

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

Mark D. Hill

University of Wisconsin-Madison Professor Emeritus

@ University of Wisconsin-Madison Computer Sciences, Sept. 2024

Computer Architecture: Big Picture of Computer HW

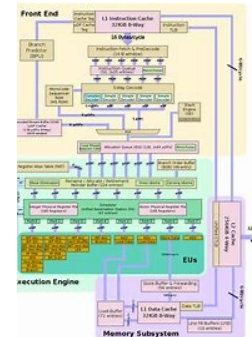
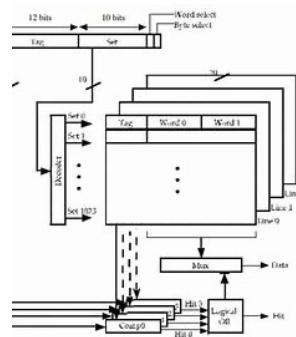
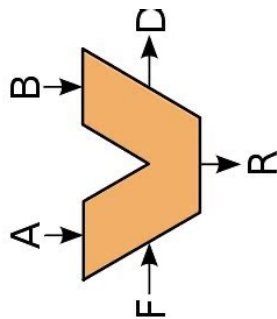
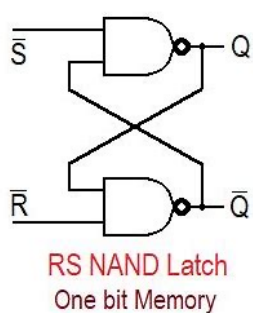
Components



Systems



Gates → ALU → Functional Block → Core → SoC → Server → Data Center



Computer Architects: Components → Systems



12/2020 – 03/2024: Hardware-software pathfinding for Azure

A View of Computing's "Stack"

Problem & Algorithms

Applications

DBMSs & Middleware

Runtime & Compiler

Operating System

(Micro) Architecture

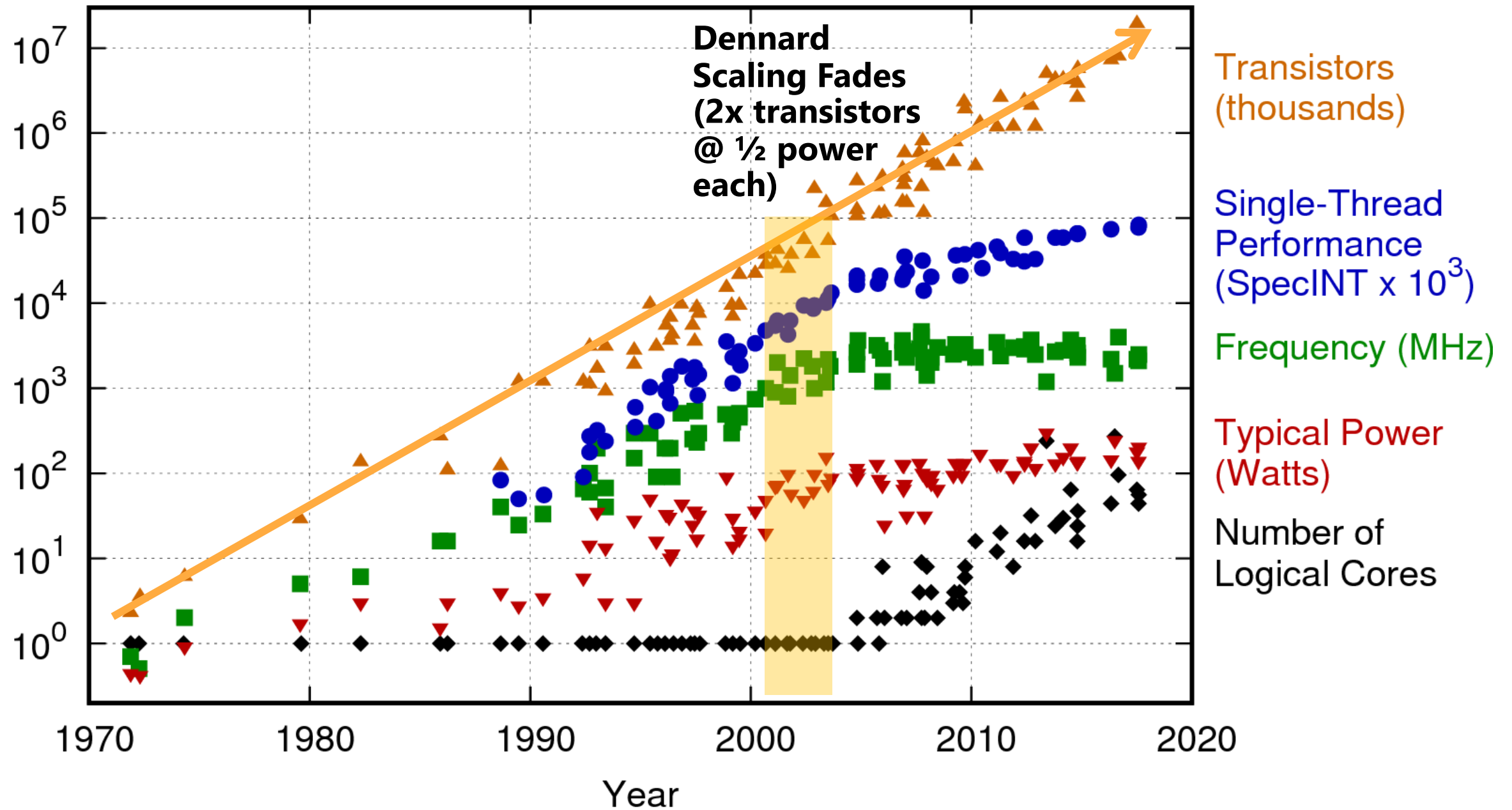
Hardware

Materials & Fabrication



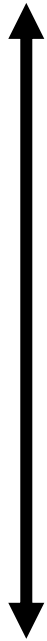
As technology scaling slows, dramatic perf/cost gains needed will require layer experts to work together!

42 Years of Microprocessor Trend Data



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp <https://www.karlrupp.net/wp-content/uploads/2018/02/42-years-processor-trend.png>

A Commercial Computing Company Helix



IBM digital

etc.

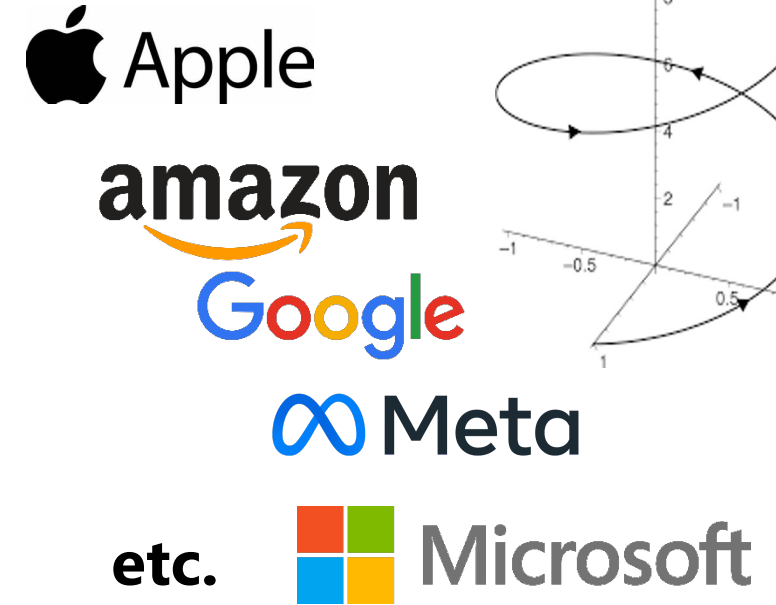
Pre-microprocessor Era

Medium tech progress
Users share
Comp layers nascent
Vertical companies



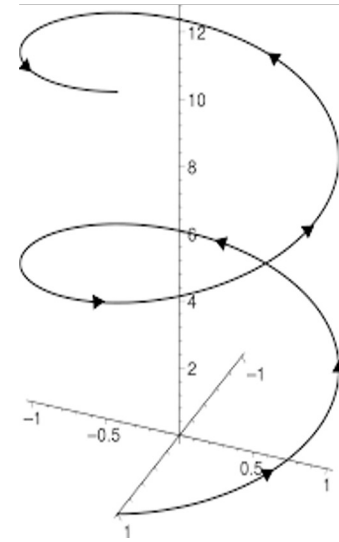
Microprocessor Era

Amazing tech progress
Per-user devices
Comp layers rigid
Horizontal companies



Cloud & Mobile Era

Medium tech progress
Users share cloud, not dev
Cross-layer opt req'd
Vertical companies



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

E.g., Single Instruction Multiple Data (SIMD): 1960s → GP-GPUs

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

Computer Architecture's Eternal Questions & Outline

How best to do these
interacting factors:

