SNIA DEVELOPER CONFERENCE

SDC 24

BY Developers FOR Developers

September 16-18, 2024
Santa Clara, CA

Big Architectural Changes Are Coming

Jim Handy, Objective Analysis
Tom Coughlin, Coughlin Associates

Outline

- Hardware Changes
- Software Changes
- Otherware Changes
- How CXL Could Go
- Persistence and Emerging Memory Types
- Processor Specialization
- Processing In Memory
- AI Everywhere
- Wrap-Up

How Hardware is Changing

- From CPUs to CPUs and GPUs/TPUs
  - CPU memory channels increasing
    - But DIMMs per channel is decreasing
  - GPUs aggressively moving from GDDR to HBM
- From local memory to memory fabrics
  - Communication bandwidth is critical bottleneck
- From centralized to edge processing
  - Reduce bandwidth requirements by reducing data size
  - Delegate tasks to edge or endpoints

Outline

- Hardware Changes
- Software Changes
- Otherware Changes
- How CXL Could Go
- Persistence and Emerging Memory Types
- Processor Specialization
- Processing In Memory
- AI Everywhere
- Wrap-Up

Software Changes

- Virtualization to composability to persistence to disaggregated memory to AI to fabrics to…
- Disaggregated memory applications & systems are coming
  - Also for memory fabrics
- Support for persistence
  - SNIA NVM Programming Model is a strong foundation
  - Persistent caches are coming
- …also AI-generated code
- …also just the fact that different languages are used for AI
  - And different talents are needed to manage it

SDC 24

AI Talent Pool

Outline

- Hardware Changes
- Software Changes
- Otherware Changes
- How CXL Could Go
- Persistence and Emerging Memory Types
- Processor Specialization
- Processing In Memory
- AI Everywhere
- Wrap-Up

Changes Coming in the Foreseeable Future

- AI-oriented networks
- Optical networks
- Widespread use of chiplets
- Widespread use of fabrics
- Machine learning at the edge
- Inference at the endpoints
- Pervasive persistence
- Coprocessors everywhere

Outline

- Hardware Changes
- Software Changes
- Otherware Changes
- How CXL Could Go
- Persistence and Emerging Memory Types
- Processor Specialization
- Processing In Memory
- AI Everywhere
- Wrap-Up

What is CXL Really For?

- Maintaining coherency?
- Eliminating stranded memory?
- Expanding memory size?
- Increasing memory bandwidth?
- Supporting persistent memory?
- Hiding DDR4/DDR5/DDR6 differences?
- Passing messages between xPUs?

CXL Supports New Memory Architectures

- Disaggregated memory
- Pooled memory
- Memory fabrics
- Shared memory
- Persistent memory

User Wants & Needs

- Microsoft Azure: Pooling can save 7% in memory costs
  - Eliminates stranded memory
- Google: Stranded memory is not important
  - VMs are efficiently packed in high-resource servers
- IBM/Georgia Tech: DDR is a poor answer
  - All DRAM should be attached by CXL or OMI
- AI Providers: We need enormous memories
  - Also fast loads of GPU HBM
    - Give us bandwidth!
- Hyperscalers: “Any-to-Any” xPU connections
- PC OEMs: CXL is not immediately useful

Optimistic, Pessimistic, & Realistic Forecasts

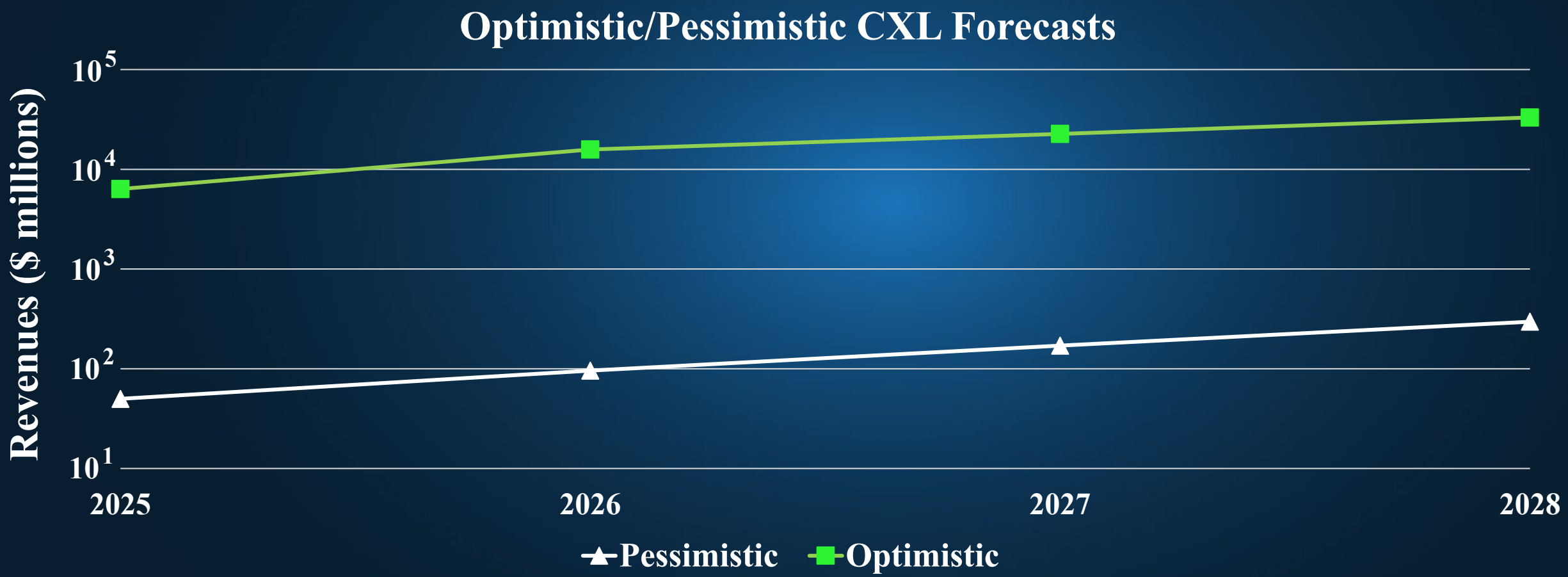
Very Optimistic Forecast

- 2 Years to 100% data center adoption
  - All DDR replaced by CXL in 5 years
- Widespread use of pooling
- Instant doubling of memory sizes
  - AI given as the reason
- CXL to re-use older DIMMs
  - MS-SSD (CXL NAND) catches on
- Switches everywhere!

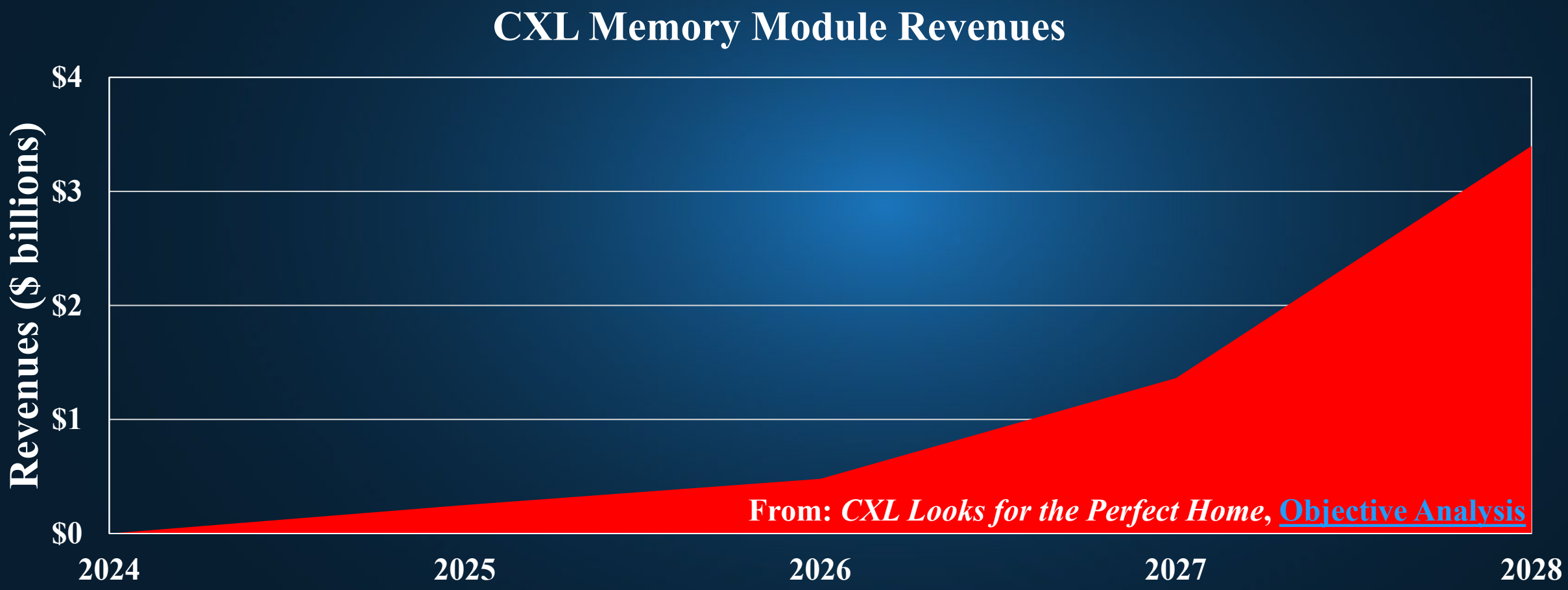
Very Pessimistic Forecast

- Extremely slow acceptance
  - No acceptance without strong software support
  - Two Olympic Cycles to create this software
  - Only popular for large-memory systems
- Large-memory servers rarely required
  - A problem that Optane faced
- Pooling not adopted
  - Switches don’t find homes

Optimistic and Pessimistic Numbers



Realistic CXL Forecast

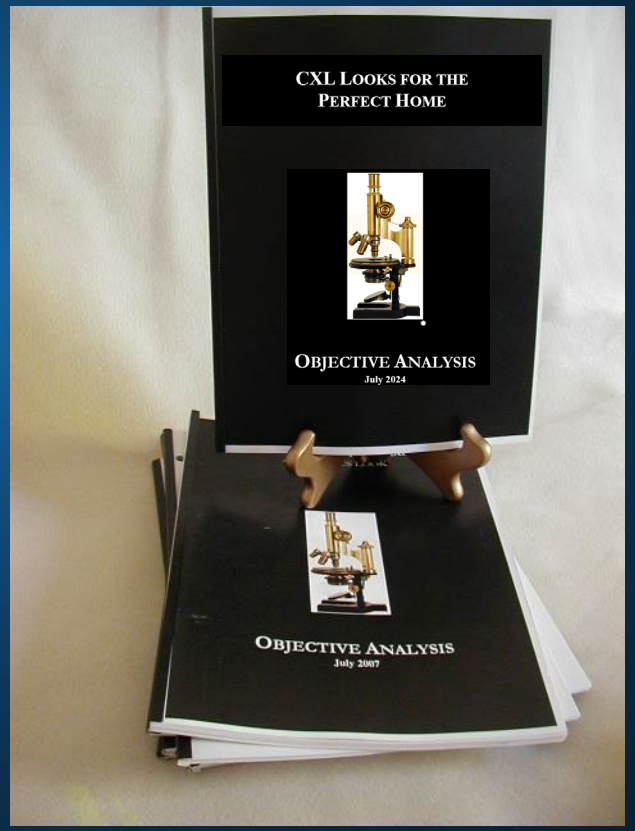


Long-Term Impact

- Re-thinking system architecture
  - Disaggregated memory
  - Processor arrays with memory fabrics
  - Memory agnostic
- Better memory bandwidth & size vs. worse latency
  - Design-arounds will optimize for this

