MVAPICH/MVAPICH2 Update

Presentation at Int'l Sonoma Workshop
(April '07)
by

Dhabaleswar K. (DK) Panda
Department of Computer Science and Engg.
The Ohio State University
E-mail: panda@cse.ohio-state.edu
http://www.cse.ohio-state.edu/~panda

Presentation Overview

- Overview of MVAPICH/MVAPICH2 Project
- Selected Features of the latest releases
  - Message Coalescing and Memory Scalability
  - ConnectX Performance
  - Congestion Avoidance with Multi-Pathing
  - Multi-core-aware Point-to-point
  - Multi-core-aware Optimized Collectives
  - Checkpoint/Restart
  - RDMA CM and iWARP
  - OSU Benchmarks
- Upcoming Features and Issues
  - Overlap of Computation and Communication
  - Automatic Path Migration (APM)
  - UD-based Design
  - Multi-Network Support using uDAPL
  - MVAPICH-PSM (QLogic) Implementation and Performance
- Conclusions

MVAPICH/MVAPICH2 Software Distribution

- High Performance and Scalable Implementations
  - MPI-1 (MVAPICH)
  - MPI-2 (MVAPICH2)
- Both are being available with OFED 1.2
  - MVAPICH 0.9.9 (released on 04/27/07)
  - MVAPICH2 0.9.8 (released on 11/10/06)
- Has enabled a large number of production IB clusters all over the world to take advantage of IB
- Have been directly downloaded and used by more than 495 organizations worldwide
- More details at (New Website)

  http://mvapich.cse.ohio-state.edu

New Features of MVAPICH 0.9.9

- Improved message coalescing:
  - Reduction of per QP send queues for reduction in memory requirement
  - Increases the small message messaging rate significantly
- Multi-core optimizations:
  - Optimized scalable shared memory design
  - Optimized, high-performance shared memory aware collective operations
  - Multi-port support for enabling user processes to bind to different IB ports for balanced communication performance
- Multi-path support for hot-spot avoidance in large scale clusters using LMC
- Memory Hook Support provided by integration with ptmalloc2 library
- Shared memory channel (for multi-processor systems without any high performance networks - clusters with serial nodes, servers, laptops, etc.)

New Features of MVAPICH2 0.9.8

- Includes most of the features of MVAPICH
- Performance and scalability comparable to MVAPICH for two-sided communication
- Added MPI-2 features (one-sided communication, collectives and datatype)
- Integrated Multi-rail support
- Multi-threading support (MPI_Thread_Multiple)
- RDMACM support for InfiniBand and iWARP
- Checkpoint/Restart support for application transparent systems-level fault tolerance

Support for Multiple Interfaces/Adapters

- OpenFabrics/Gen2-IB
  - All IB adapters supporting Gen2
  - Supports ConnectX
- uDAPL
  - Linux-IB
  - Solaris-IB
  - Neteffect 10GigE
    - support introduced in MVAPICH2 0.9.8
- OpenFabrics/Gen2-iWARP
  - Introduced in MVAPICH2 0.9.8
    - Tested with Chelsio (10GigE)
- VAPI
  - All IB adapters supporting VAPI
- TCP/IP
  - Any adapter supporting TCP/IP interface
- Support for QLogic at the PSM-level will be available soon

