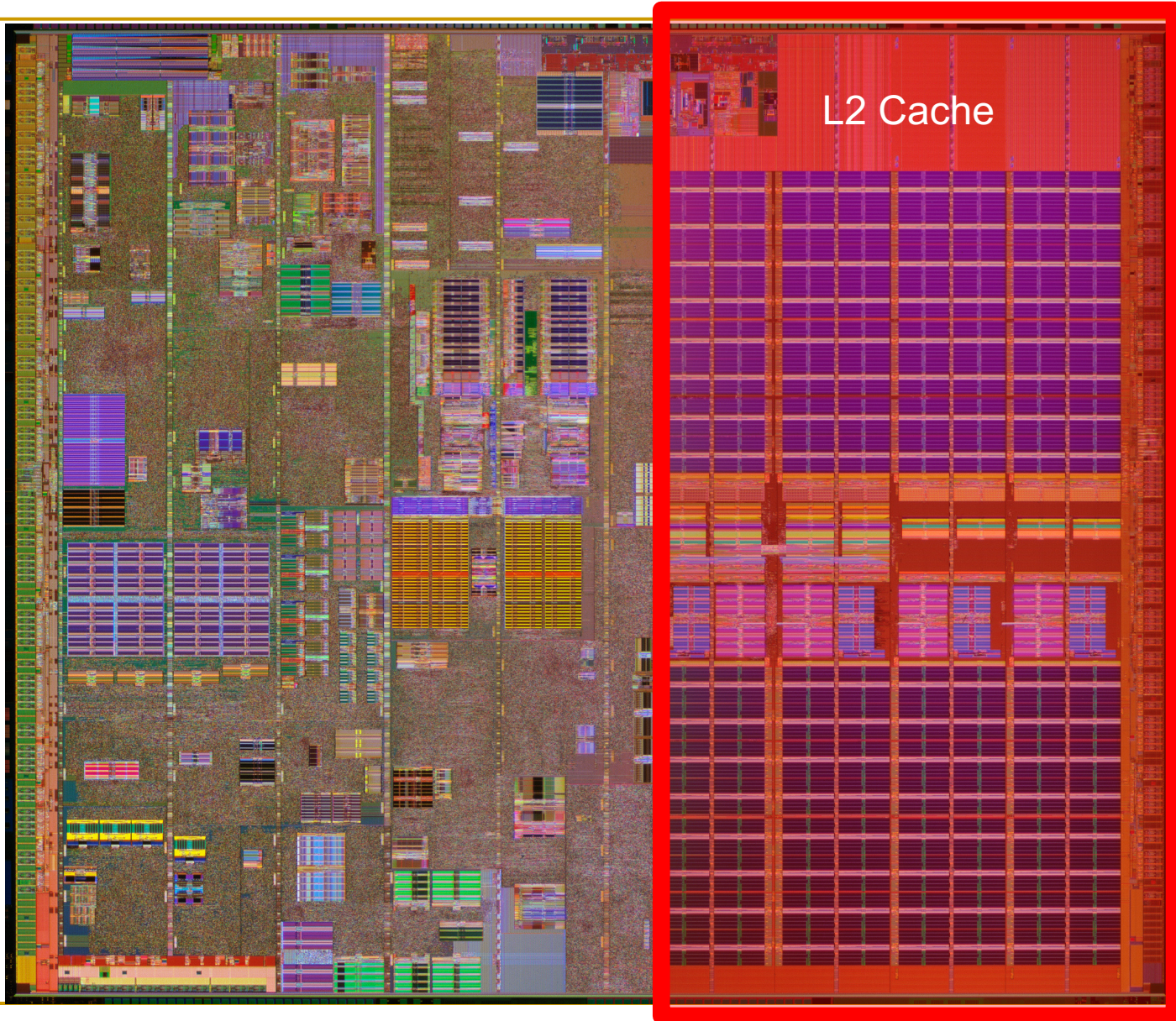
Deeper and Larger Cache Hierarchies
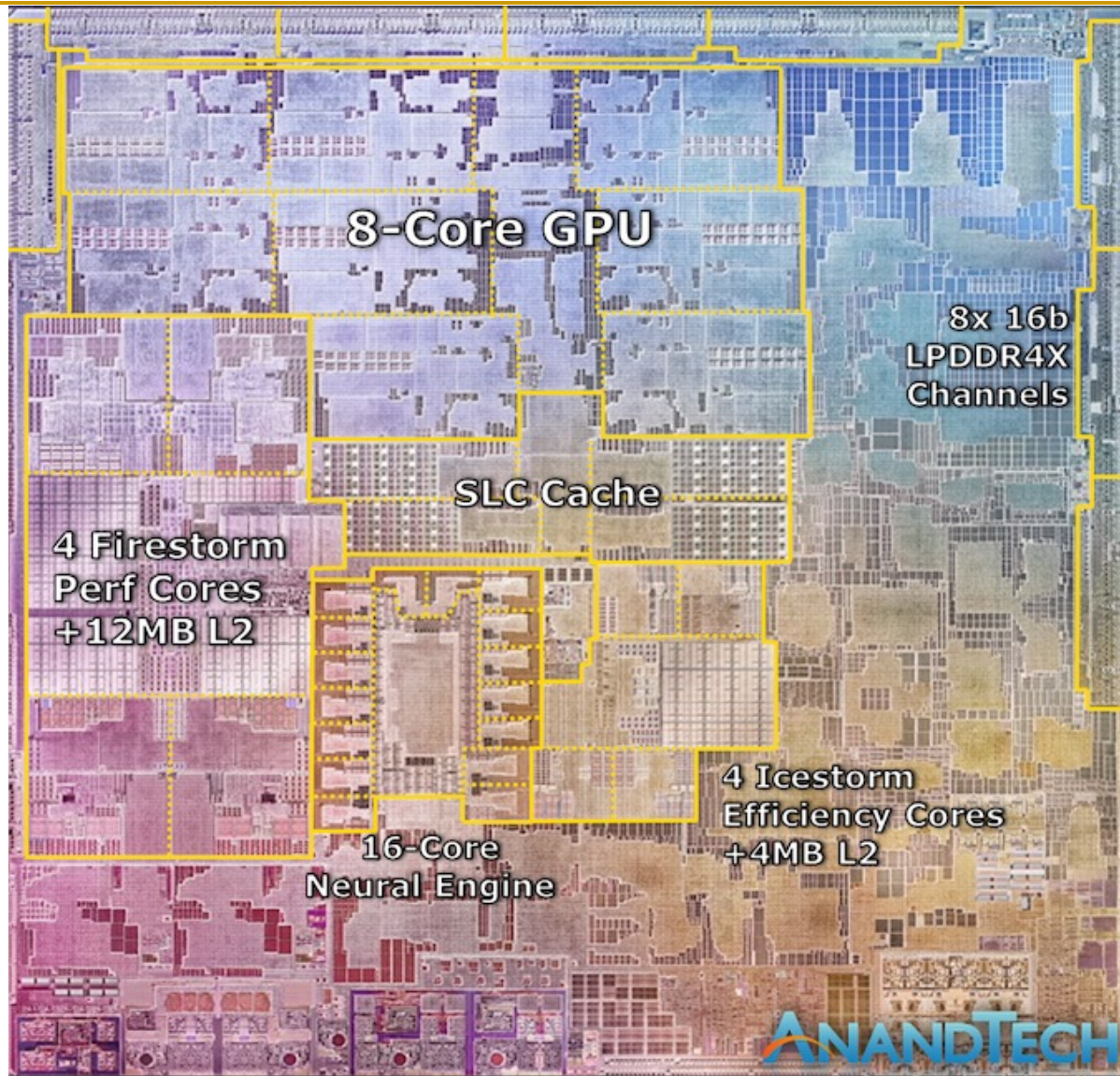
Apple M1, 2021

Deeper and Larger Cache Hierarchies
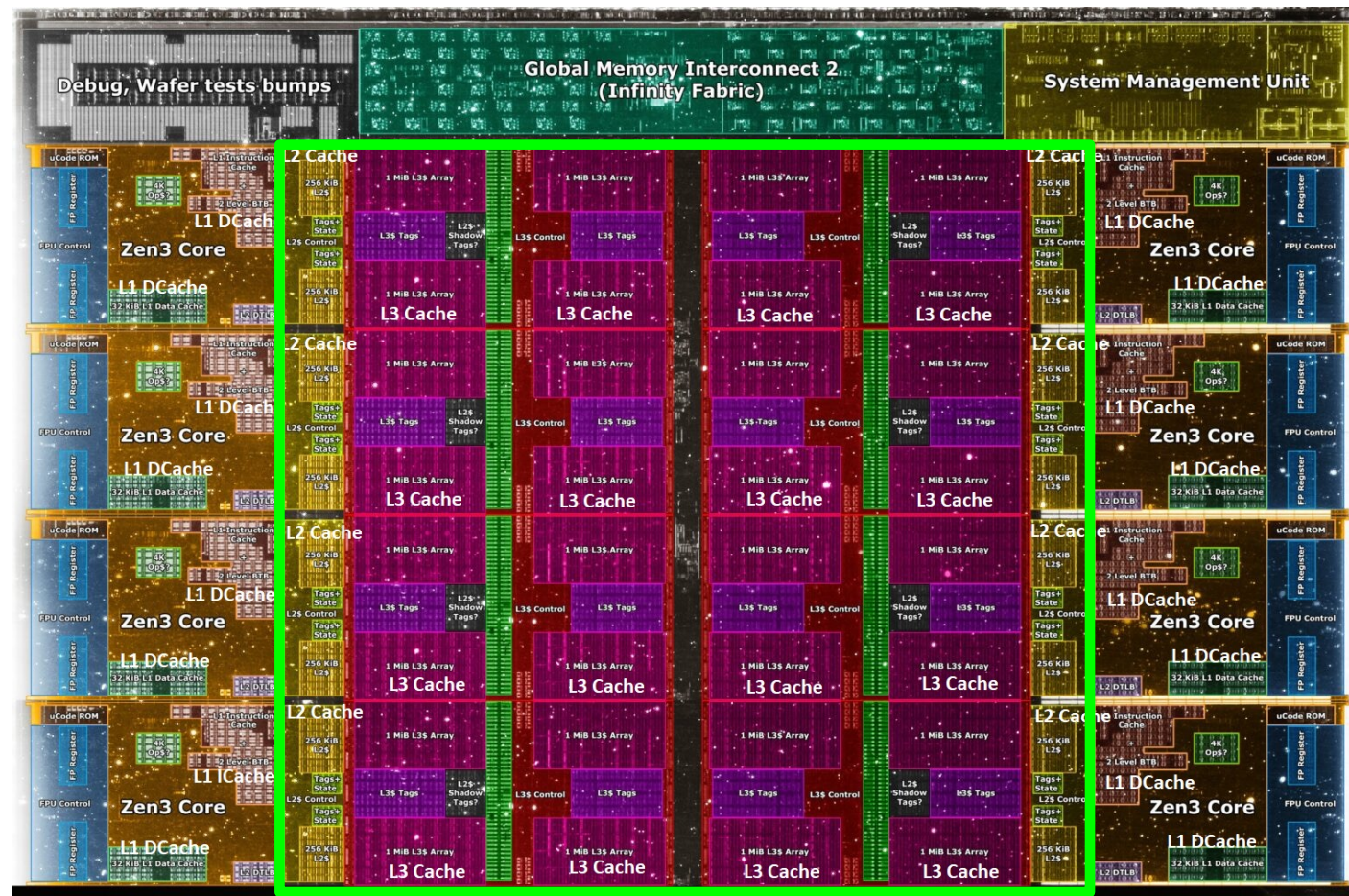
Core Count:
8 cores/16 threads

L1 Caches:
32 KB per core

L2 Caches:
512 KB per core

L3 Cache:
32 MB shared

AMD Ryzen 5000, 2020

Deeper and Larger Cache Hierarchies
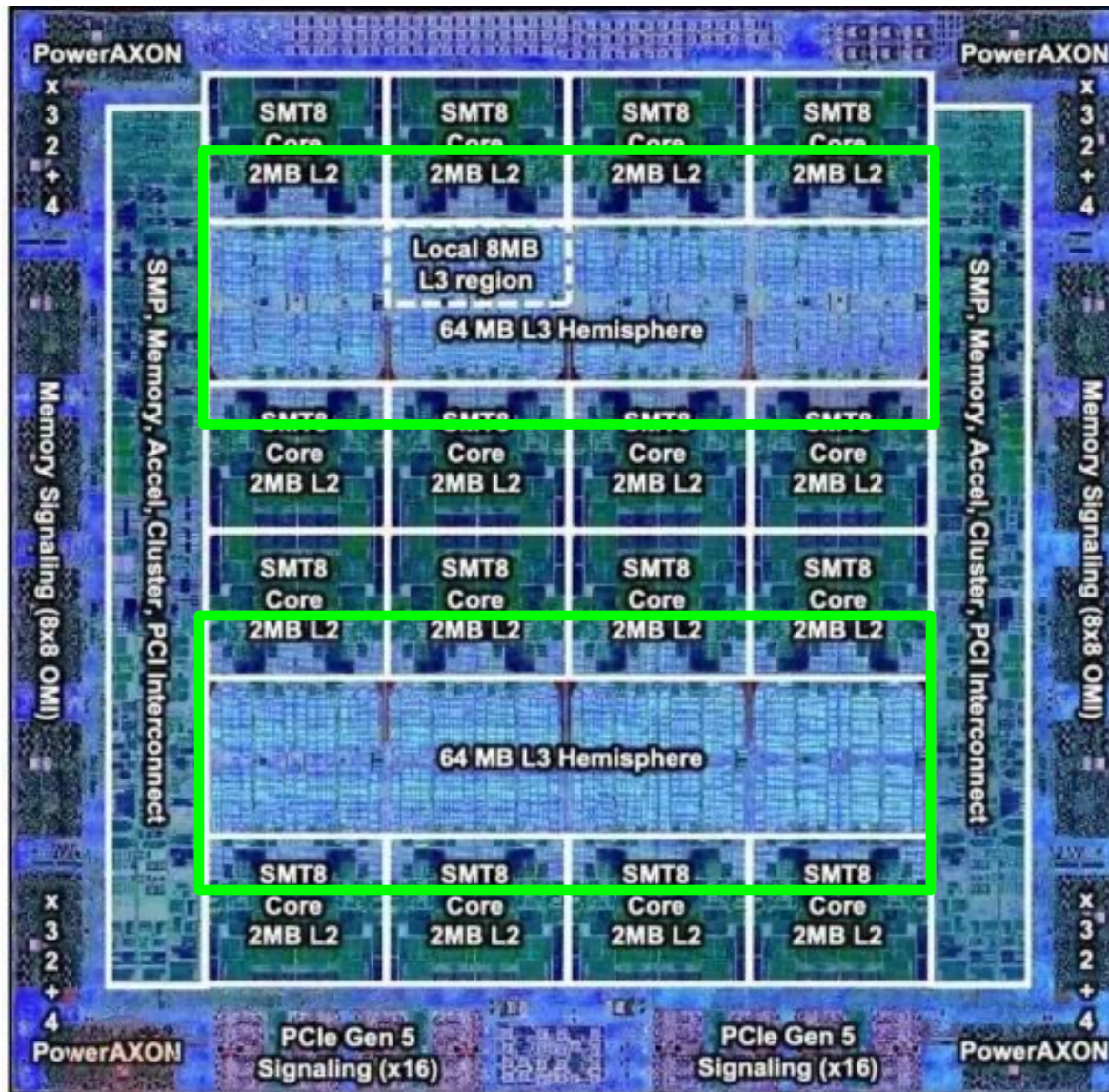
IBM POWER10, 2020

Cores:
15-16 cores,
8 threads/core

L2 Caches:
2 MB per core

L3 Cache:
120 MB shared

Deeper and Larger Cache Hierarchies
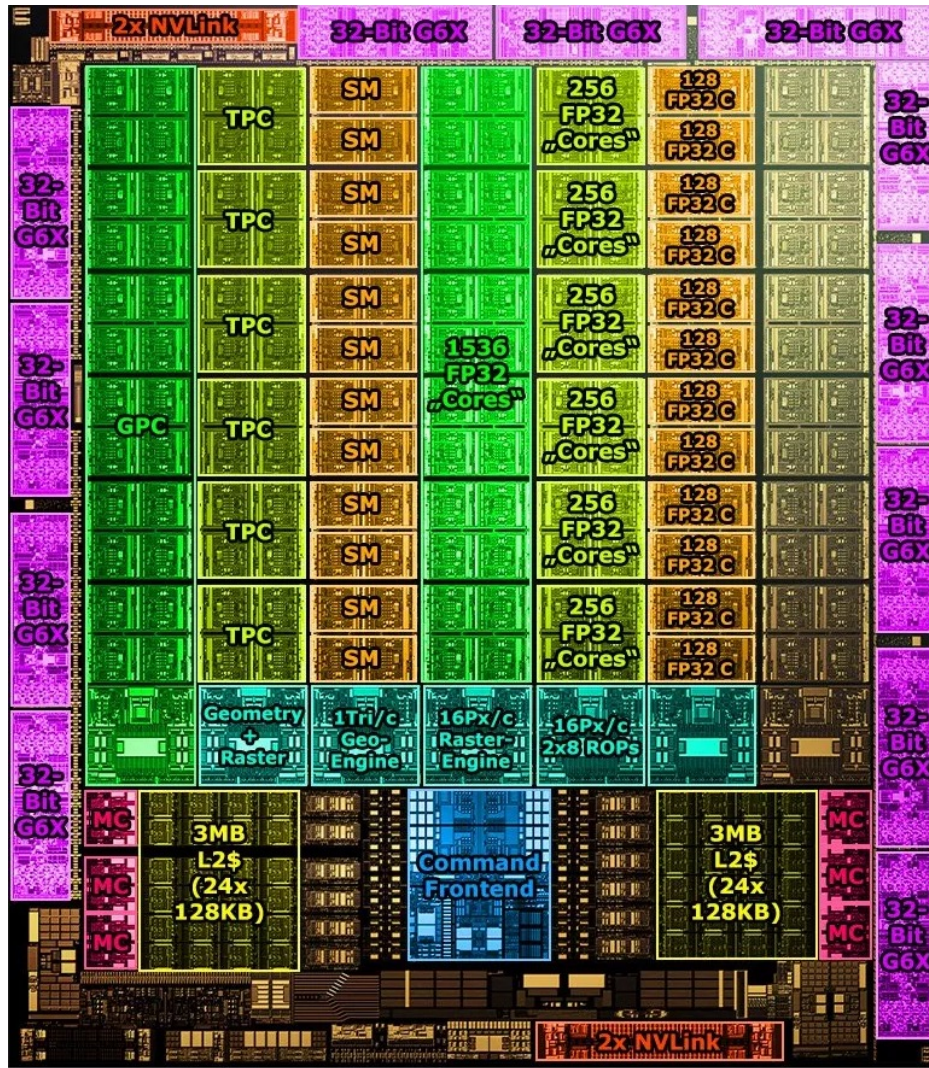
Nvidia Ampere, 2020

Cores:
128 Streaming Multiprocessors

L1 Cache or Scratchpad:
192KB per SM
Can be used as L1 Cache and/or Scratchpad

L2 Cache:
40 MB shared

AMD's 3D Last Level Cache (2021)

Zen 3 Layout

CCD

CPU Core

CPU Core

CPU Core

CPU Core

32MB
L3 Cache

CPU Core

CPU Core

CPU Core

CPU Core

https://community.microcenter.com/discussion/5134/comparing-zen-3-to-zen-2
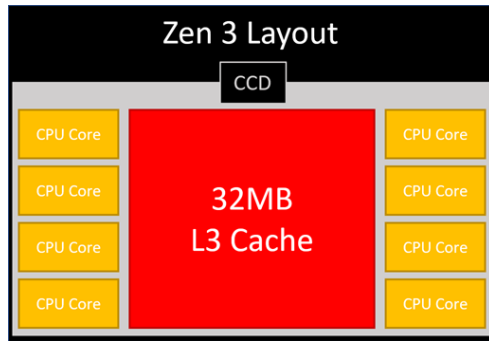
AMD increases the L3 size of their 8-core Zen 3 processors from 32 MB to 96 MB

Additional 64 MB L3 cache die stacked on top of the processor die
- Connected using Through Silicon Vias (TSVs)
- Total of 96 MB L3 cache

