Experiences with MVAPICH2 Deployment on SDSC's Expanse Supercomputer

MVAPICH2 User Group (MUG) Meeting
August 25, 2021

Mahidhar Tatineni
San Diego Supercomputer Center (SDSC)

EXPANSE
COMPUTING WITHOUT BOUNDARIES

SAN DIEGO SUPERCOMPUTER CENTER

NSF Award 1928224

Outline

- Introduction and Overview
- Expanse system architecture
- AMD EPYC Processor Architecture
  - Hardware details
  - NUMA options
  - Applications
- Expanse benchmark results with MVAPICH2/2.3.6
- Summary

InfiniBand and MVAPICH2 on SDSC Systems

Trestles (NSF)
2011-2014

- *324 nodes, 10,368 cores*
- *4-socket AMD Magny-Cours*
- *QDR InfiniBand*
- *Fat Tree topology*
- *MVAPICH2*

Gordon (NSF)
2012-2017
GordonS
(Simons Foundation)
2017- 2020

- *1024 nodes, 16,384 cores*
- *2-socket Intel Sandy Bridge*
- *Dual Rail QDR InfiniBand*
- *3-D Torus topology*
- *300TB of SSD storage - via iSCSI over RDMA (iSER)*
- *MVAPICH2 (1.9, 2.1) with 3-D torus support*

COMET (NSF)
2015-2021
COMET (CW3E)
2021-Current

- *1944 compute, 72 GPU, and 4 large memory nodes.*
- *2-socket Intel Haswell*
- *FDR InfiniBand*
- *Fat Tree topology*
- *MVAPICH2, MVAPICH2-X, MVAPICH2-GDR*
- *Leverage SRIOV for Virtual Clusters*

Expanse (NSF)
2020 (Dec) - Current

- ***784 compute, 56 GPU, and 4 large memory nodes.***
- ***2-socket AMD EPYC 7742, HDR100 InfiniBand***
- ***GPU nodes with 4 V100 GPUs + NVLINK***
- ***HDR200 Switches, Fat Tree topology with 3:1 oversubscription***
- ***MVAPICH2, MVAPICH2-GDR; MVAPICH2-X will be installed.***

Computing Without Boundaries: Cyberinfrastructure for the Long Tail of Science

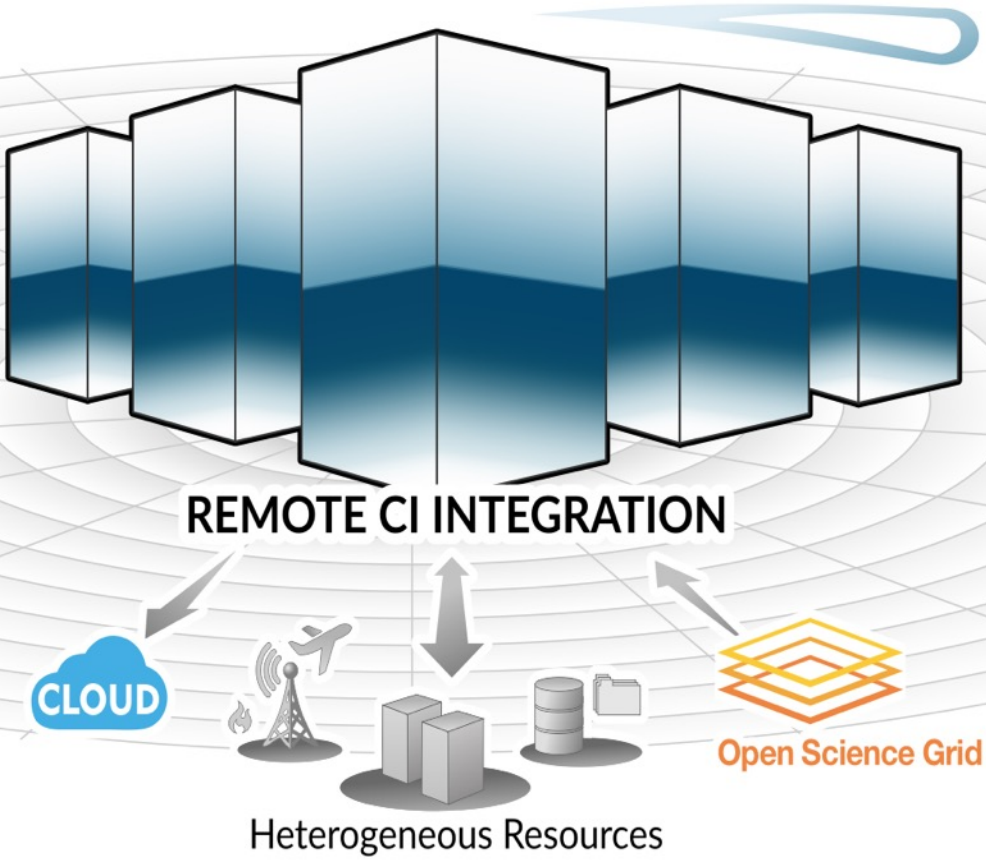
- NSF Solicitation 19-534: Advanced Computing Systems & Services: Adapting to the Rapid Evolution of Science and Engineering Research
- Category 1: Capacity System, NSF Award # 1928224
- NSF Program Officer: Robert Chadduck
- PIs: Mike Norman (PI), Ilkay Altintas, Amit Majumdar, Mahidhar Tatineni, Shawn Strande
- $10M Acquisition; Operations and Maintenance funding est. $2.5M/year
- Primary Vendors: Dell (HPC system); Aeon Computing (storage)
- Compute, interconnect, NVMe: AMD, Intel, NVIDIA, Mellanox

EXPANSE

COMPUTING WITHOUT BOUNDARIES
5 PETAFLOP/S HPC and DATA RESOURCE

HPC RESOURCE
13 Scalable Compute Units
728 Standard Compute Nodes
52 GPU Nodes: 208 GPUs
4 Large Memory Nodes

DATA CENTRIC ARCHITECTURE
12PB Perf. Storage: 140GB/s, 200k IOPS
Fast I/O Node-Local NVMe Storage
7PB Ceph Object Storage
High-Performance R&E Networking



LONG-TAIL SCIENCE
Multi-Messenger Astronomy
Genomics
Earth Science
Social Science

INNOVATIVE OPERATIONS
Composable Systems
High-Throughput Computing
Science Gateways
Interactive Computing
Containerized Computing
Cloud Bursting

Overview

- ***NSF funded***:
  - 728, 2-socket AMD-based compute nodes (2.25 GHz EPYC; 64-core/socket). 93,184 compute cores in total.
  - 52 4-way GPU nodes based on V100 w/NVLINK
- ***Industry rack:***
  - 56 2-socket AMD-based compute nodes
  - 4 4-way GPU nodes based on V100 w/NVLINK
- Based on benchmarks we've run, we expect > 2x throughput over Comet (per-core improvement over Haswell, and 2x the core counts)
- Expect a smooth transition from Intel to AMD processors. SDSC team has compiled and run many of the common software packages on Expanse.
- **In production since *December 2020,*** *with operations for 5-years*