Presentation Overview

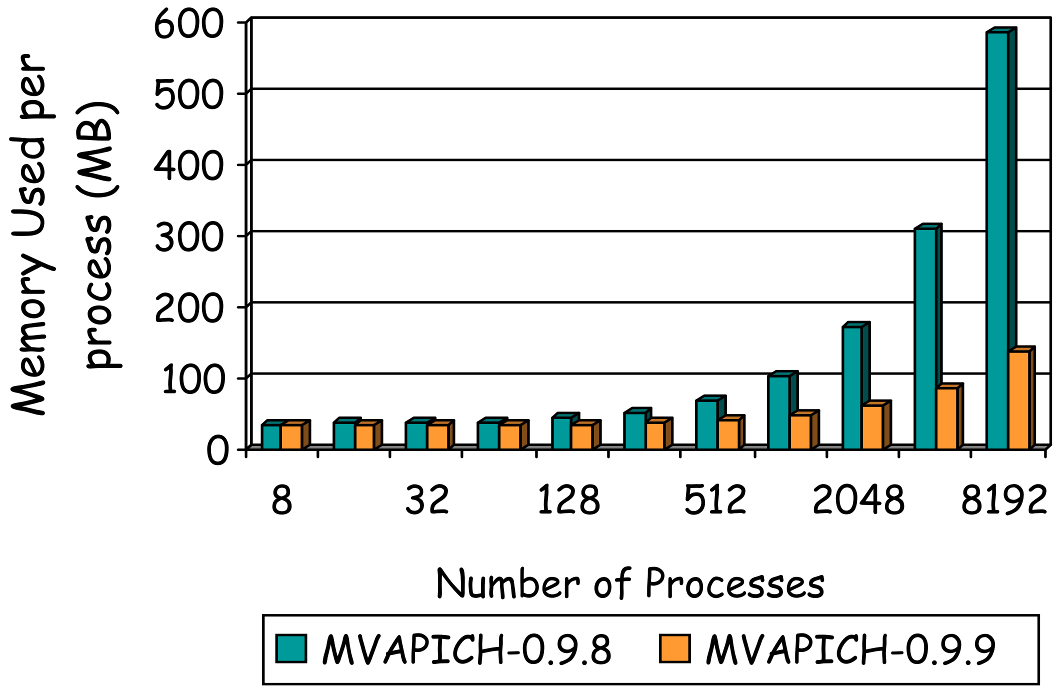
- Overview of MVAPICH/MVAPICH2 Project
- Selected Features of the latest releases
  - Message Coalescing and Memory Scalability
  - ConnectX Performance
  - Congestion Avoidance with Multi-Pathing
  - Multi-core-aware Point-to-point
  - Multi-core-aware Optimized Collectives
  - Checkpoint/Restart
  - RDMA CM and iWARP
  - OSU Benchmarks
- Upcoming Features and Issues
  - Overlap of Computation and Communication
  - Automatic Path Migration (APM)
  - UD-based Design
  - Multi-Network Support using uDAPL
  - MVAPICH-PSM (QLogic) Implementation and Performance
- Conclusions

Improved Message Coalescing

- Large-scale InfiniBand clusters are being increasingly common
- Memory usage is allocated on a per connection basis
- Improved message coalescing method
  - to allow reducing the number of allowed outstanding send operations to save memory (an order of magnitude) while maintaining performance
- Increases the small message messaging rate significantly
- Runtime environment variables
  - Enable or disable message coalescing
  - Degree of message coalescing
- Applications can be evaluated with/without coalescing

Matthew Koop, Terry Jones, and Dhabaleswar K. Panda , "Reducing Connection Memory Requirements of MPI for InfiniBand Clusters: A Message Coalescing Approach, " (CCGrid), Rio de Janeiro - Brazil, May 2007

