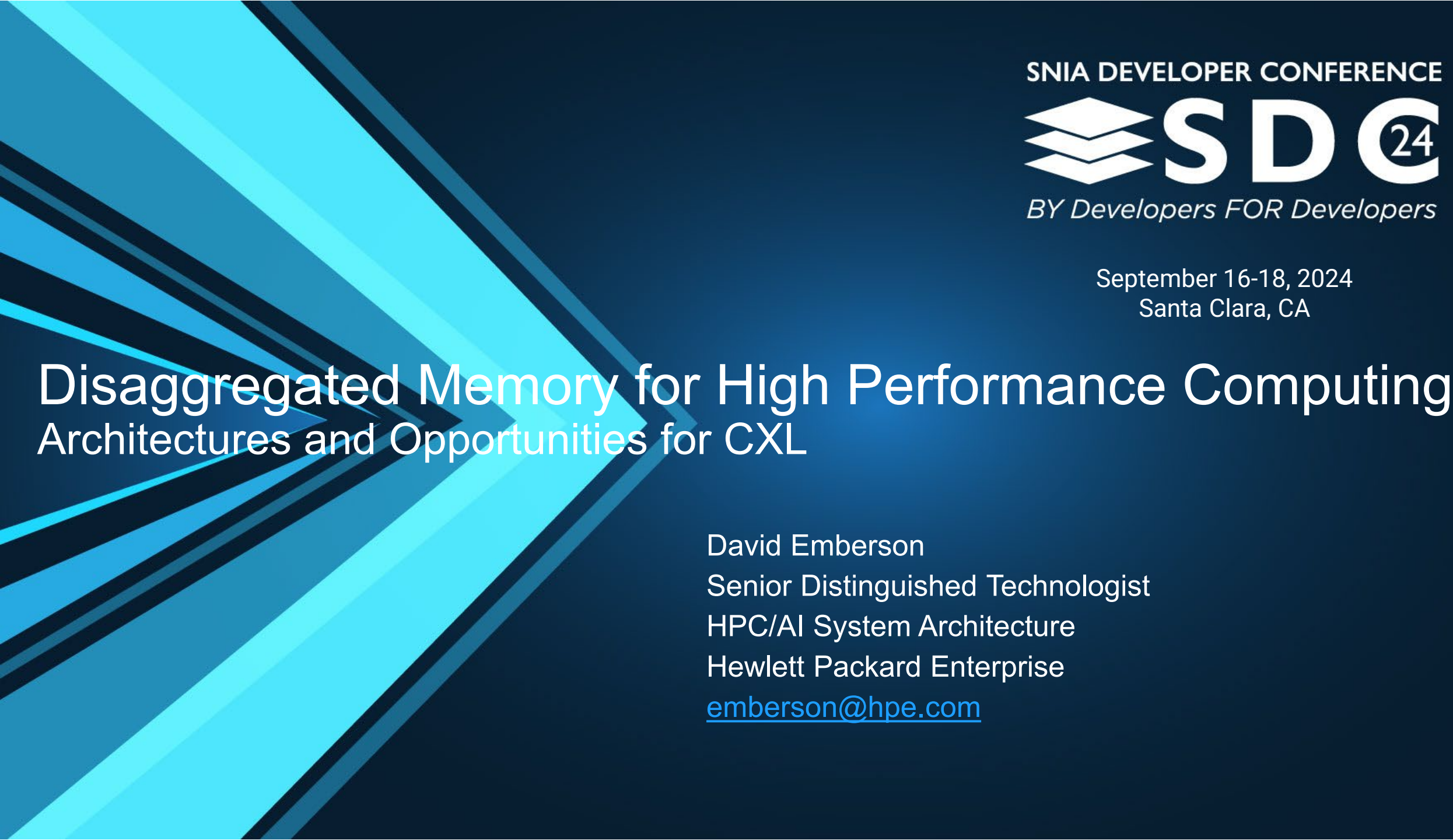


September 16-18, 2024
Santa Clara, CA

Disaggregated Memory for High Performance Computing
Architectures and Opportunities for CXL

David Emberson
Senior Distinguished Technologist
HPC/AI System Architecture
Hewlett Packard Enterprise
emberson@hpe.com

A Brief History of Disaggregated Memory Research at HPE

First, some definitions…

FAM: Fabric-Attached Memory. Memory that is disaggregated from the CPUs and accessed via the network fabric.

Persistent Memory: Memory that retains data for future use after the process that stored the data terminates. (Often confused with “resilient” memory or “non-volatile” memory.)

Non-Volatile Memory: Memory that retains its contents when the power is removed, e.g. Flash, PCM, MRAM, etc.

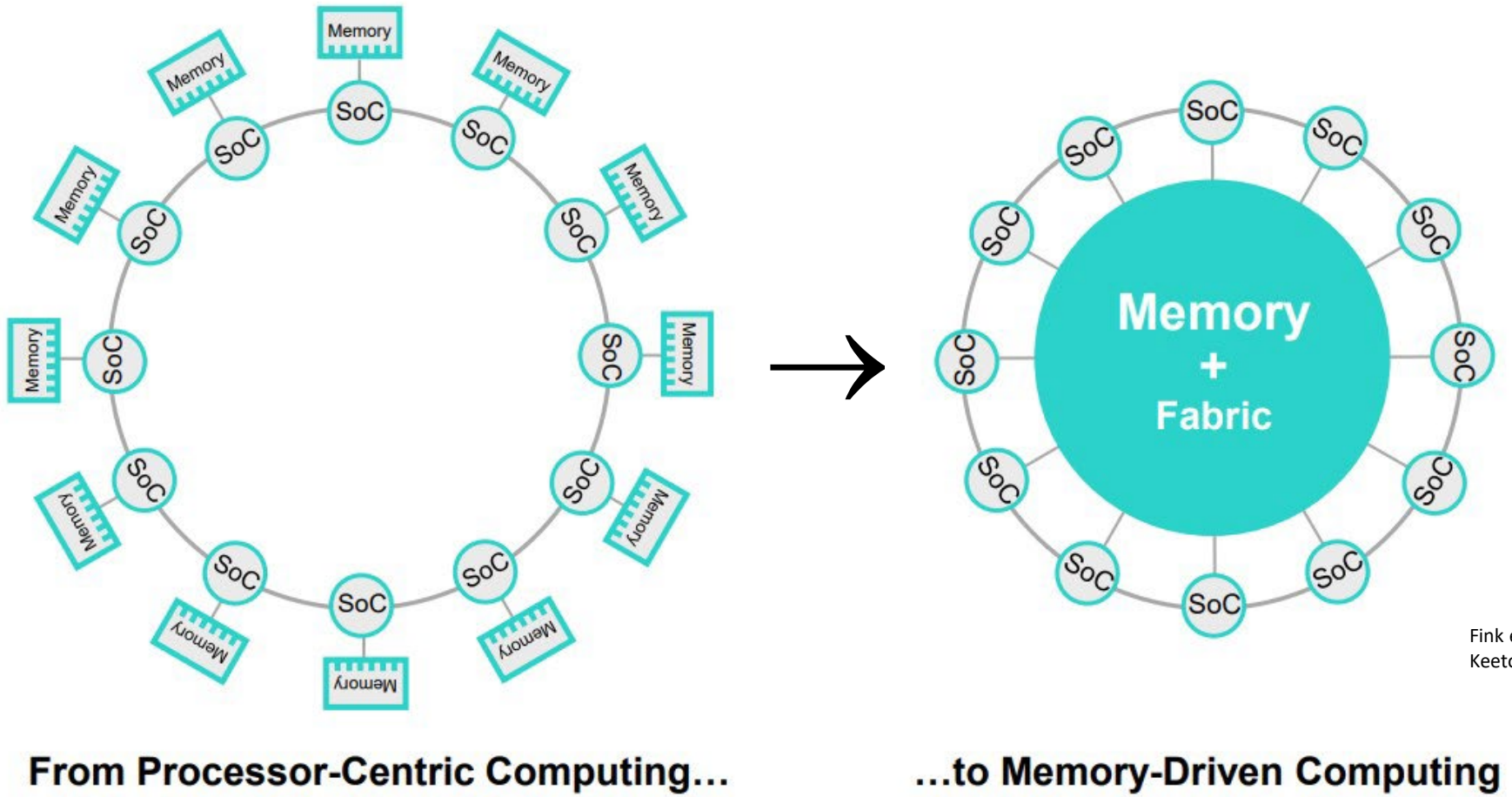
Resilient Memory: Memory that guarantees a defined level of service in the presence of failures. This “level of service” can include guaranteed data retention for a specified period of time, guaranteed recovery to a known state in the event of a crash or power failure, guaranteed write wear properties, etc.

Write Wear: The minimum number of times a memory element (cell/row/page) can be written before it fails to successfully store and retain the written data. Flash, ReRam and phase-change memory (e.g. Optane) suffer from limited write wear. Due to the nature of the underlying technology, DRAM does not suffer measurable write wear.

