Addressing the Latency Gap with Composable Architectures

Session : SOFT-201-1: Composable Infrastructure and Software Defined Storage

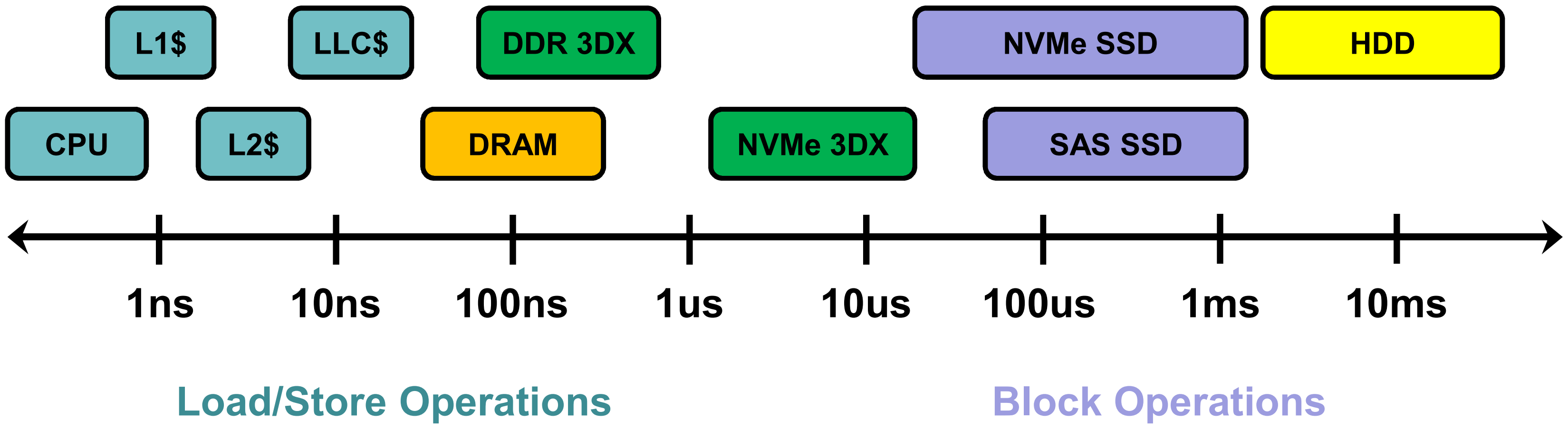
Larrie Carr

Abstract

In the past few years, the latency gap between the traditional load/store memory infrastructures and block-focused fabric and storage networks has seen the introduction of new composable architectures, CPU interfaces, processing accelerators, memory technologies and software stacks into the industry.

This presentation explores why addressing this latency gap has become so important and discusses the composable architectures enabled by technologies like OpenCAPI, CCIX, Gen-Z and CXL.

Latency Landscape



L1$

LLC$

DDR 3DX

NVMe SSD

HDD

CPU

L2$

DRAM

NVMe 3DX

SAS SSD

1ns 10ns 100ns 1us 10us 100us 1ms 10ms

Load/Store Operations

Block Operations

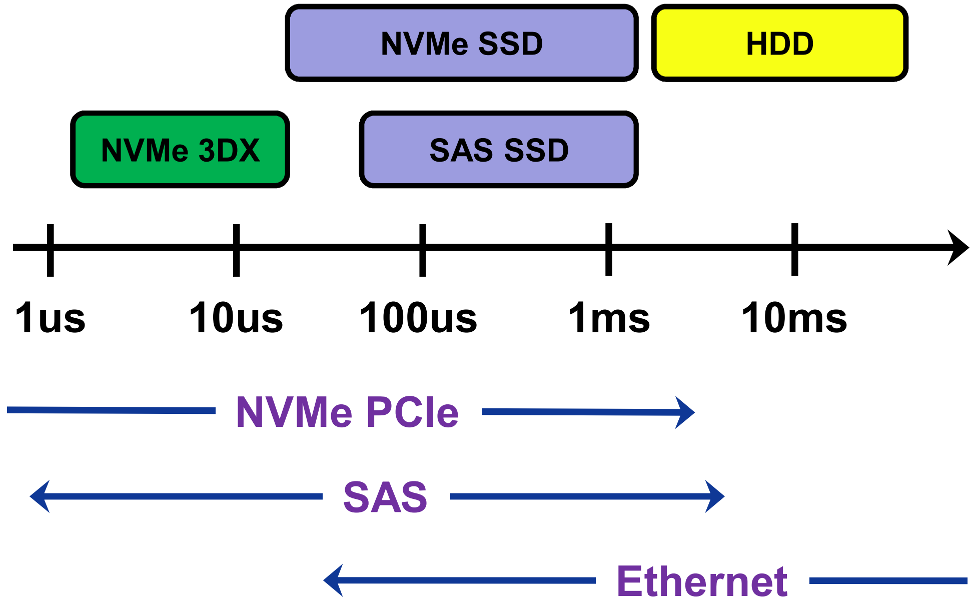
- Data needs to move from the right to the left in order to be “processed”
- Data movement and storage at the 1us “gap” is difficult for any technology