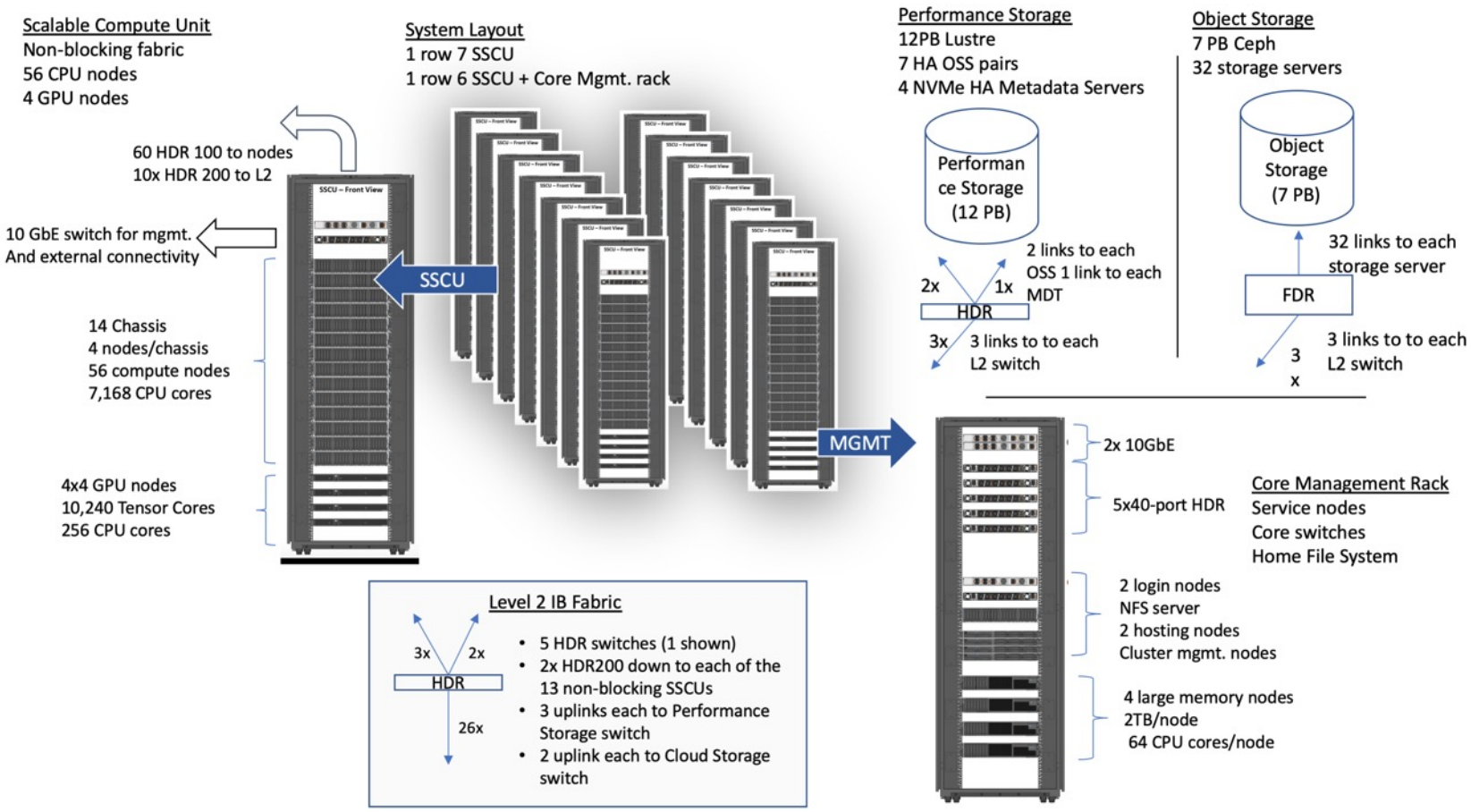
Outline

- Introduction and Overview
- Expanse system architecture
- AMD EPYC Processor Architecture
  - Hardware details
  - NUMA options
  - Applications
- Expanse benchmark results with MVAPICH2/2.3.6
- Summary

Expanse is a heterogeneous architecture designed for high performance, reliability, flexibility, and productivity

System Summary

- *14 SDSC Scalable Compute Units (SSCU)*
- *784 x 2s Standard Compute Nodes*
- *100,352 Compute Cores*
- *200 TB DDR4 Memory*
- *56x 4-way GPU Nodes w/NVLINK*
- *224 V100s*
- *4x 2TB Large Memory Nodes*
- *HDR 100 non-blocking Fabric*
- *12 PB Lustre High Performance Storage*
- *7 PB Ceph Object Storage*
- *1.2 PB on-node NVMe*
- *Dell EMC PowerEdge*
- *Direct Liquid Cooled*



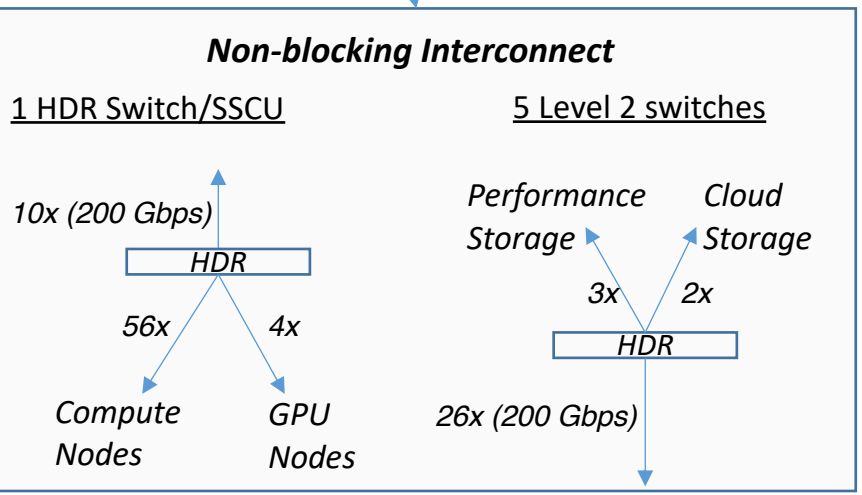
The SSCU is Designed for the Long Tail Job Mix, Maximum Performance, Efficient Systems Support, and Efficient Power and Cooling

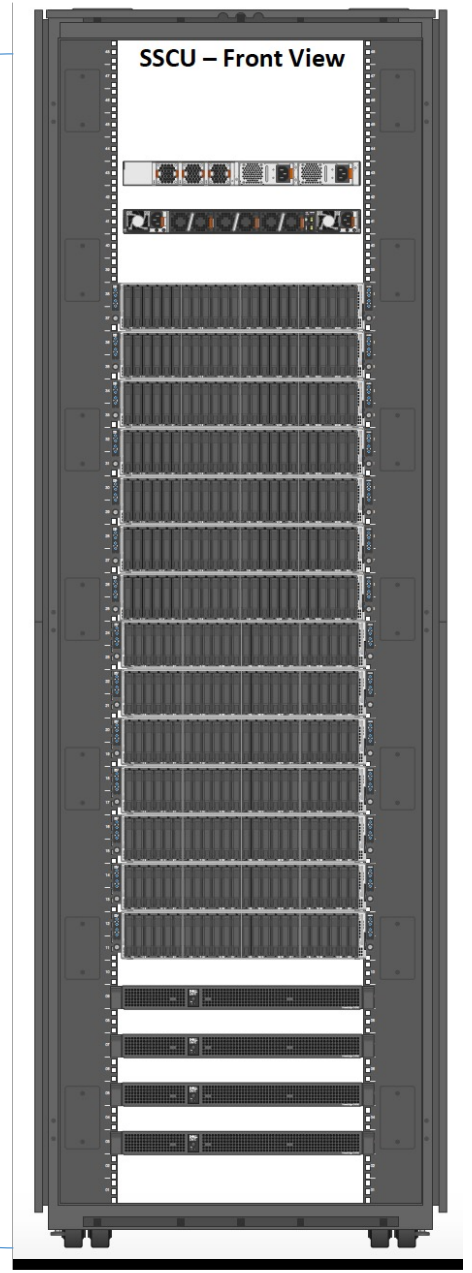
Standard Compute Nodes
- 2x AMD EPYC 7742 @2.25 GHz
- 128 Zen2 CPU cores
- PCIe Gen4
- 256 GB DDR4
- 1 TB NVME

