Memories are Driving Big Architectural Changes. Hold Onto Your Hats!

Presented by

Jim Handy, Objective Analysis

Tom Coughlin, Coughlin Associates, President, IEEE

SNIA COMPUTE, MEMORY, AND STORAGE SUMMIT

Solutions, Architectures, and Community

VIRTUAL EVENT, MAY 21-22, 2024

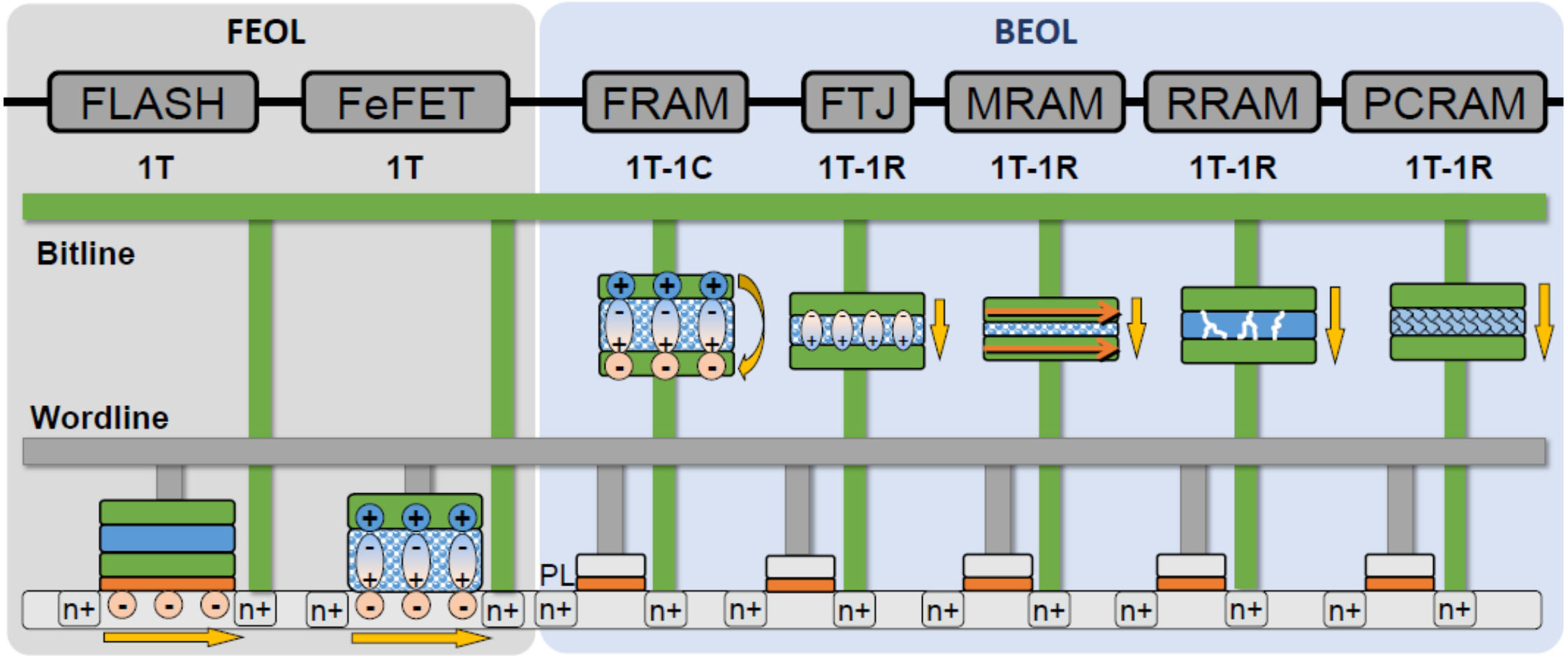
Outline

- Change 1: Emerging Memories Arrive
- Change 2: Memory Disaggregation
- Change 3: Processing in Memory
- Change 4: Chiplets
- Change 5: AI Everywhere

Outline

- Change 1: Emerging Memories Arrive
- Change 2: Memory Disaggregation
- Change 3: Processing in Memory
- Change 4: Chiplets
- Change 5: AI Everywhere

Emerging Memories Bear Strong Similarities



Graphic courtesy of Dr. Johannes Müller

What *is* an “Emerging Memory?”

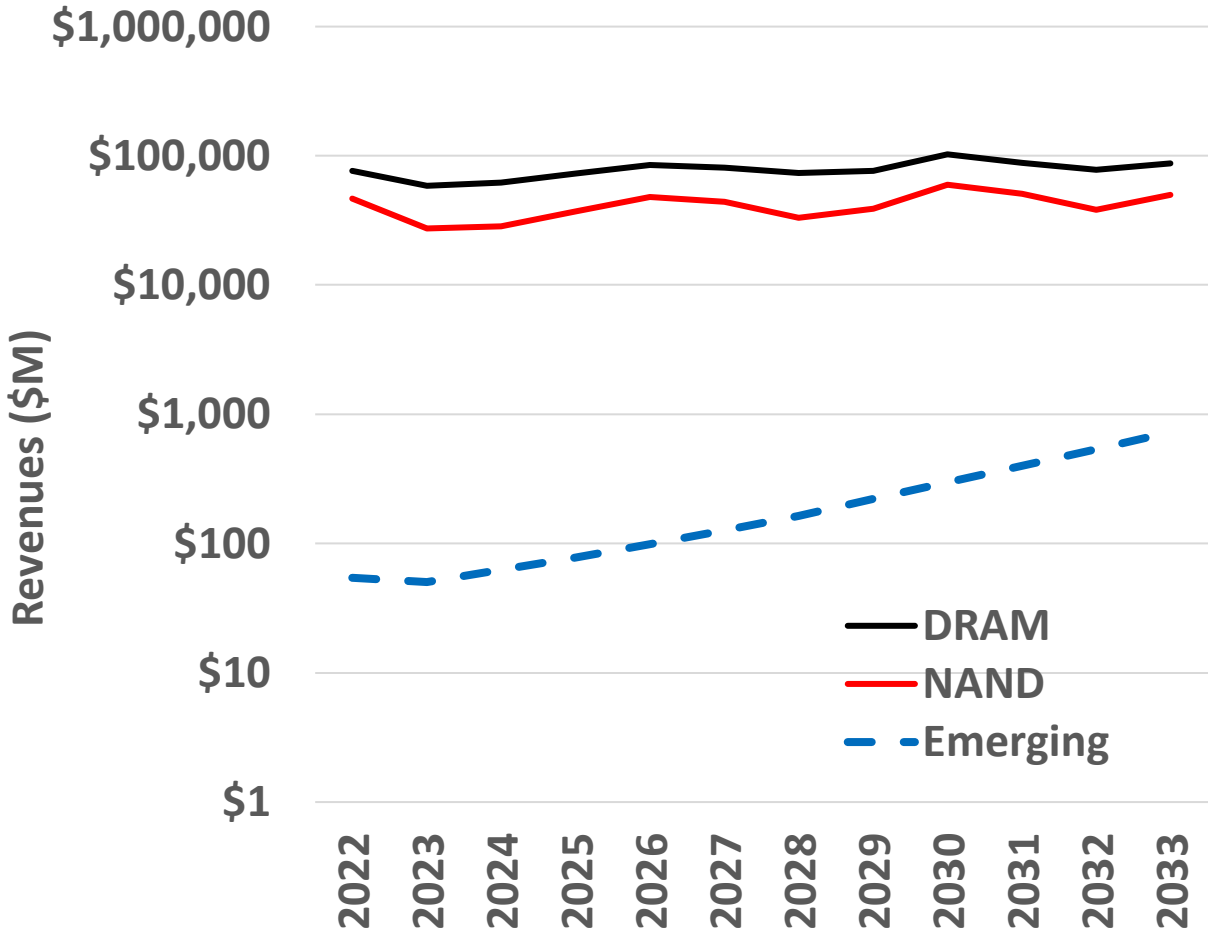
Established Memory

- Large market: $1B+
- Worrisome scaling limit
- Well-understood process
  - Based on silicon alone
- You know these names:
  - DRAM
  - NAND flash
  - NOR flash
  - SRAM
  - Specialty (FIFO, CAM…)

Emerging Memory

- Small Market: <<$1B
- Should scale significantly farther
- Process poorly understood
  - New materials almost universally used
- Do you know these names?
  - MRAM
  - ReRAM
  - PCM
  - FRAM
  - Exotics (CNT, compound semi…)
  - **ALL ARE NONVOLATILE!**

A Bright Future for Emerging Memories



- Emerging memory revenues will grow significantly faster than DRAM or NAND flash
  - DRAM: 4.5%
  - NAND: 6%
  - Emerging: 31%
- Wafers with embedded emerging memory exceed $100 billion in 2033

