Performance of Applications using MVAPICH2 and MVAPICH2-GDR on SDSC's Expanse Supercomputer

MVAPICH2 User Group (MUG) Meeting
August 23, 2022

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
- Microbenchmarks
- Applications using MVAPICH2, MVAPICH2-GDR on Expanse
  - Summary of MVAPICH2 based installs
  - Benchmark results for LAMMPS, RAxML, Neuron
- Summary

Expanse: Computing Without Boundaries: Cyberinfrastructure for the Long Tail of Science

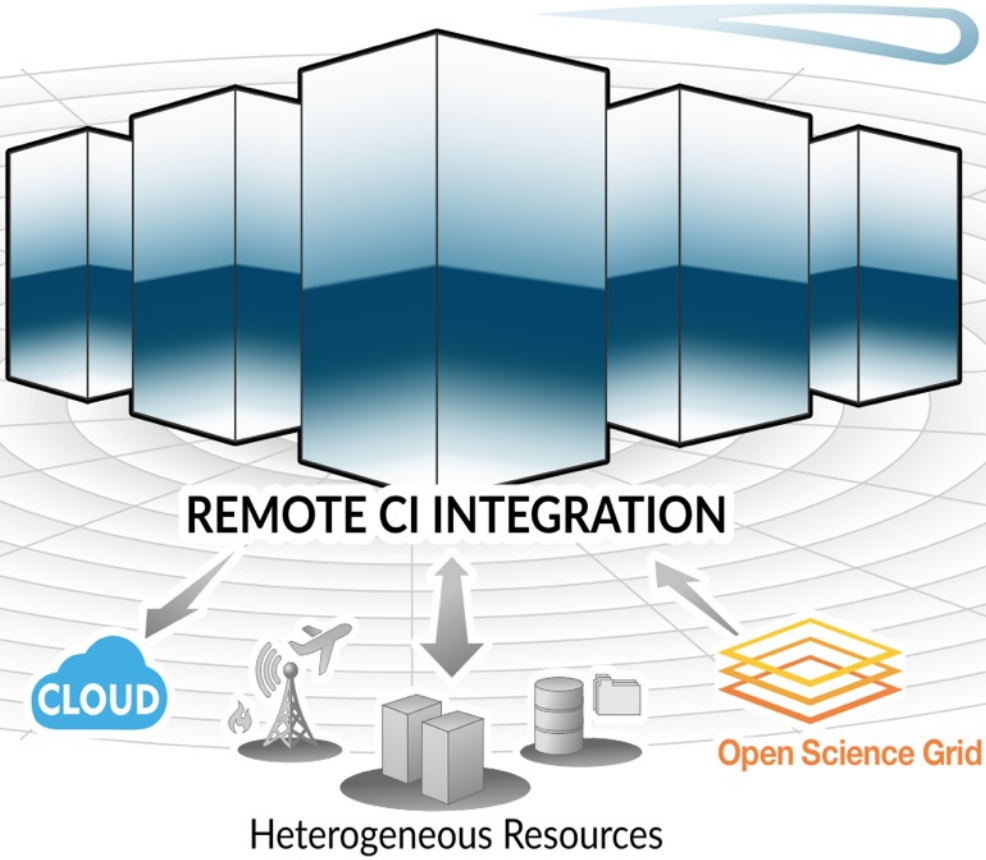
- NSF Solicitation 19-534: Advanced Computing Systems & Services: Adapting to the Rapid Evolution of Science and Engineering Research
- Category 1: Capacity System, NSF Award # 1928224
- NSF Program Officer: Robert Chadduck
- PIs: Mike Norman (PI), Ilkay Altintas, Amit Majumdar, Mahidhar Tatineni, Shawn 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

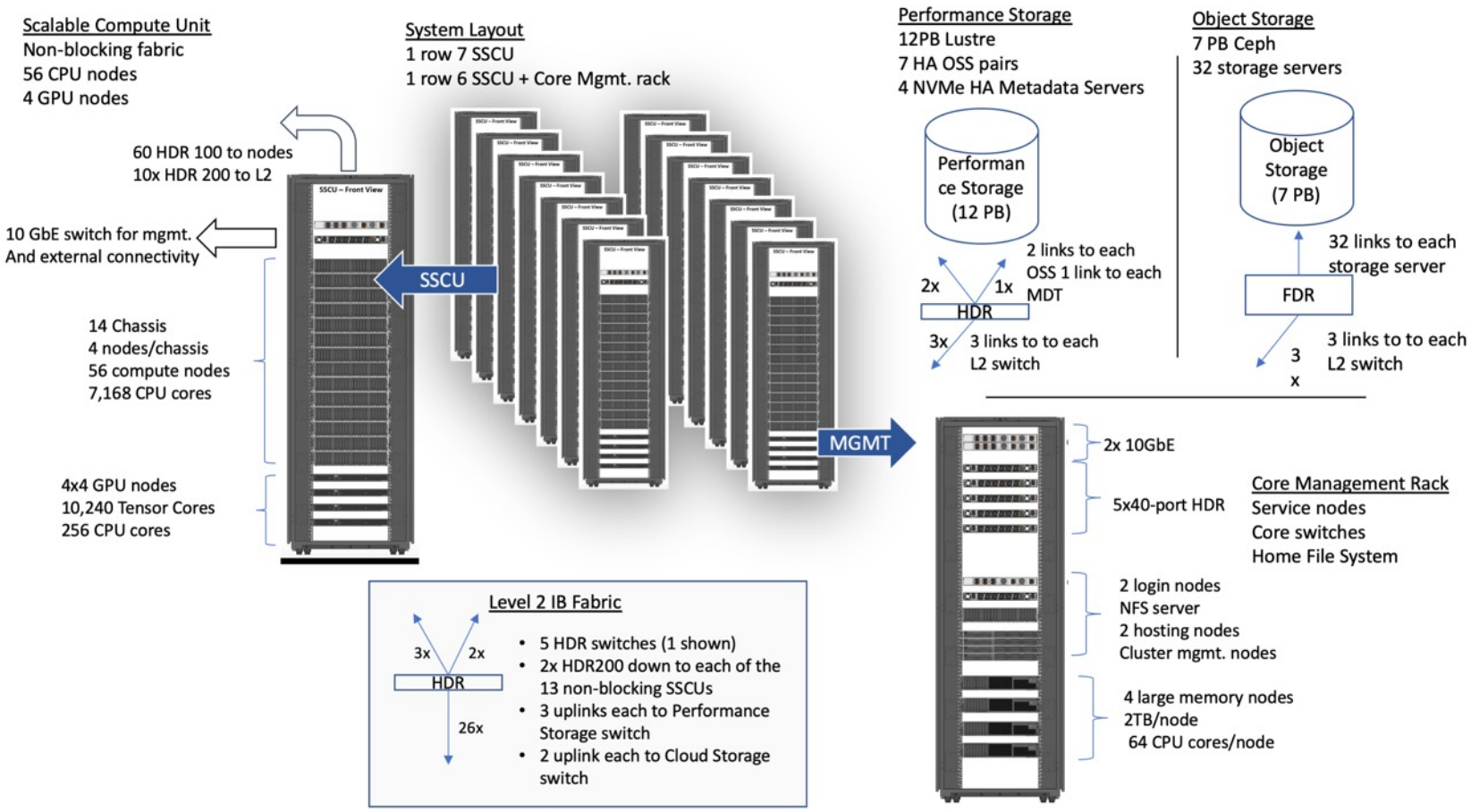
Outline

- Introduction and Overview
- Expanse system architecture
- AMD EPYC Processor Architecture
  - Hardware details
  - NUMA options
- Microbenchmarks
- Applications using MVAPICH2, MVAPICH2-GDR on Expanse
  - Summary of MVAPICH2 based installs
  - Benchmark results for LAMMPS, RAxML, Neuron
- 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*