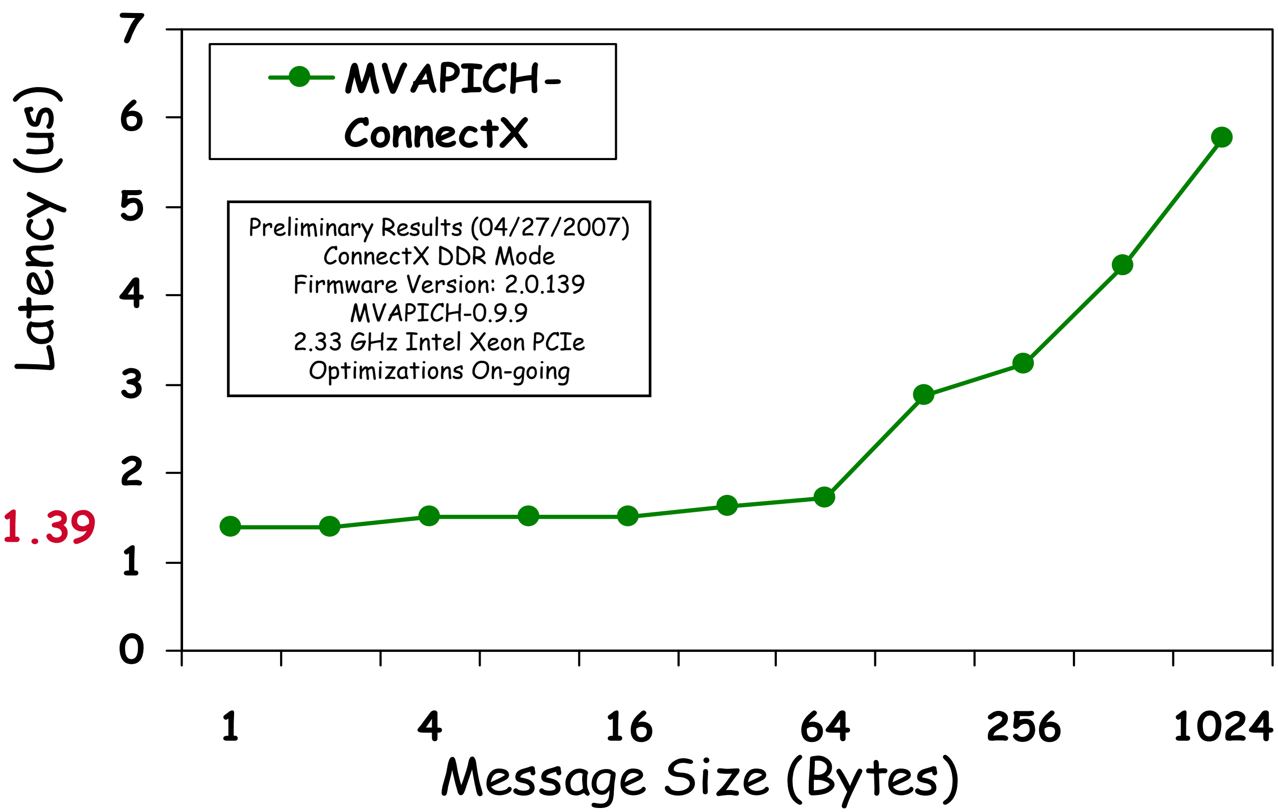
Reduced Memory Usage in MVAPICH-0.9.9



- MVAPICH-0.9.9 requires only around 140 MB per process even with all 8192 processes connected to each other using RC (worst-case scenario)
- Results taken on the IB cluster at Lawrence Livermore National Lab.

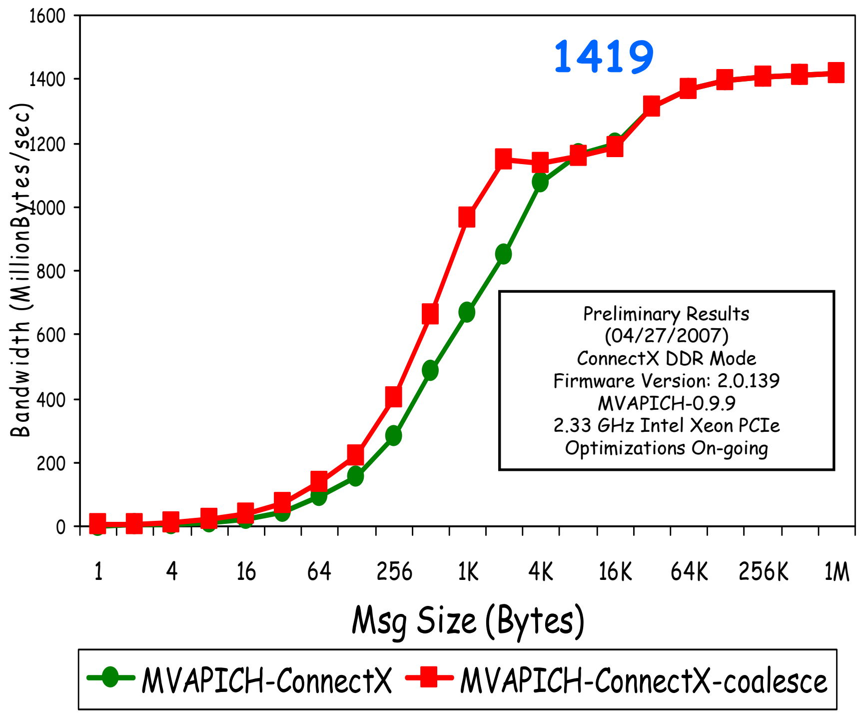
ConnectX Performance with MVAPICH 0.9.9 - Latency