Structural silicon

64MB L3 cache die

Direct copper-to-copper bond

Through Silicon Vias (TSVs) for silicon-to-silicon communication

Up to 8-core "Zen 3" CCD

Deeper and Larger Memory Hierarchies

Storage   DRAM   A lot of SRAM   DRAM   Storage
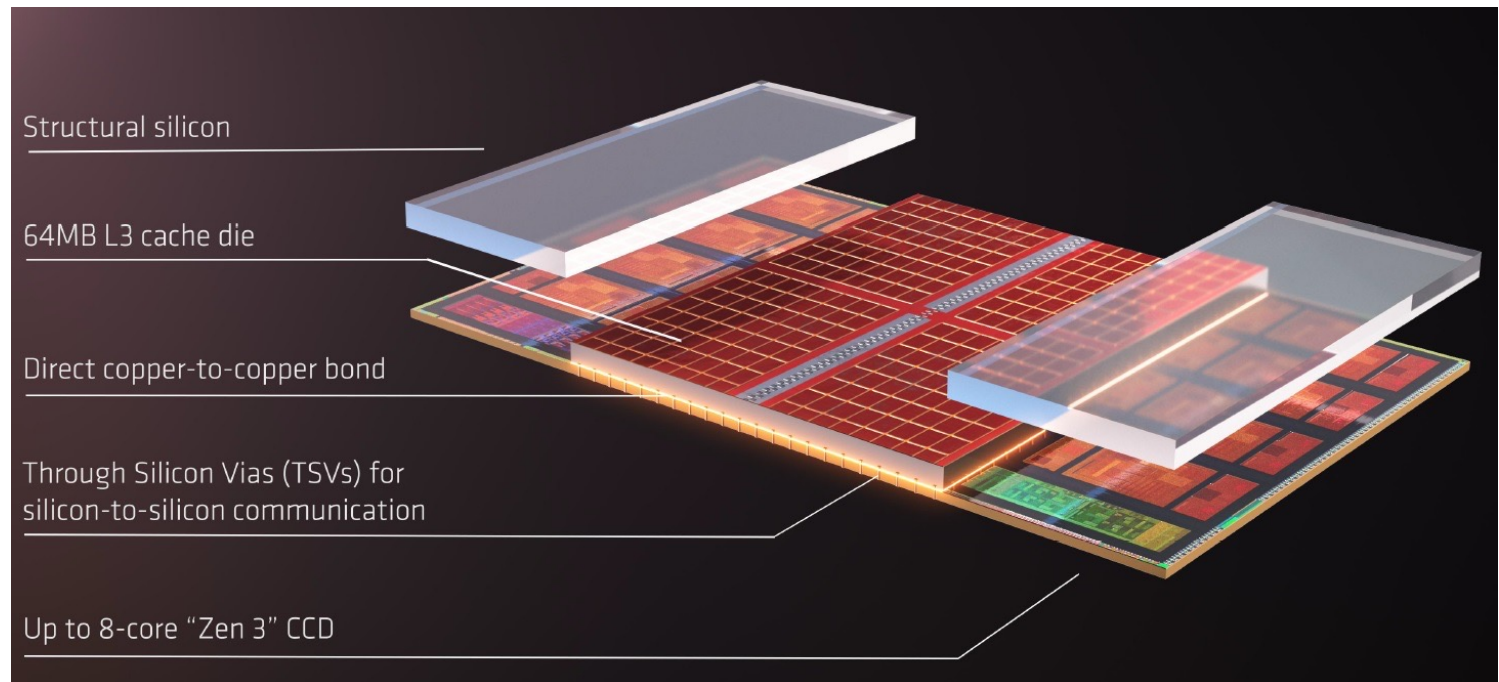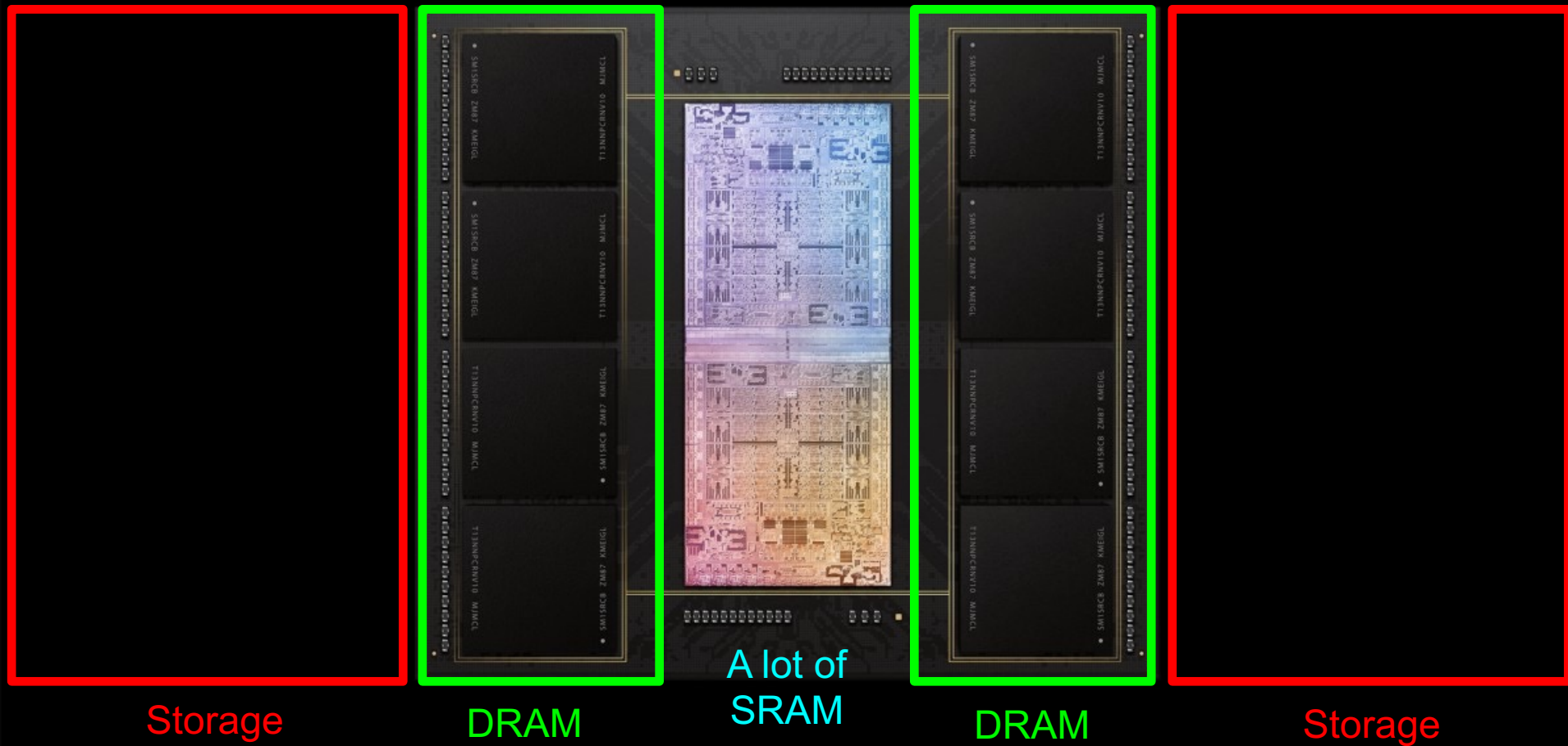
Apple M1 Ultra System (2022)

Deeper and Larger Memory Hierarchies

Announcing: The SambaNova SN40L

5nm TSMC

102B Transistors

1,040 RDU Cores

638 TFLOPS (bf16)

SambaNova
Cerulean SN40L RDU

Generative AI Training and Inference

3-tier Dataflow Memory

520 MB On-Chip Memory

64 GB High Bandwidth Memory

1.5 TB High Capacity Memory
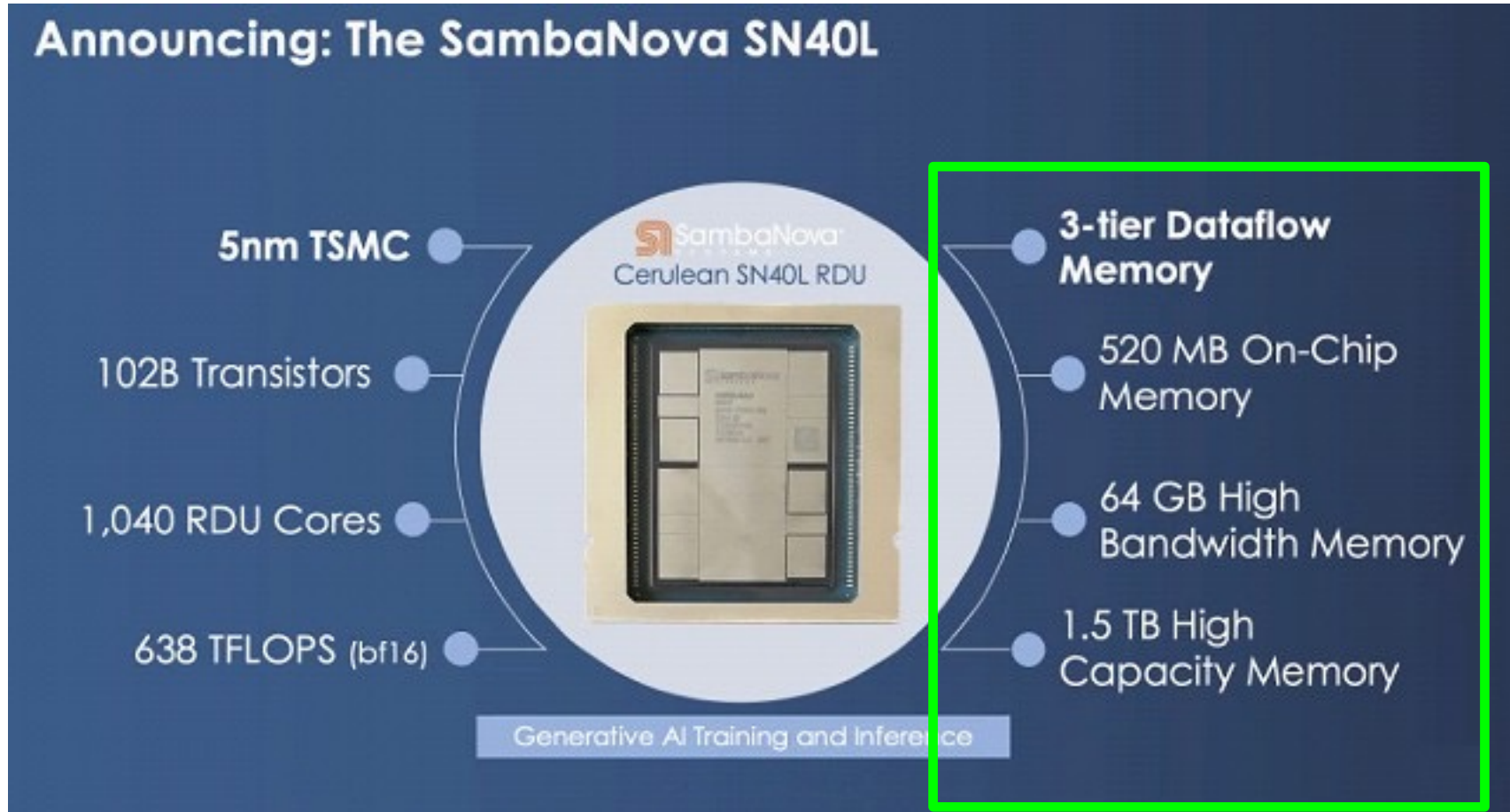
Memory System: **Most of the Platform**
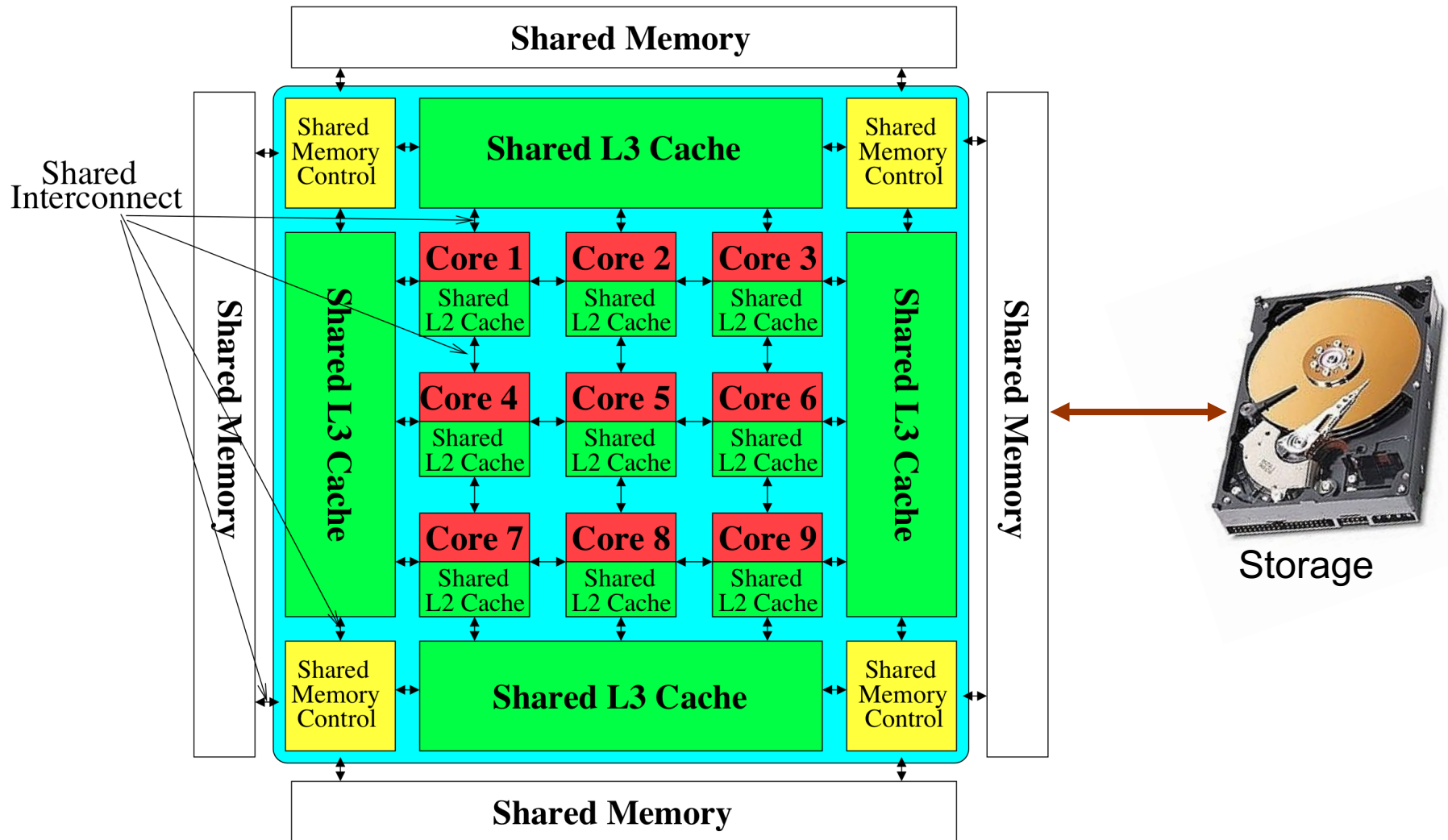
Most of the system is dedicated to storing and moving data

Yet, system is still bottlenecked by memory

Major Trends Affecting Main Memory (III)

- Need for main memory capacity, bandwidth, QoS increasing

- Main memory energy/power is a key system design concern
  - ~40-50% energy spent in off-chip memory hierarchy [Lefurgy, IEEE Computer'03] >40% power in DRAM [Ware, HPCA'10][Paul,ISCA'15]
  - DRAM consumes power even when not used (periodic refresh)

- DRAM technology scaling is ending

Data Movement vs. Computation Energy

Communication Dominates Arithmetic

Dally, HiPEAC 2015
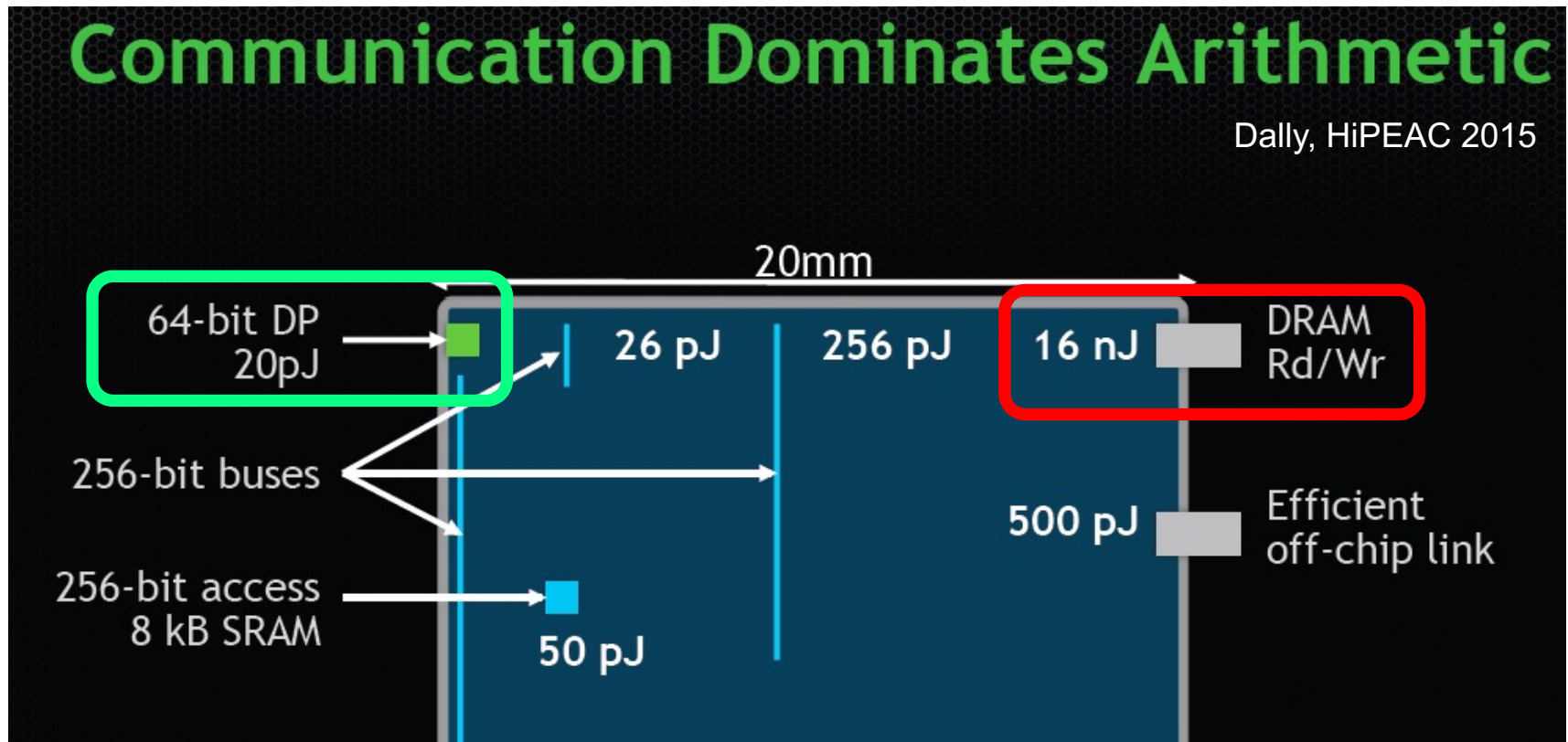
A memory access consumes ~100-1000X the energy of a complex addition