Block Latency Landscape

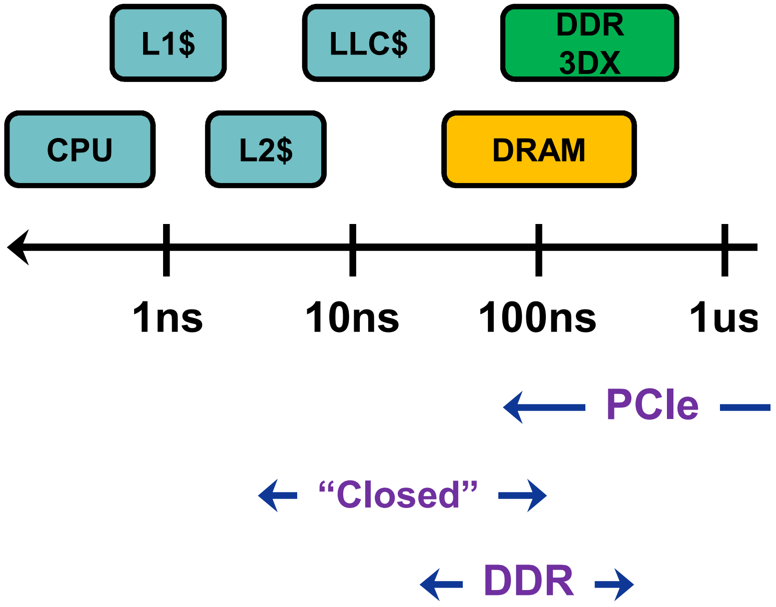
- Bit density increasing
- Cost per bit decreasing
- Connectivity rates doubling
- Latency and access times decreasing

- Lots of design flexibility and choice

- Life is pretty good….