Key Inventions – Fast(er) Bit Addressable NVM

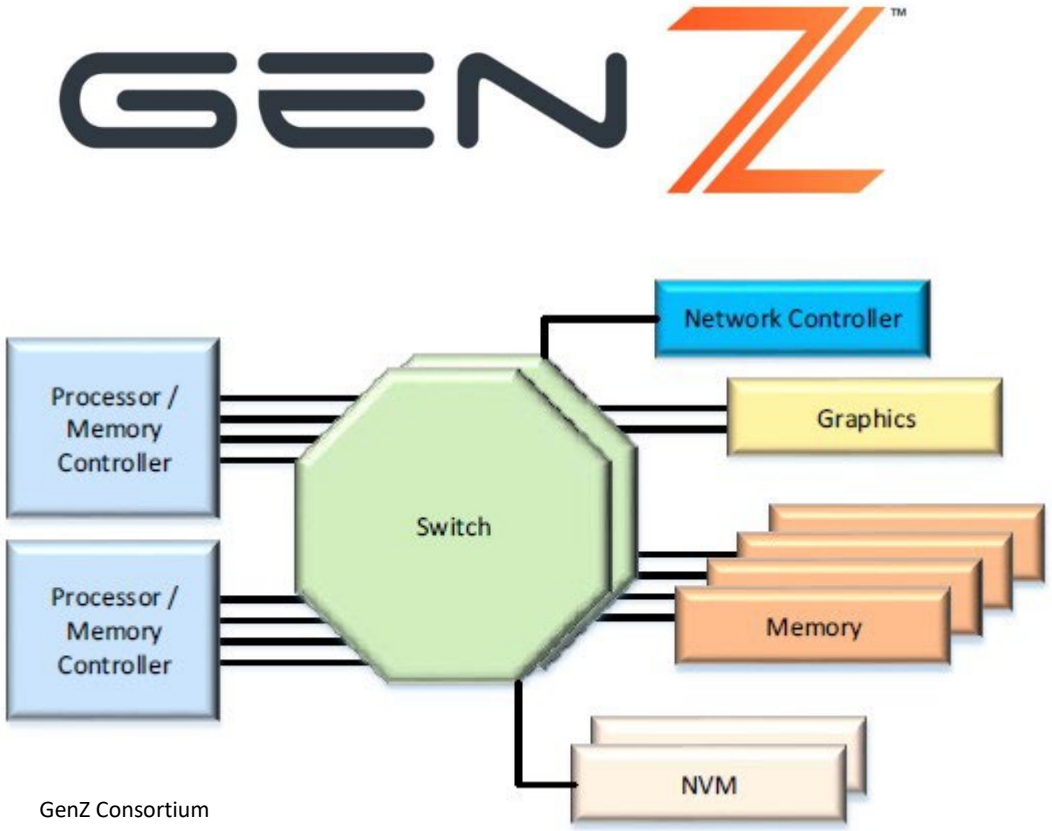
- Memristor ReRAM – Oxygen migration in TiO_2
  - Strukov, Snider, Stewart, Williams, *Nature*, May 2008, HP Labs)
- Phase Change Memory (PCM) – Amorphous/crystalline phase change in chalcogenide glass (GeSbTe)
  - Charles Sie, PhD thesis on Phase Change Memory, Iowa State, 1966
  - Shanefield, US Patent 3,448,302, June 1969 ITT
  - Ovishinsky, US Patent 5,166,758, November 1992, Energy Conversion Devices, Inc.
  - Intel/Micron Optane™ announced, 2015
- Magnetic RAM (MRAM), carbon nanotube RAM (NRAM), etc.

Memory Driven Computing – 2014



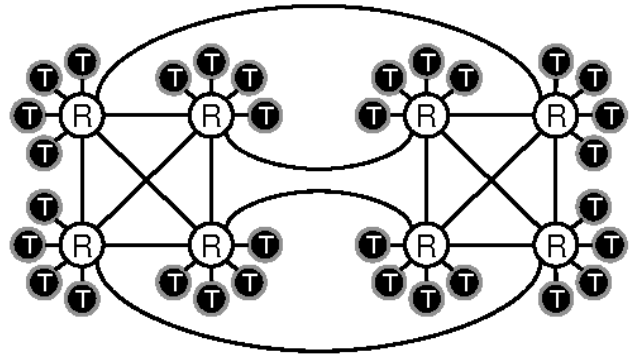
Fink et al, HP Labs 2014
Keeton et al, 2016

Gen-Z + HyperX – Memory-Semantic Interconnect - 2016

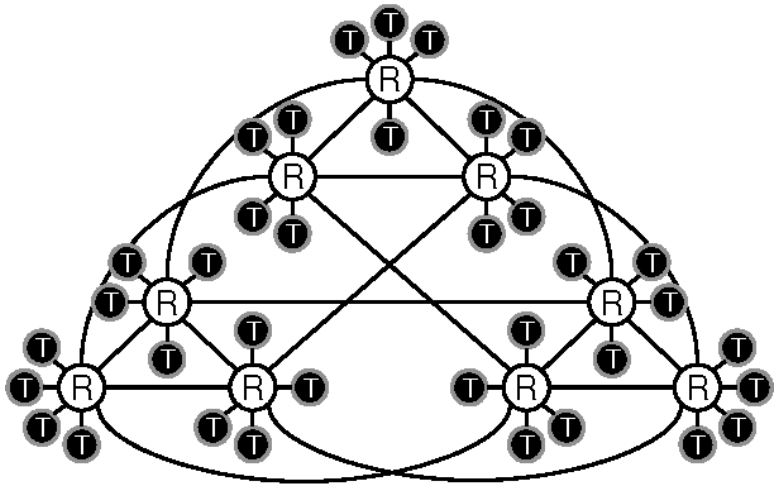


GenZ Consortium

+



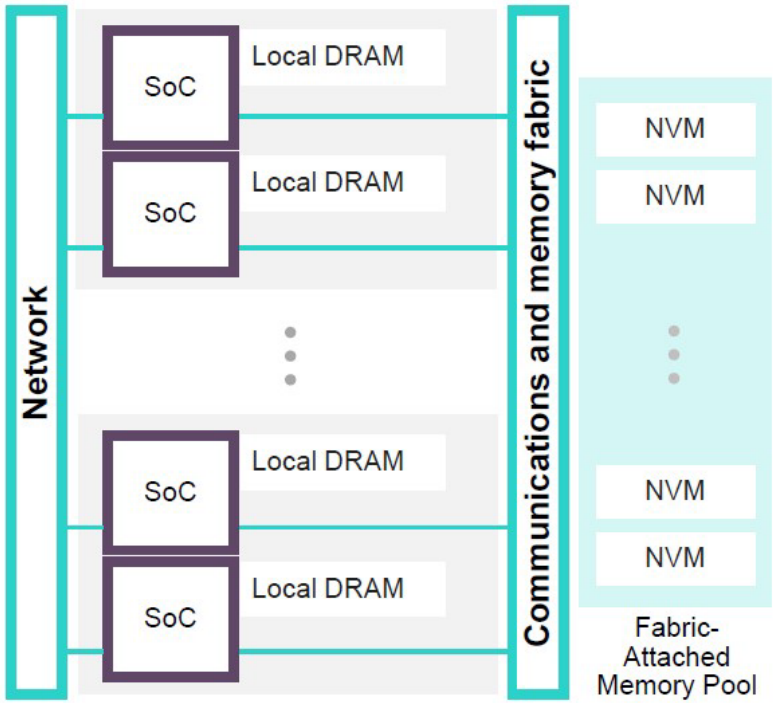
(a) $L = 2, S_1 = 2, S_2 = 4, K = 1, T = 4$

