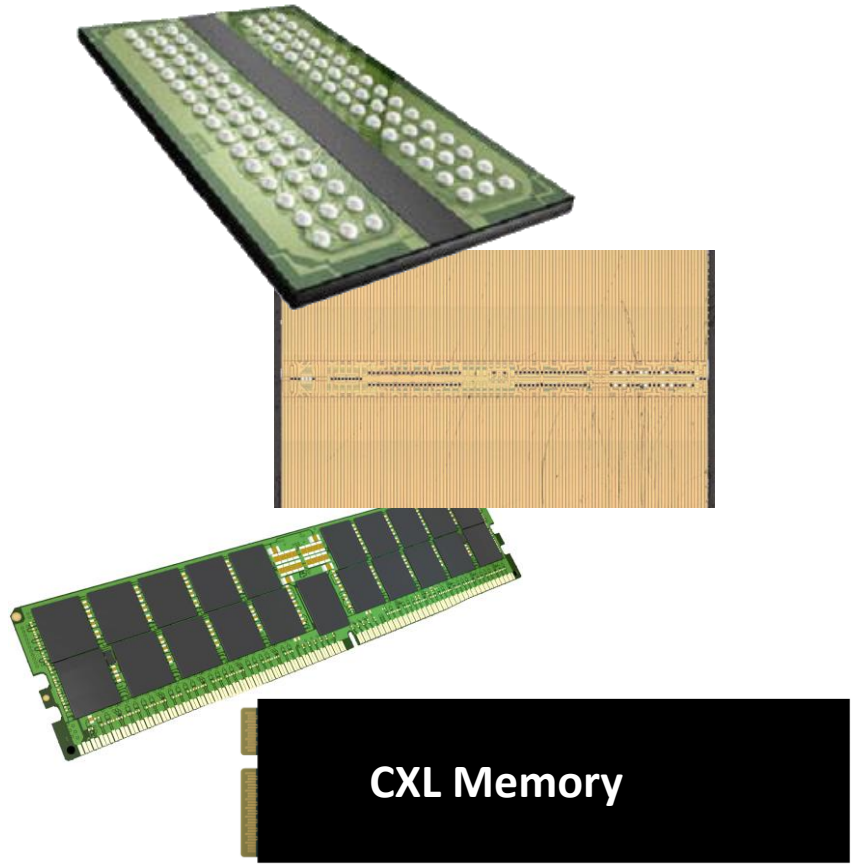
The Great Convergence; How CXL and UCIe Challenge the Memory Wall

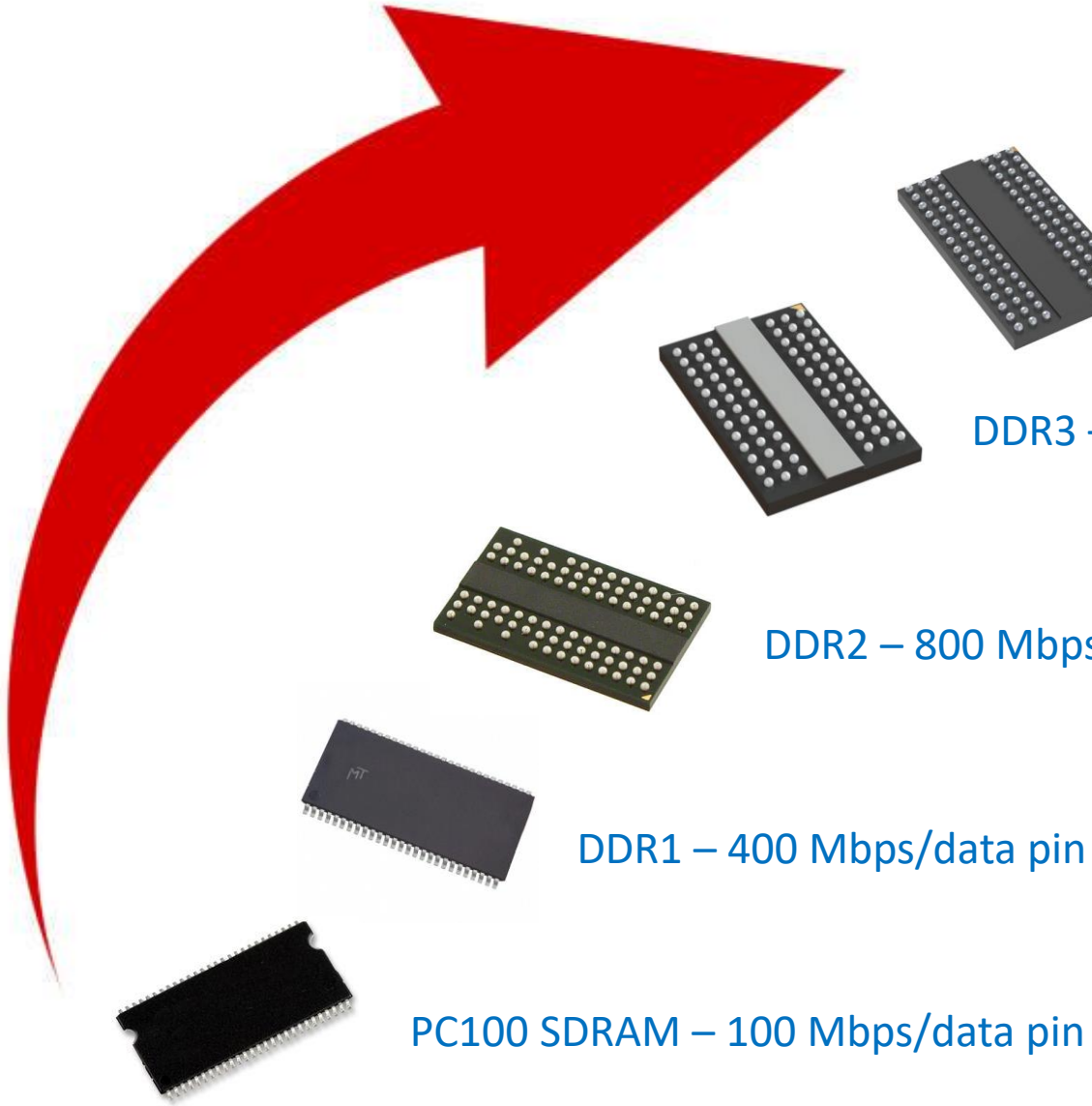


Bill Gervasi, Principal Systems Architect

Wolley Inc.

bilge@wolleytech.com

©2023 Flash Memory Summit. All Rights Reserved



DDR5 – 6400 Mbps/data pin

DDR4 – 3200 Mbps/data pin

DDR3 – 1600 Mbps/data pin

DDR2 – 800 Mbps/data pin

DDR1 – 400 Mbps/data pin

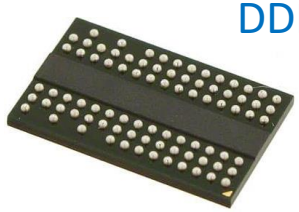
PC100 SDRAM – 100 Mbps/data pin

SDRAM has made significant gains in per-pin data throughput over the last 25 years

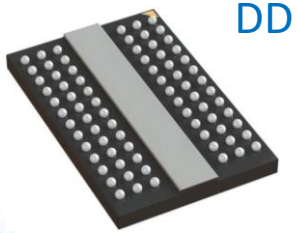
©2023 Flash Memory Summit. All Rights Reserved

PC100 SDRAM – reference synchronous main memory

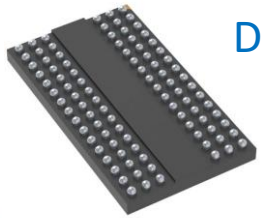
DDR1 – prefetch 2 bits, first main memory with a data strobe



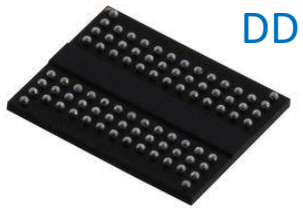
DDR2 – prefetch 4 bits, differential strobes, on-die termination