Load Store Latency Landscape

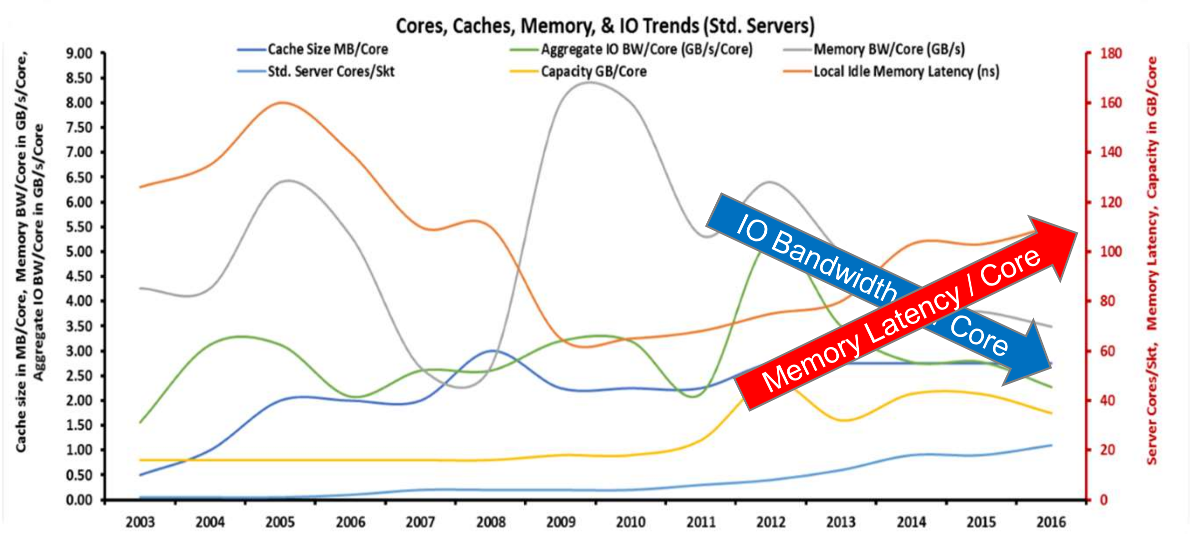


- Moore's law is dead - costs increasing
- IPC efficiency decreasing
- Latency increasing
- Connectivity rates evolving incrementally
- Closed proprietary interfaces

- Architecture flexibility limited to time of installation

- DRAM is a unique player…

Processor Memory Bottleneck

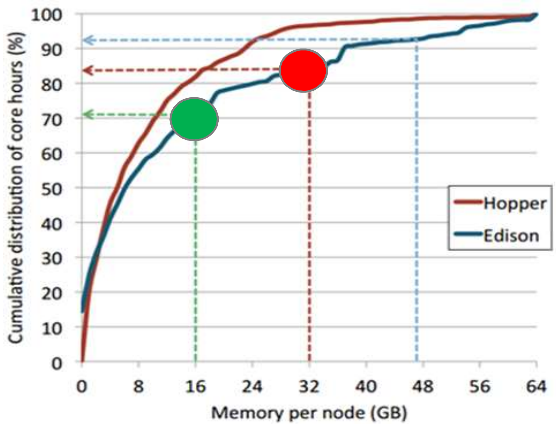


https://blog.dellemc.com/en-us/memory-centric-architecture-vision/

- Current processor DRAM buses provide limited quantity & performance scaling
- Multi-core processor devices experiencing *increased* latency per core

Flash Memory Summit

Reducing Stranded Resources



U.S. DEPARTMENT OF ENERGY | Office of Science

2014 NERSC Workload Analysis, Oct '15

NERSC HPC Datacenters	Hopper 	Edison 	Cori
Year	2011	2013	2016
Datacenter Nodes	6,384	5,576	9,300
DDR / Node	32 GB @ 54 GB/s	64 GB @ 102 GB/s	96 GB @ 90 GB/s
HBM / Node	N/A	N/A	16 GB @ 400 GB/s

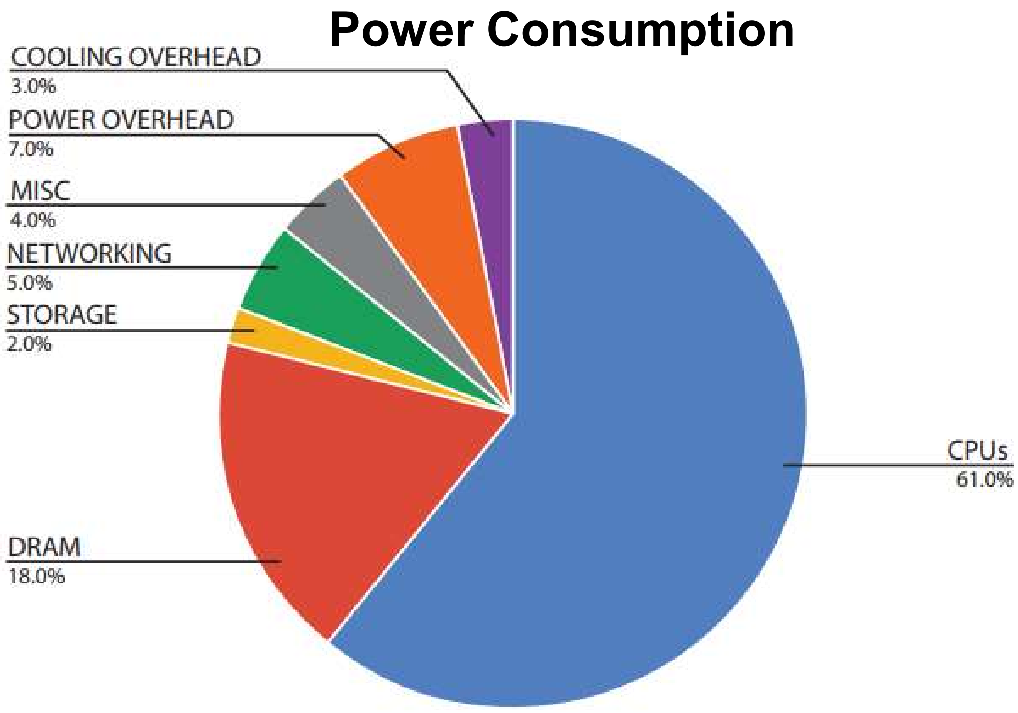
🔴 Only 16% of Edison jobs would NOT run in Hopper 32 GB nodes