From: Emerging Memories Branch Out

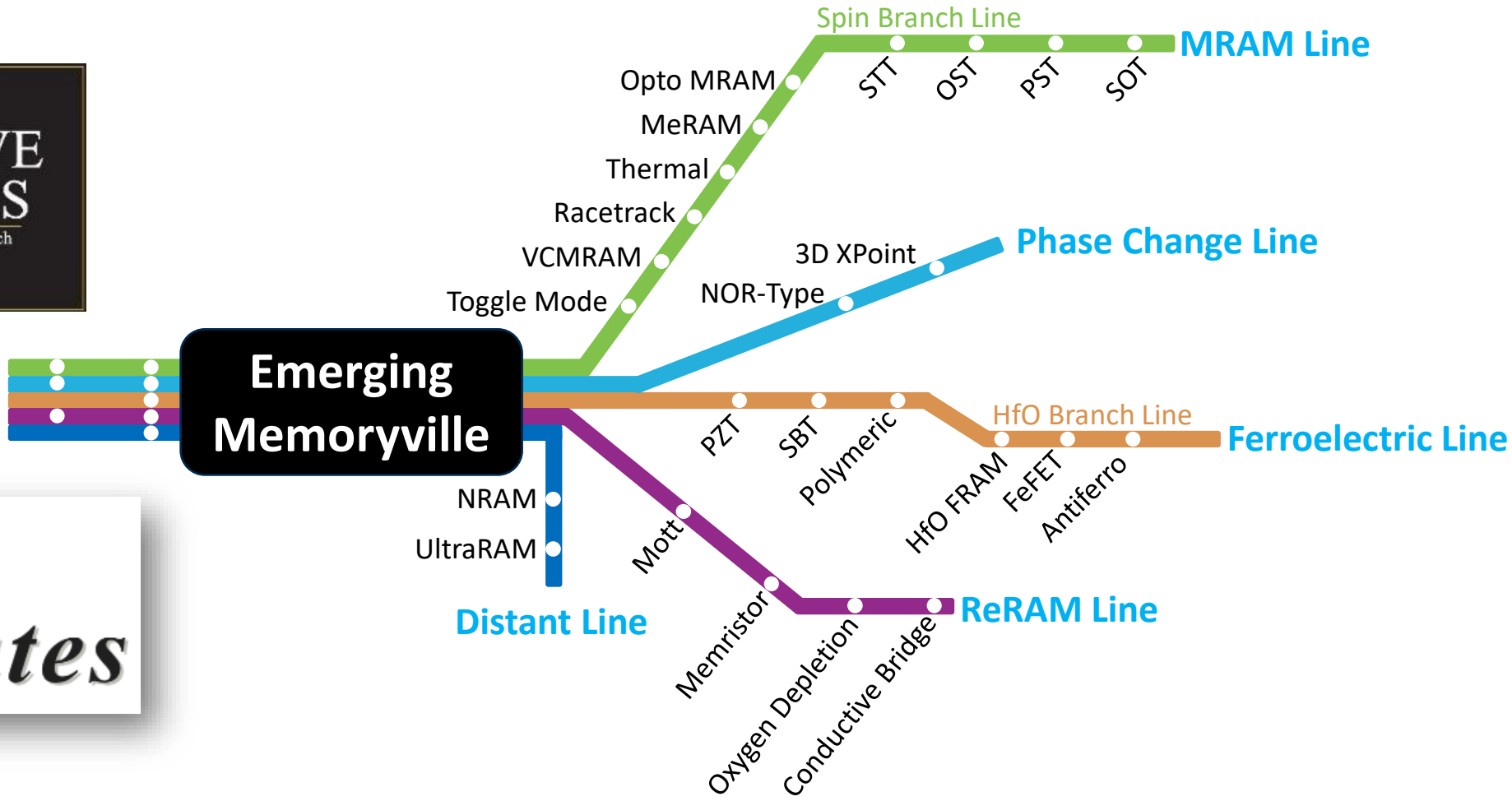
Emerging Memory's Impact on Computing

- The Optane Effect: Persistence comes closer to the processor
  - Beyond Optane: Persistence goes within the processor (more later)
  - Important speed and cost implications in certain applications
- Compelling power savings
  - No DRAM refresh
  - Nonvolatile/Persistent: Can be periodically powered down
  - Less heat to dissipate
  - Lower OpEx (energy savings) justifies higher CapEx (more expensive servers, etc.)

New Report: Emerging Memories Branch Out

OBJECTIVE ANALYSIS
Semiconductor Market Research

Coughlin Associates



Now Available!

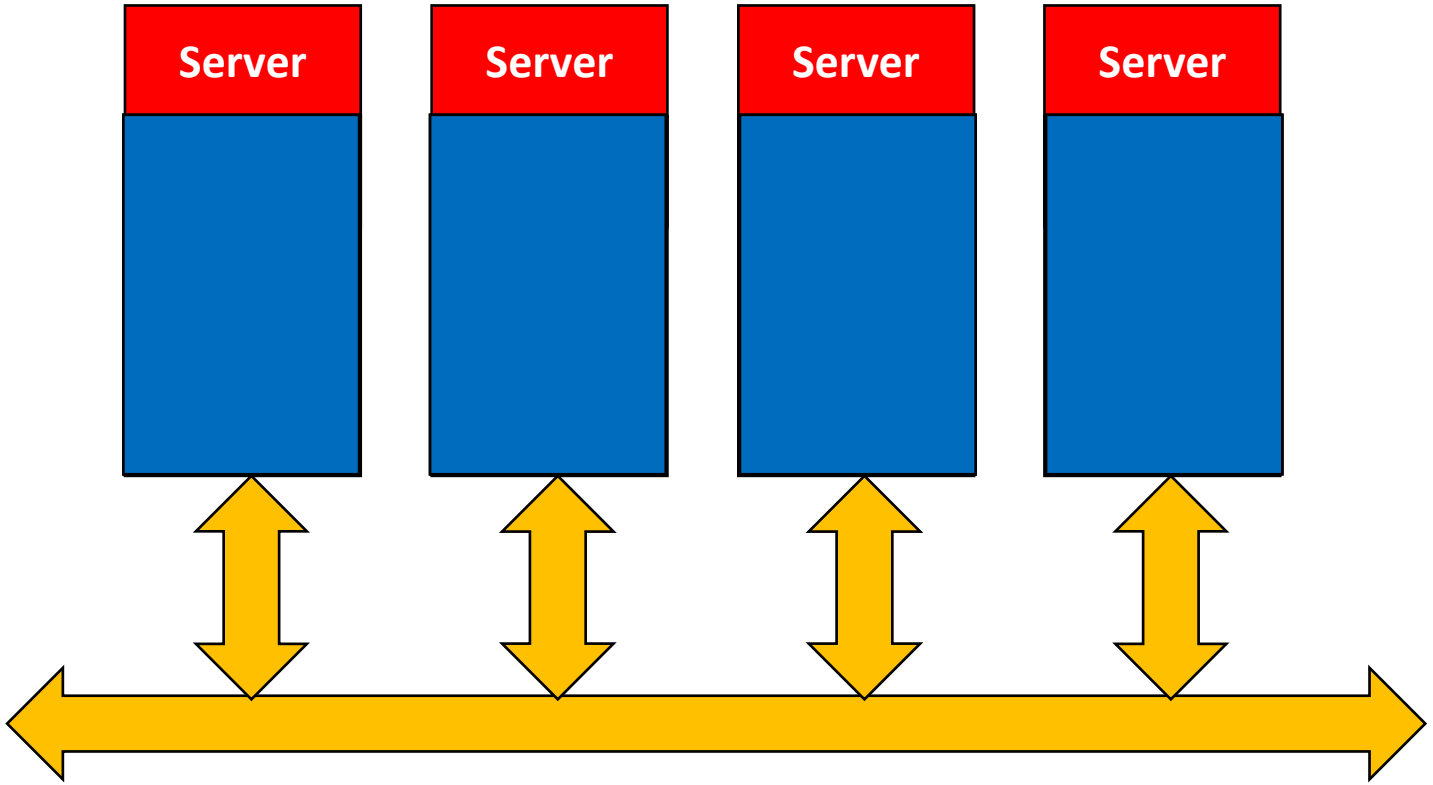
https://Objective-Analysis.com/reports/#Emerging

