Experiences with the MVAPICH2 libraries on OSC Clusters
Karen Tomko
Director of Research Software Applications
Outline

• Overview of OSC and our Resources
• OSC’s New Cluster, Pitzer
  • Goals
  • Hardware overview
  • Software Environment
  • InfiniBand Details
• INAM at OSC
• MVAPICH2 MPIs at OSC
Overview of OSC and its Resources

"640K ought to be enough for anybody." – Not Bill Gates
About OSC

• Founded in 1987
• Statewide resource for all universities in Ohio
  • high performance computing services
  • computational science expertise
  • “… propel Ohio's research universities and private industry to the forefront of computational based research.”
• Funded through the Ohio Department of Higher Education
• Reports to the Chancellor of ODHE
• Located on The Ohio State University’s (OSU) west campus
• Fiscal agent is OSU
Service Catalog

Cluster Computing
A fully scalable center with mid-range machines to match those found at National Science Foundation centers and other national labs.

Research Data Storage
High-performance, large capacity data storage spaces along with others that are perfect for a wide variety of research data.

Education
High performance computing and networking resources come together to create an exciting and innovative teaching and research environment.

Web Software Development
Our expert web development team helps you create custom web interfaces to simplify the use of powerful HPC resources.

Scientific Software Development
Deep expertise in developing and deploying software that runs efficiently and correctly on large scale cluster computing platforms.
Client Services

CY2017

- 23 academic institutions
- 48 companies
- 2,202 clients
- 256 awards made
- 23 training opportunities
- 461 trainees
- 604 projects served
- 33 courses used OSC
# OSC Computing and Storage (Q1 2019)

## Systems

<table>
<thead>
<tr>
<th></th>
<th>Ruby</th>
<th>Owens</th>
<th>Pitzer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>2014</td>
<td>2016</td>
<td>2018</td>
</tr>
<tr>
<td>Cost</td>
<td>$1.5 million</td>
<td>$7 million</td>
<td>$3.35 million</td>
</tr>
<tr>
<td>Theoretical Perf.</td>
<td>~144 TF</td>
<td>~1600 TF</td>
<td>~1300 TF</td>
</tr>
<tr>
<td>Nodes</td>
<td>240</td>
<td>824</td>
<td>260</td>
</tr>
<tr>
<td>CPU Cores</td>
<td>4800</td>
<td>23392</td>
<td>10560</td>
</tr>
<tr>
<td>RAM</td>
<td>~15.3 TB</td>
<td>~120 TB</td>
<td>~ 70.6 TB</td>
</tr>
<tr>
<td>GPUs</td>
<td>20 NVIDIA Tesla K40</td>
<td>160 NVIDIA Pascal P100</td>
<td>64 NVIDIA Volta V100</td>
</tr>
<tr>
<td>InfiniBand</td>
<td>FDR</td>
<td>EDR</td>
<td>EDR (CX-5)</td>
</tr>
</tbody>
</table>

Total compute: ~3,044 TF

## Storage

<table>
<thead>
<tr>
<th></th>
<th>Home</th>
<th>Project</th>
<th>Scratch</th>
<th>Tape Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity</td>
<td>0.8 PB</td>
<td>3.4 PB</td>
<td>1.1 PB / 40TB DDN IME</td>
<td>7+ PB</td>
</tr>
</tbody>
</table>
“To err is human, but to really foul things up you need a computer.” – Paul Ehrlich
New HPC Cluster “Pitzer”

- **Named after Russell M. Pitzer**
  - Emeritus professor of chemistry at The Ohio State University
  - Instrumental in founding both OSC and OARnet
  - Significant contributions to computational chemistry

- **Goals**
  - Complement existing systems
  - Replace Oakley cluster with a petaflop class system
    - HP SL350/Intel Xeon X5650 cluster, 2012

- **Timeline**
  - System delivery August 15, 2018
  - Early User Access October 25, 2018
  - Full production November 2018
  - Oakley decommissioning Dec 2018
Pitzer Cluster

Characteristics relative to Oakley
- Delivers 8x the processing power (1,300 vs 154 TF)
- Costs 15% less ($4M vs $3.35M)
- Provides 25% more cores (10,560 vs 8,304)
- Has 2X the memory (70.6Tb vs 33.4TB)
- Uses 20% less power

Highlights
- 10,560 processor cores, ~1.3 petaflop peak
- Latest generation: SkyLake processors, 100Gb InfiniBand
- Warm water cooling supports high density, increased performance and efficiency
## Pitzer Detailed Specifications