Data Movement vs. Computation Energy
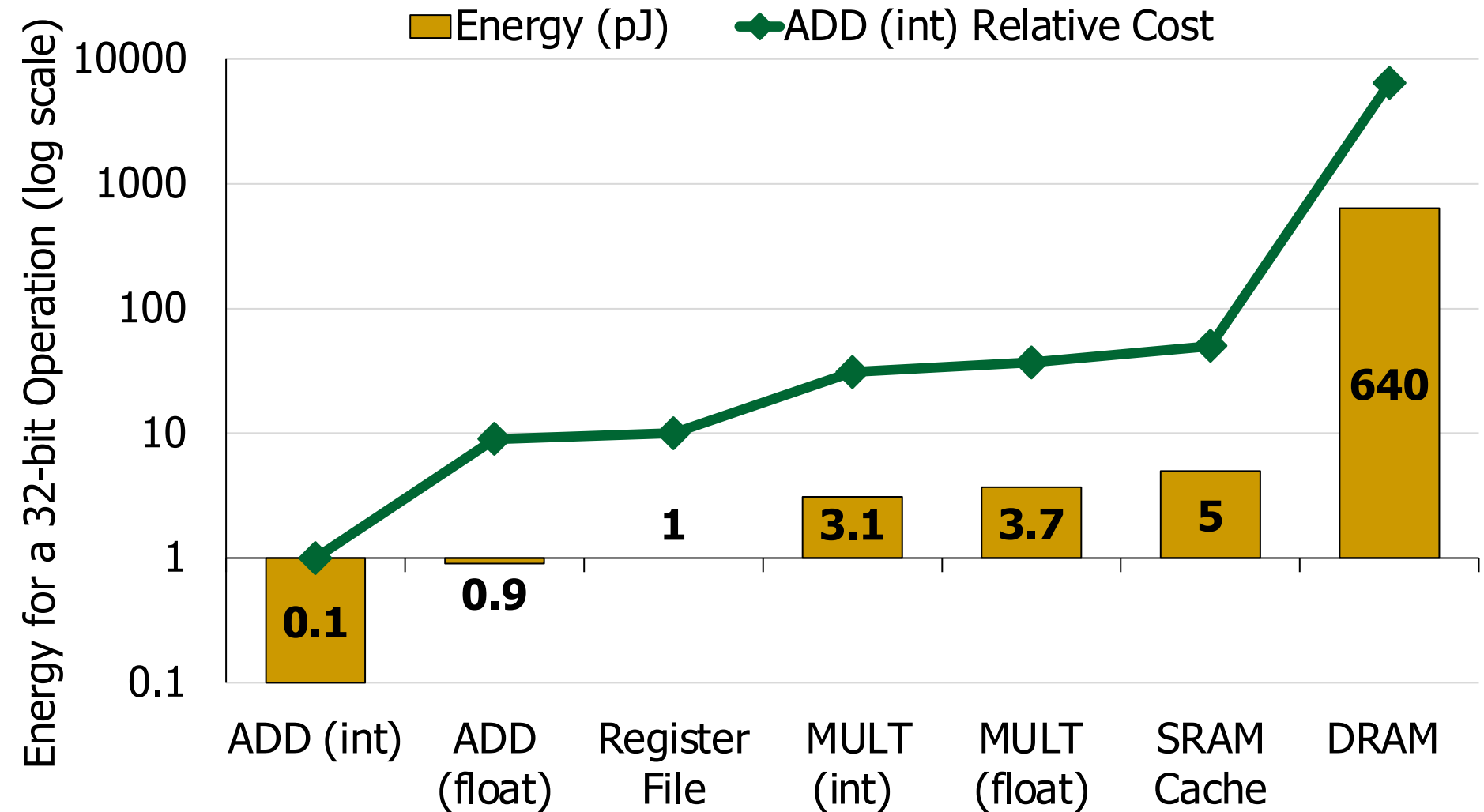
Data Movement vs. Computation Energy
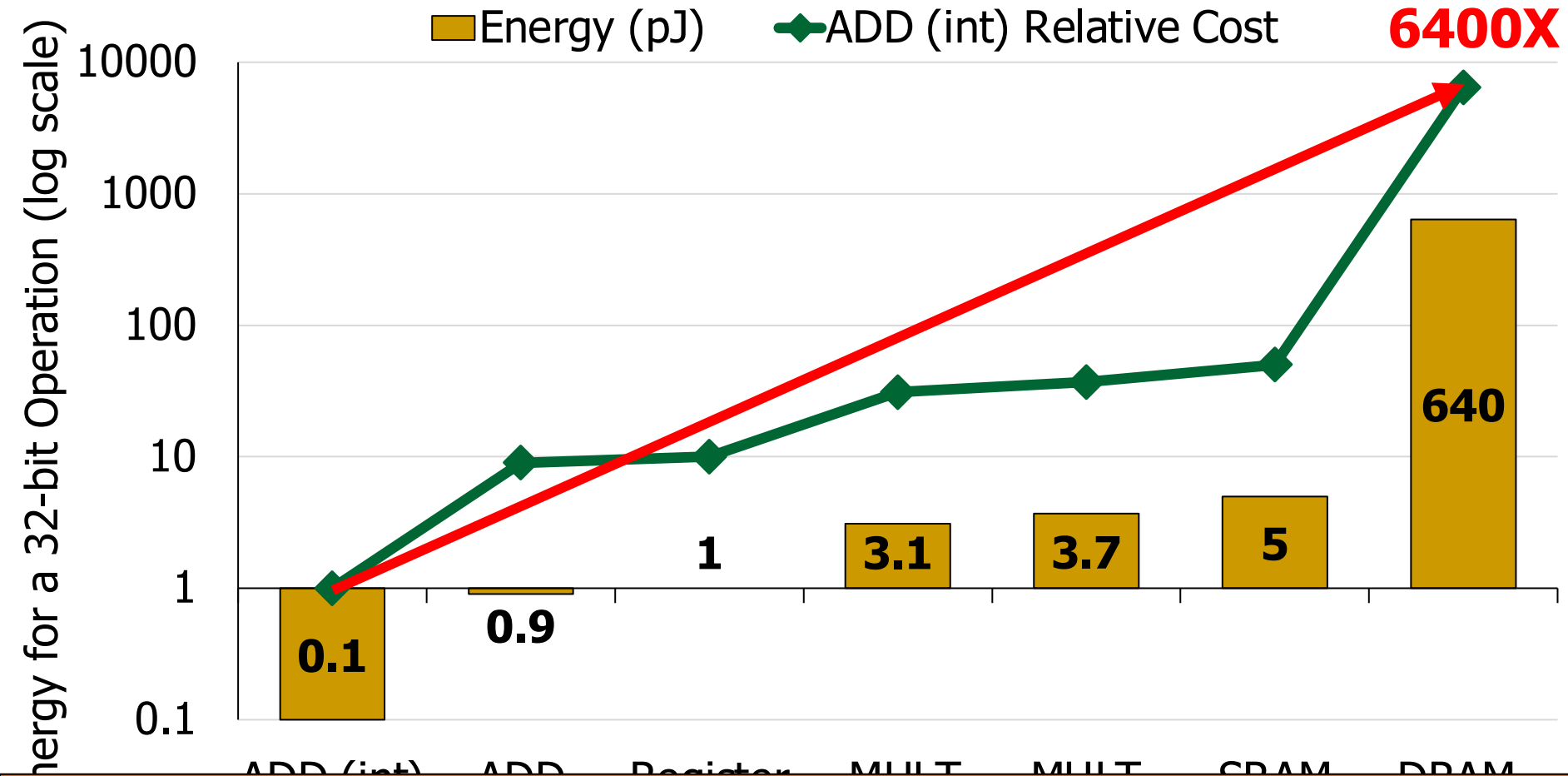
A memory access consumes 6400X the energy of a simple integer addition

Energy Waste in 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.

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]

SAFARI

Memory is Critical for Energy

- Amirali Boroumand, Saugata Ghose, Berkin Akin, Ravi Narayanaswami, Geraldo F. Oliveira, Xiaoyu Ma, Eric Shiu, and Onur Mutlu,
  "Google Neural Network Models for Edge Devices: Analyzing and Mitigating Machine Learning Inference Bottlenecks"
  Proceedings of the *30th International Conference on Parallel Architectures and Compilation Techniques* (**PACT**), Virtual, September 2021.
  [Slides (pptx) (pdf)]
  [Talk Video (14 minutes)]

> **> 90% of the total system energy
> is spent on memory in large ML models**

Google Neural Network Models for Edge Devices:
Analyzing and Mitigating Machine Learning Inference Bottlenecks

Amirali Boroumand[†◇]   Saugata Ghose[‡]   Berkin Akin[§]   Ravi Narayanaswami[§]
Geraldo F. Oliveira[★]   Xiaoyu Ma[§]   Eric Shiu[§]   Onur Mutlu[★†]

[†]*Carnegie Mellon Univ.*   [◇]*Stanford Univ.*   [‡]*Univ. of Illinois Urbana-Champaign*   [§]*Google*   [★]*ETH Zürich*

Example Energy Breakdowns

Legend: Total Static (black), PE (red), Param Buffer+NoC (gray), Act Buffer+NoC (green), Off-chip Interconnect (blue), DRAM (yellow)

Categories: LSTM1, Transd.1, Transd.2, CNN5, CNN9, CNN10, CNN12, RCNN1, RCNN3, Average — each with Baseline, Base+HB, Mensa bars.
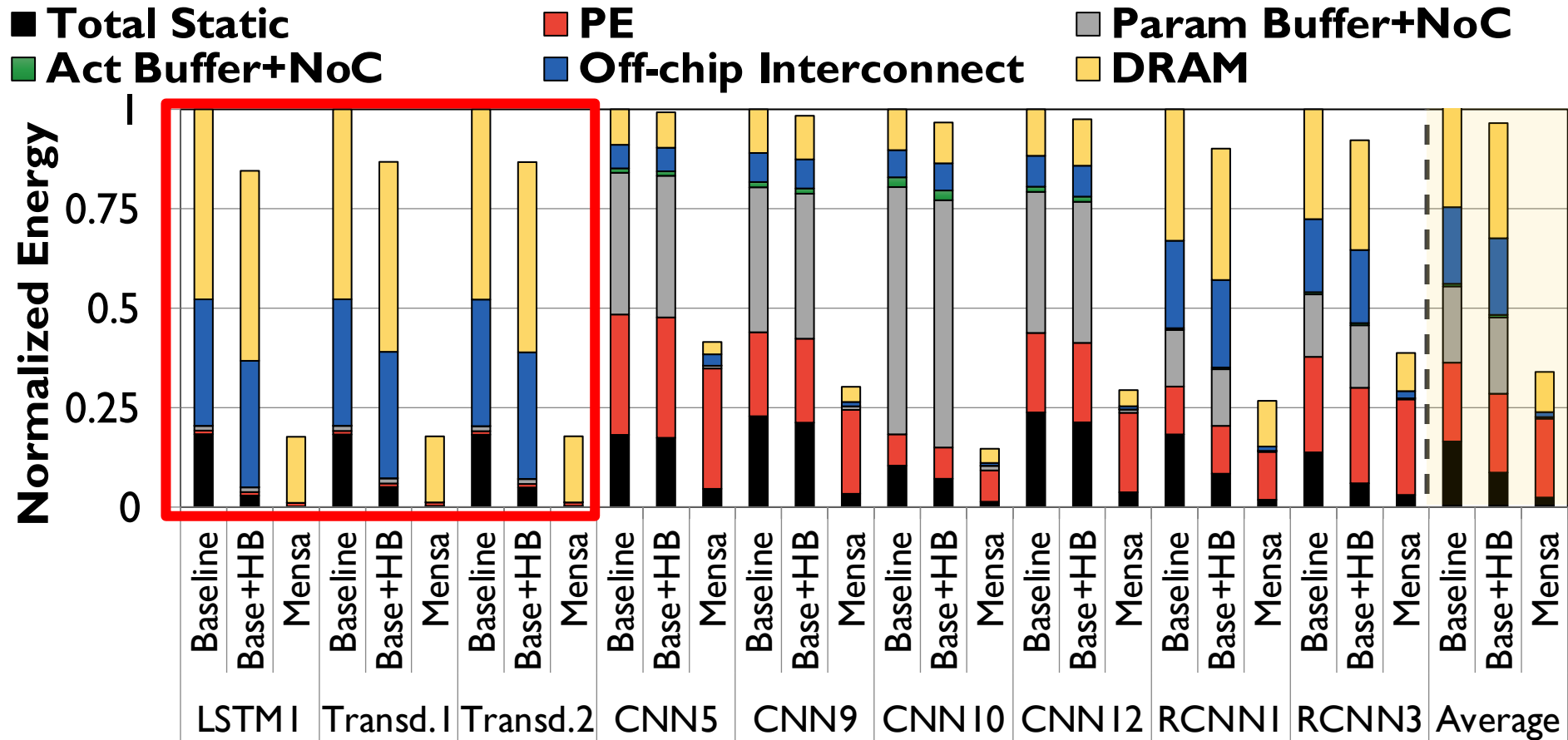
In LSTMs and Transducers used by Google, >90% energy spent on off-chip interconnect and DRAM

https://arxiv.org/pdf/2109.14320

SAFARI

We Do Not Want to Move Data!

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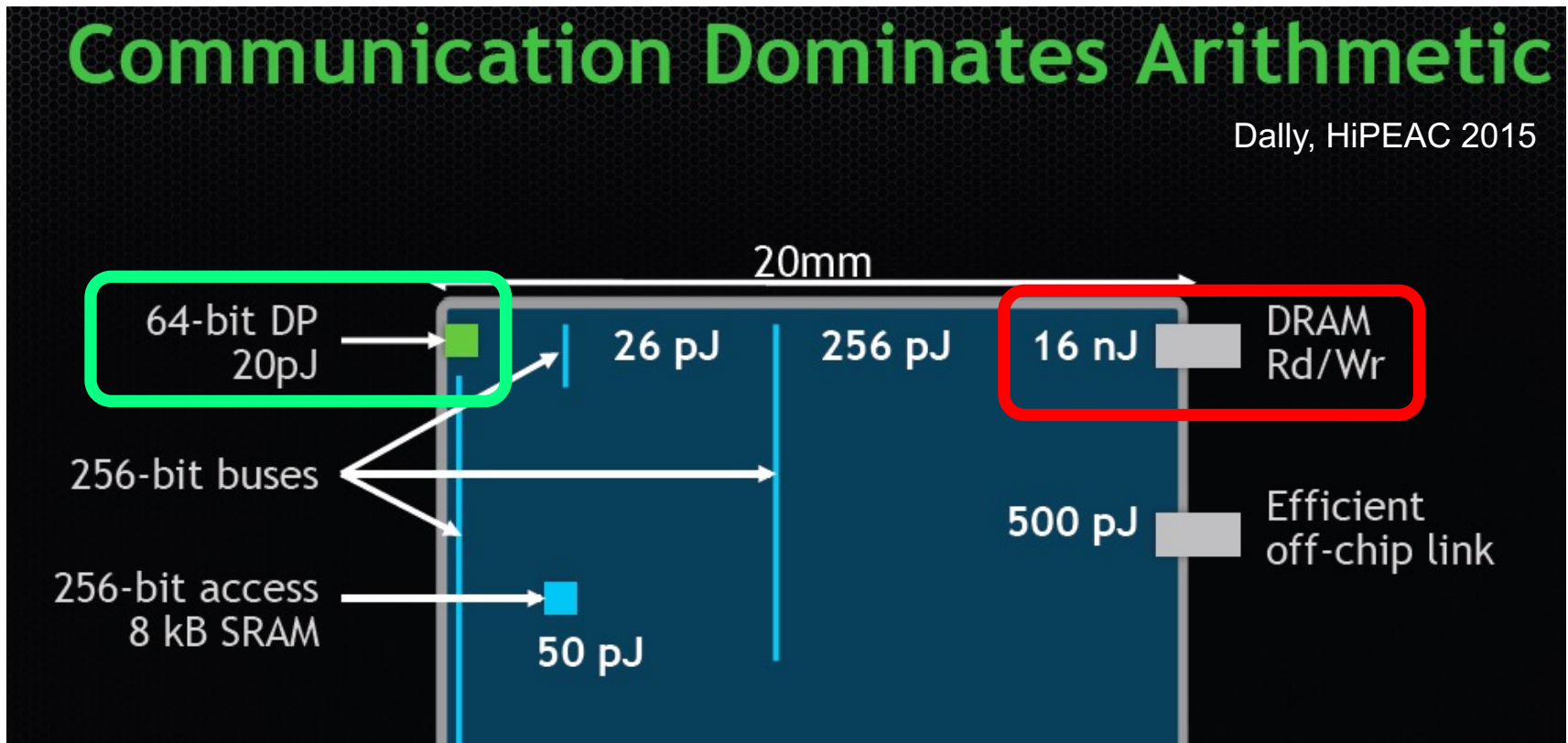
A memory access consumes ~100-1000X the energy of a complex addition

We Need A **Paradigm Shift** To …

- Enable computation with minimal data movement

- Compute where it makes sense (where data resides)

- Make computing architectures more data-centric

Process Data Where It Makes Sense

Sensors

Storage | **DRAM** | **A lot of SRAM** | **DRAM** | **Storage**
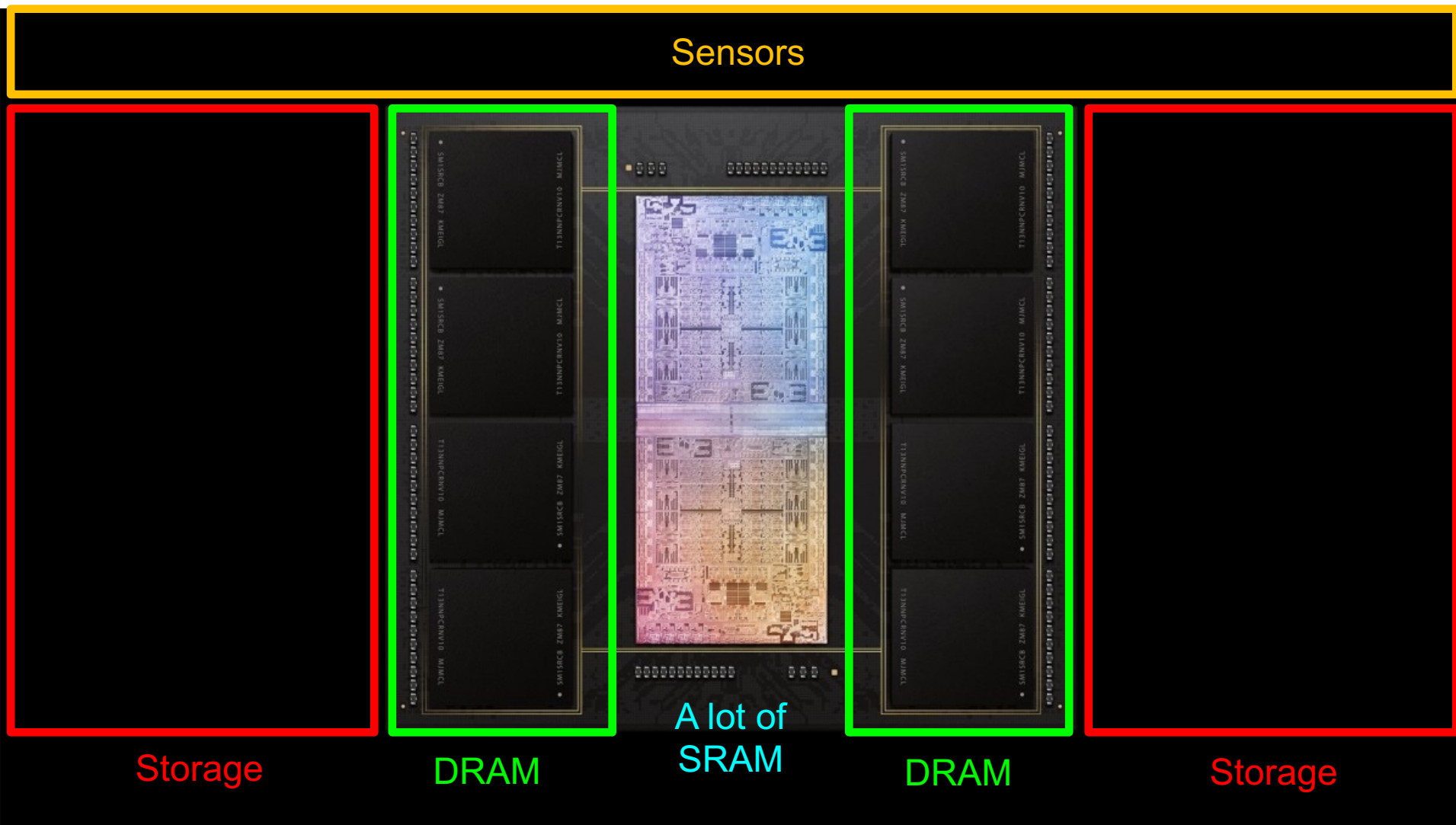
Apple M1 Ultra System (2022)