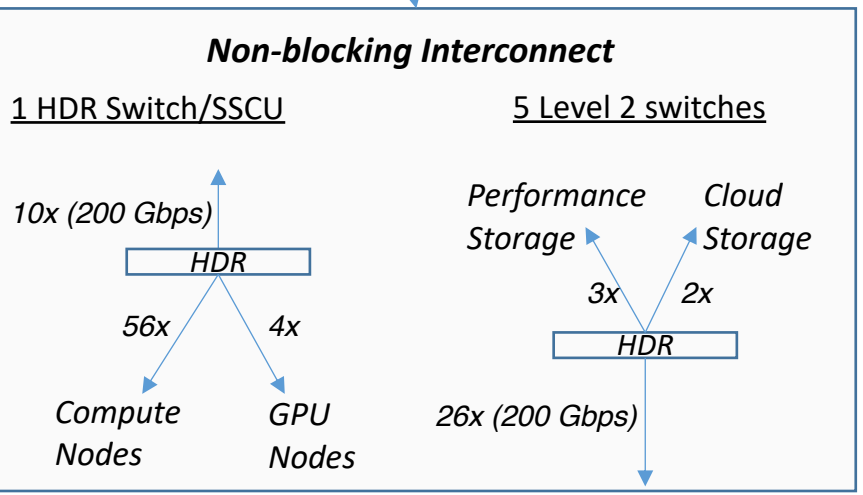
The SSCU is Designed for the Long Tail Job Mix, Maximum Performance, Efficient Systems Support, and Efficient Power and Cooling

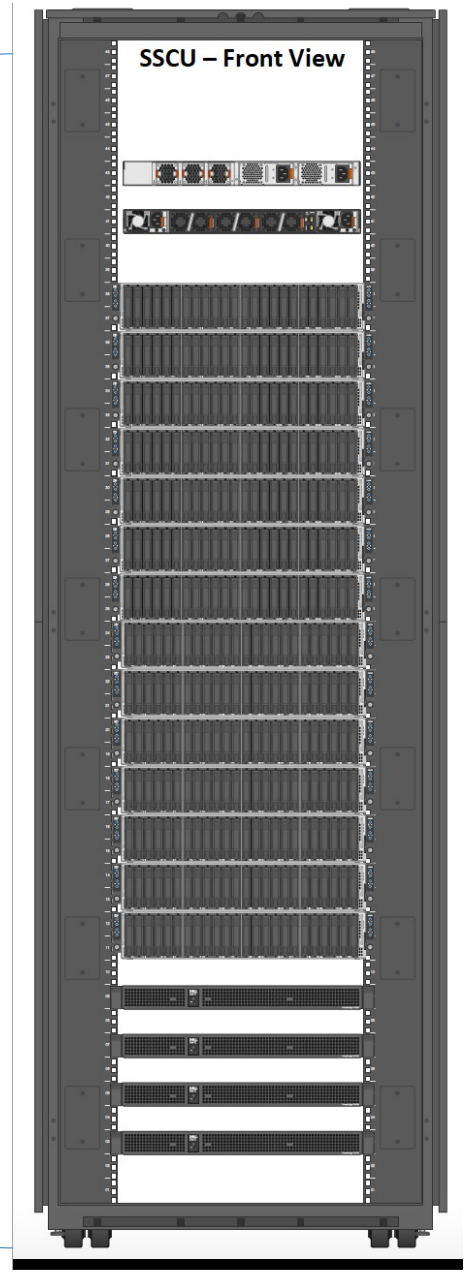
Standard Compute Nodes
- *2x AMD EPYC 7742 @2.25 GHz*
- *128 Zen2 CPU cores*
- *PCIe Gen4*
- *256 GB DDR4*
- *1 TB NVME*

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



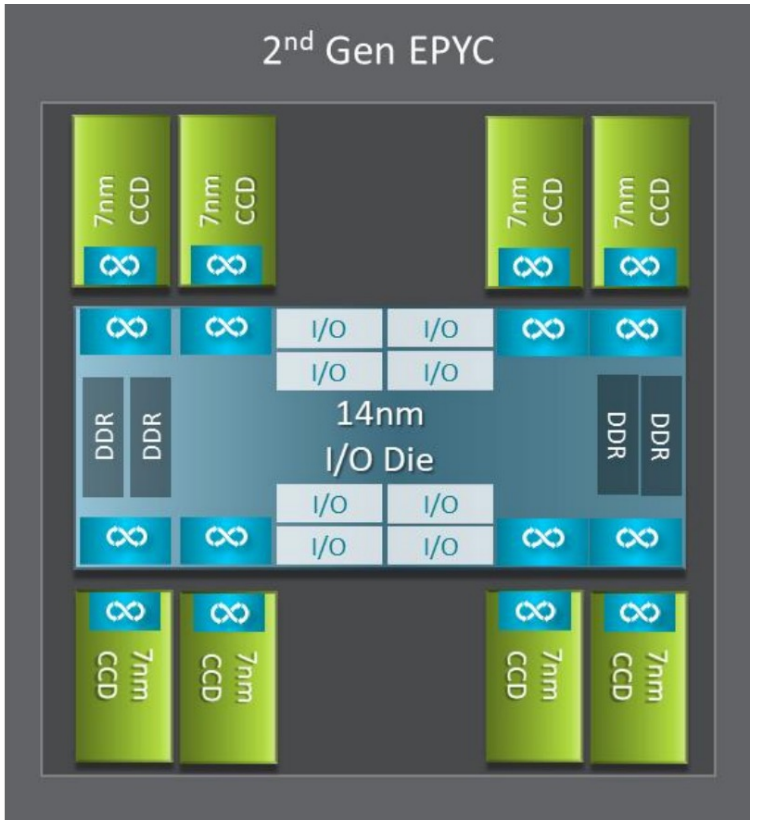


Outline

- Introduction and Overview
- Expanse system architecture
- AMD EPYC Processor Architecture
  - Hardware details
  - NUMA options
- Microbenchmarks
- Applications using MVAPICH2, MVAPICH2-GDR on Expanse
  - Summary of MVAPICH2 based installs
  - Benchmark results for LAMMPS, RAxML, Neuron
- Summary

AMD EPYC 7742 Processor Architecture

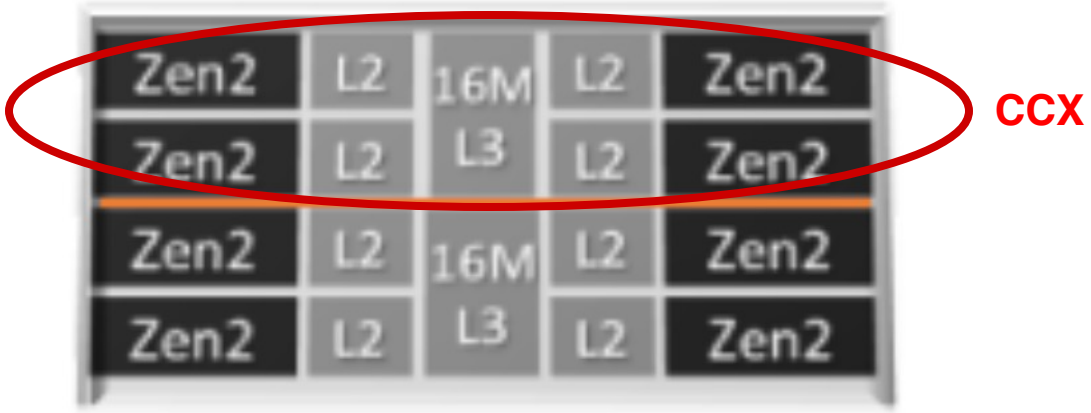
- 8 Core Complex Dies (CCDs).
- CCDs connect to memory, I/O, and each other through the I/O Die.
- 8 memory channels per socket.
- DDR4 memory at 3200MHz.
- PCI Gen4, up to 128 lanes of high speed I/O.
- Memory and I/O can be abstracted into separate quadrants each with 2 DIMM channels and 32 I/O lanes.



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

AMD EPYC 7742 Processor: Core Complex Die (CCD)

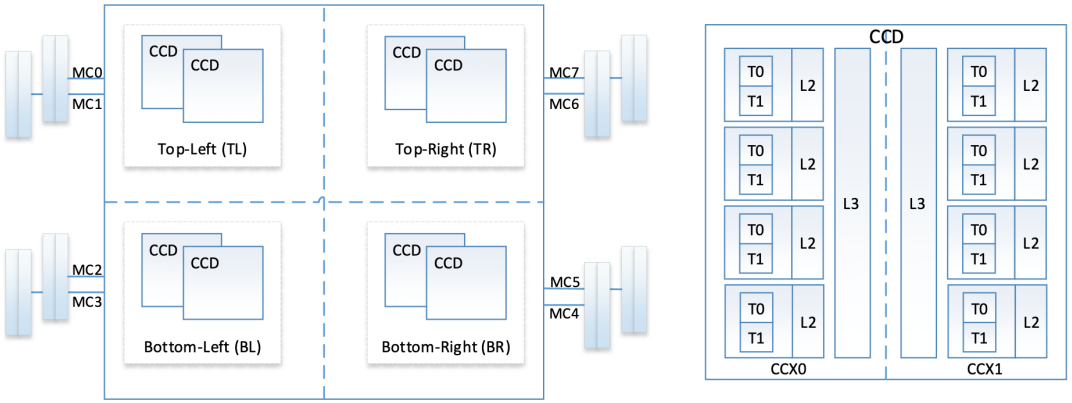
- 2 Core Complexes (CCXs) per CCD
- 4 Zen2 cores in each CCX shared a 16M L3 cache. Total of 16 x 16 = 256MB L3 cache.
- Each core includes a private 512KB L2 cache.



