In Computer Architecture, We Don't Change the Questions, We Change the Answers

Mark D. Hill, Microsoft Azure and University of Wisconsin-Madison

Abstract: When I was a new professor in the late 1980s, my senior colleague Jim Goodman told me, "On the computer architecture PhD qualifying exam, we don't change the questions, we only change the answers." More generally, I now augment this to say, *"In computer architecture, we don't change the questions, application and technology innovations change the answers, and it's our job to recognize those changes."* Eternal questions this talk will sample are how best to do the following interacting factors: compute, memory, storage, interconnect/networking, security, power, cooling and one more. The talk will not provide the answers but leave that as an audience exercise. I will dive a little more into compute and memory as in-progress trends provide both challenges and opportunities for creating tremendous value from (large) data.

Biography: Mark D. Hill is Partner Hardware Architect with Microsoft Azure (2020-present) where he leads software-hardware pathfinding. He is also the Gene M. Amdahl and John P. Morgridge Professor Emeritus of Computer Sciences at the University of Wisconsin-Madison (http://www.cs.wisc.edu/~markhill), following his 1988-2020 service in Computer Sciences and Electrical and Computer Engineering. His research interests include parallel-computer system design, memory system design, and computer simulation. Hill's work is highly collaborative with over 160 co-authors. He received the 2019 Eckert-Mauchly Award and is a fellow of AAAS, ACM, and IEEE. He served on the Computing Community Consortium (CCC) 2013-21 including as CCC Chair 2018-20, Computing Research Association (CRA) Board of Directors 2018-20, and Wisconsin Computer Sciences Department Chair 2014-2017. Hill has a PhD in computer science from the University of California, Berkeley.

Microsoft Azure

In Computer Architecture, We Don't Change the Questions, We Change the Answers

Mark D. Hill
Microsoft (Azure) & U. Wisconsin (Emeritus)

@ Data Management on New Hardware Workshop (DaMoN), June 2022

This a public, non-proprietary talk.
I speak for myself, not necessarily Microsoft or Azure.

Computer Architects: Components → Systems

My new job: Hardware-software pathfinding for Azure

A View of Computing's "Stack"

Problem & Algorthms

Applications

DBMSs & Other Middleware

Runtime & Compiler

Operating System

(Micro) Architecture

Hardware

Materials & Fabrication

As technology scaling slows, dramatic perf/cost gains needed will require layer experts to work together!

New Assistant Professor [1988]

Mark Hill:
How do we update questions for the computer architecture PhD qualifying exam?

Jim Goodman:
We don't change the questions.
We change the answers.

My Current View

In computer architecture,

We don’t change the questions

Applications & technology innovations change the answers

It’s our job to recognize those changes

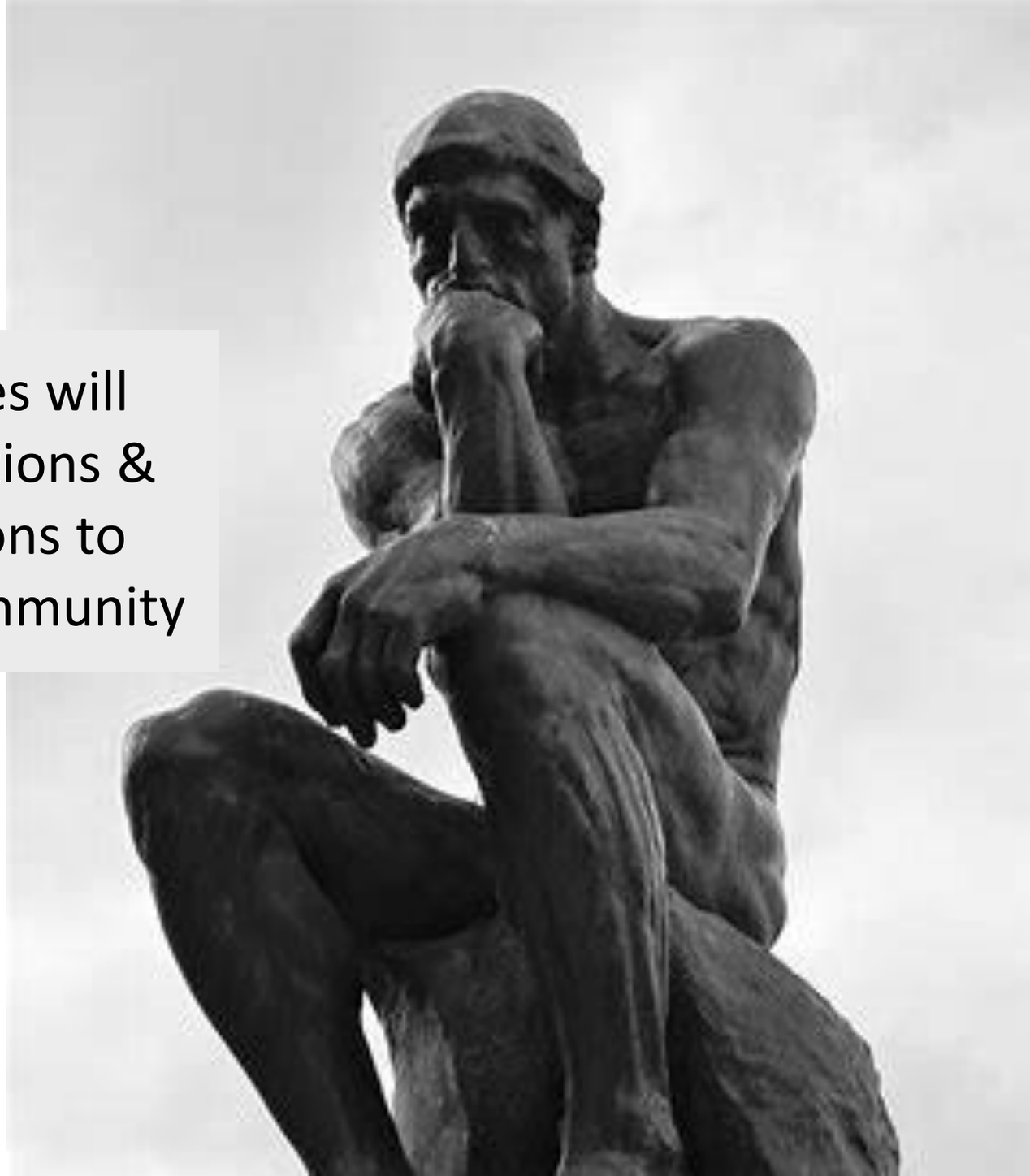
This talk discusses these eternal questions; answers TBD by you!