New Report: CXL Looks for the Perfect Home

- Released July 2024
- Covers all perspectives
  - Where CXL is useful, and where it isn't
  - Demand drivers for CXL DRAM modules
  - Opportunities outside of DRAM
  - Forecast (Revenues, units, ASP)
- Available for immediate download:
  - Objective-Analysis.com/reports



Outline

- Hardware Changes
- Software Changes
- Otherware Changes
- How CXL Could Go
- Persistence and Emerging Memory Types
- Processor Specialization
- Processing In Memory
- AI Everywhere
- Wrap-Up

RULE 1:

All Emerging Memories are Persistent

RULE 2: None Use a Charge-Based Cell

- MRAM: Magnetism
- ReRAM: Resistance
  - Either metal filament or oxygen vacancy
- PCM: Resistance, too
  - Crystalline or amorphous
- FRAM: Atom displacement

Emerging Memory Benefits

- Nonvolatile 
- Fast write compared to flash 
- Byte writeable 
- Scalable well past 28nm 
- Radiation-tolerant 
- Based on innovative materials 

The Economics Are Challenging

- A small die size isn't enough
  - Manufacturing scale determines relative cost
    - Economies of scale prevail
- Intel's Optane proved the difficulty
  - Volume never justified the cost
  - >$7B in Intel losses
    - Micron losses ~$400M/quarter

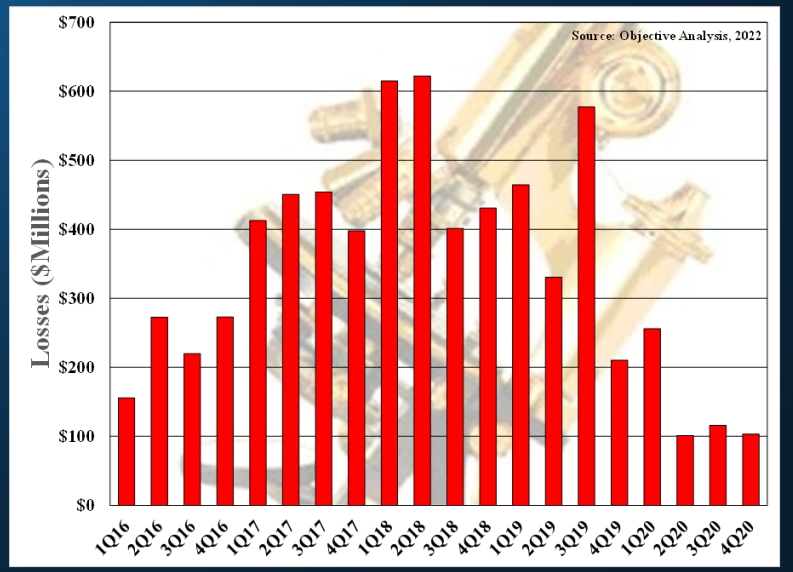
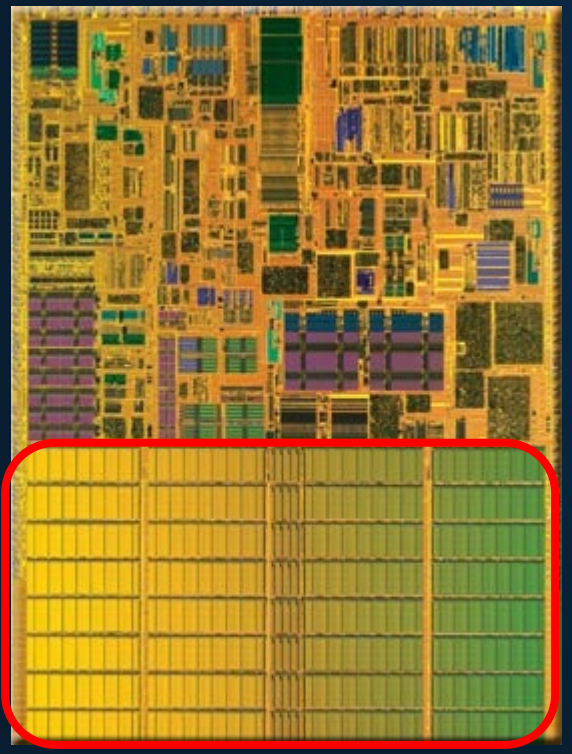


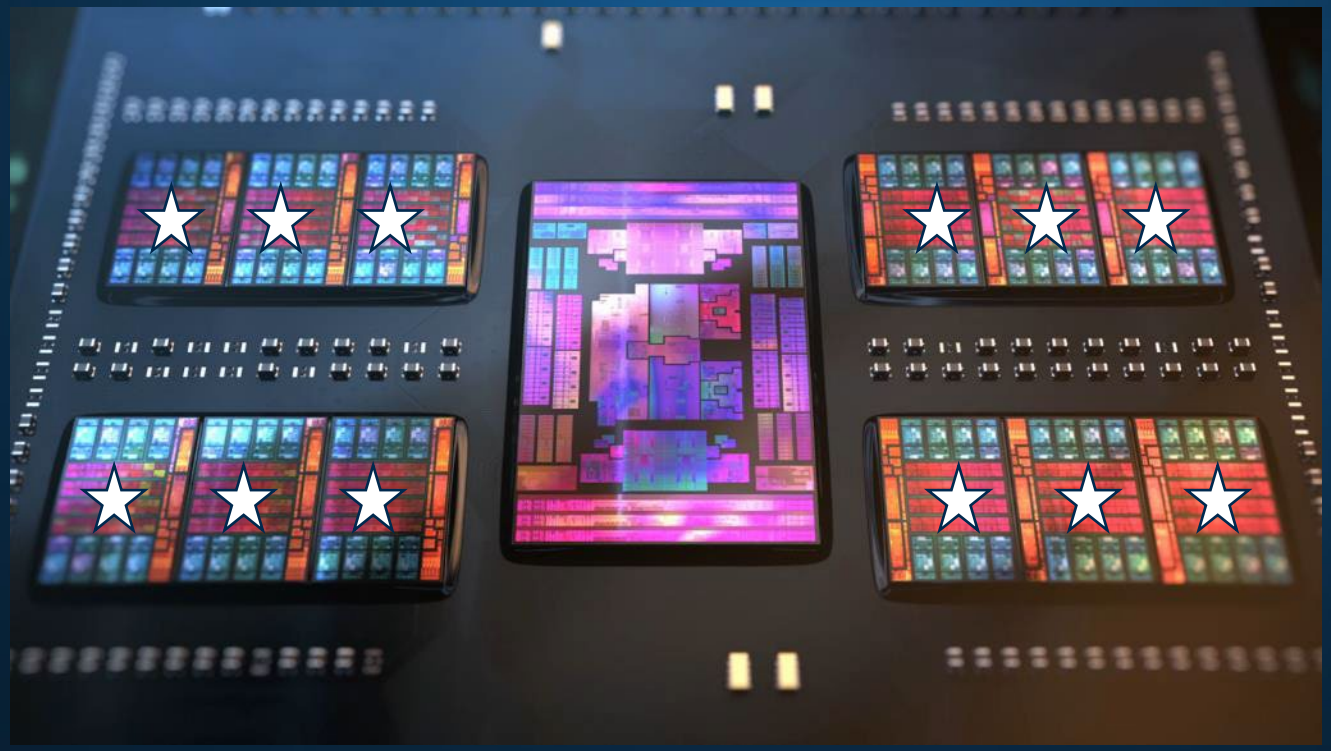
Chart Source: ***Emerging Memories Branch Out***

Where Will Persistence Reappear?