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

AMD EPYC 7742 Processor : NUMA Nodes Per Socket

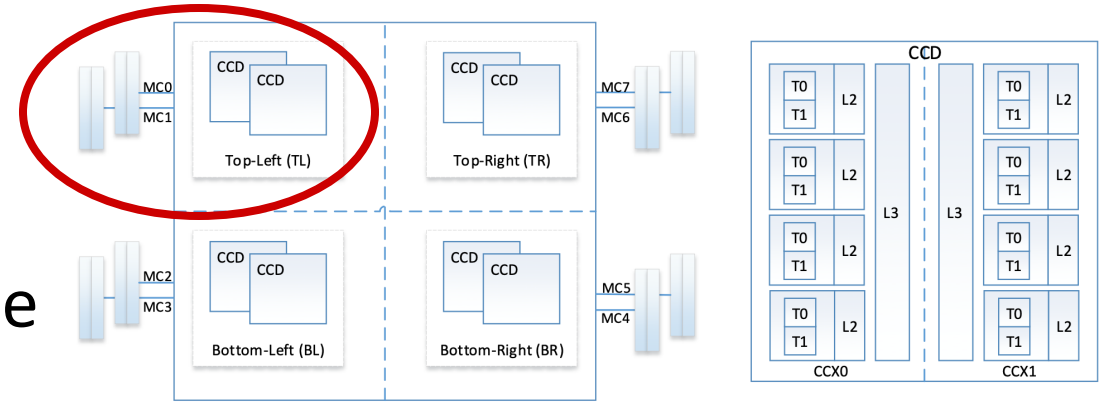
- The four logical quadrants allow the processor to be partitioned into different NUMA domains. Options set in BIOS.
- Domains are designated as NUMA per socket (NPS).
- **NPS4:** Four NUMA domains per socket is the typical HPC configuration.



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.



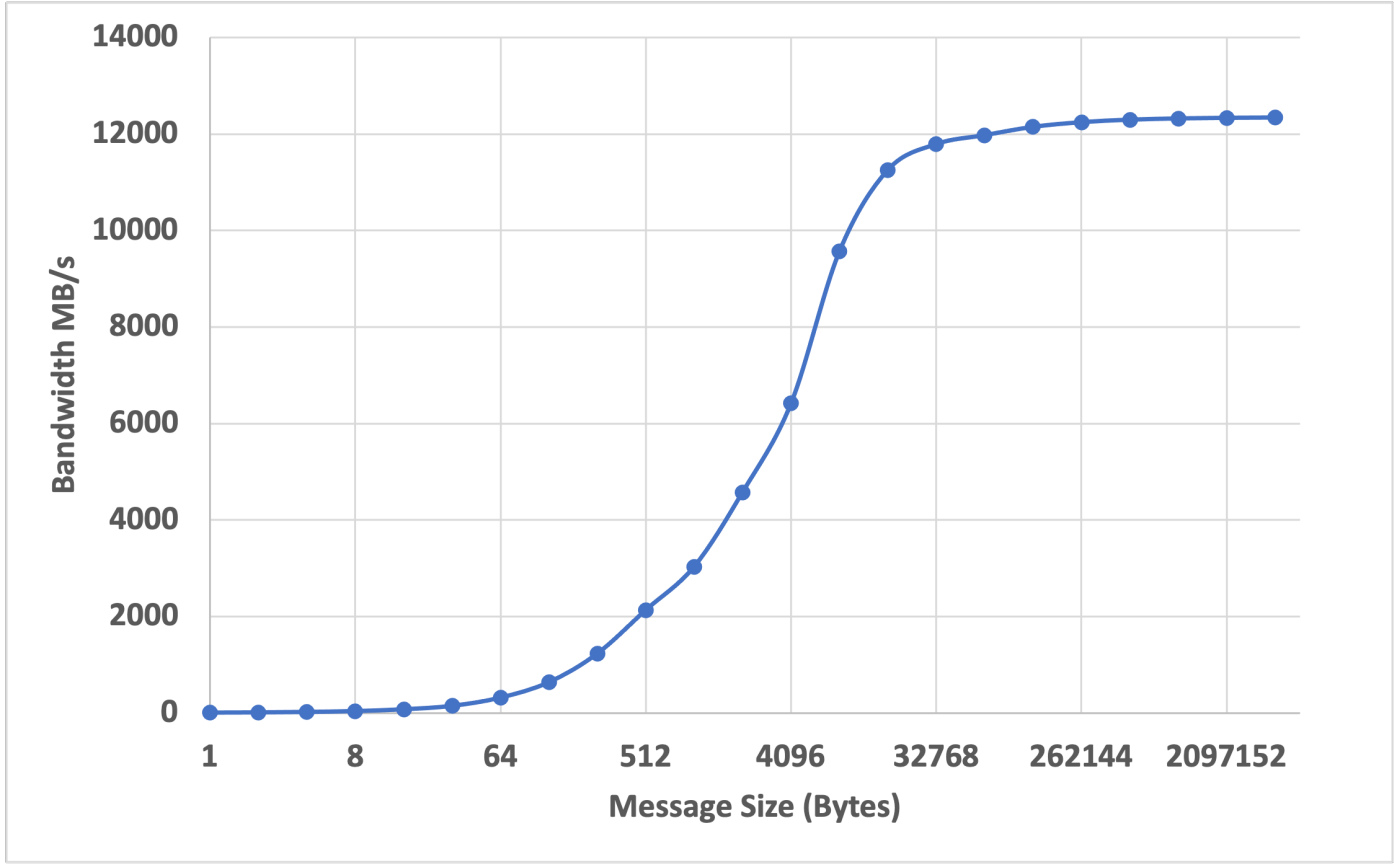
https://developer.amd.com/wp-content/resources/56338_1.00_pub.pdf

Outline

- Introduction and Overview
- Expanse system architecture
- AMD EPYC Processor Architecture
  - Hardware details
  - NUMA options
- Microbenchmarks
- Applications using MVAPICH2, MVAPICH2-GDR on Expanse
  - Summary of MVAPICH2 based installs
  - Benchmark results for LAMMPS, RAxML, Neuron
- Summary