Computer Architecture's Eternal Questions & Outline

How best to do these interacting factors:

1. Compute (longer)
2. Memory (longer)
3. Storage
4. Interconnect/networking
5. Security
6. Power
7. Cooling
8. *Bonus new question*

Gray boxes will have questions & implications to DaMoN community

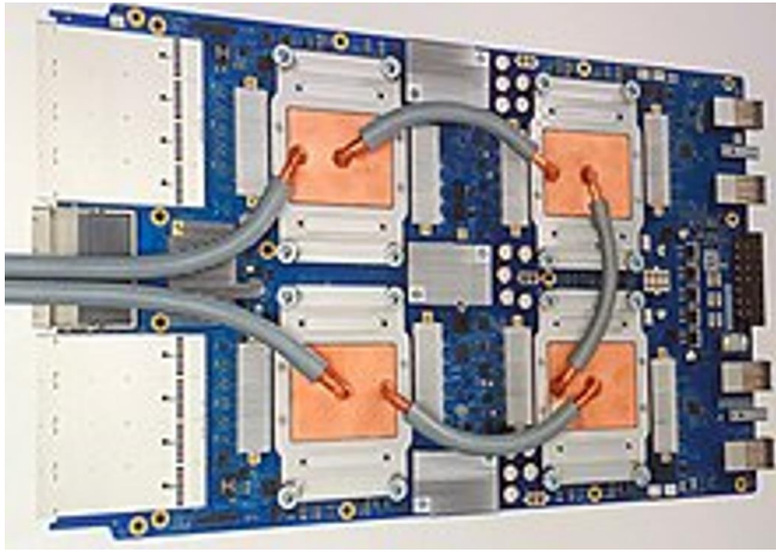


Compute: Accelerators, e.g., Deep Learning

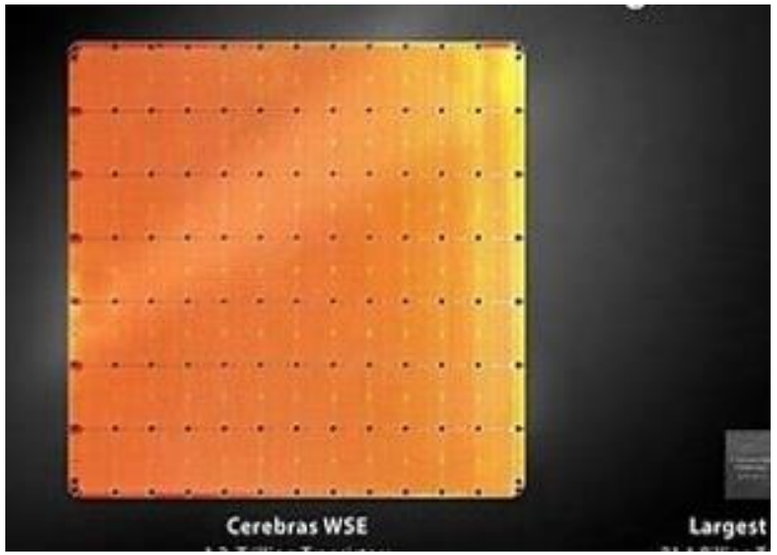
DaMoN questions on 4th compute slide

End of Dennard scaling & rise of demanding apps ➔

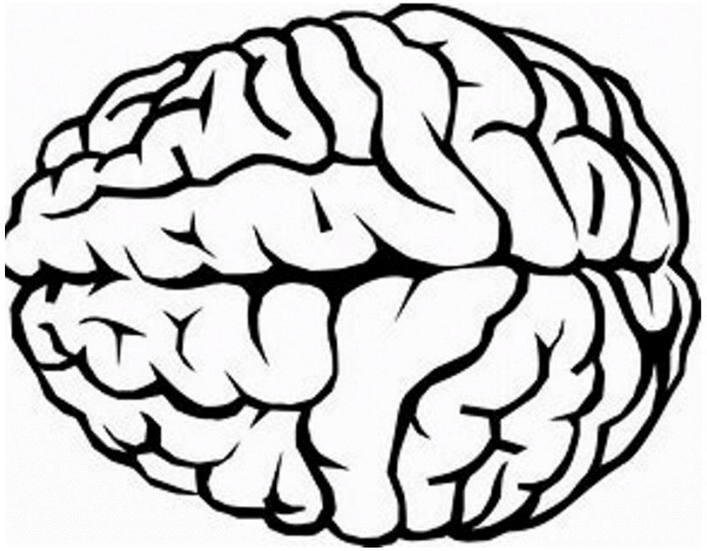
- Accelerator *is a hardware component that executes a targeted computation class faster & usually with (much) less energy.*
- Esp. Deep Neural Network Machine Learning



