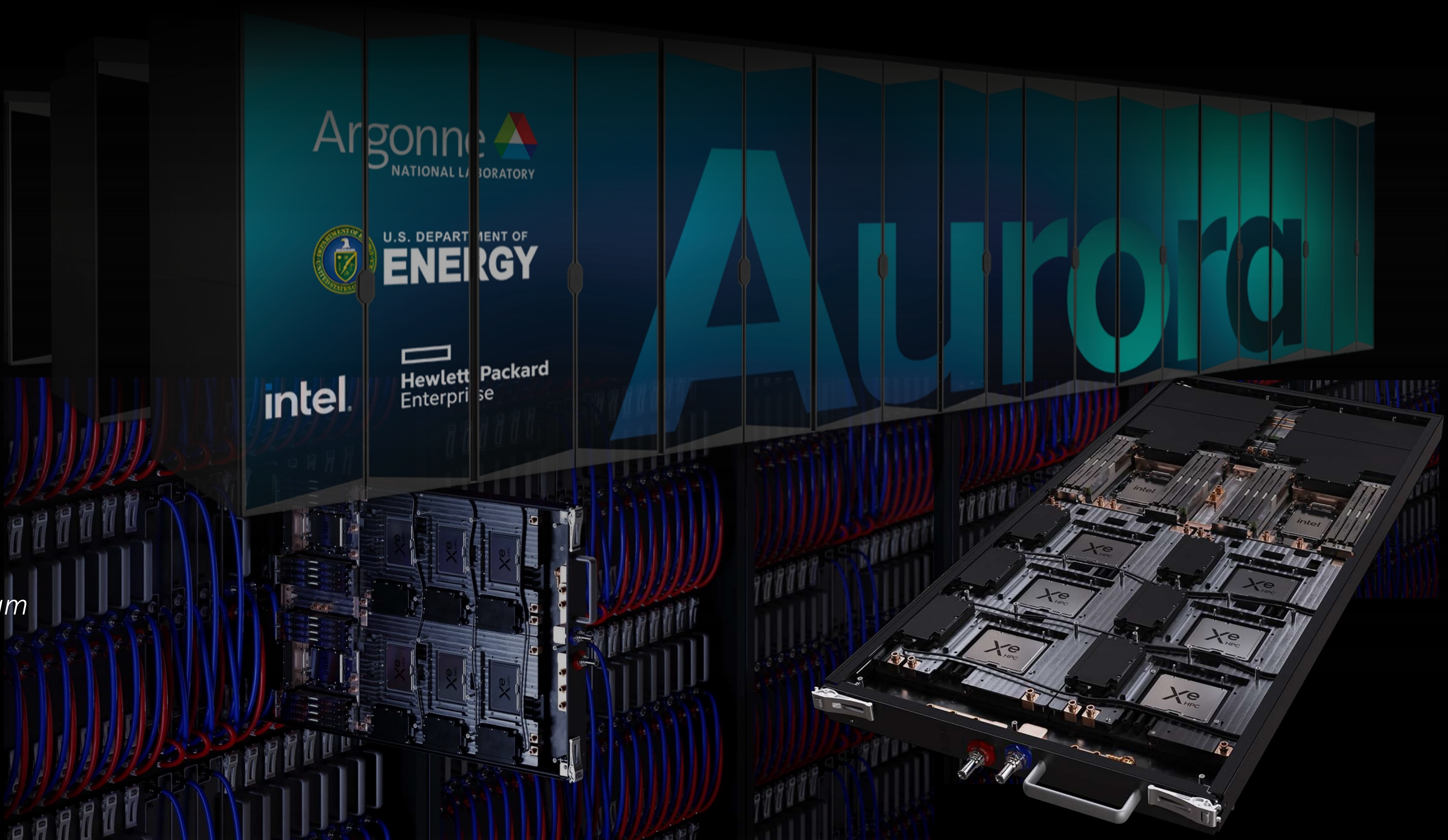
Aurora Exascale Architecture


Argonne NATIONAL LABORATORY
U.S. DEPARTMENT OF ENERGY
intel
Hewlett Packard Enterprise
Aurora
Xe HPC

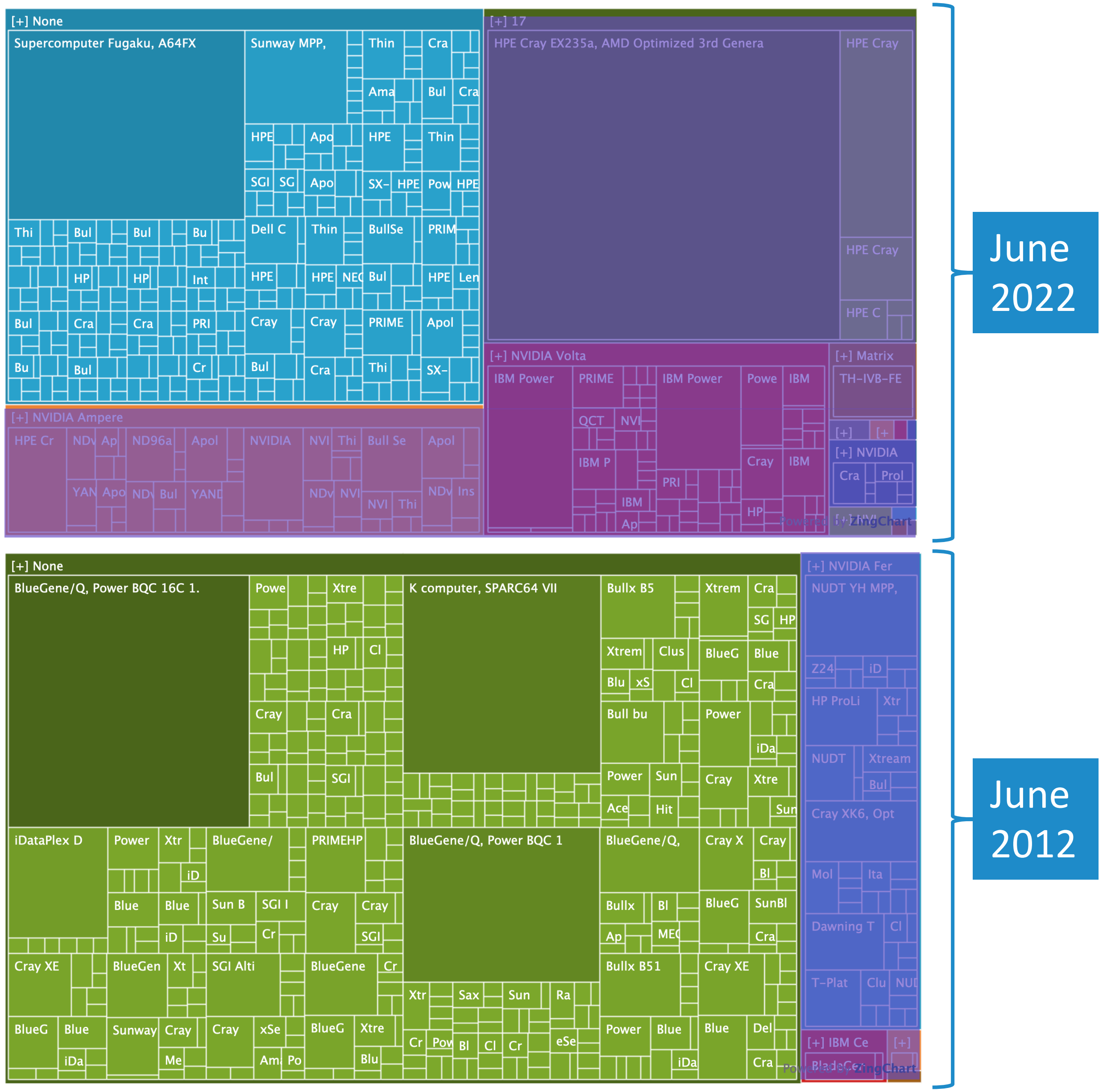
Servesh Muralidharan

Computer Scientist, Performance Engineering Team
Argonne Leadership Computing Facility

PATH TO EXASCALE

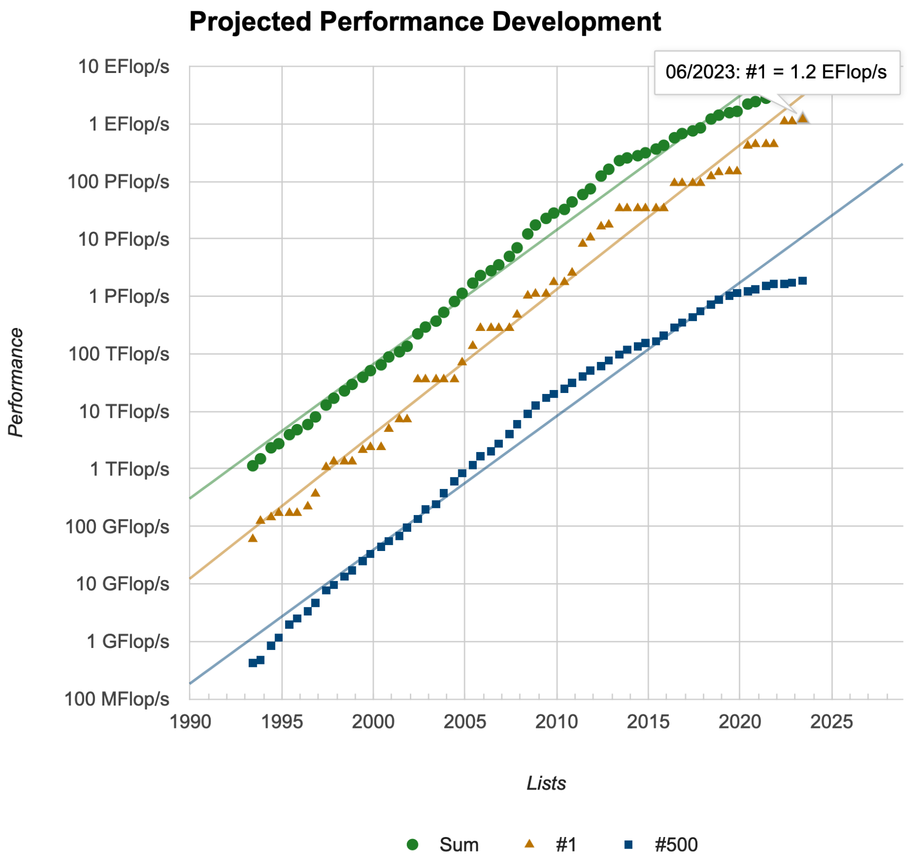
Elements of a supercomputer

- Processor
  - architecturally optimized to balance complexity, cost, performance, and power
- Memory
  - generally commodity DDR, amount limited by cost
- Node
  - may contain multiple processors, memory, and network interface
- Network
  - optimized for latency, bandwidth, and cost
- IO System
  - complex array of disks, servers, and network
- Software Stack
  - compilers, libraries, tools, debuggers, …
- Control System
  - job launcher, system management