Goal: Processing Inside Memory/Storage
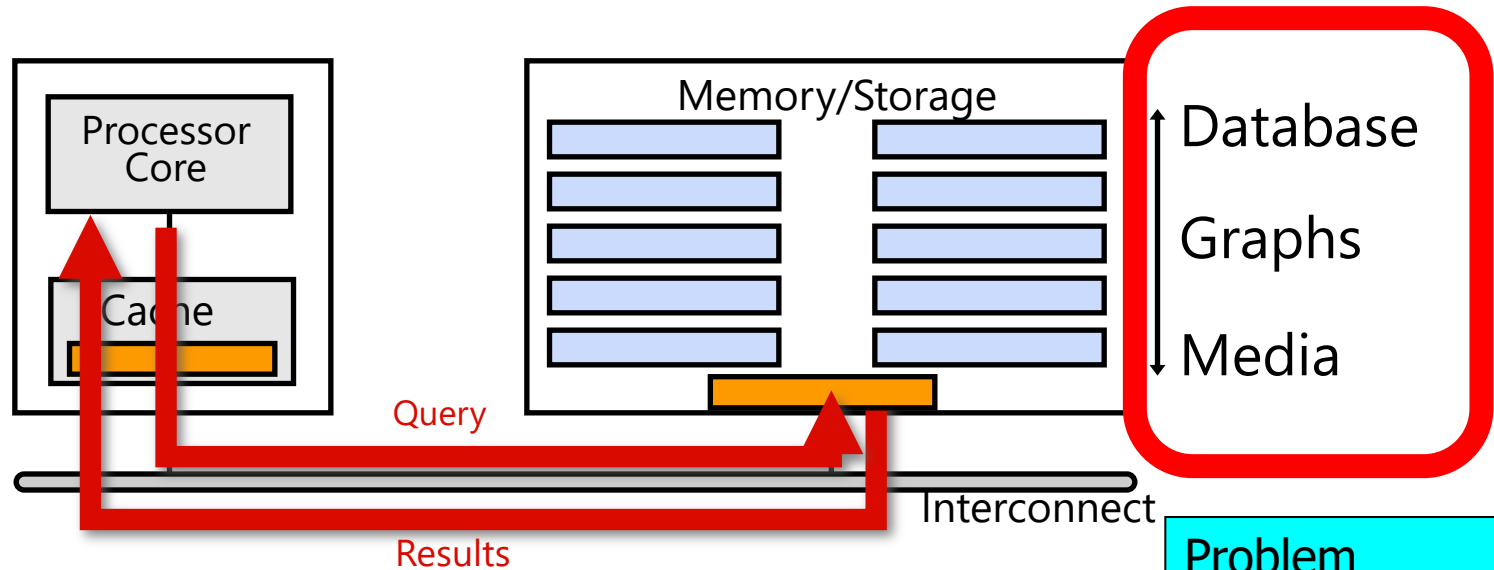
- Many questions … How do we design the:
  - compute-capable memory & controllers?
  - processors & communication units?
  - software & hardware interfaces?
  - system software, compilers, languages?
 - algorithms & theoretical foundations?

Major Trends Affecting Main Memory (IV)

- Need for main memory capacity, bandwidth, QoS increasing

- Main memory energy/power is a key system design concern

- DRAM technology scaling is ending
  - ITRS projects DRAM will not scale easily below X nm
  - Scaling has provided many benefits:
    - higher capacity (density), lower cost, lower energy
  - Difficiulties in scaling create robustness problems

SAFARI

An "Early" Position Paper [IMW'13]

- Onur Mutlu,
  "Memory Scaling: A Systems Architecture Perspective"
  Proceedings of the 5th International Memory Workshop (**IMW**), Monterey, CA, May 2013. Slides (pptx) (pdf)
  EETimes Reprint

Memory Scaling: A Systems Architecture Perspective

Onur Mutlu
Carnegie Mellon University
onur@cmu.edu
http://users.ece.cmu.edu/~omutlu/

https://people.inf.ethz.ch/omutlu/pub/memory-scaling_memcon13.pdf

The DRAM Scaling Problem

- DRAM stores charge in a capacitor (charge-based memory)
  - Capacitor must be large enough for reliable sensing
  - Access transistor should be large enough for low leakage and high retention time
  - Scaling beyond 40-35nm (2013) is challenging [ITRS, 2009]

- DRAM capacity, cost, and energy/power hard to scale

The DRAM Scaling Problem

- **DRAM stores charge in a capacitor (charge-based memory)**
  - Capacitor must be large enough for reliable sensing
  - Access transistor should be large enough for low leakage and high retention time
  - Scaling beyond 40-35nm (2013) is challenging [ITRS, 2009]

- DRAM **capacity, cost, and energy/power hard to scale**

The DRAM Scaling Problem

Cross-sectional SEM of DDR5* DRAM
*Double Data Rate 5 (1y nm generation)

https://www.youtube.com/watch?v=Bln-v9LmZ3E
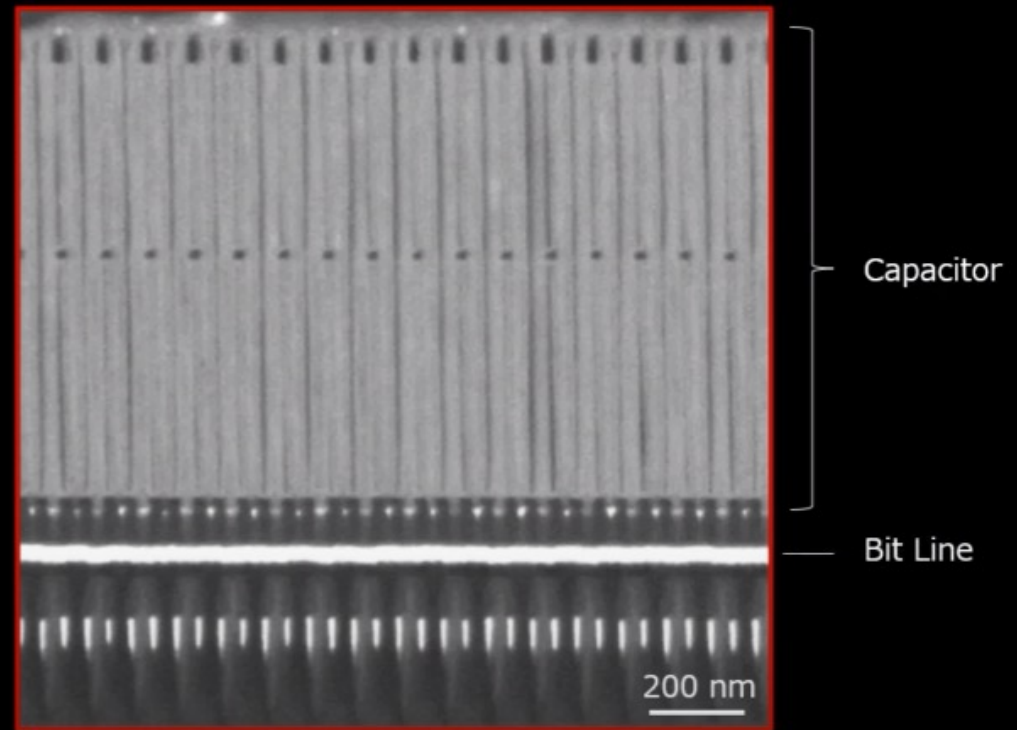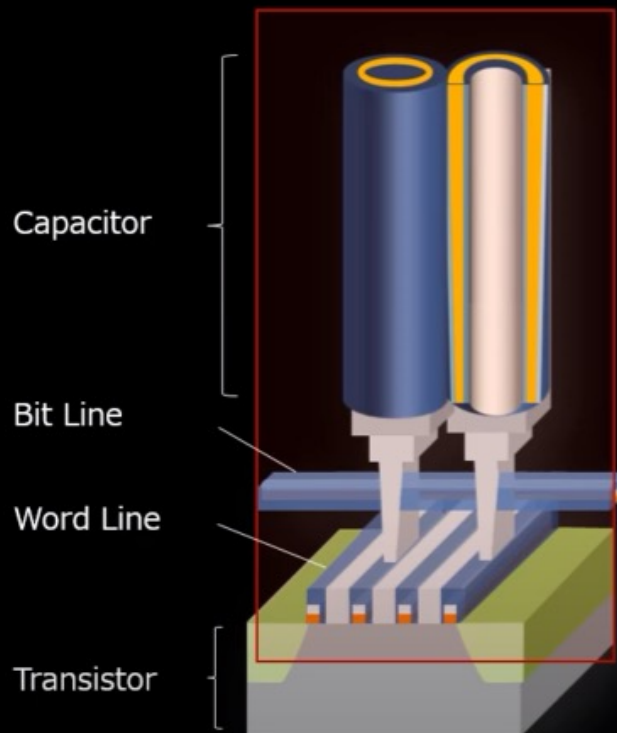
Limits of Charge Memory

- **Difficult charge placement and control**
  - Flash: floating gate charge
  - DRAM: capacitor charge, transistor leakage

- **Data retention and reliable sensing becomes difficult as charge storage unit size reduces**
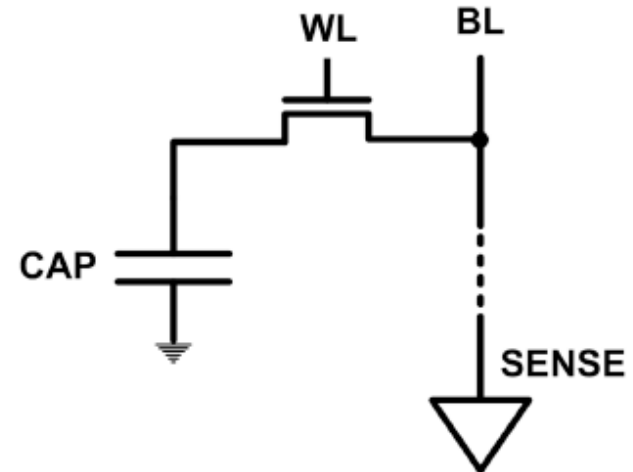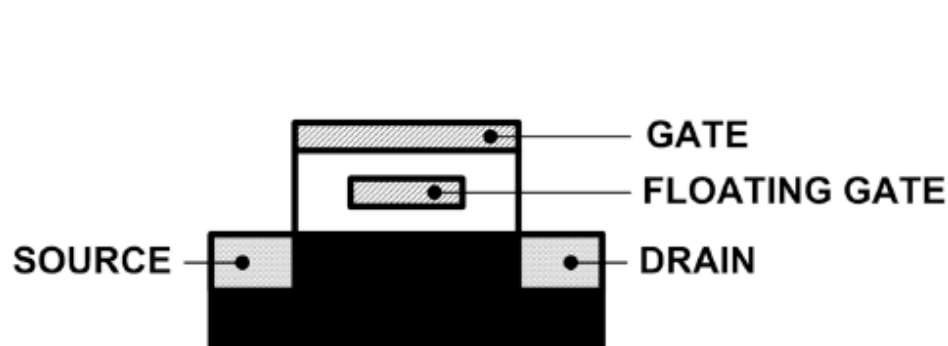
As Memory Scales, It Becomes Unreliable

- **Data from all of Facebook's servers worldwide**
- Meza+, "Revisiting Memory Errors in Large-Scale Production Data Centers," DSN'15.
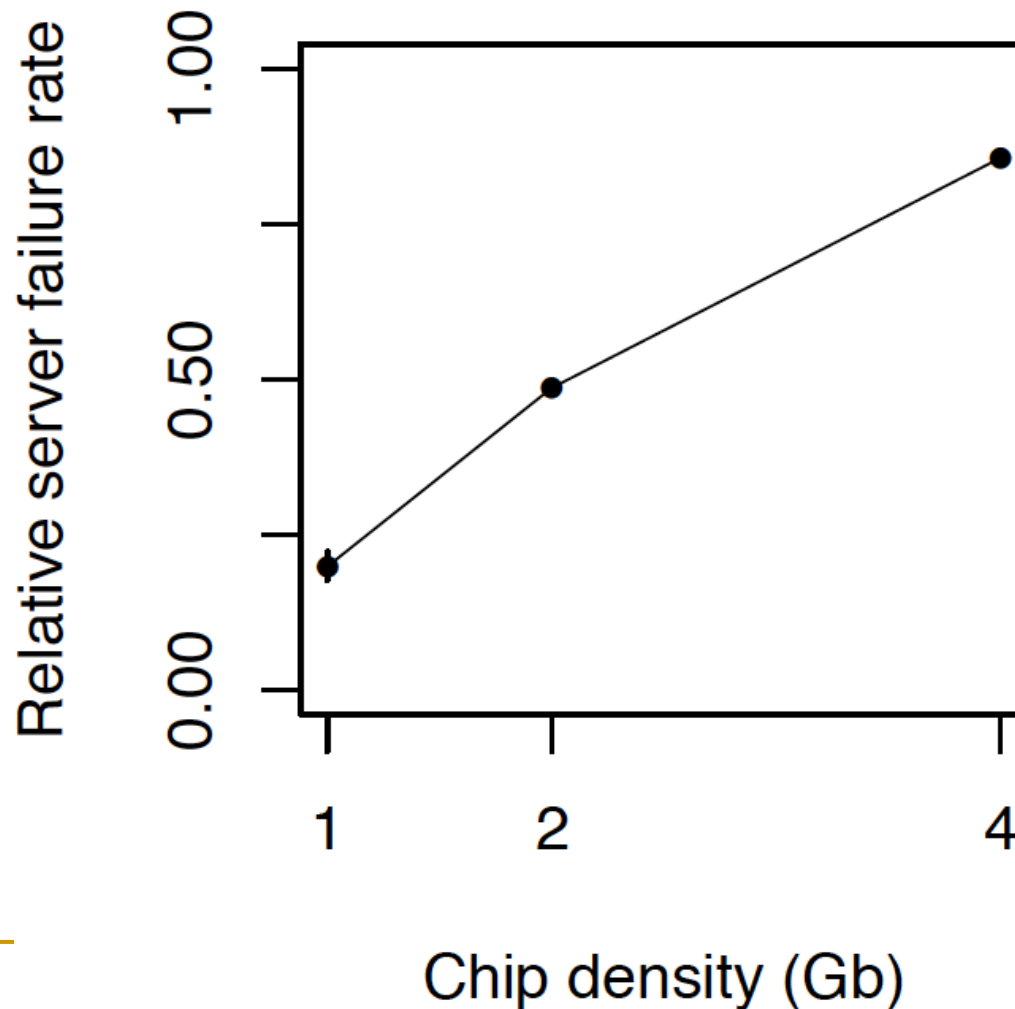
Intuition: quadratic increase in capacity

Large-Scale Failure Analysis of DRAM Chips

- Analysis and modeling of memory errors found in all of Facebook's server fleet

- Justin Meza, Qiang Wu, Sanjeev Kumar, and Onur Mutlu,
  "Revisiting Memory Errors in Large-Scale Production Data Centers: Analysis and Modeling of New Trends from the Field"
  Proceedings of the 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (**DSN**), Rio de Janeiro, Brazil, June 2015.
  [Slides (pptx) (pdf)] [DRAM Error Model]

Revisiting Memory Errors in Large-Scale Production Data Centers: Analysis and Modeling of New Trends from the Field

Justin Meza   Qiang Wu *   Sanjeev Kumar *   Onur Mutlu

Carnegie Mellon University   * Facebook, Inc.

Infrastructures to Understand Such Issues

An Experimental Study of Data Retention Behavior in Modern DRAM Devices: Implications for Retention Time Profiling Mechanisms (Liu et al., ISCA 2013)

The Efficacy of Error Mitigation Techniques for DRAM Retention Failures: A Comparative Experimental Study (Khan et al., SIGMETRICS 2014)

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

Adaptive-Latency DRAM: Optimizing DRAM Timing for the Common-Case (Lee et al., HPCA 2015)

AVATAR: A Variable-Retention-Time (VRT) Aware Refresh for DRAM Systems (Qureshi et al., DSN 2015)

SAFARI

Infrastructures to Understand Such Issues

Kim+, "Flipping Bits in Memory Without Accessing Them: An Experimental Study of DRAM Disturbance Errors," ISCA 2014.

SAFARI

186
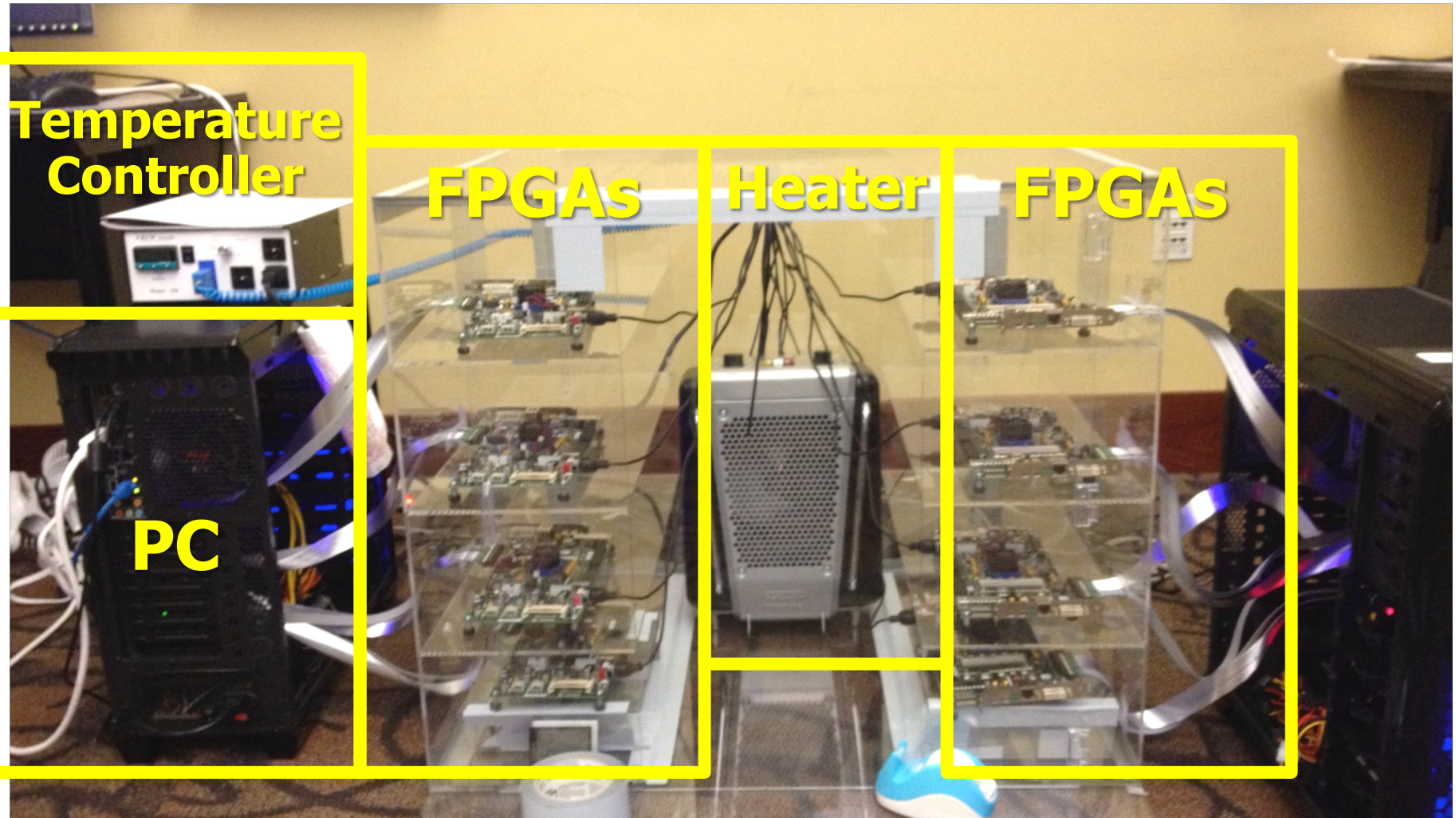
SoftMC: Open Source DRAM Infrastructure

- Hasan Hassan et al., "**SoftMC: A Flexible and Practical Open-Source Infrastructure for Enabling Experimental DRAM Studies**," HPCA 2017.