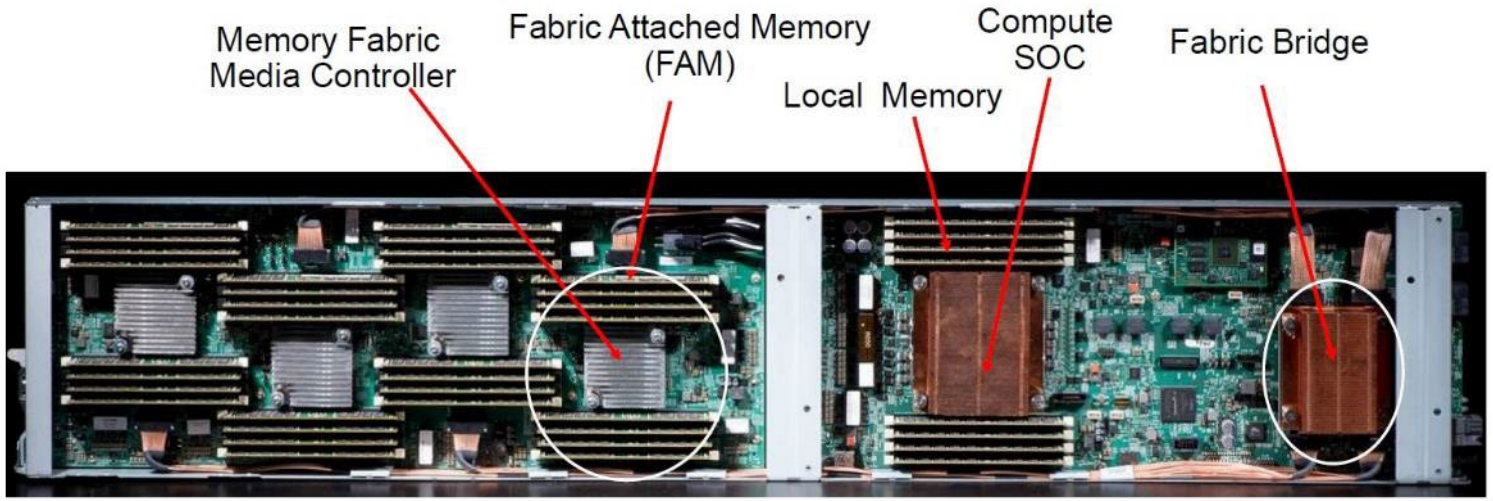


(b) $L = 2, S_1 = 3, S_2 = 3, K = 1, T = 4$

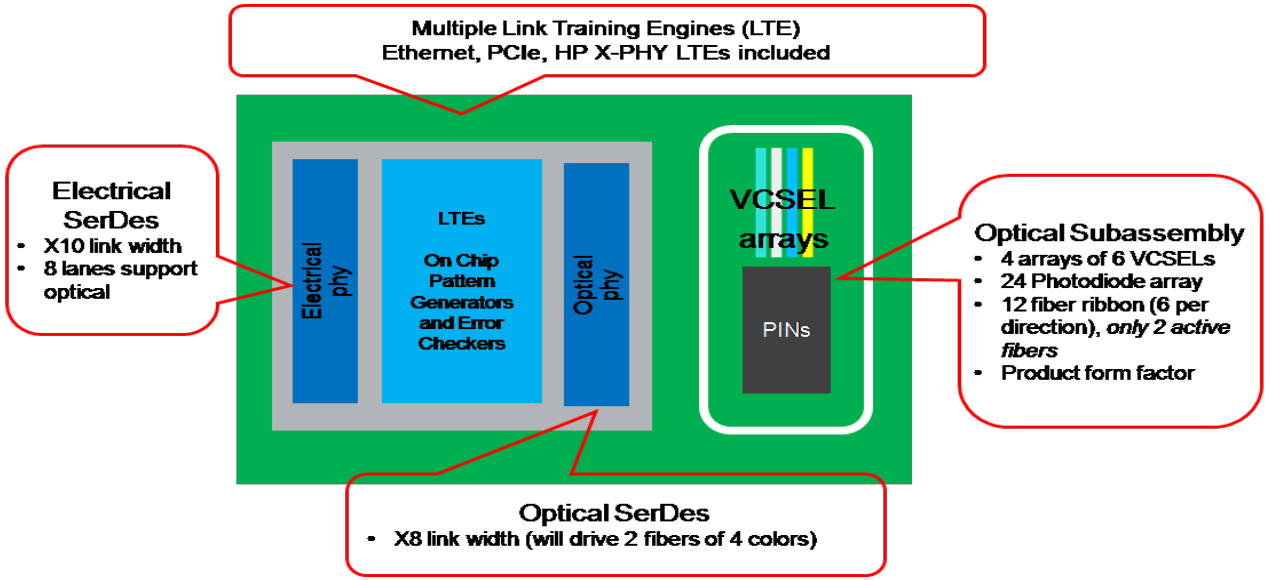
Ahn et al, 2009

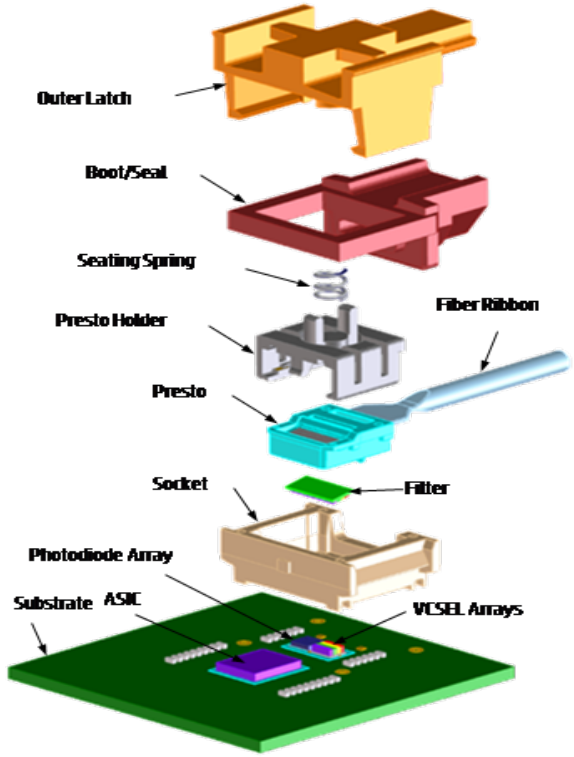
TheMachine Prototype – HP Labs 2016





Another Innovation: HP Labs “Direct Connect” Mid-Board Optics

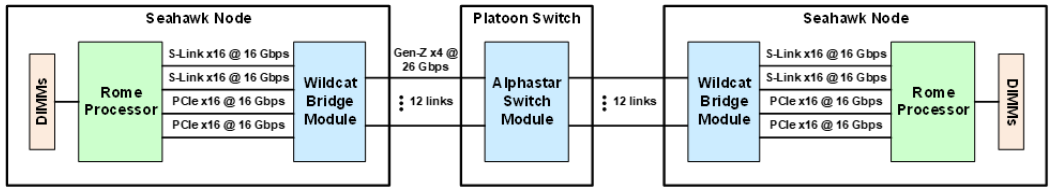


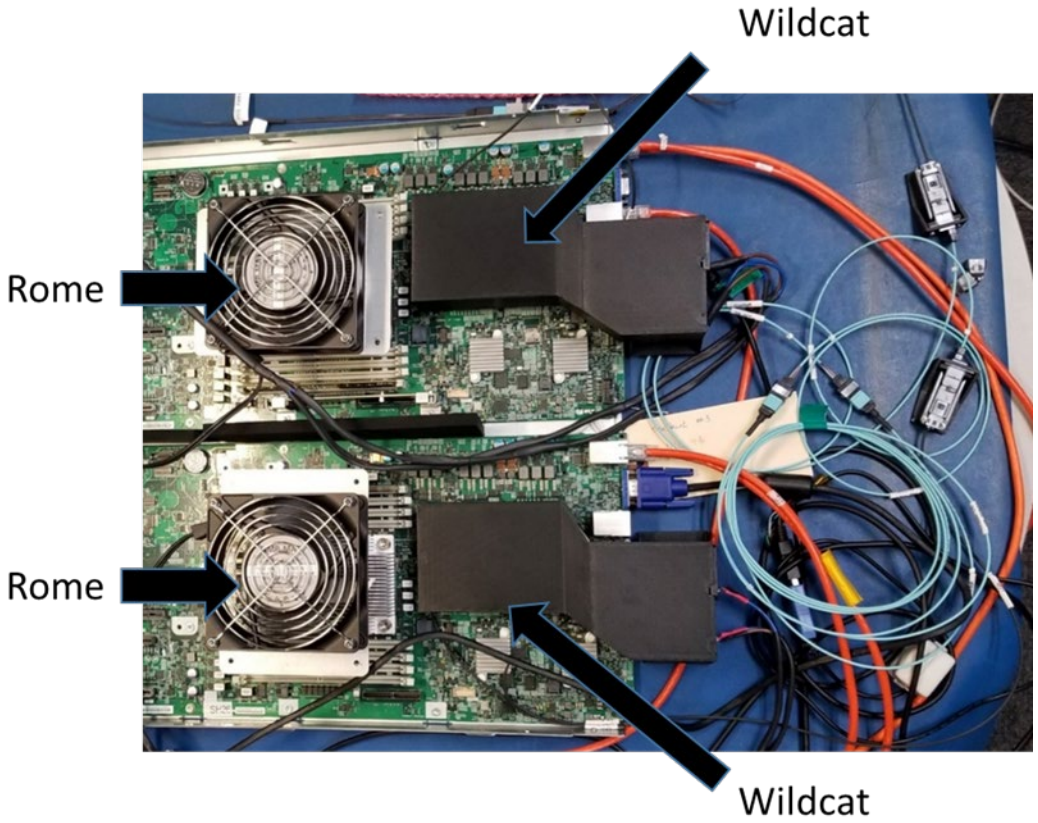


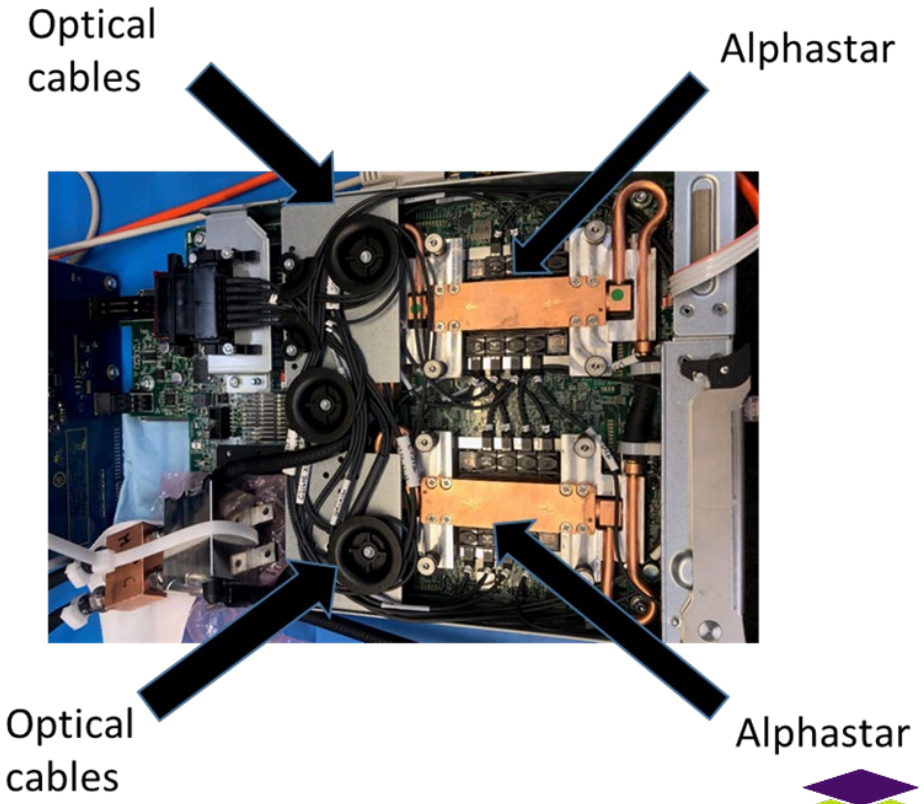
Sayedi, Fiorentino, Beausoleil, et al – HP Labs