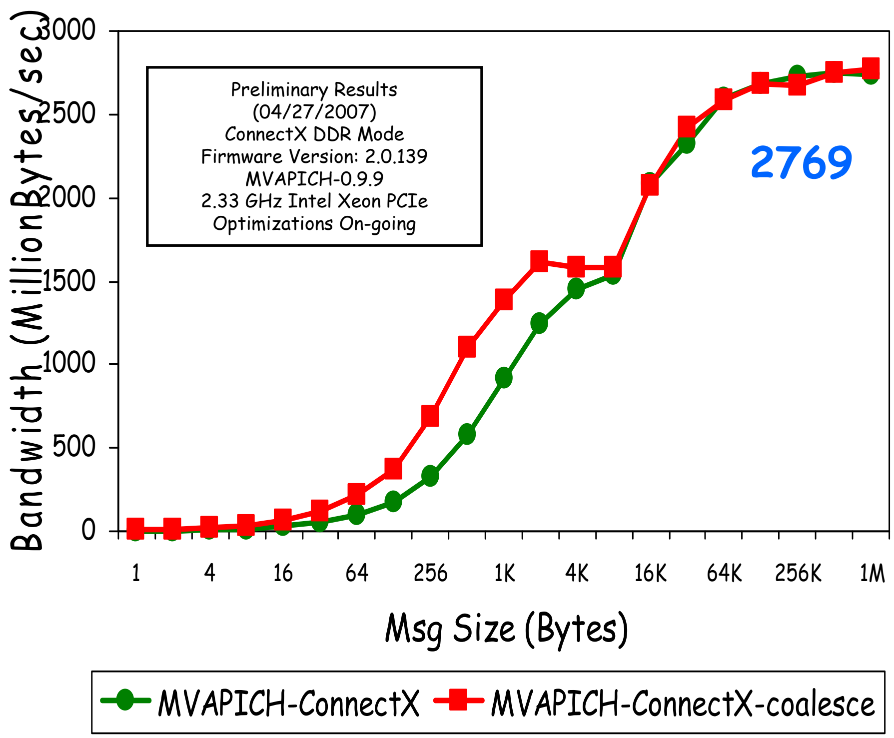
2.33 GHz Intel Quad-core, with switch, one process per node



ConnectX: Bandwidth and Bi-directional Bandwidth

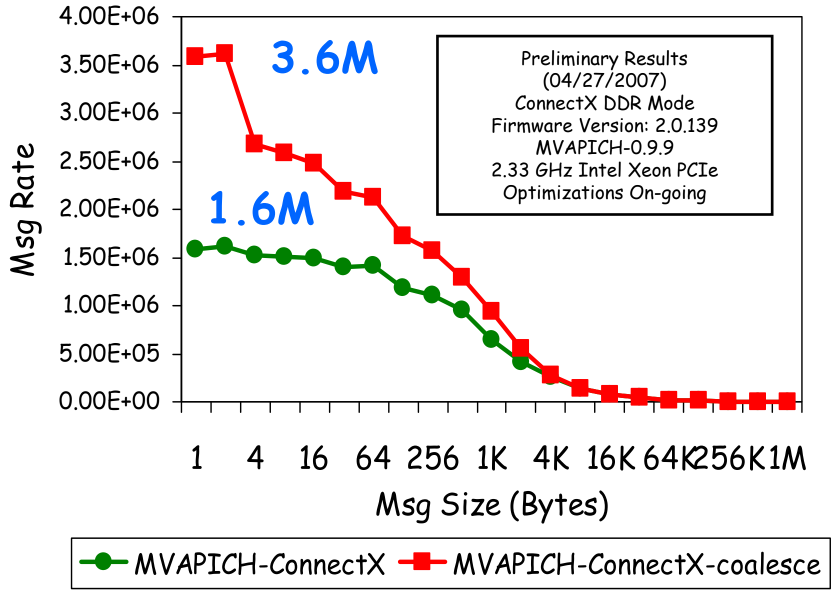
2.33 GHz Intel Quad-core, with switch, one process per node



