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

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

@ Stanford SystemX (Virtually), June 2021

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

What is Computer Architecture?

What are these eternal questions?
Computer Architecture: Big Picture of Computer HW

Components → Systems

Components:
- Gates
- ALU
- Functional Unit
- Core
- SoC
- Server
- Data Center
Computer Architecture’s Eternal Questions & Outline

How best to do these interacting factors:
1. Compute (3 slides)
2. Memory (2 slides)
3. Storage
4. Interconnect/networking
5. Security
6. Power
7. Cooling
8. *Bonus new question*
End of Dennard scaling & rise of demanding apps ➔

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

Google Tensor Processing Unit

Human Brain (esp. Neocortex)
Co-Design Nascent for Deep Learning Training/Interference

E.g., Microsoft Floating-Point
- 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?

2019 Apple A12 w/ **42 accelerators**

https://arxiv.org/abs/1907.02064 (to CACM)
DDR DRAM price not scaling ➔ poor 2D scaling

1. Use less memory: reduce waste, trade w/ compute, ...
2. Use tiering (CXL?) w/ what memory technology?
3. Need your answer here
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

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

Gokhale, Holmes, Iobst [1995]
Storage: Mind the Gaps

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/
Interconnects/Networking: Shorter Optics

Optics > Electrical at Distance D where time diminishes D

Was: Trans-Oceanic

Now: Intra Data Center

Future: Co-Packaged, But …
Security: Confidential Compute

Let $A \rightarrow B$ mean: $A$ must trust $B$

**Old:** Application $\rightarrow$ OS $\rightarrow$ Firmware $\rightarrow$ HW

**Cloud:** App $\rightarrow$ OS $\rightarrow$ VMM $\rightarrow$ Cloud Provider $\rightarrow$ FM $\rightarrow$ HW

**Confidential Compute:** App $\rightarrow$ HW subset (“TCB”)

**Need:** Root of trust, attestation, interchip comm encrypted, memory/storage w/ data/address/replay protected, ...

Many benefits but correctness/efficiency challenges

Azure CC: [https://queue.acm.org/detail.cfm?id=3456125](https://queue.acm.org/detail.cfm?id=3456125) [ACM Queue 2021]

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

Air Cooling Facing Limits
Cold Plate Coming

Then Immersion Cooling?

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

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?
- But intermittent, renewable power expanding
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?**

E.g., 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 & processing in memory
3. **Storage**: mind the gaps
4. **Interconnect/networking**: shorter optics
5. **Security**: confidential computer
6. **Power**: IoT to cloud varies
7. **Cooling**: consider immersion & its impact
8. **New**: **Sustainability**: whither emission scopes 1, 2, & 3?

Join us?
Backup Slides