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)
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**

- 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
EXPA NSE
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

REMOTE CI INTEGRATION

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

SDSC
S A N D I E G O
SUPERCOMPUTER CENTER
UC San Diego
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

**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

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

**Non-blocking Interconnect**
- 1 HDR Switch/SSCU: 10x (200 Gbps)
- 4x GPU Nodes
- 56x Compute Nodes
- 5 Level 2 switches: 26x (200 Gbps)
  - 2x Cloud Storage
  - 3x Performance Storage
  - 1 HDR Switch/SSCU

**SDSC SAN DIEGO SUPERCOMPUTER CENTER**

**UC San Diego**
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](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:

<table>
<thead>
<tr>
<th>GPU0</th>
<th>GPU1</th>
<th>GPU2</th>
<th>GPU3</th>
<th>m1x5_0</th>
<th>CPU Affinity</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>NV2</td>
<td>NV2</td>
<td>NV2</td>
<td>SYS</td>
<td>0-0, 4-4, 8-8, 12-12, 16-16, 20-20, 24-24, 28-28, 32-32, 36-36</td>
</tr>
<tr>
<td>NV2</td>
<td>X</td>
<td>NV2</td>
<td>NV2</td>
<td>SYS</td>
<td>0-0, 4-4, 8-8, 12-12, 16-16, 20-20, 24-24, 28-28, 32-32, 36-36</td>
</tr>
<tr>
<td>NV2</td>
<td>NV2</td>
<td>X</td>
<td>NV2</td>
<td>SYS</td>
<td>1-1, 5-5, 9-9, 13-13, 17-17, 21-21, 25-25, 29-29, 33-33, 37-37</td>
</tr>
<tr>
<td>NV2</td>
<td>NV2</td>
<td>NV2</td>
<td>X</td>
<td>SYS</td>
<td>1-1, 5-5, 9-9, 13-13, 17-17, 21-21, 25-25, 29-29, 33-33, 37-37</td>
</tr>
<tr>
<td>SYS</td>
<td>SYS</td>
<td>SYS</td>
<td>SYS</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

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 the 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

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

Additionally, libraries installed: e.g. hypre, fftw, hdf5, netcdf, netcdf, ncview
### LAMMPS Benchmark:
Lennard-Jones Potential with 512K atoms

**Build: gcc + MVAPICH2 compilers**

<table>
<thead>
<tr>
<th>Total #MPI Tasks</th>
<th>Expanse (Compact) (timesteps/sec)</th>
<th>Expanse (Best Memory BW) (timesteps/sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>51.5</td>
<td>56.77</td>
</tr>
<tr>
<td>32</td>
<td>104.67</td>
<td>107.67</td>
</tr>
<tr>
<td>64</td>
<td>185.69</td>
<td>204.55</td>
</tr>
<tr>
<td>128</td>
<td>340.62</td>
<td>340.62</td>
</tr>
</tbody>
</table>
LAMMPS Benchmark:
Lennard-Jones Potential with 512K atoms

Build: gcc + MVAPICH2 compilers
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

<table>
<thead>
<tr>
<th>Total #MPI Tasks</th>
<th>Expanse (Compact) Time (sec)</th>
<th>Expanse (Best Memory BW) Time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>4700</td>
<td>1397</td>
</tr>
<tr>
<td>32</td>
<td>2187</td>
<td>883</td>
</tr>
<tr>
<td>64</td>
<td>1044</td>
<td>648</td>
</tr>
<tr>
<td>128</td>
<td>477</td>
<td>477</td>
</tr>
</tbody>
</table>

---

**Diagram:**

- **Top-Left (TL):** MC0 MC1
- **Top-Right (TR):** MC2 MC3
- **Bottom-Left (BL):** MB0 MB1
- **Bottom-Right (BR):** MB2 MB3

- **CCD:**
  - L1: T0 T1
  - L2: T0 T1
  - L3: T0 T1

---

**SDSC:**
San Diego Supercomputer Center

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

<table>
<thead>
<tr>
<th>#MPI Tasks</th>
<th>Comet</th>
<th>Test Cluster AMD Rome, EDR IB</th>
<th>Expanse MVAPICH2/2.3.5</th>
<th>Expanse MAVPICH2/2.3.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>96</td>
<td>522 s</td>
<td>525 s</td>
<td>539 s</td>
<td>537 s</td>
</tr>
<tr>
<td>192</td>
<td>264 s</td>
<td>220 s</td>
<td>211 s</td>
<td>210 s</td>
</tr>
<tr>
<td>384</td>
<td>120 s</td>
<td>68 s</td>
<td>65 s</td>
<td>68 s</td>
</tr>
<tr>
<td>768</td>
<td>53 s</td>
<td>35 s</td>
<td>36 s</td>
<td>28 s</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Total tasks</th>
<th>Comet (s)</th>
<th>Stampede2 (s)</th>
<th>Expanse-Dev (s)</th>
<th>Expanse (s) (MV 2.3.7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 (5 MPI x 2 Pthreads)</td>
<td>925</td>
<td>610</td>
<td>514</td>
<td>391</td>
</tr>
<tr>
<td>20 (5 MPI x 4 Pthreads)</td>
<td>542</td>
<td>363</td>
<td>292</td>
<td>227</td>
</tr>
<tr>
<td>30 (10 MPI x 3 Pthreads)</td>
<td>433</td>
<td>332</td>
<td>247</td>
<td>181</td>
</tr>
<tr>
<td>40 (10 MPI x 4 Pthreads)</td>
<td>341</td>
<td>300</td>
<td>201</td>
<td>153</td>
</tr>
</tbody>
</table>
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!

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

San Diego Supercomputer Center

In Production October 2020