GPU Nodes
- 4x NVIDIA V100/follow-on
- 10,240 Tensor Cores
- 32 GB GDDR
- 1.6 TB NVMe
- Intel CPUs

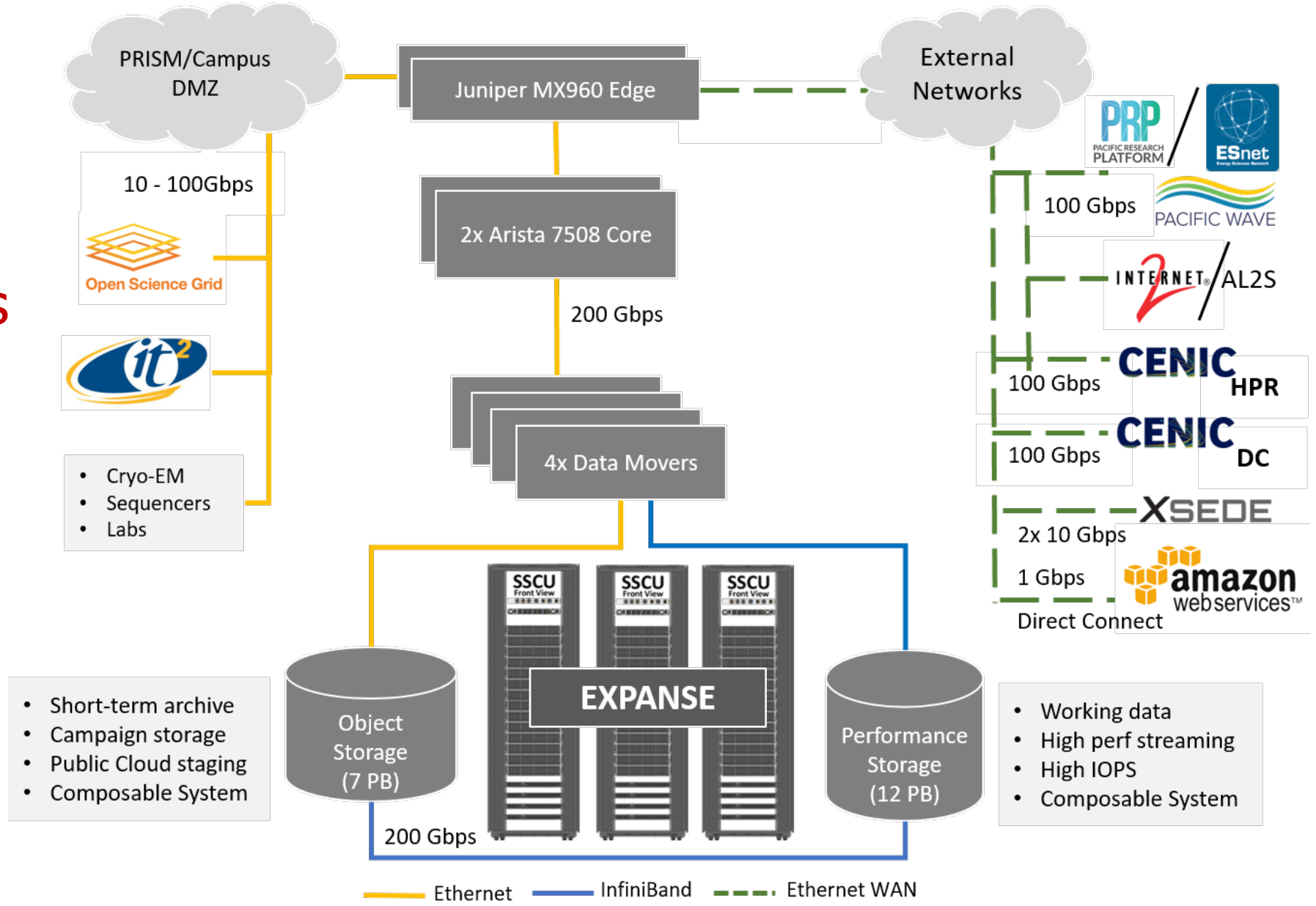
SSCU Components
- 56x CPU nodes
- 7,168 Compute Cores
- 4x GPU nodes
- 1x HDR Switch
- 1x 10GbE Switch
- HDR 100 non-blocking fabric
- Wide rack for serviceability
- Direct Liquid Cooling to CPU nodes

Non-blocking Interconnect





Connectivity to R&E Networks Facilitates Compute and Data Workflows

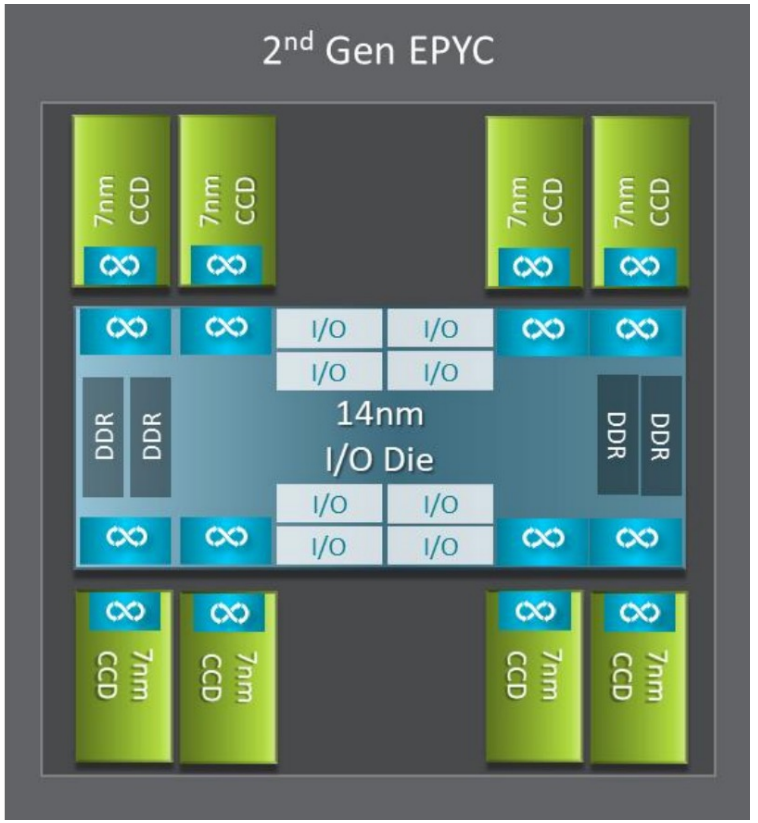


Outline

- Introduction and Overview
- Expanse system architecture
- AMD EPYC Processor Architecture
  - Hardware details
  - NUMA options
  - Applications
- Expanse benchmark results with MVAPICH2/2.3.6
- Summary

AMD EPYC 7742 Processor Architecture

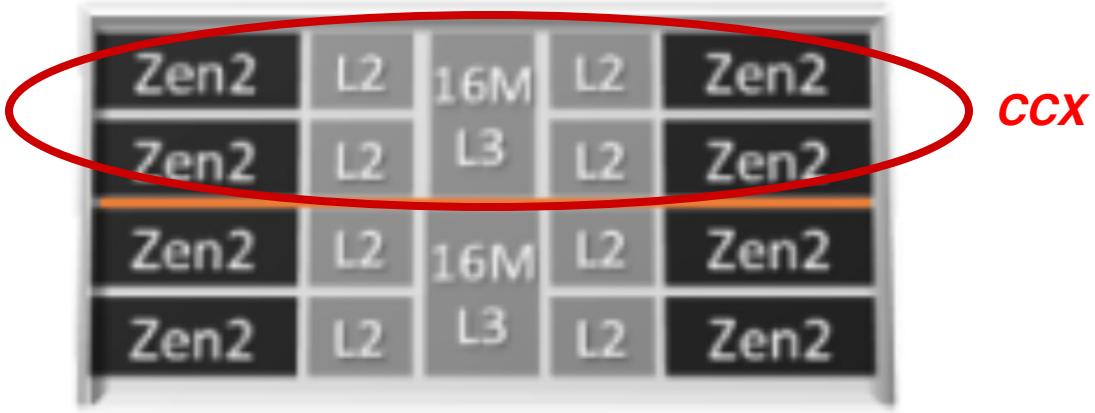
- 8 Core Complex Dies (CCDs).
- CCDs connect to memory, I/O, and each other through the I/O Die.
- 8 memory channels per socket.
- DDR4 memory at 3200MHz.
- PCI Gen4, up to 128 lanes of high speed I/O.
- Memory and I/O can be abstracted into separate quadrants each with 2 DIMM channels and 32 I/O lanes.