Google Tensor Processing Unit



Cerebras Wafer Scale Engine



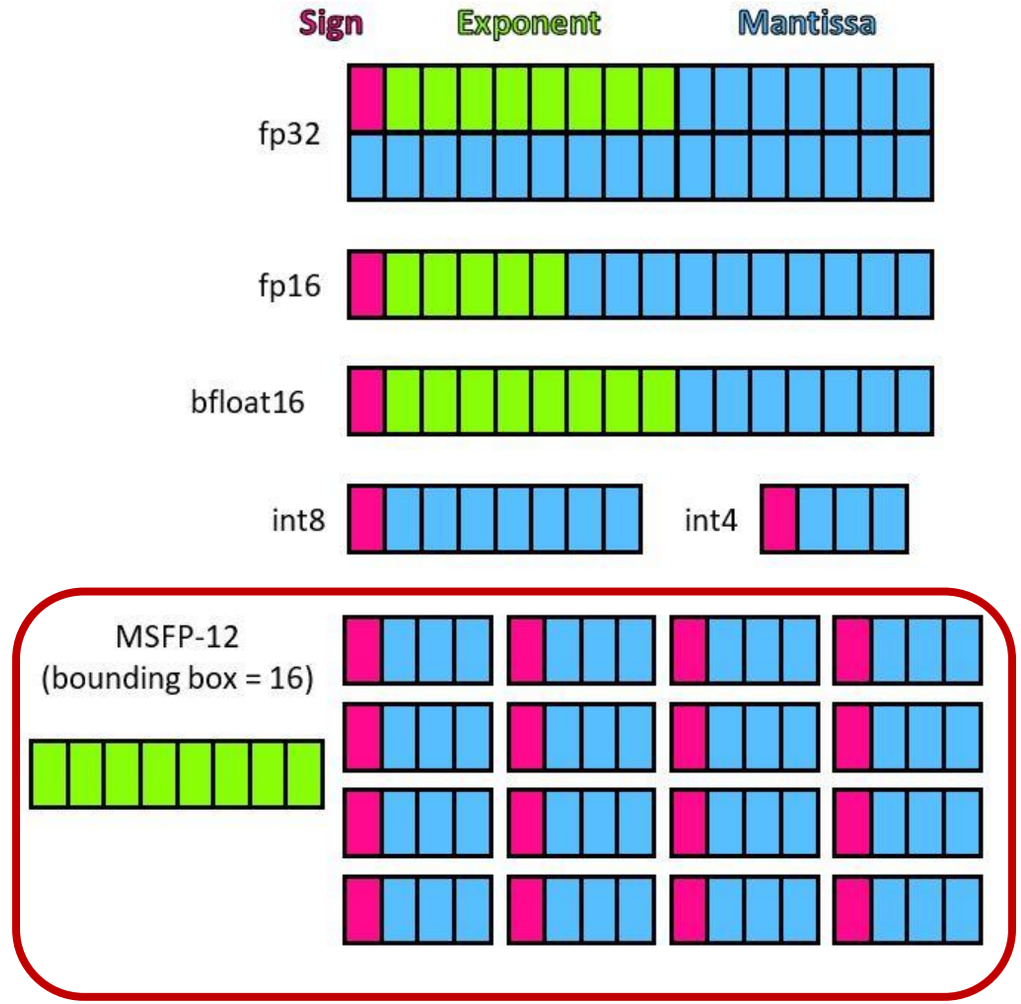
Human Brain (esp. Neocortex)