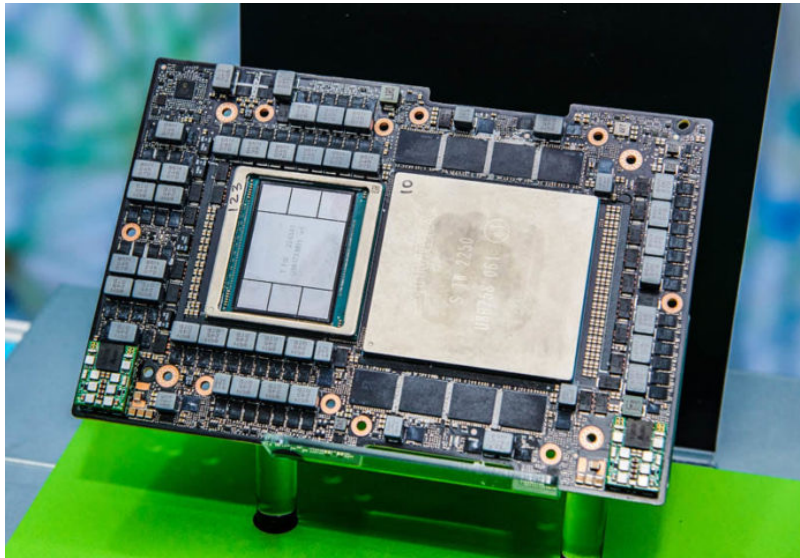
1. **Compute (longest)**
2. **Memory (longer)**
3. Interconnect/networking
4. Storage
5. Security
6. Power
7. Cooling
8. *Bonus new question*



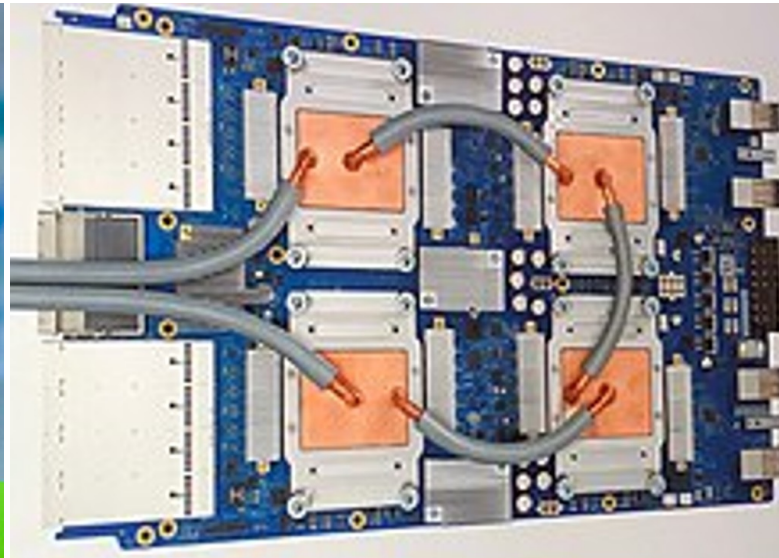
Compute: Accelerators, e.g., Deep Learning

End of Dennard scaling & rise of demanding apps →

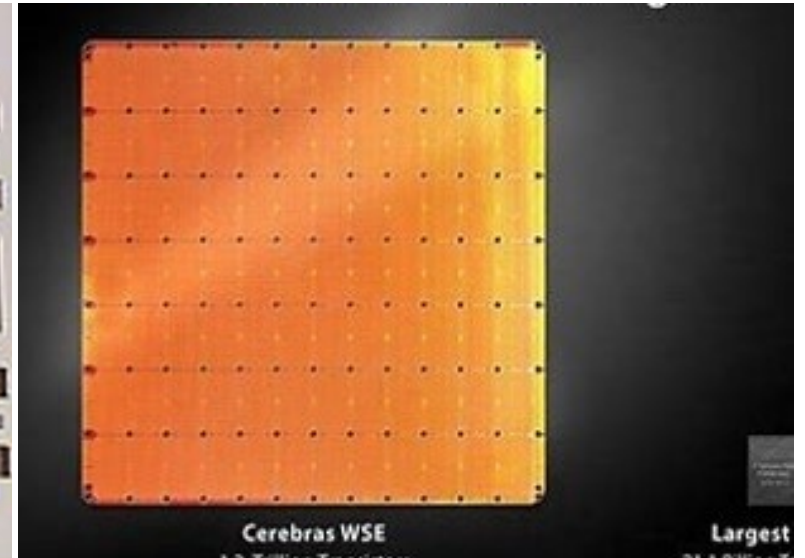
- ***Accelerator is a hardware component that executes a targeted computation class faster & usually with (much) less energy.***
- Esp. Deep Neural Network Machine Learning



Nvidia Grace-Hopper



Google Tensor Processing Unit



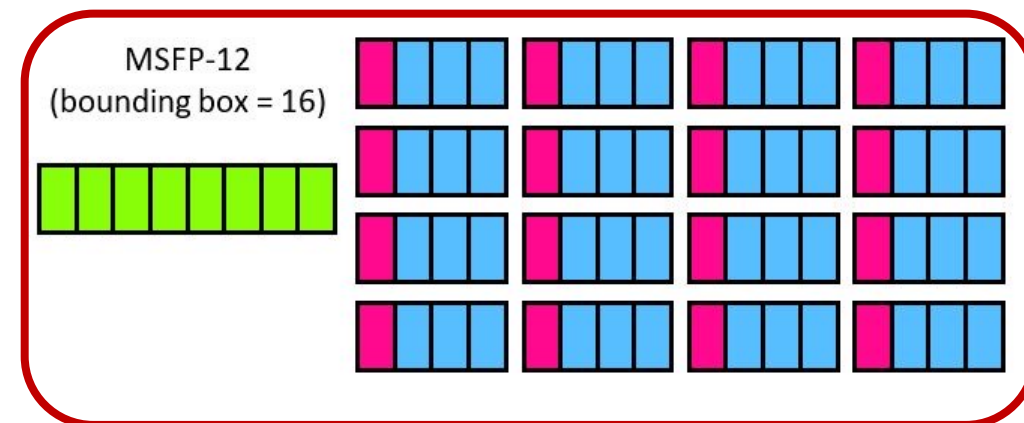
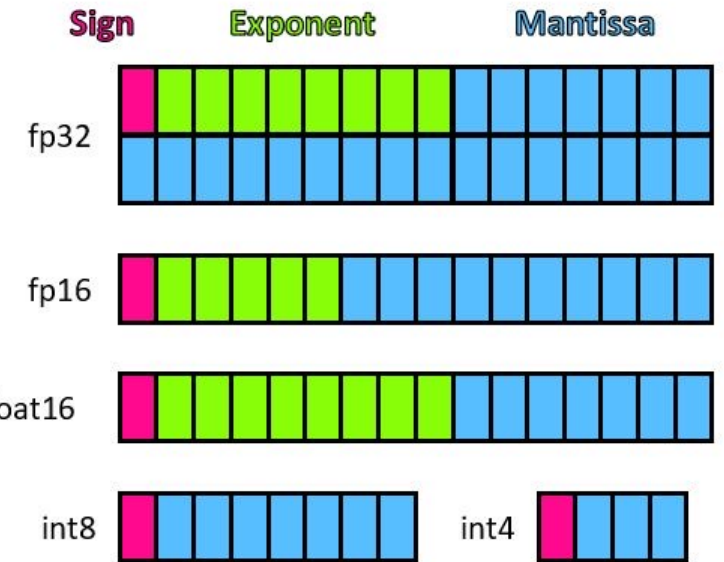
Cerebras Wafer Scale Engine

Compute: Accelerators, Deep Learning Co-design

E.g. Co-Design for Deep Learning via Number Representation

Microsoft FP → Microscaling Formats (MX)

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



2020: <https://www.microsoft.com/en-us/research/blog/a-microsoft-custom-data-type-for-efficient-inference/>

2023: <https://www.opencompute.org/blog/amd-arm-intel-meta-microsoft-nvidia-and-qualcomm-standardize-next-generation-narrow-precision-data-formats-for-ai>

Generative AI

Amazing opportunity: sum >> parts

- Foundation models == means
- Customer value provided by other apps
 - Doing now: make tedious work → faster
 - **Pot of gold**: near impossible → practical
 - Expect bumps: Gartner Hype Cycle