Reference: https://developer.amd.com/wp-content/resources/56827-1-0.pdf

AMD EPYC 7742 Processor: Core Complex Die (CCD)

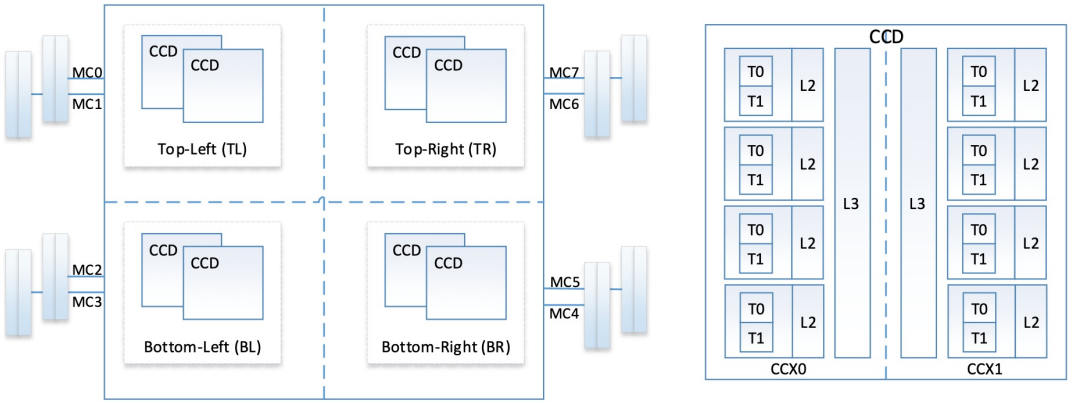
- 2 Core Complexes (CCXs) per CCD
- 4 Zen2 cores in each CCX shared a 16M L3 cache. Total of 16 x 16 = 256MB L3 cache.
- Each core includes a private 512KB L2 cache.



Reference: https://developer.amd.com/wp-content/resources/56827-1-0.pdf

AMD EPYC 7742 Processor : NUMA Nodes Per Socket

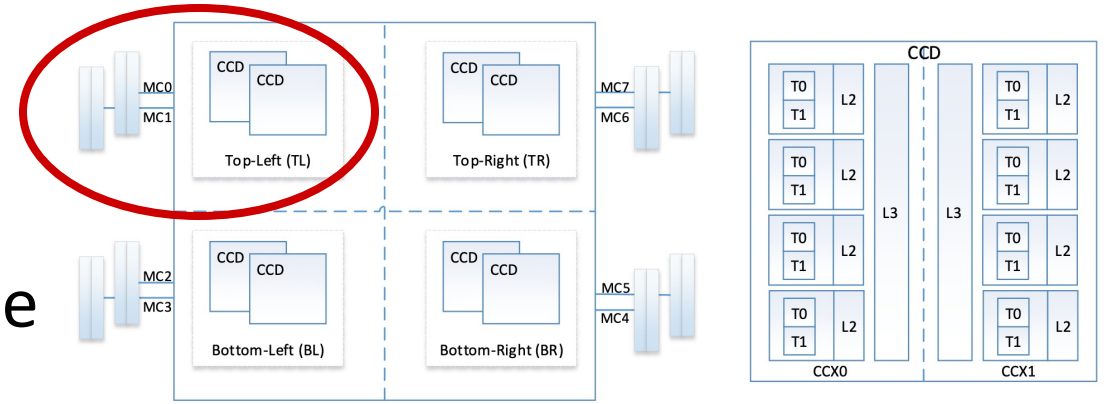
- The four logical quadrants allow the processor to be partitioned into different NUMA domains. Options set in BIOS.
- Domains are designated as NUMA per socket (NPS).
- **NPS4:** Four NUMA domains per socket is the typical HPC configuration.



https://developer.amd.com/wp-content/resources/56338_1.00_pub.pdf

NPS4 Configuration

- The processor is partitioned into four NUMA domains.
- Each logical quadrant is a NUMA domain.
- Memory is interleaved across the two memory channels
- PCIe devices will be local to one of four NUMA domains (the IO die that has the PCIe root for the device)
- ***This is the typical HPC configuration*** as workload is NUMA aware, ranks and memory can be pinned to cores and NUMA nodes.



https://developer.amd.com/wp-content/resources/56338_1.00_pub.pdf

Software Stack

- Expanse supports a broad application base with installs and modules for commonly used packages in bioinformatics, molecular dynamics, machine learning, quantum chemistry, structural mechanics, and visualization.

- Most of the application stack on Comet has been replicated on Expanse.

- Primarily Spack based installs. Active engagement with AMD under HPC User Forum activities to help with optimal Spack recipes.

- Continued support for Singularity based containerization on Expanse.

Benchmarks of Applications on Expanse

- Benchmarked CPU Applications: GROMACS, NAMD, NEURON, OpenFOAM, Quantum Espresso, RAxML, WRF, and ASTRAL.
  - MPI, Hybrid MPI/OpenMP, and Hybrid MPI/Pthreads cases. Compilers used included AOCC, gnu, and Intel.
  - Results on Expanse show performance ranges from matching on a per core basis to 1.8X faster on a per core basis compared to Comet.
  - Overall throughput is expected to be easily more than 2X of Comet.

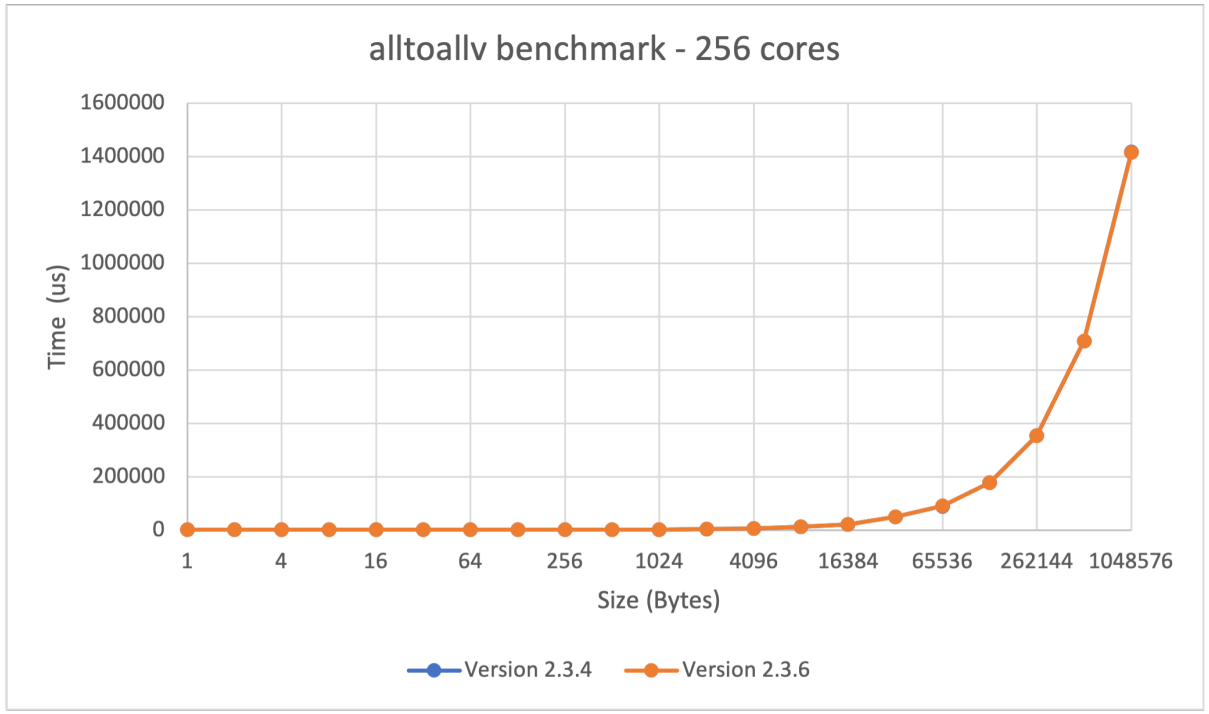
- Benchmarked GPU Applications: NAMD, AMBER, TensorFlow, PyTorch, MXNET, GROMACS, and BEAST
  - Results on Expanse show >1.5X per GPU improvement over the Comet P100 nodes.