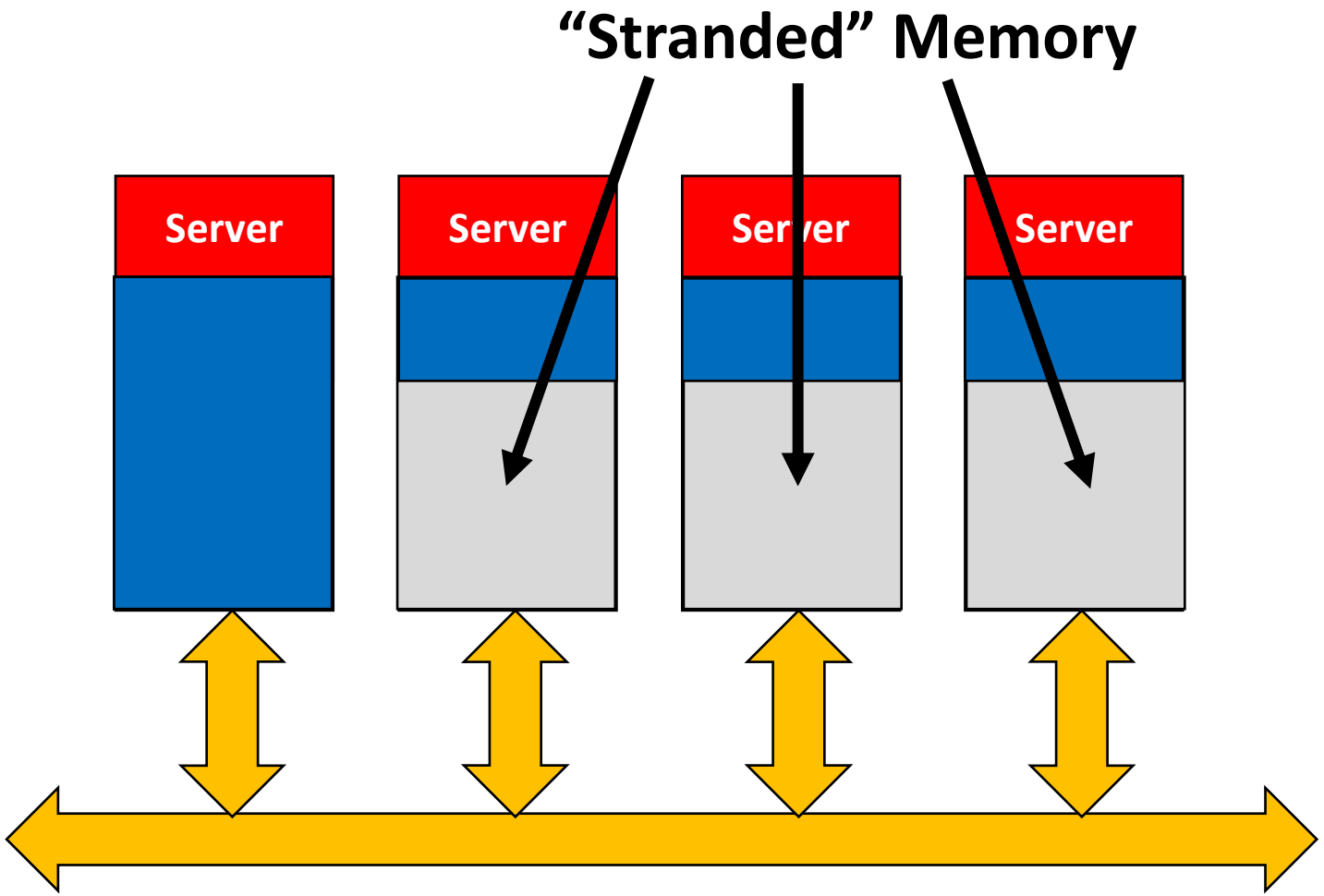
http://www.TomCoughlin.com/techpapers.htm

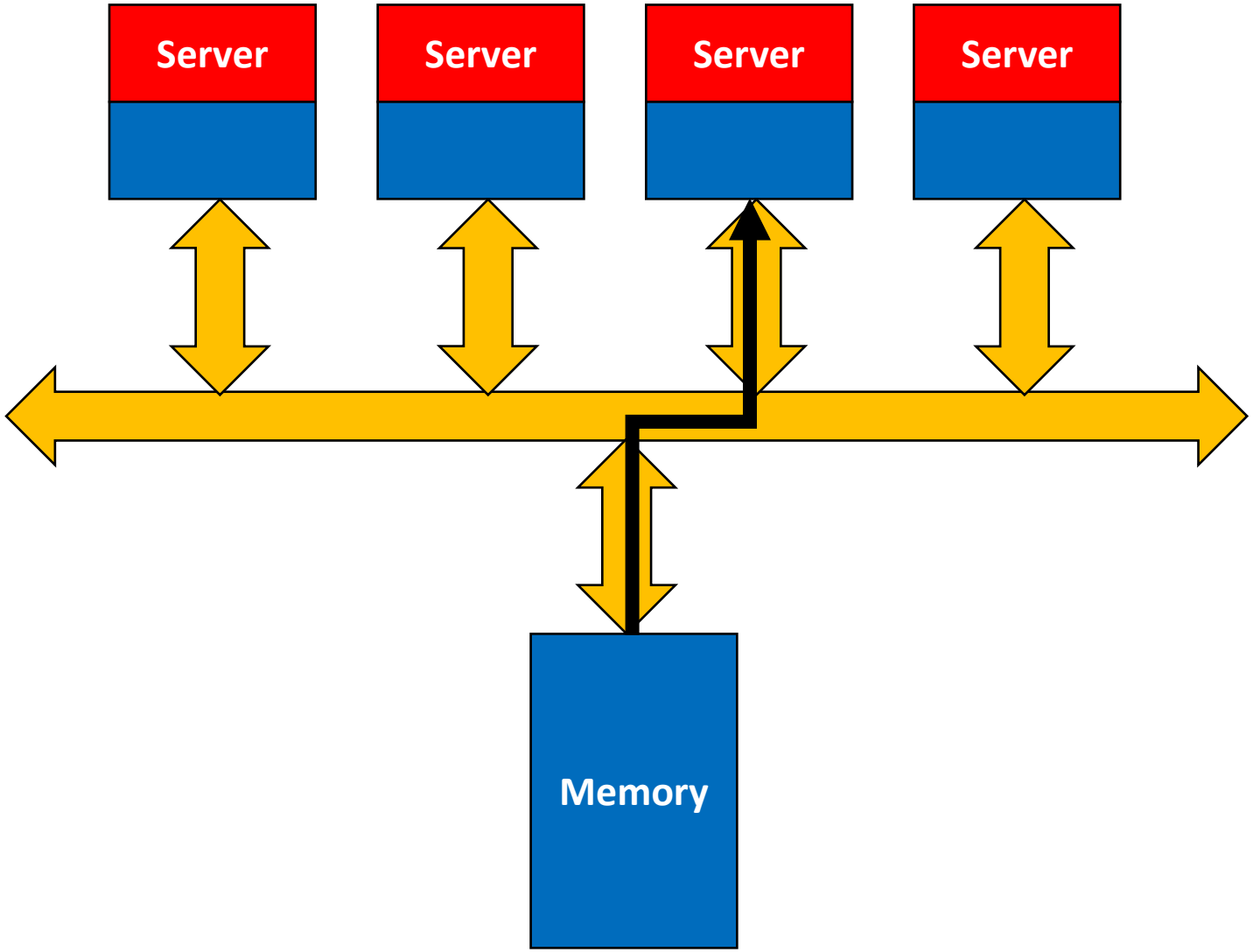
Outline

- Change 1: Emerging Memories Arrive
- Change 2: Memory Disaggregation
- Change 3: Processing in Memory
- Change 4: Chiplets
- Change 5: AI Everywhere