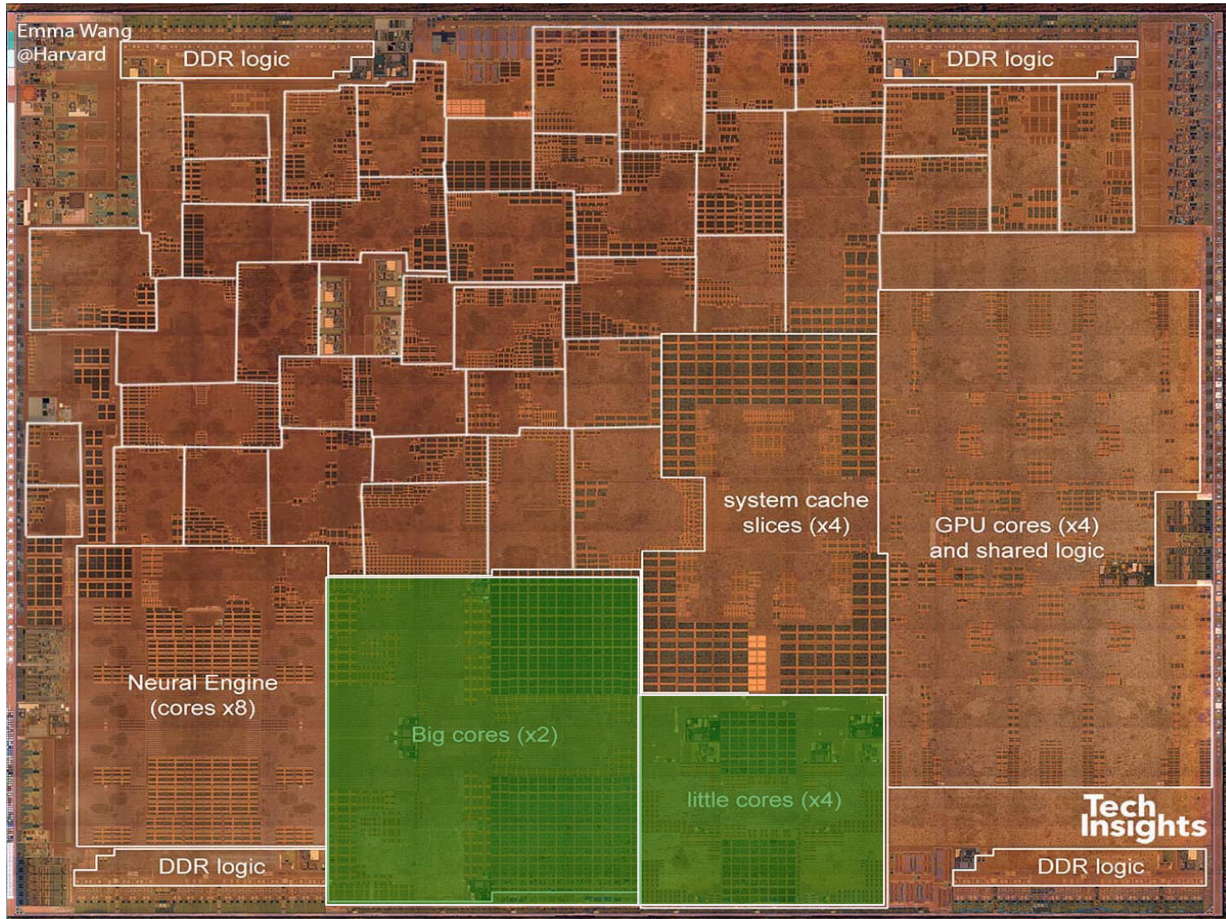


- Massive special clusters for foundational AI training: GPUs, TPUs, ...
- Growing incremental training. How & where?
- **Exploding inference: wearable, phone, laptop, edge, AND Cloud**
- How to structure AI & GP software and hardware?
- In Cloud, AI clusters will consume massive power → less for GP

New use cases are paramount, and

> Efficiency → Enables providing value to more people in more ways

Compute: Accelerator-Level Parallelism



2019 Apple A12 w/ **42 Accelerators**

Deploy Many Accelerators

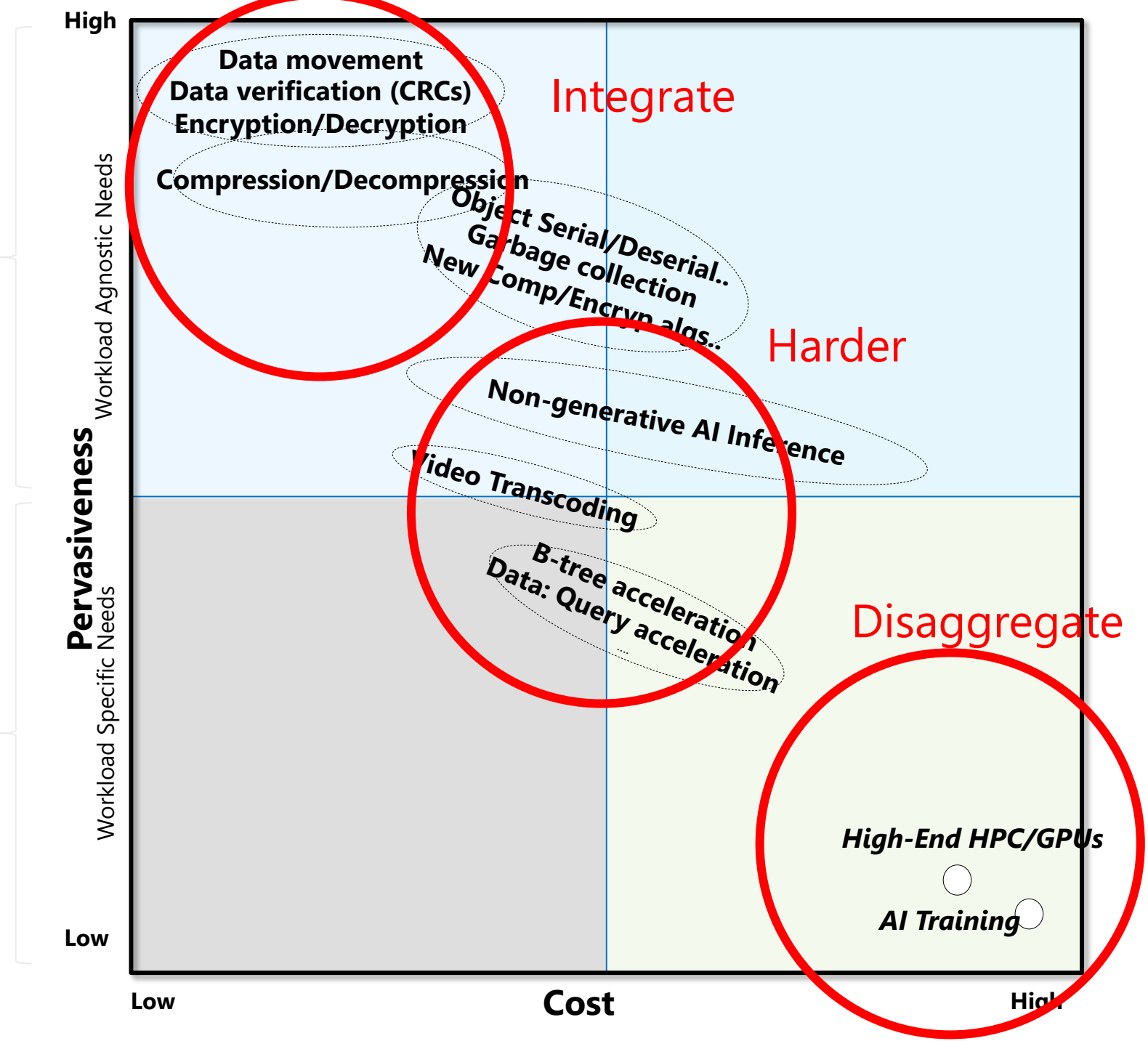
Use several concurrently

- CPUs: control plane
- Accelerator: data plane

How program, schedule, communicate, co-design?