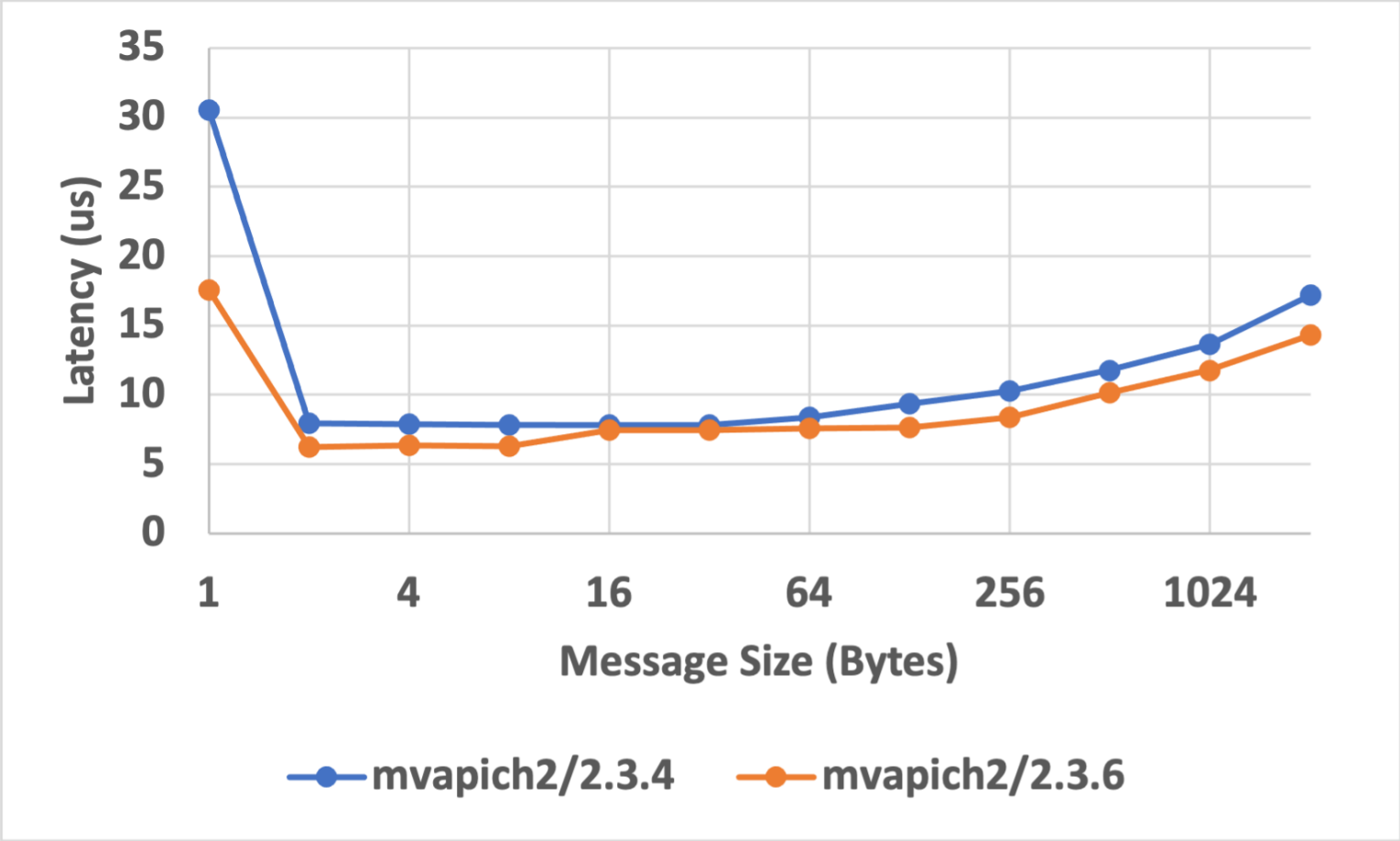
OSU Bandwidth Benchmark (osu_bw): Inter-node test

MVAPICH2 version 2.3.7



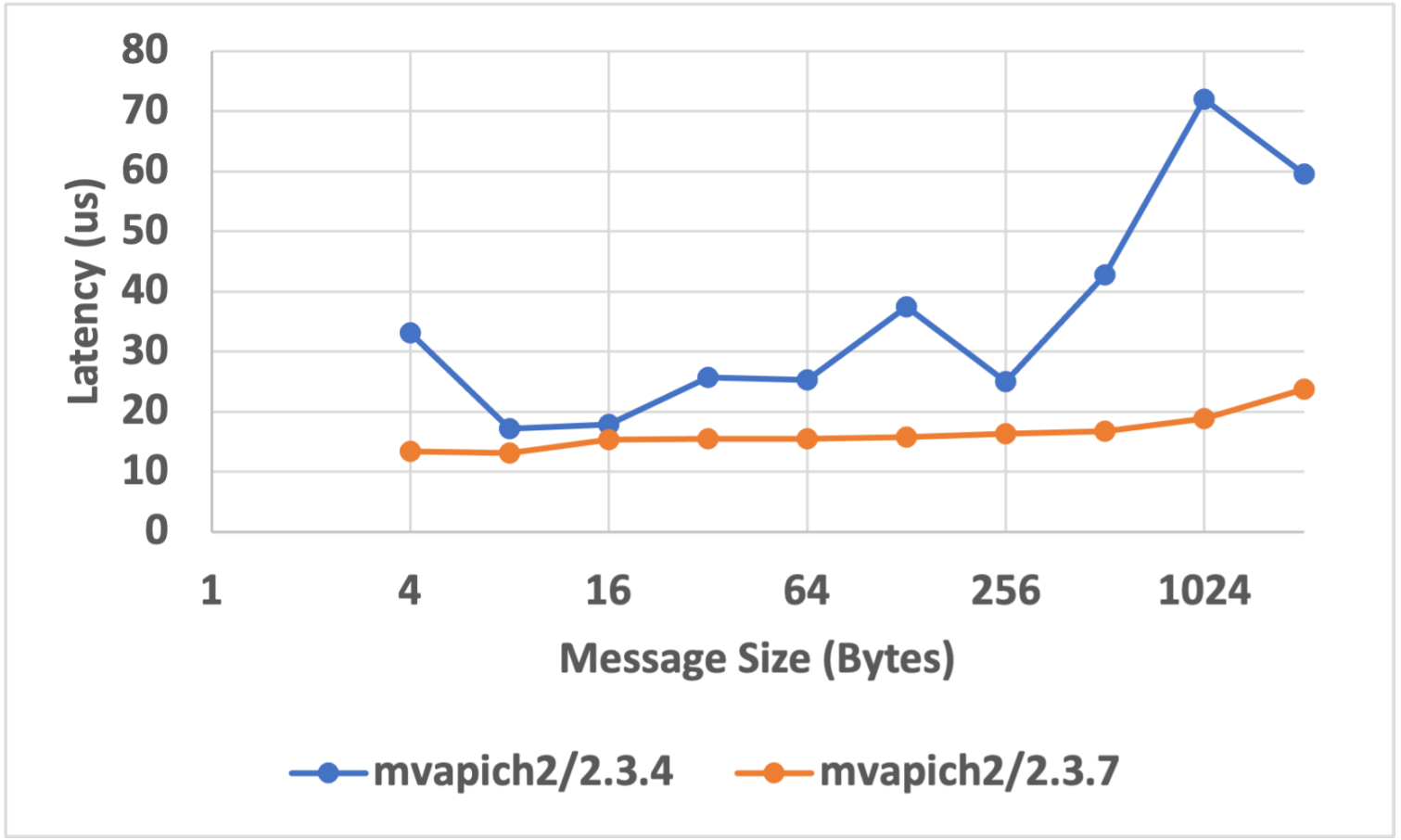
OSU bcast benchmark (osu_bcast): 2048 cores

MVAPICH2 version 2.3.4 vs 2.3.7



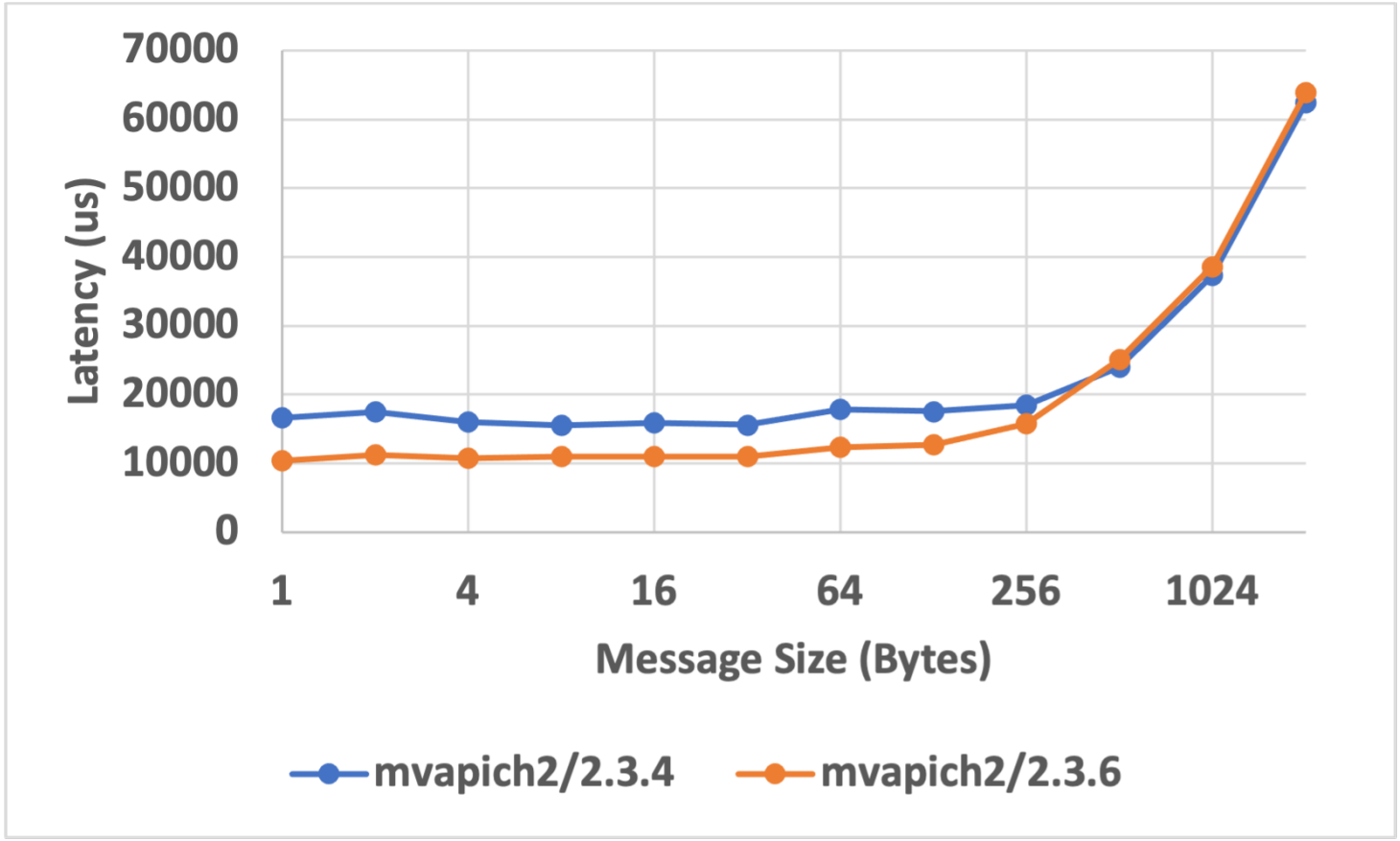
OSU Allreduce benchmark (osu_allreduce): 2048 cores