Compute: Accelerators, Deep Learning Co-design

Co-Design Nascent for Deep Learning Training/Interference

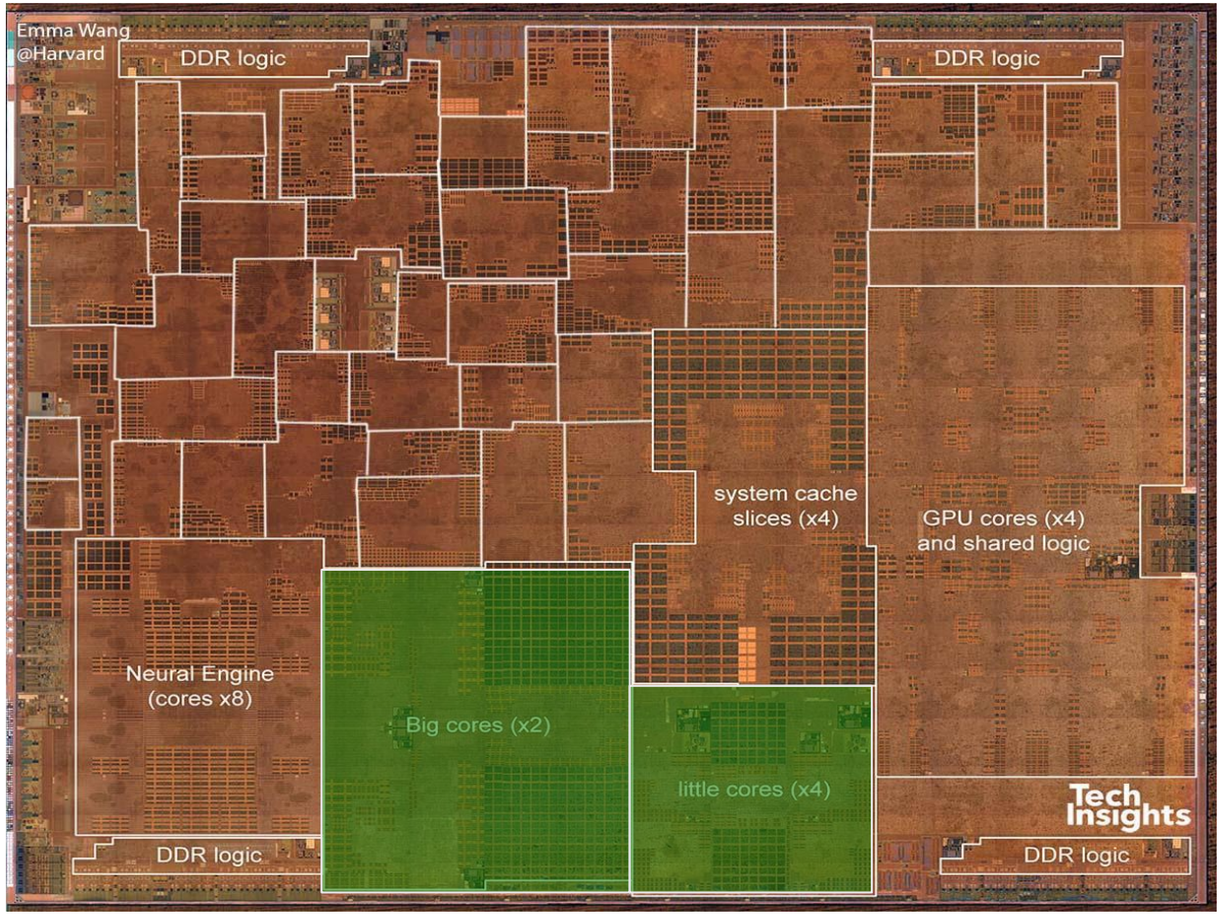
E.g., Microsoft Floating-Point

- Mantissa really small
- Multiple values share exponent
- MFFP-12: (8 + 16*4)/16
  = 4.5 bits/value
- Requires co-design



https://www.microsoft.com/en-us/research/blog/a-microsoft-custom-data-type-for-efficient-inference/

Compute: Accelerator-Level Parallelism



2019 Apple A12 w/ **42 Accelerators**

Deploy Many Accelerators

Use several concurrently

- CPUs: control plane
- Accelerator: data plane

How program, schedule, communicate, co-design?

https://cacm.acm.org/magazines/2021/12/256949-accelerator-level-parallelism

Accelerating Compute for Data-Intensive Work

Implications to DaMoN Community

What?

- Important fraction of unaccelerated execution time
- Common, stable, compute-bound & CPU unfriendly
- (De)compression, (de)encryption, columnar joins?

How? CPU SIMD (AVX), GPU, Custom (TPU for DBMSs?),

- Or FPGAs (between SW & hard HW on flexibility & speed)

Where?

- Granularity & data movement (see PIM soon)
- Location in cloud data center as computer?

Whither interactions with neural network acceleration?

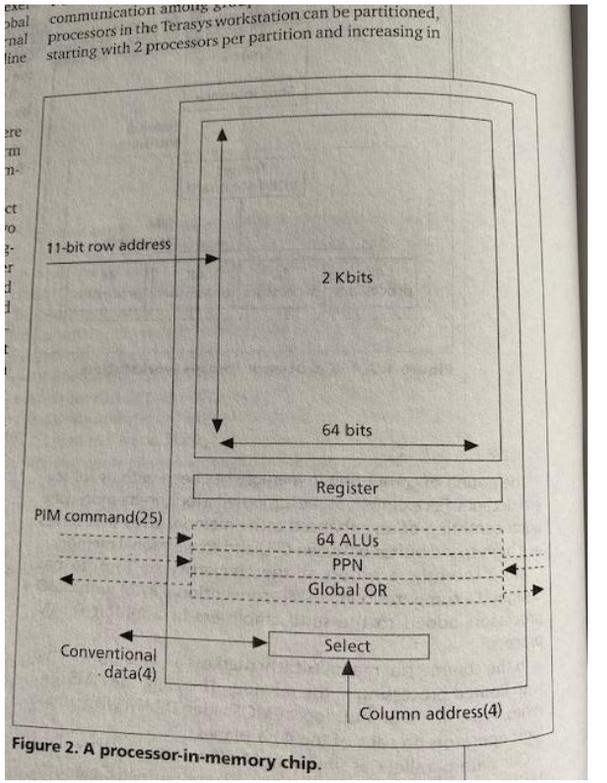
Memory: Processing In Memory (PIM)

Usually, move all data to CPU(s)

PIM: Move compute to vast data in memory

Old idea revived by

1. Conventional compute's energy problems
2. Important apps: Deep Learning & Recommendation
3. Attention from serious memory vendors