🟢 71% of the Edison jobs only need 16 GB of DRAM and fit into Cori's HBM memory

- By sizing each node's DRAM capacity to service the largest of jobs, overall DDR efficiency decreases and much of the system DRAM capacity is stranded

Microchip

DRAM Power Consumption



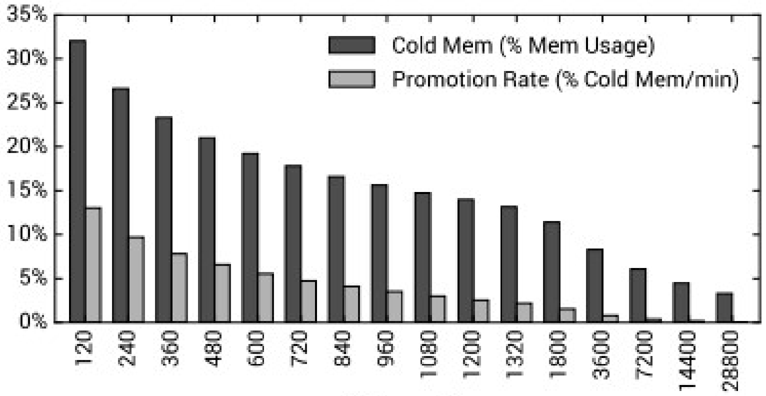
The Datacenter as a Computer, Google, 3rd Edition

- DRAM consumes 18% of Google’s WSC total power consumption
- Memory costs upwards of 50% of total server costs
- Up to 25% of the memory bandwidth consumed by “overhead”
  - Memmove, allocation, compression, ….

Compressing Memory

Software-Defined Far Memory in Warehouse-Scale Computers, Google, ASPLOS 2019

Characterize the 'Cold memory page'



Seconds since Last Access

- Google: Reduce the memory TCO by improving the density of data store in "colder" memory
  - Data gets cold after 2 minutes
  - Applications access 15% of their total cold memory every minute
  - Compress the cold memory to reduce footprint
- Results:
  - ✓ 20% of memory was compressible on average
  - ✓ 4% to 5% TCO savings on DRAM