Caches



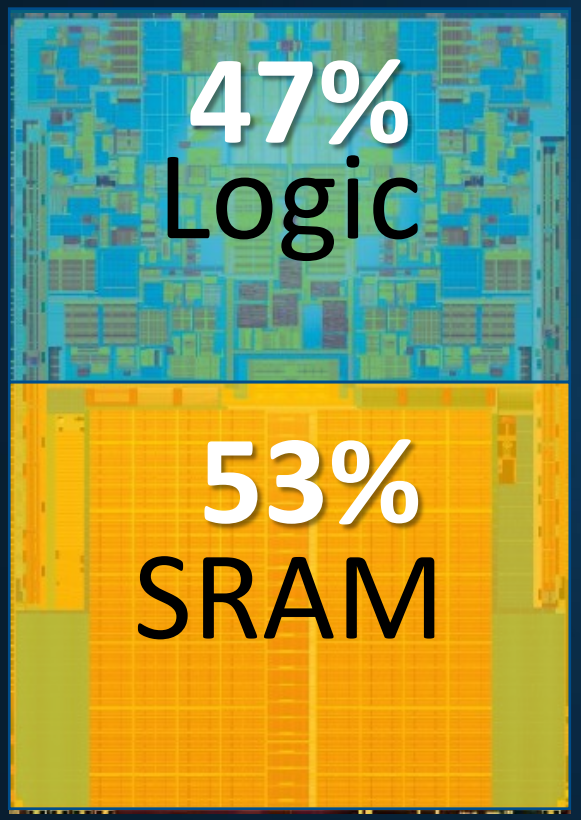
Chiplets

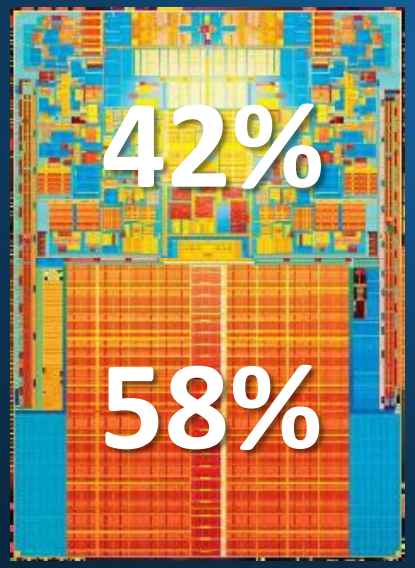


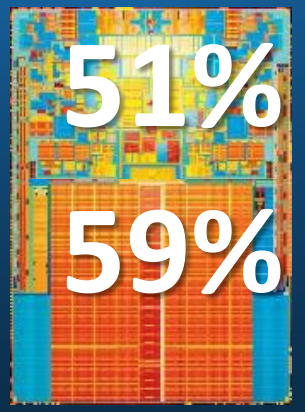
CXL

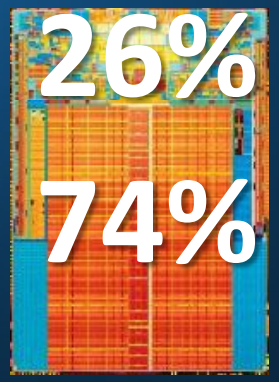
SDC 24

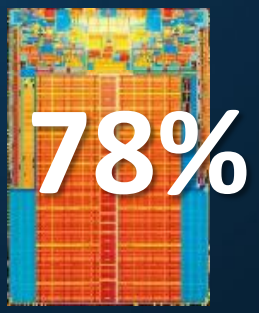
SRAM Caches Barely Shrink

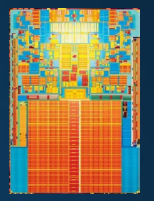








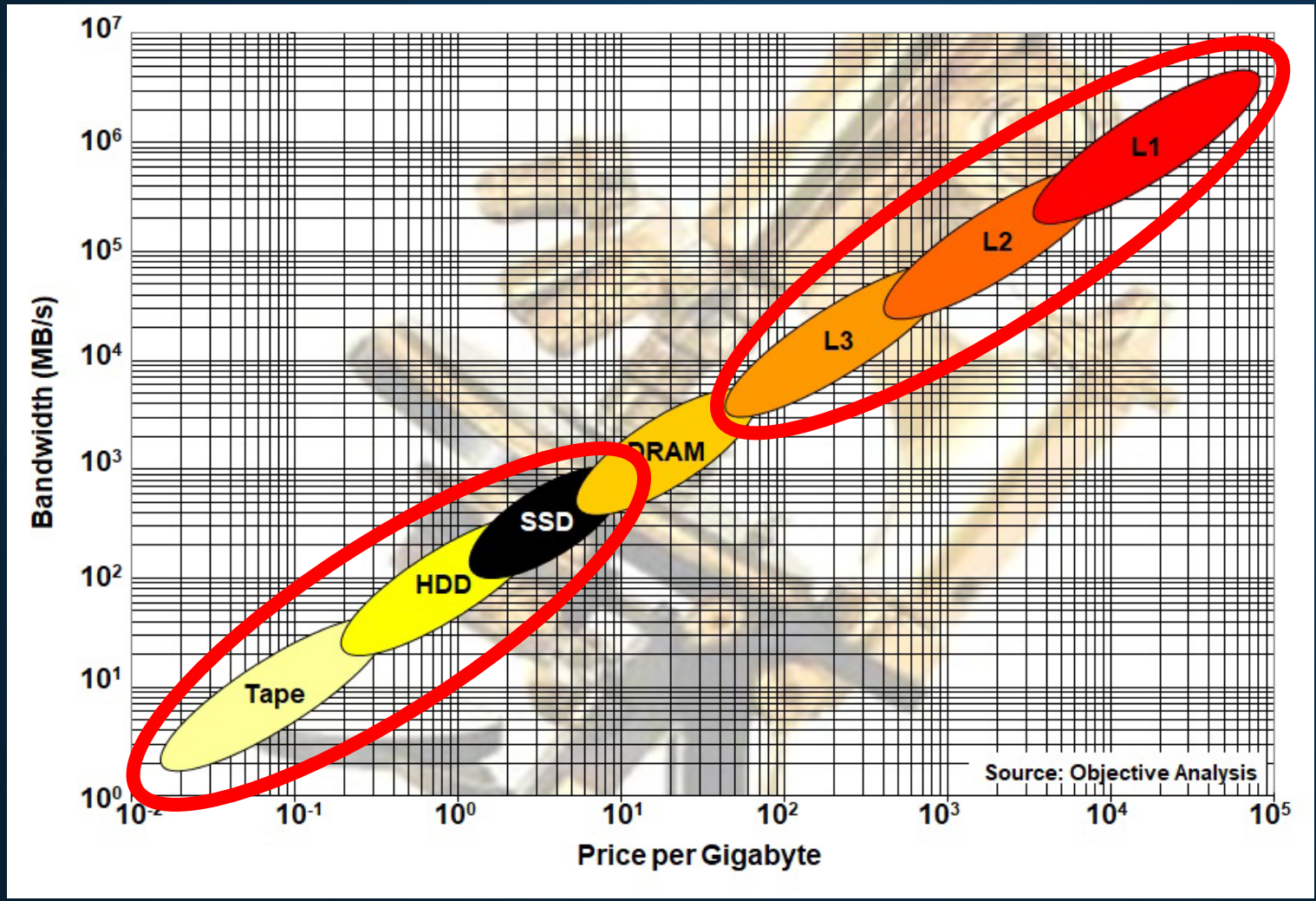




A Persistent Cache? Why Not?

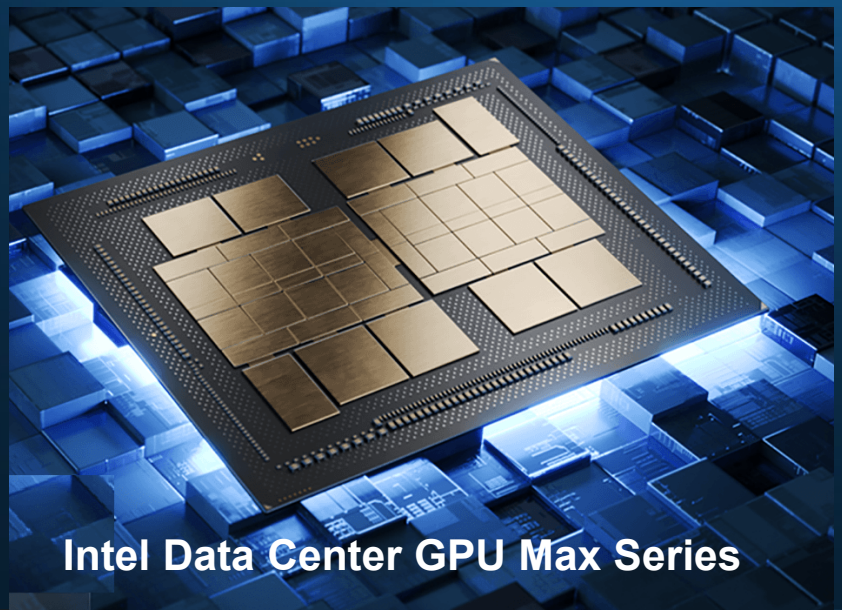
- SRAM is not shrinking with the semi process
  - Cache's share of CPU chip cost is ballooning
- Emerging (and persistent) memories scale with process
- Foundries have already developed MRAM & ReRAM processes
  - In volume production today
- There are downsides:
  - SRAM is faster than emerging memories, but far more costly
  - Software support isn't fully there, but SNIA's NVM Programming Model is helpful
  - Off-the-shelf software doesn't know what to do with persistence

What Becomes Persistent?



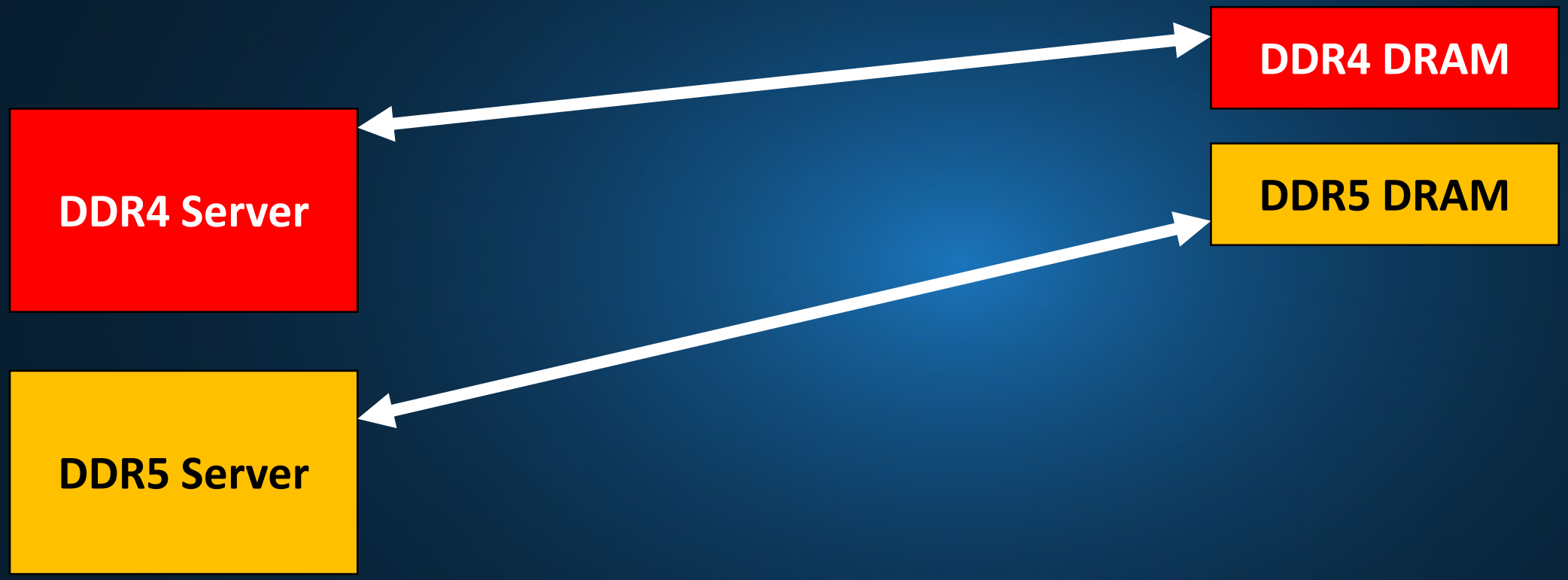
Persistent Chiplets

- Chiplets are gaining momentum
  - FPGAs have used them for over 5 years
  - Packaging techniques are well established
- Logic process for logic, memory process for memory
  - More cost-effective and faster time-to-market