- **Flexible**
- **Easy to Use (C++ API)**
- **Open-source**

  github.com/CMU-SAFARI/SoftMC
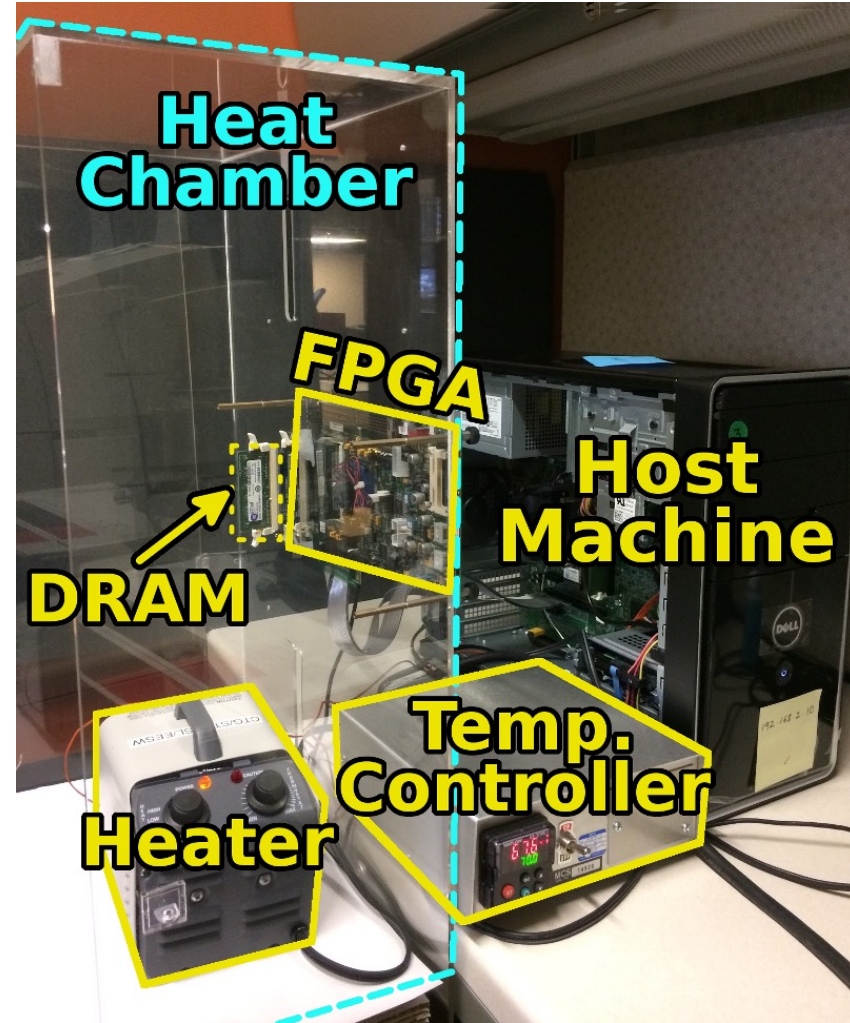
SoftMC: Open Source DRAM Infrastructure

- Hasan Hassan, Nandita Vijaykumar, Samira Khan, Saugata Ghose, Kevin Chang, Gennady Pekhimenko, Donghyuk Lee, Oguz Ergin, and Onur Mutlu,
"SoftMC: A Flexible and Practical Open-Source Infrastructure for Enabling Experimental DRAM Studies"
Proceedings of the 23rd International Symposium on High-Performance Computer Architecture (**HPCA**), Austin, TX, USA, February 2017.
[Slides (pptx) (pdf)] [Lightning Session Slides (pptx) (pdf)]
[Full Talk Lecture (39 minutes)]
[Source Code]

SoftMC: A Flexible and Practical Open-Source Infrastructure for Enabling Experimental DRAM Studies

Hasan Hassan[1,2,3]   Nandita Vijaykumar[3]   Samira Khan[4,3]   Saugata Ghose[3]   Kevin Chang[3]
Gennady Pekhimenko[5,3]   Donghyuk Lee[6,3]   Oguz Ergin[2]   Onur Mutlu[1,3]

[1]*ETH Zürich*   [2]*TOBB University of Economics & Technology*   [3]*Carnegie Mellon University*
[4]*University of Virginia*   [5]*Microsoft Research*   [6]*NVIDIA Research*

DRAM Bender

- Ataberk Olgun, Hasan Hassan, A Giray Yağlıkçı, Yahya Can Tuğrul, Lois Orosa, Haocong Luo, Minesh Patel, Oğuz Ergin, and Onur Mutlu,
"DRAM Bender: An Extensible and Versatile FPGA-based Infrastructure to Easily Test State-of-the-art DRAM Chips"
IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (**TCAD**), 2023.
[Extended arXiv version]
[DRAM Bender Source Code]
[DRAM Bender Tutorial Video (43 minutes)]

DRAM Bender: An Extensible and Versatile FPGA-based Infrastructure to Easily Test State-of-the-art DRAM Chips

Ataberk Olgun[§]   Hasan Hassan[§]   A. Giray Yağlıkçı[§]   Yahya Can Tuğrul[§†]
Lois Orosa[§⊙]   Haocong Luo[§]   Minesh Patel[§]   Oğuz Ergin[†]   Onur Mutlu[§]
[§]*ETH Zürich*   [†]*TOBB ETÜ*   [⊙]*Galician Supercomputing Center*

DRAM Bender: Prototypes

Testing Infrastructure	Protocol Support	FPGA Support
SoftMC [134]	DDR3	One Prototype
LiteX RowHammer Tester (LRT) [17]	DDR3/4, LPDDR4	Two Prototypes
DRAM Bender (this work)	**DDR3/DDR4**	**Five Prototypes**

Five out of the box FPGA-based prototypes

Xilinx Alveo U200
FPGA Board
(with DRAM Bender)

DRAM
Module

PCI-e Connection
to the Host Machine
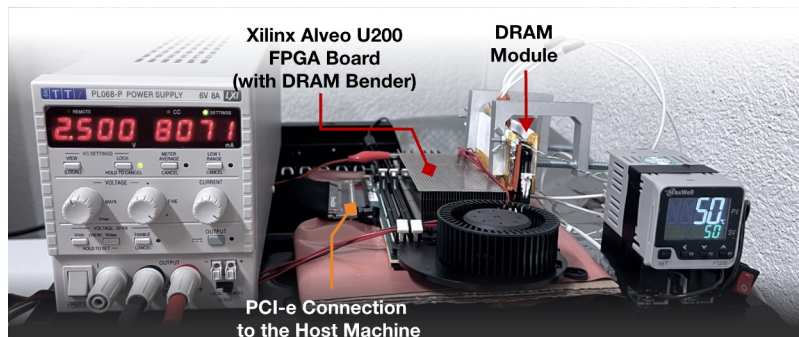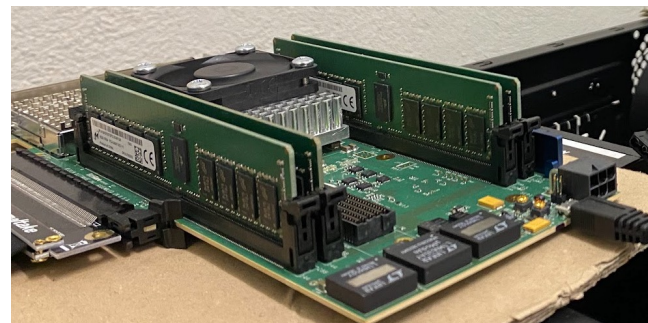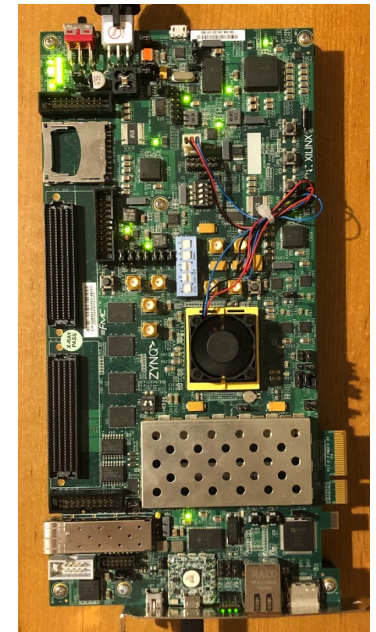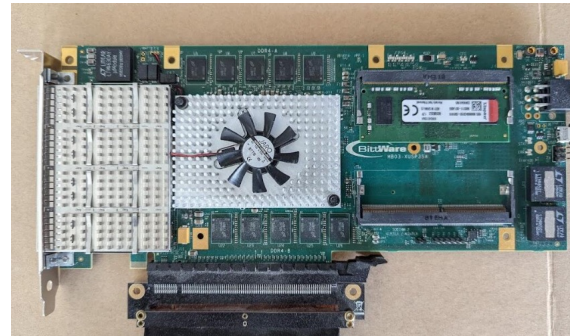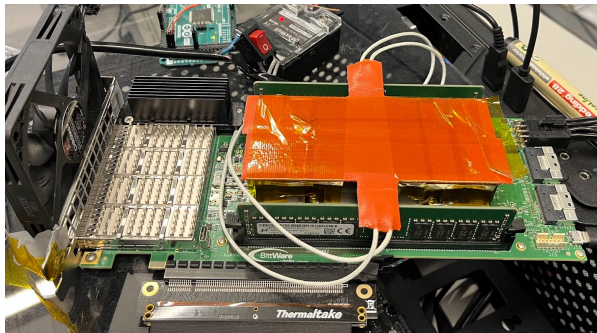
Data Retention in Memory [Liu et al., ISCA 2013]

- Retention Time Profile of DRAM looks like this:

64-128ms

>256ms

128-256ms

Location dependent
Stored value pattern dependent
Time dependent

Liu+, "RAIDR: Retention-Aware Intelligent DRAM Refresh," ISCA 2012.
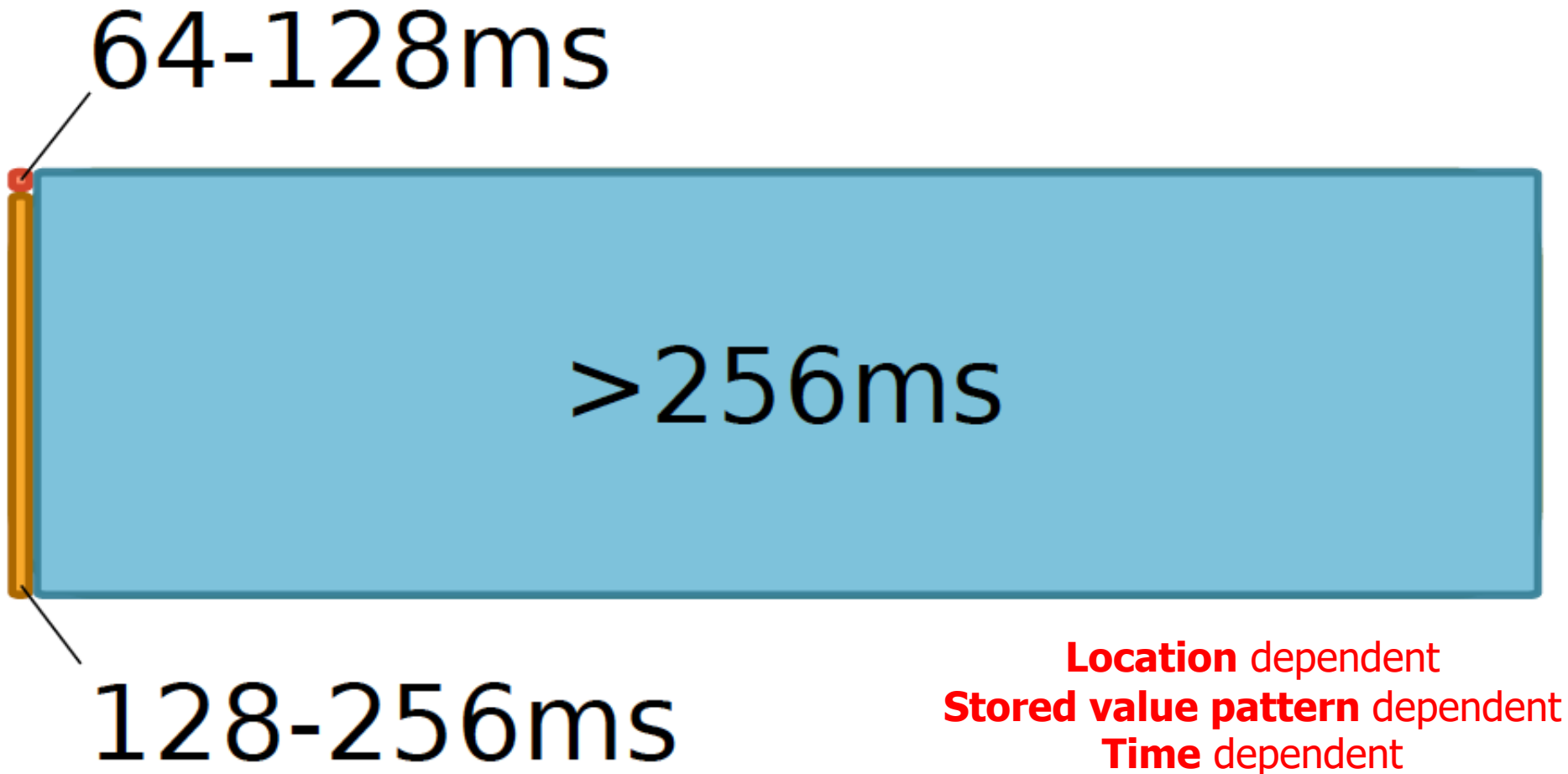
RAIDR: Heterogeneous Refresh [ISCA'12]

- Jamie Liu, Ben Jaiyen, Richard Veras, and Onur Mutlu,
  "RAIDR: Retention-Aware Intelligent DRAM Refresh"
  Proceedings of the 39th International Symposium on Computer Architecture (**ISCA**), Portland, OR, June 2012. Slides (pdf)
  [Invited Retrospective at 50 Years of ISCA, 2023 (pdf)]
  Selected to the ISCA-50 25-Year Retrospective Issue covering 1996-2020 in 2023 (Retrospective (pdf) Full Issue).

RAIDR: Retention-Aware Intelligent DRAM Refresh

Jamie Liu   Ben Jaiyen   Richard Veras   Onur Mutlu
Carnegie Mellon University

Analysis of Data Retention Failures [ISCA'13]

- Jamie Liu, Ben Jaiyen, Yoongu Kim, Chris Wilkerson, and Onur Mutlu,
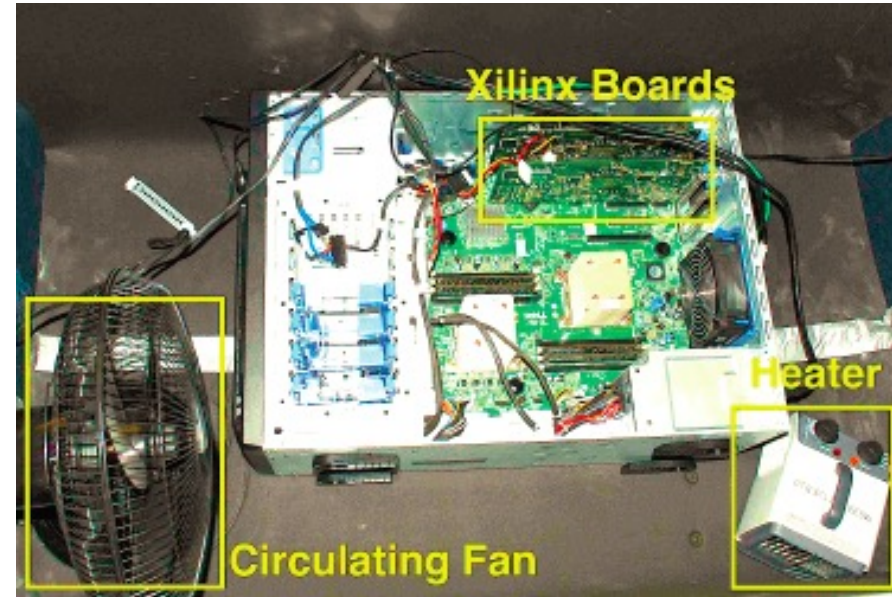 **"An Experimental Study of Data Retention Behavior in Modern DRAM Devices: Implications for Retention Time Profiling Mechanisms"**
  Proceedings of the 40th International Symposium on Computer Architecture (**ISCA**), Tel-Aviv, Israel, June 2013. Slides (ppt) Slides (pdf)
  [Invited Retrospective at 50 Years of ISCA, 2023 (pdf)]
  Selected to the ISCA-50 25-Year Retrospective Issue covering 1996-2020 in 2023 (Retrospective (pdf) Full Issue).

An Experimental Study of Data Retention Behavior in Modern DRAM Devices: Implications for Retention Time Profiling Mechanisms

Jamie Liu[*]
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213
jamiel@alumni.cmu.edu

Ben Jaiyen[*]
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213
bjaiyen@alumni.cmu.edu

Yoongu Kim
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213
yoonguk@ece.cmu.edu

Chris Wilkerson
Intel Corporation
2200 Mission College Blvd.
Santa Clara, CA 95054
chris.wilkerson@intel.com

Onur Mutlu
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213
onur@cmu.edu

SAF

Mitigation of Retention Issues [SIGMETRICS'14]

- Samira Khan, Donghyuk Lee, Yoongu Kim, Alaa Alameldeen, Chris Wilkerson, and Onur Mutlu,
  "The Efficacy of Error Mitigation Techniques for DRAM Retention Failures: A Comparative Experimental Study"
  Proceedings of the ACM International Conference on Measurement and Modeling of Computer Systems (**SIGMETRICS**), Austin, TX, June 2014. [Slides (pptx) (pdf)] [Poster (pptx) (pdf)] [Full data sets]

The Efficacy of Error Mitigation Techniques for DRAM Retention Failures: A Comparative Experimental Study

Samira Khan[†][*]
samirakhan@cmu.edu

Donghyuk Lee[†]
donghyuk1@cmu.edu

Yoongu Kim[†]
yoongukim@cmu.edu

Alaa R. Alameldeen[*]
alaa.r.alameldeen@intel.com

Chris Wilkerson[*]
chris.wilkerson@intel.com

Onur Mutlu[†]
onur@cmu.edu

[†]Carnegie Mellon University   [*]Intel Labs

Mitigation of Retention Issues [DSN'15]

- Moinuddin Qureshi, Dae Hyun Kim, Samira Khan, Prashant Nair, and Onur Mutlu,
  "AVATAR: A Variable-Retention-Time (VRT) Aware Refresh for DRAM Systems"
  Proceedings of the 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks *(**DSN**)*, Rio de Janeiro, Brazil, June 2015.
  [Slides (pptx) (pdf)]

AVATAR: A Variable-Retention-Time (VRT) Aware Refresh for DRAM Systems

Moinuddin K. Qureshi[†]   Dae-Hyun Kim[†]   Samira Khan[‡]   Prashant J. Nair[†]   Onur Mutlu[‡]
[†]Georgia Institute of Technology   [‡]Carnegie Mellon University
{moin, dhkim, pnair6}@ece.gatech.edu   {samirakhan, onur}@cmu.edu

Mitigation of Retention Issues [DSN'16]

- Samira Khan, Donghyuk Lee, and Onur Mutlu,
  "PARBOR: An Efficient System-Level Technique to Detect Data-Dependent Failures in DRAM"
  Proceedings of the 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (**DSN**), Toulouse, France, June 2016.
  [Slides (pptx) (pdf)]

PARBOR: An Efficient System-Level Technique to Detect Data-Dependent Failures in DRAM

Samira Khan[*]　　Donghyuk Lee[†‡]　　Onur Mutlu[*†]
[*]University of Virginia　　[†]Carnegie Mellon University　　[‡]Nvidia　　[*]ETH Zürich

Mitigation of Retention Issues [MICRO'17]

- Samira Khan, Chris Wilkerson, Zhe Wang, Alaa R. Alameldeen, Donghyuk Lee, and Onur Mutlu,
"Detecting and Mitigating Data-Dependent DRAM Failures by Exploiting Current Memory Content"
Proceedings of the 50th International Symposium on Microarchitecture (**MICRO**), Boston, MA, USA, October 2017.
[Slides (pptx) (pdf)] [Lightning Session Slides (pptx) (pdf)] [Poster (pptx) (pdf)]

Detecting and Mitigating Data-Dependent DRAM Failures by Exploiting Current Memory Content

Samira Khan[*] Chris Wilkerson[†] Zhe Wang[†] Alaa R. Alameldeen[†] Donghyuk Lee[‡] Onur Mutlu[*]

[*]University of Virginia   [†]Intel Labs   [‡]Nvidia Research   [*]ETH Zürich

Mitigation of Retention Issues [ISCA'17]

- Minesh Patel, Jeremie S. Kim, and Onur Mutlu,
  "The Reach Profiler (REAPER): Enabling the Mitigation of DRAM Retention Failures via Profiling at Aggressive Conditions"
  Proceedings of the 44th International Symposium on Computer Architecture (**ISCA**), Toronto, Canada, June 2017.
  [Slides (pptx) (pdf)]
  [Lightning Session Slides (pptx) (pdf)]

- First experimental analysis of (mobile) LPDDR4 chips
- Analyzes the complex tradeoff space of retention time profiling
- Idea: enable fast and robust profiling at higher refresh intervals & temperatures

The Reach Profiler (REAPER): Enabling the Mitigation of DRAM Retention Failures via Profiling at Aggressive Conditions

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

Mitigation of Retention Issues [DSN'19]

- Minesh Patel, Jeremie S. Kim, Hasan Hassan, and Onur Mutlu,
  "Understanding and Modeling On-Die Error Correction in Modern DRAM: An Experimental Study Using Real Devices"
  Proceedings of the 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (**DSN**), Portland, OR, USA, June 2019.
  [Source Code for EINSim, the Error Inference Simulator]
  Best paper award.