DDR3 – prefetch 8 bits, improved calibration, command-dependent ODT



DDR4 – improved calibration, ODT

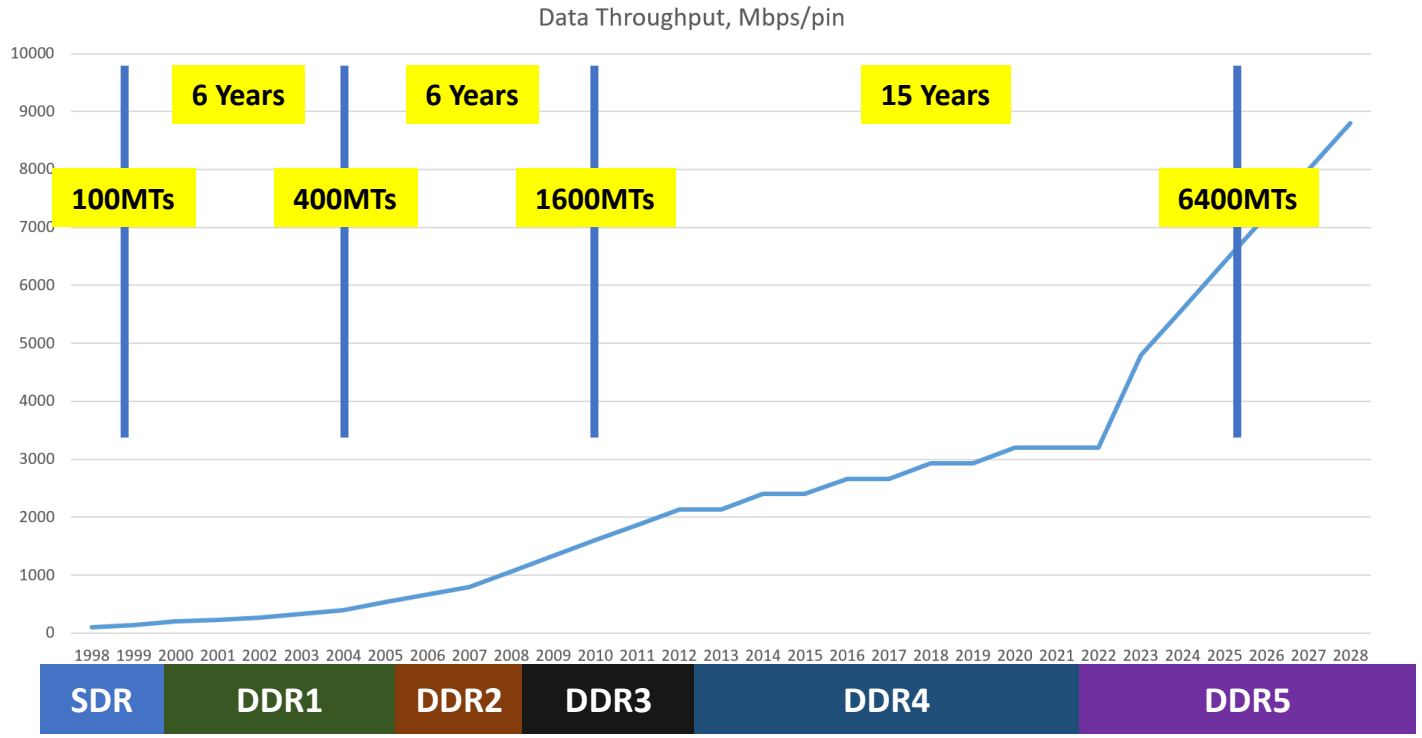


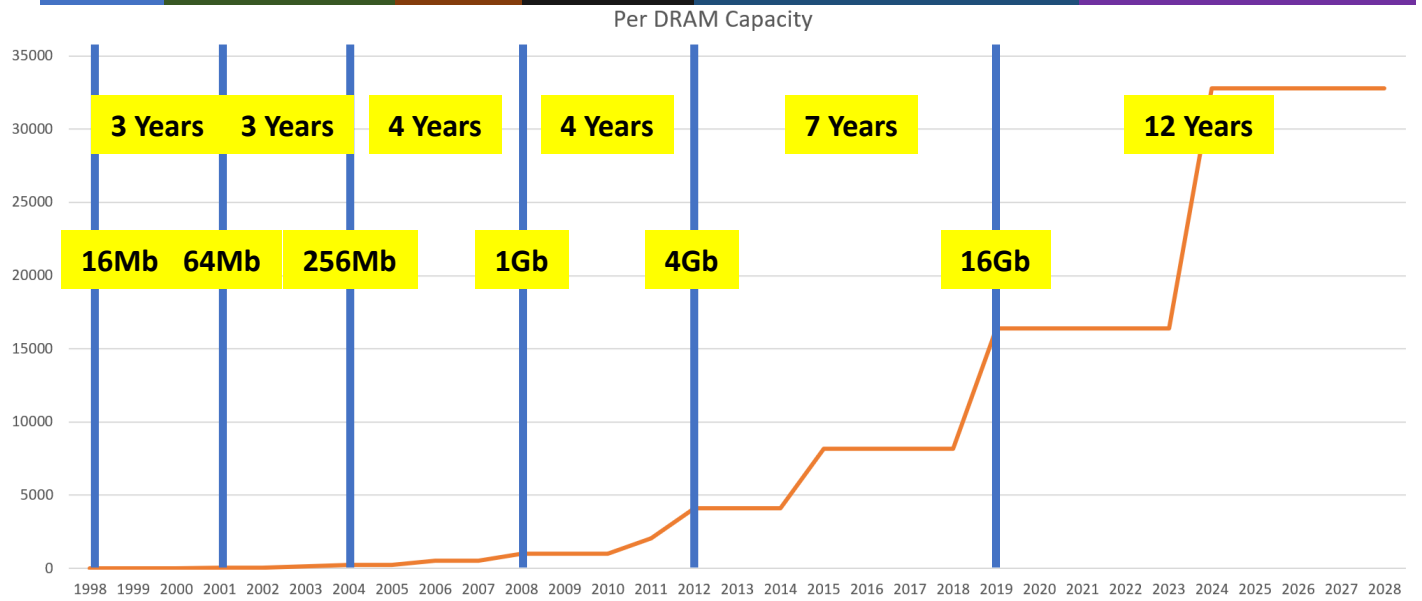
DDR5 – Prefetch 16, improved calibration, PMIC

However, random access time has only improved 28%

‘cuz I/O is cheaper than core

©2023 Flash Memory Summit. All Rights Reserved





The good news:

Data throughput has had healthy increases

DDR5 was planned for 6400 Mbps max, now extended to 8800 Mbps

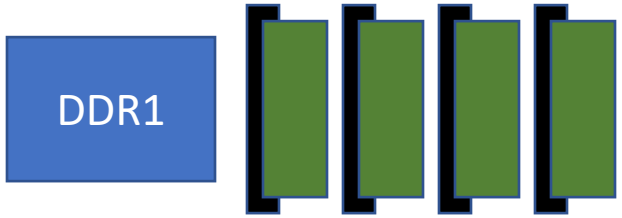
The bad news:

Speed improvements slowing

DRAM per-die capacity is taking longer with each generation

Was: quadrupling every 3 years
Is: quadrupling every 12 years

©2023 Flash Memory Summit. All Rights Reserved



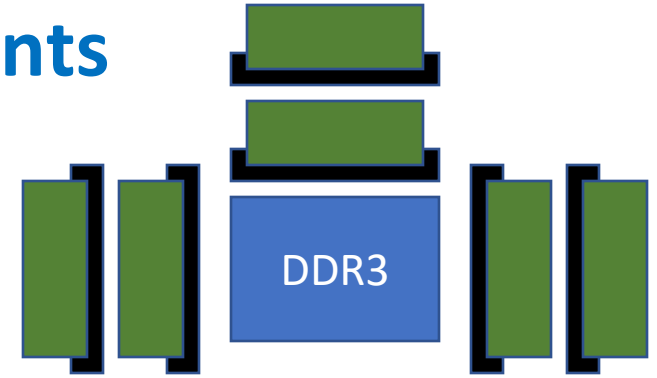
1 CH, 4 DPC, 2R/D, 256Mb = 4 GB

6X

2 CH, 3 DPC, 2R/D, 1 Gb = 24 GB

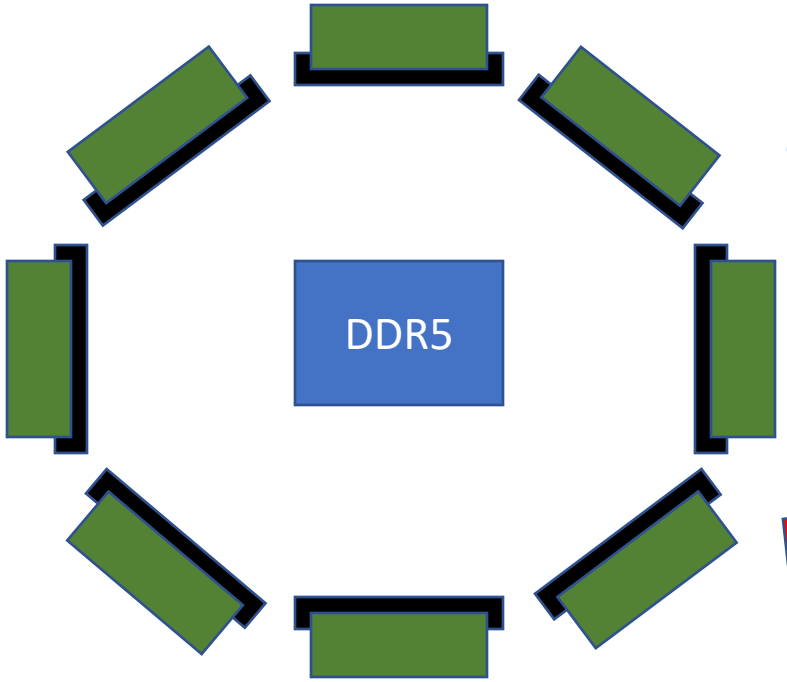
4X

Increasing frequency is slowing DIMM improvements



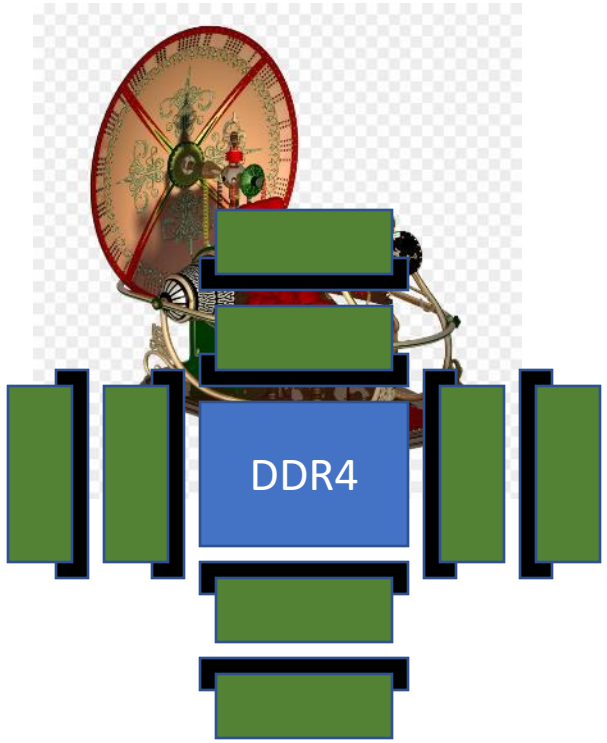
3 CH, 2 DPC, 2R/D, 4 Gb = 96 GB

5X



8 CH, 1 DPC, 2R/D, 16 Gb = 512 GB

1X



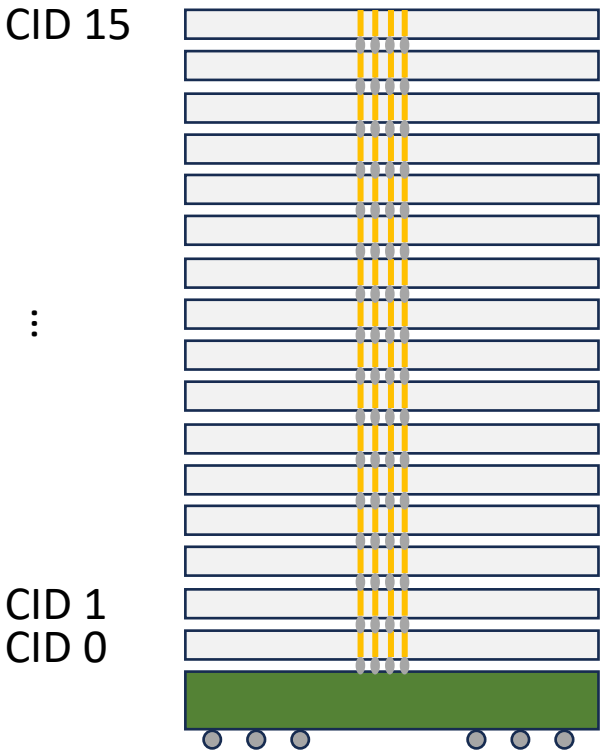
4 CH, 2 DPC, 2R/D, 16 Gb = 512 GB

CH = channel
DPC = DIMMs per channel
R/D = ranks per DIMM

Assumes no 3DS

©2023 Flash Memory Summit. All Rights Reserved

3DS to the Rescue!



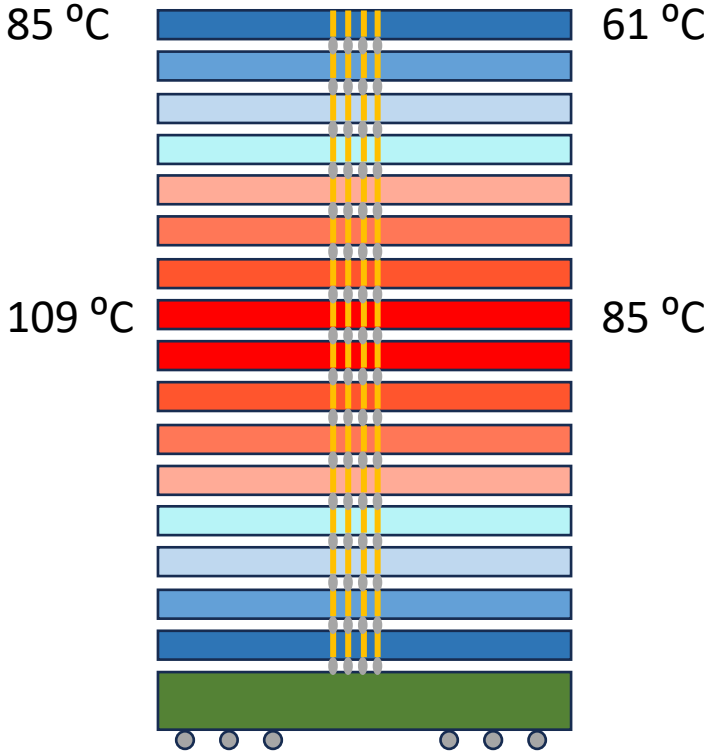
We can stack 16 DRAMs in one package!

- 16 x 16 Gb dies → 32 GIGABYTES per stack! Hooray!
- That 64 GB DIMM suddenly becomes a 1 TB DIMM!
- Thin the die to expose through-silicon vias (TSVs)!
- Microbump or high density interconnect them!
- The bottom die will proxy the stack for only 1 load!

Almost too good to be true!

