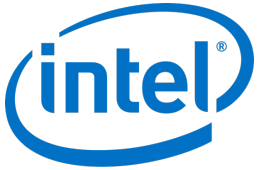
Cray User Group 2025



Scaling MPI Applications on Aurora

Huda Ibeid, Intel	Anthony-Trung Nguyen, Intel	Aditya Nishtala, Intel	Premanand Sakarda, Intel
Larry Kaplan, HPE	Nilakantan Mahadevan, HPE	Michael Woodacre, HPE	Victor Anisimov, ANL
JaeHyuk Kwack, ANL	Vitali Morozov, ANL	Servesh Muralidharan, ANL	Kalyan Kumaran, ANL

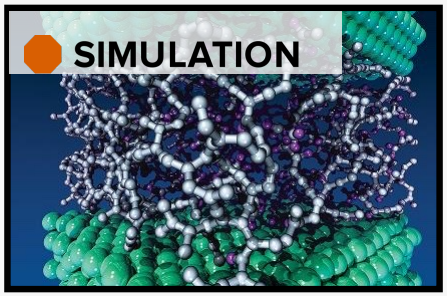
May 8, 2025

Scott Parker, Argonne National Lab

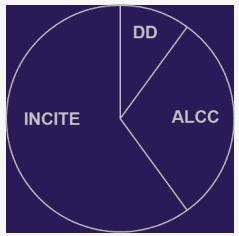
Argonne Leadership Computing Facility

The Argonne Leadership Computing Facility provides world-class computing resources to the scientific community.

- Users pursue scientific challenges
- In-house experts to help maximize results
- Resources fully dedicated to open science



ALCF offers different pipelines for different applications



Architecture supports three types of computing

- Large-scale Simulation (PDEs, traditional HPC)
- Data Intensive Applications (scalable science pipelines)
- Deep Learning and Emerging Science AI (training and inferencing)

#3 in Top500 HPL
#1 in Top500 HPL-MxP
#3 on HPCG
#6 on the IO500

Intel GPU
Intel® Data Center GPU Max Series

Intel Xeon Processor
4th Gen Intel XEON Max Series CPU **with High Bandwidth Memory**

Platform
HPE Cray-Ex

Racks - 166
Nodes - 10,624
CPUs - 21,248
GPUs – 63,744

Interconnect
HPE Slingshot 11
Dragonfly topology with adaptive routing
Network Switch:
25.6 Tb/s per switch (64 200 Gb/s ports)
Links with 25 GB/s per direction

Peak FP Performance
≧ 2 Exaflops DP

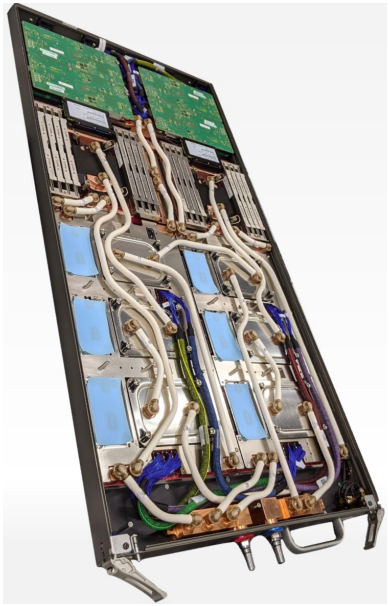
Memory
10.9PB of DDR @ 5.95 PB/s
1.36PB of CPU HBM @ 30.5 PB/s
8.16PB of GPU HBM @ 208.9 PB/s

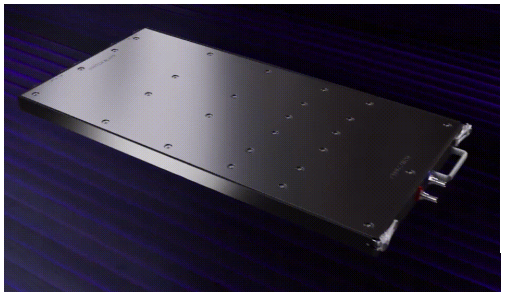
Network
2.12 PB/s Peak Injection BW
0.69 PB/s Peak Bisection BW