https://www.top500.org/statistics/treemaps/

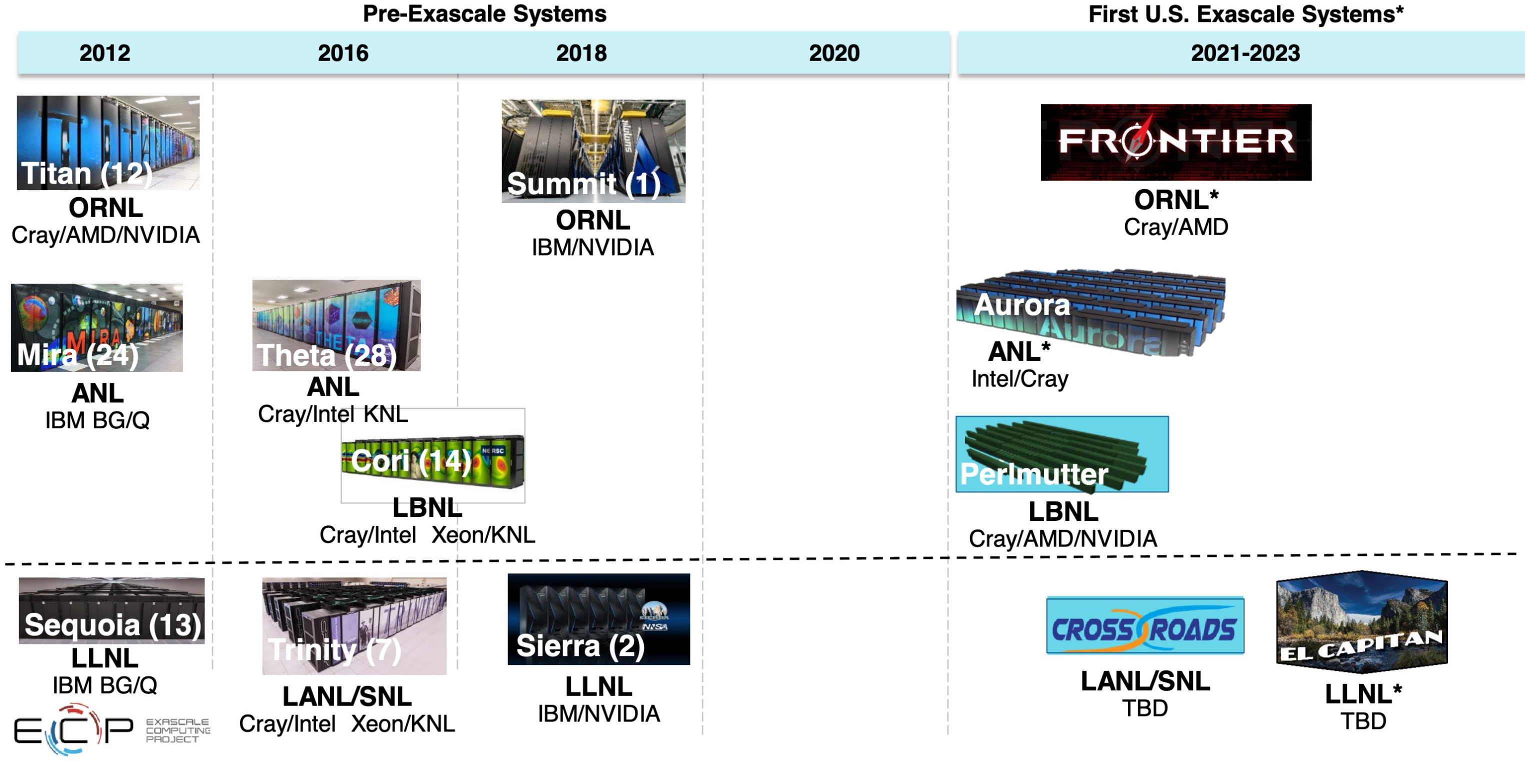
Exascale Computing Project



https://www.top500.org/statistics/perfdevel/

Department of Energy (DOE) Roadmap to Exascale Systems

An impressive, productive lineup of *accelerated node* systems supporting DOE's mission



Pre-Exascale Systems

First U.S. Exascale Systems*

2012

2016

2018

2020

2021-2023

Titan (12)
ORNL
Cray/AMD/NVIDIA

Summit (1)
ORNL
IBM/NVIDIA

FRONTIER
ORNL*
Cray/AMD

Mira (24)
ANL
IBM BG/Q

Theta (28)
ANL
Cray/Intel KNL

Aurora
ANL*
Intel/Cray

Cori (14)
LBNL
Cray/Intel Xeon/KNL

Perlmutter
LBNL
Cray/AMD/NVIDIA

Sequoia (13)
LLNL
IBM BG/Q

Trinity (7)
LANL/SNL
Cray/Intel Xeon/KNL

Sierra (2)
LLNL
IBM/NVIDIA

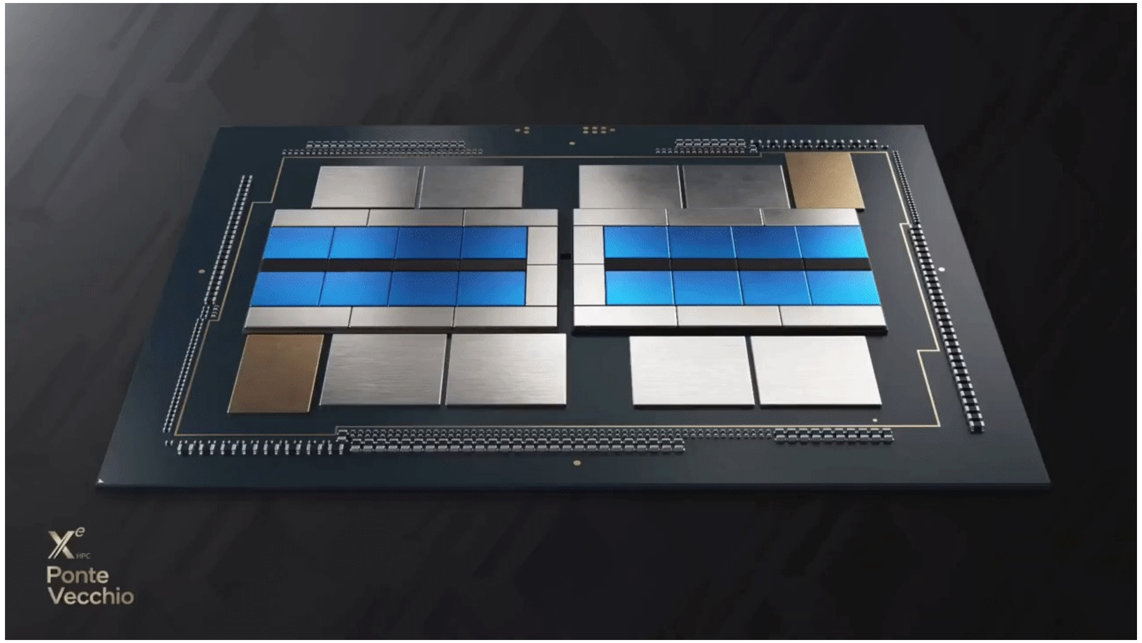
CROSSROADS
LANL/SNL
TBD

EL CAPITAN
LLNL*
TBD

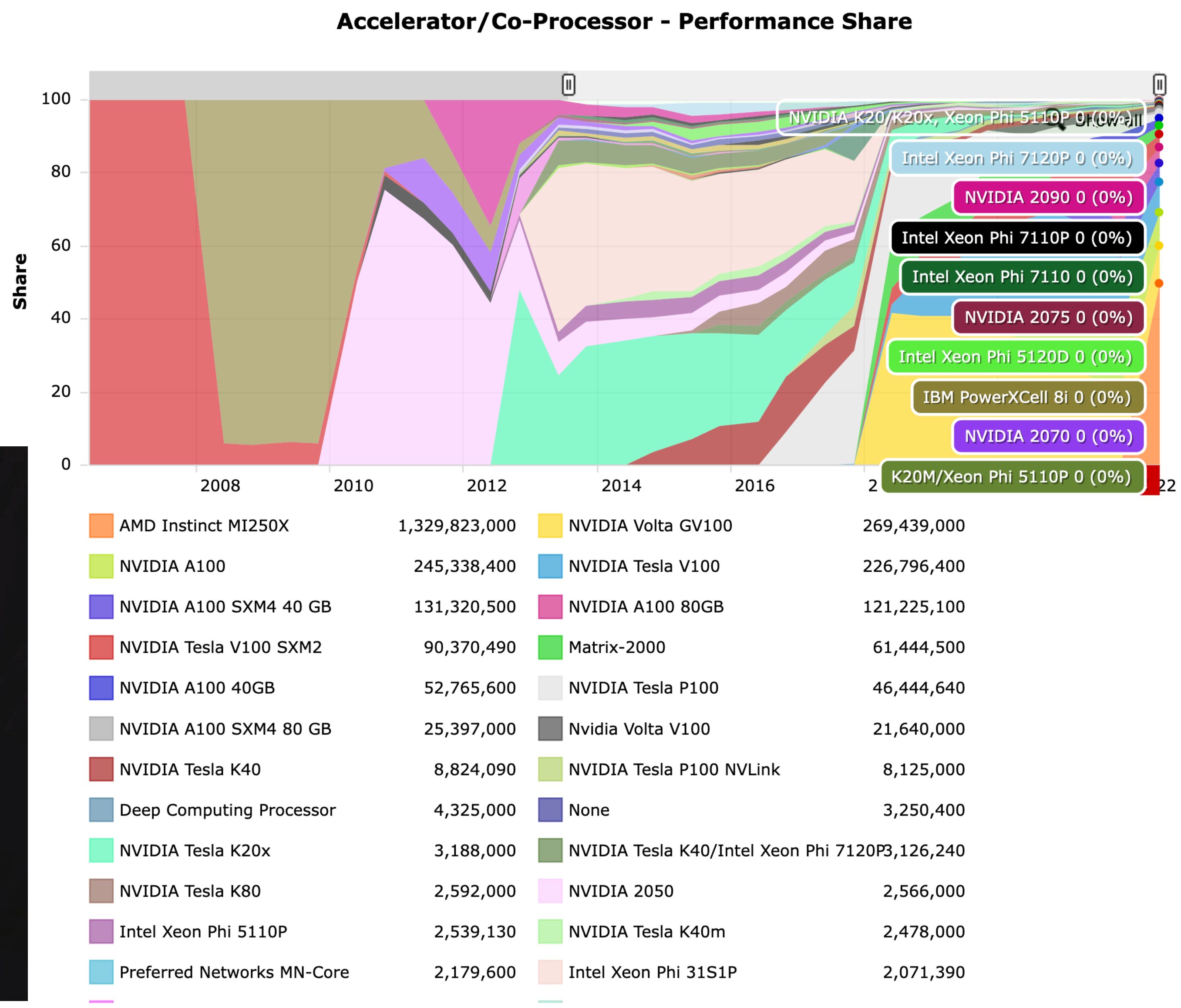
https://science.osti.gov/-/media/bes/besac/pdf/201907/1330_Diachin_ECP_Overview_BESAC_201907.pdf

Path to Exascale Computing

- Era of data parallel computing
  - —Dominated by GPUs
  - —Exploit SIMT/SIMD Parallelism
- Architectural Challenges
  - —Multichip Packaging
  - —Next generation technologies



Intel's HPC GM Trish Damkroger Keynote ISC 2021
https://www.youtube.com/watch?v=PuEcCRJLrvs
https://download.intel.com/newsroom/2021/data-center/Intel-ISC2021-keynote-presentation.pdf