©2023 Flash Memory Summit. All Rights Reserved

Ummmm…



Flash Memory Summit

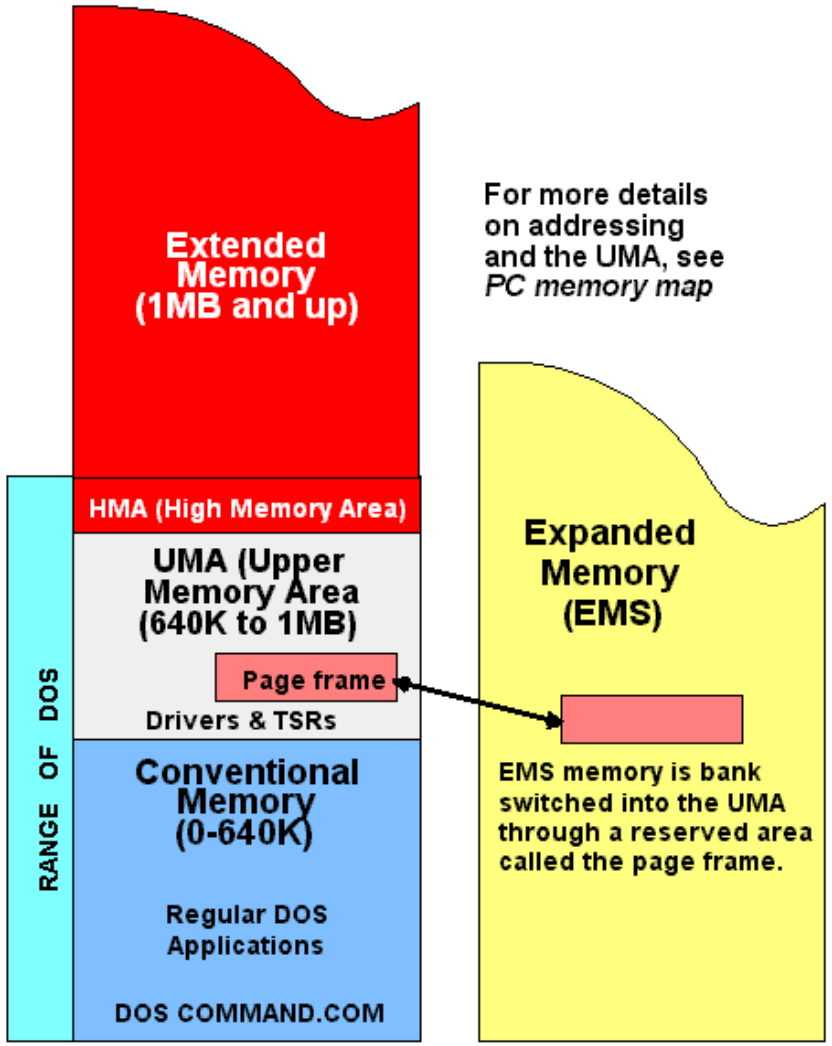
Reality check on 3DS

- The middle dies get really, really hot
- Compound die yield is a problem
  - 97% yield per die is nice
  - 0.97^{16} = 61% yield
  - And this assumes all die speed bin equivalently fast
- Manufacturability of 3DS continues to drag
- Refreshing 16 die? Really???
- 3DS dropped to 8 die, then to 4 die, then…

3DS is giving way to dual die package (DDP)

- Simpler assembly using more standard methods such as redistribution layers
- Die thinning is optional
- Requires new logic support

©2023 Flash Memory Summit. All Rights Reserved



Memory Expansion is Not New

In the 1980s, Expanded and Extended Memory were common methods to grow the memory footprint of a PC beyond the CPU limits

Real time operating systems running on such systems had to comprehend the differences in access times

©2023 Flash Memory Summit. All Rights Reserved

Memory Pooling is Also Not New

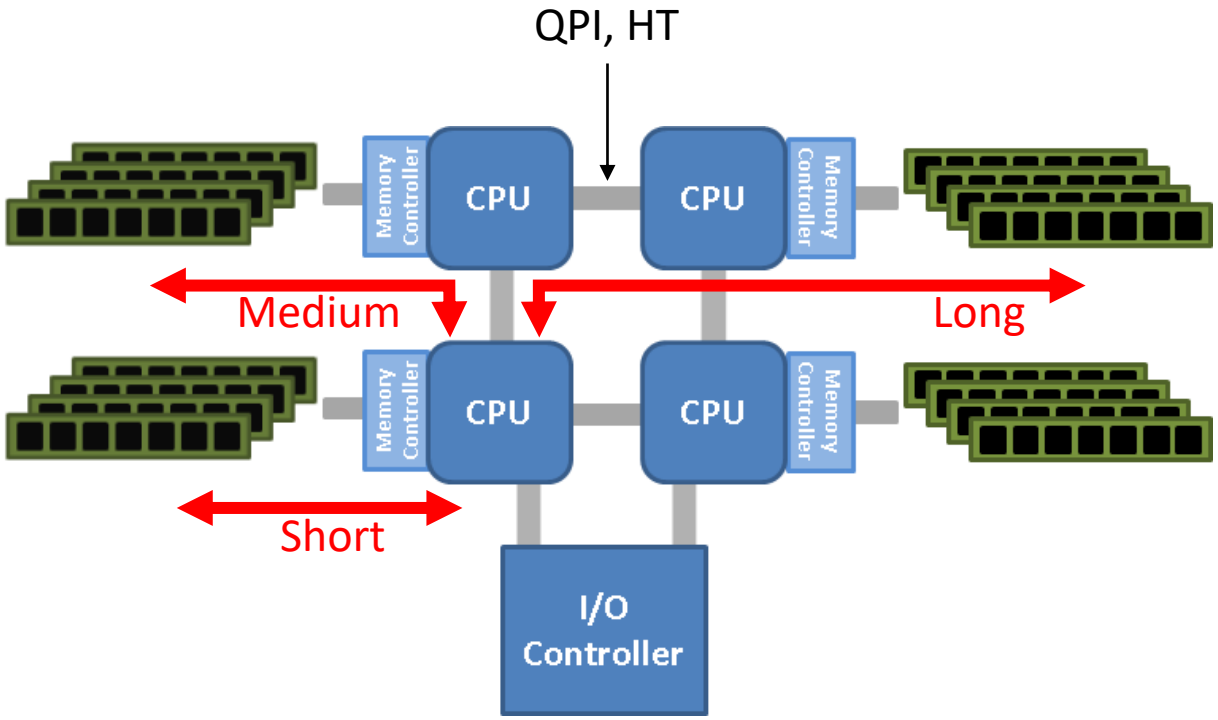
Non-Uniform Memory Architectures (NUMA) have been common ways to pool memory resources

Buses such as HyperTransport and Quick Path Interconnect have been around for decades

These NUMAs created a tier of resources

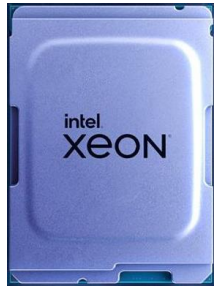
- Fastest memory attached to CPU
- Slower memory one hop away
- Slowest memory two hops away

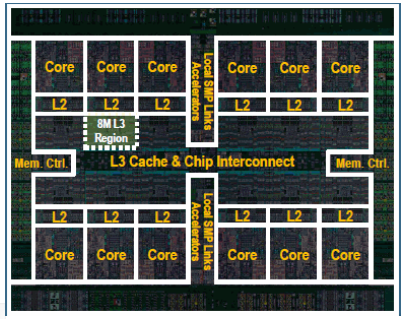
Smart software adjusted data location based on access latency

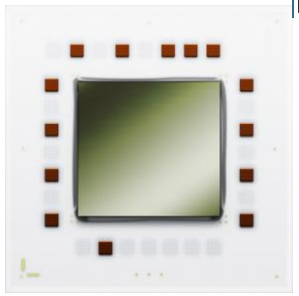


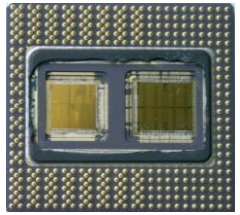
©2023 Flash Memory Summit. All Rights Reserved