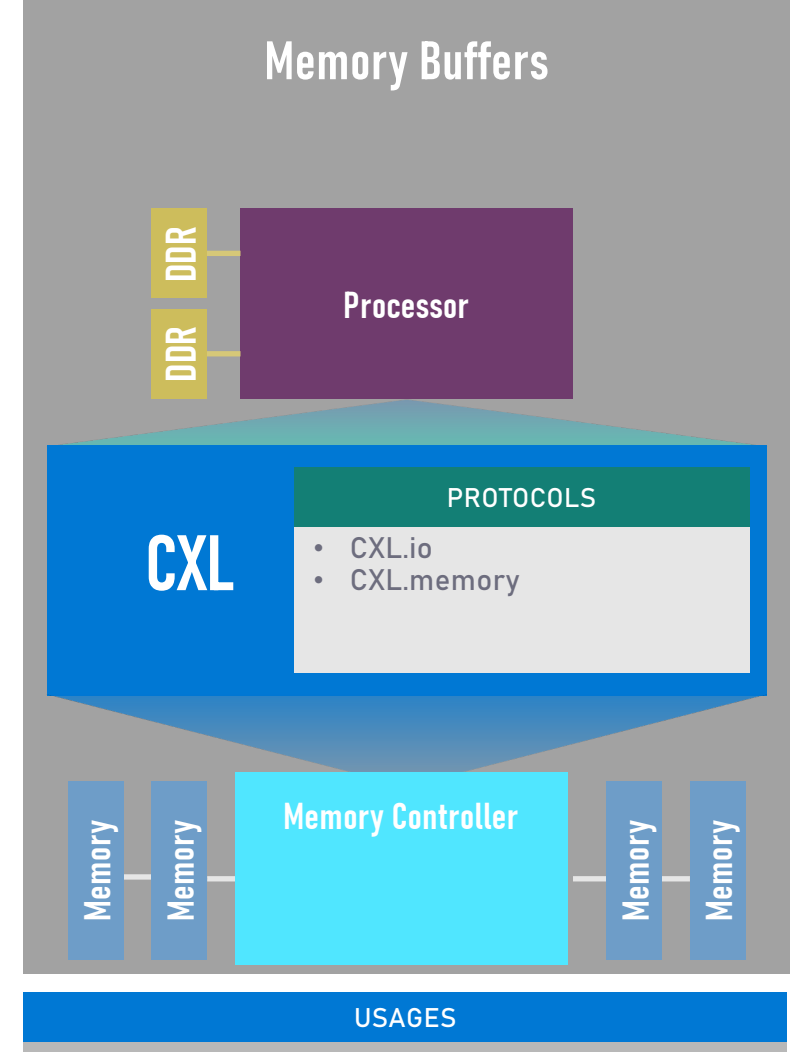
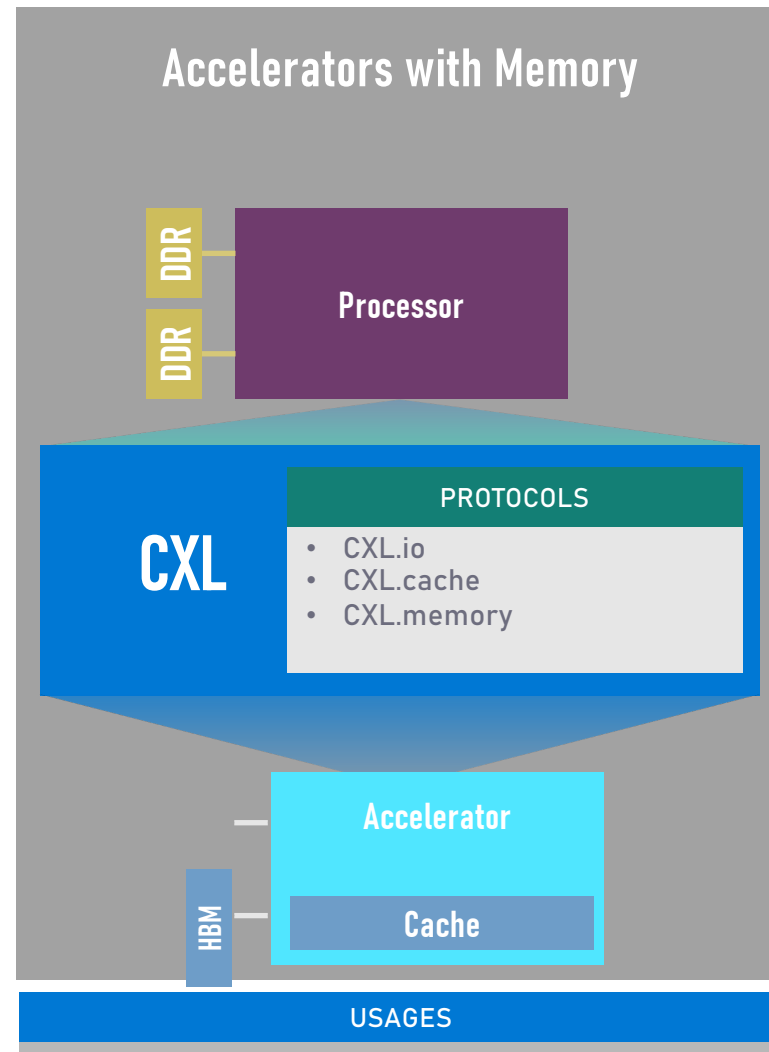
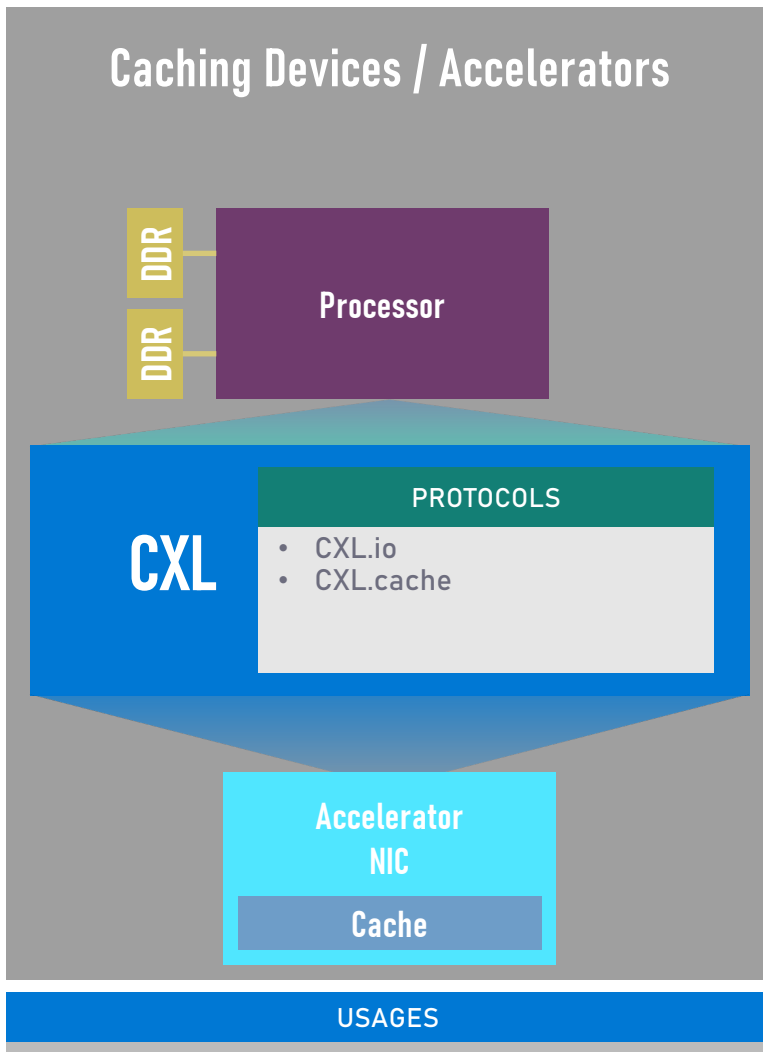
<https://cacm.acm.org/magazines/2021/12/256949-accelerator-level-parallelism>

Where to Accelerate?



Thanks to Ram Huggahalli

New Opportunity: Compute eXpress Link (CXL)



Enables accelerators “closer” than PCIe (coherent) & two-level memory

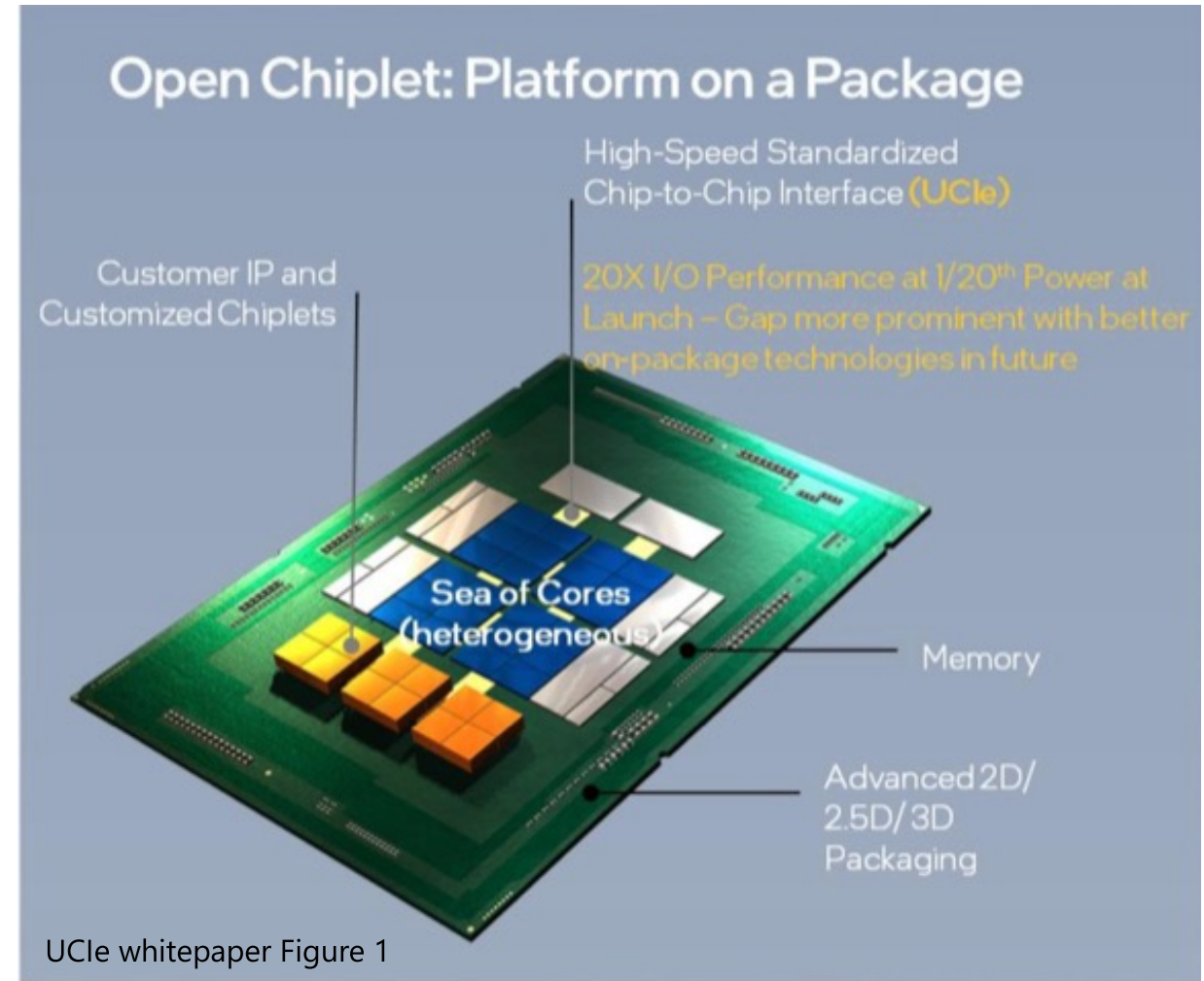
Emerging Opportunity: Universal Chiplet Interconnect Express (UCIe)