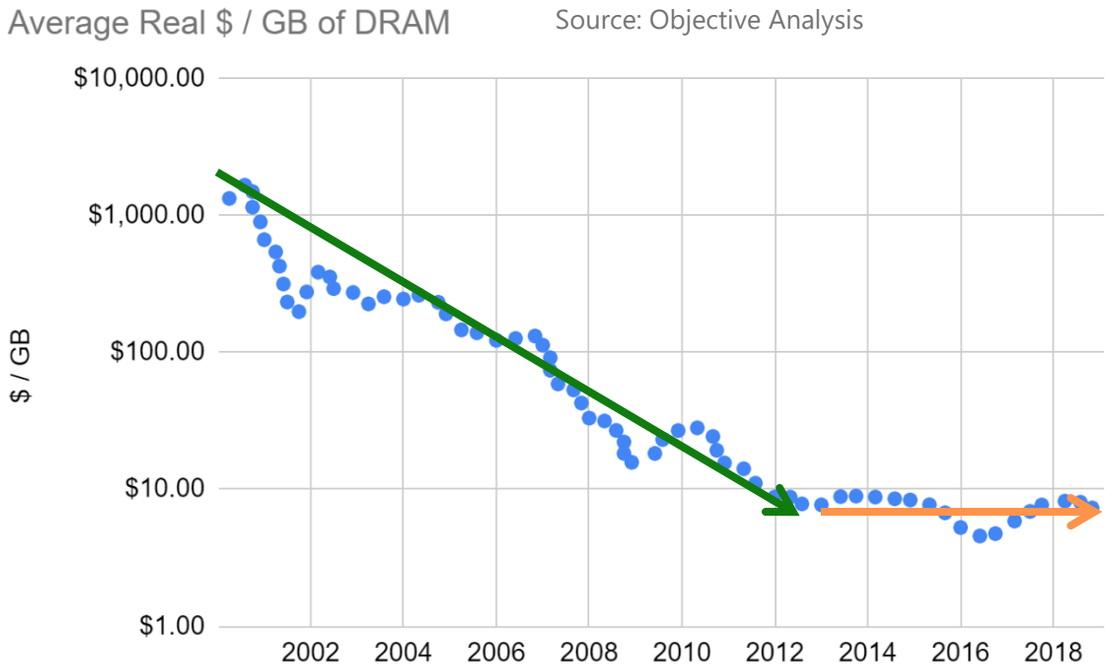


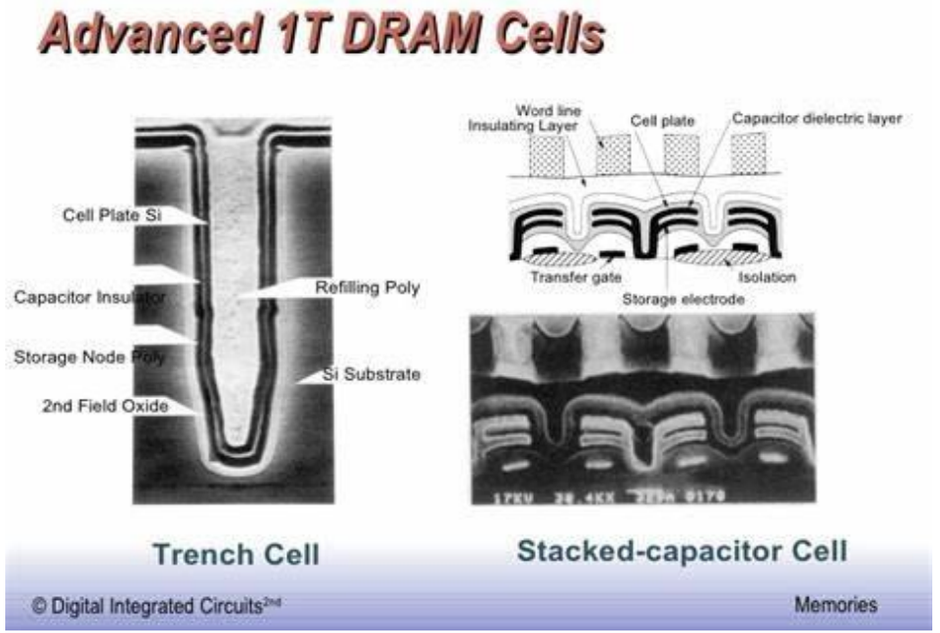
Gokhale,Holmes,Iobst [1995]

Hardware Architecture and Software Stack for PIM
Based on Commercial DRAM Technology
Sukhan Lee, et al., **Samsung**, ISCA Industrial Track, June 2021

PIM requires apps that can move small compute & data (query) to large corpus

Memory: Vast, Fast, Synchronous DDR ➔ Untenable





DDR DRAM price not scaling ➔ poor 2D scaling

Force Response: Two-Tier Memory (c.f., Multicore 20 years ago)