Understanding and Modeling On-Die Error Correction in Modern DRAM: An Experimental Study Using Real Devices

Minesh Patel[†]   Jeremie S. Kim[‡†]   Hasan Hassan[†]   Onur Mutlu[†‡]

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

Mitigation of Retention Issues [MICRO'20]

- Minesh Patel, Jeremie S. Kim, Taha Shahroodi, Hasan Hassan, and Onur Mutlu,
"Bit-Exact ECC Recovery (BEER): Determining DRAM On-Die ECC Functions by Exploiting DRAM Data Retention Characteristics"
*Proceedings of the 53rd International Symposium on Microarchitecture (**MICRO**)*, Virtual, October 2020.
[Slides (pptx) (pdf)]
[Lightning Talk Slides (pptx) (pdf)]
[Talk Video (15 minutes)]
[Lightning Talk Video (1.5 minutes)]
Best paper award.

Bit-Exact ECC Recovery (BEER): Determining DRAM On-Die ECC Functions by Exploiting DRAM Data Retention Characteristics

Minesh Patel[†]   Jeremie S. Kim[‡†]   Taha Shahroodi[†]   Hasan Hassan[†]   Onur Mutlu[†‡]

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

Mitigation of Retention Issues [MICRO'21]

- Minesh Patel, Geraldo F. de Oliveira Jr., and Onur Mutlu,
  "HARP: Practically and Effectively Identifying Uncorrectable Errors in Memory Chips That Use On-Die Error-Correcting Codes"
  Proceedings of the 54th International Symposium on Microarchitecture (**MICRO**),
  Virtual, October 2021.
  [Slides (pptx) (pdf)]
  [Short Talk Slides (pptx) (pdf)]
  [Lightning Talk Slides (pptx) (pdf)]
  [Talk Video (20 minutes)]
  [Lightning Talk Video (1.5 minutes)]
  [HARP Source Code (Officially Artifact Evaluated with All Badges)]

HARP: Practically and Effectively
Identifying Uncorrectable Errors in Memory Chips
That Use On-Die Error-Correcting Codes

Minesh Patel   Geraldo F. Oliveira   Onur Mutlu
ETH Zürich   ETH Zürich   ETH Zürich

More on DRAM Refresh & Data Retention
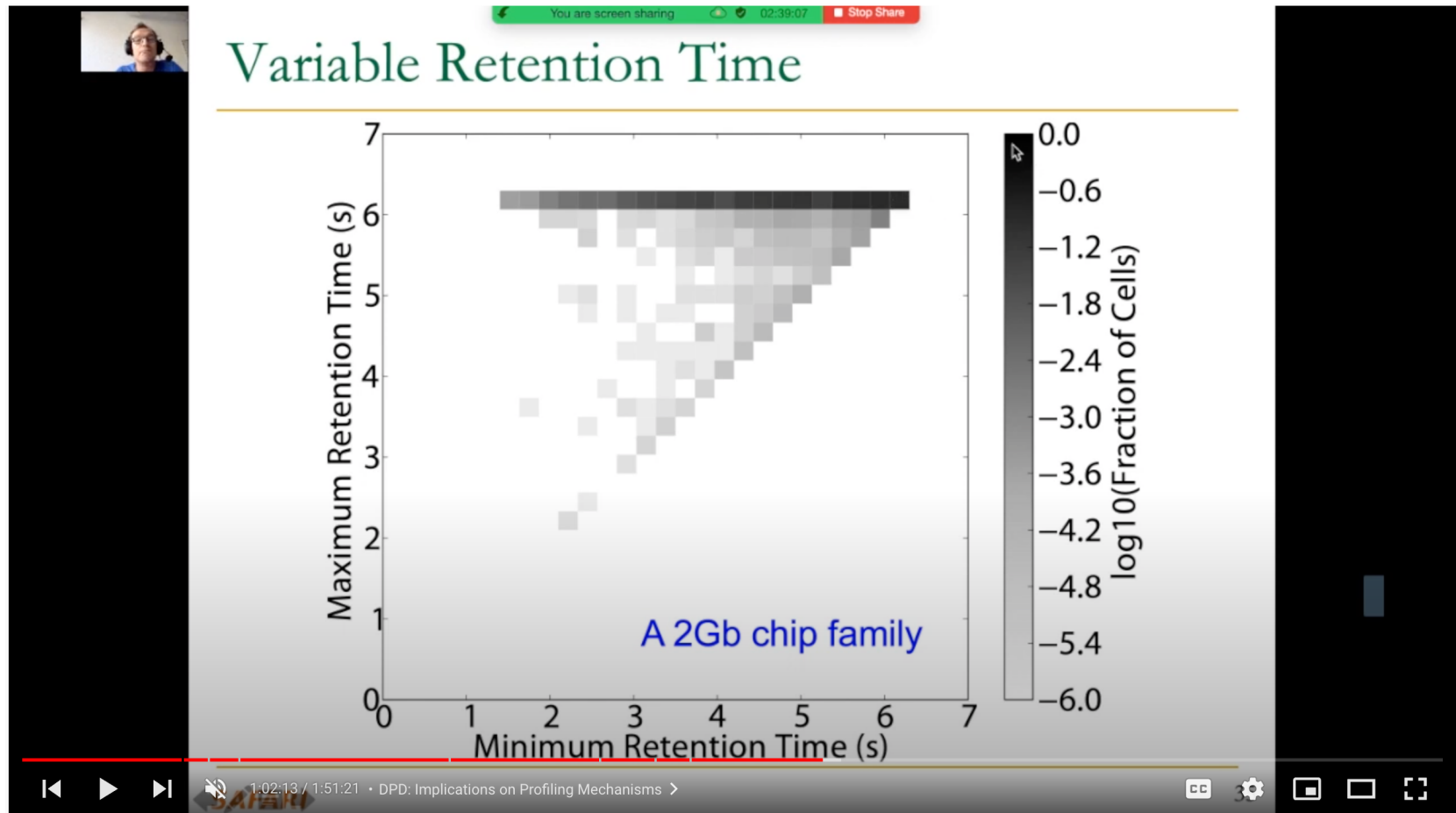
https://www.youtube.com/watch?v=v702wUnaWGE&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN&index=3

A Curious Phenomenon

One can
predictably induce errors
in DRAM memory chips

Kim+, "Flipping Bits in Memory Without Accessing Them: An Experimental Study of DRAM Disturbance Errors," ISCA 2014.

An Example: The RowHammer Problem

- One can **predictably induce bit flips** in commodity DRAM chips
  - All recent DRAM chips are fundamentally 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

FORGET SOFTWARE—NOW HACKERS ARE EXPLOITING PHYSICS

SHARE

f　SHARE 18276

🐦　TWEET

First RowHammer Analysis

- 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] [Lecture Video (1 hr 49 mins), 25 September 2020]
  One of the 7 papers of 2012-2017 selected as Top Picks in Hardware and Embedded Security for IEEE TCAD (link). Selected to the ISCA-50 25-Year Retrospective Issue covering 1996-2020 in 2023 (Retrospective (pdf) Full Issue). Winner of the 2024 IFIP Jean-Claude Laprie Award in dependable computing (link).

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

The Robustness Perspective (I)

- Onur Mutlu,
  "The RowHammer Problem and Other Issues We May Face as Memory Becomes Denser"
  Invited Paper in Proceedings of the Design, Automation, and Test in Europe Conference (**DATE**), Lausanne, Switzerland, March 2017.
  [Slides (pptx) (pdf)]

The RowHammer Problem
and Other Issues We May Face as Memory Becomes Denser

Onur Mutlu
ETH Zürich
onur.mutlu@inf.ethz.ch
https://people.inf.ethz.ch/omutlu

The Robustness Perspective (II)

- 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 (1 hr 15 minutes, with Q&A)]

RowHammer: A Retrospective

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

Major Trends Affecting Main Memory (V)

- DRAM scaling has already become very difficult
    - Increasing cell leakage current, reduced cell reliability, increasing manufacturing difficulties [Kim+ ISCA 2014], [Liu+ ISCA 2013], [Mutlu IMW 2013], [Mutlu DATE 2017]
    - **Difficult to significantly improve capacity, energy**

- **Emerging memory technologies** are promising

Major Trends Affecting Main Memory (V)

- DRAM scaling has already become very difficult
  - Increasing cell leakage current, reduced cell reliability, increasing manufacturing difficulties [Kim+ ISCA 2014], [Liu+ ISCA 2013], [Mutlu IMW 2013], [Mutlu DATE 2017]
  - **Difficult to significantly improve capacity, energy**

- **Emerging memory technologies** are promising

3D-Stacked DRAM	higher bandwidth	smaller capacity
Reduced-Latency DRAM (e.g., RL/TL-DRAM, FLY-RAM)	lower latency	higher cost
Low-Power DRAM (e.g., LPDDR3, LPDDR4, Voltron)	lower power	higher latency higher cost
Non-Volatile Memory (NVM) (e.g., PCM, STTRAM, ReRAM, 3D Xpoint)	larger capacity	higher latency higher dynamic power lower endurance

Major Trend: Hybrid Main Memory
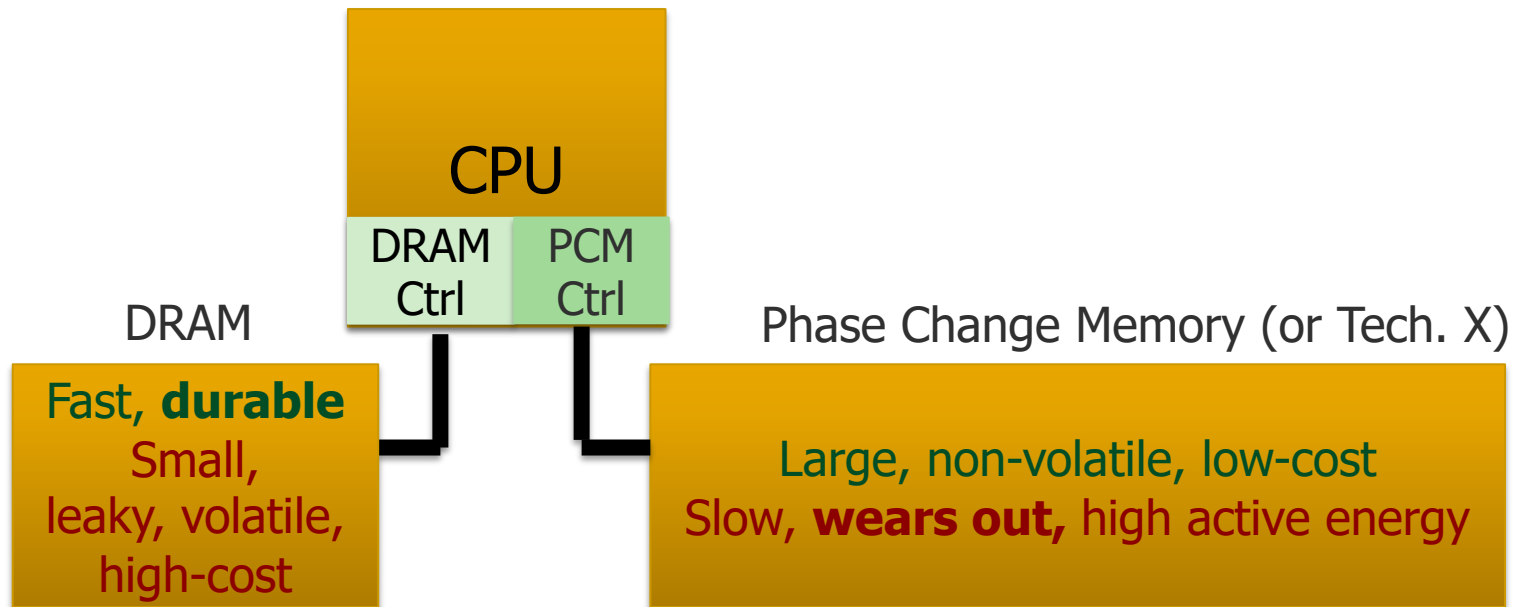
CPU

DRAM Ctrl | PCM Ctrl

DRAM

Phase Change Memory (or Tech. X)

Fast, **durable**
Small, leaky, volatile, high-cost

Large, non-volatile, low-cost
Slow, **wears out,** high active energy

Hardware/software manage data allocation and movement
to achieve the best of multiple technologies

Meza+, "Enabling Efficient and Scalable Hybrid Memories," IEEE Comp. Arch. Letters, 2012.
Yoon+, "Row Buffer Locality Aware Caching Policies for Hybrid Memories," ICCD 2012 Best Paper Award.

One Foreshadowing

Main Memory Needs
Intelligent Controllers

SAFARI

Industry Is Writing Papers About It, Too

DRAM Process Scaling Challenges

❖ **Refresh**
- Difficult to build high-aspect ratio cell capacitors decreasing cell capacitance
- Leakage current of cell access transistors increasing

❖ **tWR**
- Contact resistance between the cell capacitor and access transistor increasing
- On-current of the cell access transistor decreasing
- Bit-line resistance increasing

❖ **VRT**
- Occurring more frequently with cell capacitance decreasing

Refresh   tWR   VRT

Call for Intelligent Memory Controllers

DRAM Process Scaling Challenges

❖ **Refresh**

• Difficult to build high-aspect ratio cell capacitors decreasing cell capacitance

THE MEMORY FORUM 2014

Co-Architecting Controllers and DRAM to Enhance DRAM Process Scaling

Uksong Kang, Hak-soo Yu, Churoo Park, *Hongzhong Zheng,
**John Halbert, **Kuljit Bains, SeongJin Jang, and Joo Sun Choi

*Samsung Electronics, Hwasung, Korea / *Samsung Electronics, San Jose / **Intel*

Refresh   **tWR**   **VRT**
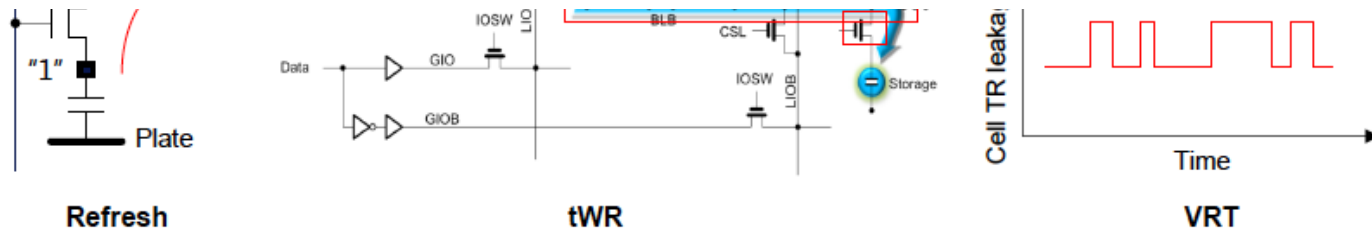
We Covered Until Here in Lecture 1

Memory Systems and Memory-Centric Computing

Lecture 1: Trends, Challenges, Opportunities

Onur Mutlu

omutlu@gmail.com

https://people.inf.ethz.ch/omutlu

15 July 2024

HiPEAC ACACES Summer School 2024

SAFARI　　　**ETH** zürich

Backup Slides
(To Be Covered in Lecture 2)

An Orthogonal Issue: Memory Interference
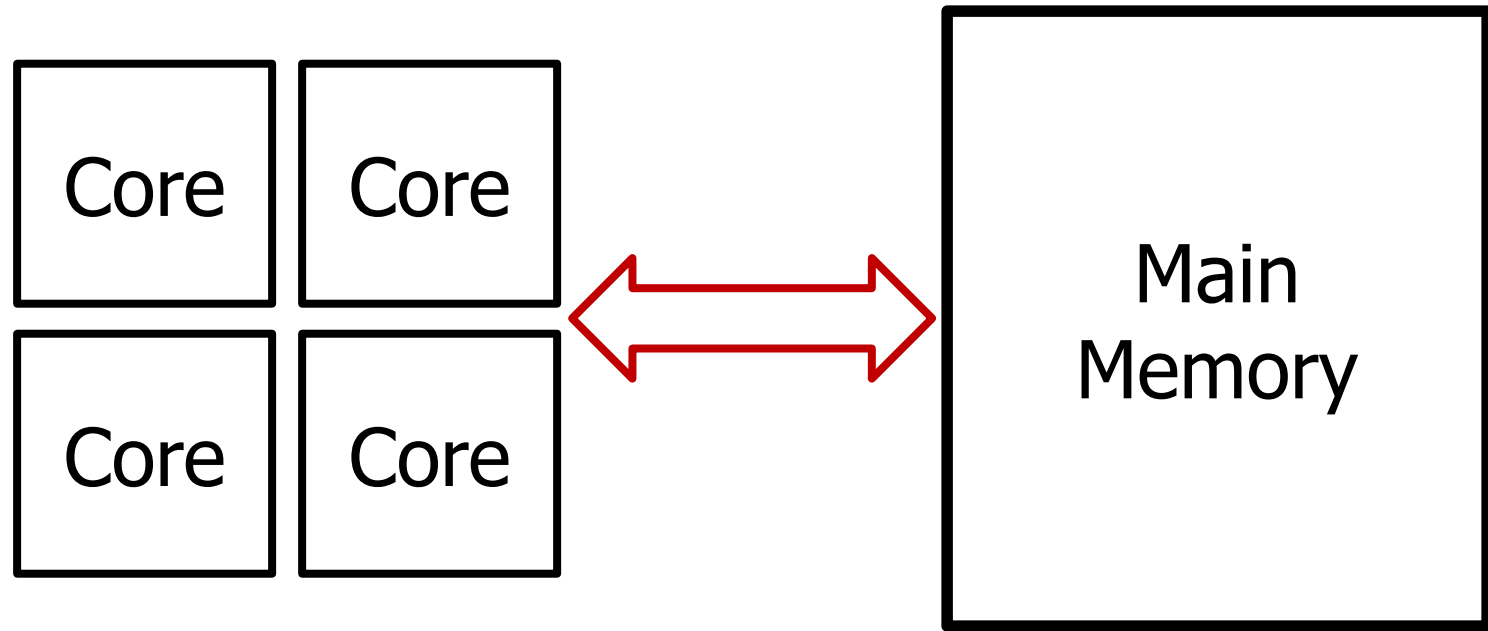
Cores' interfere with each other when accessing shared main memory
Uncontrolled interference leads to many problems (QoS, performance)

Goal: Predictable Performance in Complex Systems
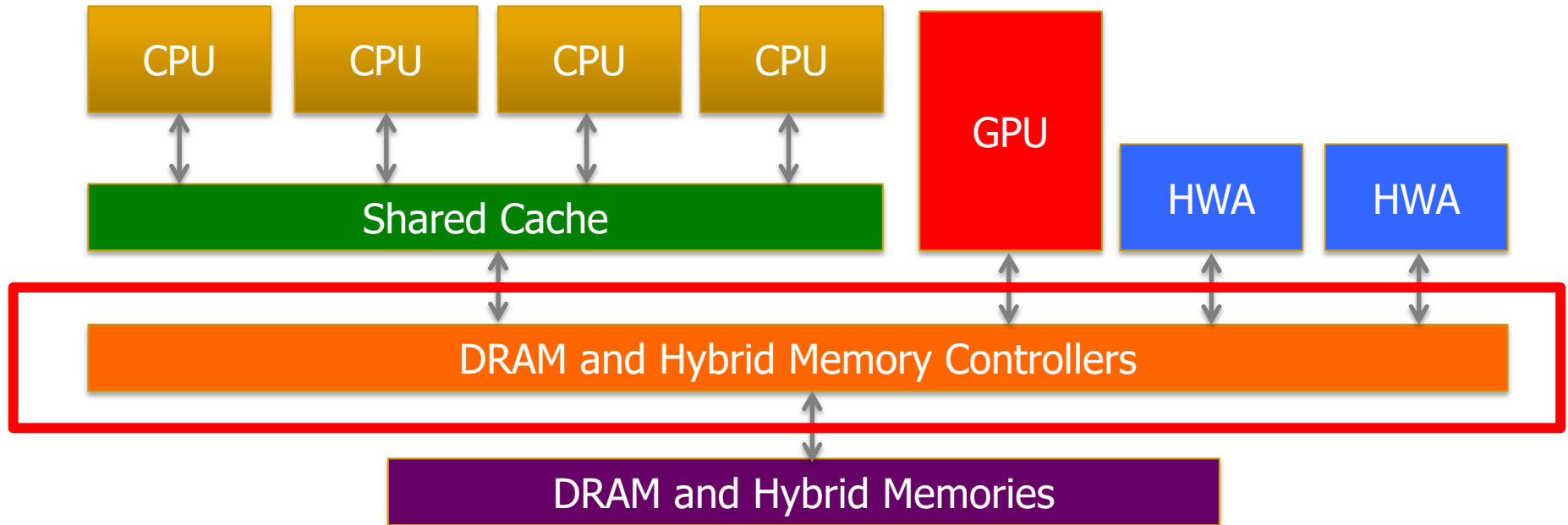
- Heterogeneous agents: CPUs, GPUs, and HWAs
- Main memory interference between CPUs, GPUs, HWAs

How to allocate resources to heterogeneous agents
to mitigate interference and provide predictable performance?