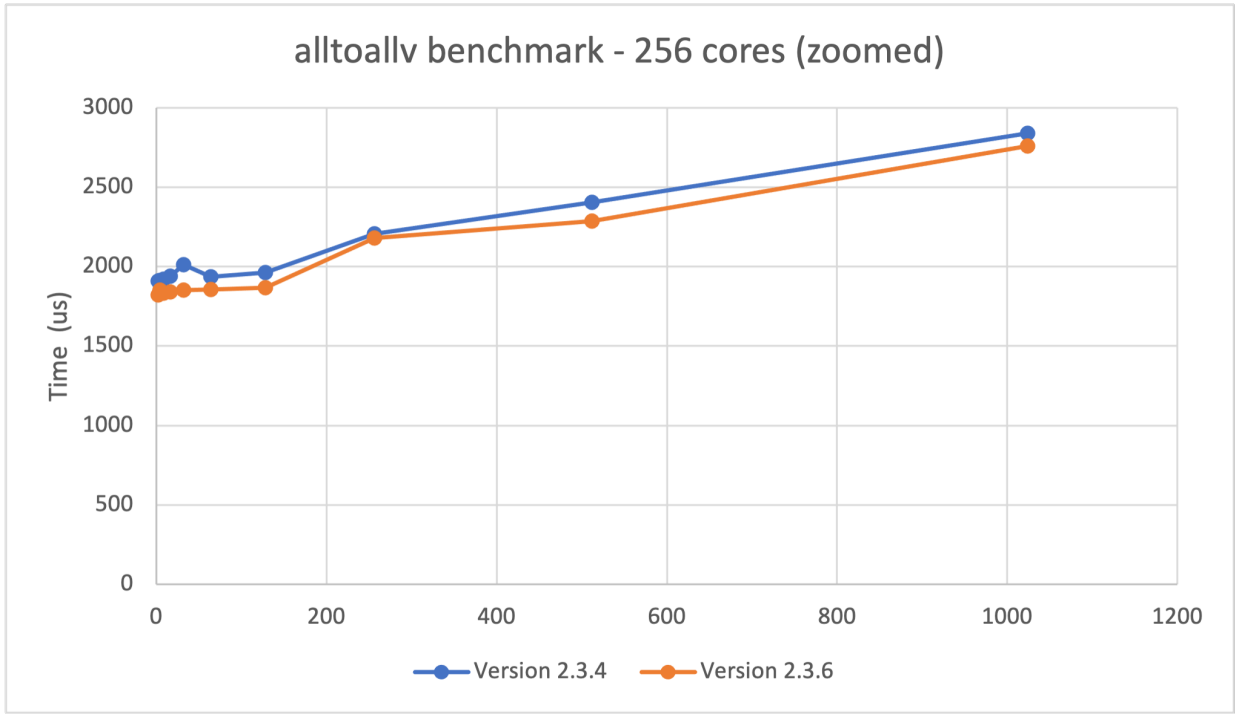
Outline

- Introduction and Overview
- Expanse system architecture
- AMD EPYC Processor Architecture
  - Hardware details
  - NUMA options
  - Applications
- Expanse benchmark results with MVAPICH2/2.3.6
- Summary