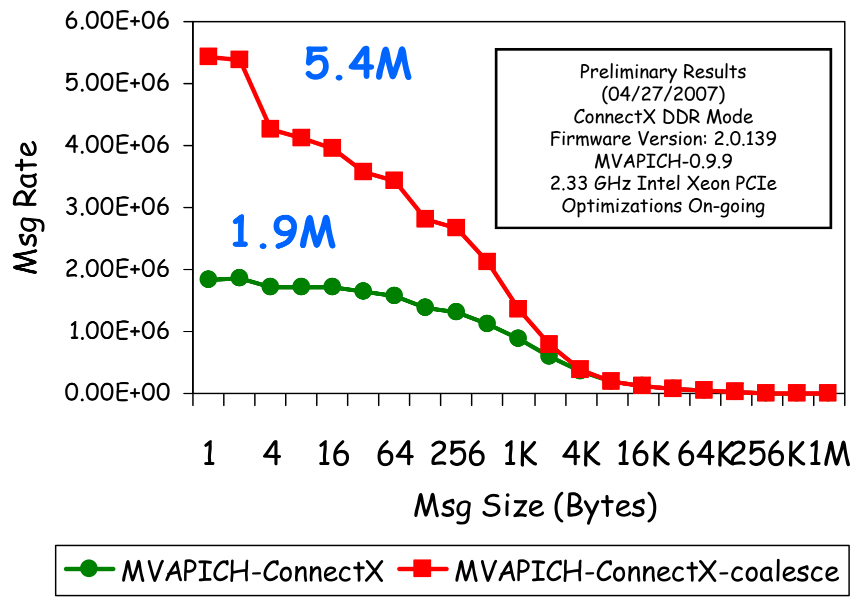


ConnectX: Messaging Rate

2.33 GHz Intel Quad-core, with switch, one process per node



Uni-directional Message Rate



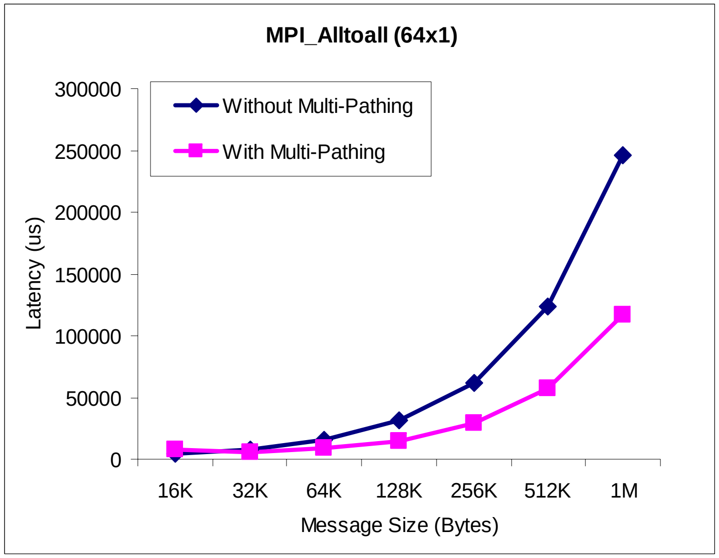
Bi-directional Message Rate

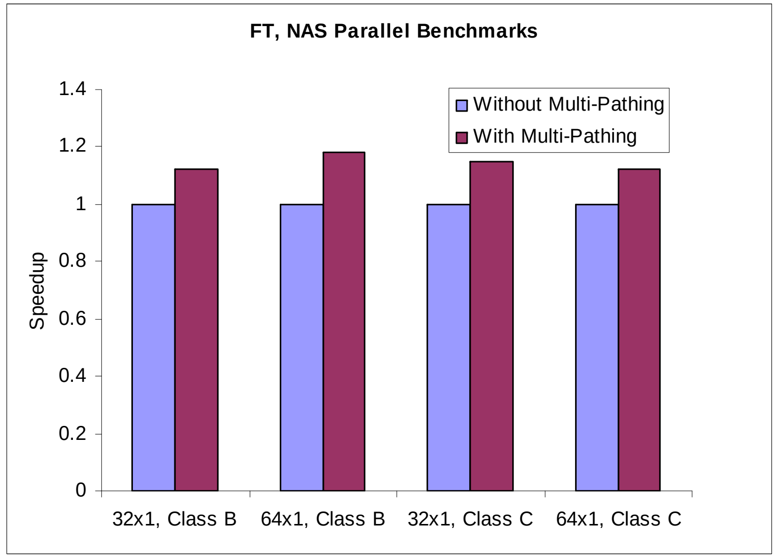
Congestion Avoidance with Multi-Pathing

- Large scale clusters may not be complete fat tree
  - Congestion due to absence of CBB
  - In the presence of CBB, routing algorithm plays an important role in the usage of these paths
- Location of MPI tasks in a job impact the overall performance
  - Static selection of paths may not work well for different MPI task allocations
