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

Stampede

A "Homogeneous" Cluster for Hetergeneous Users

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

Usage (July 2014)

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

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

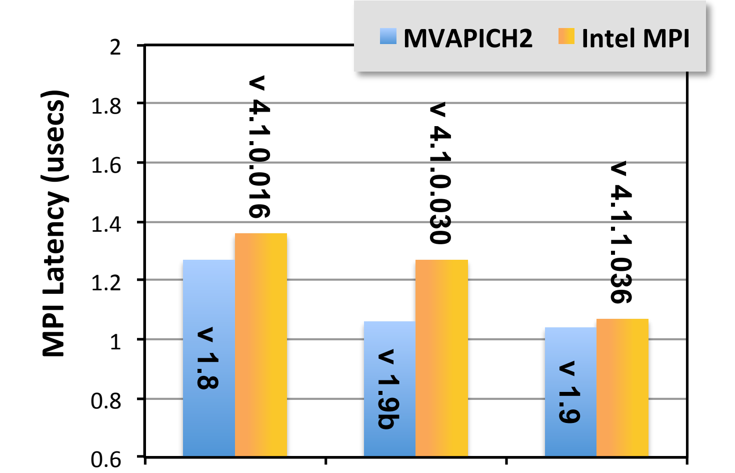
3. Multicast

4. Conclusion

Plan

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

Latency comparison (using core 8)



Plan

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

Different level of communication inside MPI libraries

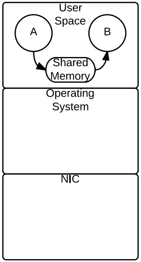
- Inter-node: Communications between nodes
- Intra-node: Communications inside the node

Growing impact of Intra-node

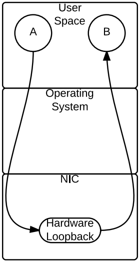
With the number of cores per node increasing in modern clusters, an efficient implementation of intra-node communications is critical for application performance.

Three different mechanisms

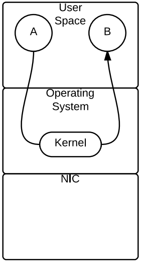
Shared Memory



Loopback

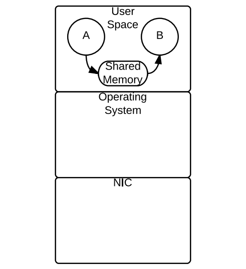


Kernel Assisted

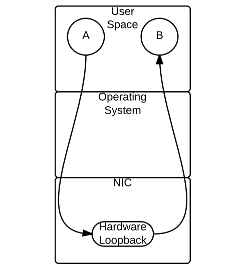


Three different mechanisms

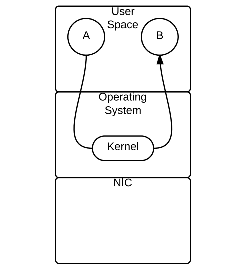
Shared Memory



Loopback



Kernel Assisted

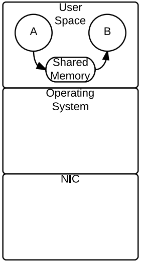


Shared Memory

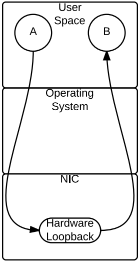
Double-copy implementation involves a shared buffer space used by local processes to exchange messages.
The sending process copies the content of the message into the shared buffer before the receiver reads from it.

Three different mechanisms

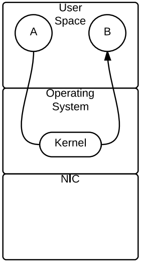
Shared Memory



Loopback



Kernel Assisted

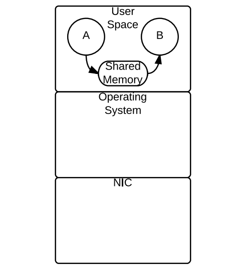


Loopback

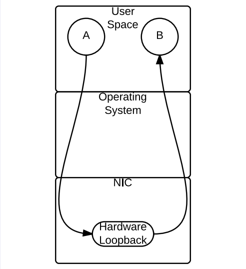
Use Direct Memory Access (DMA) to transfer data between two processes inside the node.
Two DMA operations across the I/O buses are performed by the NIC.

Three different mechanisms

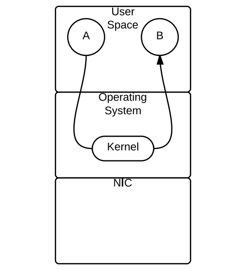
Shared Memory



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

- **Li**nux Kernel Module for **M**PI **I**ntra-Node **C**ommunication
- 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