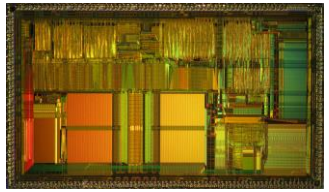
As CPUs grew hungrier

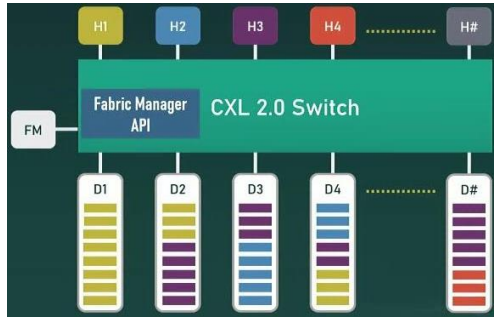


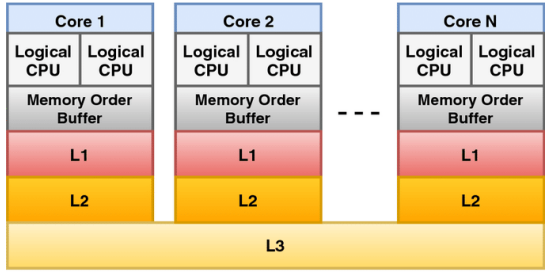






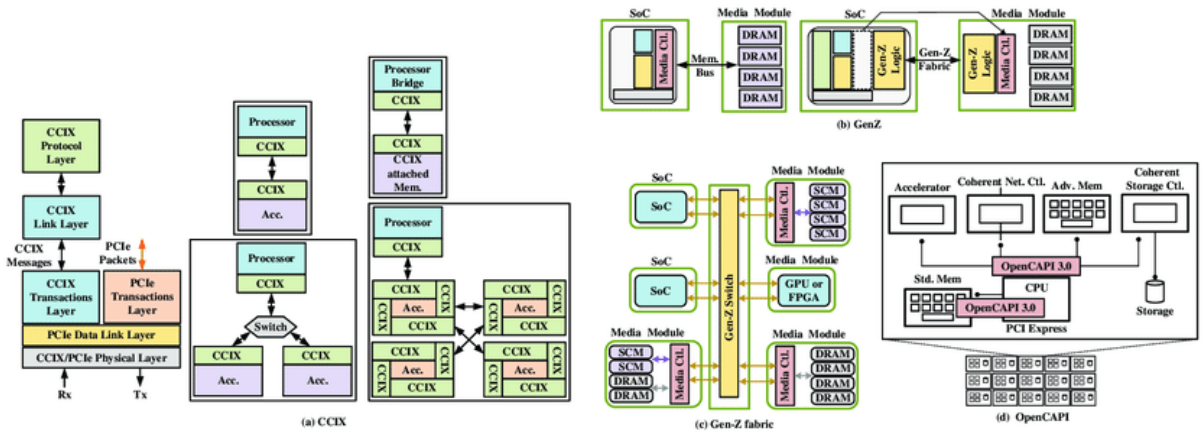






Memory solutions grew deeper and more complex

©2023 Flash Memory Summit. All Rights Reserved

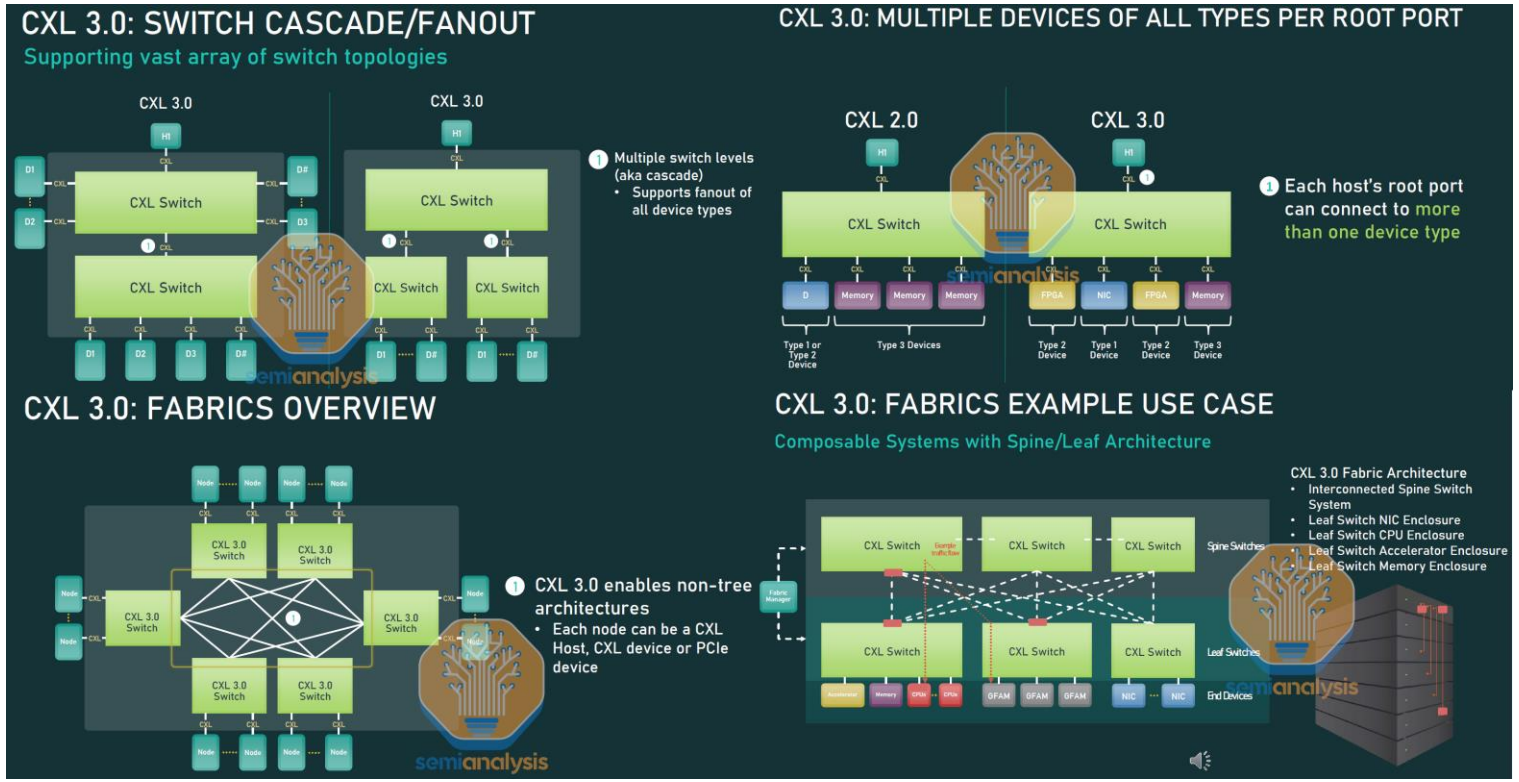


Fabric Wars

Proprietary fabrics emerged for resource sharing, however lack of standardization limited the audience

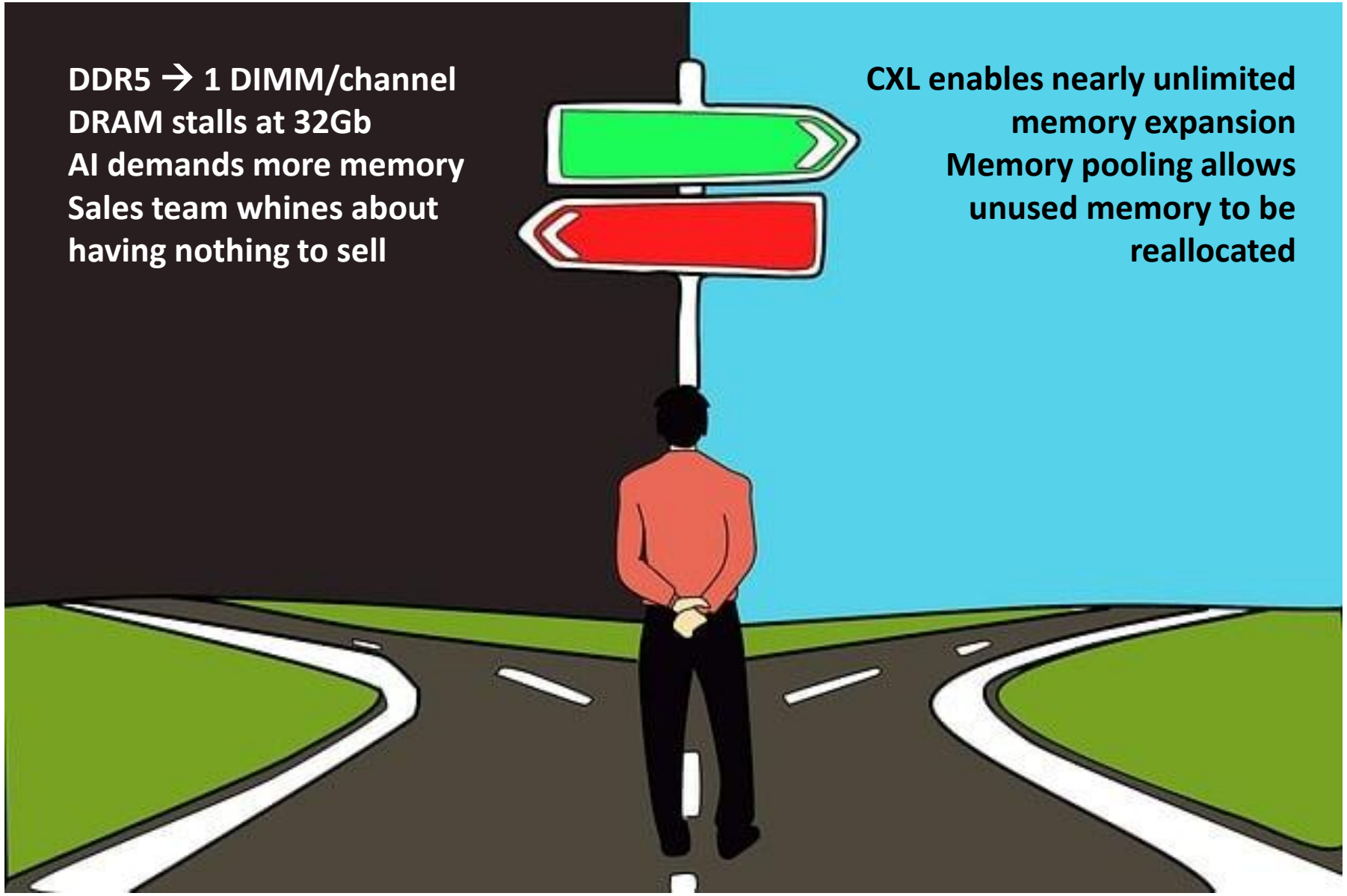
CXL Big Bang

Wide adoption of CXL allows for standardization and commoditization of expansion resources and sharing