- Microsoft: Open sourcing their Zipline memory compression RTL at OCP 2019

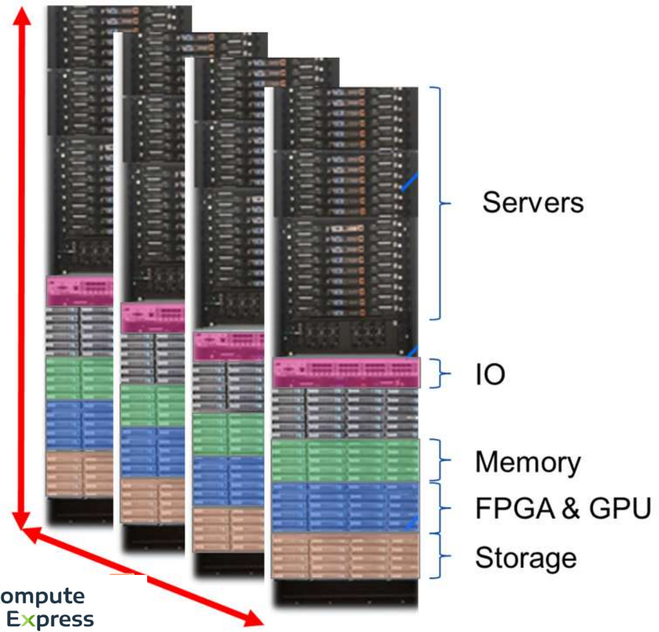
Composable Architectures

Composable architectures are about choice

- Sharing pools of resources (CPU, memory, FPGA) to reduce resource stranding
- Scaling to the server resources to match the workload
- Connecting to pools of resources

Forcing emerging memory and compute technologies behind legacy interfaces (PCIe/SATA) is inefficient

New open load/store standards provide the low-latency connectivity required to build flexibility **inside and outside the server…**