- The situation becomes more complicated
  - Different communication patterns in same application
  - Different collective communication algorithms
  - Interaction due to other jobs in the cluster
- Can we design an adaptive scheme to take care of above scenarios?

Abhinav Vishnu, Matt Koop, Adam Moody, Amith Mamidala, Sundeep Narravula and Dhabaleswar K. Panda , " Hot-Spot Avoidance With Multi-Pathing Over InfiniBand: An MPI Perspective, " (CCGrid), Rio de Janeiro - Brazil, May 2007 (Best Paper Award Nominee)

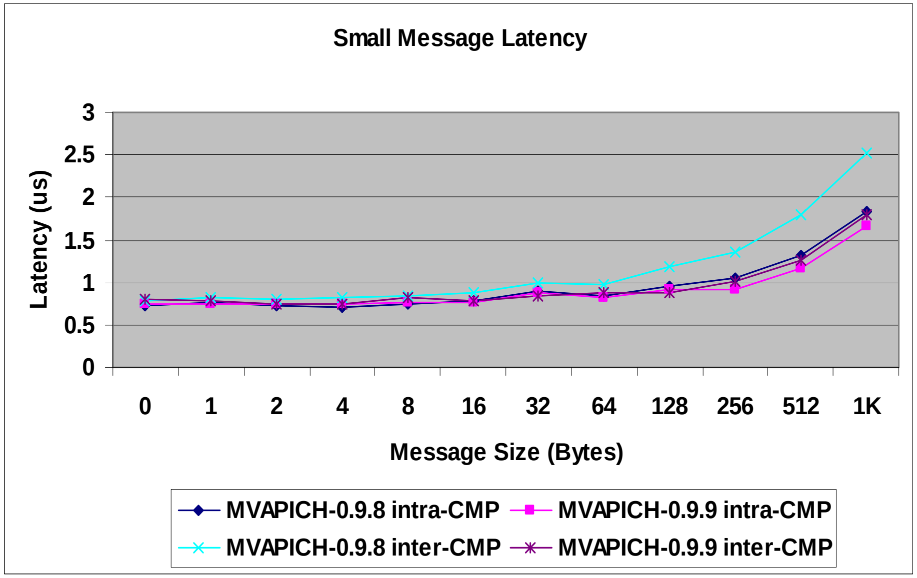
Performance Evaluation with Multi-Pathing