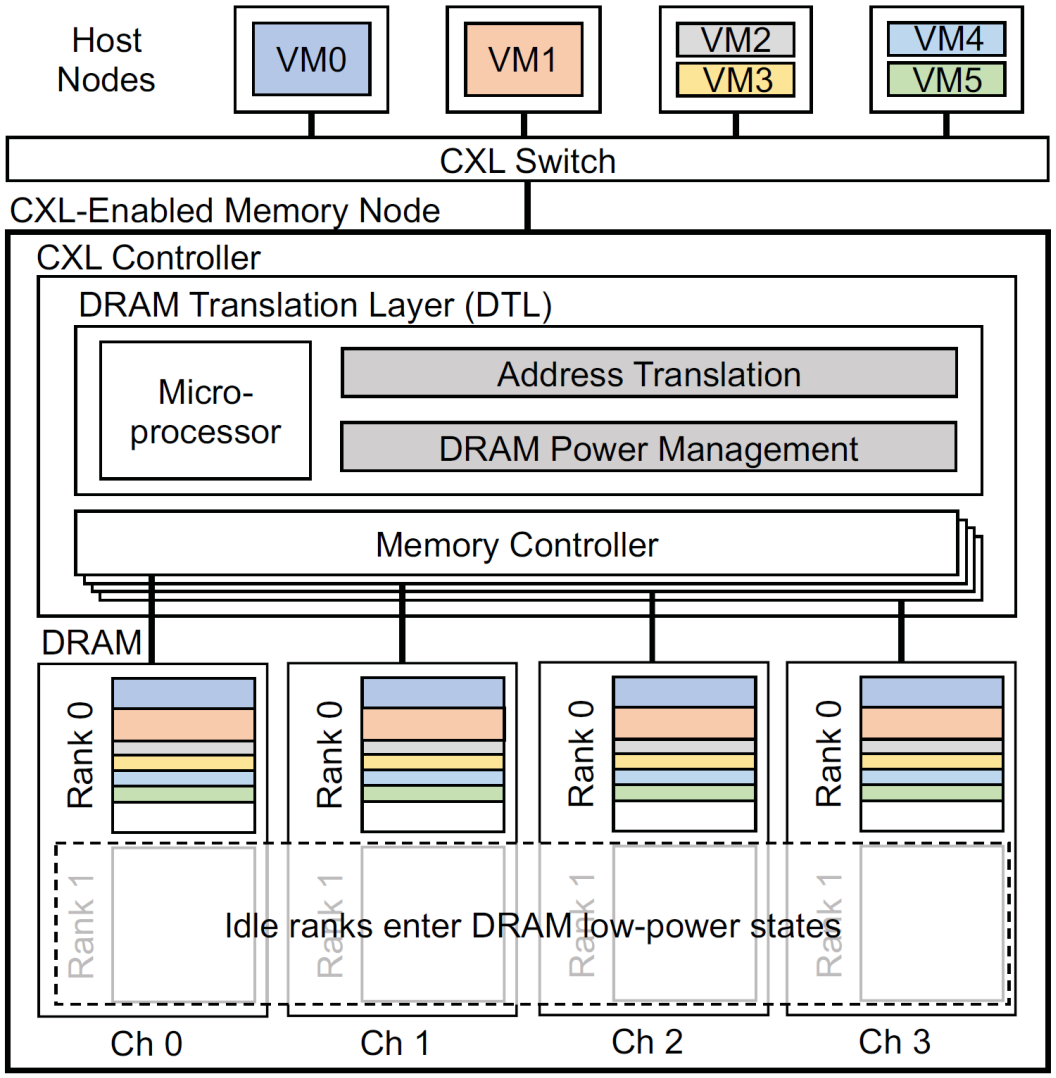
©2023 Flash Memory Summit. All Rights Reserved

Why Put DRAM on CXL?



Not to be rude, but what choice do you really have?

©2023 Flash Memory Summit. All Rights Reserved



https://dl.acm.org/doi/abs/10.1145/3579371.3589051

CXL Unifies the Fabric

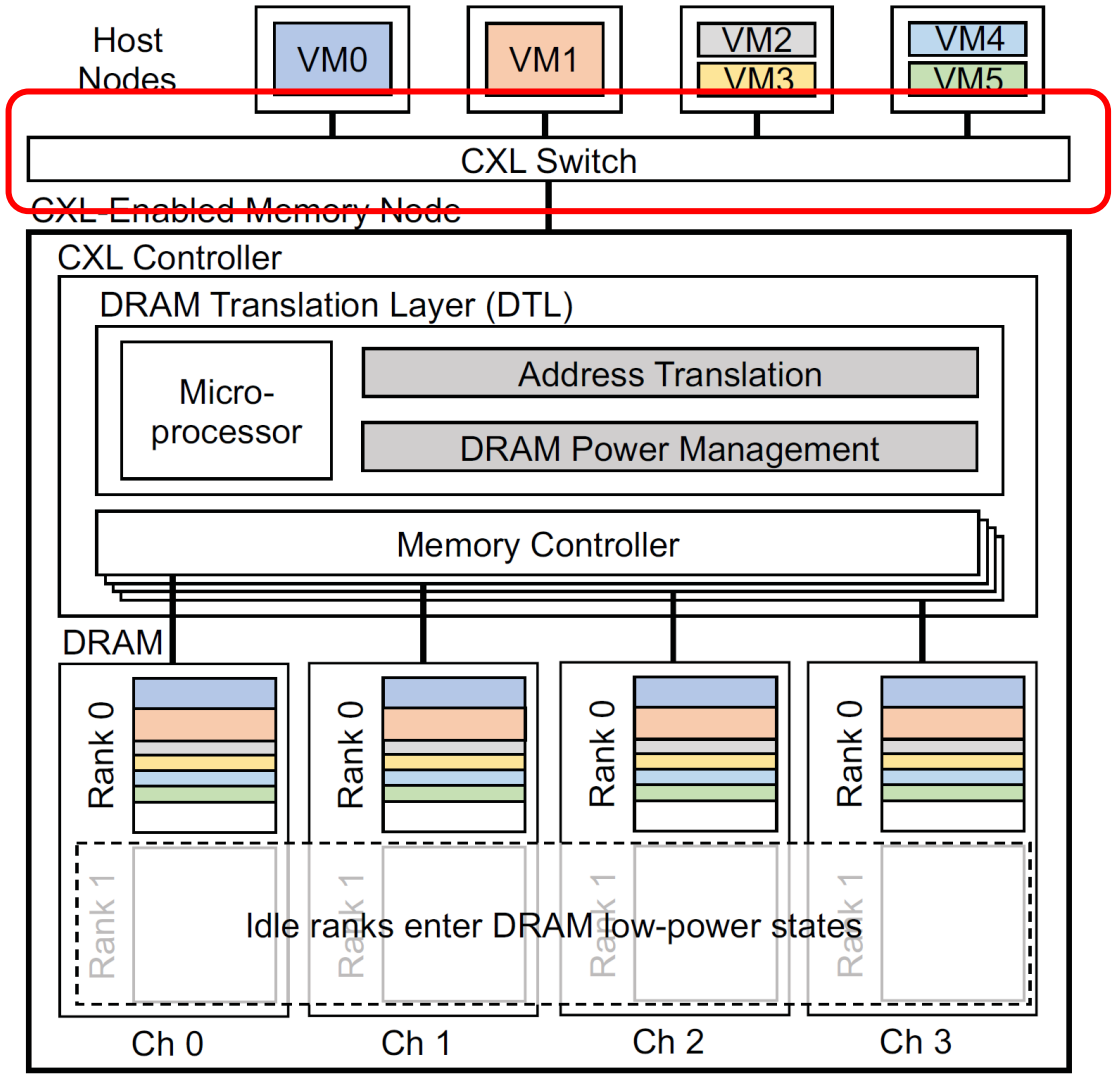
CXL is PCIe based and therefore inherits some of the features and limitations of a protocol that supports I/O or memory expansion

Legacy software only had filesystems to implement virtualization – DAX is assisting movement towards a unified addressing structure, but…

…is DAX stalled with the death of Optane?

…will CXL semantics breathe new life into a unified memory model?