OSU alltoallv benchmark: 256 cores

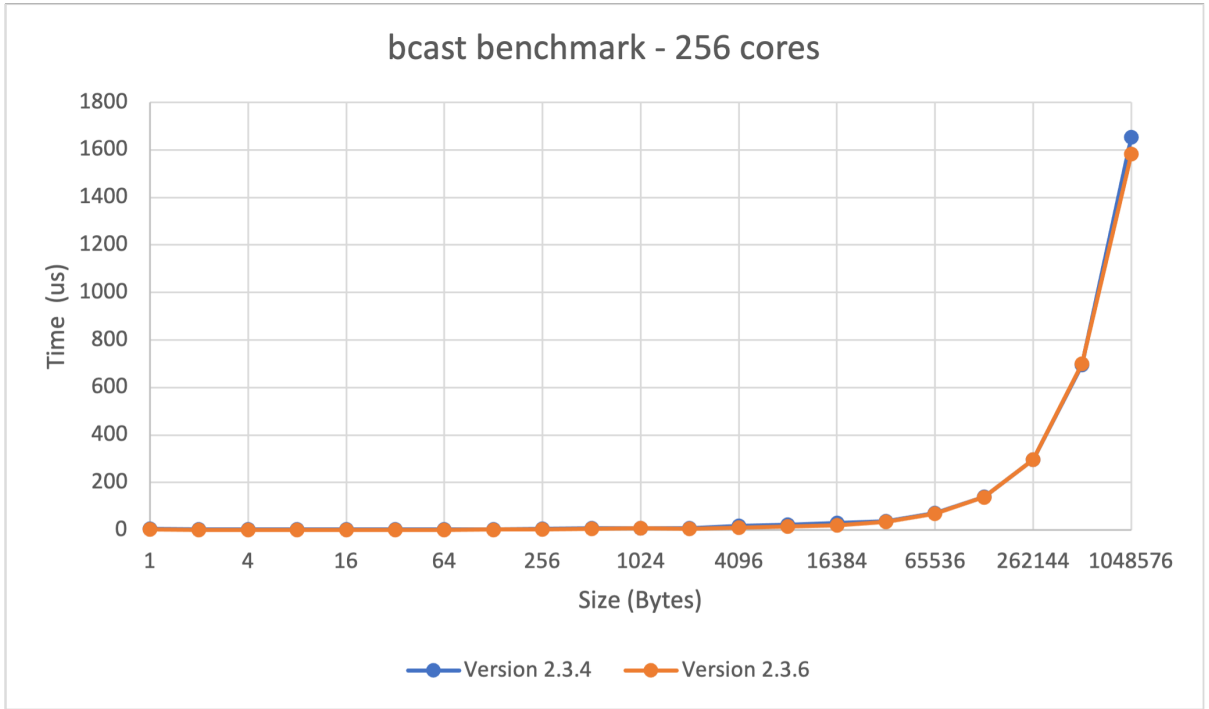
MVAPICH2 version 2.3.4 vs 2.3.6

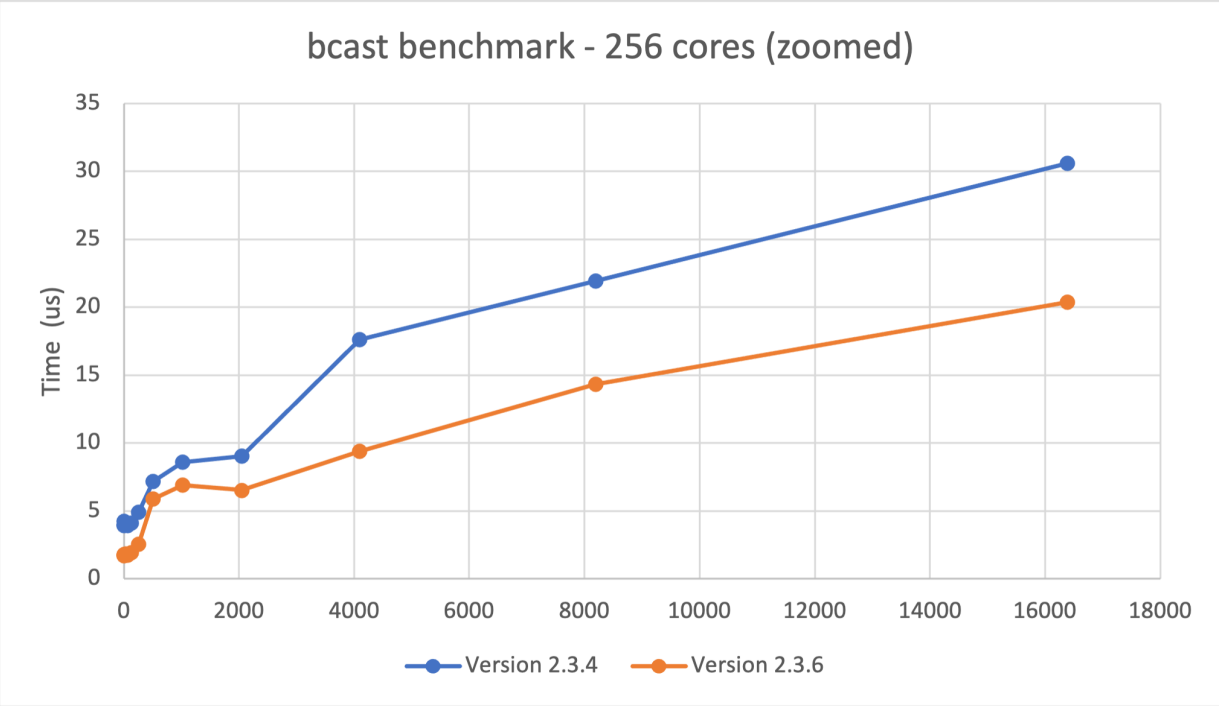




OSU bcast benchmark: 256 cores

MVAPICH2 version 2.3.4 vs 2.3.6

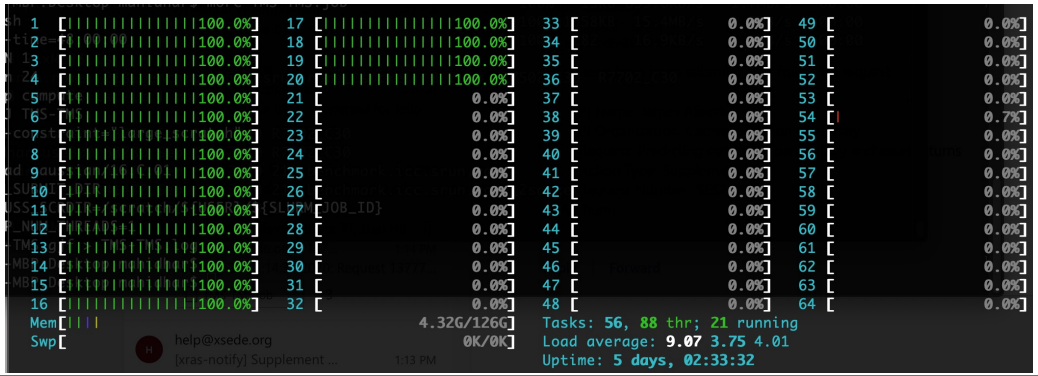


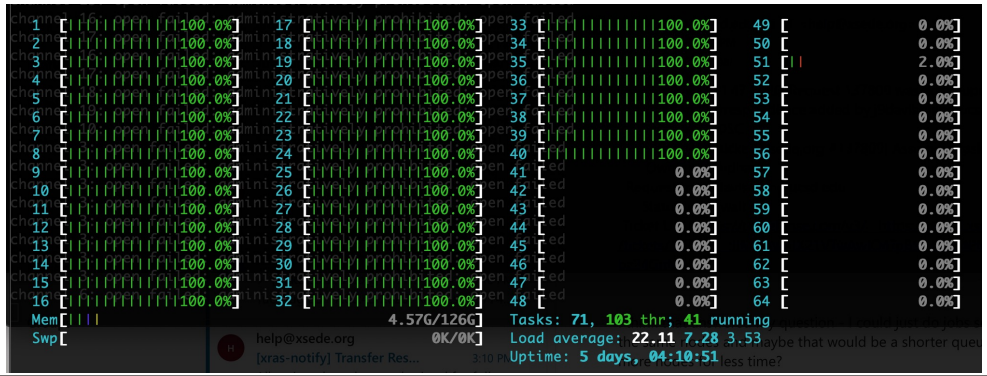


RAxML Benchmark: All-in-one analysis: 218 taxa, 2,294 DNA characters, 1,846 patterns, 100 bootstraps (MPI + Pthreads)

Build: Intel Compiler + MVAPICH2/2.3.6

Total tasks	Comet (s)	Stampede2 (s)	Expanse-Dev (s)	Expanse (s) (MV 2.3.6)
10 (5 MPI x 2 Pthreads)	925	610	514	**410**
20 (5 MPI x 4 Pthreads)	542	363	292	**249**
30 (10 MPI x 3 Pthreads)	433	332	247	**199**
40 (10 MPI x 4 Pthreads)	341	300	201	**171**



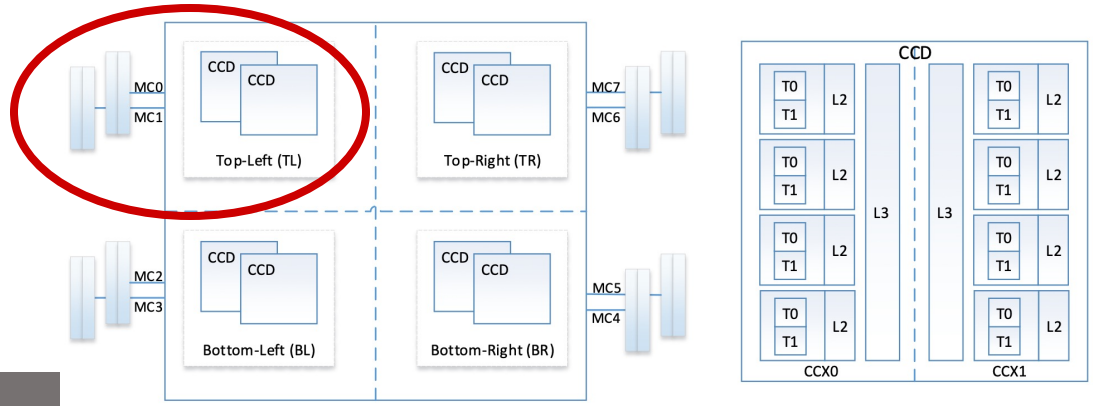


NEURON Benchmark:

Large-scale model of olfactory bulb: 10,500 cells, 40,000 timesteps

Build: Intel + Intel MPI compilers

Total #MPI Tasks	Expanse-Dev (Compact)	Expanse-Dev (Best Memory BW)
16	5004	1781
32	2336	1321
64	1130	1130