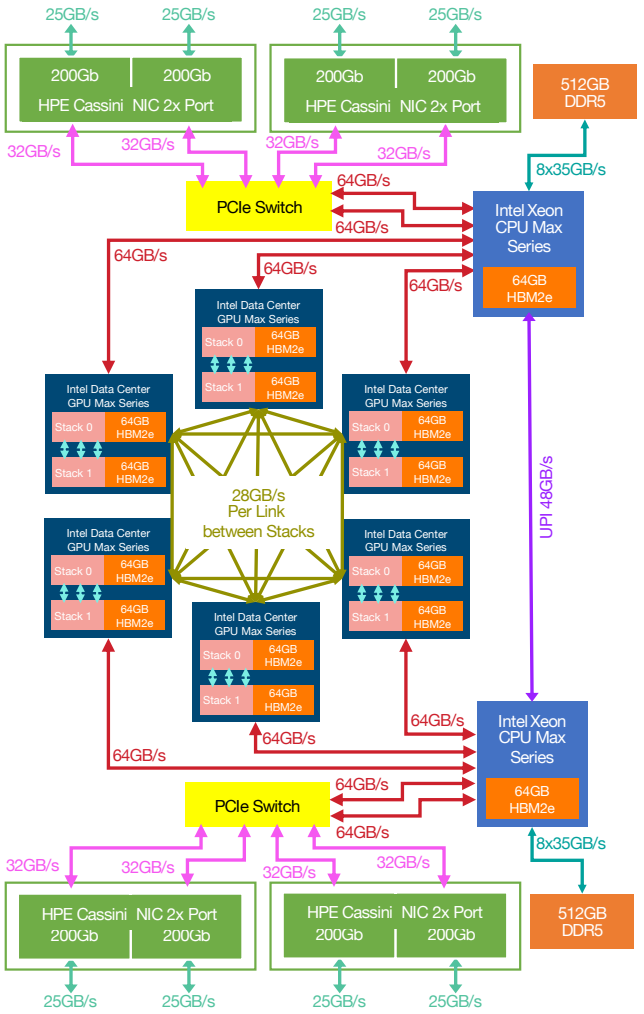
Storage
230PB DAOS Capacity
31 TB/s DAOS Bandwidth