Due to Moore's Law Challenges

- Monolithic chip → several "chiplets"
- Fast Silicon interconnect
- Currently company proprietary

Emerging UCIe Standard

- Make package like a "board"
- Standardized protocol among chiplets (physical/electrical/link/transport)
- Get closer: PCIe > CXL > UCIe
- Mix/match chiplets from different technologies/companies
- <https://doi.org/10.1038/s41928-024-01126-y>



2D then 2.5D then 3D. **3D is the frontier of tech scaling for UCIe & in general!**

Computer Architecture's Eternal Questions & Outline

How best to do these
interacting factors:

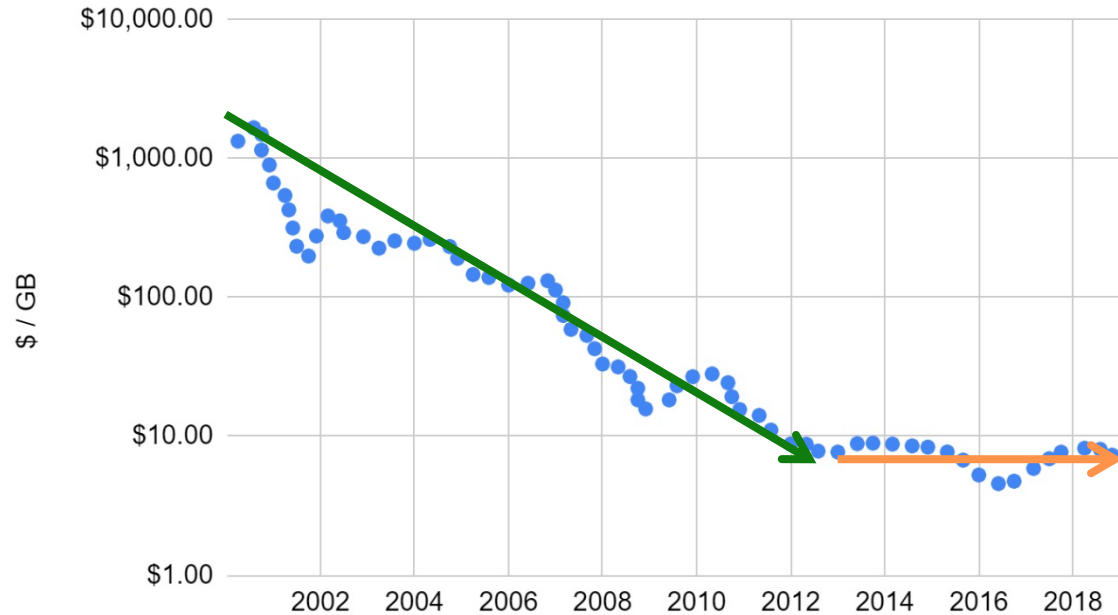
- ~~1. Compute (longest)~~
- 2. Memory (longer)**
3. Interconnect/networking
4. Storage
5. Security
6. Power
7. Cooling
8. *Bonus new question*



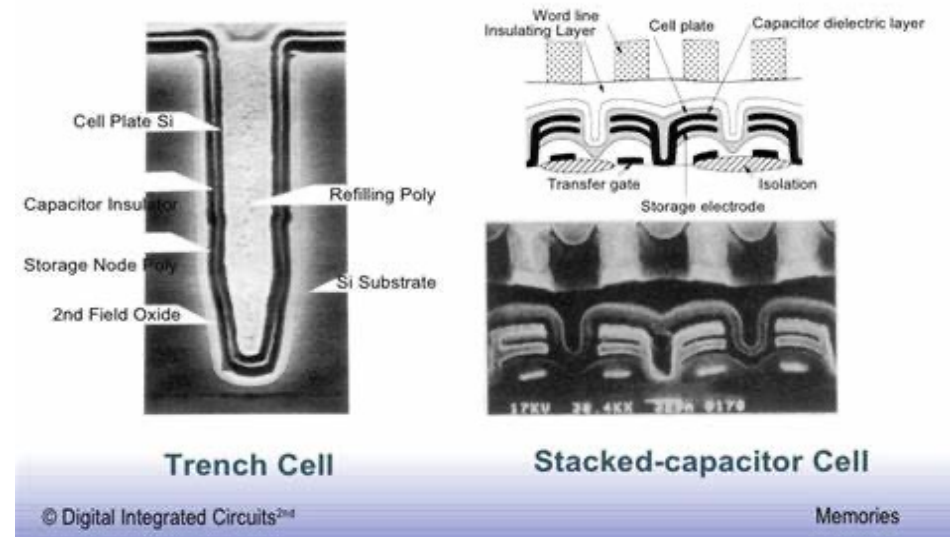
Memory: Vast, Fast, Synchronous DDR → Untenable

Average Real \$ / GB of DRAM

Source: Objective Analysis



Advanced 1T DRAM Cells



DDR DRAM price not scaling → poor 2D scaling
→ With DDR only, future cores/socket growth will slowdown

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

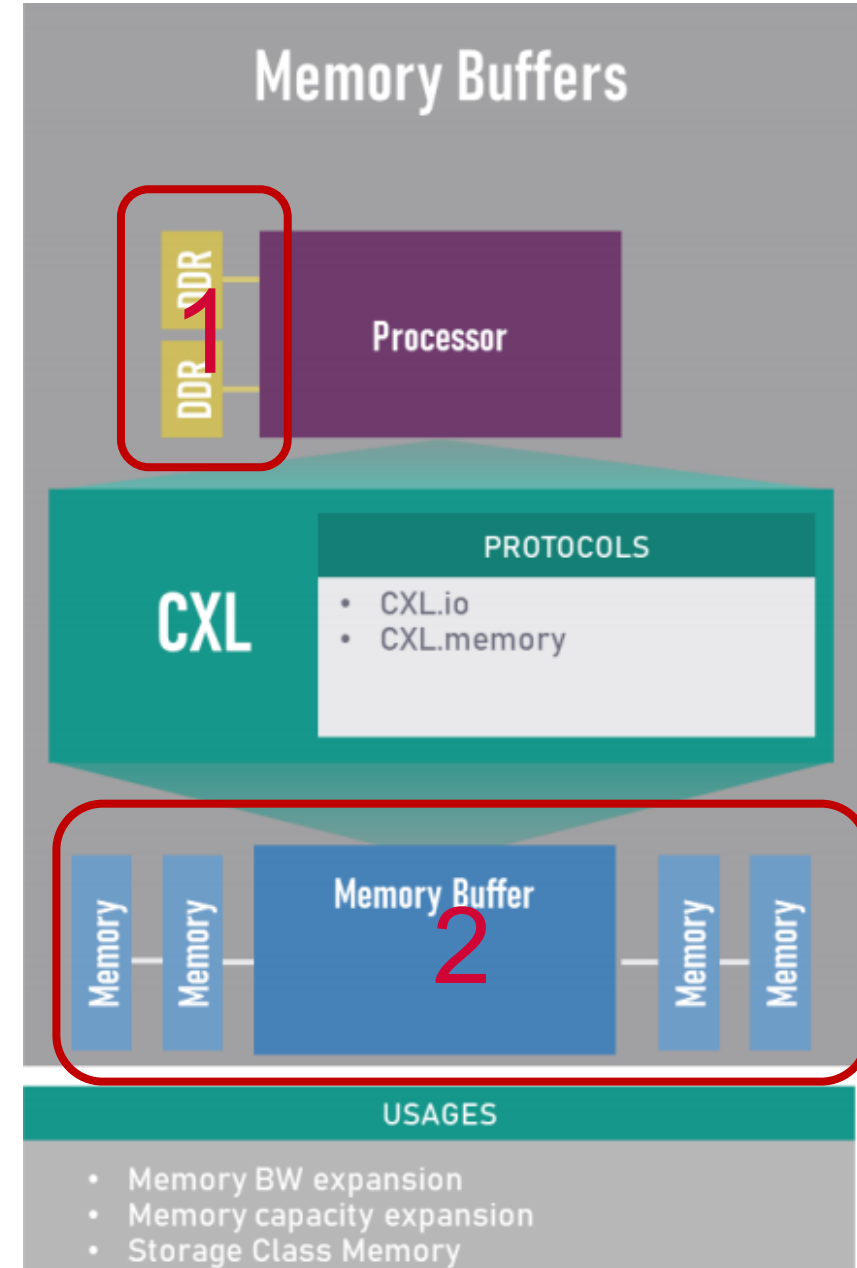
CXL Type 3 enables two-level memory

Extended Memory w/ What Tier 2 tech?

- DDR5
- Reused DDR4 (green & save money?)
- Emerging Memory Technologies

How manage?