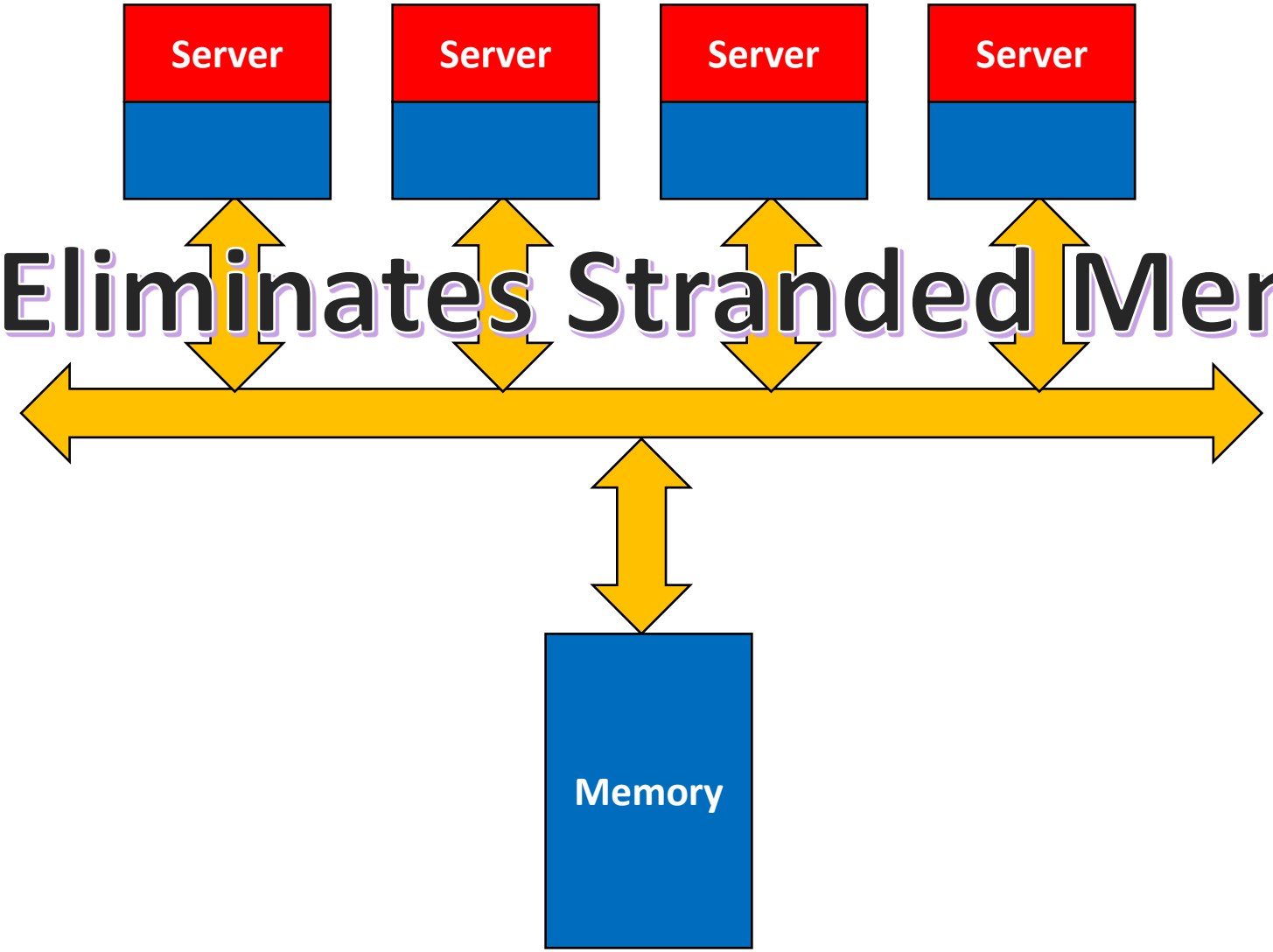
What is “Stranded Memory?”



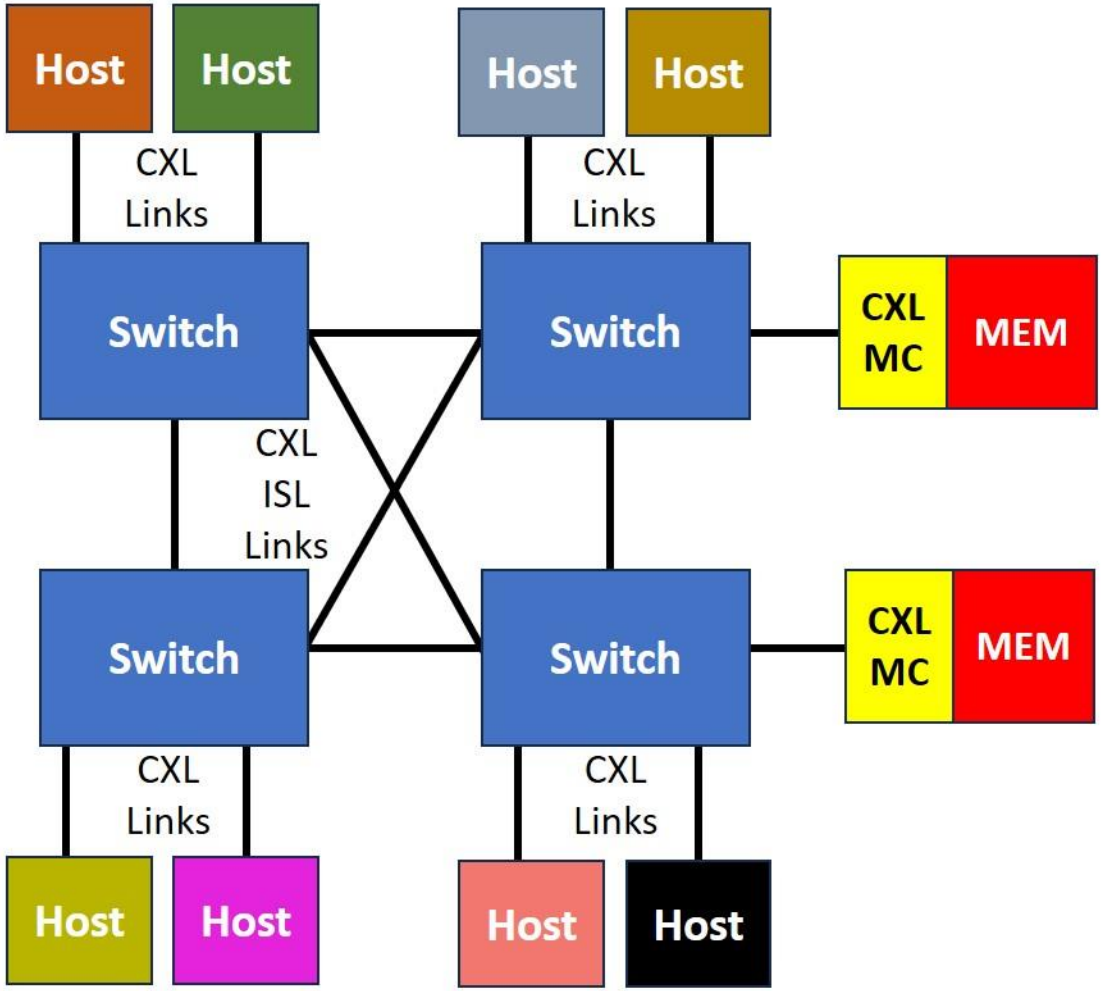




CXL Eliminates Stranded Memory



CXL Supports Memory Disaggregation



- Memory disaggregation – the Final Frontier!
  - First Servers, then Storage, now Memory
- CXL 2.0 & CXL 3.0 both support memory pools
  - No more “Stranded Memory”
  - Leads to savings in the datacenter

Computers After Memory Disaggregation

- Potential obsolescence of DIMMs
  - HBM + CXL is one approach
  - Soldered-down DRAM also possible
- Supports very uneven VMs
  - Small-memory VMs
  - Colossally-Gigantic-memory VMs

Outline

- Change 1: Emerging Memories Arrive
- Change 2: Memory Disaggregation
- Change 3: Processing in Memory
- Change 4: Chiplets
- Change 5: AI Everywhere

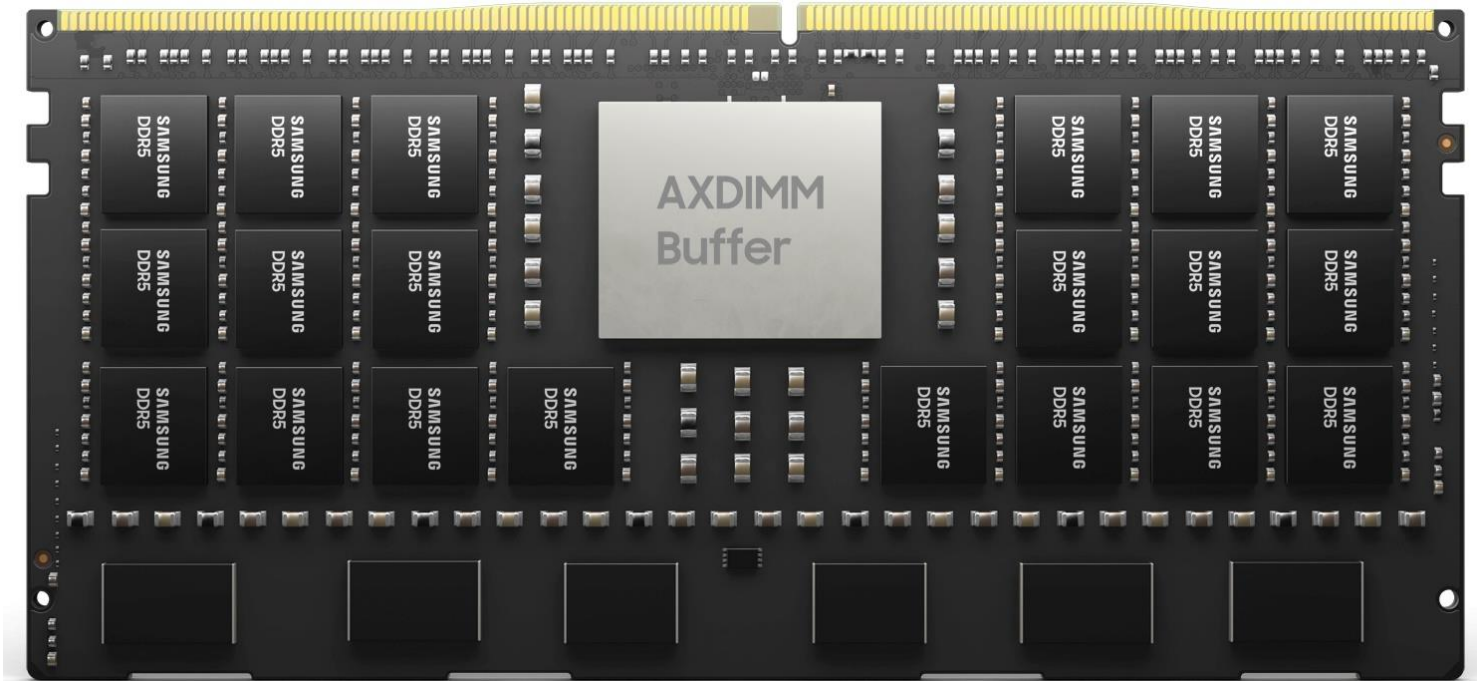
What is “Processing In Memory”?