New CXL Enables Two-Tier Memory

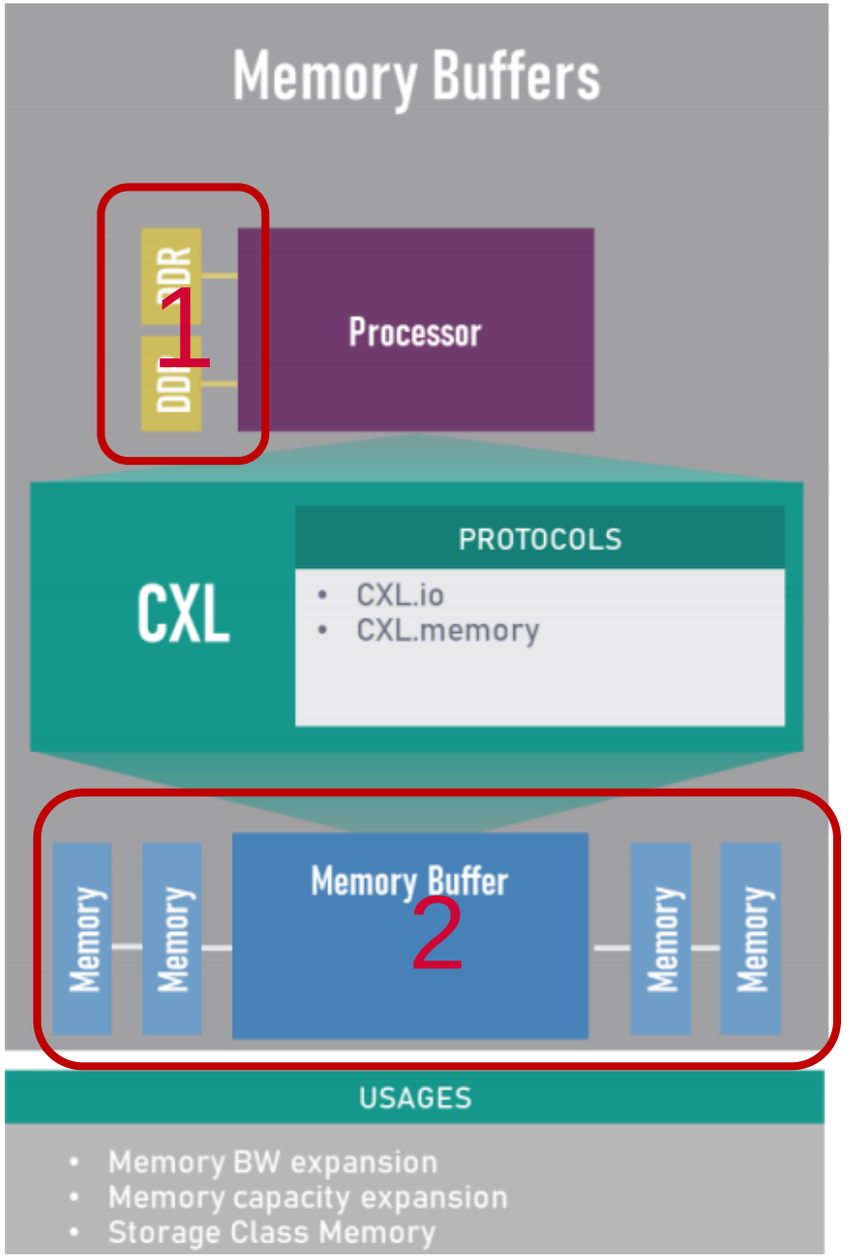
CPUs support CXL; innovation behind it

Pool Memory Among Nodes (not shown)

- Smooth per-node peaks
- Buy fewer bits; not cheaper per bit
- https://arxiv.org/abs/2203.00241 [2022]

Extended Memory

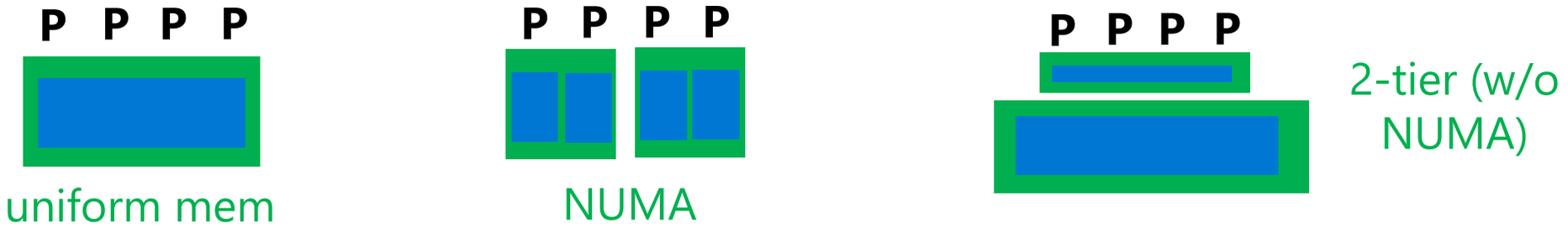
- More DRAM to avoid 3D-stacked DRAM
- New technologies (better than 3DXP)



Two-Tier Memory for Data-Intensive Work

Implications to DaMoN Community

Defense: Existing Data Apps to flourish, e.g., RDBMS buffer pool



Offense: Creative new opportunities from Tier 2 memory?

- DRAM << Capacity << SSD
- DRAM < Latency << SSD
- DRAM < Energy/Bit << SSD

Storage: Mind the Gaps

Without new HW, look at hard disk structures for cold & archival storage?

Solid State Drive

Hard Disk Drive

Tapes (2 vendors)

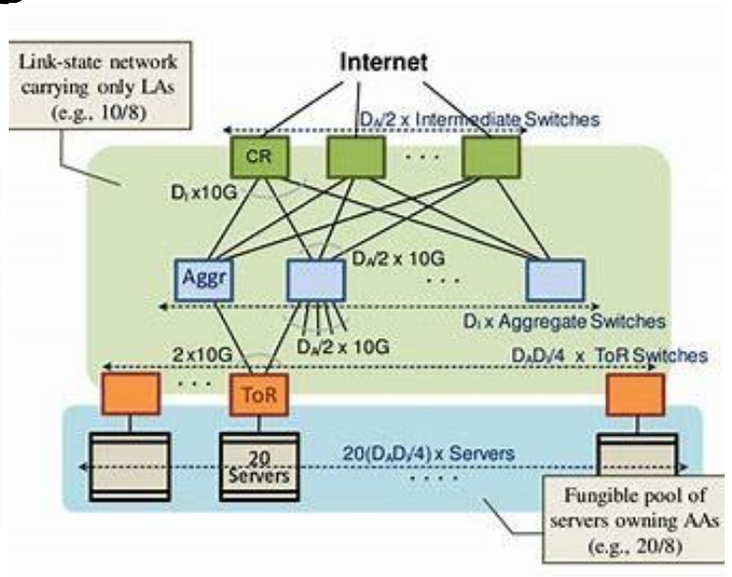
Persistent Memory?