DOE PathForward – 2017-2019

- "Seahawk" Compute Node and "Platoon" Optical Switch
- 100% Optical Gen-Z Fabric

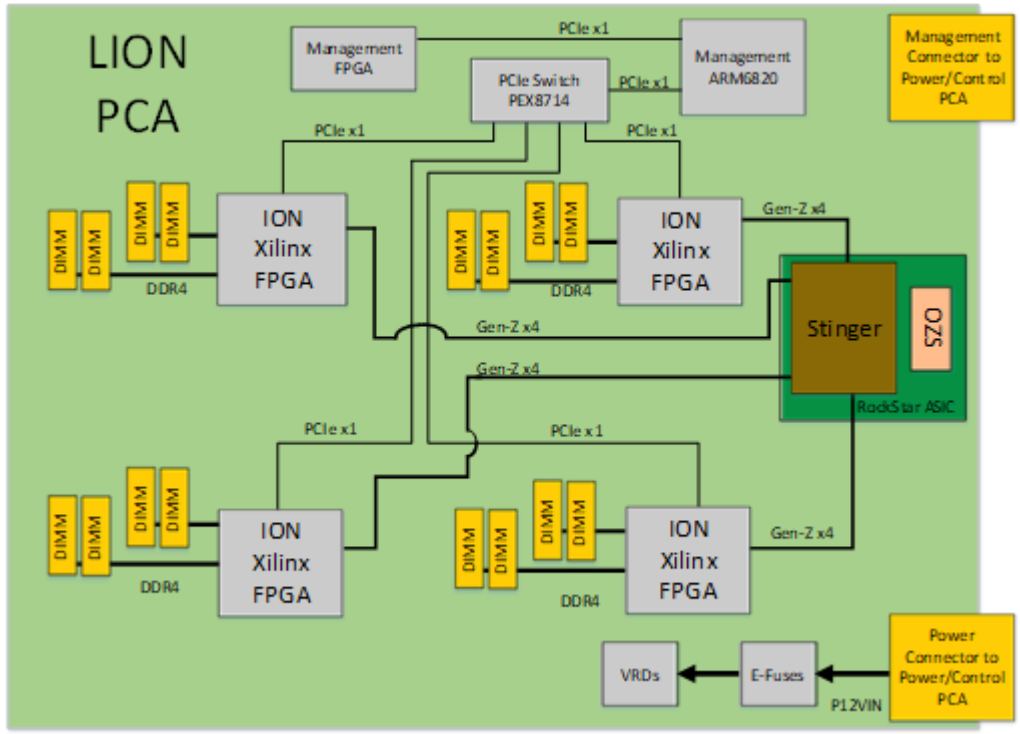


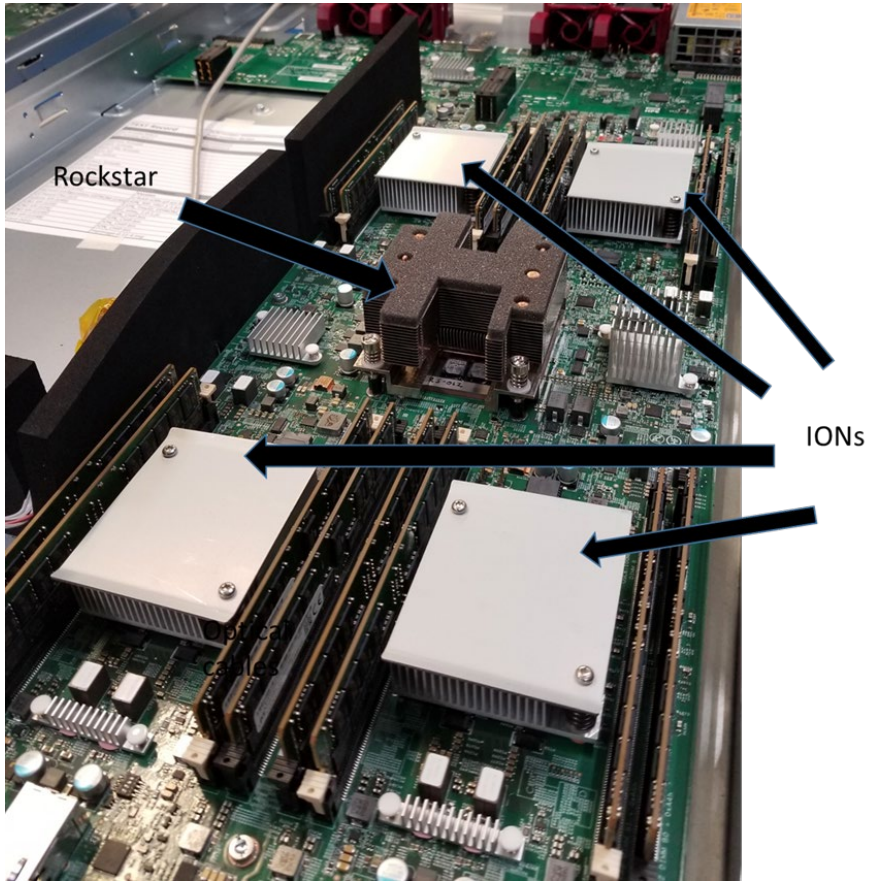




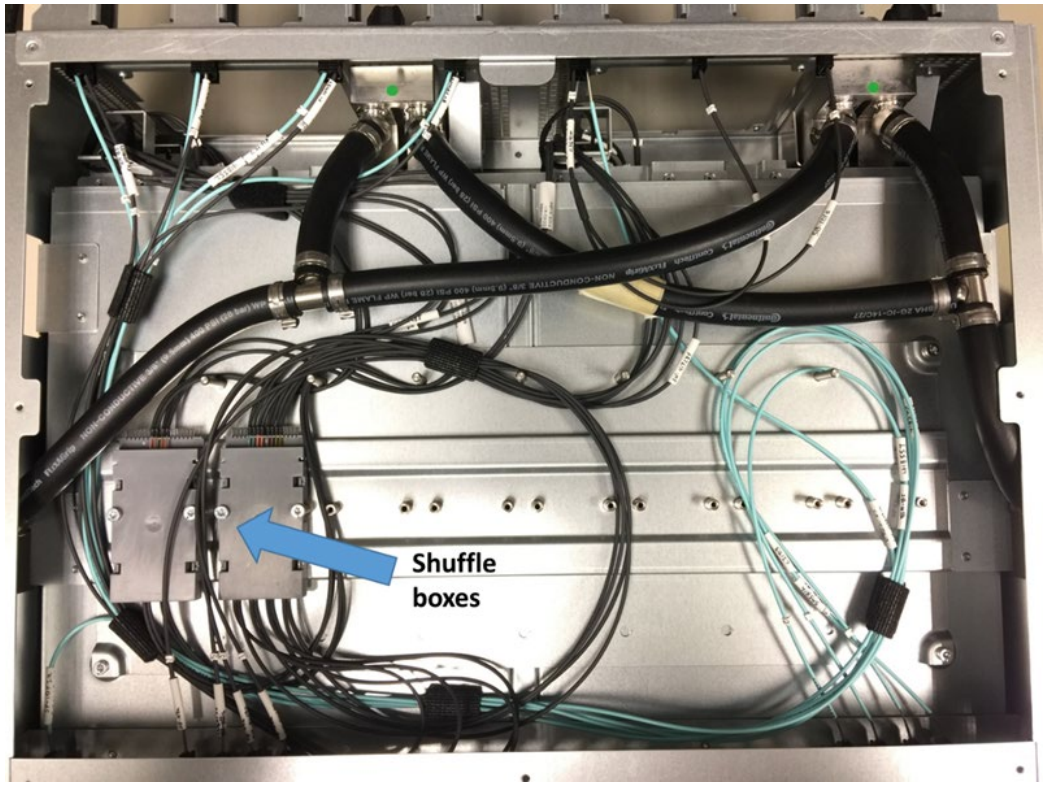
PathForward FAM Module

- "Lion" Optical Gen-Z Attached FAM Module
- ION FPGA memory controller
- "Rockstar" ASIC with "Stinger" optical interface and "OZS" Gen-Z switch optical tile





“Battalion” Liquid-Cooled Enclosure w/Optical Cable Management



Cray Awarded Coral 2 - 2018

THENEXTPLATFORM

HOME | COMPUTE | STORE | CONNECT | CONTROL | CODE | AI | HPC | ENTERPRISE | HYPERSCALE | CLOUD

LATEST > Where Retail Meets The Intelligent Edge, Great Things Are Happening ▸ EDGE

Search ...