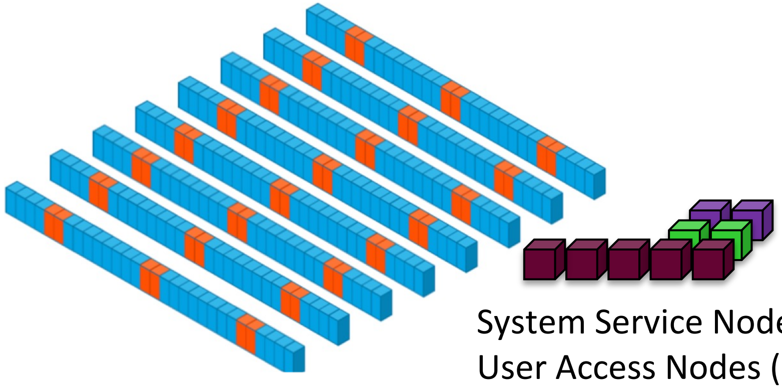


AMD Instinct MI250X	1,329,823,000	NVIDIA Volta GV100	269,439,000
NVIDIA A100	245,338,400	NVIDIA Tesla V100	226,796,400
NVIDIA A100 SXM4 40 GB	131,320,500	NVIDIA A100 80GB	121,225,100
NVIDIA Tesla V100 SXM2	90,370,490	Matrix-2000	61,444,500
NVIDIA A100 40GB	52,765,600	NVIDIA Tesla P100	46,444,640
NVIDIA A100 SXM4 80 GB	25,397,000	Nvidia Volta V100	21,640,000
NVIDIA Tesla K40	8,824,090	NVIDIA Tesla P100 NVLink	8,125,000
Deep Computing Processor	4,325,000	None	3,250,400
NVIDIA Tesla K20x	3,188,000	NVIDIA Tesla K40/Intel Xeon Phi 7120P	3,126,240
NVIDIA Tesla K80	2,592,000	NVIDIA 2050	2,566,000
Intel Xeon Phi 5110P	2,539,130	NVIDIA Tesla K40m	2,478,000
Preferred Networks MN-Core	2,179,600	Intel Xeon Phi 31S1P	2,071,390

https://www.top500.org/statistics/overtime/

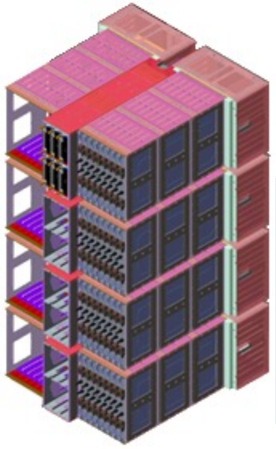
AURORA: HARDWARE

Aurora High-level System Overview

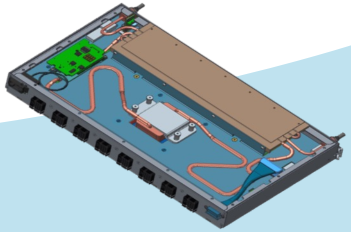


System Service Nodes (SSNs)
User Access Nodes (UANs)
DAOS Nodes (DNs)
Gateway Nodes (GNs)
IOF service, scalable library loading
DAOS <-> Lustre data mover

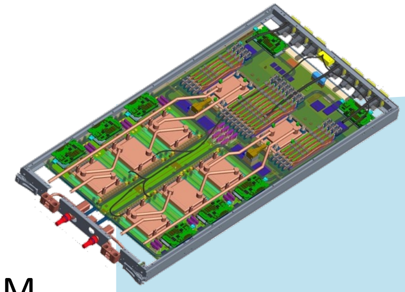
AURORA SYSTEM
166 Compute racks
10,624 Nodes
GPU: 8.16 PB HBM
CPU: 1.36 PB HBM, 10.9 PB DDR5
DAOS: 64 racks, 1024 nodes
230 PB (usable), 31 TB/s



COMPUTE RACK
64 Compute blades
32 Switch blades
GPU: 49.1 TB HBM
CPU: 8.2 TB HBM, 64 TB DDR5



SWITCH BLADE
1 Slingshot switch
64 ports
Dragonfly topology

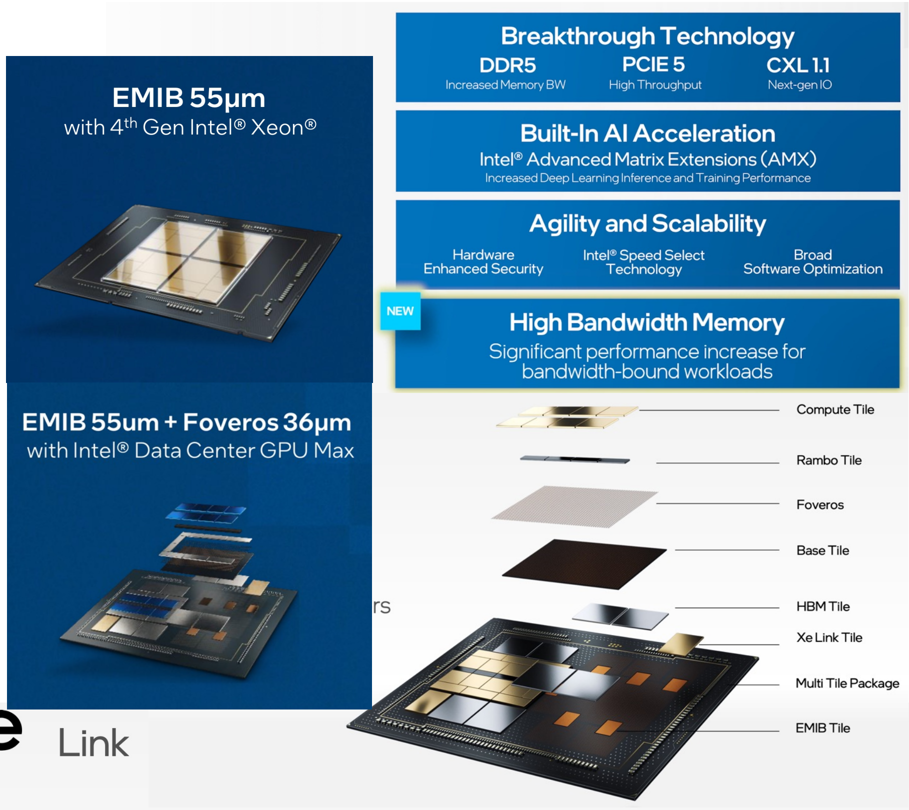


COMPUTE BLADE
2x Intel Xeon Max Series w HBM
6x Intel Data Center GPU Max Series
GPU: 768 GB HBM
CPU: 128 GB HBM, 1024 GB DDR5

Aurora Exascale Compute Blade - Components

- Intel Xeon Max Series CPU w HBM
  - DDR5 and HBM
  - PCIe Gen5
- Intel Data Center Max Series GPU
  - Multi Tile architecture
    - Compute Tile
      - Xe Cores
      - L1 Cache
    - Base Tile
      - PCIe Gen5
      - HBM2e Main Memory
      - MDFI
      - EMIB
- GPU – GPU Interconnect
  - Xe Link

https://download.intel.com/newsroom/2021/data-center/Intel-ISC2021-keynote-presentation.pdf



EMIB 55µm
with 4th Gen Intel® Xeon®

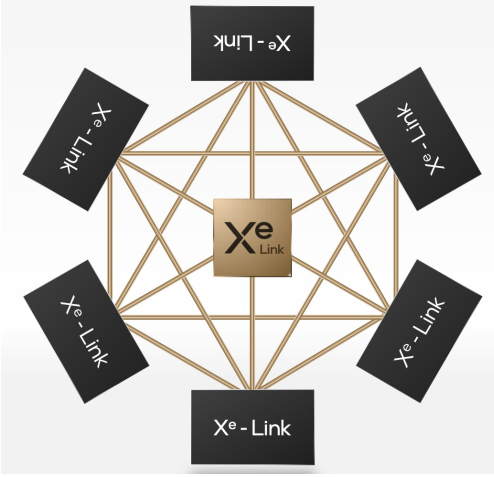
Breakthrough Technology
DDR5 Increased Memory BW
PCIE 5 High Throughput
CXL 1.1 Next-gen IO

Built-In AI Acceleration
Intel® Advanced Matrix Extensions (AMX)
Increased Deep Learning Inference and Training Performance

Agility and Scalability
Hardware Enhanced Security
Intel® Speed Select Technology
Broad Software Optimization

NEW
High Bandwidth Memory
Significant performance increase for bandwidth-bound workloads

EMIB 55um + Foveros 36µm
with Intel® Data Center GPU Max



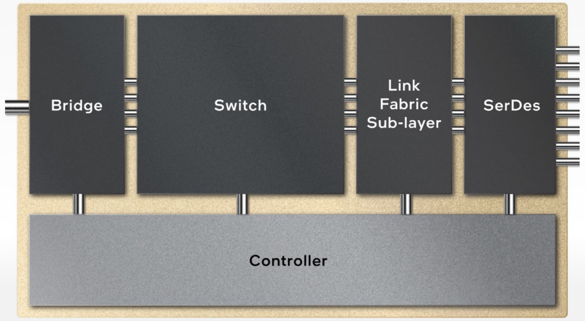
Xe Link

https://www.intel.com/content/www/us/en/newsroom/resources/press-kit-architecture-day-2021.html

High Speed Coherent Unified Fabric (GPU to GPU)

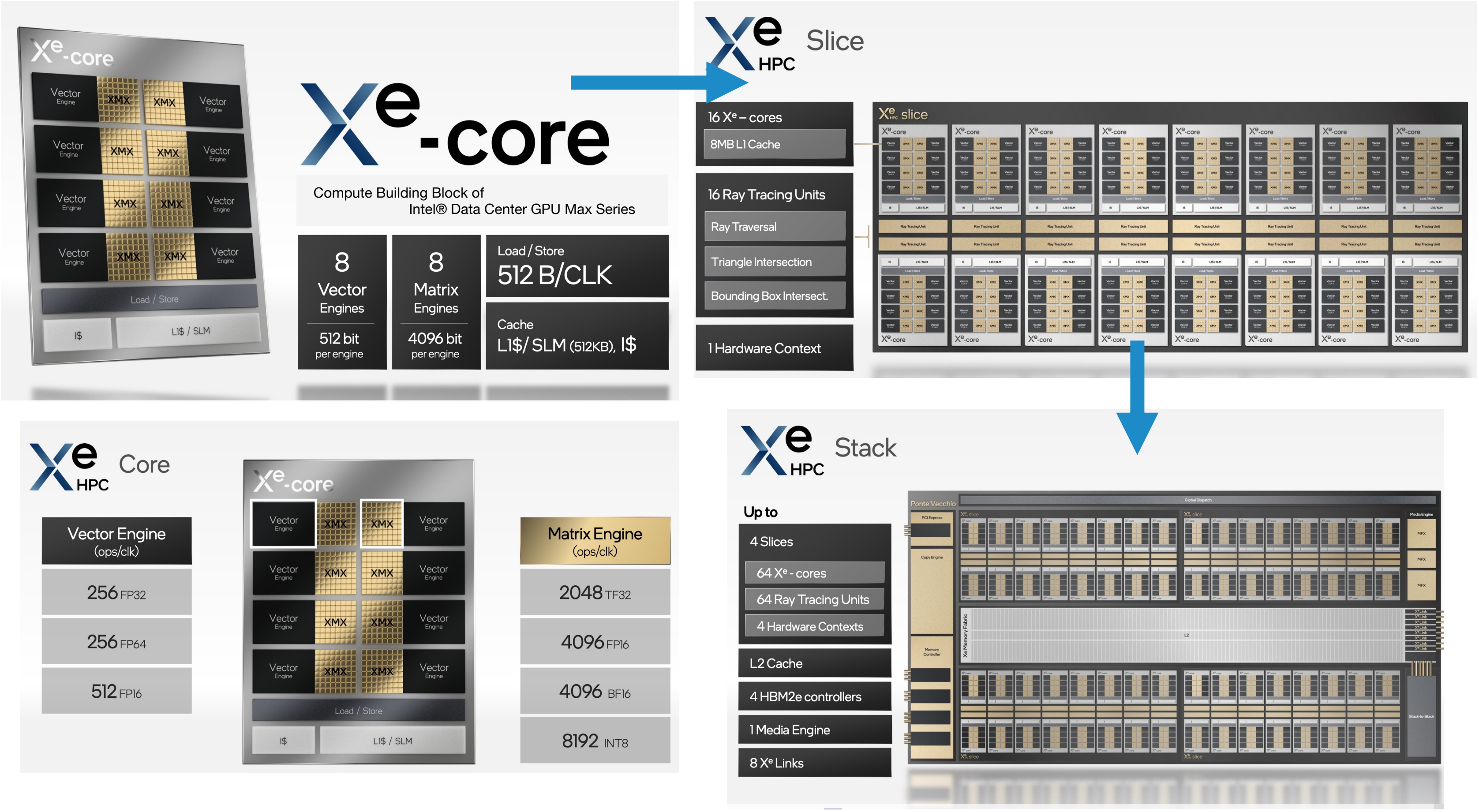
Load/Store, Bulk Data Transfer & Sync Semantics