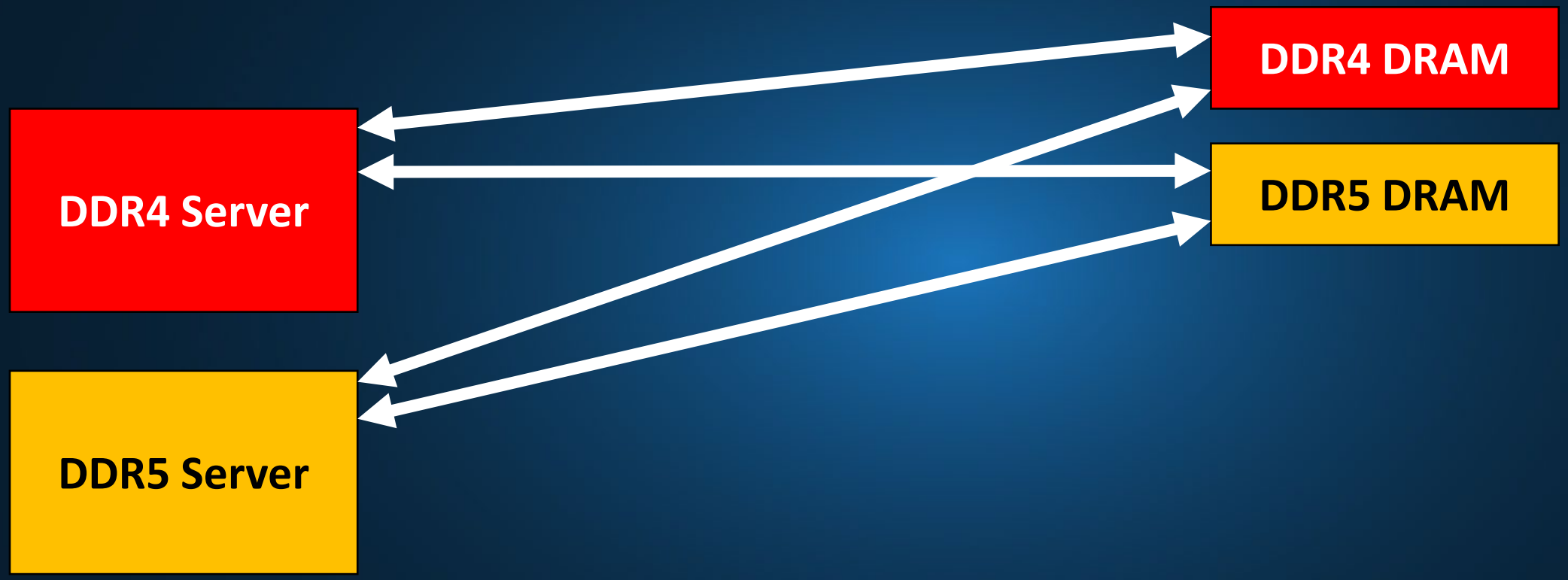


Intel Data Center GPU Max Series

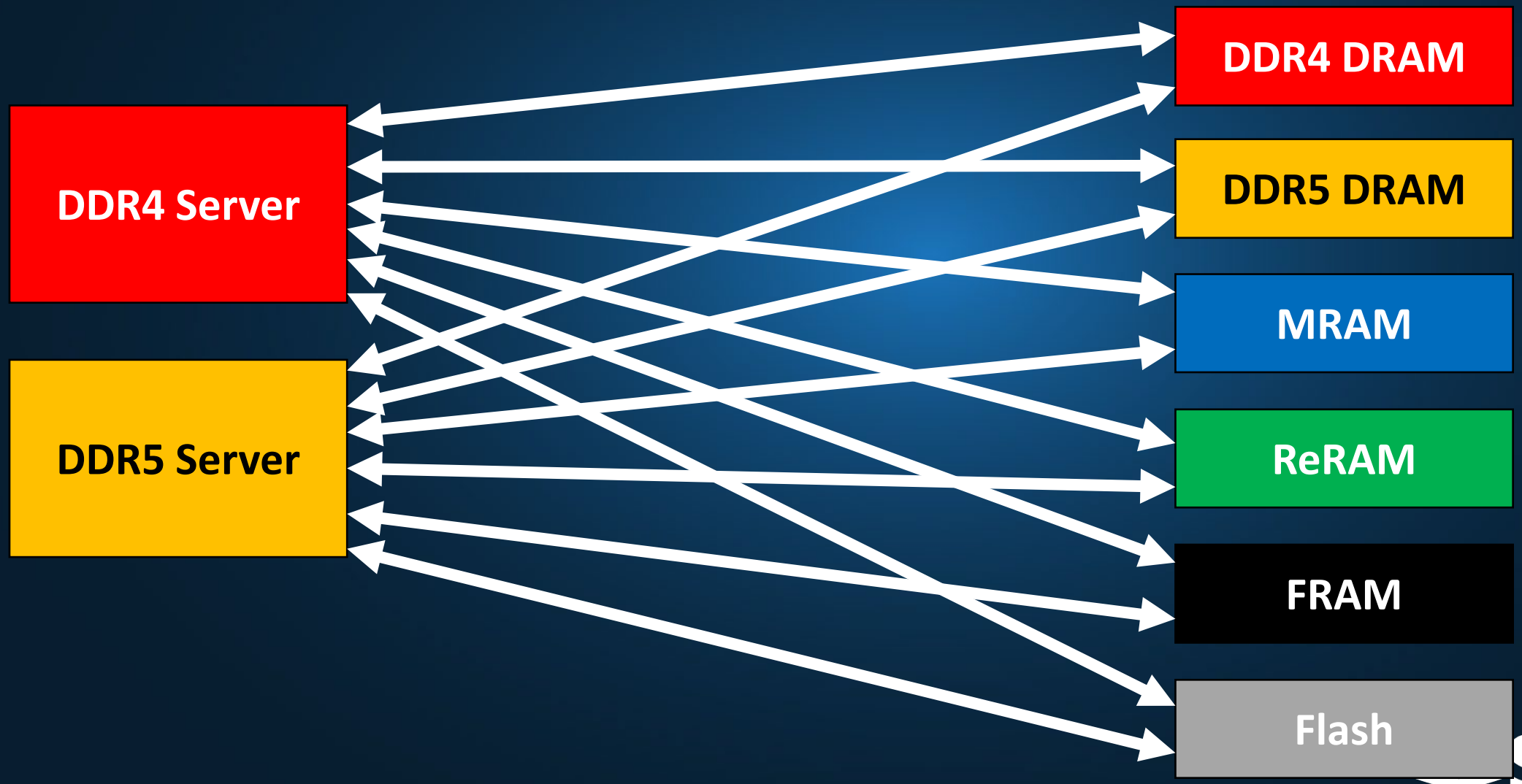
CXL Supports Any Memory, Volatile or Persistent



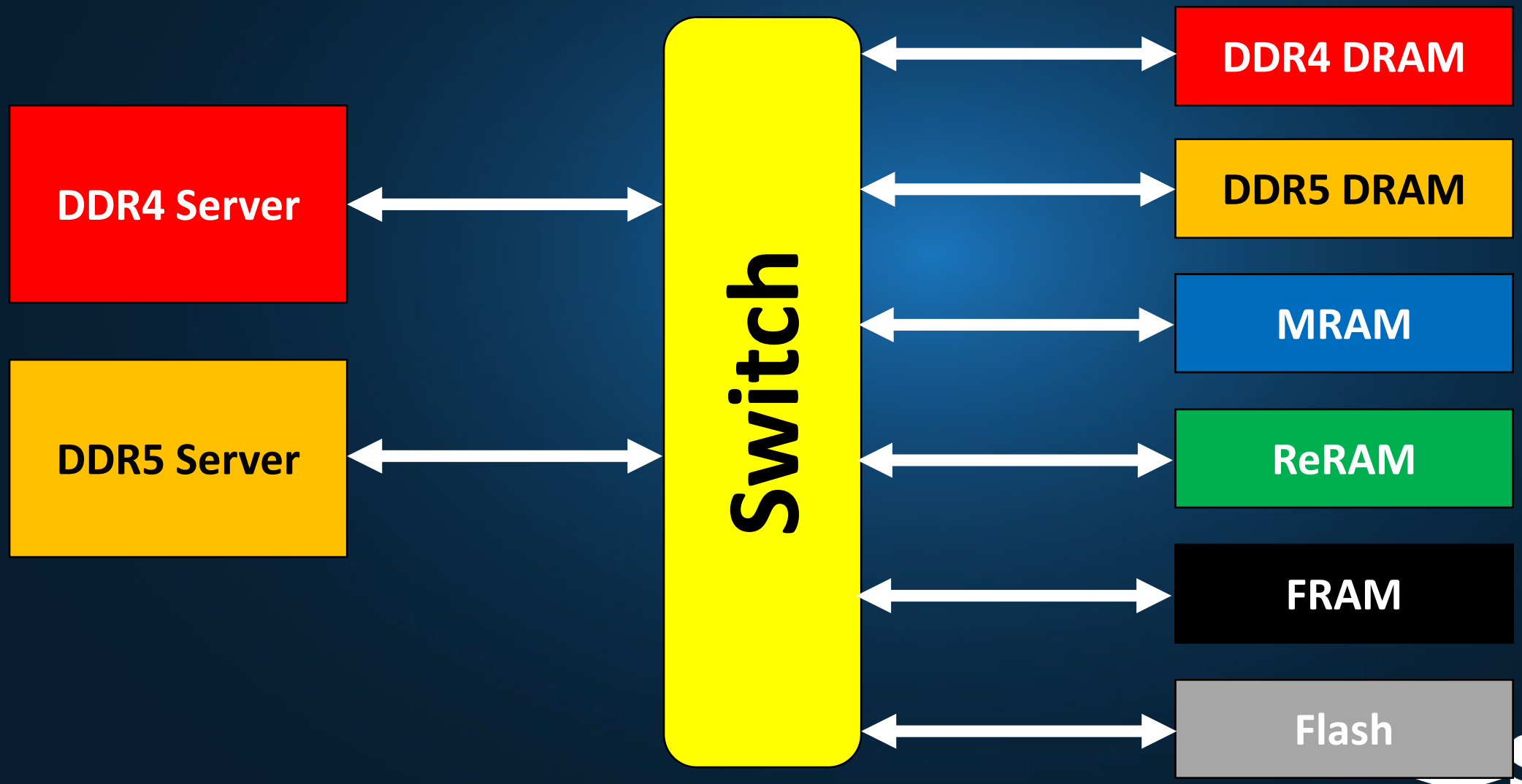
CXL Supports Any Memory, Volatile or Persistent



CXL Supports Any Memory, Volatile or Persistent



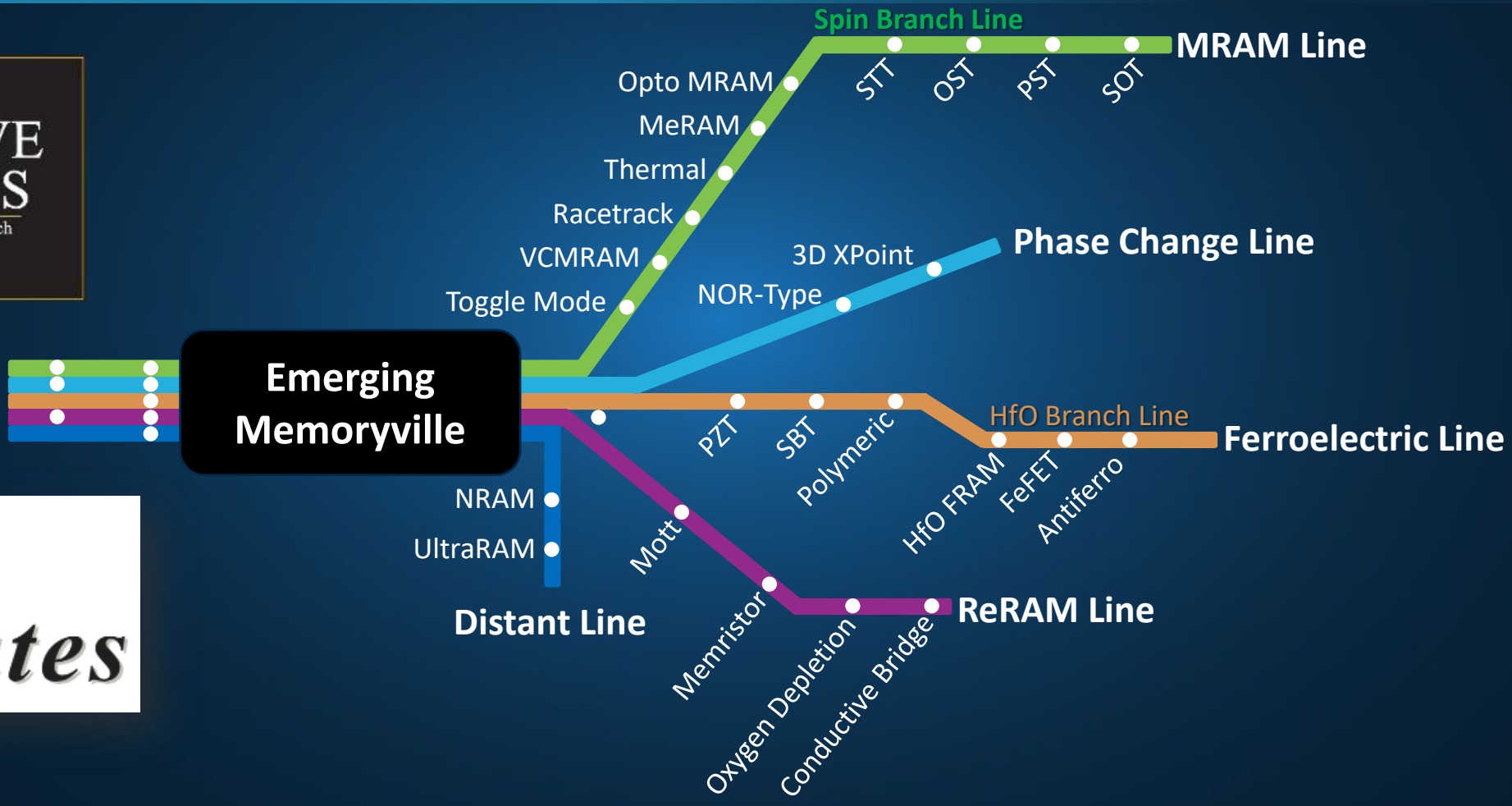
CXL Supports Any Memory, Volatile or Persistent



Report: Emerging Memories Branch Out

OBJECTIVE ANALYSIS
Semiconductor Market Research

Coughlin Associates



Now Available!