Aurora Exascale Compute Blade

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

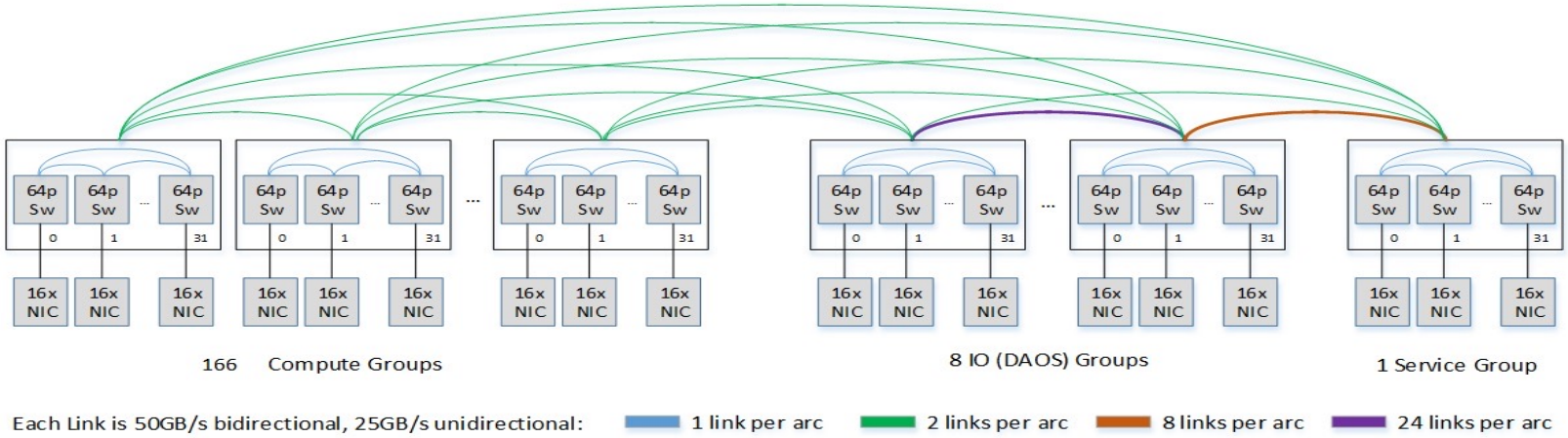






4

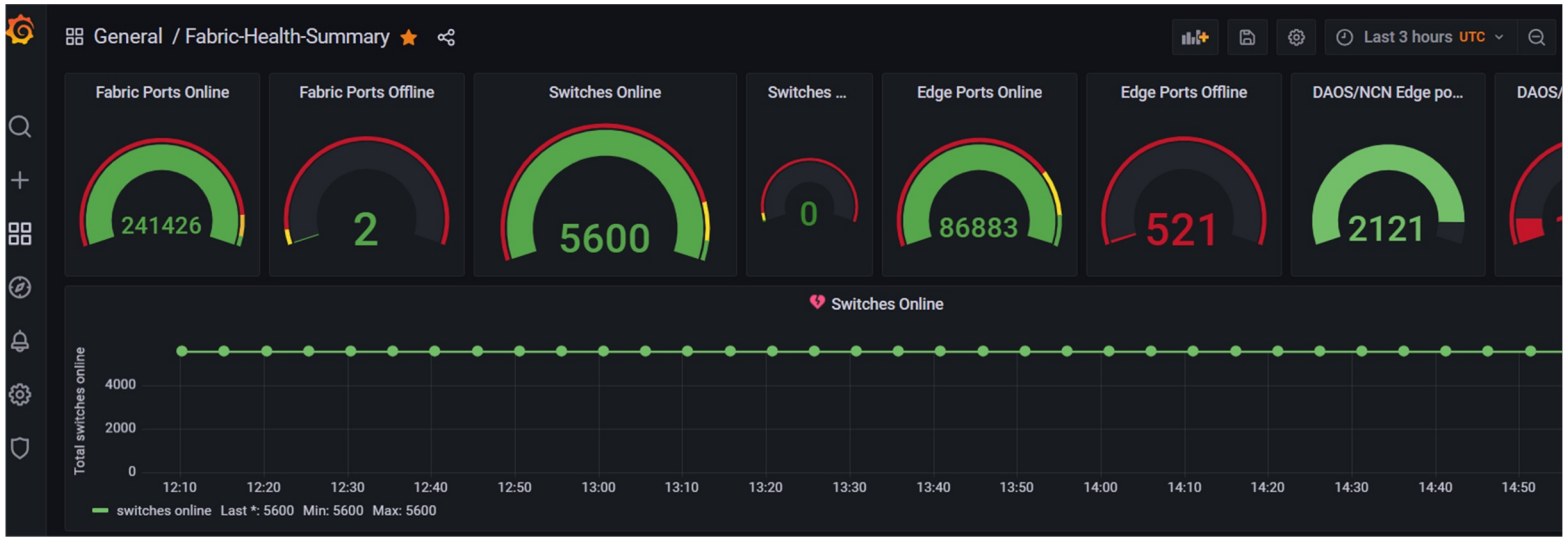
Aurora Network Overview



- Aurora is the largest deployment of a Slingshot-11 interconnect with over 300,000 ports and
  - 166 compute groups (8 NICS per node for a total of 84,992 NICS)
  - 8 storage groups
  - 1 service group
- Aurora's network architecture is a single dimension dragonfly topology with 3 hops max between end points
  - uses all-to-all local groups connected to each other through global links
  - compute group is a single HPE Cray EX cabinet with 32 switches that are all-to-all connected
- All-to-all connectivity at the global level:
  - The 8 I/O DAOS groups have 24 links between each pair
  - The 166 compute groups have 2 links between each pair

5

Aurora Network Overview



- System bandwidths:
  - 2.12 PB/s of injection bandwidth.
  - 1.38 PB/s of global bandwidth
  - 0.69 PB/s of global bisection bandwidth

6

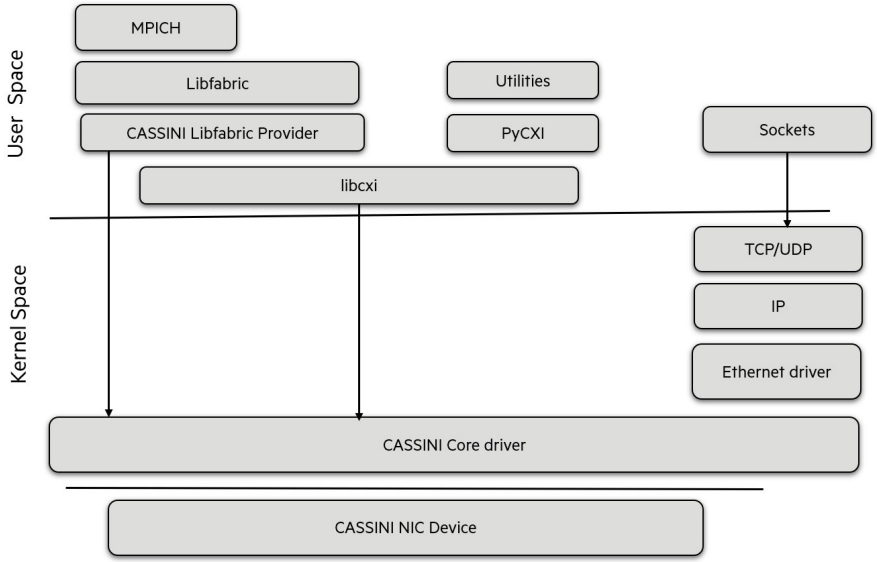
Slingshot Routing and Traffic Management