Up to 8 Fully Connected GPUs through Embedded Switch



Intel Data Center GPU Max Series Architectural Components

- Xe Cores
  - Vector Engine
    - Traditional compute pipeline
  - Matrix Engine
    - Low precision systolic pipeline
  - L1 Data Cache
    - Shared Local Memory
  - Instruction Cache
- Xe Slice
  - —Hardware Context
  - —Offload Units
- Xe Stack
  - —LLC
  - —HBM2e controllers
  - —Xe link
  - —Cache Memory Fabric
  - —PCIe Endpoint
  - —Hardware specific engines
  - —Stack to Stack Interconnect
  - —Xe links
    - Multi GPU Interconnect



X^e-core

Compute Building Block of Intel® Data Center GPU Max Series

8 Vector Engines, 512 bit per engine

8 Matrix Engines, 4096 bit per engine

Load / Store 512 B/CLK

Cache L1$/ SLM (512KB), I$

X^e HPC Slice

16 X^e – cores
- 8MB L1 Cache

16 Ray Tracing Units
- Ray Traversal
- Triangle Intersection
- Bounding Box Intersect.

1 Hardware Context

X^e HPC Core

Vector Engine (ops/clk)	Matrix Engine (ops/clk)
256 FP32	2048 TF32
256 FP64	4096 FP16
512 FP16	4096 BF16
	8192 INT8

X^e HPC Stack

Up to

4 Slices
- 64 X^e - cores
- 64 Ray Tracing Units
- 4 Hardware Contexts

L2 Cache

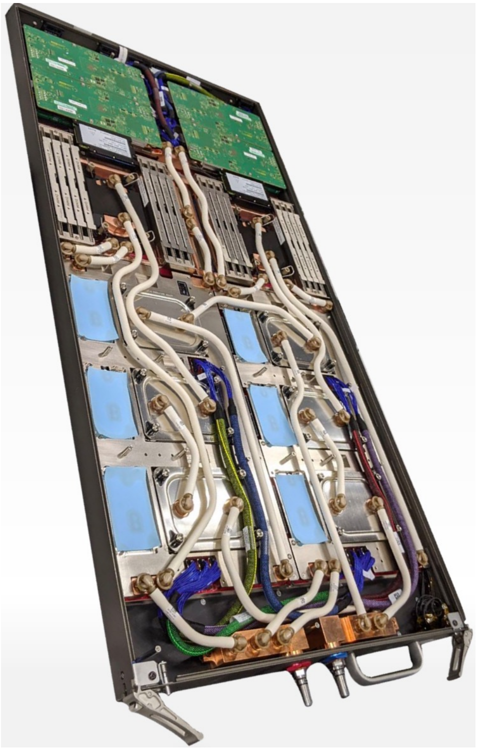
4 HBM2e controllers

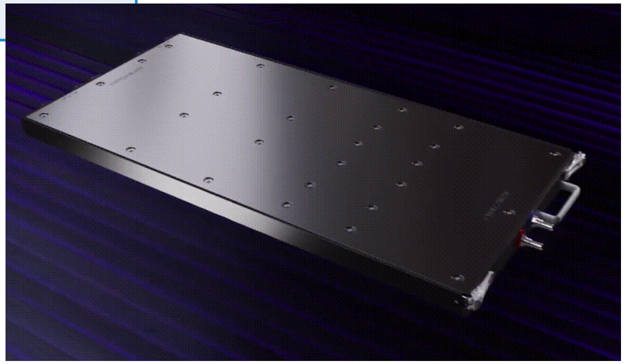
1 Media Engine

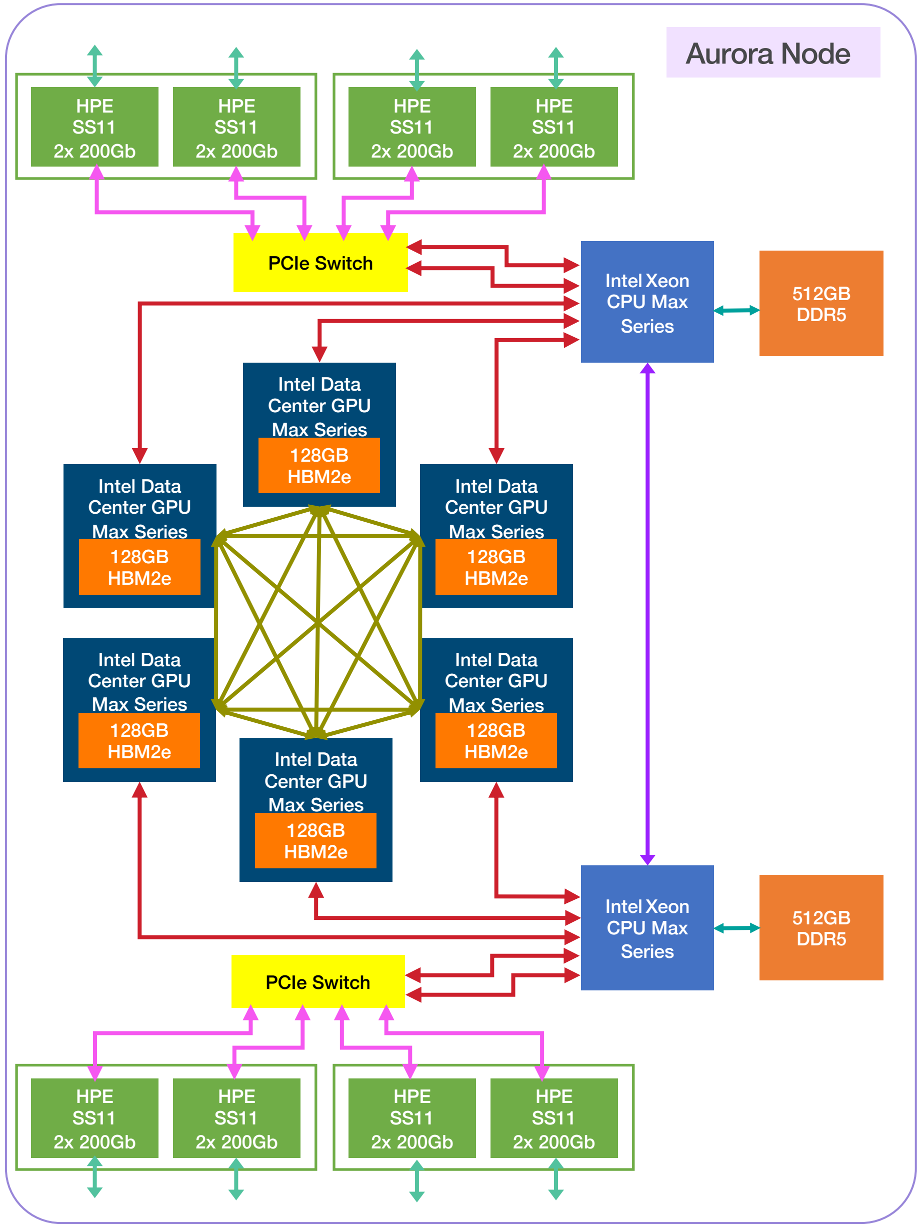
8 X^e Links

Aurora Exascale Compute Blade

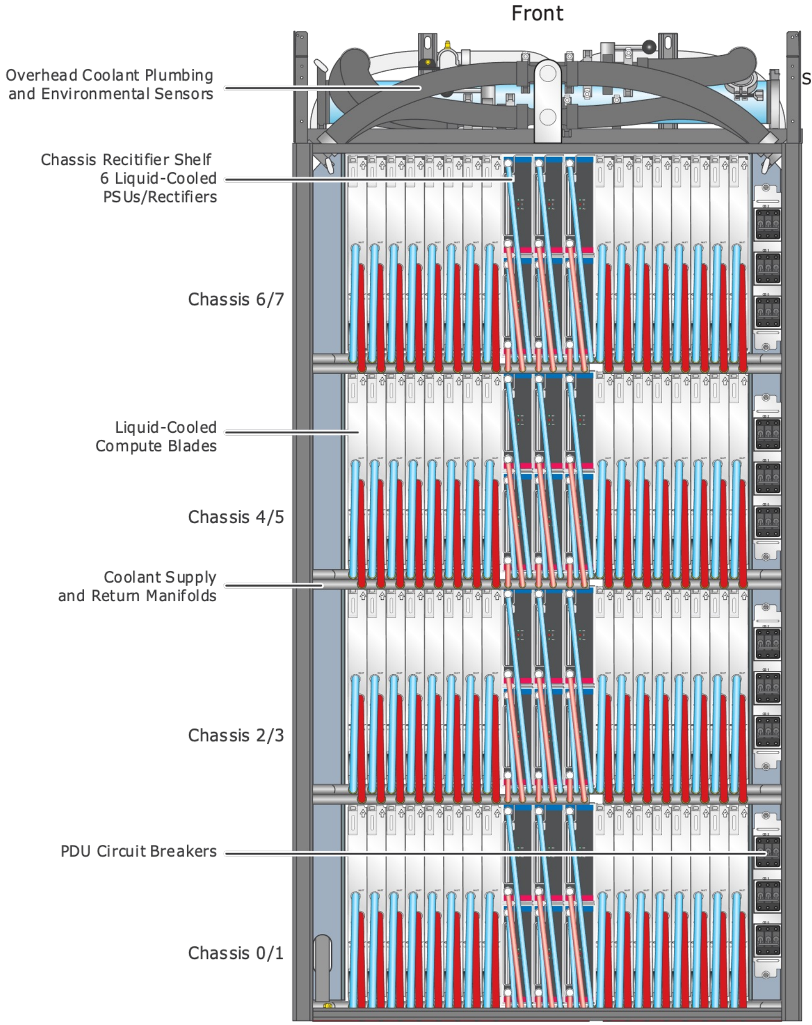
NODE CHARACTERISTICS	
6	GPU - Intel Data Center GPU Max Series (#)
2	CPU - Intel Xeon CPU Max Series (#)
768	GPU HBM Memory (GB)
19.66	Peak GPU HBM BW (TB/s)
128	CPU HBM Memory (GB)
2.87	Peak CPU HBM BW (TB/s)
1024	CPU DDR5 Memory (GB)
0.56	Peak CPU DDR5 BW (TB/s)
≧ 130	Peak Node DP FLOPS (TF)
200	Max Fabric Injection (GB/s)
8	NICs (#)