©2023 Flash Memory Summit. All Rights Reserved



https://dl.acm.org/doi/abs/10.1145/3579371.3589051

CXL Switches

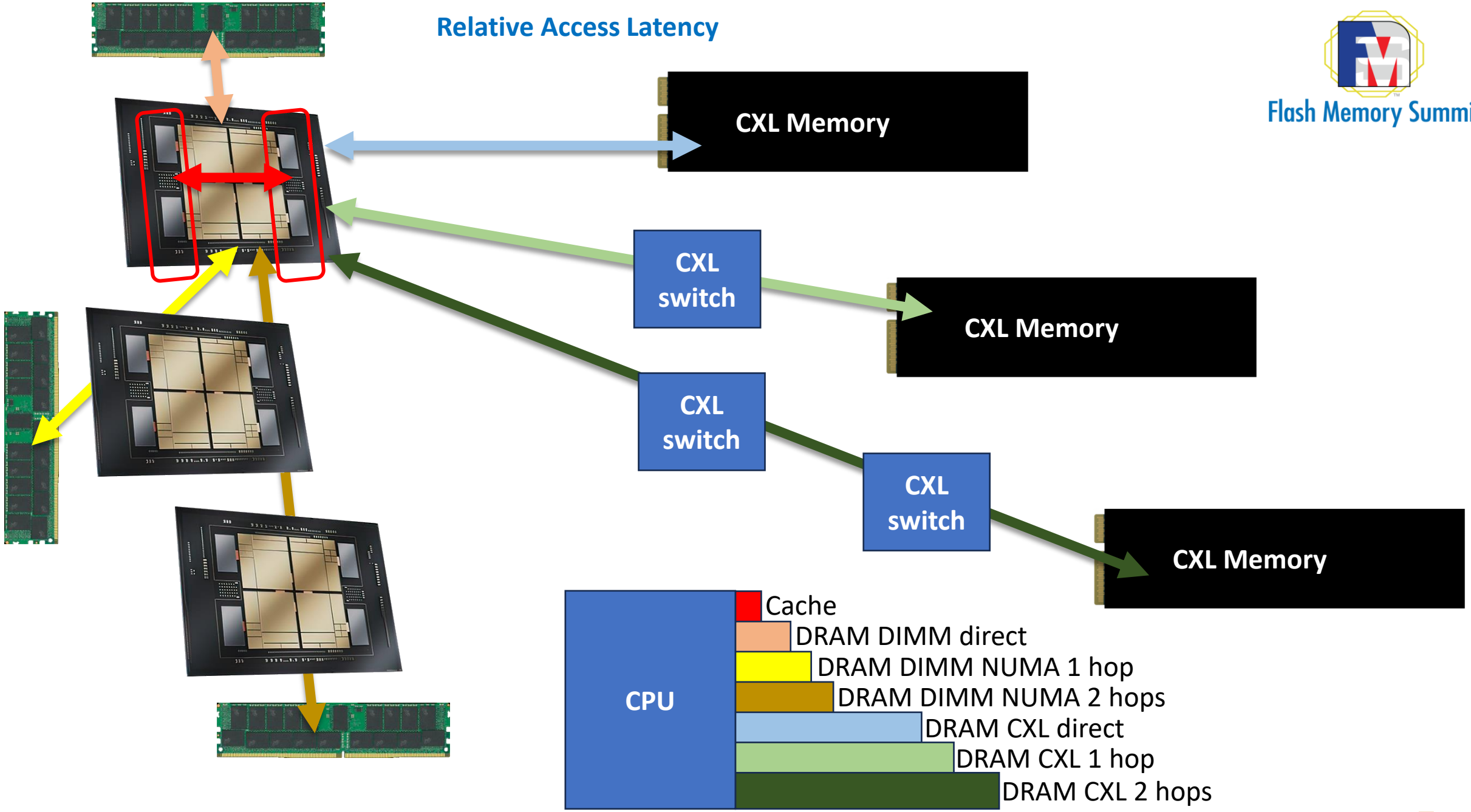
CXL switches are likely going to be the next “fabric war” as it fragments into dumb hubs versus highly intelligent controllers

A big hole in CXL 3.0 is the lack of definition of a “fabric manager”

For now, except in a few places, we can ignore the switch

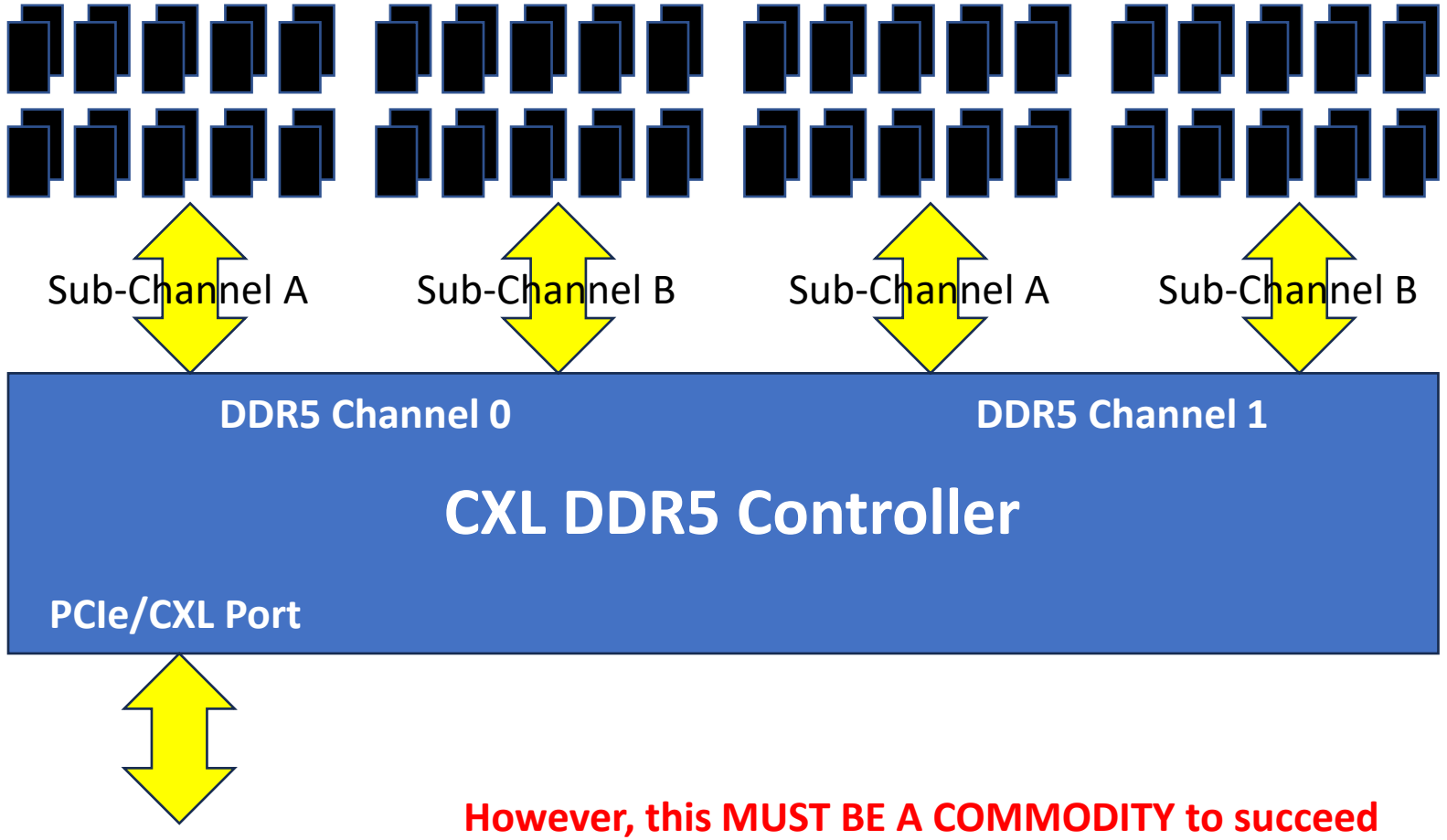
©2023 Flash Memory Summit. All Rights Reserved

Relative Access Latency



©2023 Flash Memory Summit. All Rights Reserved

Anatomy of a CXL to DRAM Bridge



However, this MUST BE A COMMODITY to succeed
Standardization required for PLUG AND PLAY compatibility