- Auto-HW, e.g., Intel Flat Memory Mode
- App Aware (Explicit)



Auto-Magic Extended Memory Mgmt

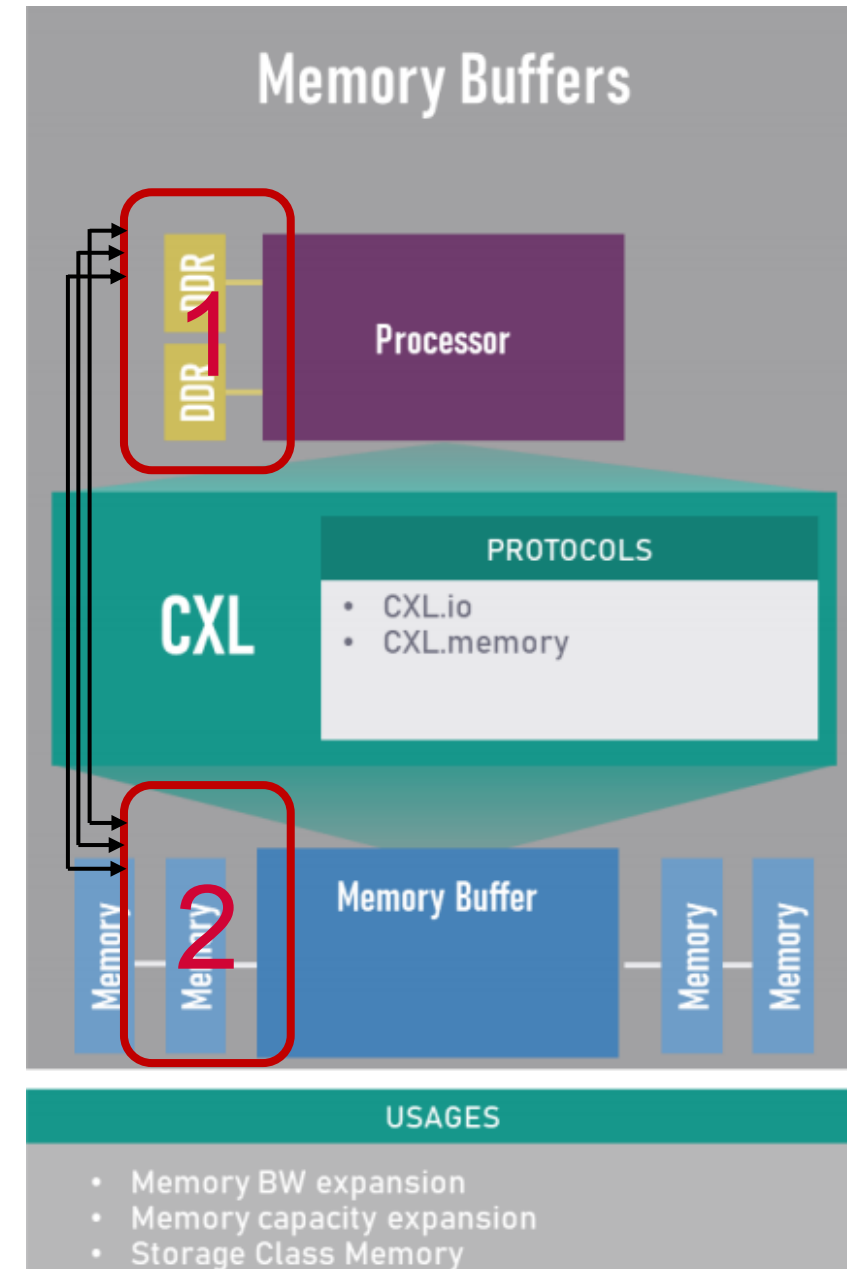
Intel Flat Memory Mode [HotChips'23]

- HW managed & SW transparent
→ Like a HW cache
- BUT SW sees Tier 1 + Tier 2 capacity
→ Like explicit memory

Details

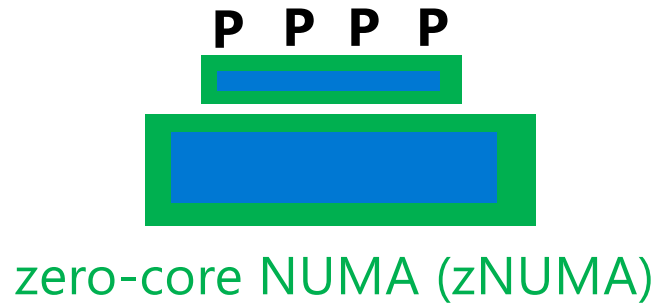
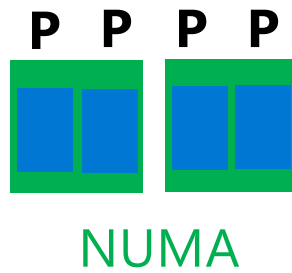
- Easiest if Tier 1 == Tier 2 capacity
- Memory access to Tier 1; swap 64 bytes on miss
- HW logically has “swap” bit per line
- Like a direct-mapped cache (behind SoC caches)

See: Managing Memory Tiers with CXL ... [OSDI 2024]



Explicit Extended Two-Tier Memory Mgmt

Applicable to important apps that care about performance
E.g., Relational DBMS **buffer pool**



Not shown:
zNUMA & NUMA

Key: P = CPU core

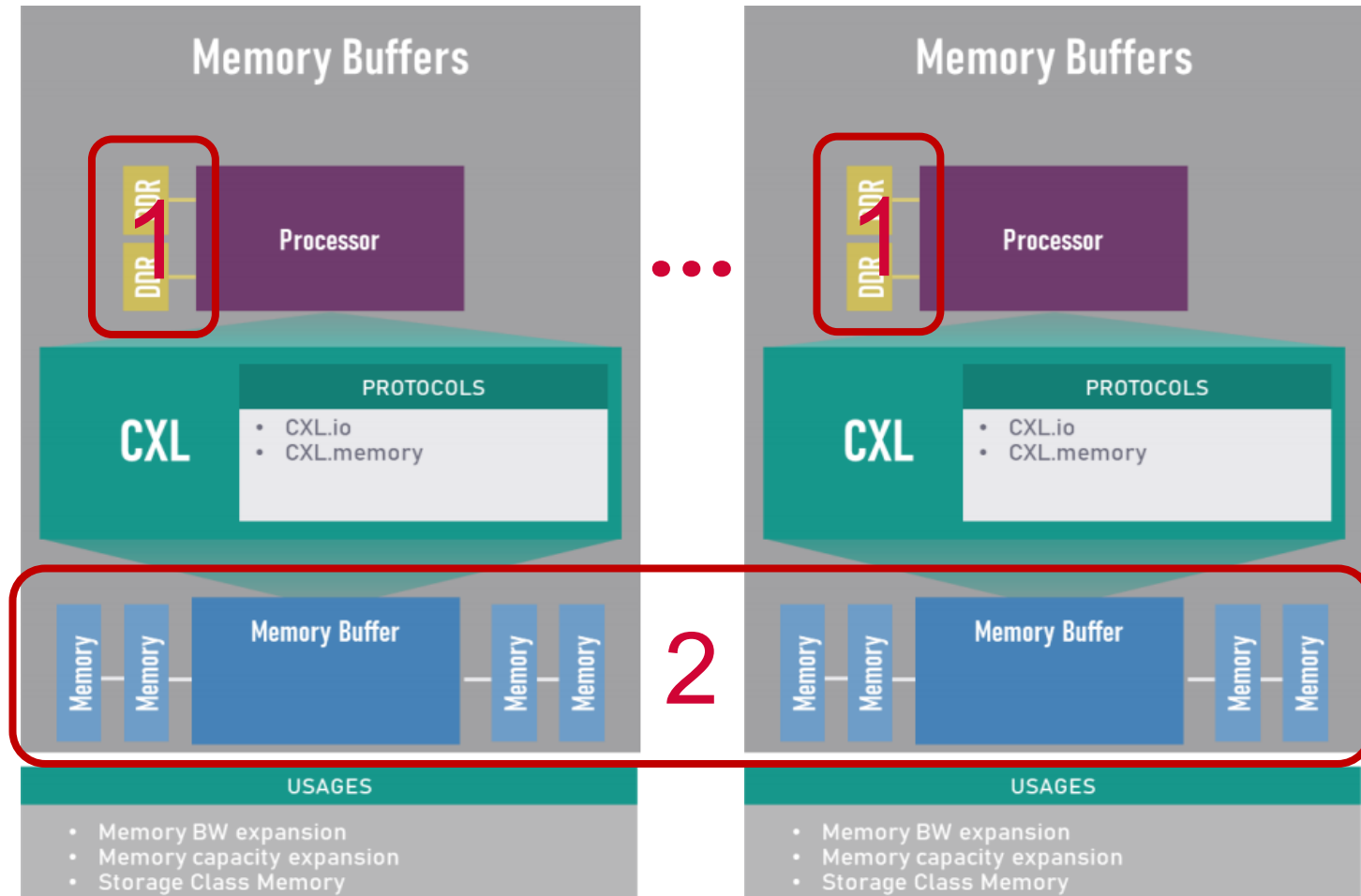
However, improving existing apps is playing **defense**.
What about playing **offense** with new CXL opportunities?

After CXL extended memory: Pooling & Sharing

Many-socket HW coherence **support withering**. What about analytic databases?

CXL Opportunity

- Connect several sockets to same CXL memory
1. **Pooling**: dynamic region accessed by one socket
 2. **Structured Sharing** with limited HW coherence (caching & messaging?)