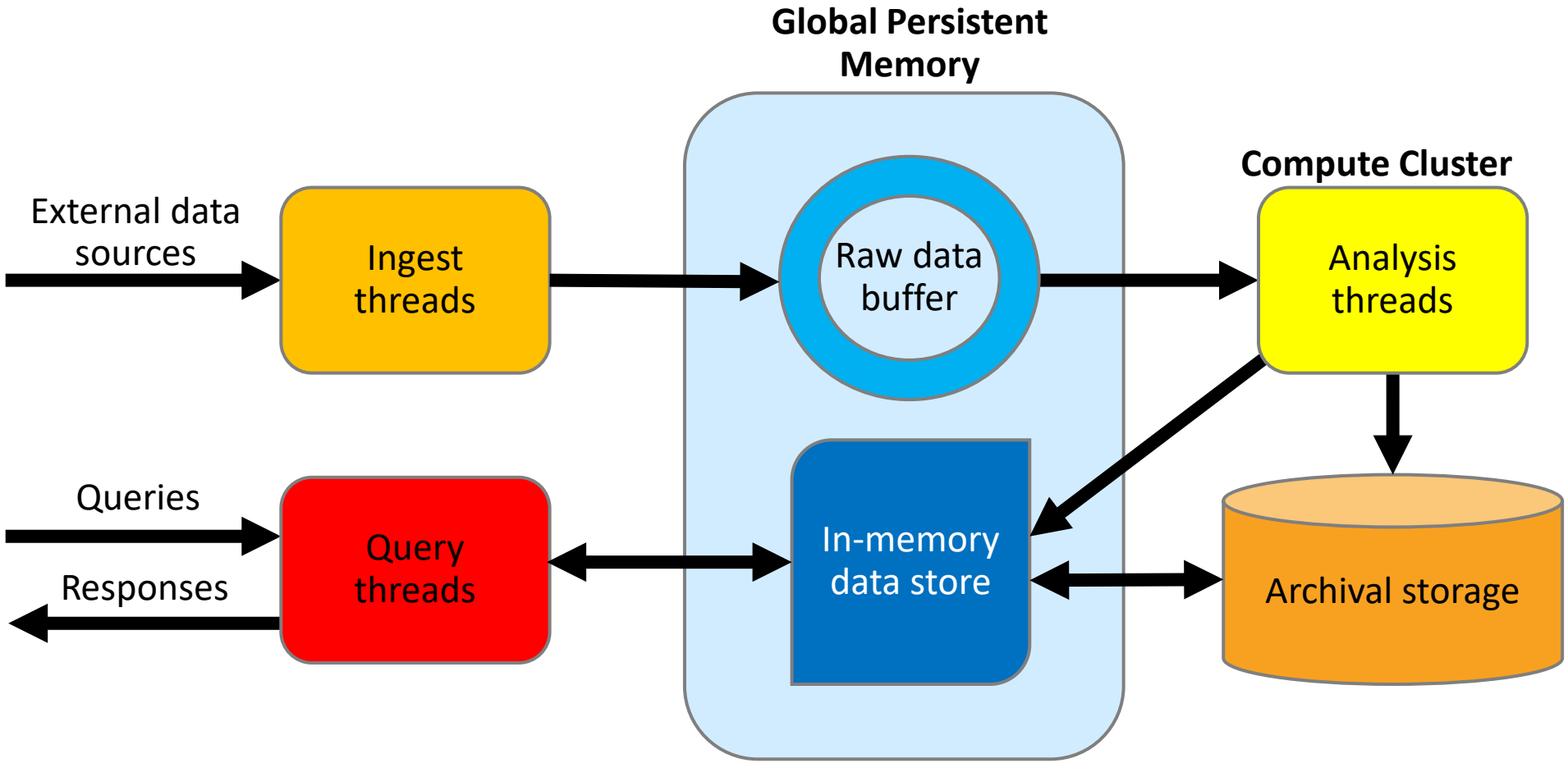
HOME > HPC > Cray Runs The Exascale Table In The United States