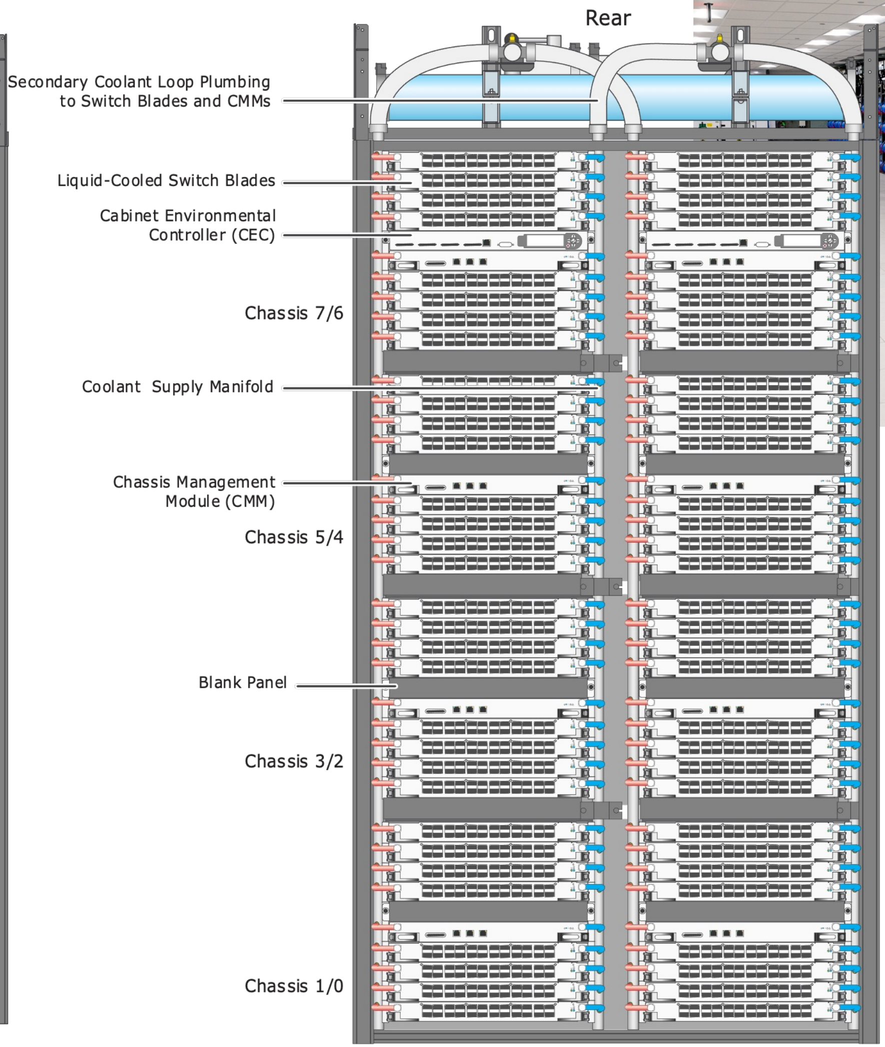





Aurora Node
HPE SS11 2x 200Gb
HPE SS11 2x 200Gb
HPE SS11 2x 200Gb
HPE SS11 2x 200Gb
PCIe Switch
Intel Xeon CPU Max Series
512GB DDR5
Intel Data Center GPU Max Series 128GB HBM2e
Intel Data Center GPU Max Series 128GB HBM2e
Intel Data Center GPU Max Series 128GB HBM2e
Intel Data Center GPU Max Series 128GB HBM2e
Intel Data Center GPU Max Series 128GB HBM2e
Intel Data Center GPU Max Series 128GB HBM2e
Intel Xeon CPU Max Series
512GB DDR5
PCIe Switch
HPE SS11 2x 200Gb
HPE SS11 2x 200Gb
HPE SS11 2x 200Gb
HPE SS11 2x 200Gb

Aurora Cabinets at Argonne


Front
Overhead Coolant Plumbing and Environmental Sensors
Chassis Recitifier Shelf 6 Liquid-Cooled PSUs/Rectifiers
Chassis 6/7
Liquid-Cooled Compute Blades
Chassis 4/5
Coolant Supply and Return Manifolds
Chassis 2/3
PDU Circuit Breakers
Chassis 0/1


Rear
Secondary Coolant Loop Plumbing to Switch Blades and CMMs
Liquid-Cooled Switch Blades
Cabinet Environmental Controller (CEC)
Chassis 7/6
Coolant Supply Manifold
Chassis Management Module (CMM)
Chassis 5/4
Blank Panel
Chassis 3/2
Chassis 1/0

Network Switch

Consistent, Repeatable Application Performance

- Advanced congestion control
- Fine grained adaptive routing
- Very low average and tail latency

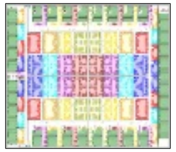
Extremely Scalable RDMA Performance

- Connectionless protocol
- Fine grained flow control
- MPI HW tag matching & progress engine
- Dragonfly topology – 3 switch hops (typical)

Native Ethernet

- Native IP – no encapsulation
- High-scale bandwidth integration to campus

HPE Slingshot Switches - 64 ports @ 200 Gbps



HPE Switch ASIC

Rack switches

100% DLC Switches

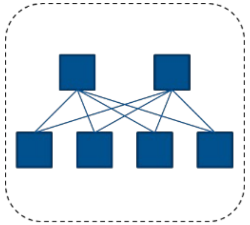
HPE Slingshot NICs - 200 Gbps

HPE NIC ASIC

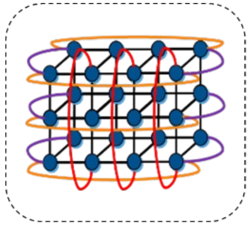
PCIe Adapters

100% DLC NIC Mezz

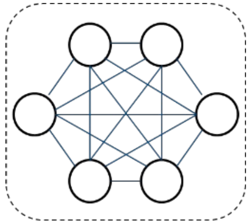
Interconnect Topology



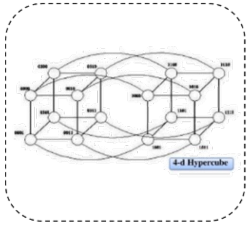
Fat Tree



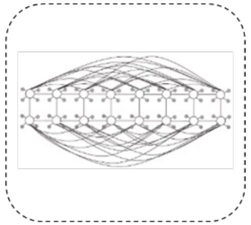
Torus



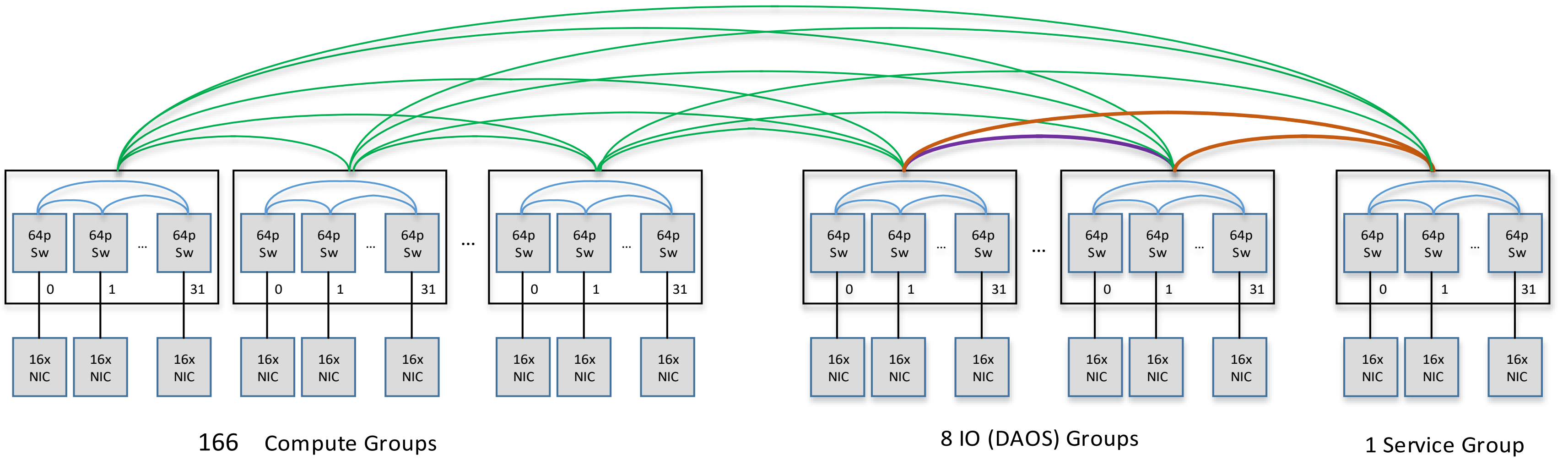
Dragonfly



Hypercube



HyperX



Each Link is 50GB/s bidirectional, 25GB/s unidirectional: 🟦 1 link per arc 🟩 2 links per arc 🟧 8 links per arc 🟪 24 links per arc

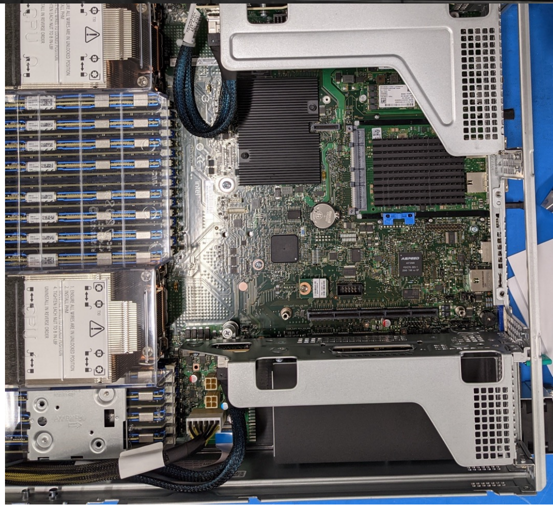
- 1-D Dragonfly Topology - 175 total groups (166 compute + 8 IO + 1 Service),
- All the global links are optical, all the local links in compute groups are electrical
- 2 global links between any two compute groups
- 24 links between any two IO groups, 8 links between the Service group and each IO group
- Total injection bandwidth: 2.12PB/s
- Total bisection bandwidth: 0.69PB/s