Pond pooling [ASPLoS'23]
<https://arxiv.org/abs/2203.00241>

Memory: Processing In Memory (PIM)

Usually, move all data to CPU(s)

PIM: Move compute to vast data in memory

- **A high pain, high grain opportunity**

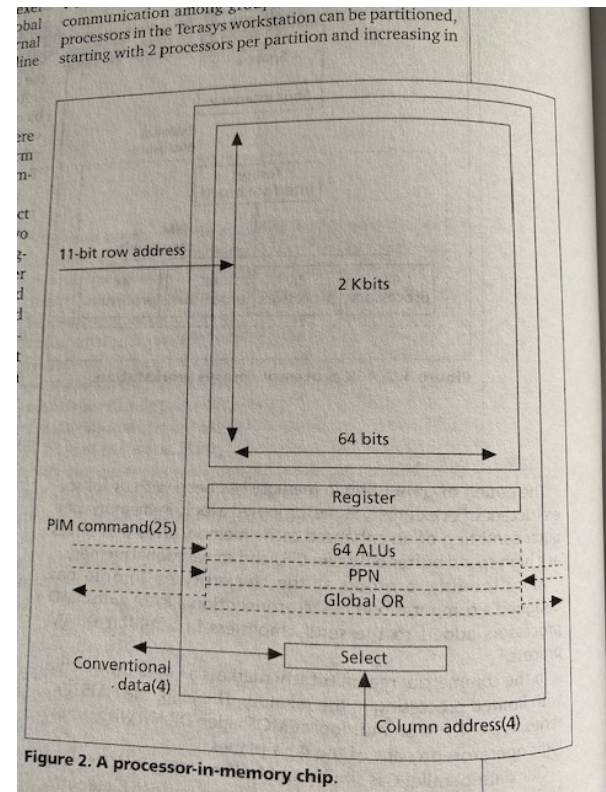
Old idea revived by

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

Alternatives: Processing {In, Near} Memory

Hardware Architecture and Software Stack for PIM
Based on Commercial DRAM Technology

(CO MATH) Sukhan Lee, et al., **Samsung**, ISCA Industrial Track, June 2021



Gokhale, Holmes, Iobst [1995]

PIM requires use cases with
small compute large corpus

Consider kernel → workflow

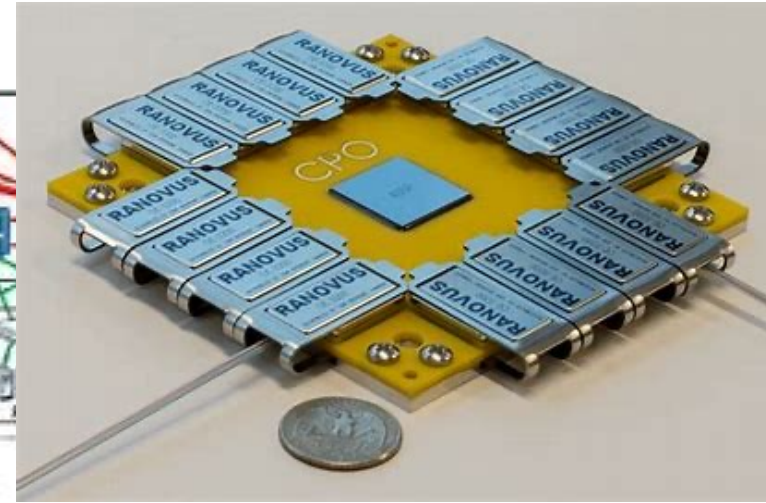
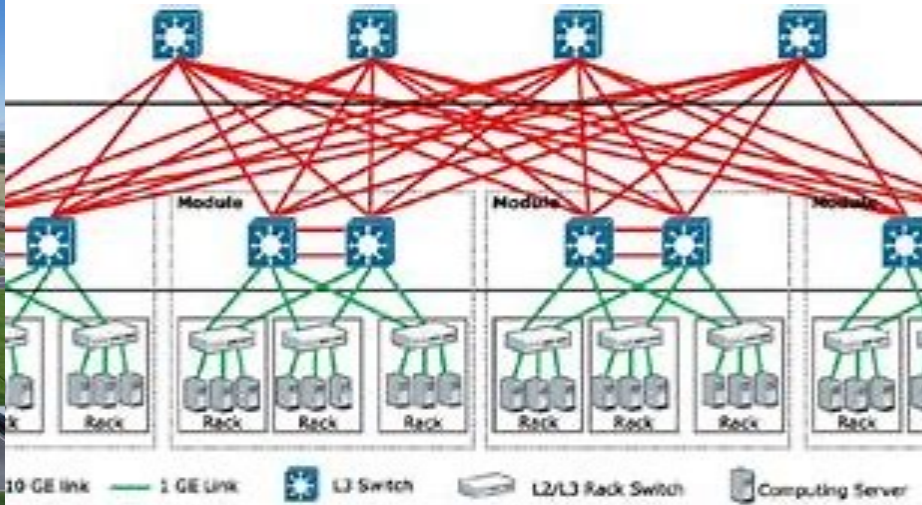
Computer Architecture's Eternal Questions & Outline

How best to do these
interacting factors:

- ~~1. Compute (longest)~~
- ~~2. Memory (longer)~~
3. Interconnect/networking ← CXL & UCIe
already covered
4. Storage
5. Security
6. Power
7. Cooling
8. *Bonus new question*



Data Center Networking: Main & Specialized (e.g., AI)



Want: Inexpensive, High Bandwidth, Reliable, Low Jitter, Low latency

Have: Inexpensive Ethernet & High-cost, single-source Infiniband

New Protocols: **Ultra Ethernet** (next slide)

New Technology

- **Optics** already used above top-of-rack switch (ToR)
- Evolving to replace electrical within rack then host maybe package
- **First use to replace in existing systems; later enable new systems**

Ultra Ethernet (<https://ultraethernet.org/>)



Started by Alphabet, AMD, Arista, Atos, Broadcom, Cisco, HPE, Intel, Meta, Microsoft, Oracle, but more now

Goal: Provide a high-performance low-cost Ethernet-based solution for emerging AI and other high-bandwidth low-latency workloads

Insight: Improve Ethernet by focusing

- Workloads in a data center
- Rather than arbitrary Internet

Targeted solutions: packet spraying, relaxed ordering, phase-aware congestion control, improved telemetry, refined software interfaces,....

Multiple switches & network interface cards (NICs) under development

Storage: Mind the Gaps



Solid State Drive



Hard Disk Drive



Tapes (2 vendors)



Persistent Memory?



Many-bit Cell, Appliance?



Microsoft Silica?

<https://www.microsoft.com/en-us/research/publication/project-silica-towards-sustainable-cloud-archival-storage-in-glass/> [SOSP2023]

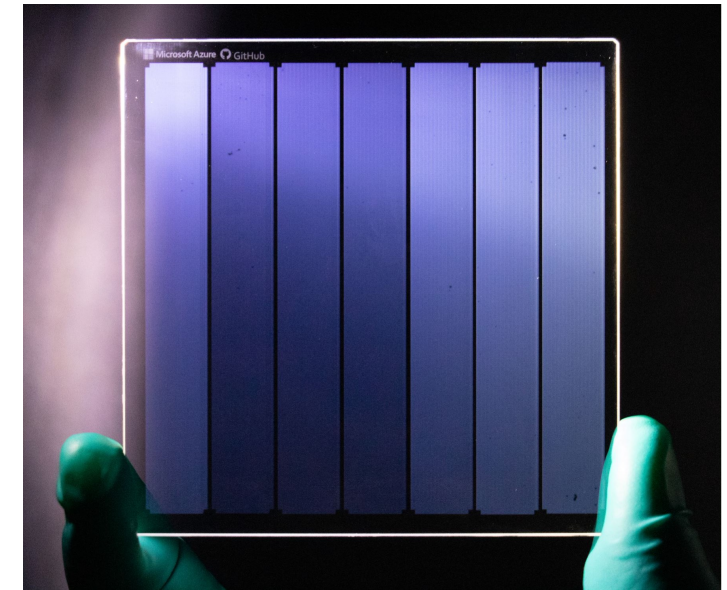
Microsoft Project Silica

Low-cost material: fused silica (quartz glass)

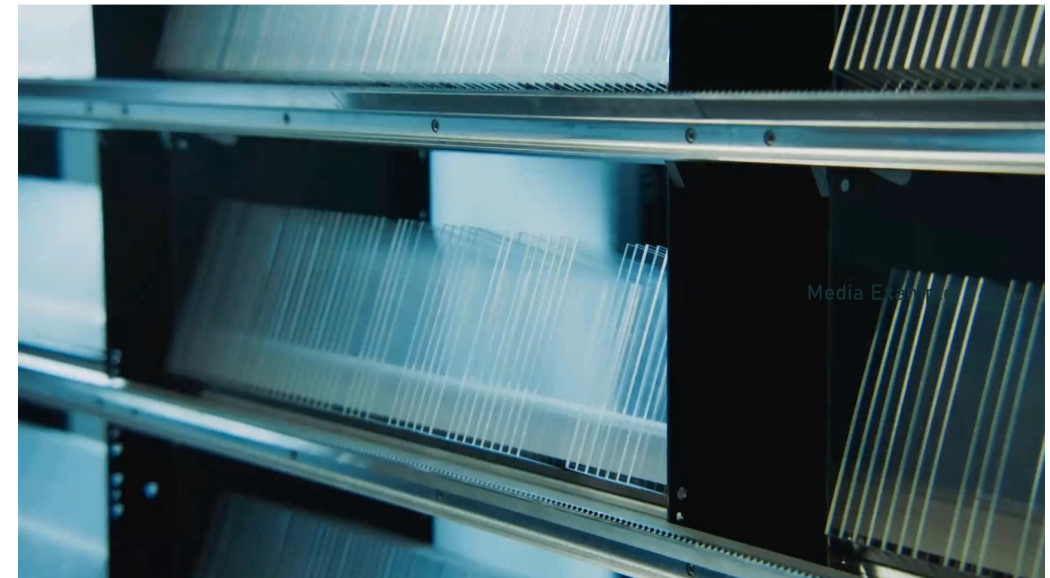
- Durable media
- Electromagnetic field-proof
- Write Once Read Many (WORM) media
- No bit/media rot
- Data lifetimes > 1,000s years
- No scrubbing required!