- Dynamic Routing
  - In the absence of contention, all traffic is routed minimally
  - In the presence of congestion, packets are routed non-minimally to avoid the congestion
- Congestion Management
  - Congestion management reduces interference between jobs
  - Switch hardware will detect congestion and identify its causes
  - Switch hardware applies back pressure to congesting traffic, limiting injections to fair share of bandwidth
  - Traffic not contributing to the congestion is unaffected by the back-pressure and is free to pass blocked packets.
- QoS
  - On all network traffic for performance isolation across different classes of traffic
  - Network traffic is tagged with a traffic class according to the issuing queue
  - Traffic shaping in the NICs and the switches operate on these classes
  - Arbiters can select packets to forward based on their traffic class and the credits available to that class.
  - Supports priority-based traffic scheduling, prioritizing traffic from a given class above that of other classes

7

Network Software

- Messaging Stack:
  - Cassini Drivers – Operates the NIC
  - CXI provider - Interface between libfabric and Cassini
  - Libfabric – Low level communication API to network
  - MPICH – Open source MPI implementation supporting direct access to Intel GPUs and large scale system optimizations
- Fabric Manager
  - Configuration and Initialization of network components
  - Monitoring:
    - Identifies local and global links that are unhealthy and require hardware action
    - Identifies switches that exhibit any hardware errors and required hardware actions
    - Provides insights into performance anomalies requiring analysis of counters from different subsystems.
    - Fabric components identified as problematic can be put to maintenance mode for diagnosis



8

Network Settings

- Routing Bias and Group Load Setting
  - Using the group load setting adaptive routing is able to use an lightly loaded intermediate group
  - Ensures that non-minimal adaptive routing is able to optimally utilize all the available global links
- Fabric Manager Sweep Interval Settings
  - Maintenance routines occur at predefined cadences which can be fine-tuned or changed for optimization
  - In large systems setting sweep intervals very aggressively can increase the load on the Fabric Manager
  - Setting a large value will result in delays in handling the system events and can cause performance impacts.
- QoS classes
  - Quality of Service (QoS) is used to control traffic and help ensure the performance of applications.
  - The profile used is LlBeBdEt (Profile 2) - HPC low latency, HPC bulk data, HPC best effort, and Ethernet (LlBeBdEt) :

9

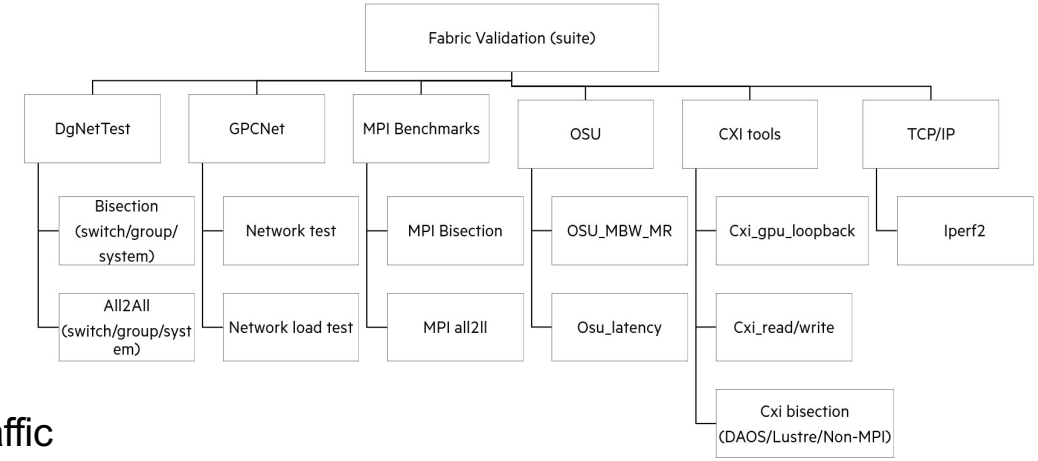
Complexities in Large Scale HPC Systems

- Deployment and operation of large scale HPC systems has become increasingly complex
- System have more components which are increasingly complex with more and smaller transistors making them more susceptible to manufacturing defects and transient errors (cosmic rays, etc)
- Aurora has over:
  - —60,000 GPUs
  - —20,000 CPUs
  - —80,000 NICs
  - —5,600 switches
  - —200,000 network links
- Components are often designed for smaller scale or commodity markets and failure rates acceptable at these scale become problematic in systems with ten or hundreds of thousands of devices
- All of these components must be functioning correctly in order to perform full scale runs on Aurora.
- Applications typically utilize 20% or more of the system (2,000 nodes) for multi-hour runs and need to be able to effectively utilize the CPUs and GPUs and closely synchronous and share data across the nodes using the slingshot network