CRAY RUNS THE EXASCALE TABLE IN THE UNITED STATES

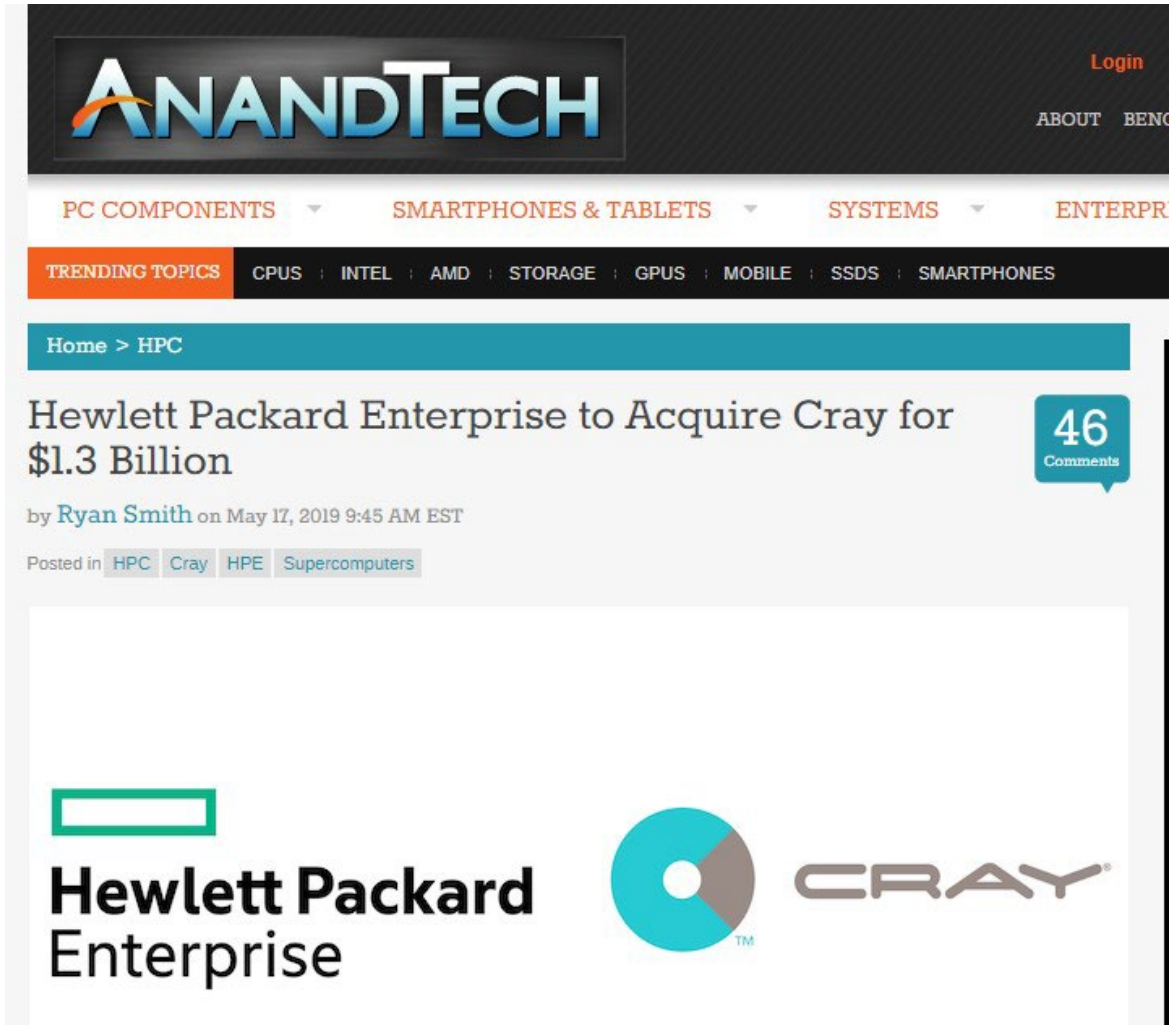
August 13, 2019 Timothy Prickett Morgan

SDC 24

Idealized Workflow for HPC and Data Analysis



HPE Acquires Cray – 2019-2020



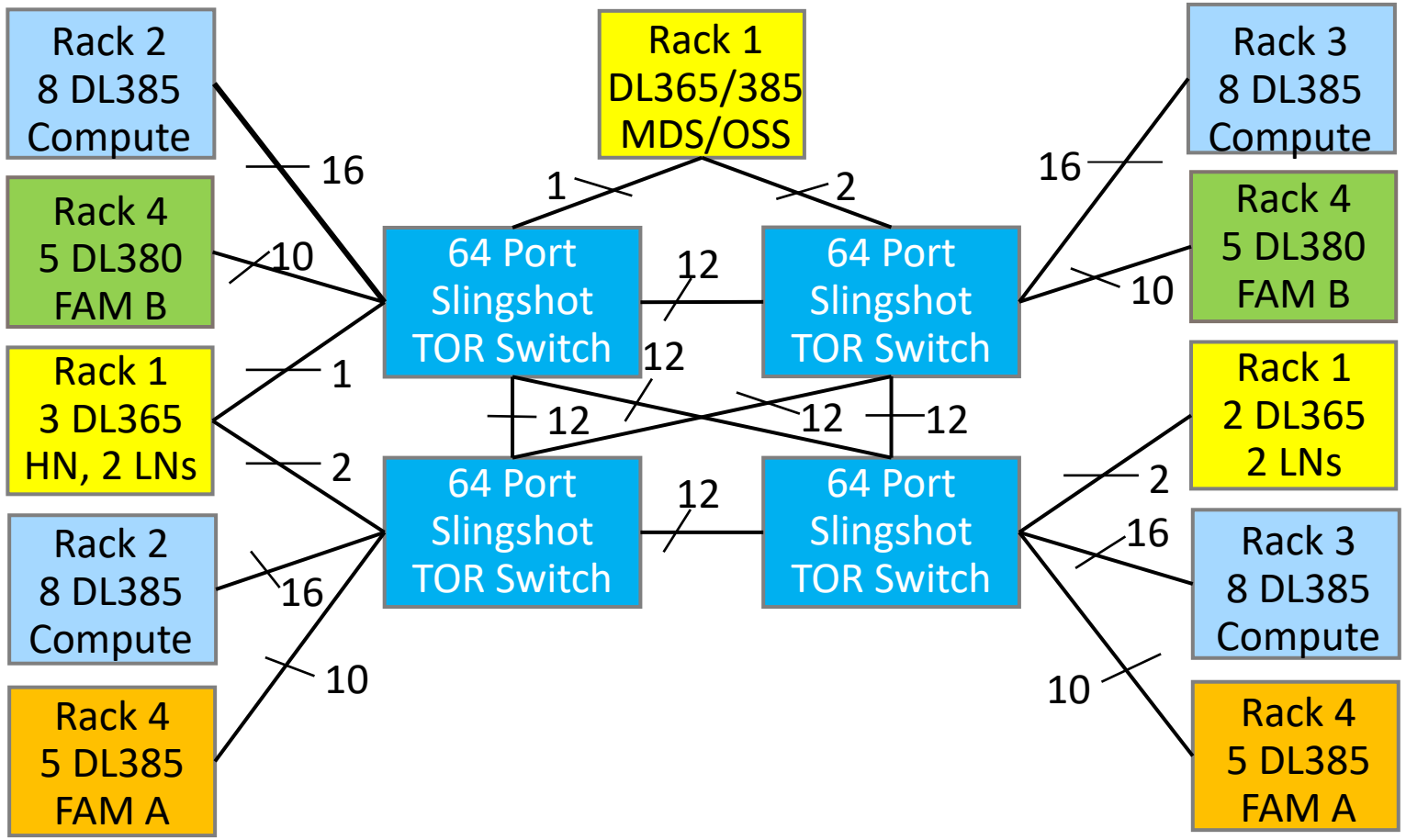
GoldenTicket – 2018-2022

- 32 Compute Nodes
  - DL385 Dual AMD “Milan” CPUs
  - 1024GB memory
  - Dual Slingshot NICs
- FAM Partition A (10 Nodes)
  - DL385 Dual AMD “Milan” CPUs
  - 4TB memory
  - Dual Slingshot NICs
  - Four 6.4TB NVMe SSDs
  - Dolphin PCIe interconnect
- FAM Partition B (10 Nodes)
  - DL380 Dual Intel “Ice Lake”
  - 1TB DRAM
  - Dual Slingshot NICs
  - 8TB “Barlow Pass” SCM
  - Four 6.4TB NVMe SSDs
  - Dolphin PCIe Interconnect

- *84 Epyc 7763, 40 Xeon 8380s*
- *6,176 Total Cores*
- *400 Gbps Slingshot per Node*
- *82 TB Total DRAM*
- *80 TB Optane*

Front view

Prototype Slingshot Network – Combined CN/FAM Groups



GoldenTicket: Converged Data Driven Computing

Prototype & Design Study

GT Notional Prototype
- Running OpenFAM Slingshot memory servers
- Platform for research, benchmarking, and software prototyping

Memory Node Software Stack
- Libfabric, atomics
- Access to FAM from Chapel and OpenSHMEM applications
- Built-in security
- User-defined memory-side operations

Simulation and Benchmarking
- Simulation at scale
- Latency, bandwidth, capacity experiments

Arkouda System Demonstration
- Open source, written in Chapel
- Performance-optimized for FAM
- Interactivity via Jupyter/Python→ NumPy/Arkouda
- Chapel-based development environment

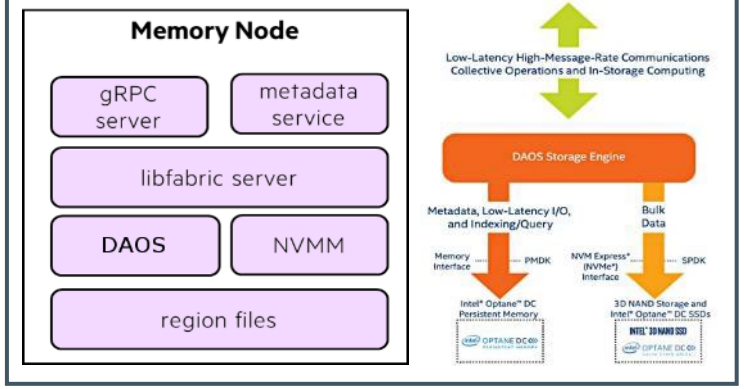
OpenSHMEM/OpenFAM Refinement
- SpMV proxy application
- Sort proxy application

GT Prototype

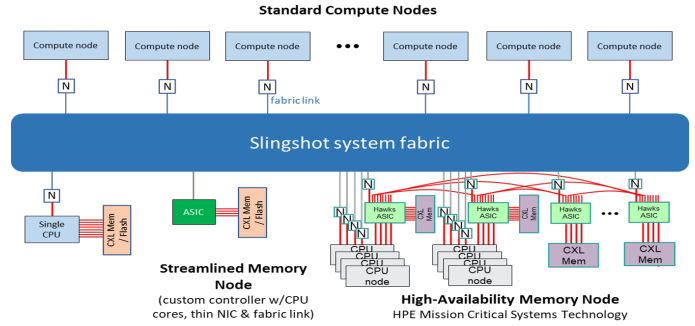
- 32 Compute Nodes
  - DL385 Dual AMD "Milan" CPUs
  - 1024GB memory
  - Dual Slingshot NICs
- FAM Partition A (10 Nodes)
  - DL385 Dual AMD "Milan" CPUs
  - 4TB memory
  - Dual Slingshot NICs
  - Four 6.4TB NVMe SSDs
  - Dolphin PCIe interconnect
- FAM Partition B (10 Nodes)
  - DL380 Dual Intel "Ice Lake"
  - 1TB DRAM
  - Dual Slingshot NICs
  - 8TB "Barlow Pass" SCM
  - Four 6.4TB NVMe SSDs
  - Dolphin PCIe Interconnect

- *84 Epyc 7763, 40 Xeon 8380s*
- *6,176 Total Cores*
- *400 Gbps Slingshot per Node*
- *82 TB Total DRAM*
- *80 TB Optane*

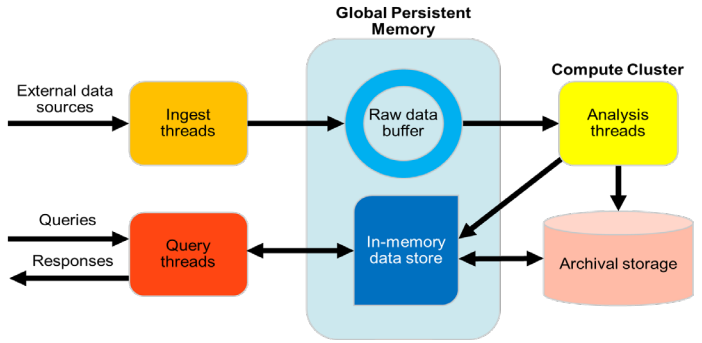
Memory Node Software Stack



GoldenTicket Architecture

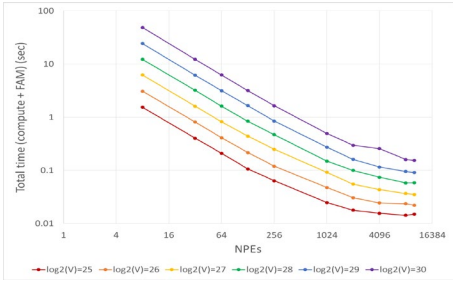


Arkouda Demonstration



Simulation at Scale

- Validation vs. measured prototype performance
- Simulations at scale to 10,000 nodes
- Slingshot Dragonfly fabric