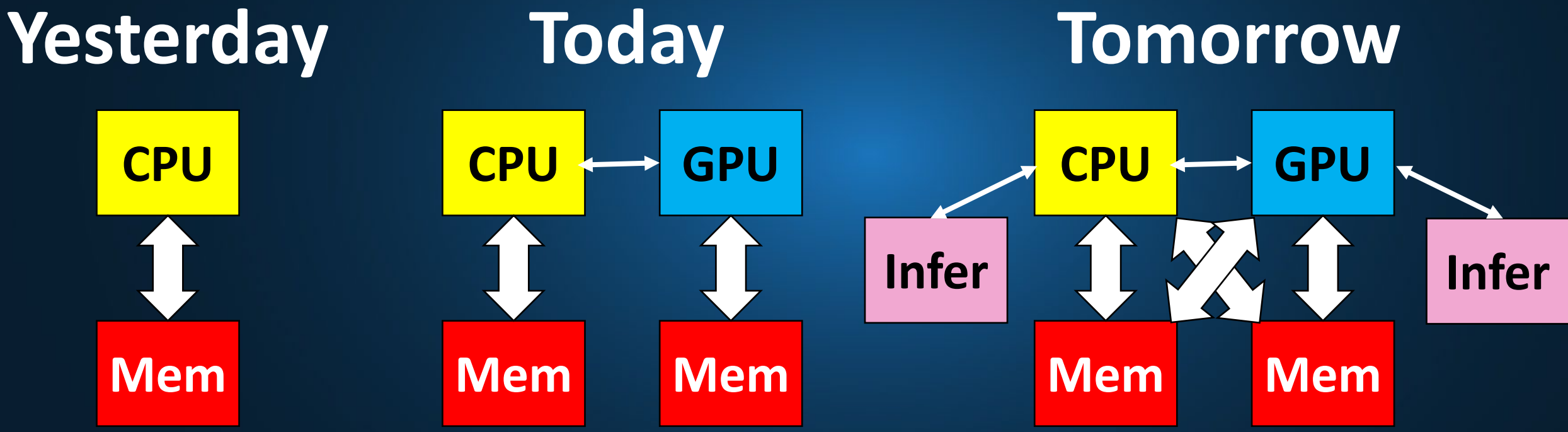
https://Objective-Analysis.com/reports/#Emerging
http://www.TomCoughlin.com/techpapers.htm

SDC 24

Outline

- Hardware Changes
- Software Changes
- Otherware Changes
- How CXL Could Go
- Persistence and Emerging Memory Types
- Processor Specialization
- Processing In Memory
- AI Everywhere
- Wrap-Up

Whither the Processor?



Outline

- Hardware Changes
- Software Changes
- Otherware Changes
- How CXL Could Go
- Persistence and Emerging Memory Types
- Processor Specialization
- Processing In Memory
- AI Everywhere
- Wrap-Up

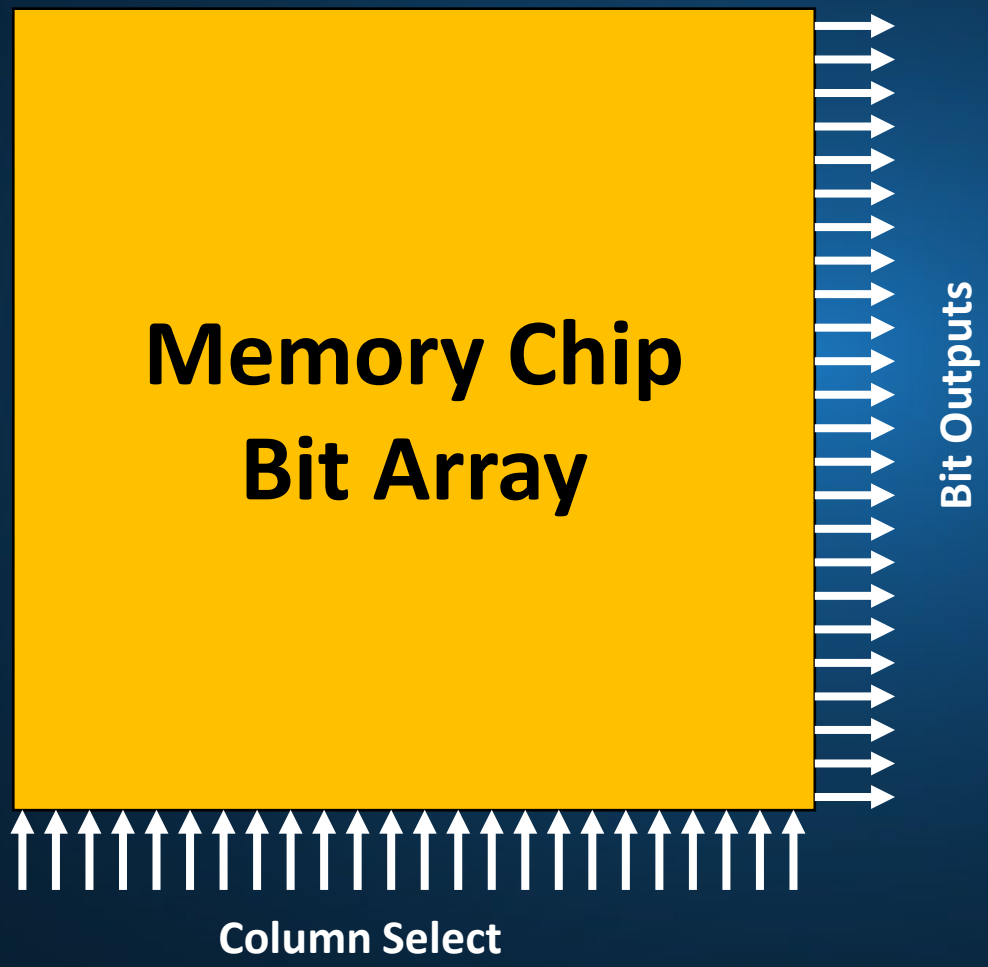
What is “Processing In Memory”?

It depends on who you talk to

- DIMMs with DRAM and a processor chip
  - Hints of HBMs with processor on logic chip
- DRAM chips with an internal processor
- Processing logic within the memory bit cells
- Analog neural net chips

Goal is to reduce data movement

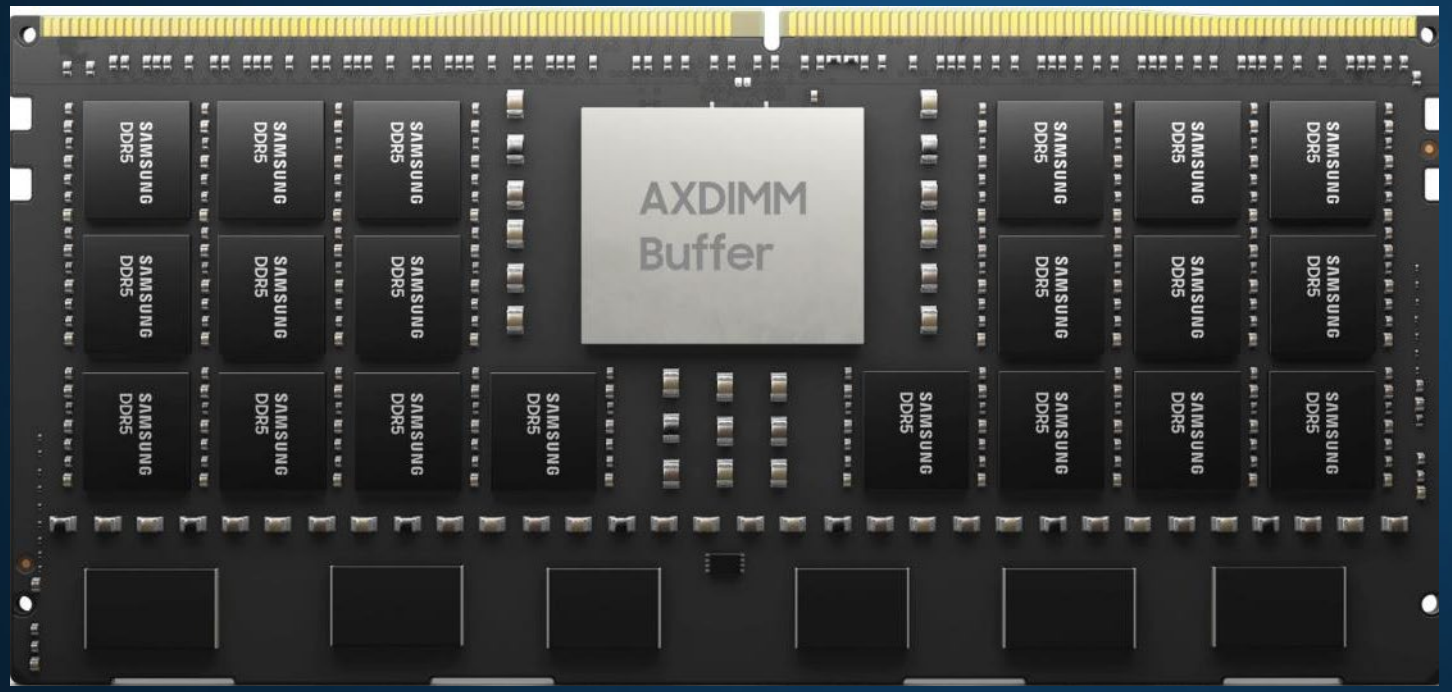
Goal: Wider Buses, Greater Processing Bandwidth



Tens of thousands of bits ALL AT ONCE!