MVAPICH2 version 2.3.4 vs 2.3.7



OSU Alltoallv benchmark (osu_alltoallv): 2048 cores

MVAPICH2 version 2.3.4 vs 2.3.7



Expanse GPU Node Architecture

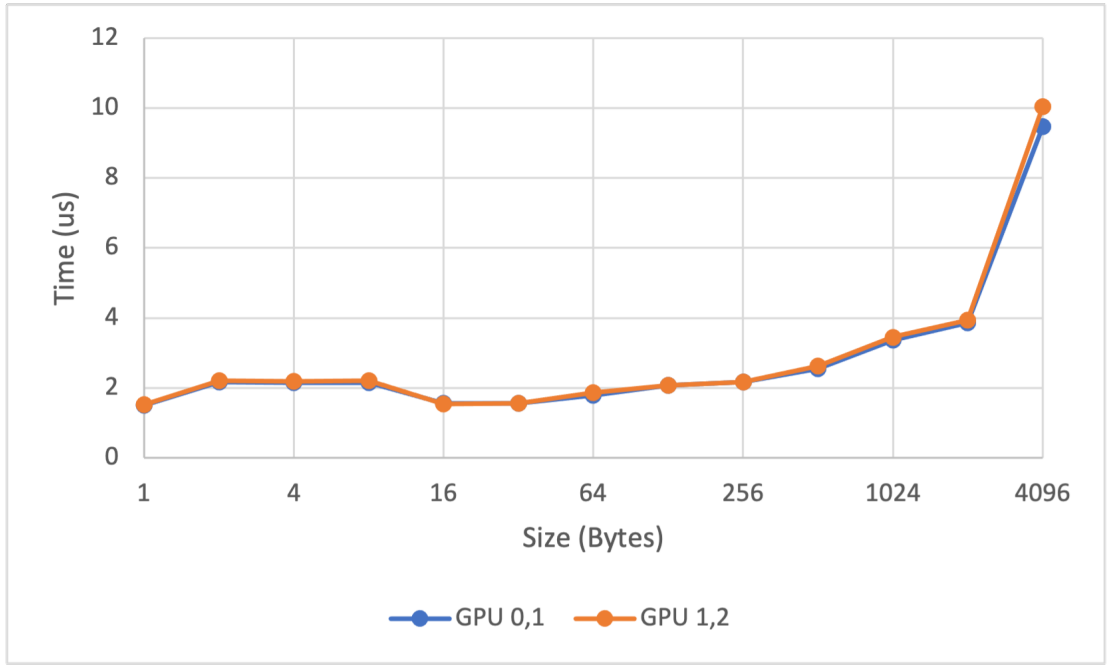
- 4 V100 32GB SMX2 GPUs
- 384 GB RAM, 1.6 TB PCIe NVMe
- 2 Intel Xeon 6248 CPUs
- Topology:

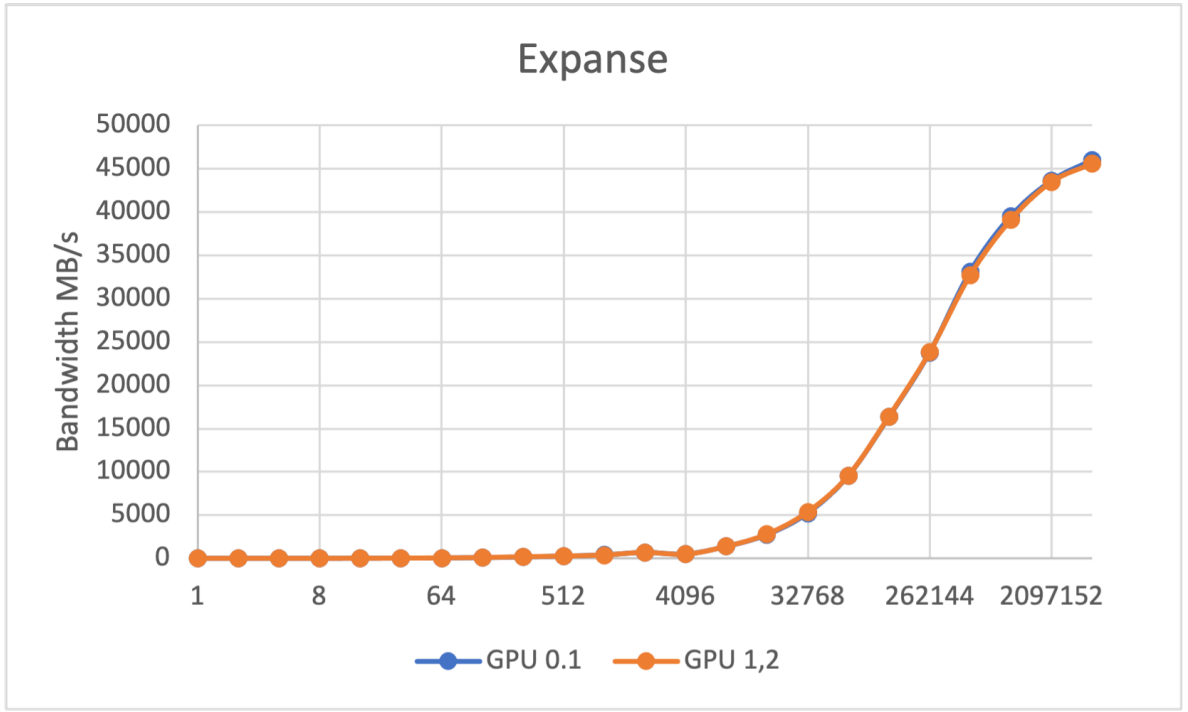
```
        GPU0    GPU1    GPU2    GPU3    mlx5_0  CPU Affinity
GPU0     X      NV2     NV2     NV2     SYS     0-0,4-4,8-8,12-12,16-16,20-20,24-24,28-28,32-32,36-36
GPU1    NV2      X      NV2     NV2     SYS     0-0,4-4,8-8,12-12,16-16,20-20,24-24,28-28,32-32,36-36
GPU2    NV2     NV2      X      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
```

