Advantages to Using MVAPICH2 on TACC HPC Clusters

Jérôme VIENNE
viennej@tacc.utexas.edu

Texas Advanced Computing Center (TACC)
University of Texas at Austin

Wednesday 27th August, 2014
A "Homogeneous" Cluster for Heterogeneous Users

- Large base of users
- Wishes: Runs need to be Simple, Efficient, Reliable

### Usage (July 2014)

<table>
<thead>
<tr>
<th>Job Size (Nb nodes)</th>
<th>Jobs Count</th>
<th>Su Charge</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-16 (1)</td>
<td>87,724</td>
<td>4,392,044</td>
</tr>
<tr>
<td>17-32 (2)</td>
<td>19,097</td>
<td>2,555,694</td>
</tr>
<tr>
<td>33-64 (3-4)</td>
<td>15,961</td>
<td>4,091,764</td>
</tr>
<tr>
<td>65-128 (5-8)</td>
<td>41,636</td>
<td>6,930,632</td>
</tr>
<tr>
<td>129-256 (9-16)</td>
<td>20,187</td>
<td>11,933,678</td>
</tr>
<tr>
<td>257-512 (17-32)</td>
<td>8,991</td>
<td>10,724,880</td>
</tr>
<tr>
<td>513-1024 (33-64)</td>
<td>3,157</td>
<td>11,249,581</td>
</tr>
<tr>
<td>1024+ (65+)</td>
<td>4,289</td>
<td>13,141,299</td>
</tr>
</tbody>
</table>
## Requirements

### For our MPI libraries
- Simplicity
- Reliability
- Performance
- Scalability

**MVAPICH2 is the perfect library for us**
Over the years, we saw that MVAPICH2 was able to fill all our needs.
Plan

1. A continuous improvement
2. Intra-node optimization
3. Multicast
4. Conclusion
Plan

1. A continuous improvement
2. Intra-node optimization
3. Multicast
4. Conclusion
Advantages to Using MVAPICH2 on TACC HPC Clusters
Plan

1. A continuous improvement
2. Intra-node optimization
3. Multicast
4. Conclusion
## Different level of communication inside MPI libraries

- **Inter-node**: Communications between nodes
- **Intra-node**: Communications inside the node

## Growing impact of Intra-node

With the number of cores per node increasing in modern clusters, an efficient implementation of intra-node communications is critical for application performance.
Three different mechanisms

**Shared Memory**

- User Space
- Operating System
- NIC

**Loopback**

- User Space
- Operating System
- NIC
- Hardware Loopback

**Kernel Assisted**

- User Space
- Operating System
- Kernel
- NIC
Three different mechanisms

**Shared Memory**

- Double-copy implementation involves a shared buffer space used by local processes to exchange messages.
- The sending process copies the content of the message into the shared buffer before the receiver reads from it.
Three different mechanisms

Shared Memory

Loopback

Kernel Assisted

Advantages to Using MVAPICH2 on TACC HPC Clusters

Loopback

Use Direct Memory Access (DMA) to transfer data between two processes inside the node.
Two DMA operations across the I/O buses are performed by the NIC.
Three different mechanisms

Shared Memory

Loopback

Kernel Assisted

Kernel Assisted

CMA and kernel modules like LiMIC enable single copy mechanisms for intra-node communication in MPI libraries.
Kernel Assisted

**LiMIC**
- Linux Kernel Module for MPI Intra-Node Communication
- Available on Stampede and Lonestar
- Allows a process to map and access contiguous portions of a remote process’s virtual address space.
- Need a MPI library configured with LiMIC support

**CMA**
LiMIC

CMA

- Cross Memory Attach
- Introduced with Linux kernel 3.2 and has been back-ported to some Linux distribution
- Available on Stampede and Maverick
- Since 2.0, MVAPICCH2 is configured with CMA support automatically (if available).
- CMA will be used for large messages
Advantages to Using MVAPICH2 on TACC HPC Clusters
Advantages to Using MVAPICH2 on TACC HPC Clusters
MPI Collectives on Large Mem Node

IMB Alltoall with MVAPICH2 with 32 MPI tasks

Advantages to Using MVAPICH2 on TACC HPC Clusters
### NAS results on large mem node with 32 cores, MV2 1.9

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Class</th>
<th>Shared (s)</th>
<th>CMA (s)</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG</td>
<td>C</td>
<td>10.29</td>
<td>9.66</td>
<td>+6.12%</td>
</tr>
<tr>
<td>EP</td>
<td>C</td>
<td>3.89</td>
<td>3.88</td>
<td>+0%</td>
</tr>
<tr>
<td>FT</td>
<td>C</td>
<td>16.04</td>
<td>12.07</td>
<td>+24.75%</td>
</tr>
<tr>
<td>IS</td>
<td>C</td>
<td>1.37</td>
<td>1.04</td>
<td>+24.08%</td>
</tr>
<tr>
<td>CG</td>
<td>D</td>
<td>381.95</td>
<td>382.03</td>
<td>-0.02%</td>
</tr>
<tr>
<td>EP</td>
<td>D</td>
<td>62.07</td>
<td>62.08</td>
<td>+0.8%</td>
</tr>
<tr>
<td>FT</td>
<td>D</td>
<td>365.84</td>
<td>289.32</td>
<td>+20.91%</td>
</tr>
<tr>
<td>IS</td>
<td>D</td>
<td>26.1</td>
<td>20.92</td>
<td>+19.8%</td>
</tr>
</tbody>
</table>

Advantages to Using MVAPICH2 on TACC HPCClusters
Plan

1. A continuous improvement
2. Intra-node optimization
3. Multicast
4. Conclusion
Stampede/MVAPICH2 Multicast Features

Hardware support for multicast in new generation of IB

- MVAPICH2 has support to use this
- Large MPI_Bcast, MPI_Scatter and MPI_Allreduce can be much more efficient
- Dramatic improvement with increasing node count
- factors of 3-5X reduction at 16k cores.

Requirements:

- Need MVAPICH2 1.9a or higher
- Configure flag: --enable-mcast (Enabled by default)
- Runtime: MV2_MCAST_COMM_INIT_TIMEOUT=20000 MV2_USE_MCAST=1 (Disabled by default)
Advantages to Using MVAPICH2 on TACC HPC Clusters
Advantages to Using MVAPICH2 on TACC HPC Clusters
Plan

1. A continuous improvement
2. Intra-node optimization
3. Multicast
4. Conclusion
Conclusion

- Each release brings new features and performance optimization.
- LiMIC and CMA bring a boost for intra-node communication.
- Multicast can help at large scale, it worse to try
- Don't forget to update your MVAPICH2 install
- Thank you to the MVAPICH2 team for the hard work!