DIMM with an Internal Processor

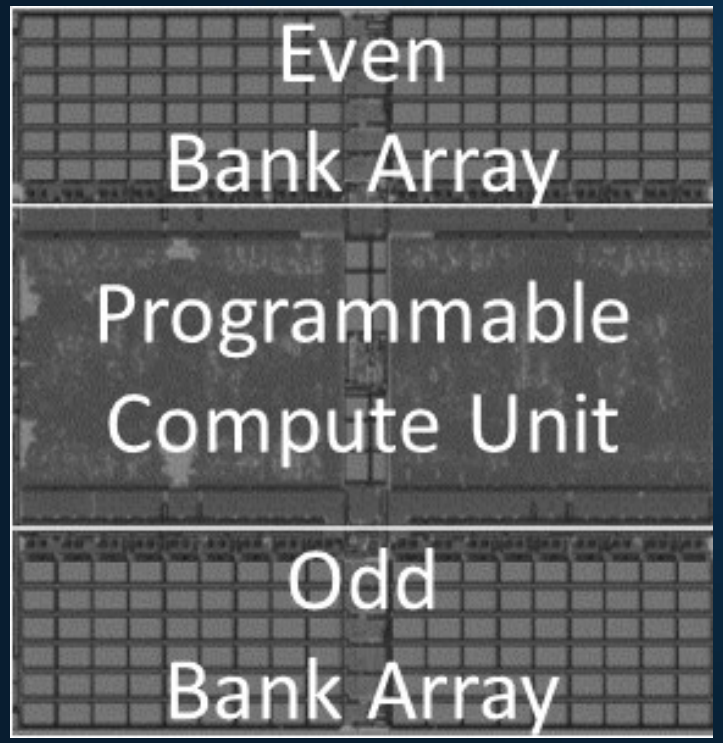
Samsung AXDIMM



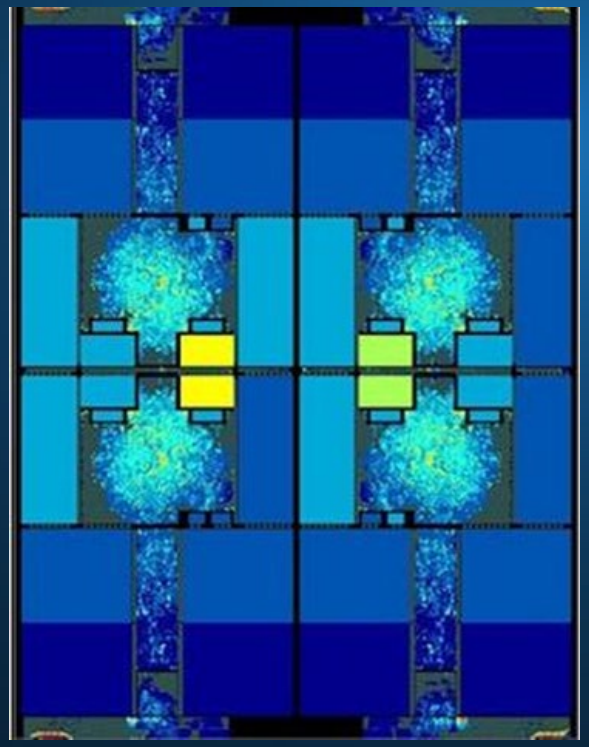
CXL can replicate this approach

DRAM Chips with Internal Processor

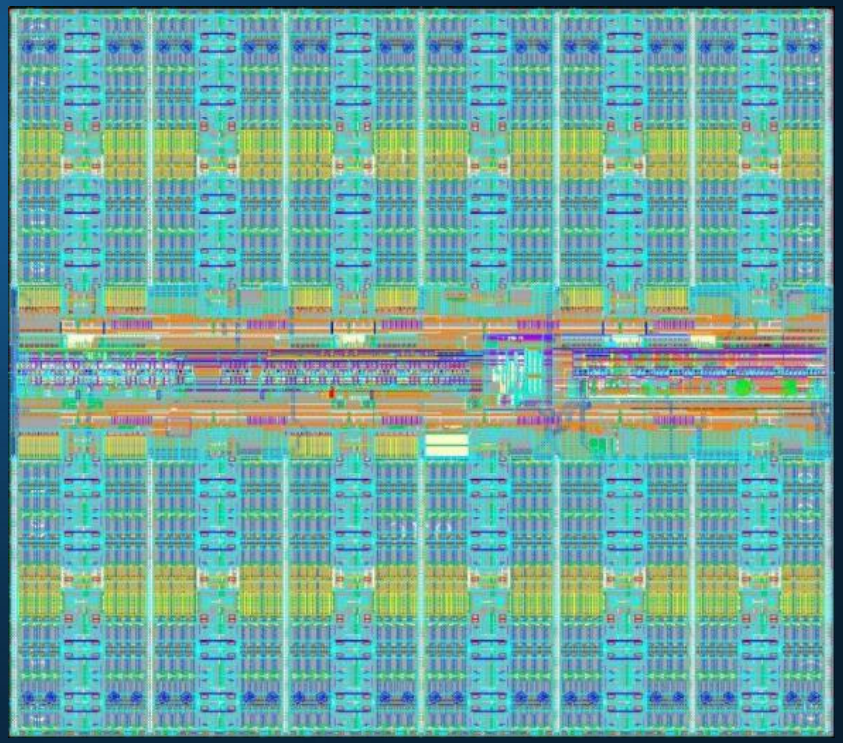
Samsung Aquabolt AX



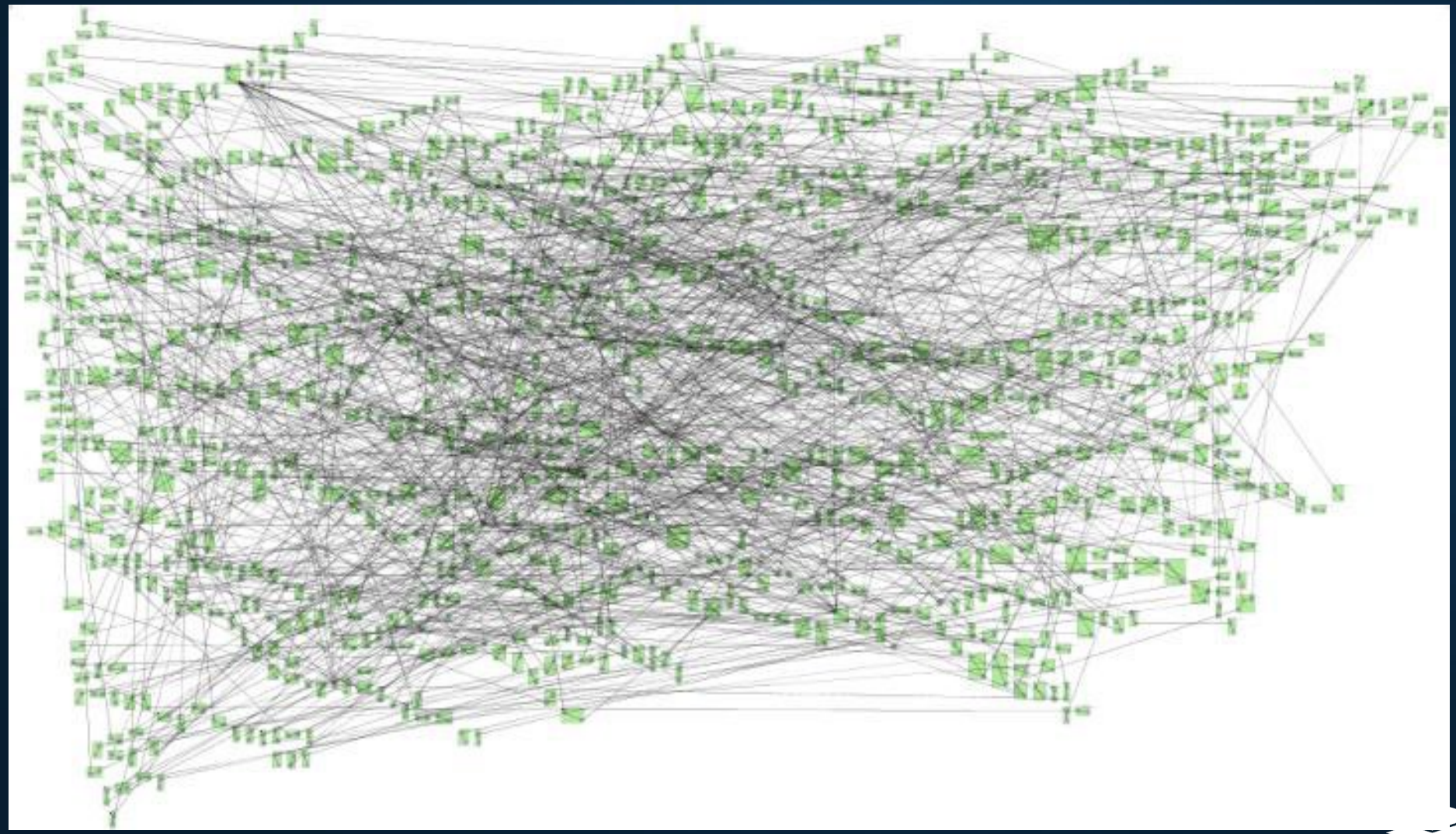
Upmem DPU



Natural Intelligence Automaton

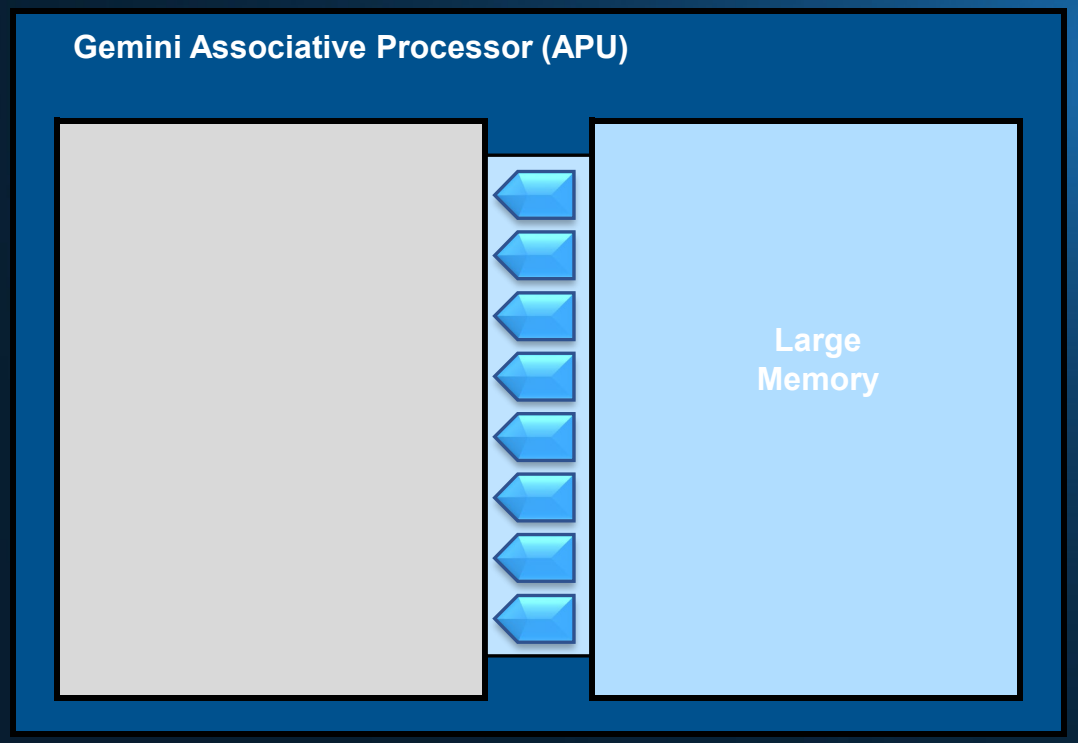


DRAM Chips with Internal Processor

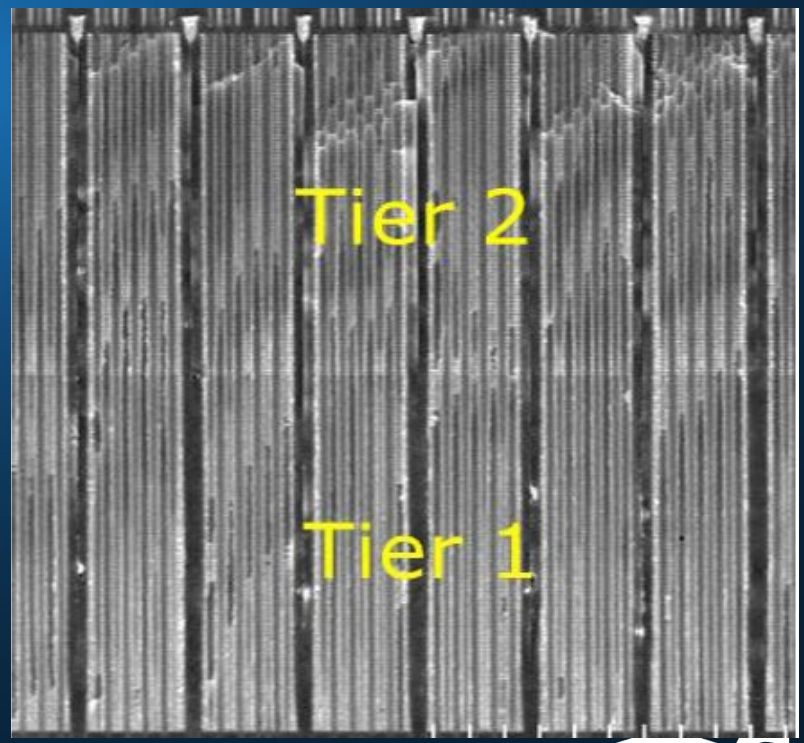


Processing Within the Memory Bit Cell

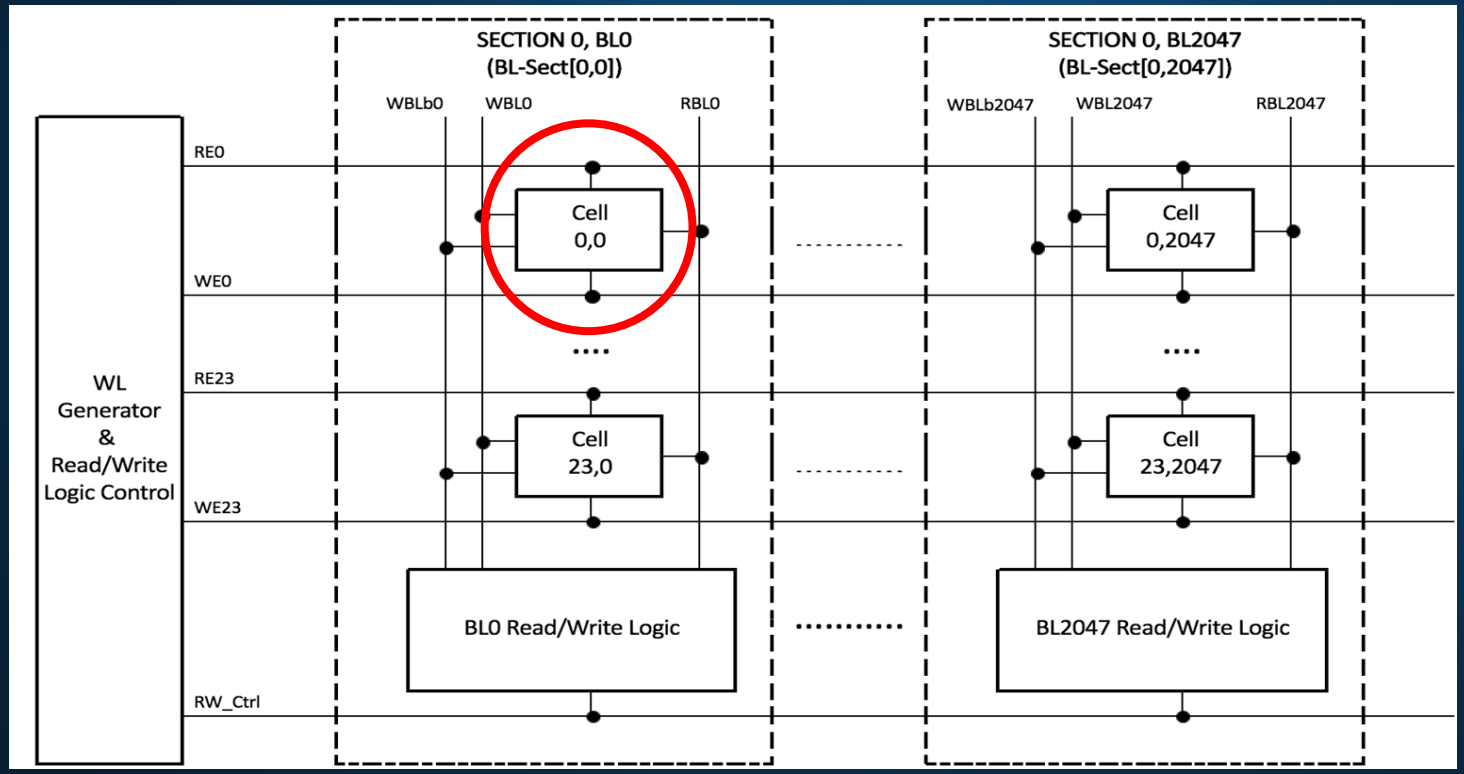
GSI Gemini APU



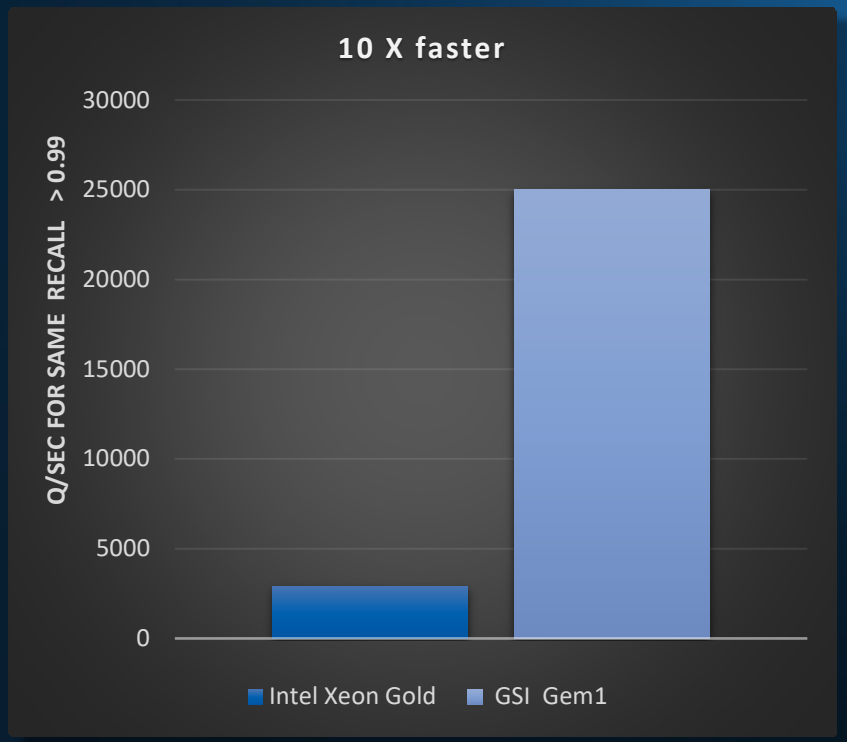
Macronix FortiX

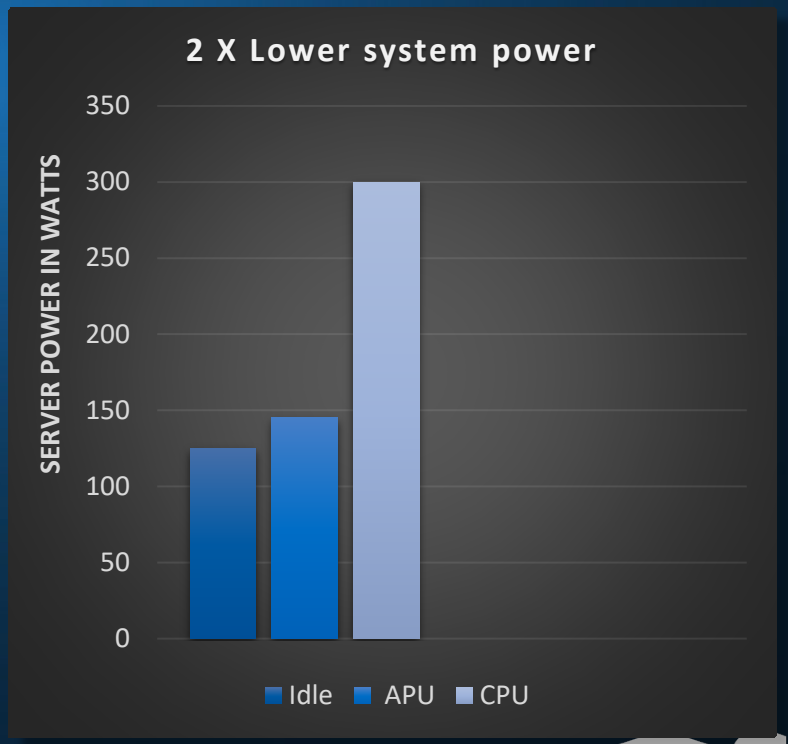


Processing Within the Memory Bit Cell



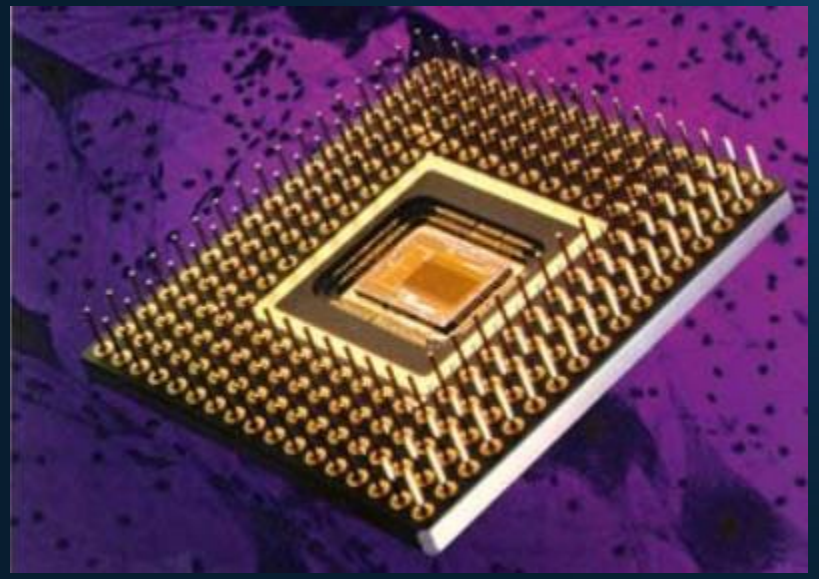
Processing Within the Memory Bit Cell

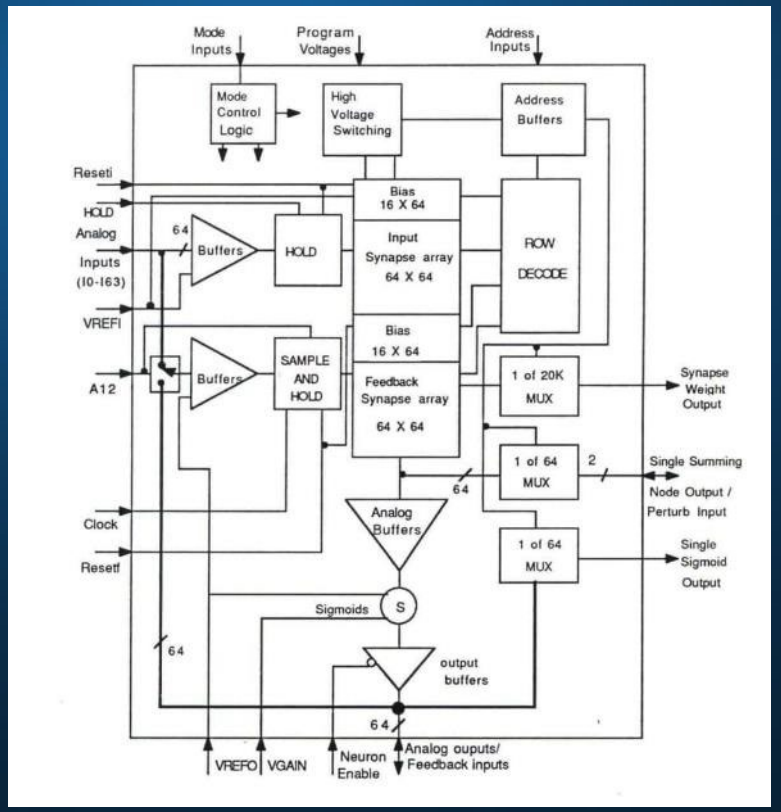




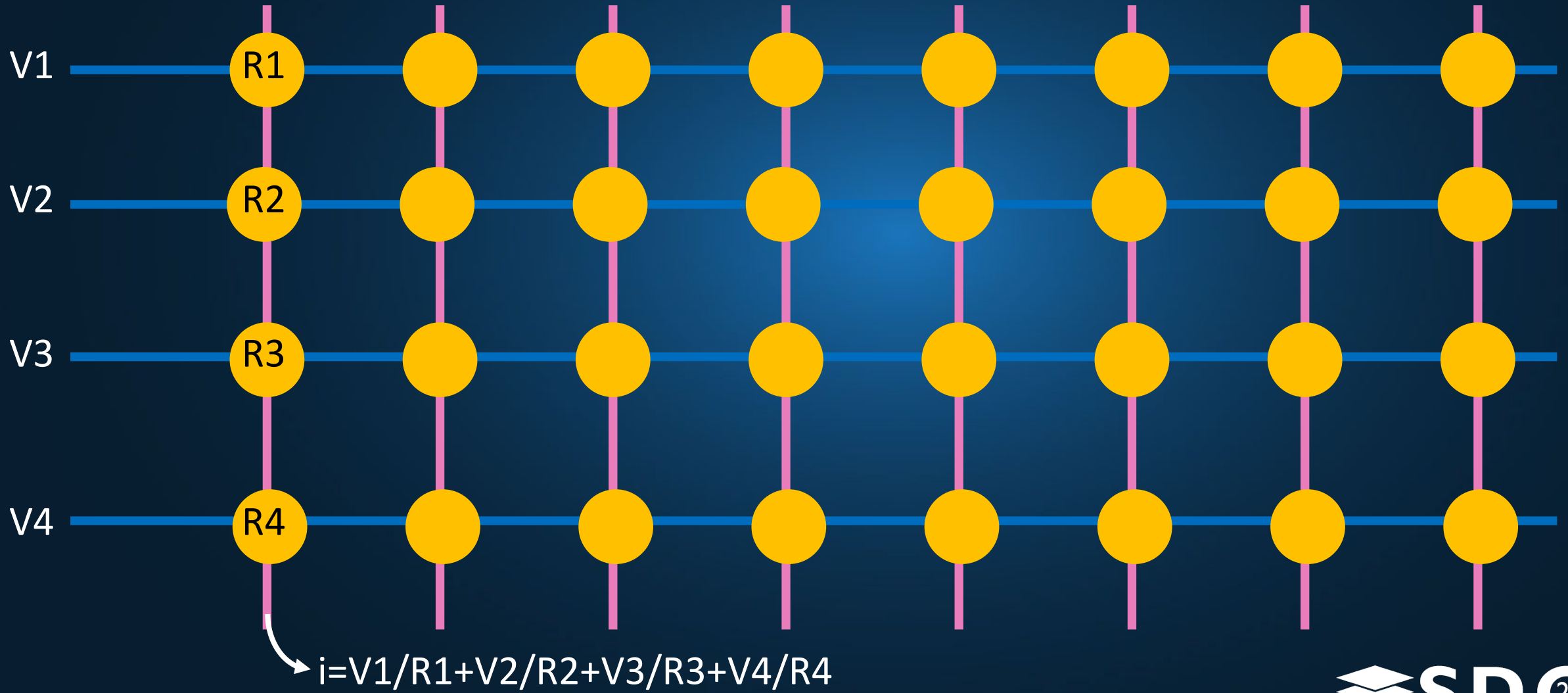
Neural Networks: Anything But New!