10

Fabric Validation

- HPE fabric validation suite identifies failures and low performing nodes or groups
- Run on hierarchy of hardware
  - Node → Switch → Group → System
- Consists of multiple components including:
  - **MPI All2all** – connectivity issues, hardware issues, performance bottlenecks
  - **GCPNet** – natural ring pattern, random ring pattern. Runs isolated and with congesting traffic
  - **OSU** - osu_mbw_mr test is used for bandwidth and osu_multi_lat test
  - **CXI Counters** – reported by Cray MPI
  - **Job Prolog** – cxi_healthcheck, cxi_gpu_loopback, slingshot-diag
  - **Job Epilog** – Cassini flaps, Cassini service cleanup, Node hardware errors
- Types of issues identified
  - Network timeouts caused by fabric or node events
  - Low performing nodes
    - Node level hardware issues
    - Link flaps – transient issues that cause a link to reset

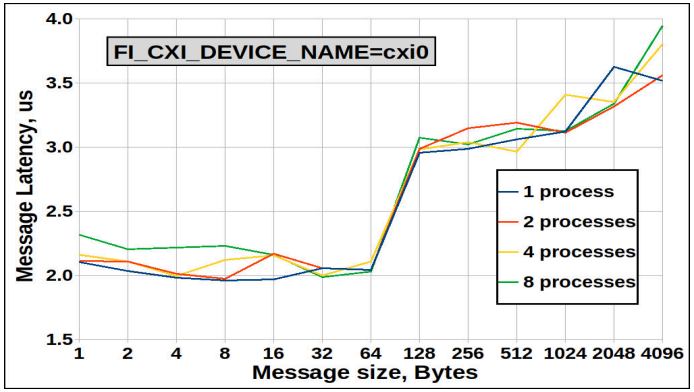


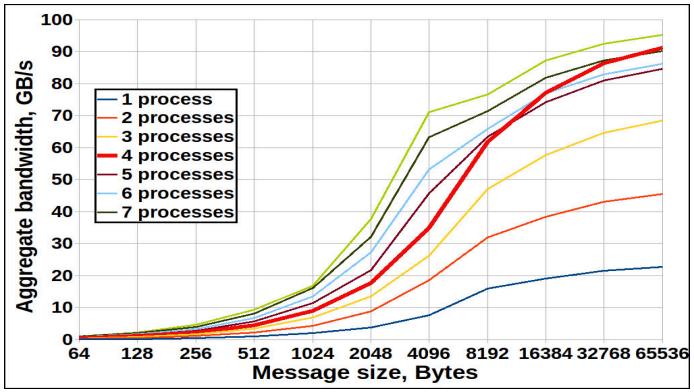
11

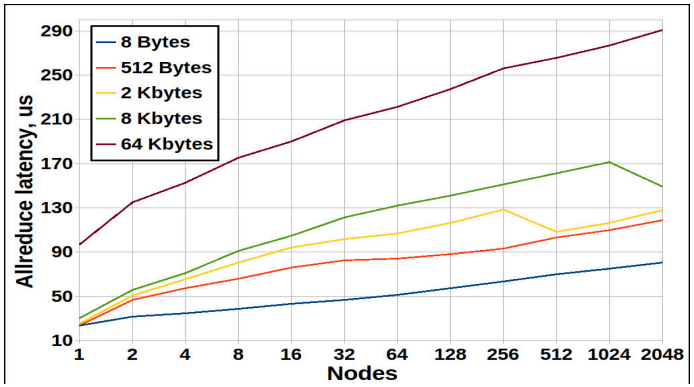
MPI Benchmarks

MPI benchmarks are used to evaluate communication patterns of typical workloads:

- **Single hop latency -** From host memory using a single NIC between two nodes in group
  - The latency stays constant showing that a NIC can effectively multiplex with little penalty
  - Jump from 64 Bytes to 128 Bytes shows when message buffering uses DRAM instead of NIC's SRAM
- **Off node bandwidth -** Multiple MPI processes are placed on a single socket and assigned to the four NICs attached to that socket in round-robin fashion.
  - Bandwidth is linearly increasing as the processes are added up to 4 ranks beyond which the processes start sharing the NICs.
  - NIC performance cannot be saturated by one process per NIC and adding the second process achieves higher bandwidth
- **All Reduce -** An important element of applications performance at larger scales is the cost of MPI collective operations.
  - Performance at different node counts up to 2048 nodes and for different message sizes
  - Less than linear latency growth typical for a recursive-doubling tree algorithm
  - A switch from a ring algorithm to a tree algorithm is seen on the curves