- It depends on who you talk to
  - DIMMs with DRAM and a processor chip
  - Chips with DRAM & an internal processor
  - Chips with processing logic inside the memory bit cells
  - Analog neural net chips

Goal is to reduce data movement

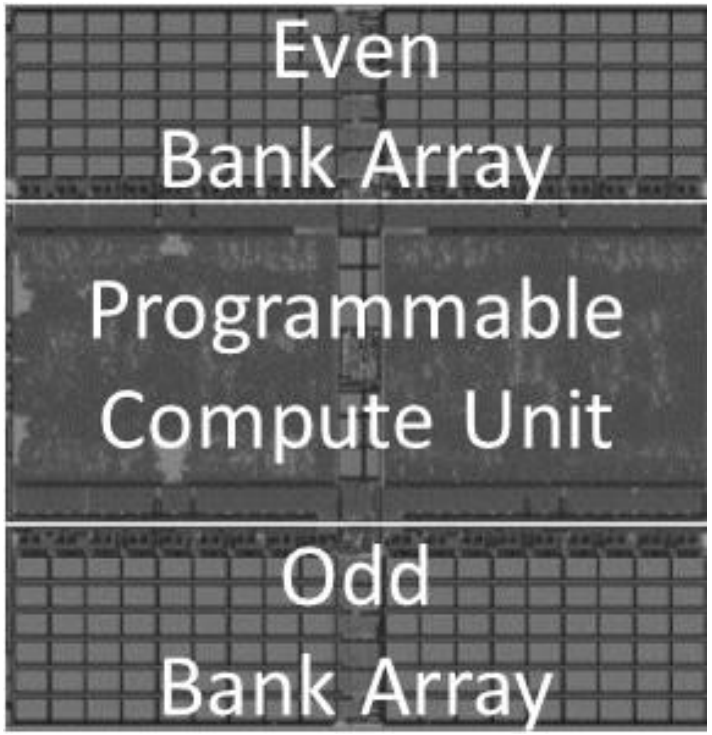
DIMMs with an Internal Processor

Samsung AXDIMM

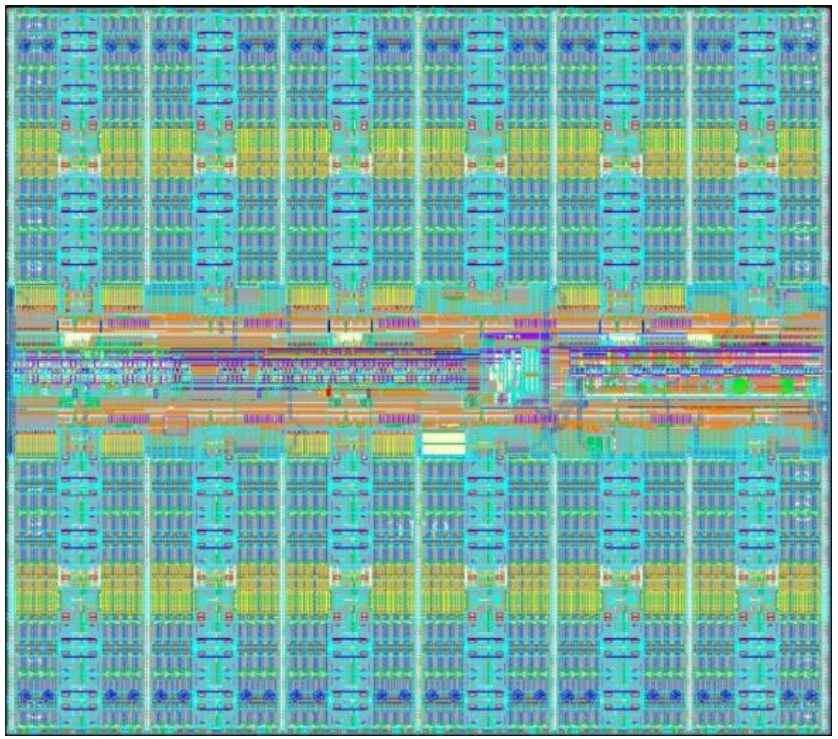


DRAM Chips with Internal Processor

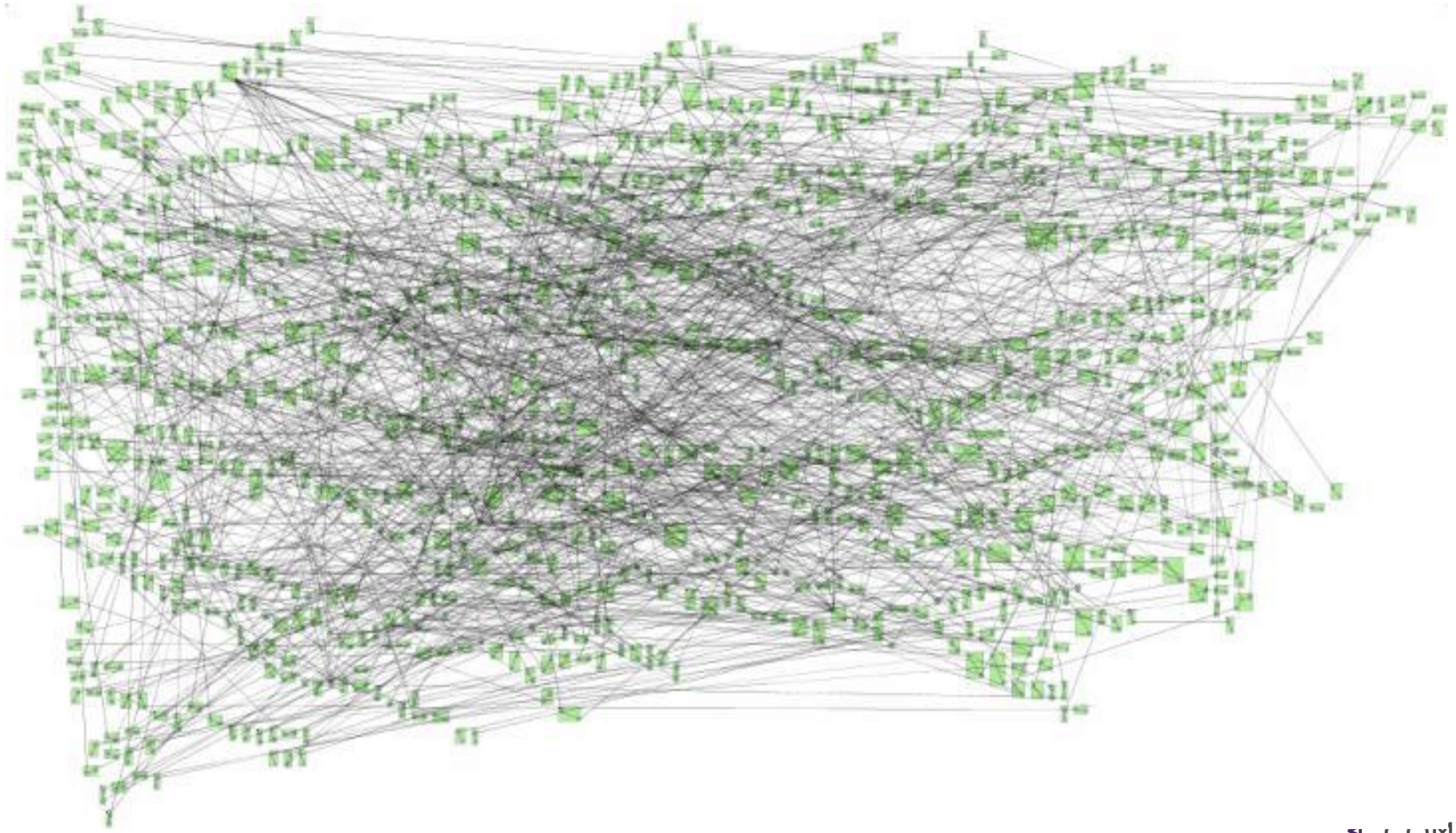
Samsung AXDIMM



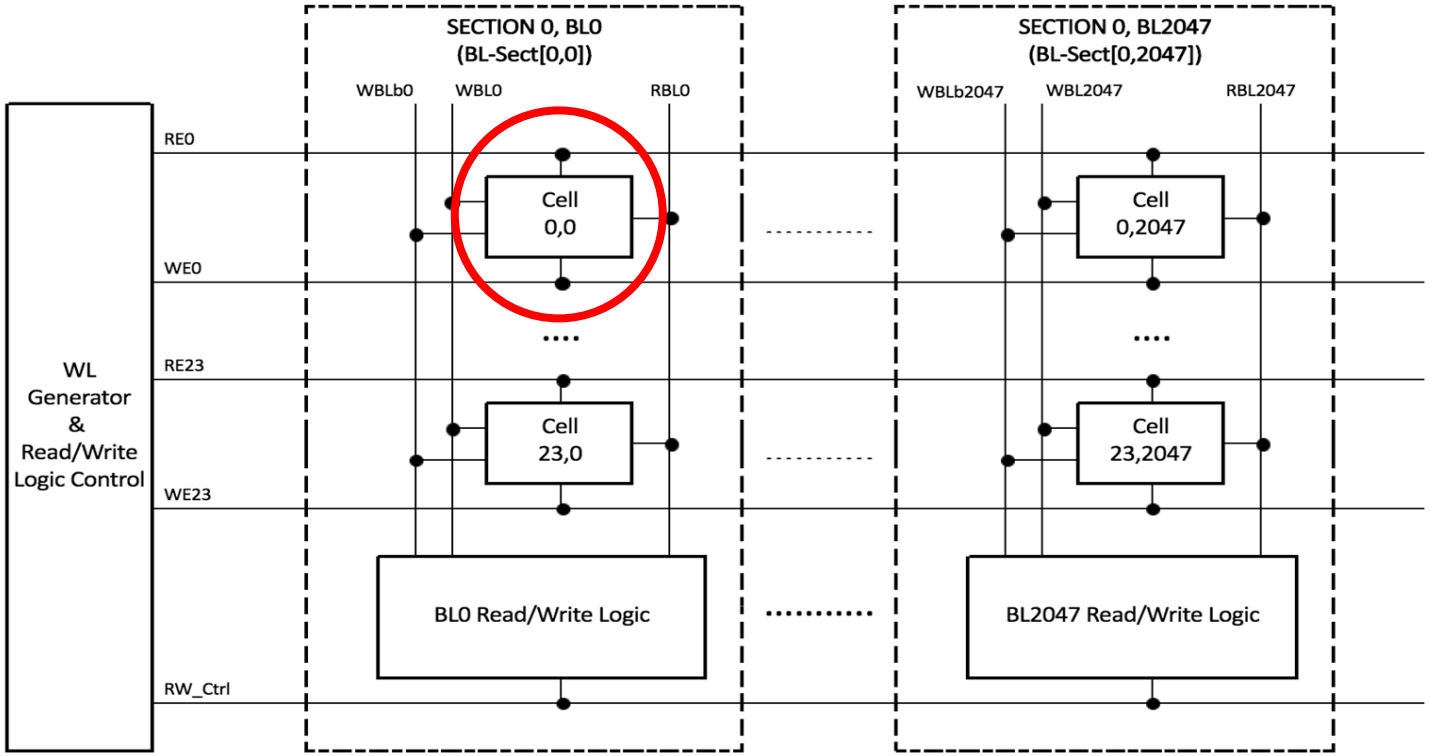
Natural Intelligence Automaton



DRAM Chips with Internal Processor

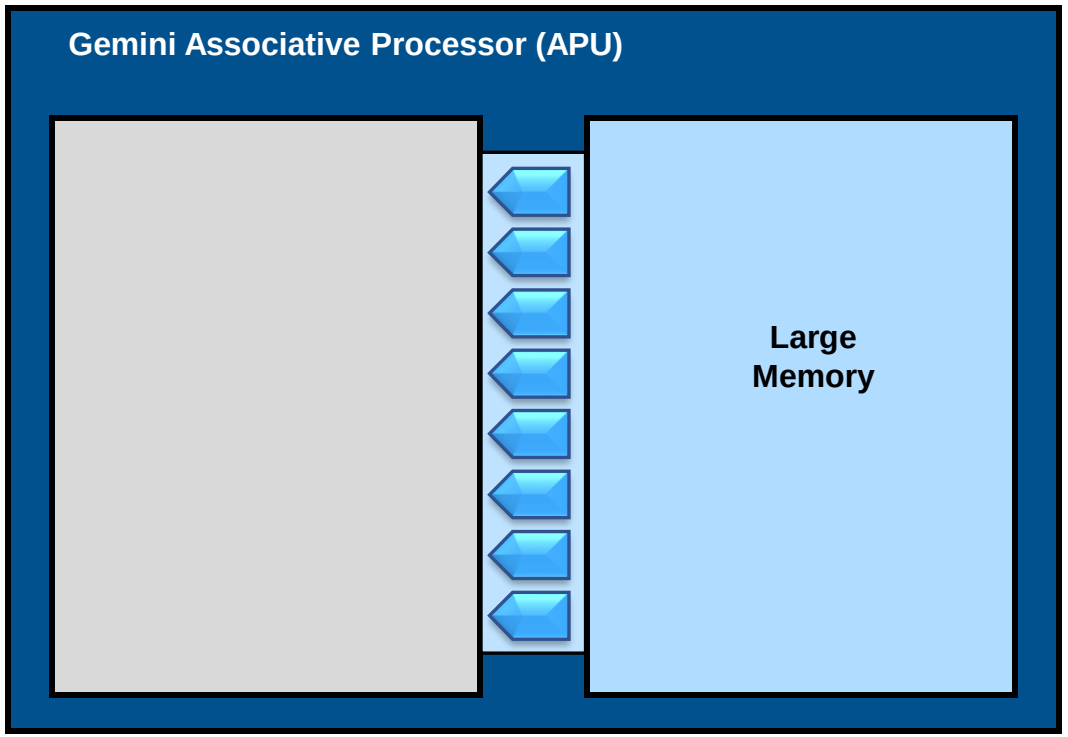


Processing Within the Memory Bit Cell

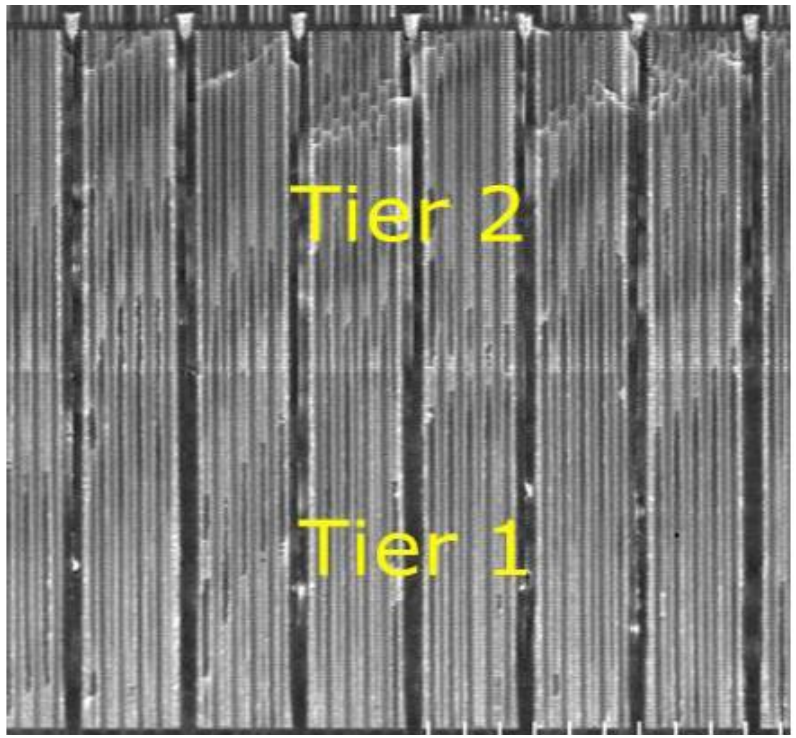


Processing Within the Memory Bit Cell

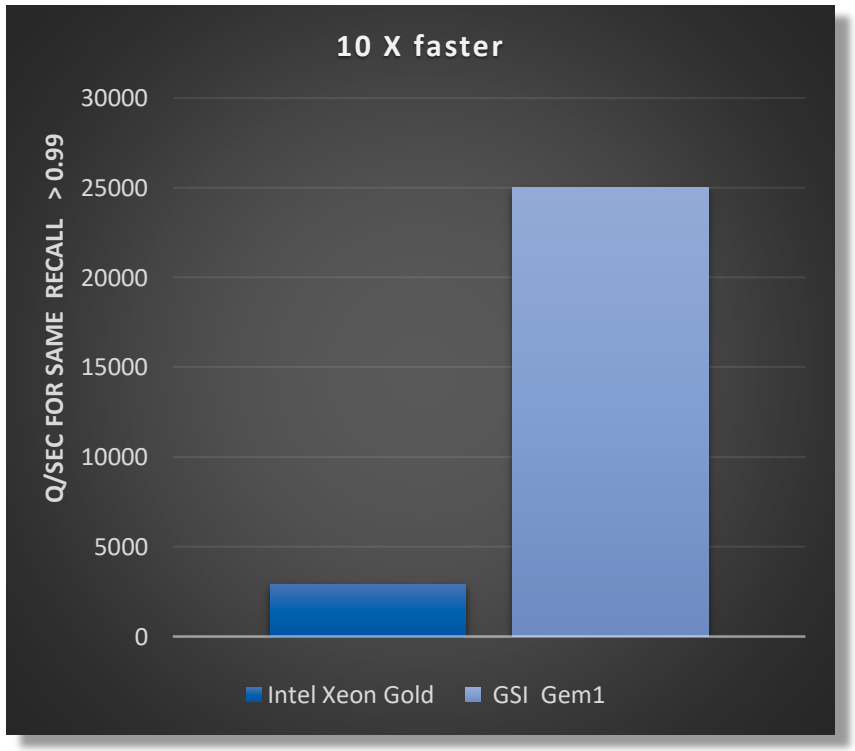
GSI Gemini APU

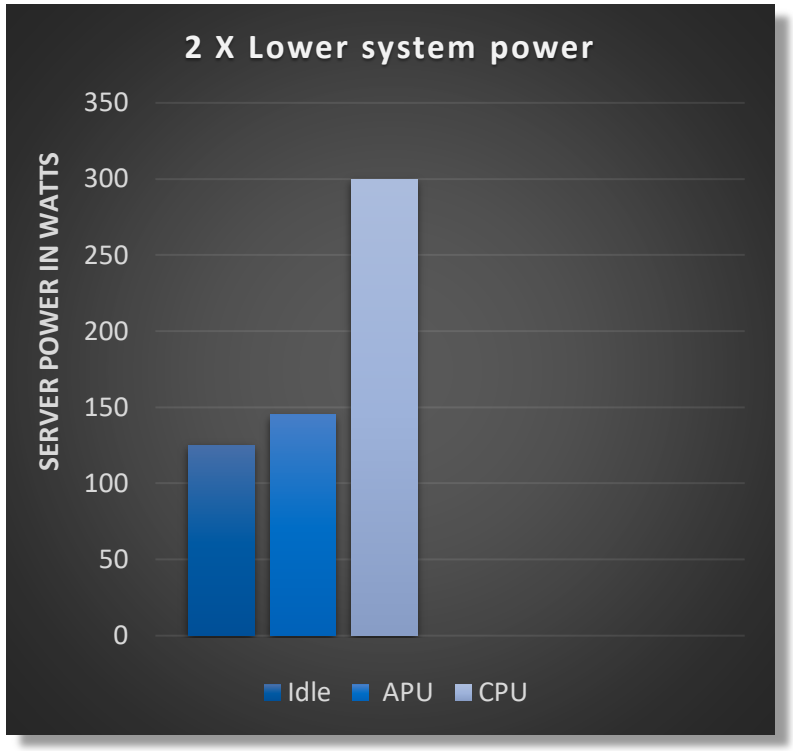