Many goals, many constraints, many metrics ...

The Future

Memory Controllers

are critical to research

They will become

even more important

Memory Control w/ Machine Learning [ISCA'08]

- Engin Ipek, Onur Mutlu, José F. Martínez, and Rich Caruana,
  "Self Optimizing Memory Controllers: A Reinforcement Learning Approach"
  Proceedings of the 35th International Symposium on Computer Architecture
  (**ISCA**), pages 39-50, Beijing, China, June 2008. Slides (pptx)
  Selected to the ISCA-50 25-Year Retrospective Issue covering 1996-2020 in 2023 (Retrospective (pdf) Full Issue).

Self-Optimizing Memory Controllers: A Reinforcement Learning Approach

Engin İpek[1,2]   Onur Mutlu[2]   José F. Martínez[1]   Rich Caruana[1]

[1]Cornell University, Ithaca, NY 14850 USA
[2] Microsoft Research, Redmond, WA 98052 USA

Solving the Memory Problem

How Do We Solve **The Memory Problem?**

- **Fix it**: Make memory and controllers more intelligent
  - New interfaces, functions, architectures: system-mem codesign

- **Eliminate or minimize it**: Replace or (more likely) augment DRAM with a different technology
  - New technologies and system-wide rethinking of memory & storage

- **Embrace it**: Design heterogeneous memories (none of which are perfect) and map data intelligently across them
  - New models for data management and maybe usage

- ...

How Do We Solve **The Memory Problem?**

- Fix it: Make memory and controllers more intelligent
  - New interfaces, functions, architectures: system-mem codesign

- Eliminate or minimize it: Replace or (more likely) augment DRAM with a different technology
  - New technologies and system-wide rethinking of memory & storage

- Embrace it: Design heterogeneous memories (none of which are perfect) and map data intelligently across them
  - New models for data management and maybe usage

Solutions (to memory scaling) require software/hardware/device cooperation

How Do We Solve **The Memory Problem?**

- Fix it: Make memory and controllers more intelligent
  - New interfaces, functions, architectures: system-mem codesign

- Eliminate or minimize it: Replace or (more likely) augment DRAM with a different technology
  - New technologies and system-wide rethinking of memory & storage

- Embrace it: Design heterogeneous memories (none of which are perfect) and map data intelligently across them
  - New models for data management and maybe usage

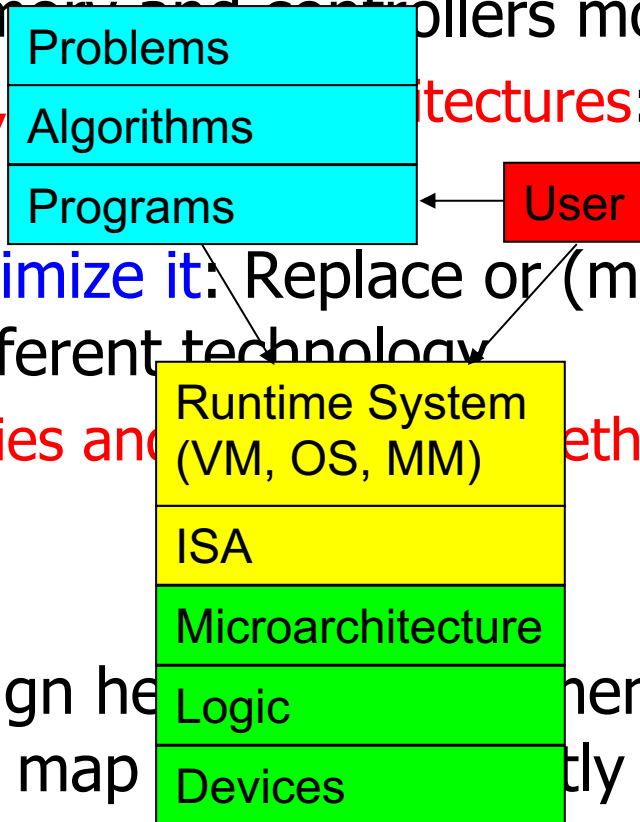| Problems |
|---|
| Algorithms |
| Programs |

User

Runtime System (VM, OS, MM)
ISA
Microarchitecture
Logic
Devices

Solutions (to memory scaling) require software/hardware/device cooperation

Solution 1: New Memory Architectures

- Overcome memory shortcomings with
  - Memory-centric system design
  - Novel memory architectures, interfaces, functions
  - Better waste management (efficient utilization)

- Key issues to tackle
  - Enable reliability at low cost → high capacity
  - Reduce energy
  - Reduce latency
  - Improve bandwidth
  - Reduce waste (capacity, bandwidth, latency)
  - Enable computation close to data

SAFARI

Solution 2: Emerging Memory Technologies

- Some emerging **resistive** memory technologies seem more scalable than DRAM (and they are non-volatile)

- Example: Phase Change Memory
  - Data stored by changing phase of material
  - Data read by detecting material's resistance
  - Expected to scale to 9nm (2022 [ITRS 2009])
  - Prototyped at 20nm (Raoux+, IBM JRD 2008)
 - Expected to be denser than DRAM: can store multiple bits/cell
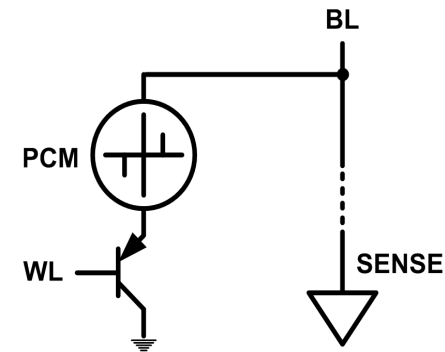
- But, emerging technologies have (many) shortcomings
  - Can they be enabled to replace/augment/surpass DRAM?

Solution 2: Emerging Memory Technologies

- Lee+, "Architecting Phase Change Memory as a Scalable DRAM Alternative," ISCA'09, CACM'10, IEEE Micro'10.
- Meza+, "Enabling Efficient and Scalable Hybrid Memories," IEEE Comp. Arch. Letters 2012.
- Yoon, Meza+, "Row Buffer Locality Aware Caching Policies for Hybrid Memories," ICCD 2012.
- Kultursay+, "Evaluating STT-RAM as an Energy-Efficient Main Memory Alternative," ISPASS 2013.
- Meza+, "A Case for Efficient Hardware-Software Cooperative Management of Storage and Memory," WEED 2013.
- Lu+, "Loose Ordering Consistency for Persistent Memory," ICCD 2014.
- Zhao+, "FIRM: Fair and High-Performance Memory Control for Persistent Memory Systems," MICRO 2014.
- Yoon, Meza+, "Efficient Data Mapping and Buffering Techniques for Multi-Level Cell Phase-Change Memories," TACO 2014.
- Ren+, "ThyNVM: Enabling Software-Transparent Crash Consistency in Persistent Memory Systems," MICRO 2015.
- Chauhan+, "NVMove: Helping Programmers Move to Byte-Based Persistence," INFLOW 2016.
- Li+, "Utility-Based Hybrid Memory Management," CLUSTER 2017.
- Yu+, "Banshee: Bandwidth-Efficient DRAM Caching via Software/Hardware Cooperation," MICRO 2017.
- Tavakkol+, "MQSim: A Framework for Enabling Realistic Studies of Modern Multi-Queue SSD Devices," FAST 2018.
- Tavakkol+, "FLIN: Enabling Fairness and Enhancing Performance in Modern NVMe Solid State Drives," ISCA 2018.
- Sadrosadati+. "LTRF: Enabling High-Capacity Register Files for GPUs via Hardware/Software Cooperative Register Prefetching," ASPLOS 2018.
- Salkhordeh+, "An Analytical Model for Performance and Lifetime Estimation of Hybrid DRAM-NVM Main Memories," TC 2019.
- Wang+, "Panthera: Holistic Memory Management for Big Data Processing over Hybrid Memories," PLDI 2019.
- Song+, "Enabling and Exploiting Partition-Level Parallelism (PALP) in Phase Change Memories," CASES 2019.
- Liu+, "Binary Star: Coordinated Reliability in Heterogeneous Memory Systems for High Performance and Scalability," MICRO'19.
- Song+, "Improving Phase Change Memory Performance with Data Content Aware Access," ISMM 2020.
- Yavits+, "WoLFRaM: Enhancing Wear-Leveling and Fault Tolerance in Resistive Memories using Programmable Address Decoders," ICCD 2020.
- Song+, "Aging-Aware Request Scheduling for Non-Volatile Main Memory," ASP-DAC 2021.

SAFARI

Combination: Hybrid Memory Systems
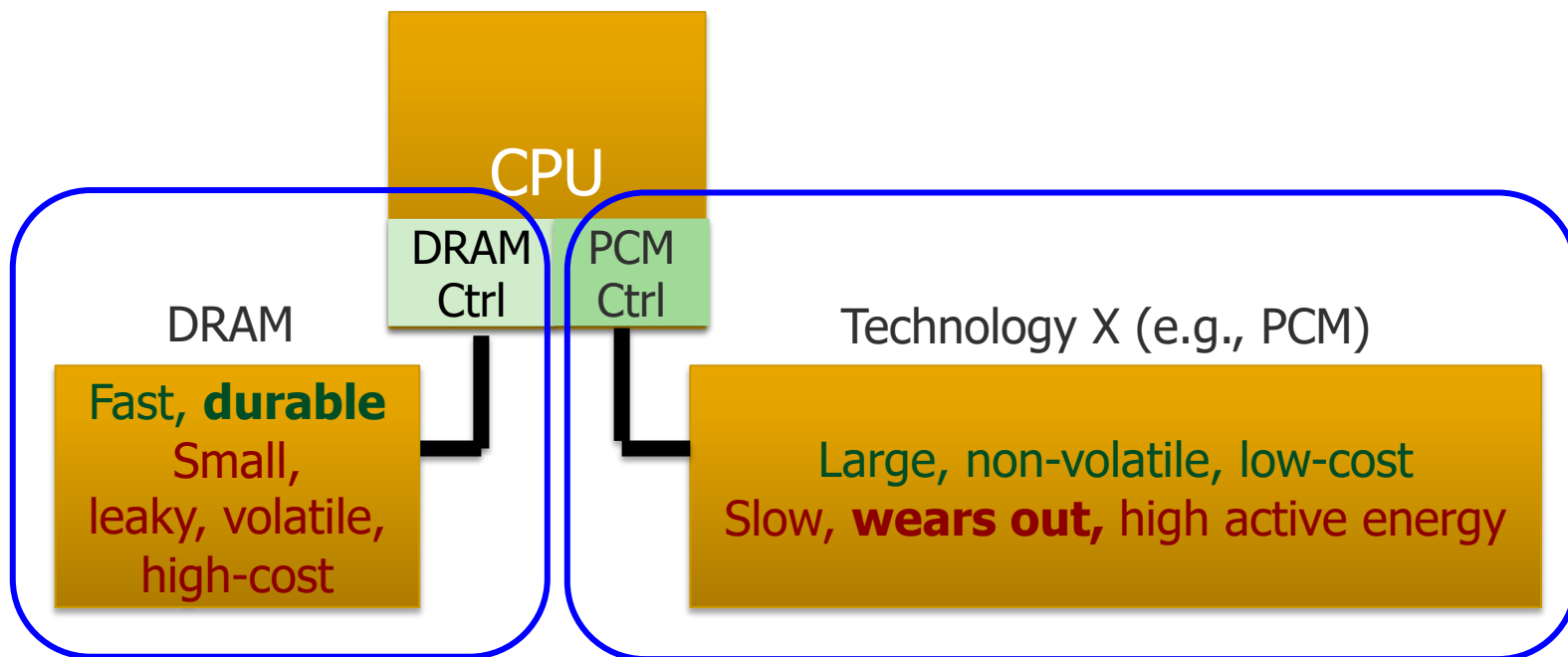
Hardware/software manage data allocation and movement
to achieve the best of multiple technologies

Meza+, "Enabling Efficient and Scalable Hybrid Memories," IEEE Comp. Arch. Letters, 2012.
Yoon, Meza et al., "Row Buffer Locality Aware Caching Policies for Hybrid Memories," ICCD 2012 Best Paper Award.

Vulnerable data

Tolerant data

Reliable memory

Low-cost memory
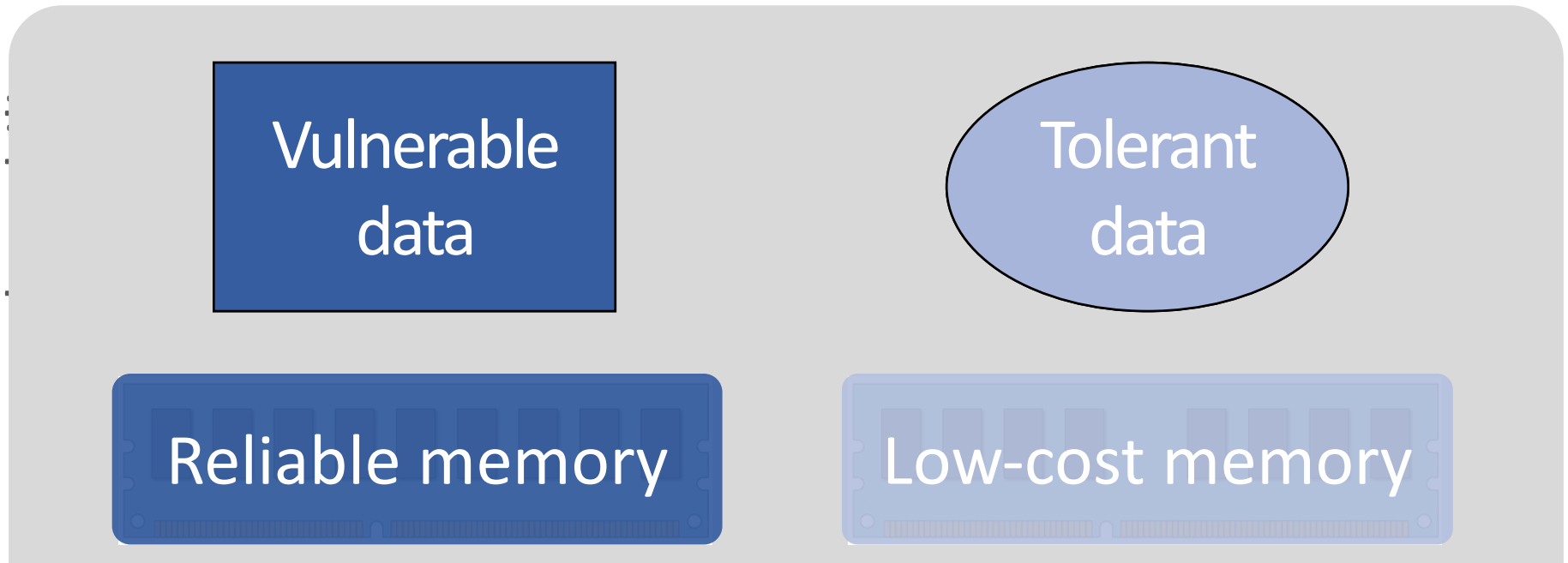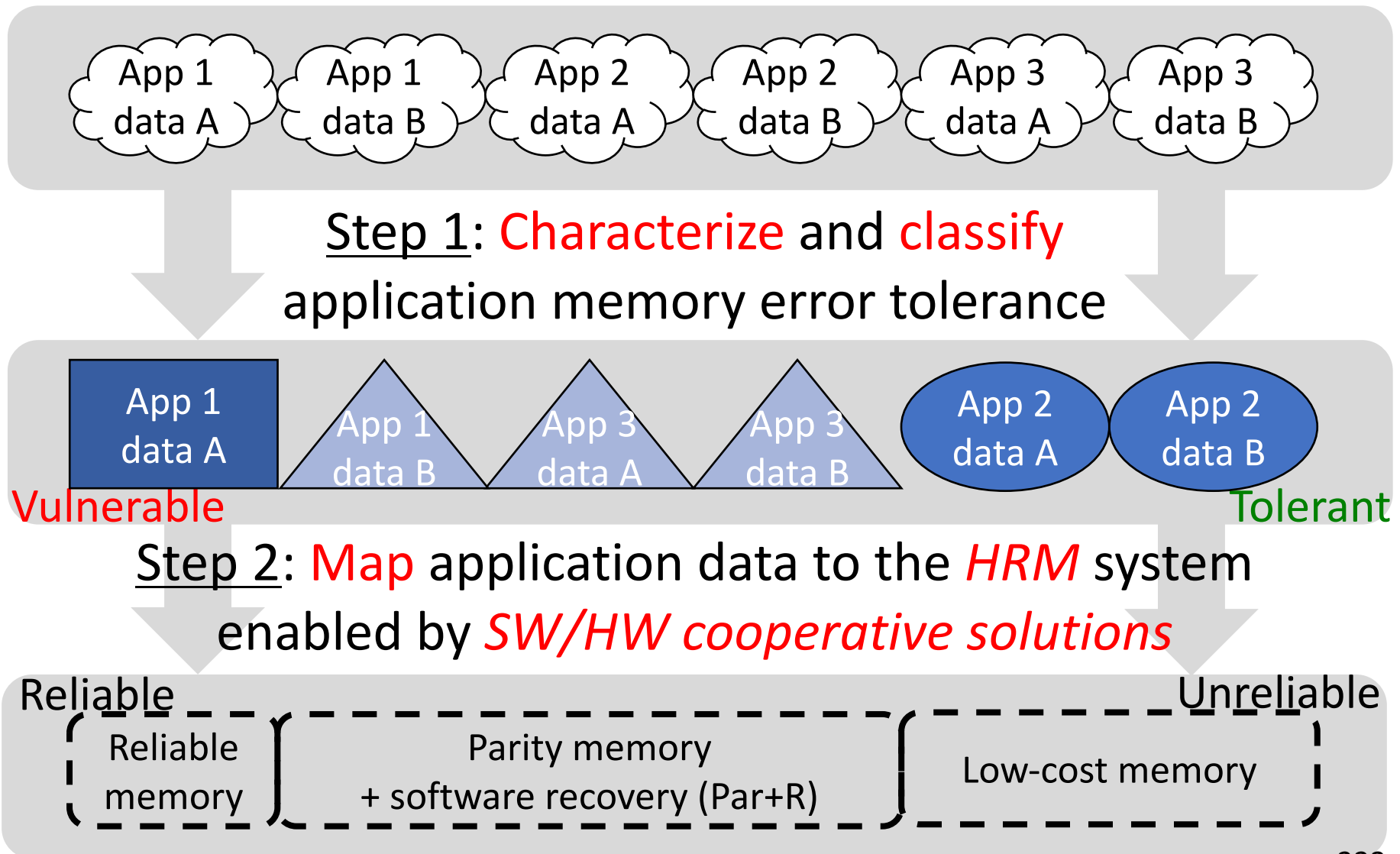
On Microsoft's Web Search workload