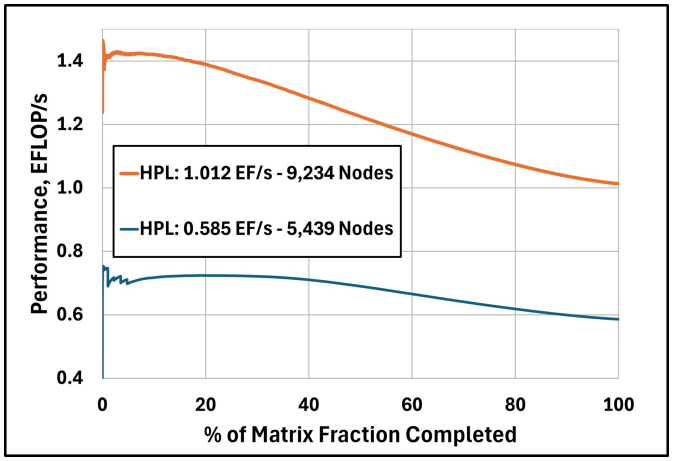


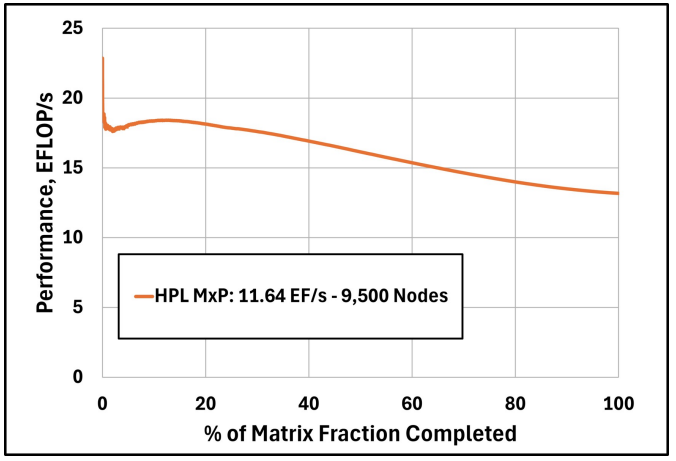
12

Performance Benchmarks

Four benchmarks were used as part of the Aurora bring-up:

- HPL -
  - 1.012 EF/s using 9,234 nodes, rank #3
  - Scaling efficiency of 78.84% over a runtime of 4 hours, 21 minutes, and 54 seconds
- HPL-MxP - mixed-precision variant of HPL designed to better represent the performance of systems on AI-oriented workloads.
  —11.64 EF/s using 9,500 nodes, ranks #1
- Graph500 - targets data-intensive workloads and emphasizes memory access patterns and graph traversal rather than floating-point operations.
  —69,373 GTEPS with 8,192 nodes, ranks #6
- HPCG - computational benchmark focused on memory bandwidth, data access patterns, and communication
  —5.613 PFLOPS with 4,096 nodes, ranks #3