Aurora Storage Systems

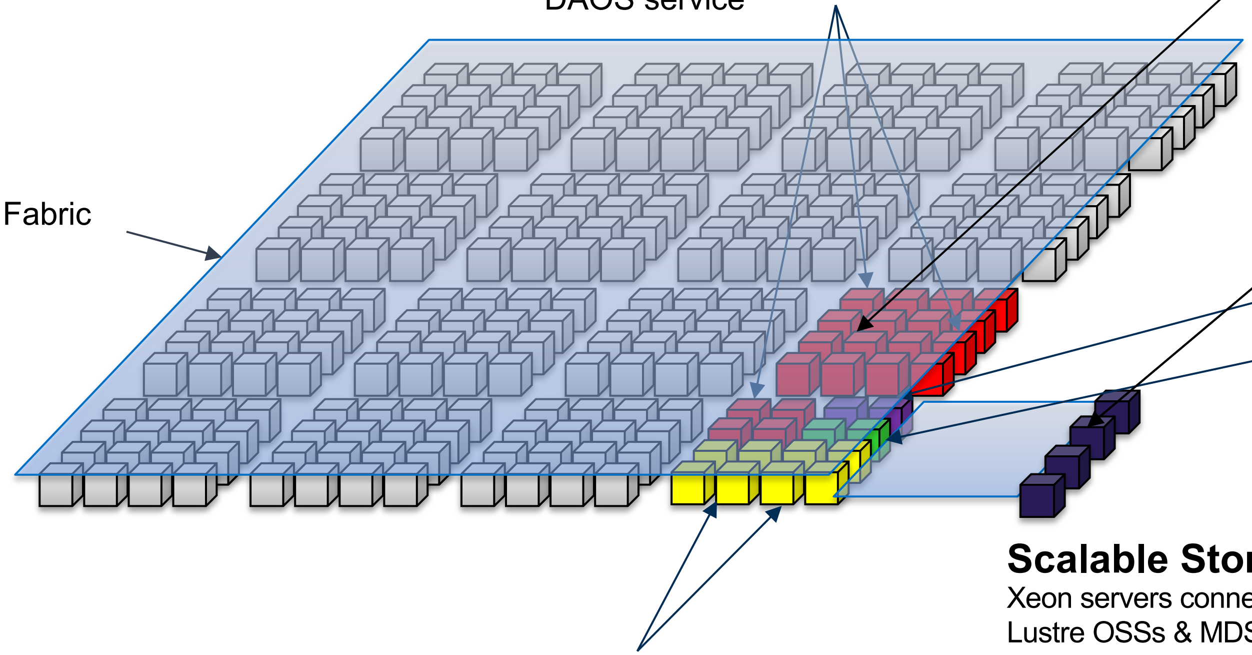
- DAOS provides Aurora's main "platform" high performance storage system
- Aurora leverages existing Lustre storage systems, Grand and Eagle, for center-wide data access and data sharing

System	Capacity	Performance
Aurora DAOS	230 PB @ EC16+2 ▪ 250 PB NVMe ▪ 8 PB Optane PMEM	31 TB/s Read & Write
Eagle	100 PB @ RAID6 ▪ 8480 HDD ▪ 40 Lustre MDT	> 650 GB/s Read & Write
Grand	100 PB @ RAID6 ▪ 8480 HDD ▪ 40 Lustre MDT	> 650 GB/s Read & Write

- Intel Coyote Pass System
  - (2) Xeon 5320 CPU (Ice Lake)
  - (16) 32GB DDR4 DIMMs
  - (16) 512GB Intel Optane Persistent Memory 200
  - (16) 15.3TB Samsung PM1733
  - (2) HPE Slingshot NIC
- 1024 Total Servers
  - Each node will run 2 DAOS engines
  - 2048 DAOS engines



Aurora Storage Overview



DAOS Nodes (DNs)
1024 Xeon servers
(16) 512GB NVRAM
(16) 15TB NVMe attached storage
DAOS service

DAOS Performance
230 PB capacity @ EC16+2
31TB/s

Lustre Performance
Grand – 100 PB @ 650 GB/s
Eagle – 100 PB @ 650 GB/s

Slingshot Fabric

System Service Nodes (SSNs)

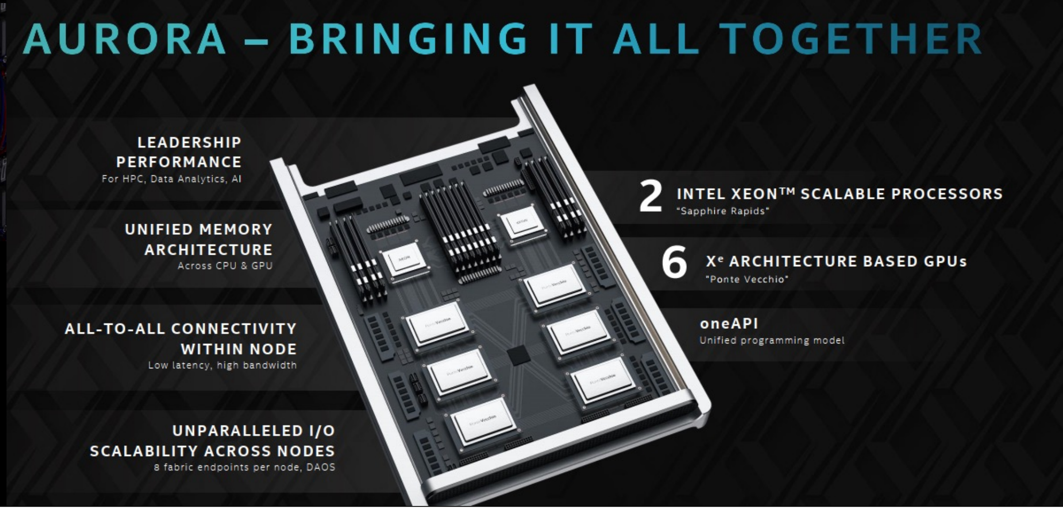
User Access Nodes (UANs)

Scalable Storage Cluster (SSC)
Xeon servers connected to JBOD
Lustre OSSs & MDSs

Gateway nodes

Gateway Nodes (GNs)
100 Xeon servers with no local storage
Access to external storage

Argonne NATIONAL LABORATORY
U.S. DEPARTMENT OF ENERGY
Aurora
intel
Hewlett Packard Enterprise



AURORA – BRINGING IT ALL TOGETHER

LEADERSHIP PERFORMANCE
For HPC, Data Analytics, AI

UNIFIED MEMORY ARCHITECTURE
Across CPU & GPU

ALL-TO-ALL CONNECTIVITY WITHIN NODE
Low latency, high bandwidth

UNPARALLELED I/O SCALABILITY ACROSS NODES
8 fabric endpoints per node, DAOS

2 INTEL XEON™ SCALABLE PROCESSORS
"Sapphire Rapids"

6 Xe ARCHITECTURE BASED GPUs
"Ponte Vecchio"

oneAPI
Unified programming model

Peak Performance
≧ 2 Exaflops DP

Intel GPU
Intel® Data Center GPU Max Series 1550

Intel Xeon Processor
Intel® Xeon Max Series 9470C CPU with High Bandwidth Memory

Platform
HPE Cray-Ex

Compute Node
2x Intel® Xeon Max Series processors
6x Intel® Data Center GPU Max Series
8x Slingshot11 fabric endpoints

GPU Architecture
Intel XeHPC architecture
High Bandwidth Memory

Node Performance
>130 TF

System Size
166 Cabinets
10,624 Nodes
21,248 CPUs
63,744 GPUs

System Memory
1.36PB HBM CPU Capacity
10.9PB DDR5 Capacity
8.16PB HBM GPU Capacity

System Memory Bandwidth
30.58PB/s Peak HBM BW CPU
5.95PB/s Peak DDR5 BW
208.9PB/s Peak HBM BW GPU

High-Performance Storage
230PB
31TB/s DAOS bandwidth
1024 DAOS Nodes

System Interconnect
HPE Slingshot 11
Dragonfly topology with adaptive routing

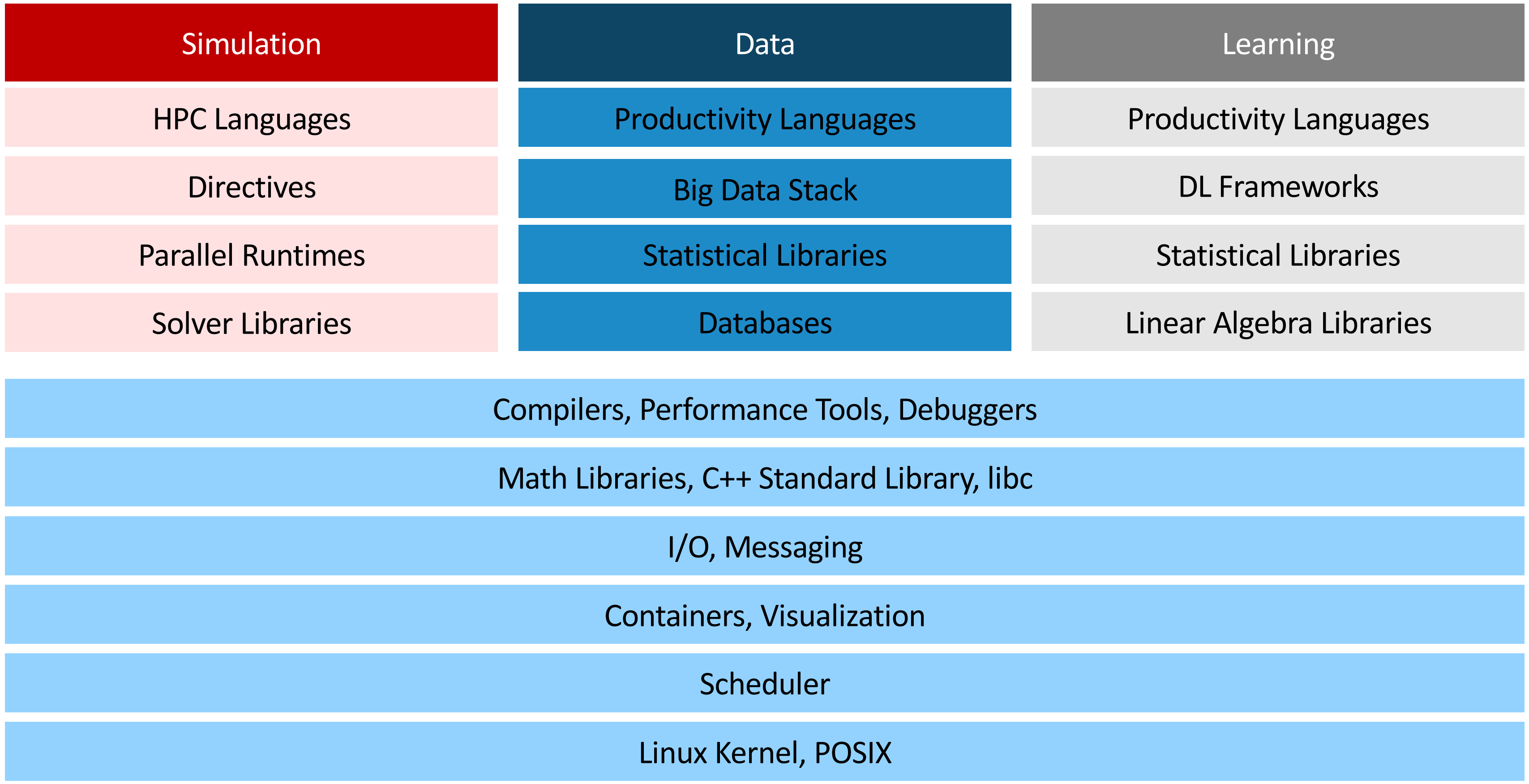
System Interconnect BW
Peak Injection BW 2.12PB/s
Peak Bisection BW 0.69PB/s

Network Switch
25.6 Tb/s per switch (64x 200 Gb/s ports)
Links with 25 GB/s per direction

Programming Environment
- C/C++, Fortran
- SYCL/DPC++
- OpenMP 5.0
- Kokkos, RAJA

AURORA: SOFTWARE

Three Pillars of Aurora



Simulation	Data	Learning
HPC Languages	Productivity Languages	Productivity Languages
Directives	Big Data Stack	DL Frameworks
Parallel Runtimes	Statistical Libraries	Statistical Libraries
Solver Libraries	Databases	Linear Algebra Libraries

Compilers, Performance Tools, Debuggers

Math Libraries, C++ Standard Library, libc

I/O, Messaging

Containers, Visualization

Scheduler

Linux Kernel, POSIX

Introducing oneAPI Ecosystem

“**oneAPI** is a cross-industry, open, standards-based unified programming model that delivers a common developer experience across accelerator architectures—for faster application performance, more productivity, and greater innovation.”