Data can be left in situ forever!

<https://www.microsoft.com/en-us/research/publication/project-silica-towards-sustainable-cloud-archival-storage-in-glass/> [SOSP2023]



Media Example



Library Concept

Security: Confidential Compute (CC)



Cloud Providers Now:

- Promise to protect your data/code from outsider/insider threats

With Confidential Compute

- **Your data/code is cryptographically protected from both threats**
- Hard: Root of trust, attestation, inter-package comm encrypted, memory/storage w/ data/address/replay protected, ...
- Can expand markets, but correctness/efficiency challenges

CC: <https://queue.acm.org/detail.cfm?id=3456125> [ACM Queue'21] & ACM Queue Jul/Aug'23 issue
OpenSource Root-of-Trust: <https://petri.com/microsoft-caliptra-open-source-root-of-trust/>
Azure Sphere (IoT): <https://aka.ms/7properties>

New ideas & accelerators must be compatible with CC.
E.g., accelerator trusted to manage tenant crypto keys

Power: IoT to Cloud Varies

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

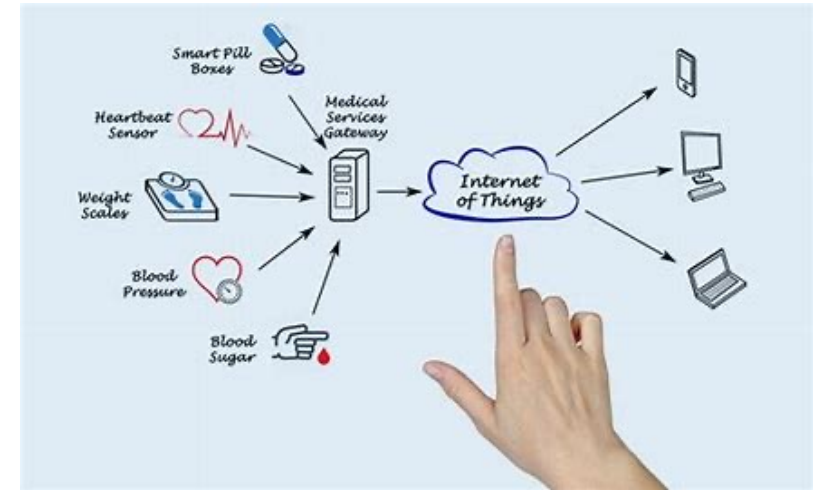
Wearables/IoT/Mobile: Energy (battery life)

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

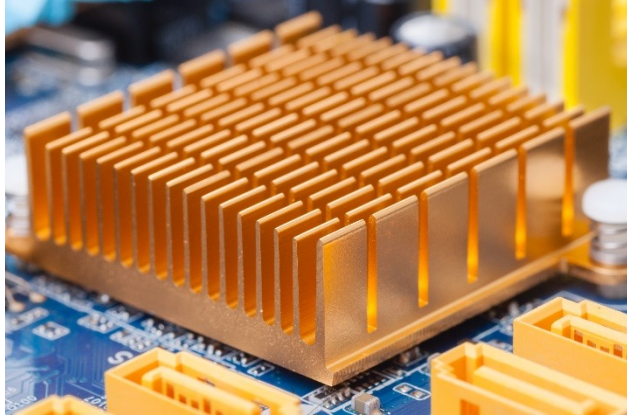
...

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
- MSFT[5/2023] contracts w/ Helion Fusion



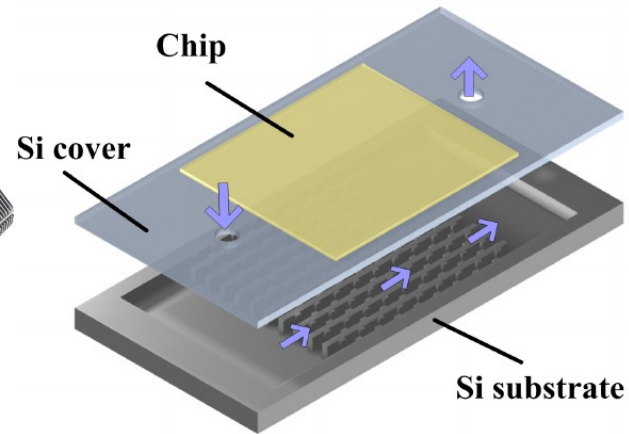
Cooling



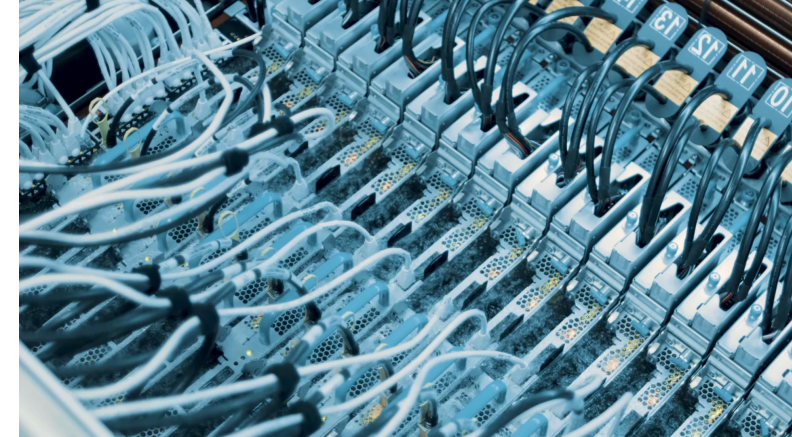
Air w/ heat sink



Cold plate



Microfluidics



Immersion (1 or 2 phase)

Data Centers are becoming gigantic supercomputers!

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

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

(Bonus) Sustainability!

How to reduce provisioned power (scope 2) & Si area (scope)? 3)?



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

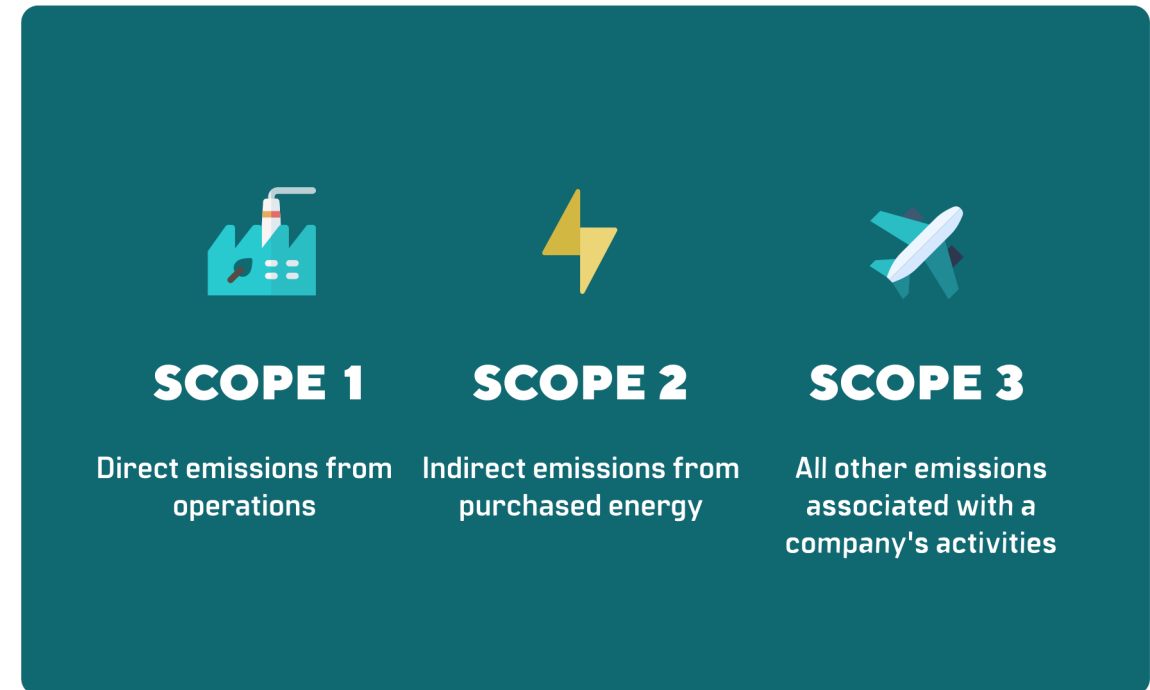
New: Make Computing More Sustainable?

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 Harvard & Facebook/Meta HPCA 2021 (<https://ieeexplore.ieee.org/document/9407142/>)
& ISCA 2022 (<https://dl.acm.org/doi/epdf/10.1145/3470496.3527408>)



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. **Interconnect/network:** protocols/optics
4. **Storage:** mind the gaps
5. **Security:** confidential compute
6. **Power:** IoT to cloud varies
7. **Cooling:** consider cold plate & its impact
8. **New: Sustainability:** whither emission scopes 1, 2, & 3?