Macronix FortiX



Processing Within the Memory Bit Cell

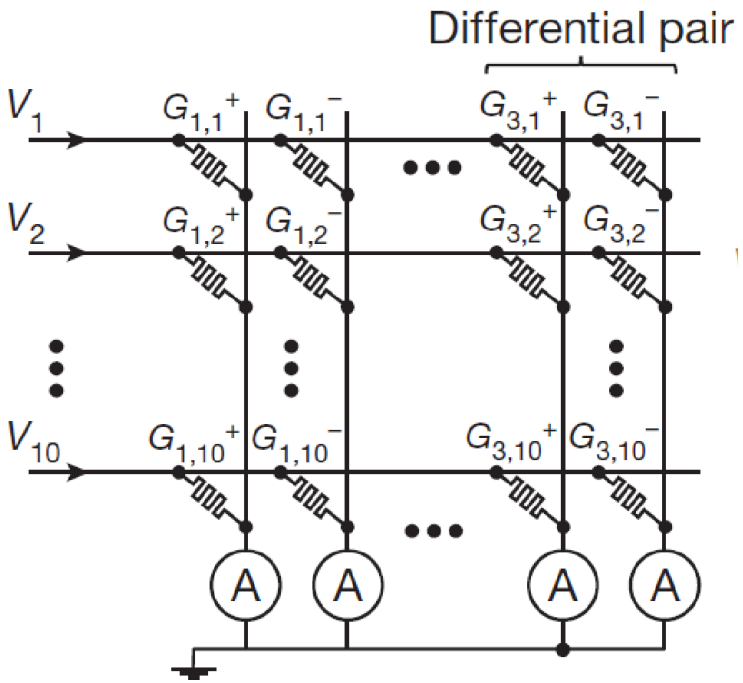




Analog Neural Nets

5.5E-07	3.7E-07	4.2E-07	3.5E-07	5.0E-07	4.7E-07	4.4E-07	5.0E-07
3.7E-07	3.6E-07	6.3E-07	3.7E-07	4.1E-07	4.2E-07	5.3E-07	3.3E-07
4.6E-07	5.7E-07	5.4E-07	4.9E-07	4.9E-07	4.2E-07	5.6E-07	6.0E-07
4.6E-07	4.2E-07	3.6E-07	3.1E-07	2.7E-07	3.7E-07	4.4E-07	3.7E-07
3.5E-07	4.0E-07	5.8E-07	4.8E-07	6.5E-07	4.1E-07	4.0E-07	4.4E-07
4.5E-07	3.8E-07	5.4E-07	4.7E-07	5.9E-07	4.6E-07	4.7E-07	4.8E-07
3.6E-07	4.1E-07	4.5E-07	3.9E-07	5.0E-07	3.6E-07	5.6E-07	4.8E-07
3.5E-07	4.0E-07	4.3E-07	4.1E-07	3.5E-07	4.4E-07	4.6E-07	3.7E-07
4.9E-07	3.7E-07	6.0E-07	3.6E-07	3.3E-07	5.1E-07	3.9E-07	4.2E-07
4.4E-07	3.3E-07	3.3E-07	4.0E-07	3.9E-07	4.5E-07	4.3E-07	4.4E-07

- Old idea seeing renewed interest
- Instant Matrix Algebra
  - Single cycle math for entire matrix multiply