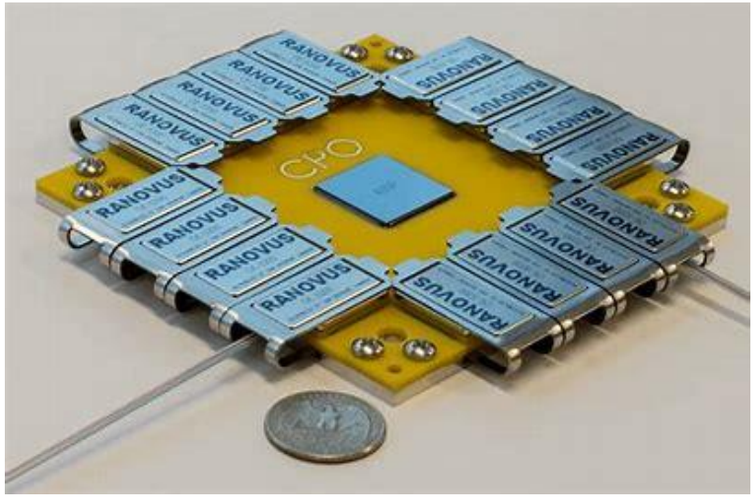
Many-bit Cell, Appliance?

Microsoft Research:
DNA & Silica

https://www.microsoft.com/en-us/research/project/dna-storage/
https://www.microsoft.com/en-us/research/project/project-silica/

Data Center Networking





Protocols

- Cost-effectively (Ethernet) → high perf computer interconnect
- BW, latency, & jitter matter (see Infiniband & Cray/HPE Slingshot)

- Technology
- Optics already used above top-of-rack switch (ToR)
- Cost-effective move closer to node? To what effect?

What new Data apps or app structure?

Security: Confidential Compute

Cloud Providers Now:

- We **promise** to protect your data/code from outsider/insider threats

With Confidential Compute

- **Your data/code is cryptographically protected from both threats**

But

- Hard: Root of trust, attestation, interchip comm encrypted, memory/storage w/ data/address/replay protected, …
- Can expand markets, but correctness/efficiency challenges

Azure CC: https://queue.acm.org/detail.cfm?id=3456125 [ACM Queue 2021]
Azure Sphere (IoT): https://aka.ms/7properties

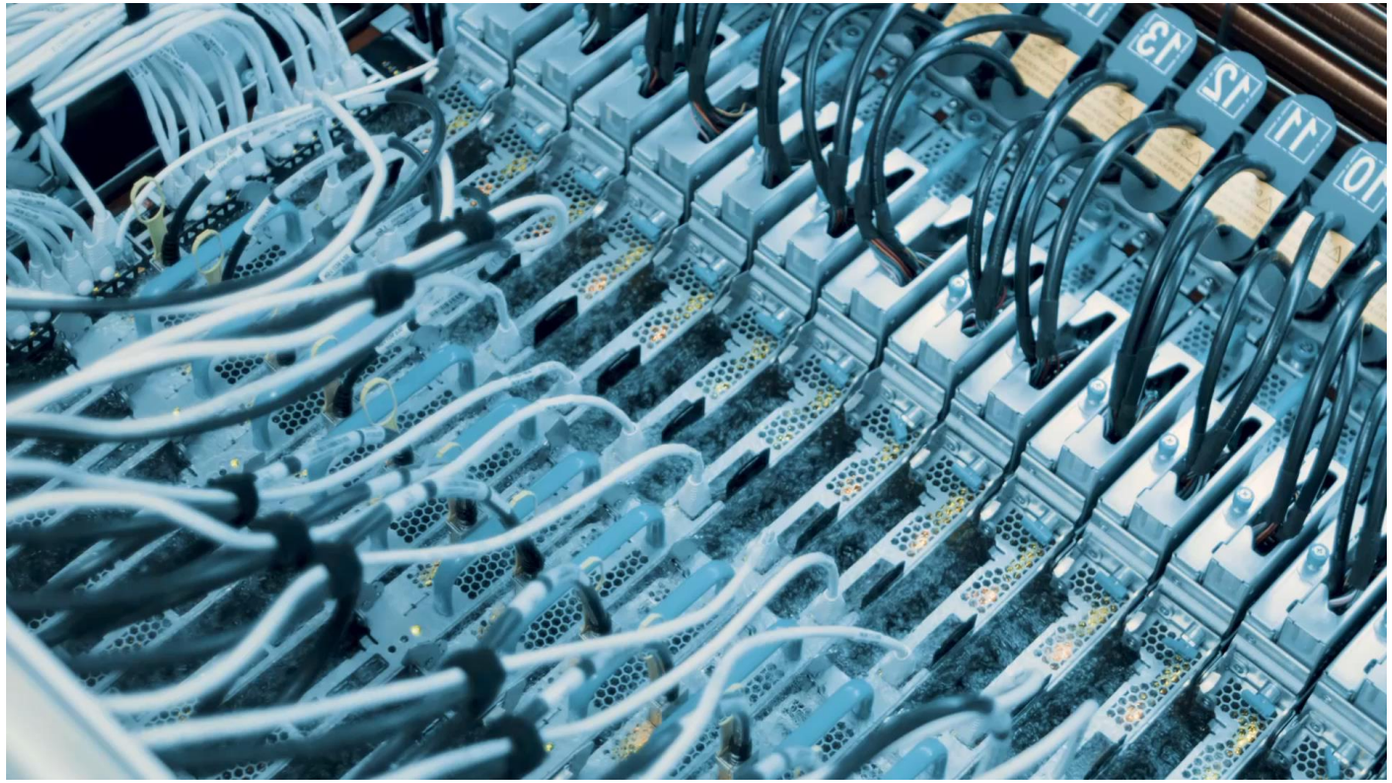
Must apps trust data middleware? How?
Homomorphic encryption still absurd