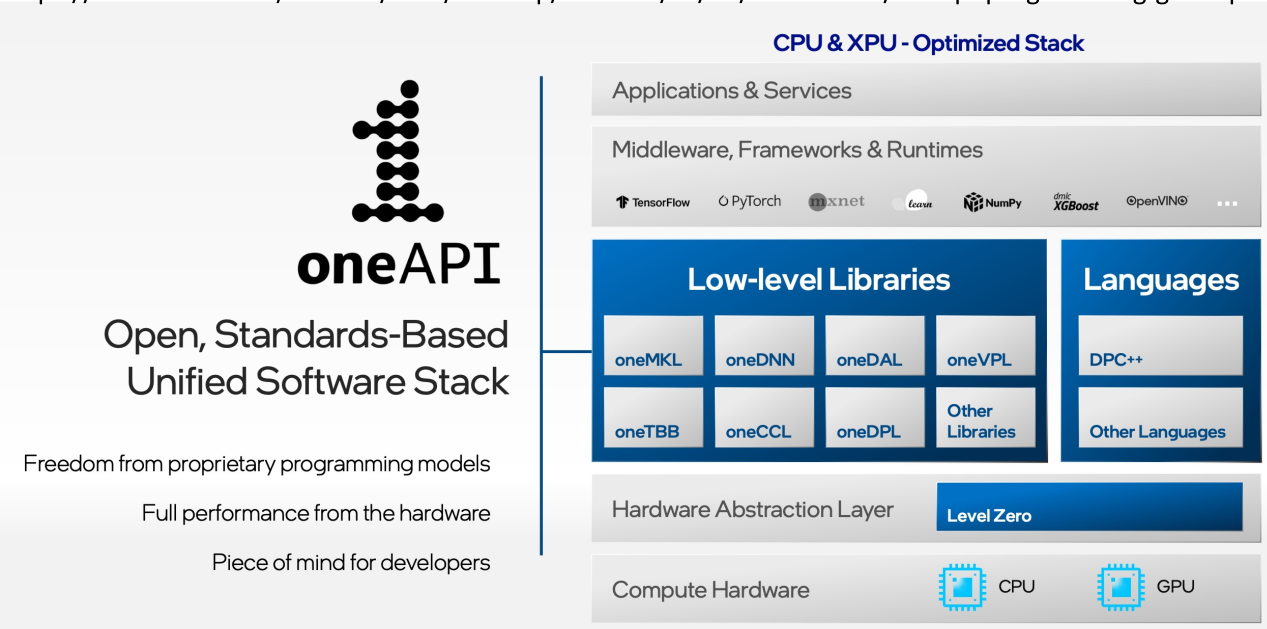
<u>Three Components</u>

- Language
  - DPC++
- Libraries
  - oneMKL, oneDAL, ...
- Hardware Abstraction Layer
  - Level Zero (L0)

Set of specifications that any one can implement

Intel has their own implementations
https://software.intel.com/ONEAPI

https://www.intel.com/content/dam/develop/external/us/en/documents/oneapi-programming-guide.pdf



oneAPI

Open, Standards-Based Unified Software Stack

Freedom from proprietary programming models

Full performance from the hardware

Piece of mind for developers

CPU & XPU - Optimized Stack

Applications & Services

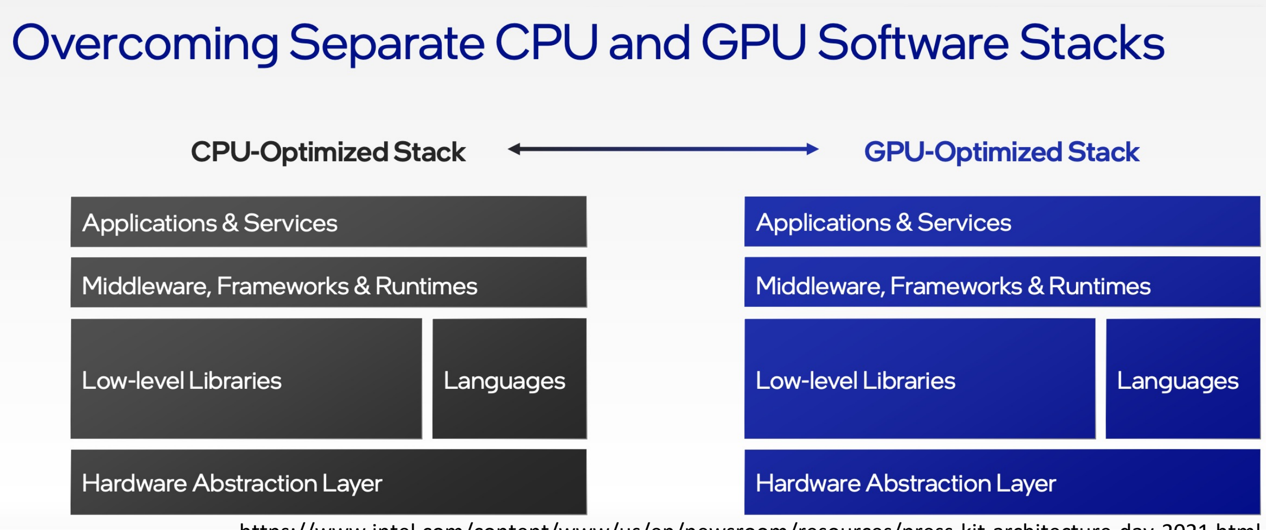
Middleware, Frameworks & Runtimes

Low-level Libraries: oneMKL, oneDNN, oneDAL, oneVPL, oneTBB, oneCCL, oneDPL, Other Libraries

Languages: DPC++, Other Languages

Hardware Abstraction Layer: Level Zero

Compute Hardware: CPU, GPU



Overcoming Separate CPU and GPU Software Stacks

CPU-Optimized Stack ⟷ GPU-Optimized Stack

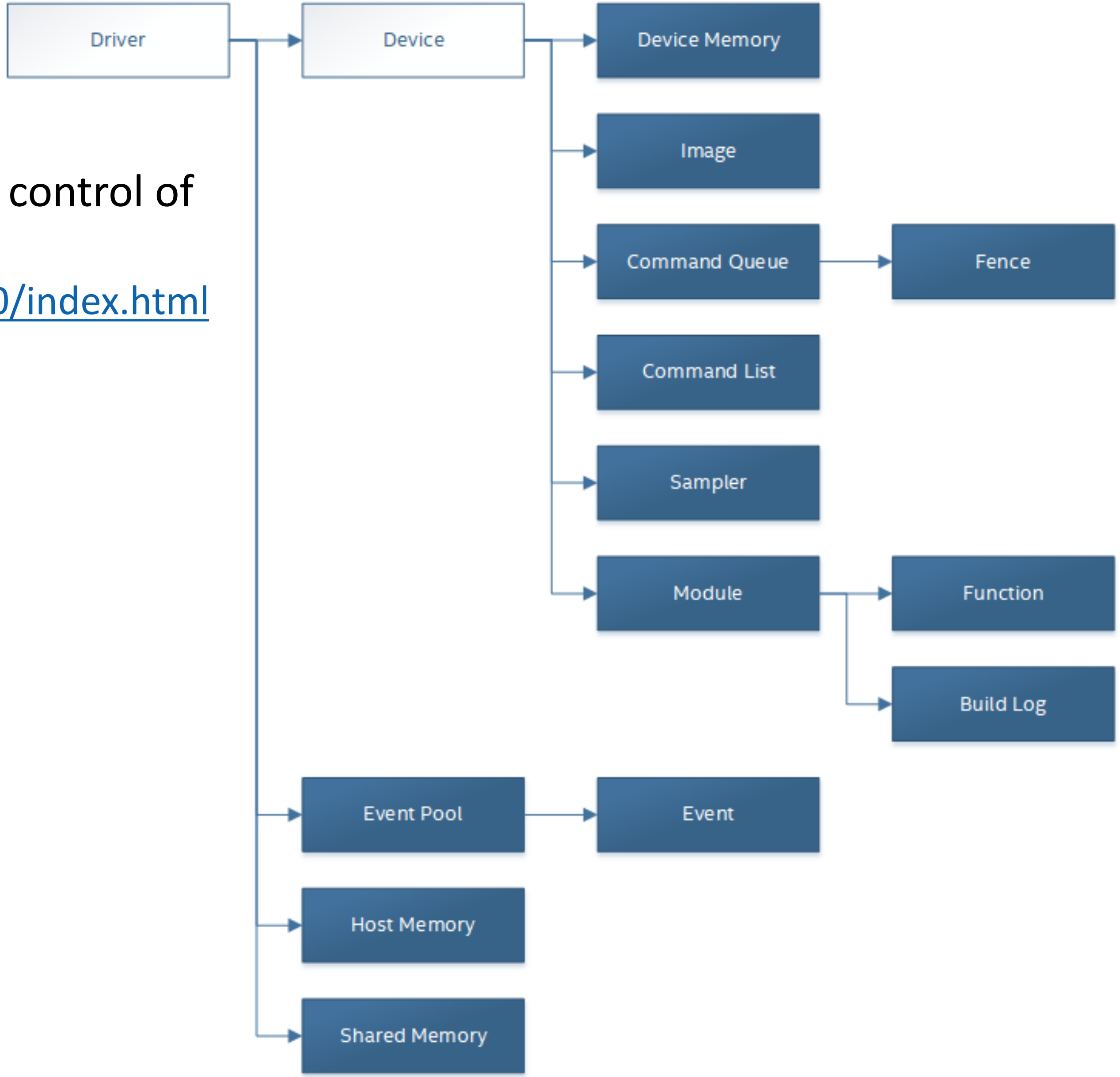
CPU-Optimized Stack		GPU-Optimized Stack	
Applications & Services		Applications & Services	
Middleware, Frameworks & Runtimes		Middleware, Frameworks & Runtimes	
Low-level Libraries	Languages	Low-level Libraries	Languages
Hardware Abstraction Layer		Hardware Abstraction Layer	

https://www.intel.com/content/www/us/en/newsroom/resources/press-kit-architecture-day-2021.html

Level Zero (L0)