- Intel's 80170NX ETANN
  - Electrically-Trainable Analog Neural Network
- Introduced in 1989
- Not a commercial success

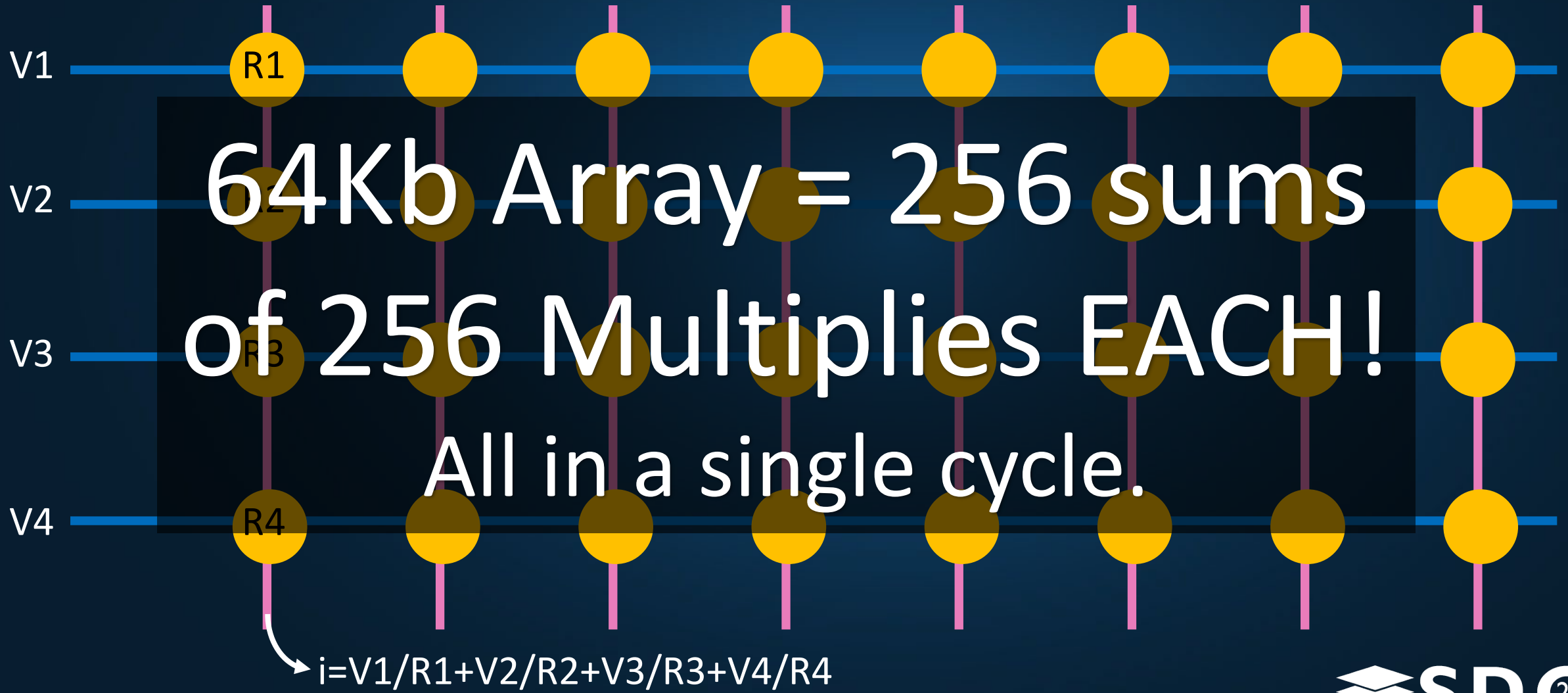




Neural Networks Fit Emerging Memories



Neural Networks Fit Emerging Memories



64Kb Array = 256 sums of 256 Multiplies EACH!

All in a single cycle.

PIM Challenges

- Lack of software support
  - Few tools
  - Few applications programs
- Lack of existing talent

It’s a game of catch-up

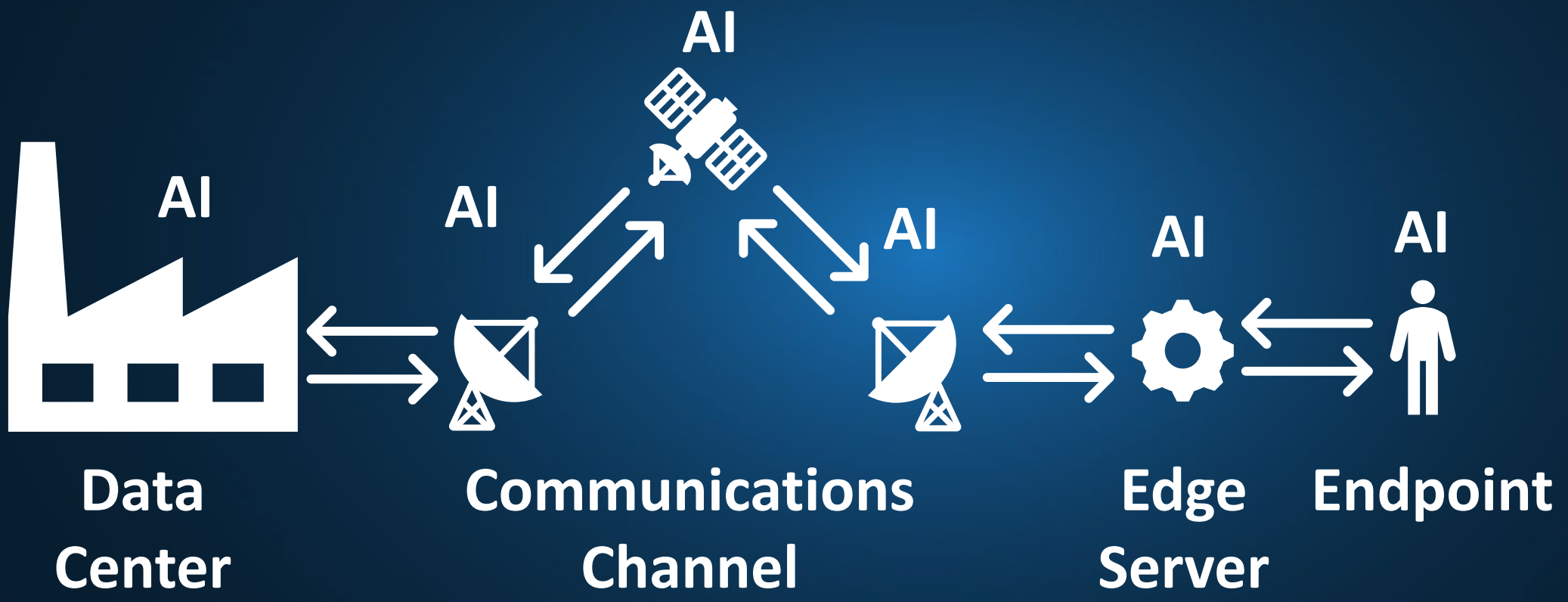
Outline

- Hardware Changes
- Software Changes
- Otherware Changes
- How CXL Could Go
- Persistence and Emerging Memory Types
- Processor Specialization
- Processing In Memory
- AI Everywhere
- Wrap-Up

AI Without Limits

- Today: GPUs in the data center
- Tomorrow: Neural nets at the edge
- Later: AI manages parts of the AI system, like networks?
  - AI already manages some SSD internals
- Some CMOS Image Sensors already include an AI chip
  - Used for image recognition
- AI-generated code in use today
- AI could configure datacenters
  - AI's great at evaluating numerous options

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

Protocol standards will be required

Outline

- Hardware Changes
- Software Changes
- Otherware Changes
- How CXL Could Go
- Persistence and Emerging Memory Types
- Processor Specialization
- Processing In Memory
- AI Everywhere
- Wrap-Up

Summary

- Hardware changes: CPUs, GPUs, fabrics, & edge processing
- Software changes: Disaggregated Memory & Fabric Support, Persistence, AI
- Otherware changes: Edge processors & ML, AI, network advances
- CXL: On its way, but how quickly?
- Persistence: Coming to a cache near you!
- Diverse processor types: More skill sets required
- PIM at the edge: Condenses communications
- Ubiquitous AI: Small doses to reduce bandwidth & delays

QUESTIONS?