Cooling

Air Cooling Facing Limits

Cold Plate Coming

Then Immersion Cooling?



How might this interact with computer architecture's other eternal questions?

https://news.microsoft.com/innovation-stories/datacenter-liquid-cooling/

Exploit compact access to vast compute, memory, & storage?

Power: IoT to Cloud Varies

Can data apps do batch work when power plentiful to be ready for power throttling?

IoT/Mobile: Energy (battery life)

- Save energy: Use little energy ~idle
- Add energy: E.g., harvesting

…

Cloud: Constant Power

- Mega-datacenters pay for fixed power
- **Using less power doesn't save money**
- **How to use constant power well?**
- Intermittent, renewable power expanding

https://www.microsoft.com/en-us/research/uploads/prod/2020/10/Per-VM-Capping-ATC21.pdf

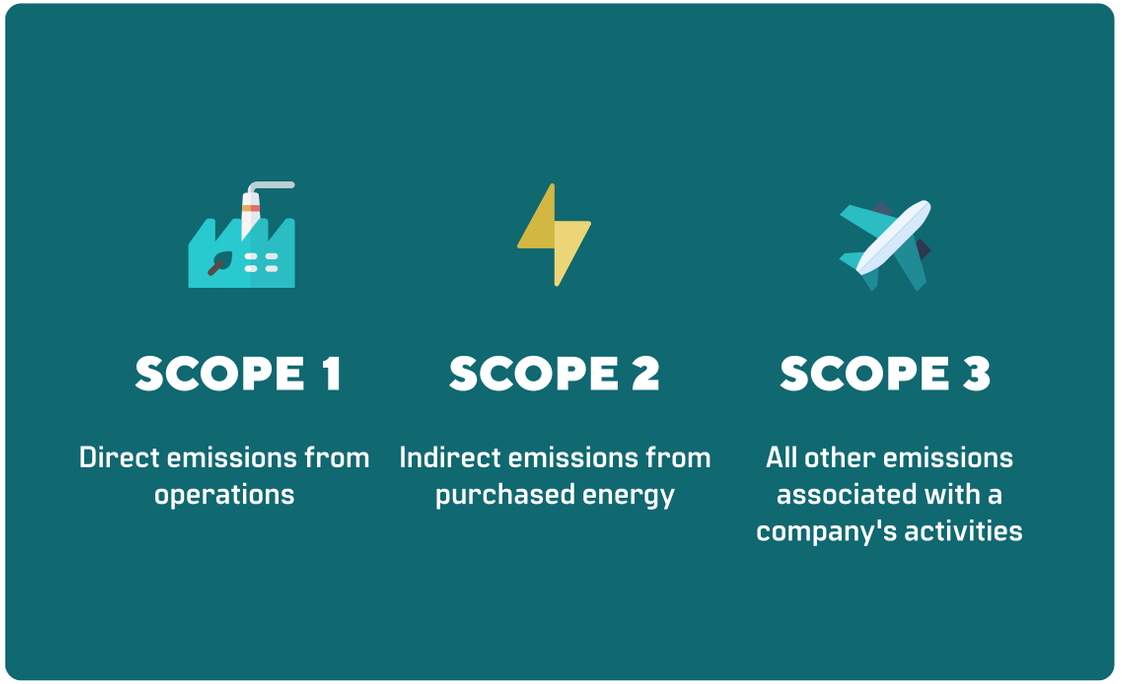
(Bonus) Sustainability!

Where do data apps spend carbon in CAPEX & OPEX?

I said comp arch's questions don't change but
George Box: *All models are wrong, but some are useful.*

New: Make Computing More Sustainable?

Green House Gas Emission Scopes



US EPA: https://www.epa.gov/ghgemissions

Microsoft seeks carbon negative by 2030, https://www.microsoft.com/en-us/corporate-responsibility/sustainability
See also: Chasing Carbon: The Elusive Environmental Footprint of Computing
Udit Gupta, et al., Harvard & Facebook, HPCA 2021 Industrial Track, https://arxiv.org/abs/2011.02839

Computer Architecture's Eternal Questions & Outline

How best to do these interacting factors:

1. ***Compute:*** accelerators, deep learning, & many
2. ***Memory:*** 2D scaling dead & processing in memory
3. ***Storage:*** mind the gaps
4. ***Interconnect/network:*** protocols/optics
5. ***Security:*** confidential compute
6. ***Power:*** IoT to cloud varies
7. ***Cooling:*** consider immersion & its impact
8. New: ***Sustainability:*** whither emission scopes 1, 2, & 3?

Let's work together to increase the value data apps provide!

Invent with purpose.

https://careers.microsoft.com