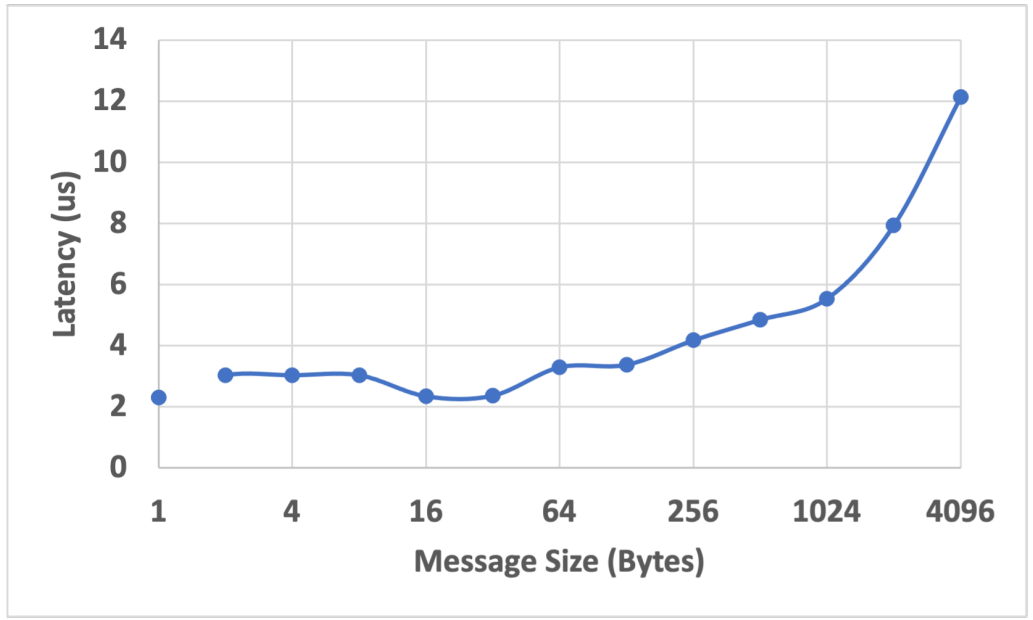
OSU Latency and Bandwidth (osu_latency, osu_bw) Benchmark Intra-node, V100 nodes on Expanse



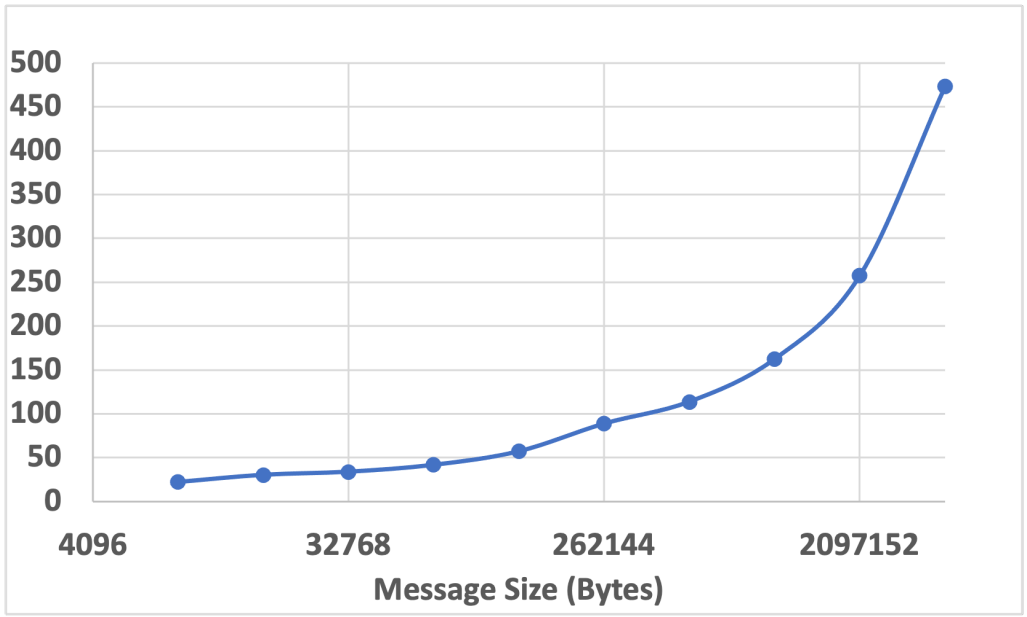


- *Expanse - V100 nodes*
- *Latency between GPU 0 , GPU 1: 1.51 µs*
- *Latency between GPU 1 , GPU 2: 1.53 µs*
- *MVAPICH2 GDR 2.3.6, GCC 8.3.1*

OSU Latency Benchmark: Inter-node, V100 nodes on Expanse MVAPICH2-GDR v2.3.7 w/ gcc/8.5.0



MPI Latency - Small Messages



MPI Latency - Large Messages

Outline

- Introduction and Overview
- Expanse system architecture
- AMD EPYC Processor Architecture
  - Hardware details
  - NUMA options
- Microbenchmarks
- Applications using MVAPICH2, MVAPICH2-GDR on Expanse
  - Summary of MVAPICH2 based installs
  - Benchmark results for LAMMPS, RAxML, Neuron
- Summary

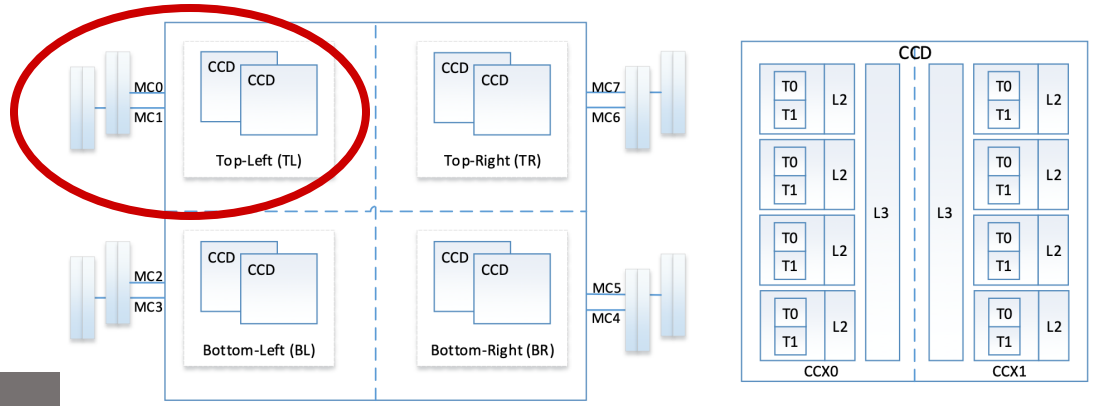
Summary of MVAPICH2 based Application Installs

Application	Description
RAxML	Code for sequential and parallel Maximum Likelihood based inference of large phylogenetic trees
Q-Chem	Commercial package for comprehensive ab initio quantum chemistry software for accurate predictions of molecular structures,
AMBER	Suite of biomolecular simulation programs
LAMMPS	Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) Classical molecular dynamics code with a focus on materials modeling
NAMD	Parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems
ABINIT	Open-source software suite to calculate the optical, mechanical, vibrational, and other observable properties of materials
NEURON	Simulation environment for modeling individual and networks of neurons
TensorFlow w/ Horovod	Open-source platform for machine learning
PyTorch w/ Horovod	Open-source machine learning framework

Additionally, libraries installed: e.g. hypre, fftw, hdf5, netcdf, netcdf, ncview

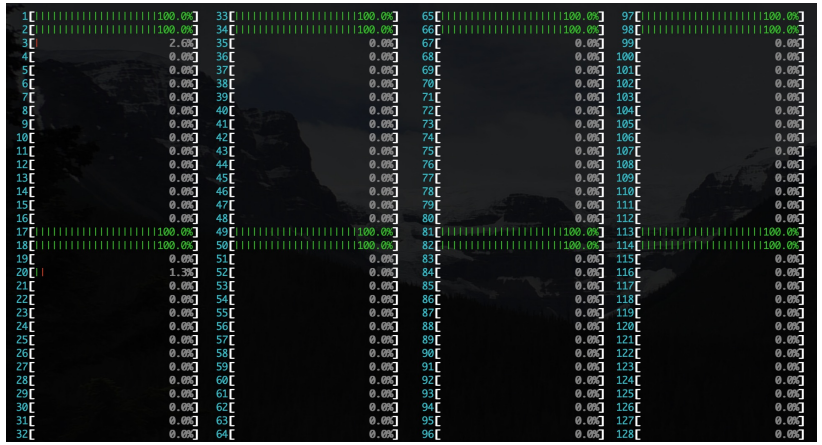
LAMMPS Benchmark:
Lennard-Jones Potential with 512K atoms
Build: gcc + MVAPICH2 compilers