NEURON Benchmark:
Large-scale model of olfactory bulb: 10,500 cells, 40K timesteps

#MPI Tasks	Comet	Test Cluster AMD Rome, EDR IB	Expanse MVAPICH2/2.3.5	Expanse MAVPICH2/2.3.6
96	522 s	525 s	539 s	537 s
192	264 s	220 s	211 s	211 s
384	120 s	68 s	65 s	66 s
768	**53 s**	**35 s**	**36 s**	**29 s**

Comet P100 node architecture

```
        GPU0    GPU1    GPU2    GPU3    mlx4_0  CPU Affinity
GPU0     X      PIX     SOC     SOC     PHB     0-0,2-2,4-4,6-6,8-8,10-10,12-12,14-14,16-16,18-18,20-20,22-22,24-24,26-26
GPU1    PIX      X      SOC     SOC     PHB     0-0,2-2,4-4,6-6,8-8,10-10,12-12,14-14,16-16,18-18,20-20,22-22,24-24,26-26
GPU2    SOC     SOC      X      PIX     SOC     1-1,3-3,5-5,7-7,9-9,11-11,13-13,15-15,17-17,19-19,21-21,23-23,25-25,27-27
GPU3    SOC     SOC     PIX      X      SOC     1-1,3-3,5-5,7-7,9-9,11-11,13-13,15-15,17-17,19-19,21-21,23-23,25-25,27-27
mlx4_0  PHB     PHB     SOC     SOC      X

Legend:

  X    = Self
  SOC  = Connection traversing PCIe as well as the SMP link between CPU sockets(e.g. QPI)
  PHB  = Connection traversing PCIe as well as a PCIe Host Bridge (typically the CPU)
  PXB  = Connection traversing multiple PCIe switches (without traversing the PCIe Host Bridge)
  PIX  = Connection traversing a single PCIe switch
  NV#  = Connection traversing a bonded set of # NVLinks
```

- *4 GPUs per node*
- *GPUs (0,1) and (2,3) can do P2P communication*
- *Mellanox InfiniBand adapter associated with first socket (GPUs 0, 1)*

Expanse GPU Node Architecture

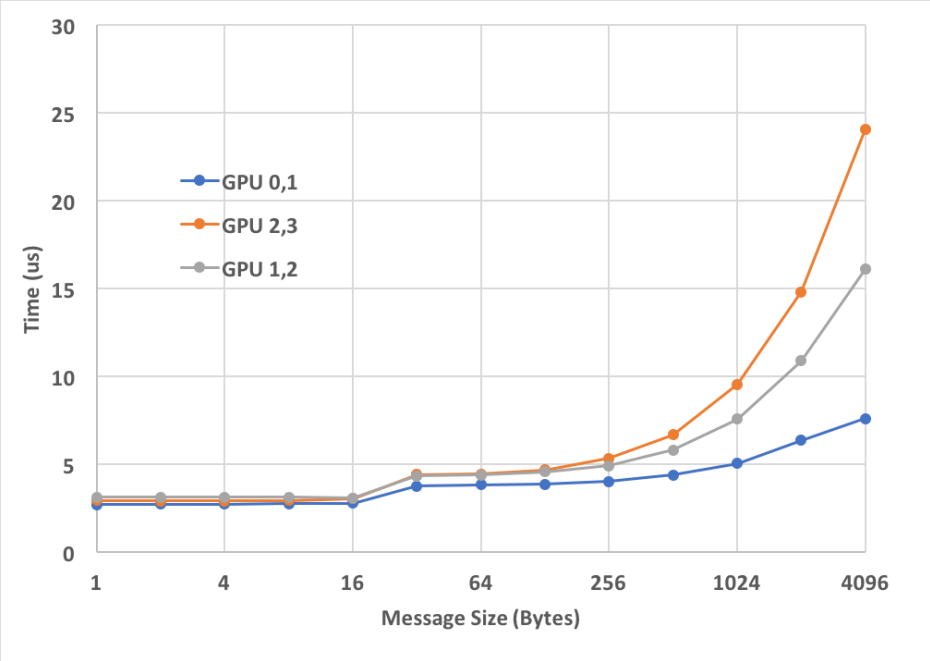
- 4 V100 32GB SMX2 GPUs
- 384 GB RAM, 1.6 TB PCIe NVMe
- 2 Intel Xeon 6248 CPUs
- Topology:

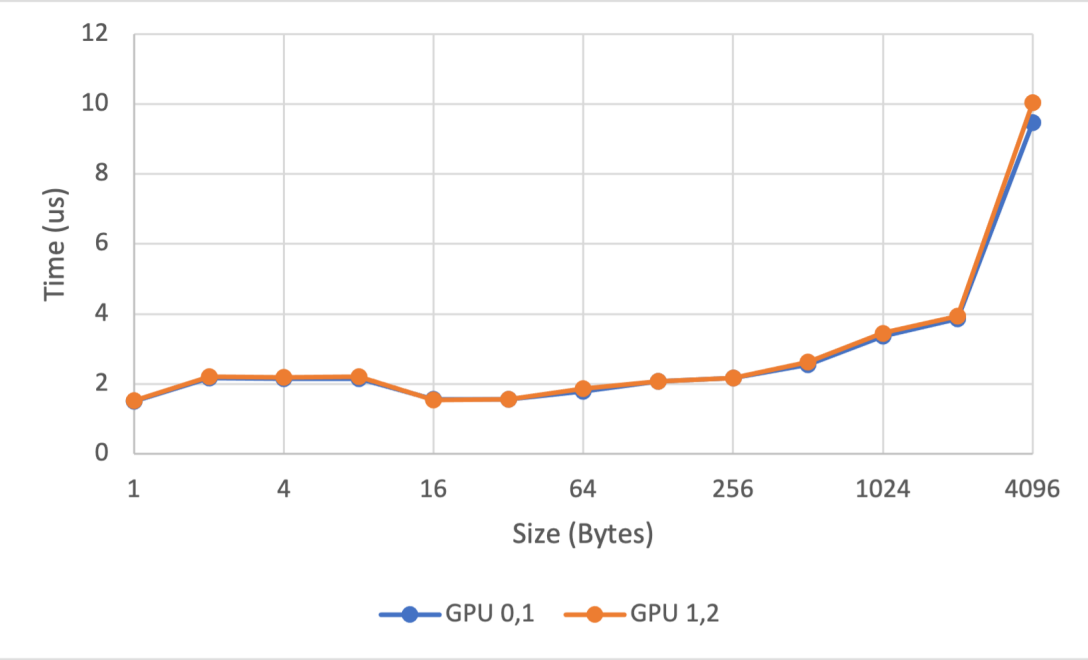
```
        GPU0    GPU1    GPU2    GPU3    mlx5_0  CPU Affinity
GPU0     X      NV2     NV2     NV2     SYS     0-0,4-4,8-8,12-12,16-16,20-20,24-24,28-28,32-32,36-36
GPU1    NV2      X      NV2     NV2     SYS     0-0,4-4,8-8,12-12,16-16,20-20,24-24,28-28,32-32,36-36
GPU2    NV2     NV2      X      NV2     SYS     1-1,5-5,9-9,13-13,17-17,21-21,25-25,29-29,33-33,37-37
GPU3    NV2     NV2     NV2      X      SYS     1-1,5-5,9-9,13-13,17-17,21-21,25-25,29-29,33-33,37-37
mlx5_0  SYS     SYS     SYS     SYS      X

Legend:

  X    = Self
  SYS  = Connection traversing PCIe as well as the SMP interconnect between NUMA nodes (e.g., QPI/UPI)
  NODE = Connection traversing PCIe as well as the interconnect between PCIe Host Bridges within a NUMA node
  PHB  = Connection traversing PCIe as well as a PCIe Host Bridge (typically the CPU)
  PXB  = Connection traversing multiple PCIe bridges (without traversing the PCIe Host Bridge)
  PIX  = Connection traversing at most a single PCIe bridge
  NV#  = Connection traversing a bonded set of # NVLinks
```