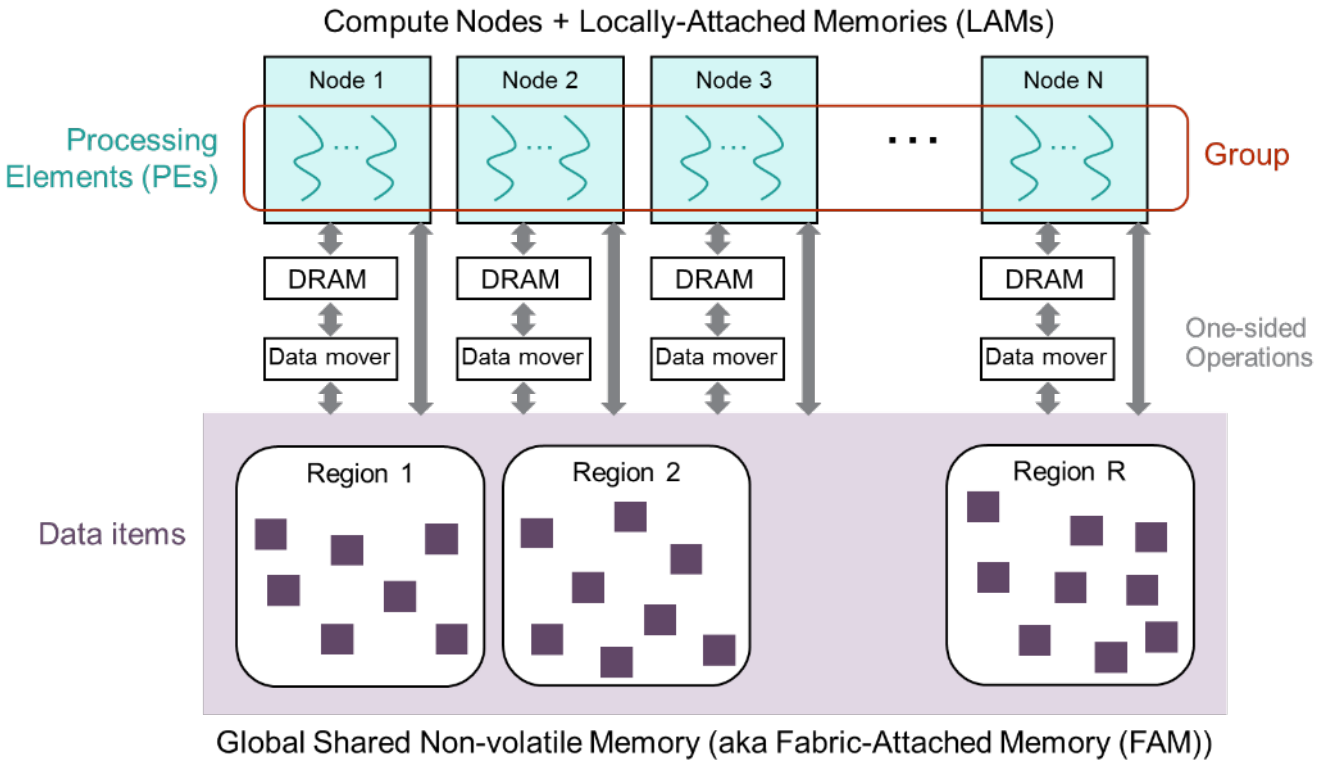
GoldenTicket Prototype – Spring, TX

GoldenTicket Prototype – Spring, TX

- This was the first machine I ever built over the telephone.
- These are the guys who actually did the work! Thanks to:
  - Daniel Moore – Hardware Technician (left)
  - Binoy Arnold – HP Labs IT (right)
- I never got to meet them or see the machine until February, 2024—two years after the machine was brought up.
- GT is now a major HP Labs resource
  - Is being upgraded with GPU cards for AI research
  - Is being used for DAOS development, and CXL prototyping

OpenFAM API – 2018

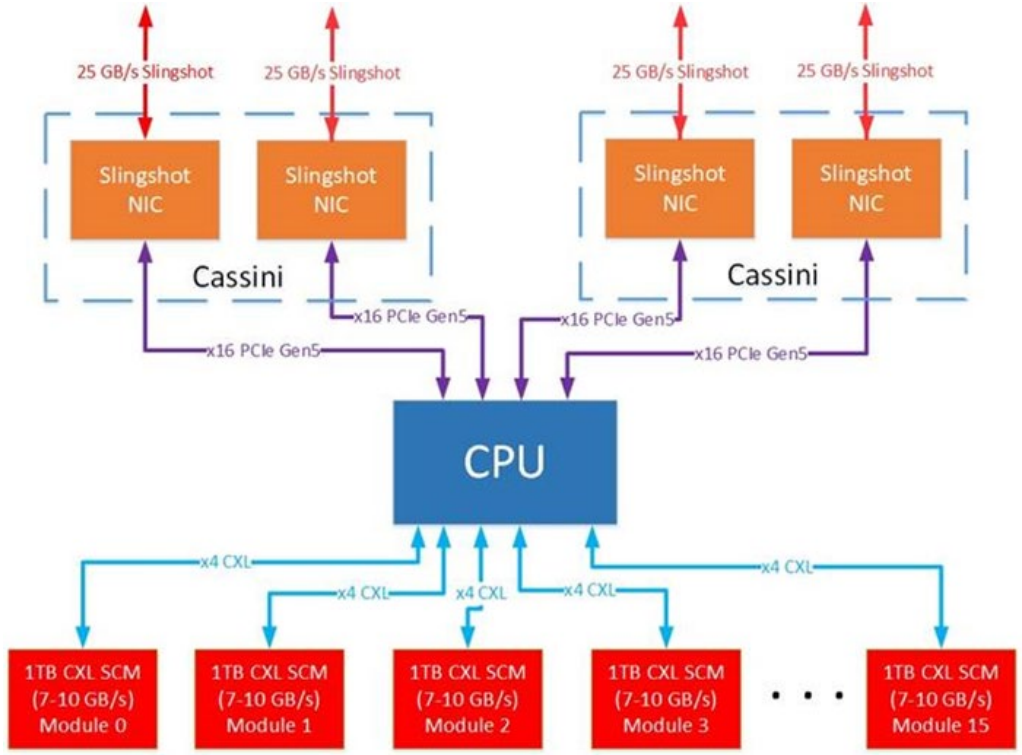
- Open source API to manage and access FAM
- https://github.com/OpenFAM



CXL, finally something the CPU guys can agree on! – 2019

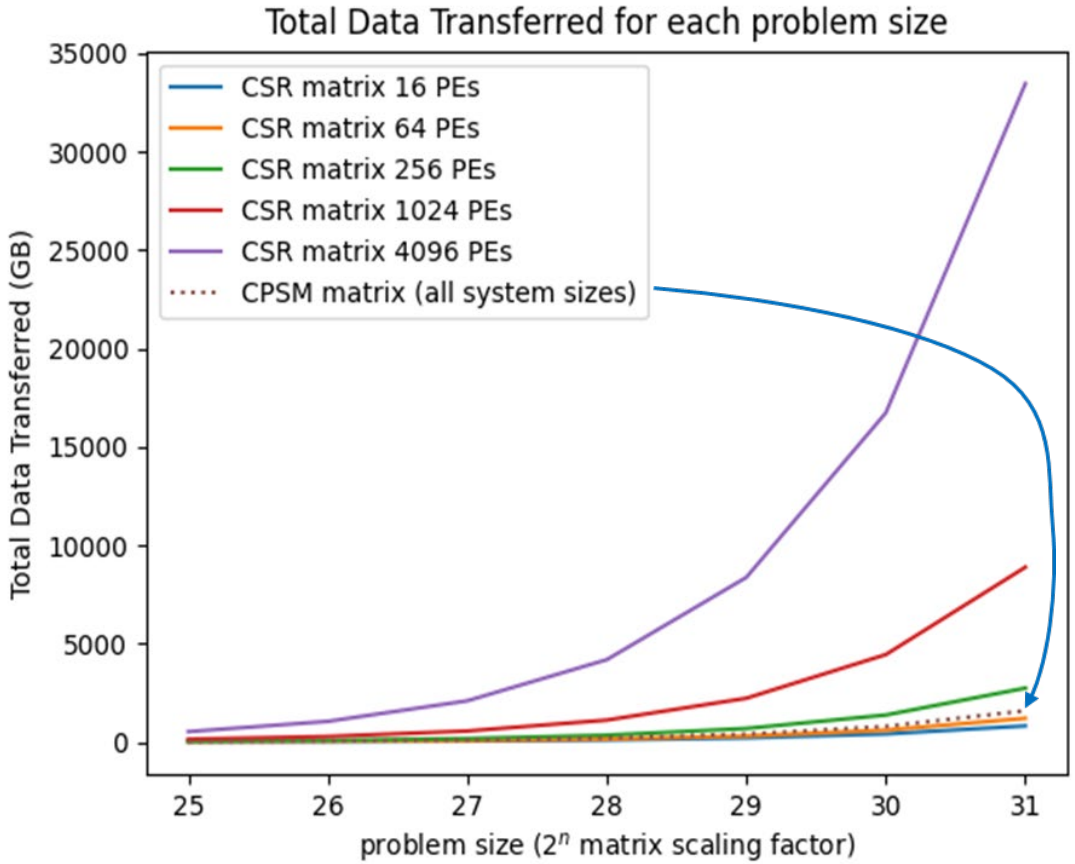
GoldenTicket Notional System – 2020

- Granite Rapids Node
  - 100 GB/s injection bandwidth per node
  - 16 1TB Donahue Pass CXL Optane™ modules
  - The node that could have been…



SpMV – Sparse Matrix-Vector multiply Y=Ax

- Sparsity = ⟨number of non-zero values⟩ / ⟨size of matrix⟩
- CSR – Compressed Sparse Row
  - Tuples { value, column } and count of non-zero values per row
  - Sequential rows of CSR matrix possibly intersect any elements of multiplication vector – essentially random
  - Each PE preemptively accesses entire vector
  - Sparsity challenges reuse of multiplication vector cache lines
  - Similar variants (CSC, C-CSR and C-CSC) have same issue
- CPSM – Column Partitioned Sparse Matrix
  - Sets of tuples {row, column, value} span non-zero elements for all rows within partitions of columns
  - Each set spans a number of columns with approximately same number of non-zero values intersecting a range of multiplication vector similar in capacity to processor core local cache
  - Vector partitions only transferred to one PE
- For small system sizes, CSR's redundant transfer of multiplication vector
  does not exceed CPSM's explicit inclusion of row per element