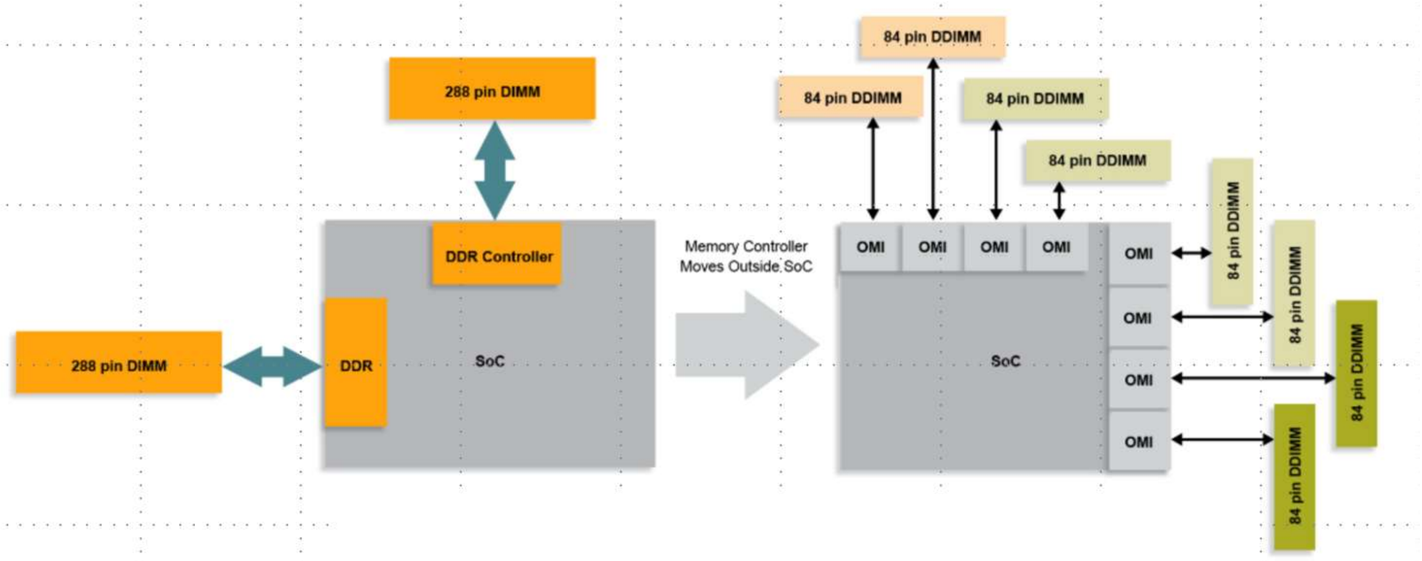


CXL Compute Express Link

GEN Z

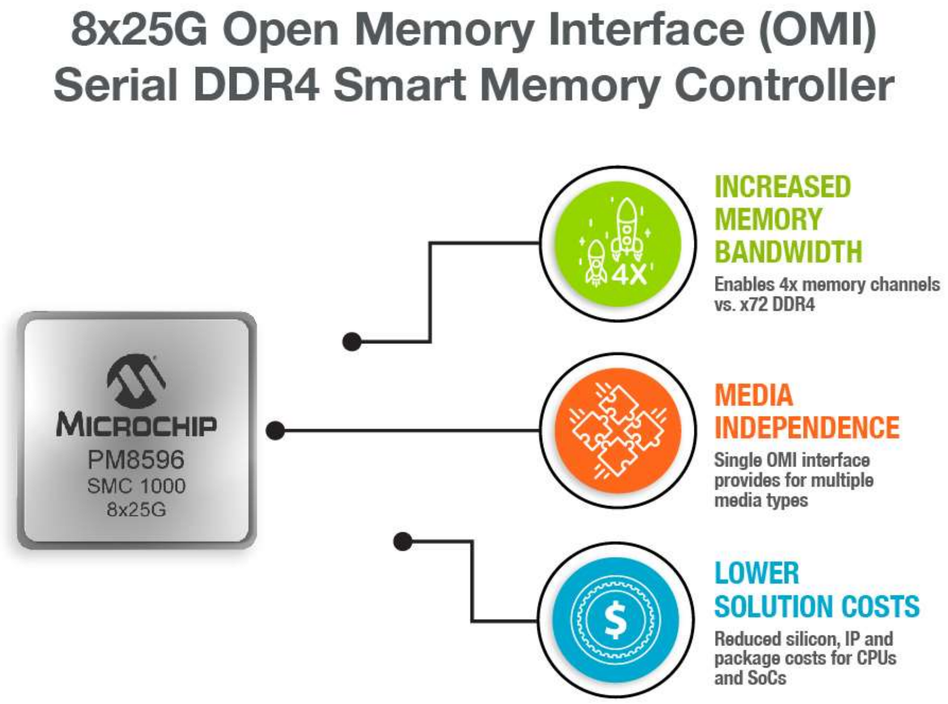
OpenCAPI™

Serial Memory Controllers



- Dramatic bandwidth expansion with significant reduction in pin count
- Media independence - enable multiple memory types
- Lower cost SoC packaging by using serial memory interfaces

The Smart Memory Controller



8x25G Open Memory Interface (OMI)
Serial DDR4 Smart Memory Controller

INCREASED MEMORY BANDWIDTH

Enables 4x memory channels vs. x72 DDR4

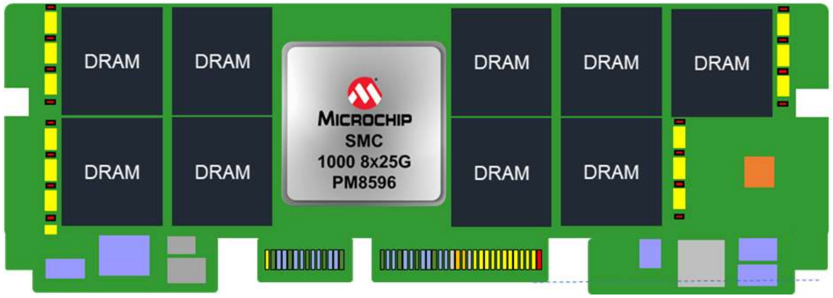
MEDIA INDEPENDENCE

Single OMI interface provides for multiple media types

LOWER SOLUTION COSTS

Reduced silicon, IP and package costs for CPUs and SoCs

Smart Memory Controller PM8596



Open Memory Interface

- OIF-28G-MR
- Dynamic low-power modes

DDR Interface

- x72 bit DDR4-3200
- Up to 4 ranks, 3D stack
- 16Gbit DRAM support

Persistent Memory

- NVDIMM-N module support

On-Chip Processor

- Initialization, monitoring and diagnostics
- Open source firmware

Security

- Hardware root-of trust
- SECDEC with memory scrub

Applications

- JEDEC DDIMM applications
- Chip-down with conventional DIMMs
- FPGA accelerator memory fanout
- ML memory bandwidth expansion (versus HBM architectures)

Flash Memory Summit

Summary

- New low-latency load/store memory interfaces allow new compute and memory architectures
- New low-latency non-persistent load/store memory technology will play a role with these new architectures
  - Less expensive and power-hungry than DRAM
- Come see our Serial Memory Controller demo at our booth…

MICROCHIP

Thank You