- Simple operation
  - Difficult to set up
- A good accelerator to a standard CPU
- Fits emerging memories well
- Lots of research, but few products



One or More of these May Succeed

- Prospective co-processor to offload tasks from a standard CPU
- Possibly the main processor in smaller edge applications
  - Low-end CPU or MPU will perform housekeeping
- Could slow CPU performance increases
  - CPU only needed for less significant & slower tasks
  - More complicated tasks offloaded to the PIM chip
  - Overall cost savings result
- Performance scales with the addition of PIM chips
  - One CPU per 10/100/1,000 PIMs

Outline

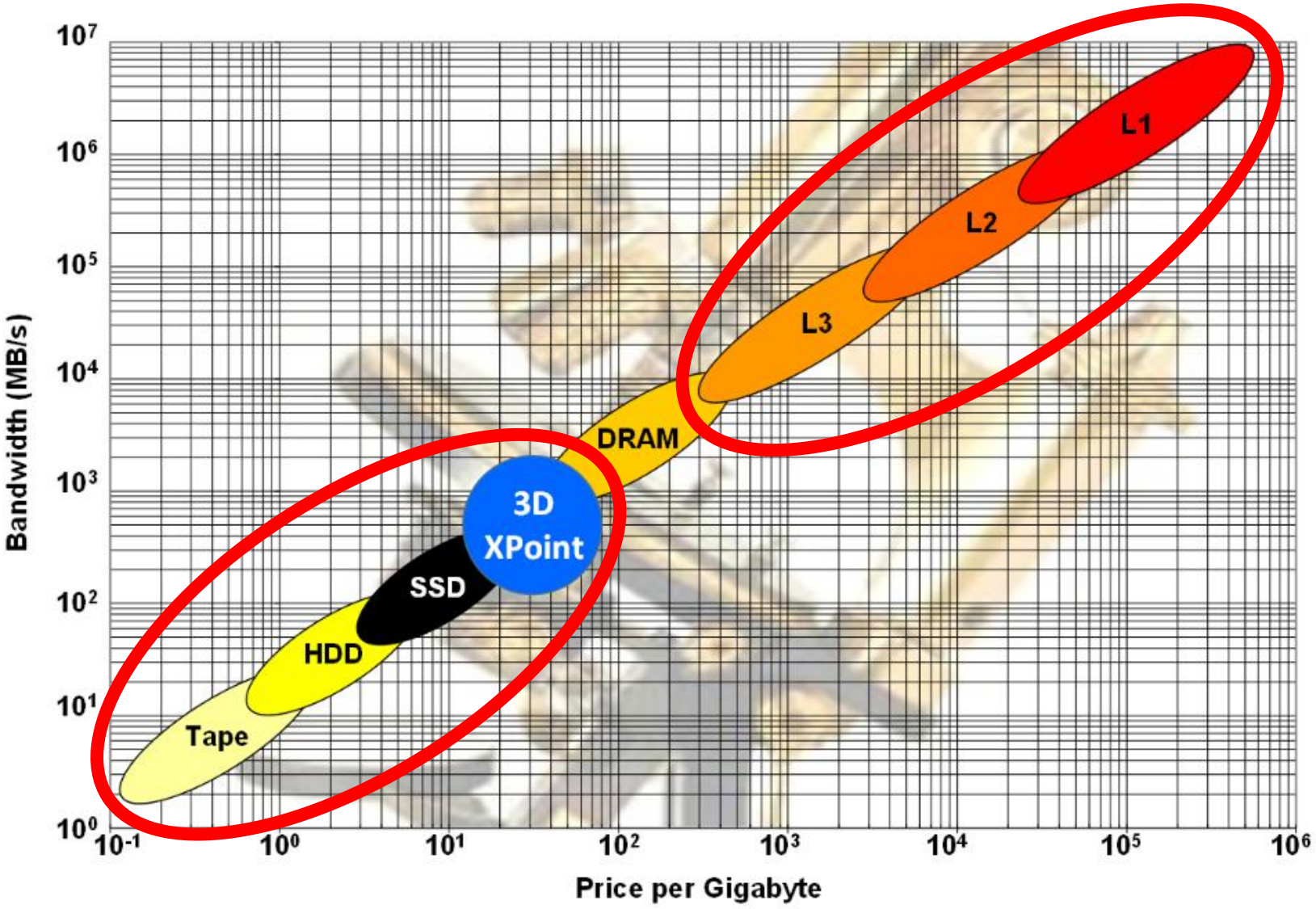
- Change 1: Emerging Memories Arrive
- Change 2: Memory Disaggregation
- Change 3: Processing in Memory
- Change 4: Chiplets
- Change 5: AI Everywhere

