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

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

@ Database Community Five-Year Meeting @ MIT, October 2023

This a public, non-proprietary talk.
I speak for myself, not necessarily Microsoft, Azure, or U. Wisconsin.
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 170 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.
Computer Architecture: Big Picture of Computer HW

Components → Systems

Gates → ALU → Functional Block → Core → SoC → Server → Data Center
Computer Architects: Components → Systems

My job: Hardware-software pathfinding for Azure
A View of Computing’s “Stack”

- Problem & Algorithms
- Applications
- DBMSs & Other Middleware
- Runtime & Compiler
- Operating System/Hypervisor
- (Micro) Architecture
- Hardware
- Materials & Fabrication

As technology scaling slows, dramatic perf/cost gains needed will require layer experts to work together (more)!
42 Years of Microprocessor Trend Data

Dennard Scaling Fades (2x transistors @ 1/2 power each)

Transistors (thousands)

Single-Thread Performance (SpecINT x 10^3)

Frequency (MHz)

Typical Power (Watts)

Number of Logical Cores

Year


A Commercial Computing Company Helix

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

For Data: cross-layer opt, embrace tech (ML & accelerators), which .com innovate?
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. Storage
4. Interconnect/networking
5. Security
6. Power
7. Cooling
8. *Bonus new question*

Gray boxes will have questions & implications to Database community
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

Google Tensor Processing Unit  Cerebras Wafer Scale Engine  Nvidia Grace-Hopper
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
- MFFP-12: \((8 + 16\times4)/16\)
  \(= 4.5\) bits/value
- Requires co-design

Compute: Accelerator-Level Parallelism

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

2019 Apple A12 w/ 42 Accelerators
Where to Accelerate?

Thanks to Ram Huggahalli
New Opportunity: Compute eXpress Link (CXL)

Caching Devices / Accelerators
- Processor
- CXL
  - CXL.io
  - CXL.cache
- Accelerator
- NIC
- Cache

Accelerators with Memory
- Processor
- CXL
  - CXL.io
  - CXL.cache
  - CXL.memory
- Accelerator
- HBM
- Cache

Memory Buffers
- Processor
- CXL
  - CXL.io
  - CXL.memory
- Memory Controller
- Memory
- Memory
- Memory

Enables DB accelerators “closer” than PCIe (coherent) & two-level memory
Not There Yet 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
• Standardized protocol among chiplets (physical/electrical/link/transport)
• Get closer: PCIe > CXL > UCIe
• Mix/match chiplets from different technologies/companies

➡ Enable innovation “on package” that was previously limited to chips on boards
Accelerating Compute for Data-Intensive Work

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

How? CPU SIMD (AVX) / Matrix (AMX), GPU, Custom (DB TPU?)
• Or FPGAs (between SW & hard HW on flexibility & speed)

Where?
• Granularity & data movement (see PIM later)
• Location in cloud data center as computer?

Key: Interactions with generative AI (accelerators)
Cloud + Data Apps + Generative AI

Amazing opportunity: sum >> parts
- Foundation models == means
- Customer value provided by data apps, etc.
  - Doing now: make tedious work → faster
  - Pot of gold: near impossible → practical

HW especially at Hyperscalar Clouds
- Massive special clusters for foundational AI training: GPUs, TPUs, ...
- Growing incremental training. Where?
- Exploding generative inference computation at low latency. Where?
- How connect to general-purpose (GP) systems that run data apps?
- AI clusters will consume massive power → less for GP & data apps

New use cases are paramount, and
> Efficiency → Enables providing value to more people in more ways
Computer Architecture’s Eternal Questions & Outline

How best to do these interacting factors:

1. **Compute** (longest)
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 Database community
Memory: Vast, Fast, Synchronous DDR ➔ Untenable

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)
New CXL Enables Two-Tier Memory

CPUs support CXL; innovation behind it

Pool Memory Among Nodes (not shown)
- Smooth per-node (overcommit) peaks
- Buy fewer bits; not cheaper per bit
- [https://arxiv.org/abs/2203.00241][ASPLOS’23]

Extended Memory
- More DRAM to avoid 3D-stacked DRAM
- Reuse older DRAM?
- How Manage? Auto-magic or 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)

Maybe good for general workloads; BUT…
**Two-Tier Memory for Data-Intensive Work**

**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

Implications to Database Community

Not shown: zNUMA & NUMA
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

Alternatives: Processing {In, Near} Memory

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
Computer Architecture’s Eternal Questions & Outline

How best to do these interacting factors:

1. **Compute** (longest)
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 Database community
Storage: Mind the Gaps

- Solid State Drive
- Hard Disk Drive
- Tapes (2 vendors)

Persistent Memory? Many-bit Cell, Appliance?

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

https://www.microsoft.com/en-us/research/project/dna-storage/
https://www.microsoft.com/en-us/research/project/project-silica/
Data Center Networking: Main & Specialized (e.g., AI)

Protocols
- Cost-effectively (Ethernet) → high perf computer interconnect
- BW, latency, jitter matter (Infiniband & Ultra Ethernet Consortium)

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

Azure Sphere (IoT): https://aka.ms/7properties

Even if business apps secure, how trust data middleware?
Niche for now: homomorphic encryption to compute encrypted
Power: IoT to Cloud Varies

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

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

Cooling

Data Centers are becoming supercomputers!

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

(Bonus) Sustainability!

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](https://www.epa.gov/ghgemissions)


Computer Architecture’s Eternal Questions & Outline

How best to do these interacting factors:

1. **Compute**: accelerators, deep learning, & many
2. **Memory**: 2D scaling dead, CXL, & 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 cold plate & its impact
8. **New**: **Sustainability**: whither emission scopes 1, 2, & 3?

Invent with purpose.

[https://careers.microsoft.com](https://careers.microsoft.com)