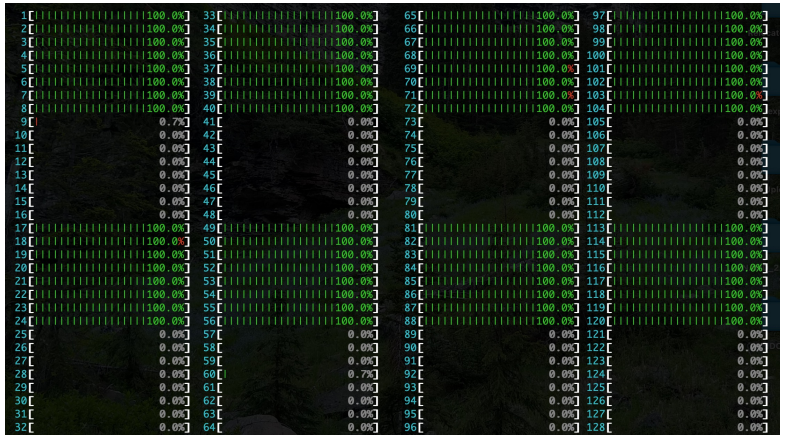
Total #MPI Tasks	Expanse (Compact) (timesteps/sec)	Expanse (Best Memory BW) (timesteps/sec)
16	51.5	56.77
32	104.67	107.67
64	185.69	204.55
128	340.62	340.62



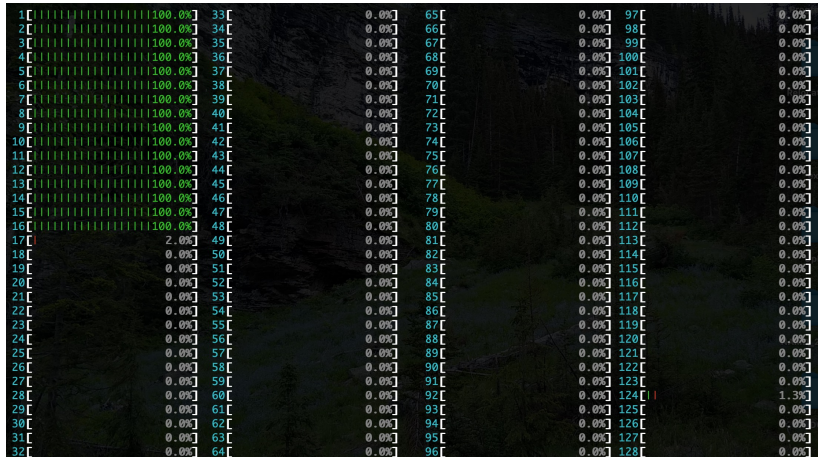
LAMMPS Benchmark:
Lennard-Jones Potential with 512K atoms
Build: gcc + MVAPICH2 compilers

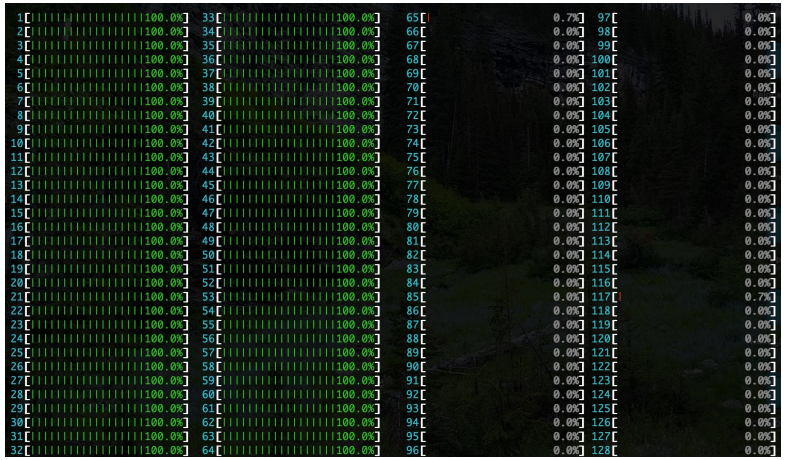
Max Memory Bandwidth



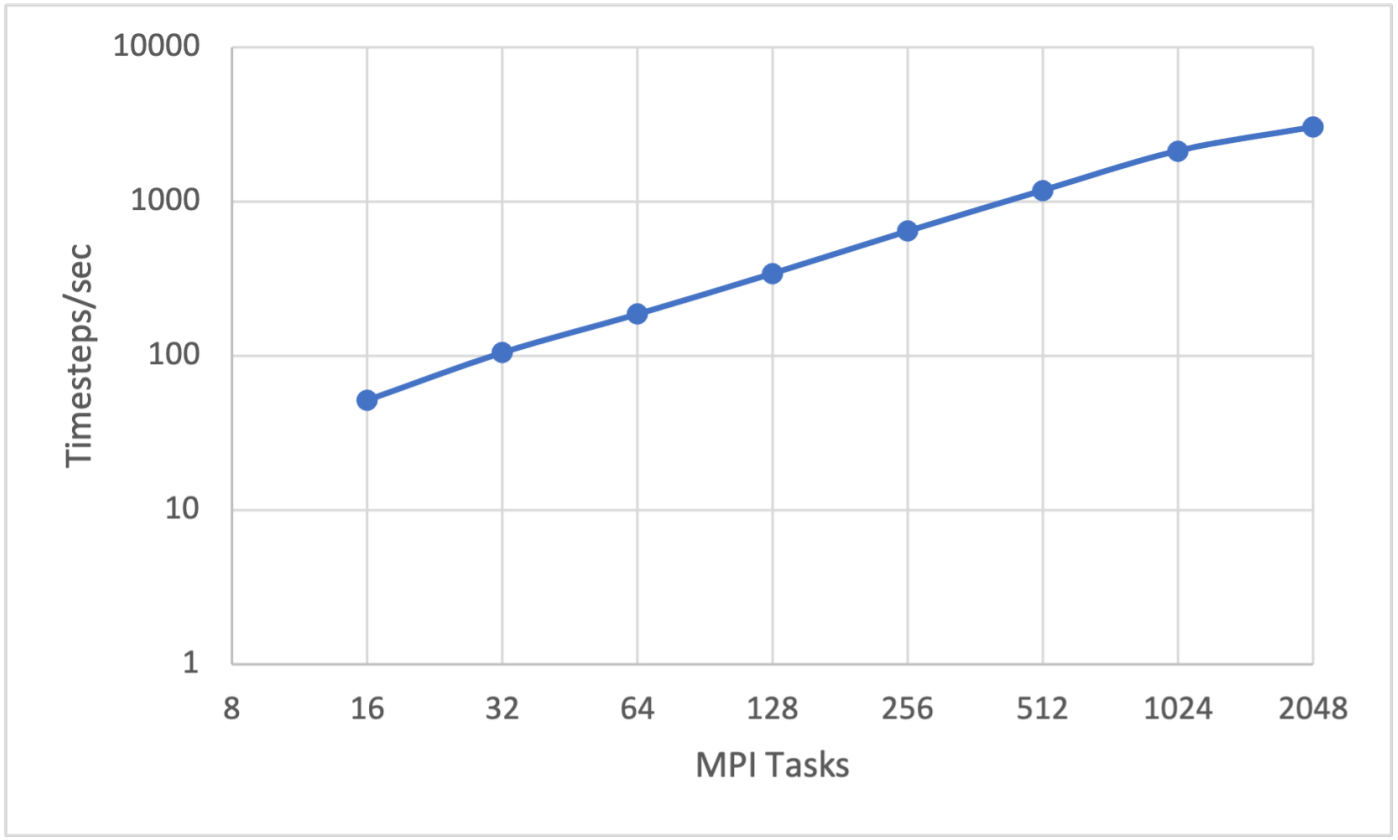


Compact



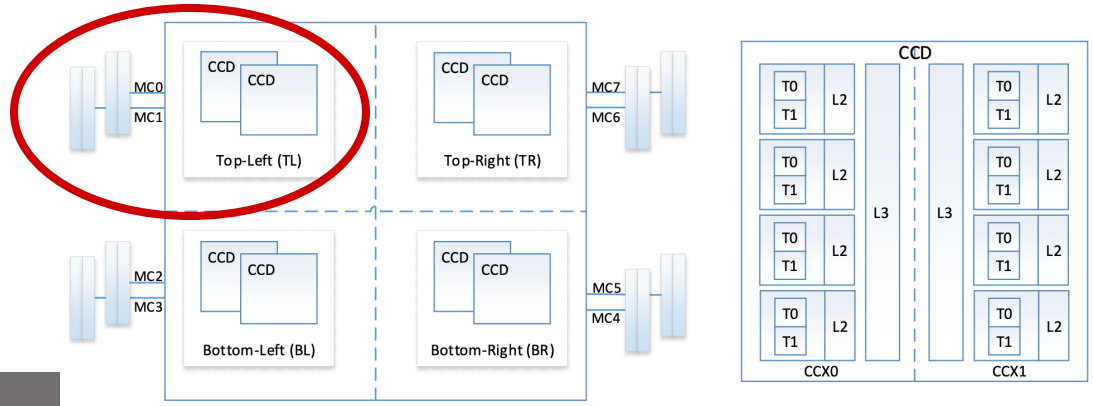


LAMMPS Benchmark:
Lennard-Jones Potential with 512K atoms
Build: gcc + MVAPICH2 compilers