NOR Flash Scaling has Stopped
SRAM Scaling is Slowing

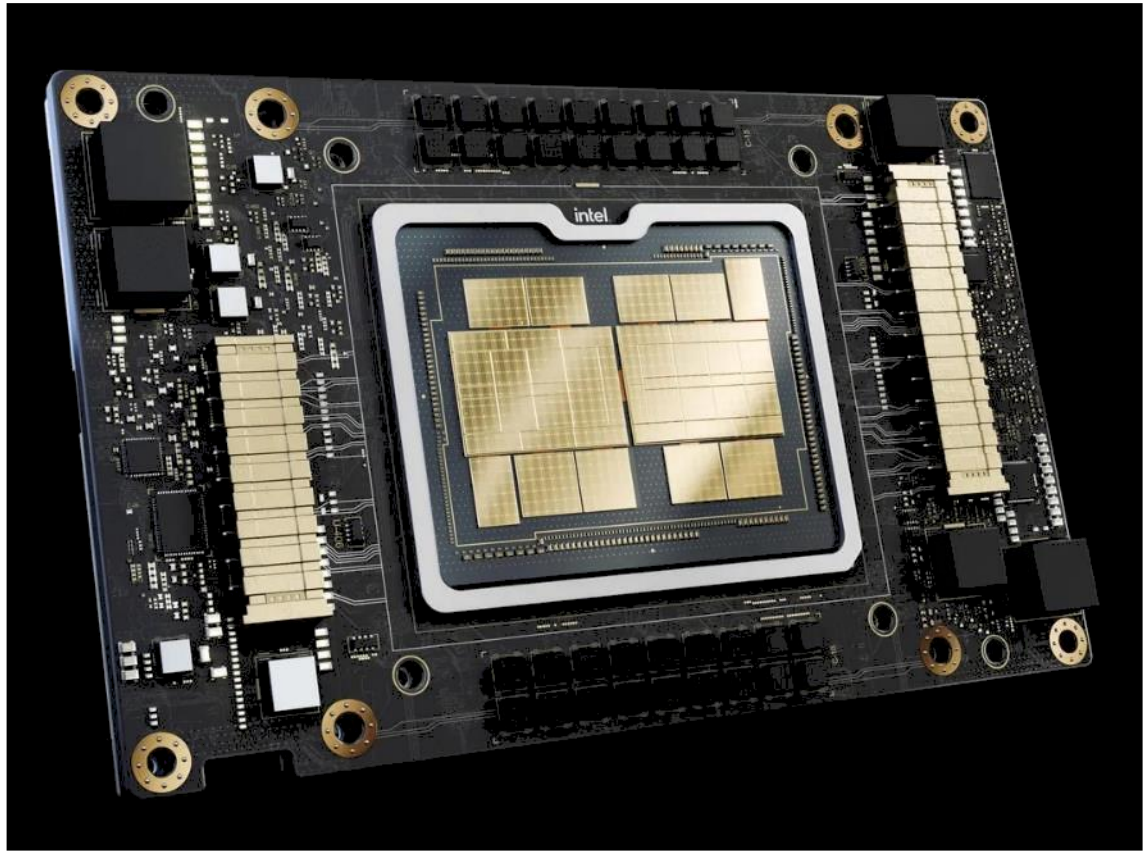
Embedded PROM is moving to an emerging memory

Embedded SRAM (including caches) will become persistent

What Becomes Persistent?



UCIe: A Standardized Chiplet Interface



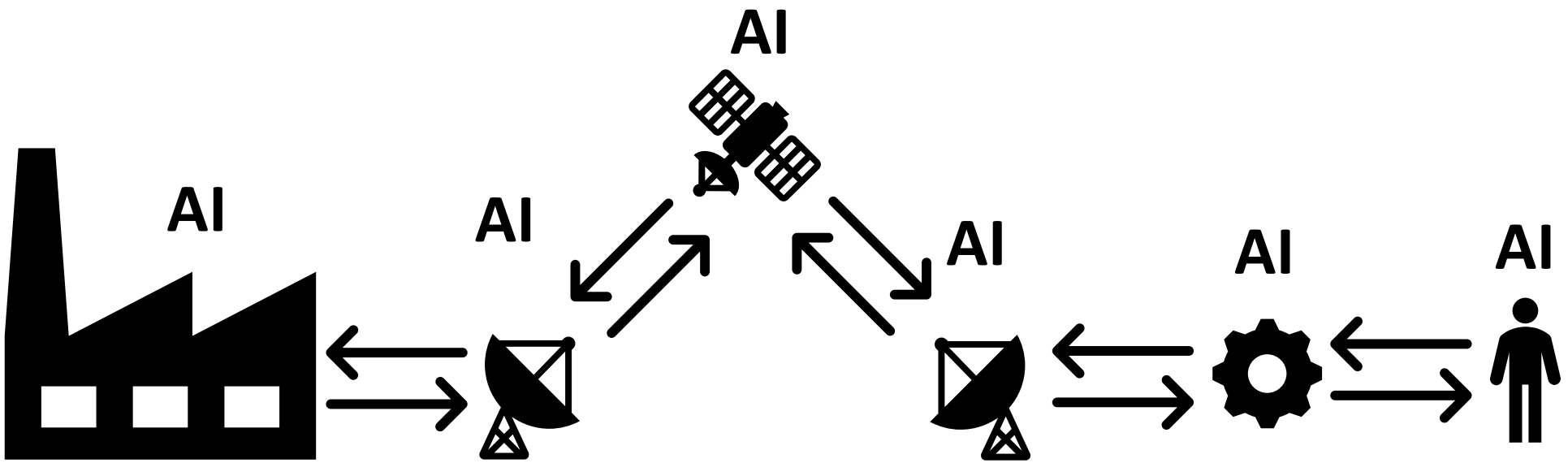
UCIe and Memories

- Mixed processes optimize cost/performance
  - Logic in a CMOS logic process
    - In logic SRAM & NOR flash are the only options for on-die memory
  - Memory chiplet in a memory process
    - DRAM, MRAM, ReRAM, FRAM, PCM…
  - Significant die area & cost reductions
- Commoditizes chiplets
  - One memory chiplet can be used by multiple logic companies
    - Increases volume & lowers costs
  - All vendors' parts equivalent
    - Vendors compete on price

Outline

- Change 1: Emerging Memories Arrive
- Change 2: Memory Disaggregation
- Change 3: Processing in Memory
- Change 4: Chiplets
- Change 5: AI Everywhere

Where does AI fit in Tomorrow's World?



AI eases bandwidth requirements

What AI Brings to the Party

- Faster response times
- Reduced bandwidth requirements
- Higher data integrity
- Improved security
- Better user experience

Summary: You Have a New Bag of Tricks

- Emerging memories for persistence & power savings
- CXL for pooling, persistence, and more
- Compute in Memory for scale & performance
- Chiplets for amazing new processor types
- AI to accelerate response while reducing bandwidth and centralized requirements

Hold Onto Your Hats!

Please take a moment to rate this session.

Your feedback is important to us.