OSU Latency (osu_latency) Benchmark
Intra-node, P100 nodes on Comet, V100 nodes on Expanse

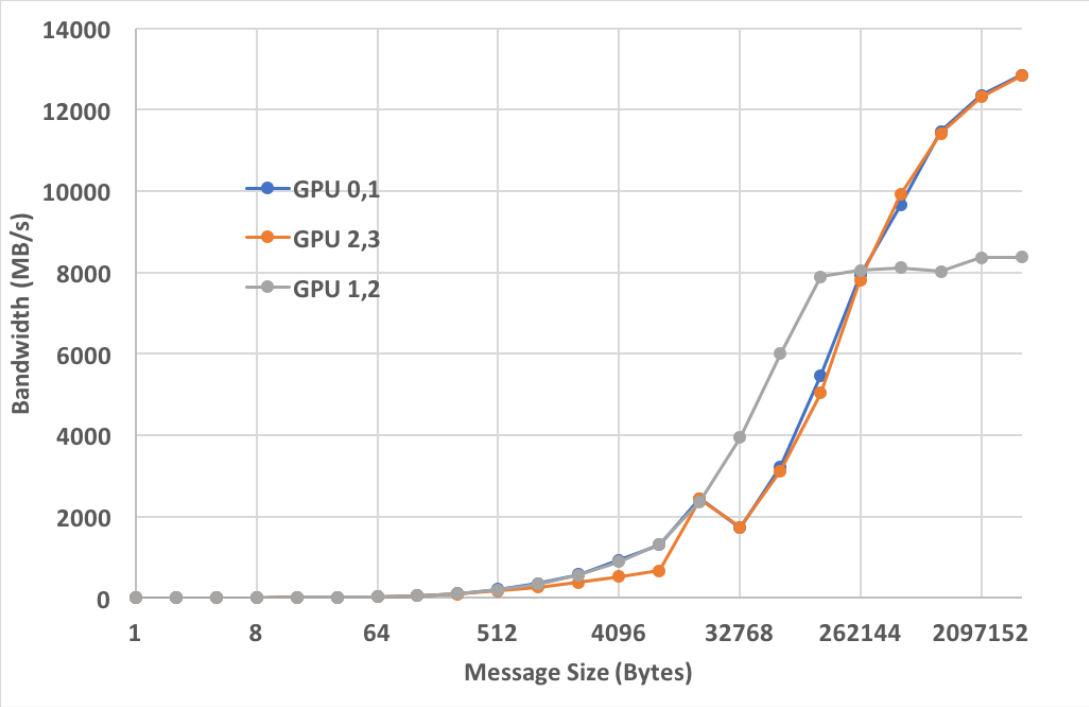


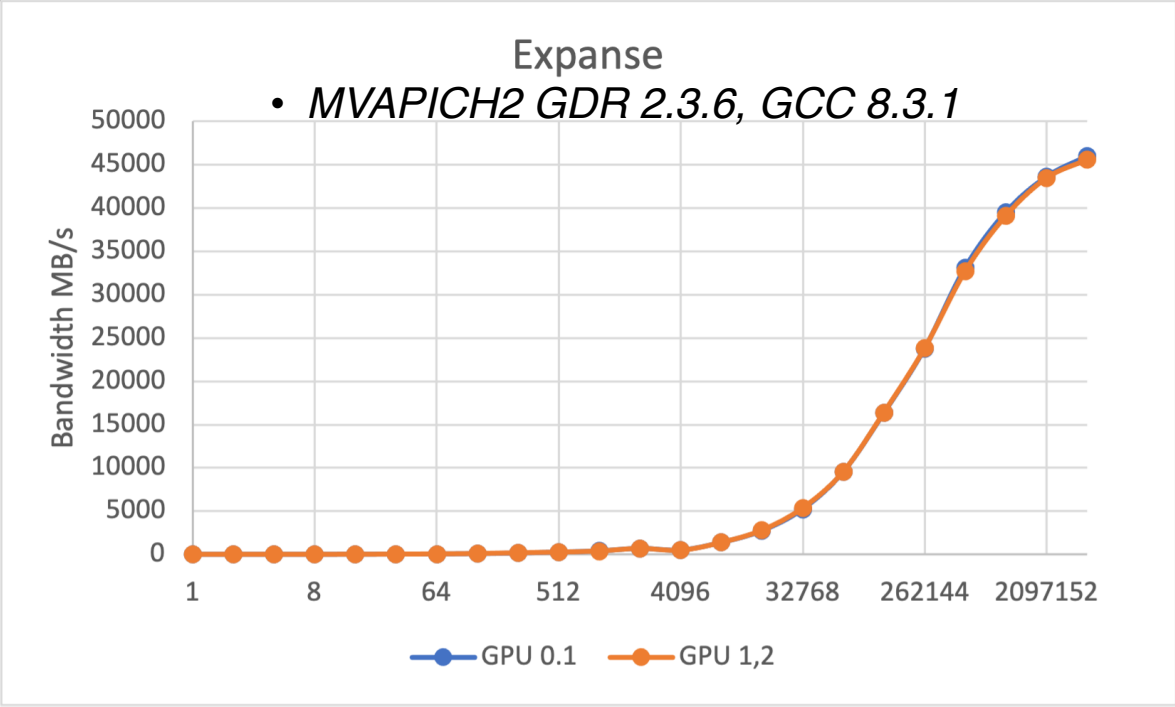


- *COMET - P100 nodes*
- *Latency between GPU 0 , GPU 1: 2.73 µs*
- *Latency between GPU 2 , GPU 3: 2.95 µs*
- *Latency between GPU 1 , GPU 2: 3.13 µs*

- *Expanse - V100 nodes*
- *Latency between GPU 0 , GPU 1: 1.51 µs*
- *Latency between GPU 1 , GPU 2: 1.53 µs*
- *MVAPICH2 GDR 2.3.6, GCC 8.3.1*

OSU Bandwidth (osu_bw) Benchmark
Intra-node, P100 nodes on Comet, V100 nodes on Expanse





Summary

- Expanse will provide a substantial increase in the performance and throughput compared to the highly successful, NSF-funded Comet supercomputer.
- 728, 2-socket AMD-based compute nodes (2.25 GHz EPYC; 64-cores/socket) and 52 4-way GPU nodes based on V100 w/NVLINK. Industry rack has an additional 56 compute nodes and 4 GPU nodes.
- HDR InfiniBand interconnect – HDR100 to the nodes and HDR200 switches.
- MVAPICH2 2.3.6 has shown improvements in performance for several benchmarks and applications on AMD CPUs. Application testing with MVAPICH2 GDR ongoing.
- In production since December 2020. Follow all things Expanse at https://expanse.sdsc.edu !

Thank you to our collaborators, partners, users, and the SDSC team!

Ilkay Altintas
Haisong Cai
Amit Chourasia
Trevor Cooper
Jerry Greenberg
Eva Hocks
Tom Hutton
Christopher Irving
Marty Kandes
Amit Majumdar
Dima Mishin
Sonia Nayak
Mike Norman
Wayne Pfeiffer
Scott Sakai
Fernando Silva
Bob Sinkovits
Subha Sivagnanam
Michele Strong
Shawn Strande
Mahidhar Tatineni
Mary Thomas
Nicole Wolter
Frank Wuerthwein

EXPANSE
COMPUTING WITHOUT BOUNDARIES

SAN DIEGO SUPERCOMPUTER CENTER

IN PRODUCTION OCTOBER 2020