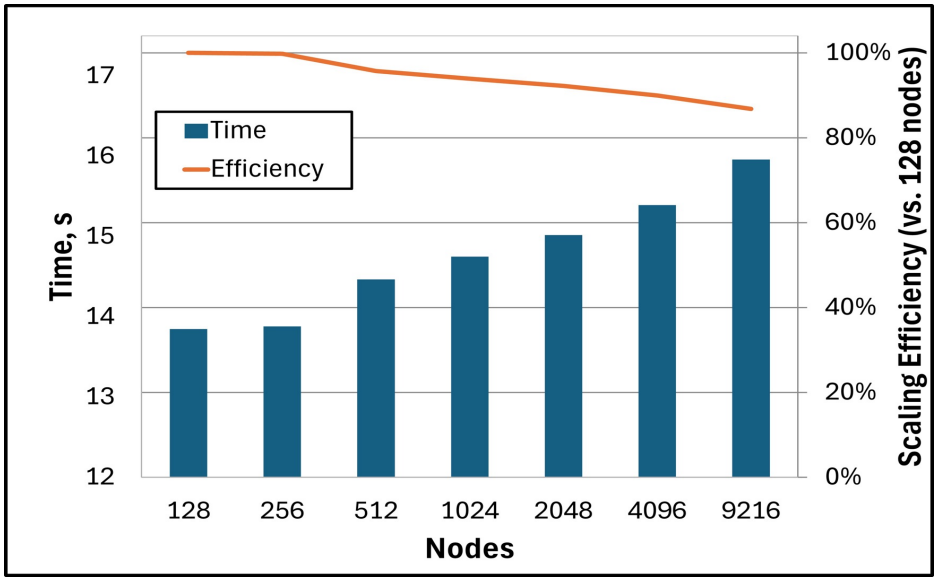




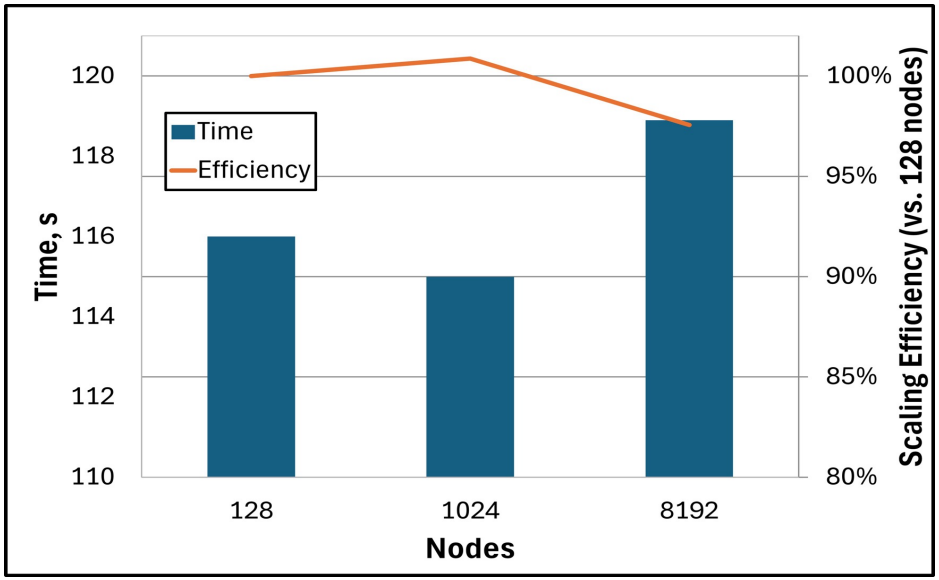
13

Applications: LAMMPS and HACC

- LAMMPS is a highly scalable molecular dynamics code designed for the simulation of materials and biomolecular
- The LAMMPS Rhodopsin benchmark in the largest configuration spans 254 billion atoms across 9,216 nodes
- Figure shows the weak scaling performance as the system scales from 128 to 9,216 nodes with over 85% parallel efficiency



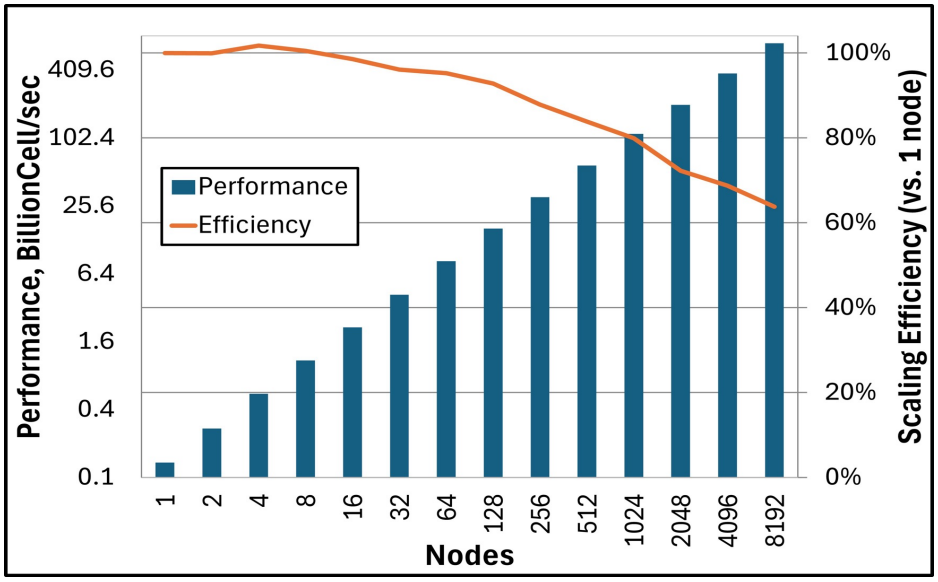
- HACC is a high-performance simulation framework designed to solve large-scale cosmological problems efficiently
- It is widely used for simulating dark matter evolution, galaxy formation, and large-scale cosmic structures.
- Weak scaled across 128, 1,024, and 8,192 nodes 97% efficiency at 8,192 nodes