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

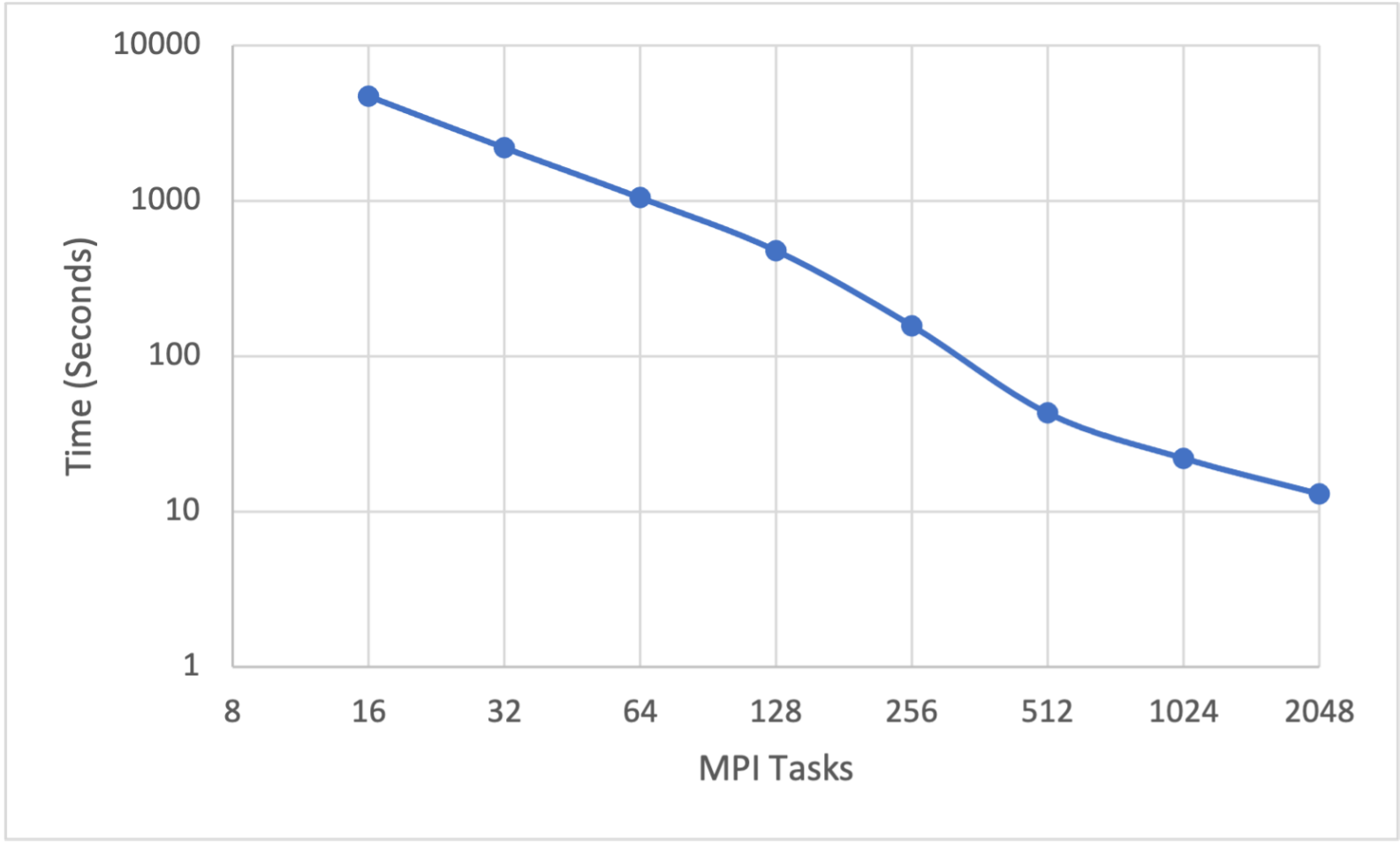
Total #MPI Tasks	Expanse (Compact) Time (sec)	Expanse (Best Memory BW) Time (sec)
16	4700	1397
32	2187	883
64	1044	648
128	477	477



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 MAVPICH2/2.3.7
96	522 s	525 s	539 s	537 s
192	264 s	220 s	211 s	210 s
384	120 s	68 s	65 s	68 s
768	**53 s**	**35 s**	**36 s**	**28 s**

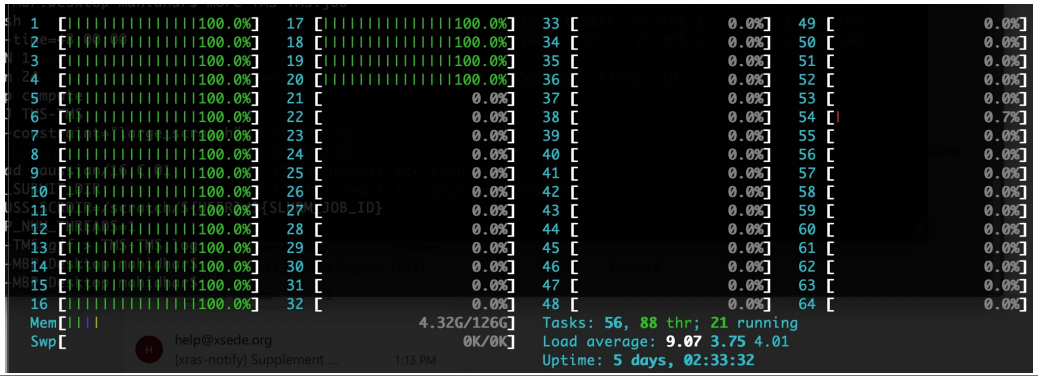
NEURON Benchmark:
Large-scale model of olfactory bulb: 10,500 cells, 40K timesteps

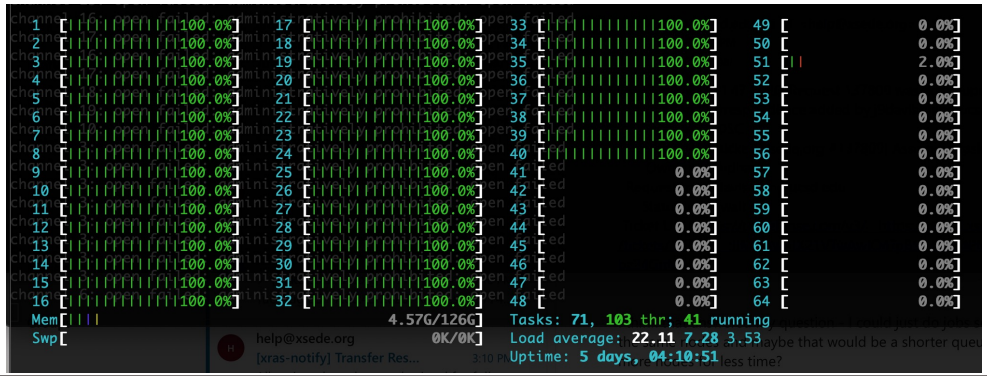


RAxML Benchmark: All-in-one analysis: 218 taxa, 2,294 DNA characters, 1,846 patterns, 100 bootstraps (MPI + Pthreads)

Build: OneAPI Compiler + MVAPICH2/2.3.7

Total tasks	Comet (s)	Stampede2 (s)	Expanse-Dev (s)	Expanse (s) (MV 2.3.7)
10 (5 MPI x 2 Pthreads)	925	610	514	**391**
20 (5 MPI x 4 Pthreads)	542	363	292	**227**
30 (10 MPI x 3 Pthreads)	433	332	247	**181**
40 (10 MPI x 4 Pthreads)	341	300	201	**153**





Summary

- Expanse: 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 used for several application installs on Expanse. Versions used include 2.3.4, 2.3.6, and 2.3.7.
- Continuous improvement in performance with newer versions as optimizations are ongoing.
- Ongoing performance testing with MVAPICH2-GDR v2.3.7. Includes tests with AMBER, HOOMD-Blue, TensorFlow, PyTorch applications.

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

EXPANSE

COMPUTING WITHOUT BOUNDARIES

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