<table>
<thead>
<tr>
<th></th>
<th>Standard Compute Node, Dell PowerEdge C6420</th>
<th>GPU Compute Node, Dell PowerEdge R740</th>
<th>Huge Memory Compute Node, Dell PowerEdge R940</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of nodes</td>
<td>224</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>CPUs per node/Cores per node</td>
<td>2/40</td>
<td>2/40</td>
<td>4/80</td>
</tr>
<tr>
<td>Processor</td>
<td>Intel Xeon Gold 6148</td>
<td>Intel Xeon Gold 6148</td>
<td>Intel Xeon Gold 6148</td>
</tr>
<tr>
<td>Memory (GB)</td>
<td>192</td>
<td>384</td>
<td>3072</td>
</tr>
<tr>
<td>GPUs</td>
<td>0</td>
<td>2 NVIDIA V100s, 16GB per GPU</td>
<td>0</td>
</tr>
<tr>
<td>High Speed Interconnect</td>
<td>Mellanox IB EDR 100Gb ConnectX-5</td>
<td>Mellanox IB EDR 100Gb Socket Direct ConnectX-5</td>
<td>Mellanox IB EDR 100Gb ConnectX-5</td>
</tr>
<tr>
<td>Internal Disk</td>
<td>1TB hard drive</td>
<td>1TB hard drive</td>
<td>1TB hard drive</td>
</tr>
<tr>
<td>Cooling</td>
<td>Liquid direct to chip</td>
<td>Liquid direct to chip</td>
<td>Air</td>
</tr>
</tbody>
</table>

Four login nodes: identical to GPU compute nodes but with no GPUs and air cooled
Software Environment

- **Linux**: RHEL 7.5
- **Batch Scheduler**: Torque/MOAB
- **Compilers**: Intel, gnu, PGI
- **MPI**: MVAPICH2, IntelMPI, OpenMPI
- **GPU**: CUDA 9.x, OpenACC (PGI compilers)
- **High Level Languages**: Python, R, Julia, Matlab
- **Performance and debug tools**: Arm, Intel and opensource tools
- **Containers**: singularity to run docker containers, with support for GPU
Deploying INAM at OSC

“In God we trust. All others must bring data.” – W. Edwards Deming
FAMII Project Collaboration

Central Question:
Can a high performance and scalable tool be designed which is capable of analyzing and correlating the communication on the fabric with behavior of HPC/Big Data applications through tight integration with the communication runtime and the job scheduler?

Project Team

**OSU**: Pouya Kousha, Meagan Haupt, Hari Subramoni, Mark Arnold, DK Panda

**OSC**: Trey Dockendorf, Heechang Na, Karen Tomko
Not your lab’s fabric

OSC has a single integrated IB fabric
• Currently 4 compute clusters
• GPFS Scratch and Project filesystems
• Fabric Size: 1,720 compute nodes, 184 switches
• Multiple generations of IB
  • QDR
  • FDR
  • EDR
  • EDR with CX-5
OSC’s Fabric

- Multiple clusters
- Multiple generations of IB
- One fabric
OSC’s Fabric (more detail)
Impact on INAM

Some resulting improvements and configuration:
  • Caching of Rendered Fabric Diagram
    • Time reduced from ~2 minutes to just a few seconds
  • Data Base Optimizations
    • Time for insertion operations reduced 2-4x

Data collection at OSC
  • Collection rate – 5 sec intervals
  • Can only keep a few days worth of data

Next steps
  • Integration with Torque/MOAB
"Alone we can do so little; together we can do so much." – Helen Keller
MVAPICH2 is OSC’s default MPI library

- Oakley  mvapich2  1.7 - 2.2
- Ruby    mvapich2  1.9 - 2.3
- Owens   mvapich2  2.2 - 2.3
- Pitzer  mvapich2  2.3

With builds for Intel, gnu and PGI compilers and with cuda enabled for gpu api.
Pitzer InfiniBand details

• Latest generation InfiniBand
  • ConnectX-5 EDR cards, Switch-IB 2 based switches
  • Same throughput as first generation EDR, 100Gb/s
  • 33% higher message rate relative to first generation EDR
  • Tag matching (asynchronous/one-sided optimization)
  • Full support for GDRCopy on GPU nodes
  • SHARP collective optimizations
  • SHIELD/adaptive routing and network resiliency

• Oversubscription ratio of 2:1

• Integrated into existing OSC InfiniBand fabric
  • RDMA access to file systems
Upcoming

• mvapich2-GDR 2.3
  • In testing on Owens and Pitzer
  • Full support for GDRCopy on Pitzer

• mvapich2-X 2.3
  • XPMEM kernel module available on Owens and Pitzer
  • Targeting Pitzer EDR Connect-X 5 features SHArP, Core-Direct