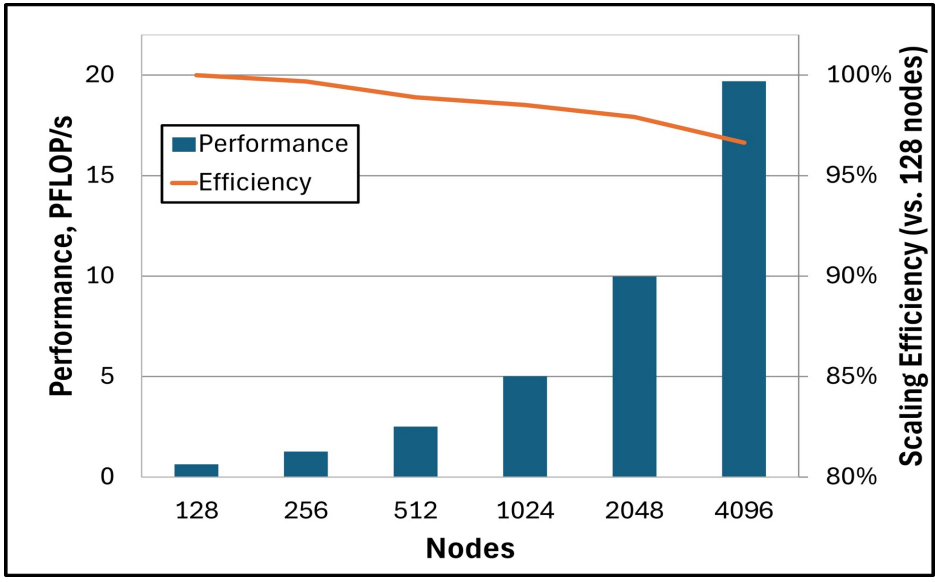
14

Applications: AMR-Wind & Nekbone

- AMR-Wind is a block-structured adaptive-mesh, incompressible flow solver for wind farm simulations
- Primary applications is performing large-eddy simulations of atmospheric boundary layer flows
- Built on top of the AMReX library with a SYCL backend for Intel GPUs.
- Weak scaling from 1 to 8192 nodes with ~62% efficiency



- Nekbone is a proxy application derived from Nek5000 which is high-order spectral element CFD code for solving the incompressible Navier--Stokes equations.
- The code is implemented in Fortran and C, with Fortran handling most numerical operations and C handling communication routines.
- Nekbone demonstrates excellent weak scaling with over 95% parallel efficiency up to 4,096 nodes.



15

Application - FMM

- The Fast Multipole Method (FMM) is a part of electronic structure theory package NWChemEx and computes long-range electrostatic interactions between charged particles in a linear-scaling fashion.
- FMM has an irregular communication pattern where each process requests a large number of data pieces sparsely populated in the local memory of numerous remote ranks
- One-sided model ideally fits into this pattern (MPI_Get, MPI_Put)
- One-sided communication it has historically been less frequently utilized in HPC than two-sided communication methods and has required more testing and debugging to make work at scale
- Benchmarks have been developed and utilized to test the functionality and performance of one-sided communication
- One issue identified is that the PVC GPU has been found to be unable to provide RMA support in hardware and instead it has been required to be implemented in software
- MPI_Get is significantly faster than that of MPI_Put
- Overall initial tests demonstrate satisfactory performance

N Nodes	with HMEM	without HMEM
1 × 8	0.9	24.6
1 × 16	1.1	17.1
1 × 32	1.6	13.0
9 × 16	14.5	

Time, sec to complete data transfer by using MPI_Get.

N Nodes	with HMEM	without HMEM
1 × 8	14.2	28.4
1 × 16	17.6	38.9
1 × 32	20.7	49.7

Time, sec to complete data transfer by using MPI_Put.

16

Conclusion

- Aurora was deployment occurred 2024 and began production computing with scientific applications in 2025
- It is the largest deployment to date of the HPE Slingshot Fabric and the first large system using Intel's first discrete data center GPU
- Machine capabilities demonstrated through benchmarks including:
  - HPL – 1.012 EF on 9234 nodes, 3rd on the Top500 list
  - HPL-MxP – 11.64 EF/s on 9500 nodes, #1 on Top500 list
  - Graph500 – 69, 373 GTEPS on 8,192 nodes, #6 on Graph500 list
  - HPCG – 5.6 PFLOPS on 4096 nodes, #3 HPCG list
- Several workload applications have been demonstrated weak scaling:
  - LAMMPS: 85% parallel efficiency on 9,216 nodes
  - HACC: 97% parallel efficiency on 8,192 nodes
  - AMR-Wind: 63% parallel efficiency on 8,192 nodes
  - Nekbone: 95% parallel efficiency on 4,096 nodes

17