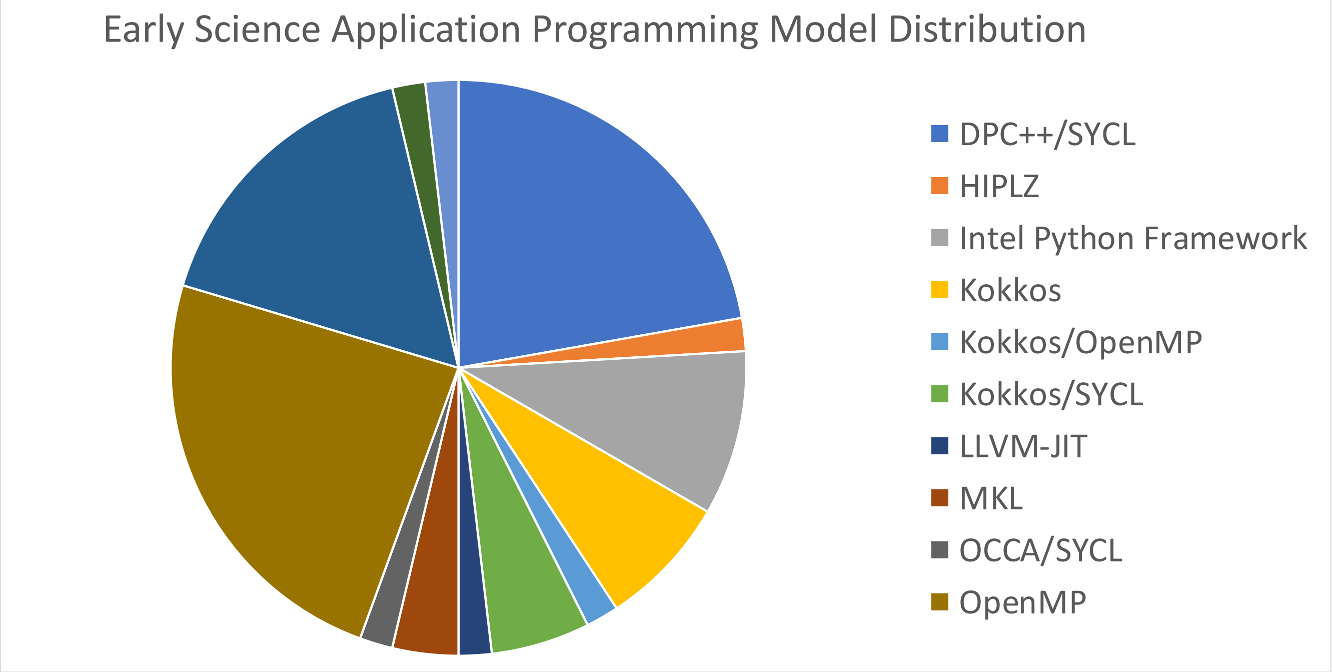
- Low-level programming model for fine grained control of device
  - https://spec.oneapi.com/versions/latest/oneL0/index.html

- Management of:
  - Device memory
  - Synchronization
  - Command queue and command lists
  - And more



Aurora Programming Models

- Aurora applications may use
  - DPC++/SYCL
  - OpenMP
  - Kokkos
  - Raja
  - OpenCL
- Experimental
  - HIP
- Not available on Aurora
  - CUDA
  - OpenACC



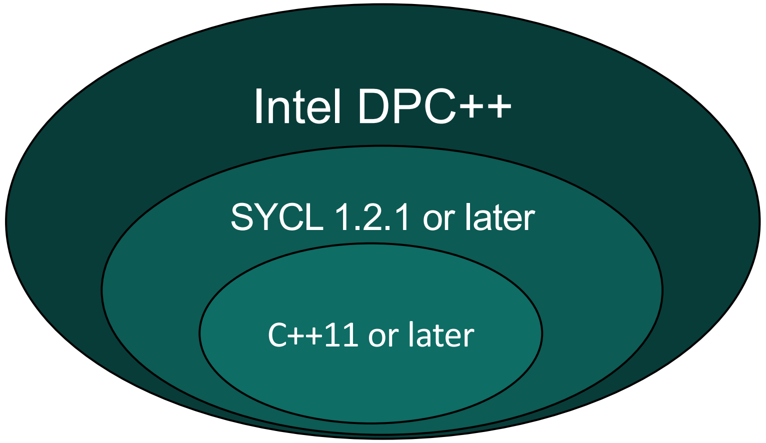
DPC++ (SYCL)

DPC++

- Intel implementation of SYCL standard
- Add language or runtime extensions as needed to meet user needs
- Incorporates SYCL 1.2.1 specification and Unified Shared Memory
- Part of Intel oneAPI specification

SYCL

- Khronos standard specification
- SYCL is a C++ based abstraction layer (standard C++11)
- Based on OpenCL concepts (but single-source)
- *SYCL is designed to be as close to standard C++ as possible*
- Current Implementations of SYCL:
  - ComputeCPP™ (www.codeplay.com)
  - Intel SYCL (github.com/intel/llvm)
  - triSYCL (github.com/triSYCL/triSYCL)
  - hipSYCL (github.com/illuhad/hipSYCL)
  - **Runs on today's CPUs and nVidia, AMD, Intel GPUs**



Extensions	Description
Unified Shared Memory (USM)	defines pointer-based memory accesses and management interfaces.
In-order queues	defines simple in-order semantics for queues, to simplify common coding patterns.
Reduction	provides reduction abstraction to the ND-range form of parallel_for.
Optional lambda name	removes requirement to manually name lambdas that define kernels.
Subgroups	defines a grouping of work-items within a work-group.
Data flow pipes	enables efficient First-In, First-Out (FIFO) communication (FPGA-only)

https://spec.oneapi.com/oneAPI/Elements/dpcpp/dpcpp_root.html#extensions-table

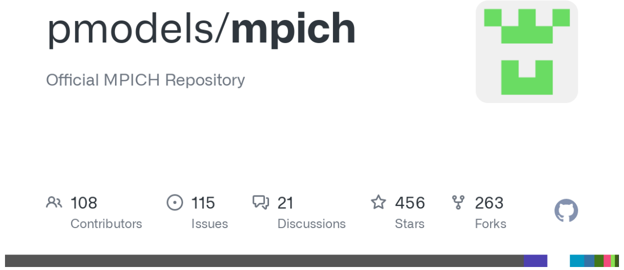
OpenMP 4.5/5

- OpenMP 5 constructs will provide directives based programming model for Intel GPUs
- Available for C, C++, and Fortran
- A portable model expected to be supported on a variety of platforms (Aurora, Frontier, Perlmutter, …)
- Optimized for Aurora
- Integration with MKL for GPU offload

https://www.openmp.org/

MPI

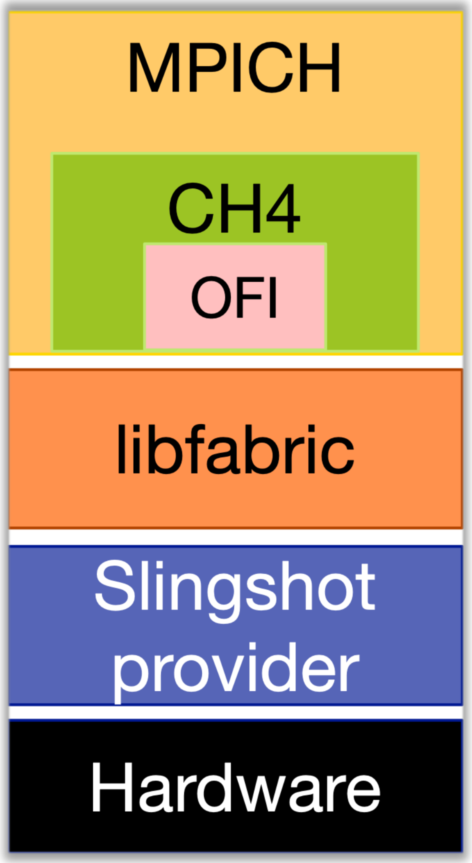
- Based on open source MPICH with new features to support Aurora
- Uses OFI (Open Fabrics Interface) to communicate with the Slingshot Interconnect
- Redesigned to reduce instruction counts and remove non-scalable data structures
- Innovative collective algorithms optimized for Dragonfly network topology
- GPU aware for Intel GPUs
  - It is built on top of oneAPI Level Zero
  - It supports point to point, one-sided, and collectives
  - Support for different data types through the Yaksa library
- Intel GPUs and all-to-all connectivity across the GPUs inside the node
- Multiple NICs on the same node
  - Distribution of processes to NICs
  - Striping (a single rank distributes a single message across multiple NICS)
  - Hashing (a single rank sends different messages through different NICs, e.g., depending on the communicator or the target rank)
  - Efficient multithreading support to use multiple NICs



pmodels/**mpich**

Official MPICH Repository

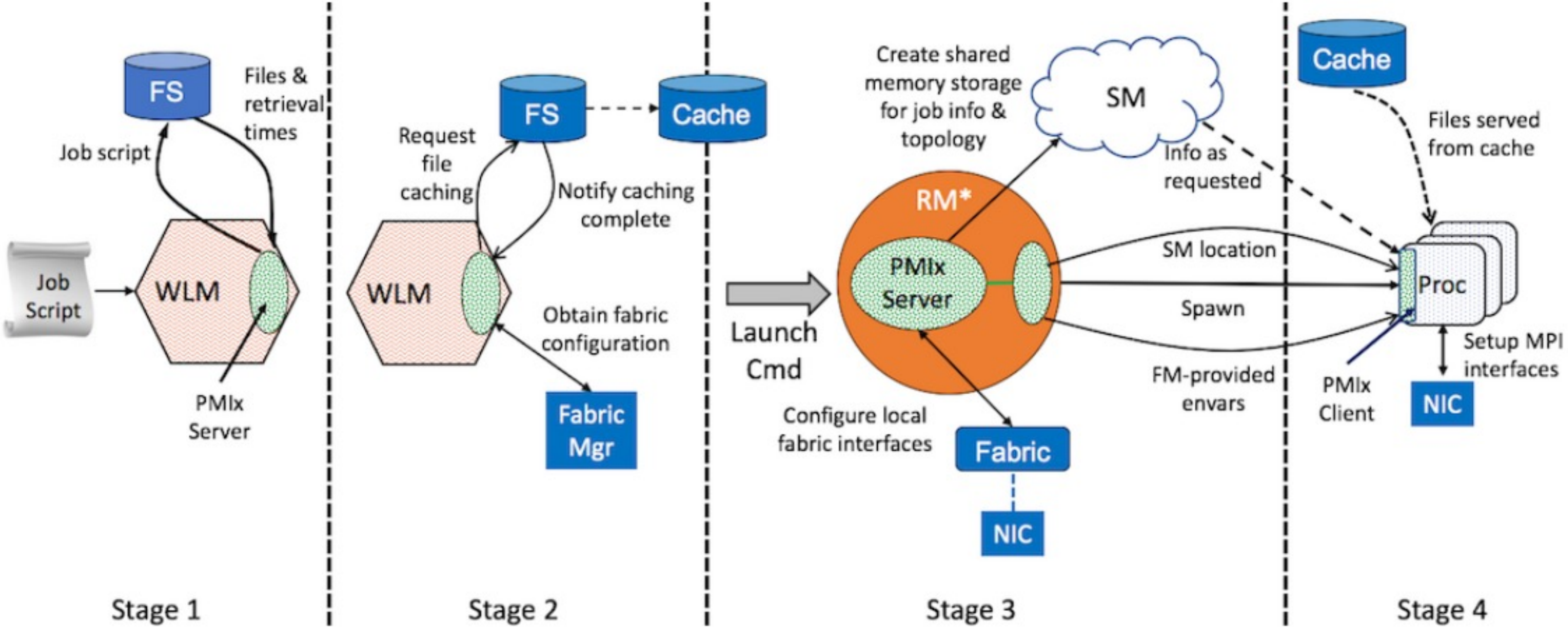
108	115	21	456	263
Contributors	Issues	Discussions	Stars	Forks



Launching jobs on Aurora

- Workload manager (WLM)
  - Handles allocations of nodes to Jobs
  - PBS Pro
- Application Launcher
  - Provides a service to launch applications on the allocated nodes
  - HPE PALS
- Process Management
  - Process Management Interface - Exascale (PMIx)
    - Scalable workflow orchestration by defining an abstract set of interfaces

PMIx 10^{18}



PBS Professional®

HPE Parallel Application Launch Service (PALS)

QUESTIONS?

www.anl.gov

Argonne
NATIONAL LABORATORY