Some of the Lessons Learned

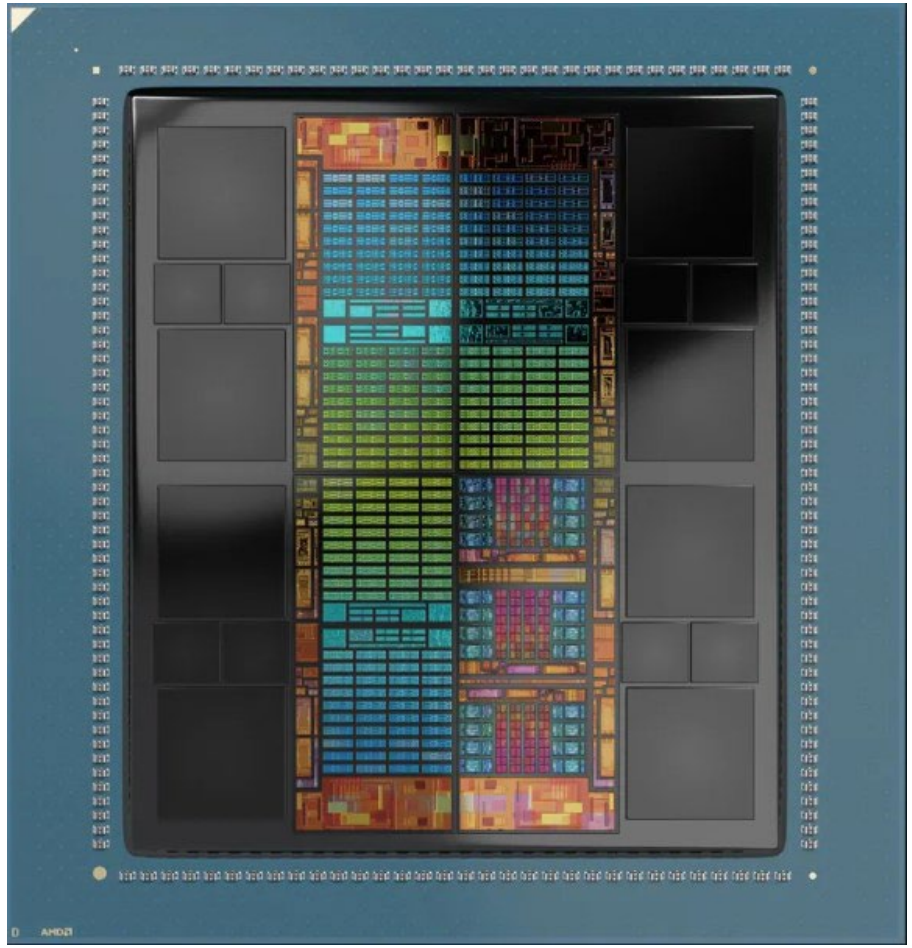
- Lack of memory semantics in Slingshot, Ethernet, and InfiniBand means that references to FAM must be converted into messages using RDMA
  - FAM latency means that most computing must be done in compute node memory anyway
  - Atomics, small messages, rarely referenced data accesses suffer
- FAM is not great memory, but it is really fast storage!
  - What we really want is higher capacity and low cost-per-bit
- Design of data structures stored in FAM is critical!
  - HPC is lots and lots of linear algebra
  - Higher performance requires more memory for compressed sparse matrices

SDC 24

CXL in HPC

Case Study: MI300A Compute Node in El Capitan

- 8 stacks HBM3
- Peak memory bandwidth 5.3TB/s
- 4x16 Infinity™ scale up links
- 4x16 PCIe Gen5/CX L2.0 or Infinity scale out links
  - 2 x16 PCIe/CXL 2.0 links available in Cray EX
  - One x16 PCIe link used for Slingshot NIC
  - Only one x16 or two x8 CXL 2.0 links available → 64 GB/s
  - 64/5300 = 1.2% of HBM memory bandwidth!



CXL DRAM Modules are not Great for HPC

- CXL does not significantly improve memory bandwidth for leadership class HPC nodes (or anything with HBM or lots of LPDDR channels)
- Next generation of leadership class HPC nodes will not have significant memory capacity limitations
  - Yes, everyone can always use more memory, but the minimum configuration for a device with 16 DDR5 channels is 512GB.
  - HBM stacks will be 64GB each!
  - Memory is (normally) a huge cost component of a large HPC system.
- CXL link bandwidth will improve to 64GT/s with PCIe Gen6, but that still is not enough
  - Bleeding edge SerDes will run at 224Gbps in the Gen6 timeframe!
  - A single HBM4/4e device will run deliver 2-4 TB/s!

What CXL Modules Might Help HPC?

- First tier storage case
  - DRAM cost-per-bit is very high compared to Flash (and the CXL controller is not free)
  - HPC storage systems would benefit from large-capacity, non-volatile, byte-addressable memory
  - Capacity has to be O(8TB) per E3.S module at the price of a DRAM module
  - 100ns access time, acceptable write wear (need 5-year lifetime)
  - There are promising technologies out there, but equipment for process development is extremely expensive
- Compute accelerator and memory
  - Potentially high bandwidth per core for CXL module
  - Power is a problem (E3.S spec is ~40W, could do better with liquid cooling)
  - Still have the low-bandwidth CXL interface, but CXL 3.1 enables scalable fabrics. Max number of endpoints is 4,096.
- Unconventional accelerators
  - Memristor-based neuromorphic accelerators
  - Inference engines requiring limited I/O bandwidth
  - <your idea here>

CXL Has Competition: UAL

HPC wire

Since 1987 - Covering the Fastest Computers In the World and the People Who Run Them

- Home
- Topics
- Sectors
- Exascale
- Specials
- Resource Library
- Podcast
- Events
- Job Bank
- About
- Subscribe

Everyone Except Nvidia Forms Ultra Accelerator Link (UALink) Consortium

By Doug Eadline

May 30, 2024

Consider the GPU. An island of SIMD greatness that makes light work of matrix math. Originally designed to rapidly paint dots on a computer monitor, it was then found to be quite useful in large numbers by HPC practitioners. Enter GenAI, and now these little matrix mavens are in huge demand, so much so that we call it the GPU Squeeze.

The well-known and dominant market leader, Nvidia, has charted much of the pathway for GPU technology. For HPC, GenAI, and a raft of other applications, connecting GPUs provides a way to solve bigger problems and improve your application's performance

THENEXTPLATFORM

HOME | COMPUTE | STORE | CONNECT | CONTROL | CODE | AI | HPC | ENTERPRISE | HYPERSCALE | CLOUD

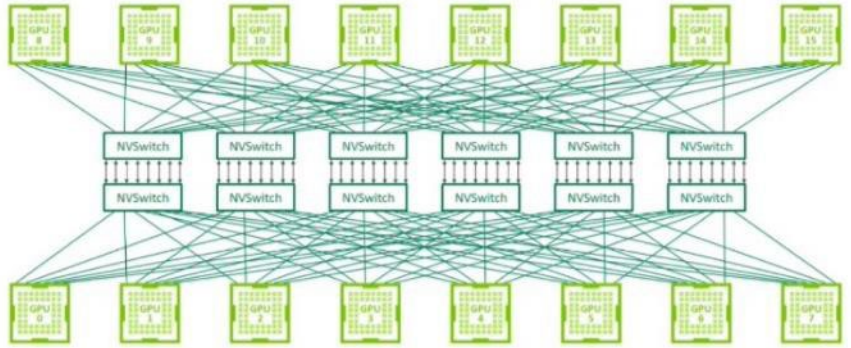
LATEST > NOAA Gets $100 Million Windfall For "Rhea" Research Supercomputer ▸ HPC

Search ...

HOME > CONNECT > Key Hyperscalers And Chip Makers Gang Up On Nvidia's NVSwitch Interconnect

KEY HYPERSCALERS AND CHIP MAKERS GANG UP ON NVIDIA'S NVSWITCH INTERCONNECT

May 30, 2024 Timothy Prickett Morgan



The generative AI revolution is making strange bedfellows, as revolutions and emerging monopolies that

SDC 24

What About Networking?: UltraEthernet

HPC wire

Since 1987 - Covering the Fastest Computers in the World and the People Who Run Them

- Home
- Topics
- Sectors
- Exascale
- Specials
- Resource Library
- Podcast

Industry-Leading Consortium to Redefine Ethernet Performance for AI and HPC

July 19, 2023

SAN FRANCISCO, July 19, 2023 — Announced today, Ultra Ethernet Consortium (UEC) is bringing together leading companies for industry-wide cooperation to build a complete Ethernet-based communication stack architecture for high-performance networking.

Artificial Intelligence (AI) and High-Performance Computing (HPC) workloads are rapidly evolving and require best-in-class functionality, performance, interoperability and total cost of ownership, without sacrificing

Ultra Ethernet Consortium

THENEXTPLATFORM

HOME | COMPUTE | STORE | CONNECT | CONTROL | CODE | AI | HPC | ENTERPRISE | HYPERSCALE | CLOUD | EDGE

LATEST > Green Acres Is The Place For Larry > AI

Search ...

HOME > CONNECT > Ethernet Consortium Shoots For 1 Million Node Clusters That Beat InfiniBand

ETHERNET CONSORTIUM SHOOTS FOR 1 MILLION NODE CLUSTERS THAT BEAT INFINIBAND

July 20, 2023 Timothy Prickett Morgan

SDC 24

Thank you!

Please take a moment to rate this session.

Your feedback is important to us.