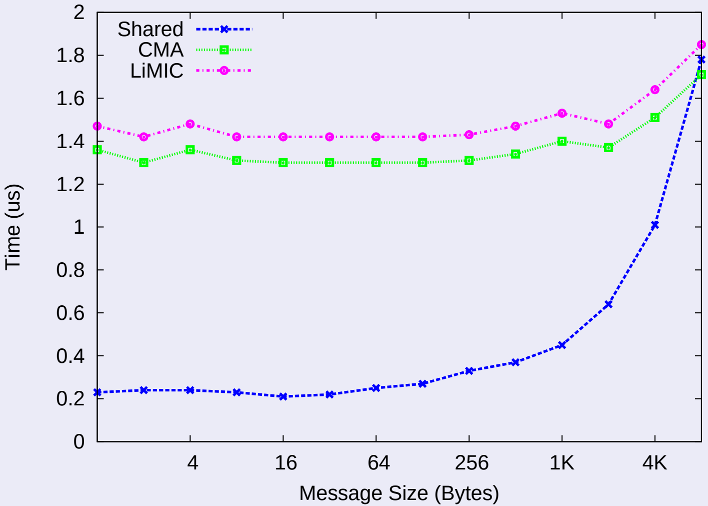
Kernel Assisted

LiMIC

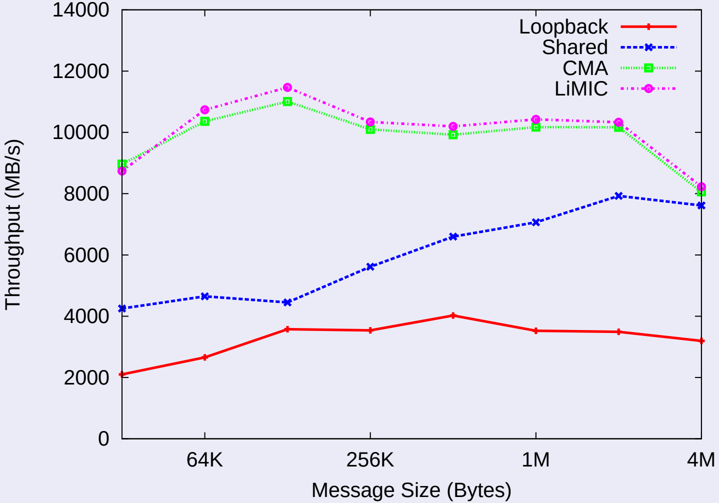
CMA

- **C**ross **M**emory **A**ttach
- Introduced with Linux kernel 3.2 and has been back-ported to some Linux distribution
- Available on Stampede and Maverick
- Since 2.0, MVAPICH2 is configured with CMA support automatically (if available).
- CMA will be used for large messages

IMB Pingpong on Compute Node, Intra-socket, MV2 1.9

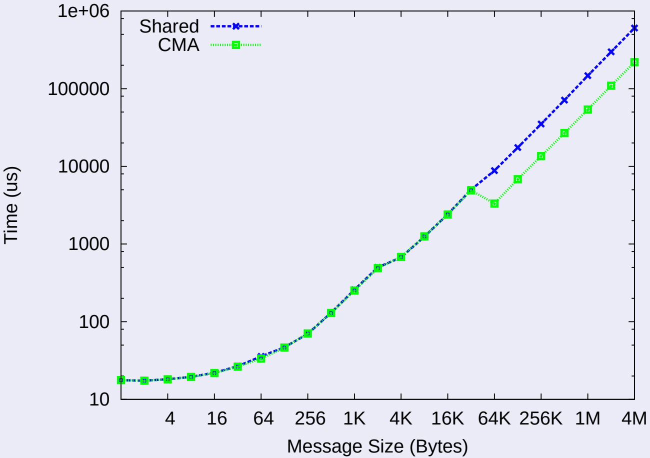


Intra-socket on Compute Node: Large Messages, MV2 1.9



MPI Collectives on Largemem Node

IMB Alltoall with MVAPICH2 with 32 MPI tasks



NAS results on large mem node with 32 cores, MV2 1.9

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

Plan

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

Stampede/MVAPICH2 Multicast Features

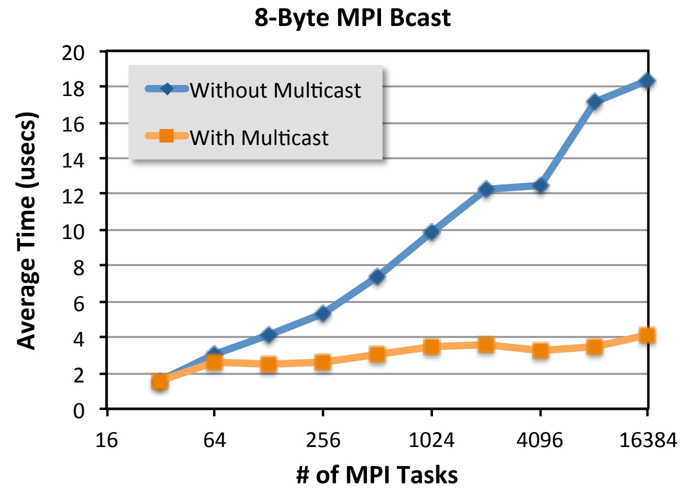
Hardware support for multicast in new generation of IB

- 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.

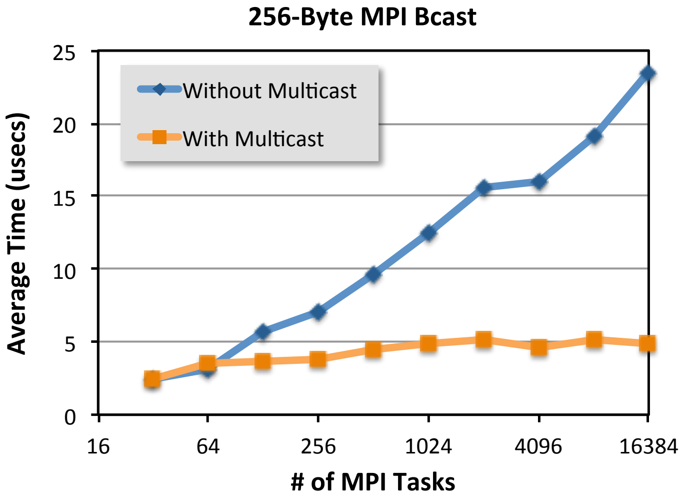
Requirements:

- 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)

Multicast on Stampede



Multicast on Stampede



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 !