Reduces server hardware cost by 4.7 %

Achieves single server availability target of 99.90 %

Heterogeneous-**R**eliability **M**emory **[DSN 2014]**

Heterogeneous-Reliability Memory

App 1 data A   App 1 data B   App 2 data A   App 2 data B   App 3 data A   App 3 data B

Step 1: Characterize and classify application memory error tolerance

App 1 data A   App 1 data B   App 3 data A   App 3 data B   App 2 data A   App 2 data B

Vulnerable   Tolerant

Step 2: Map application data to the *HRM* system enabled by *SW/HW cooperative solutions*

Reliable   Unreliable

Reliable memory   Parity memory + software recovery (Par+R)   Low-cost memory

Evaluation Results

Legend:
- 🔵 ***Typical Server***
- 🔵 ***Consumer PC***
- 🟢 ***HRM***
- 🟡 ***Less-Tested (L)***
- 🔴 ***HRM/L***

Crashes/server/month

Server HW cost savings (%)

Memory cost savings (%)

Inner is worse — *Outer is better*
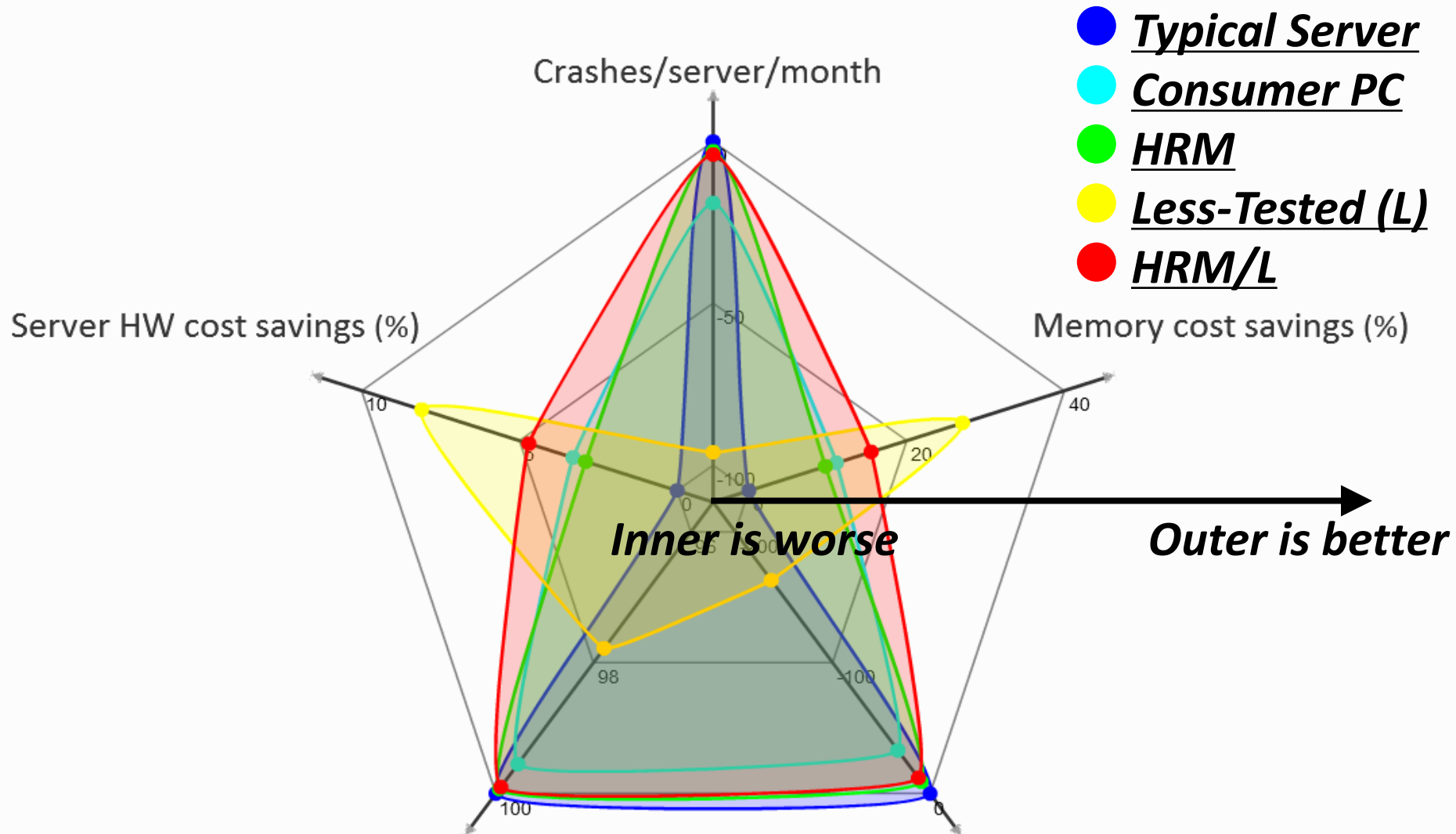
🟢 🔴 Bigger area means better tradeoff

More on Heterogeneous Reliability Memory

- Yixin Luo, Sriram Govindan, Bikash Sharma, Mark Santaniello, Justin Meza, Aman Kansal, Jie Liu, Badriddine Khessib, Kushagra Vaid, and Onur Mutlu,
"Characterizing Application Memory Error Vulnerability to Optimize Data Center Cost via Heterogeneous-Reliability Memory"
Proceedings of the 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (**DSN**), Atlanta, GA, June 2014. [Summary] [Slides (pptx) (pdf)] [Coverage on ZDNet]

Characterizing Application Memory Error Vulnerability to Optimize Datacenter Cost via Heterogeneous-Reliability Memory

Yixin Luo   Sriram Govindan[*]   Bikash Sharma[*]   Mark Santaniello[*]   Justin Meza
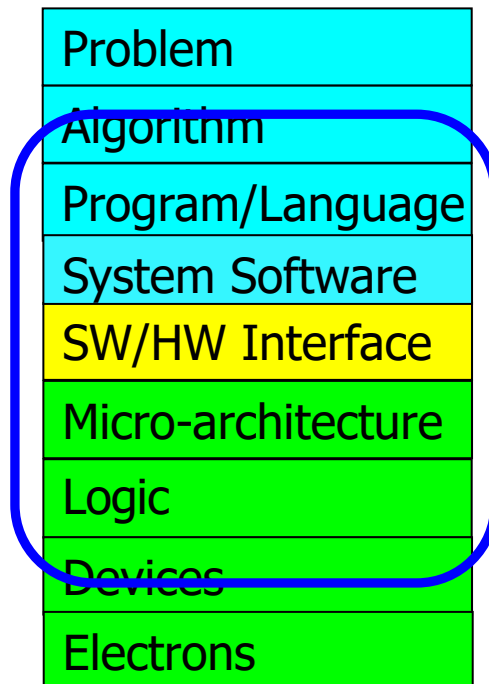Aman Kansal[*]   Jie Liu[*]   Badriddine Khessib[*]   Kushagra Vaid[*]   Onur Mutlu
Carnegie Mellon University, yixinluo@cs.cmu.edu, {meza, onur}@cmu.edu
[*]Microsoft Corporation, {srgovin, bsharma, marksan, kansal, jie.liu, bkhessib, kvaid}@microsoft.com

HRM is an Example of Our 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**

Another Example: EDEN for DNNs

- Deep Neural Network evaluation is very DRAM-intensive (especially for large networks)

1. Some data and layers in DNNs are very tolerant to errors

2. Reduce DRAM latency and voltage on such data and layers

3. While still achieving a user-specified DNN accuracy target by making training DRAM-error-aware

Data-aware management of DRAM latency and voltage for Deep Neural Network Inference

Example DNN Data Type to DRAM Mapping
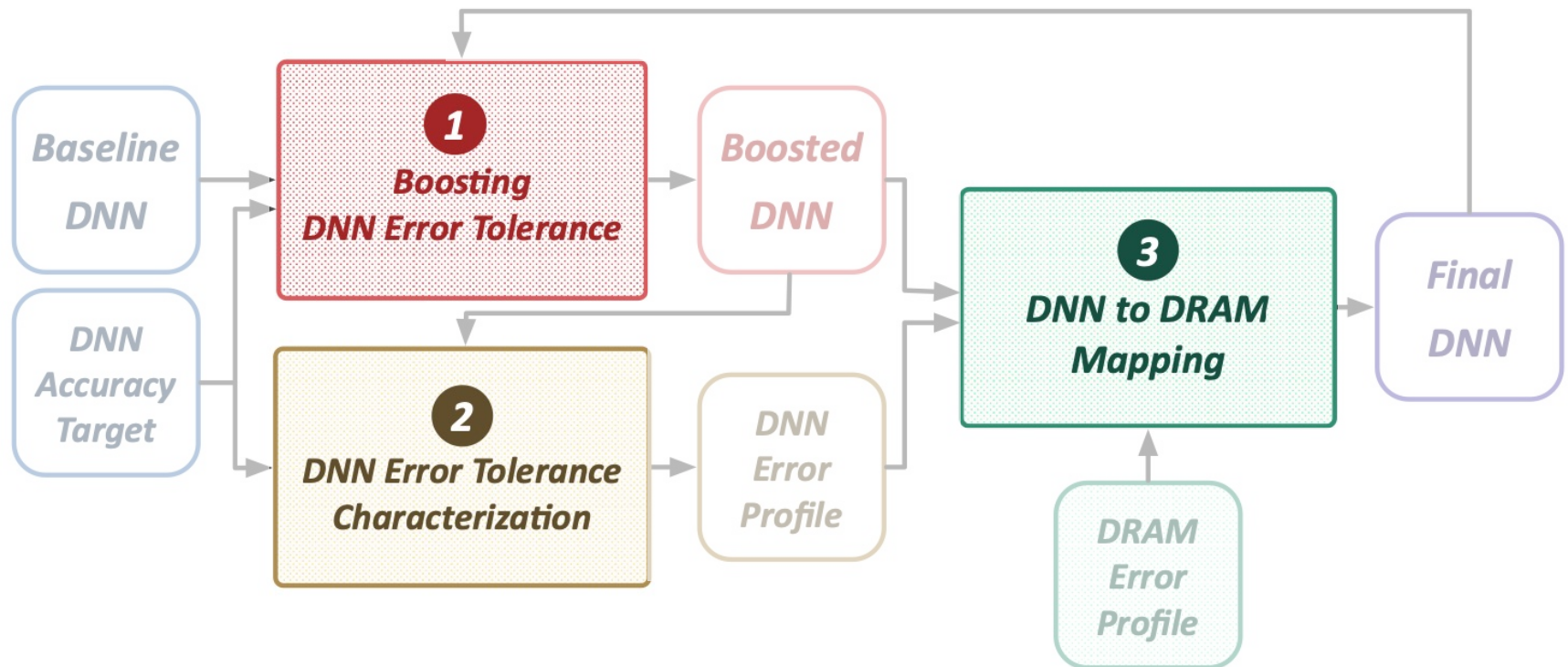
Mapping example of ResNet-50:
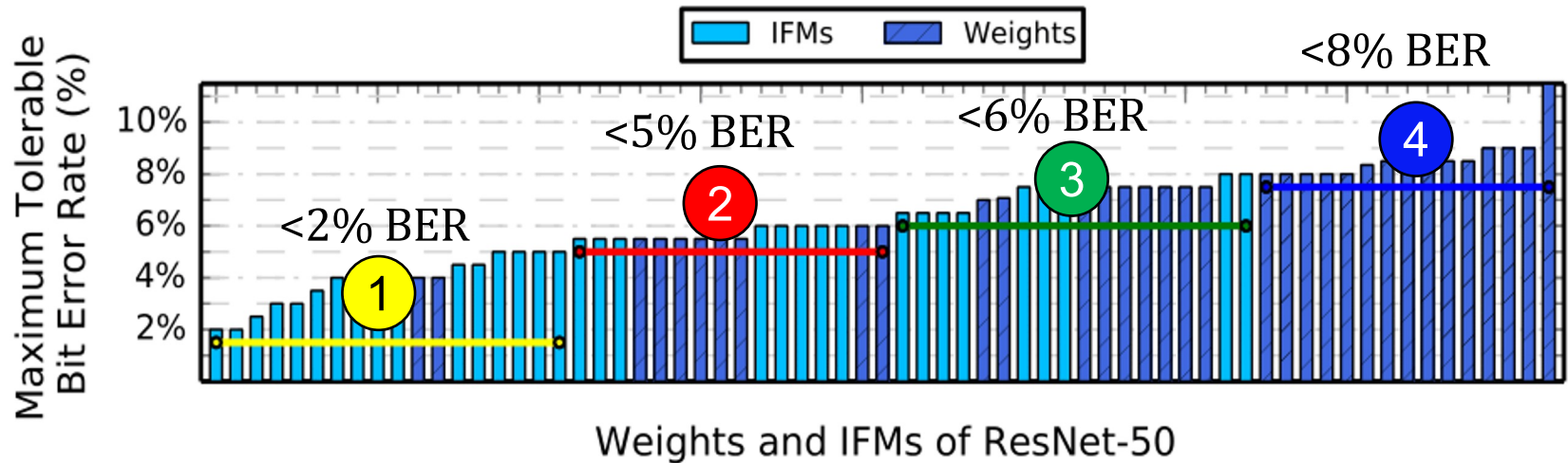
Map more error-tolerant DNN layers
to DRAM partitions **with lower voltage/latency**

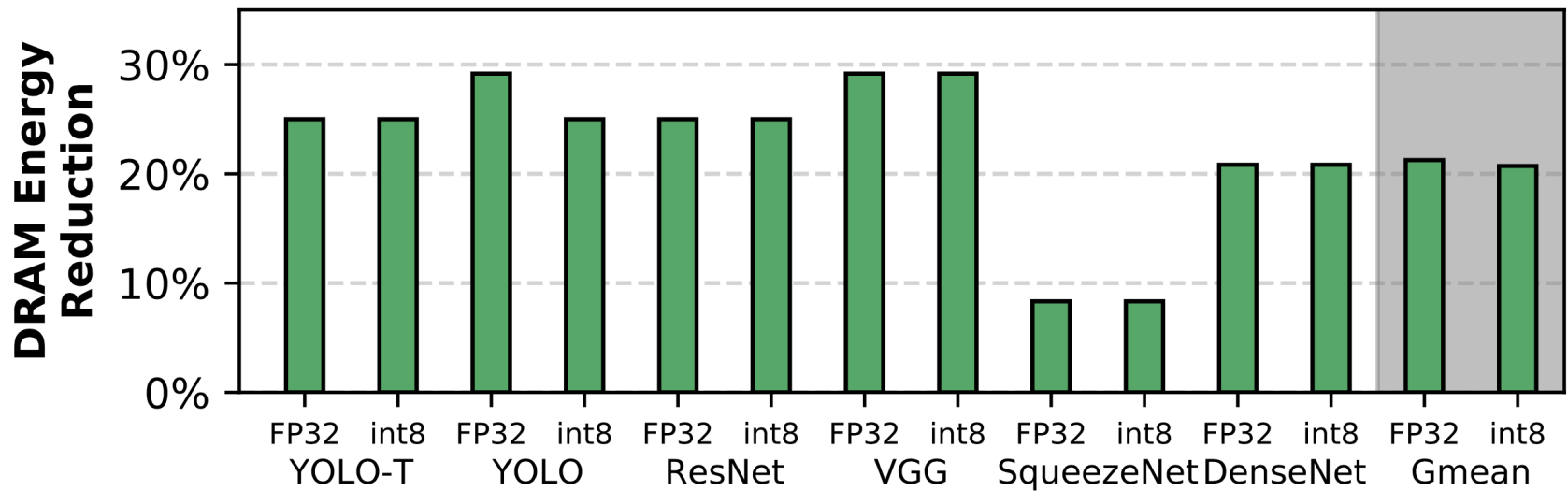**4 DRAM partitions** with different error rates

SAFARI

EDEN: Overview

Key idea: Enable **accurate, efficient** DNN inference using **approximate DRAM**
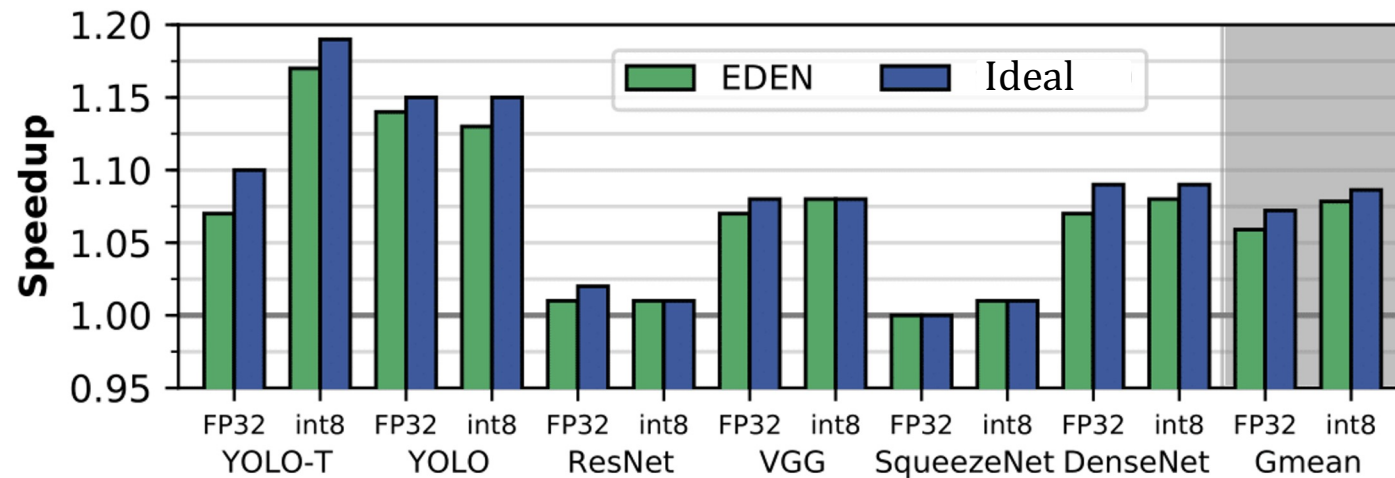
EDEN is an **iterative** process that has **3 key steps**

CPU: DRAM Energy Evaluation

Average 21% DRAM energy reduction
maintaining accuracy within 1% of original

SAFARI

CPU: Performance Evaluation

Average 8% system speedup
Some workloads achieve **17% speedup**

EDEN achieves **close to the ideal** speedup
possible via tRCD latency reduction

SAFARI

GPU, Eyeriss, and TPU: Energy Evaluation

- **GPU**: average **37% energy reduction**

- **Eyeriss**: average **31% energy reduction**

- **TPU**: average **32% energy reduction**

SAFARI

EDEN: Data-Aware Efficient DNN Inference

- Skanda Koppula, Lois Orosa, A. Giray Yaglikci, Roknoddin Azizi, Taha Shahroodi, Konstantinos Kanellopoulos, and Onur Mutlu,
"EDEN: Enabling Energy-Efficient, High-Performance Deep Neural Network Inference Using Approximate DRAM"
Proceedings of the 52nd International Symposium on Microarchitecture (**MICRO**), Columbus, OH, USA, October 2019.
[Lightning Talk Slides (pptx) (pdf)]
[Lightning Talk Video (90 seconds)]

EDEN: Enabling Energy-Efficient, High-Performance Deep Neural Network Inference Using Approximate DRAM

Skanda Koppula   Lois Orosa   A. Giray Yağlıkçı
Roknoddin Azizi   Taha Shahroodi   Konstantinos Kanellopoulos   Onur Mutlu

ETH Zürich