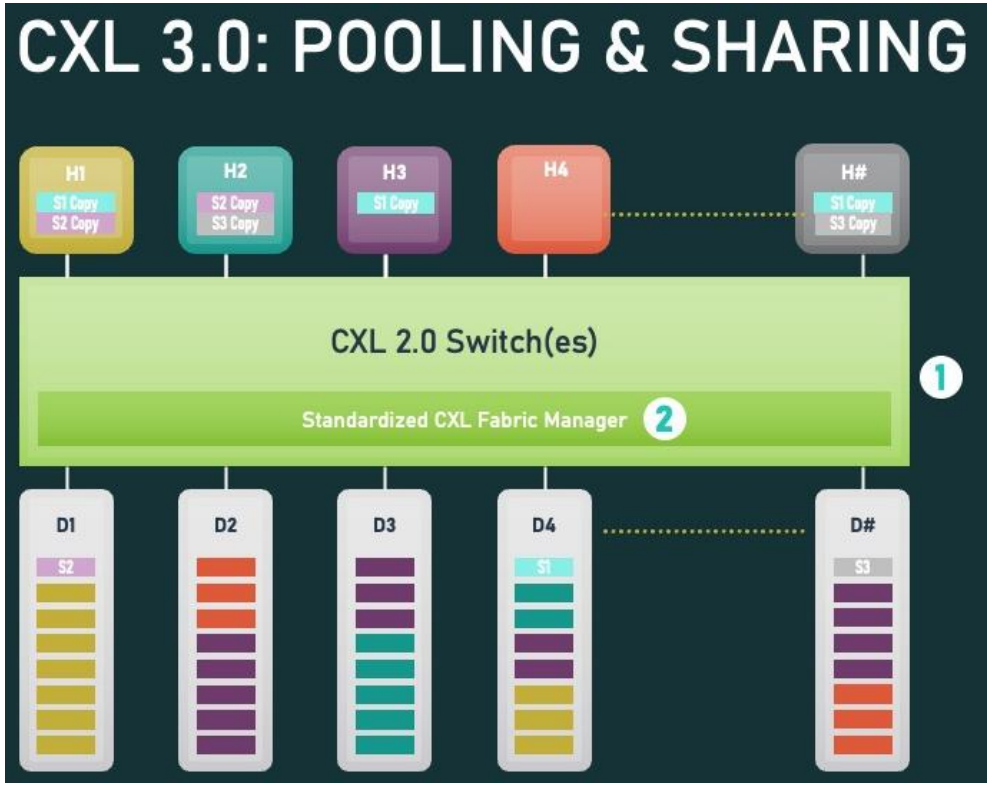
KISS: Just Do Writes and Reads

CXL is a non-deterministic protocol which allows the CXL module to operate independently

- Refresh
- Error check scrub
- Post-package repair

CXL 3+ incorporates some additional functions such as coherency

©2023 Flash Memory Summit. All Rights Reserved



It's a Brave New World with CXL Memory

CXL memory modules may be dedicated to a single processor

CXL memory modules may be allocated in chunks to different processors

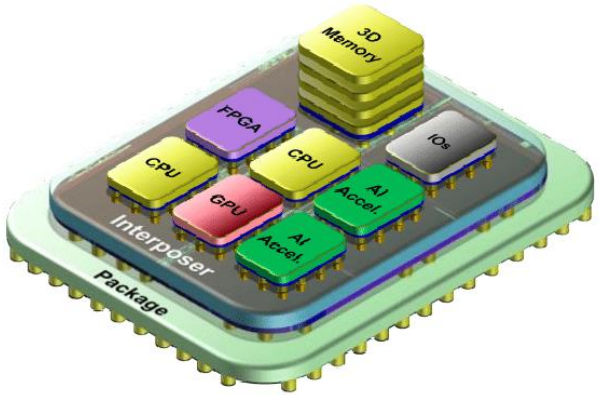
CXL memory modules may be shared by multiple processors

But What About Cache Coherency Via Back Invalidation?

Someone smarter than me needs to explain how back invalidation works if a CXL memory region is shared by a random number of CPUs…

©2023 Flash Memory Summit. All Rights Reserved

UCIe™
Universal Chiplet Interconnect Express



Another industry standards effort to pay attention to

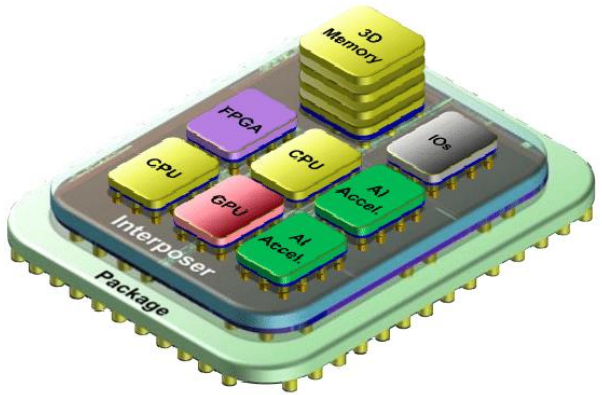
Ideally, extends one common view of resources from die to system

UCIe Quick Summary

Die to die connectivity over a substrate
Multi-lane differential signaling to 32 GT/s
Width up to 64 lanes
256 GB/s peak
Silicon or organic substrates supported
- Silicon: 25 to 55 um pitch
  - Maximum reach 2 mm
- Organic: 100 to 130 um pitch
  - Maximum reach 25 mm

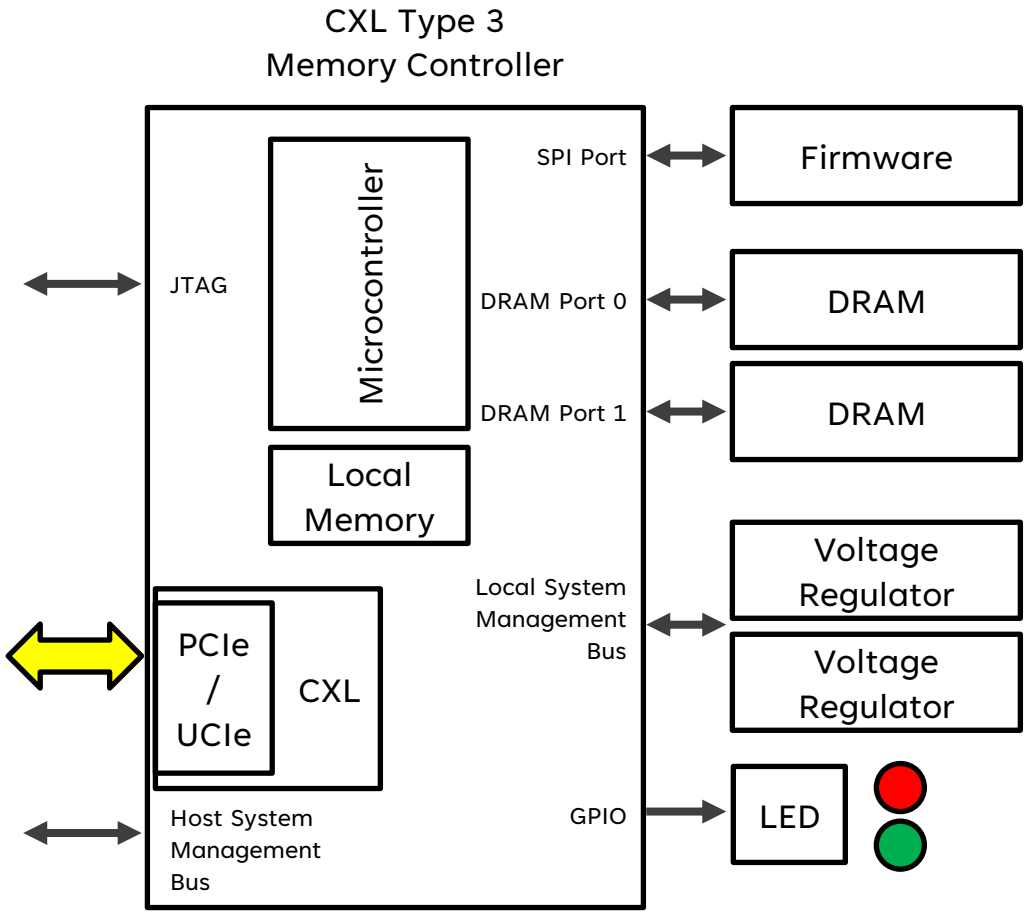
PCIe 6.0
CXL 3.0

©2023 Flash Memory Summit. All Rights Reserved



It's not "CXL on Silicon" but close enough to allow re-use of controller designs

Reality of UCIe is that it is still in diapers
For now, UCIe methods will be controlled by one supplier
We're probably 20 years from ordering die on Mouser and Digikey



Connection to DRAM can be external, on-substrate, or 3D connected

©2023 Flash Memory Summit. All Rights Reserved