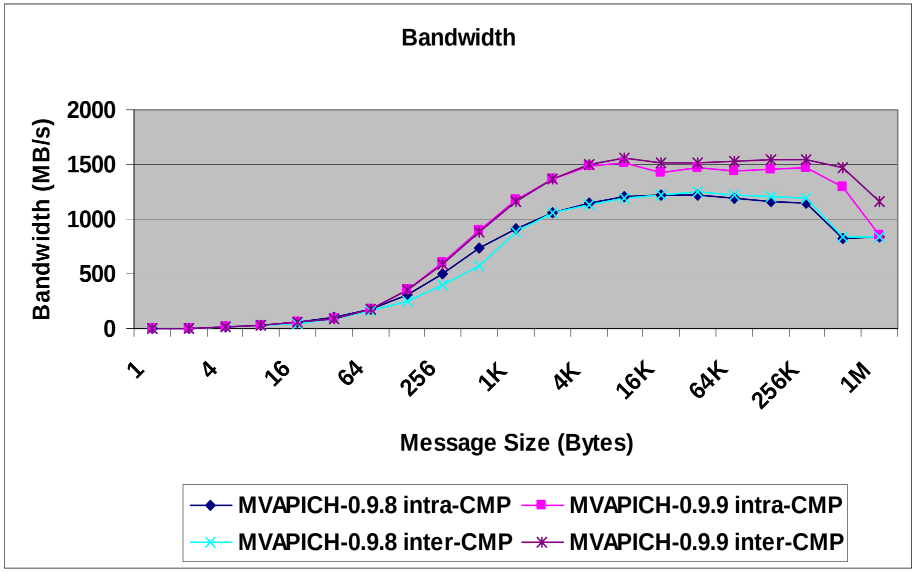




- Multi-pathing with LMC improves the performance of MPI_Alltoall
  - Reduces the latency to half for 64x1 configuration
- Fourier Transform benefits speedup with multi-pathing
  - 11% with 32x1 and 18% with 64x1 for Class B
  - 11% with 32x1 and 14% with 64x1 for class C
- For clusters with multi-thousand scale, more benefits are expected

Improvement of Intra-node Communication in MVAPICH-0.9.9

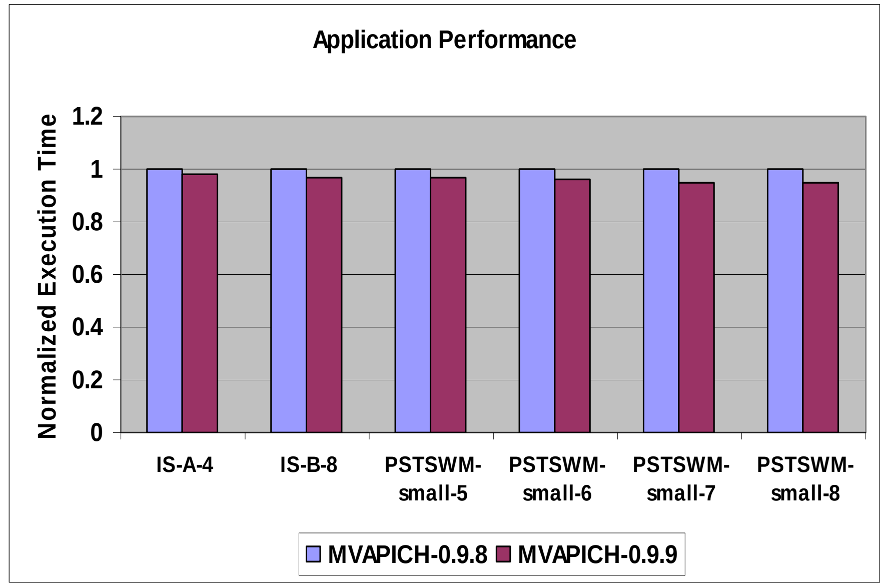




- Dual-core Opteron, 2.4GHz, 1M L2 cache
- Latency improved by up to 30%
- Bandwidth improved by up to 50%

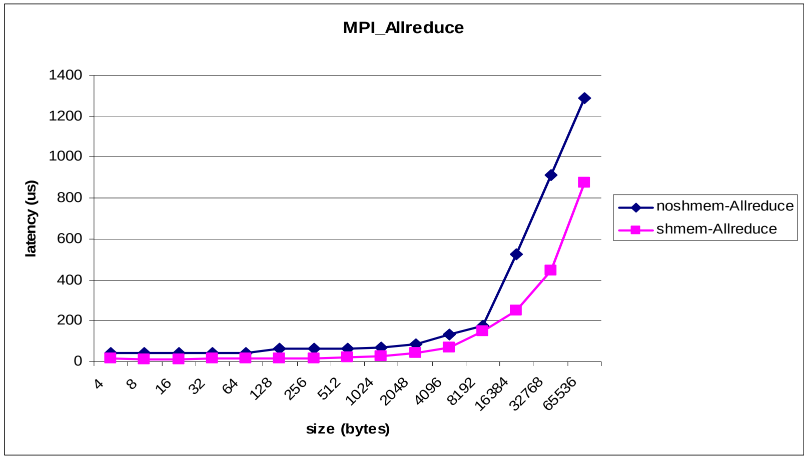
Intra-node Application Performance

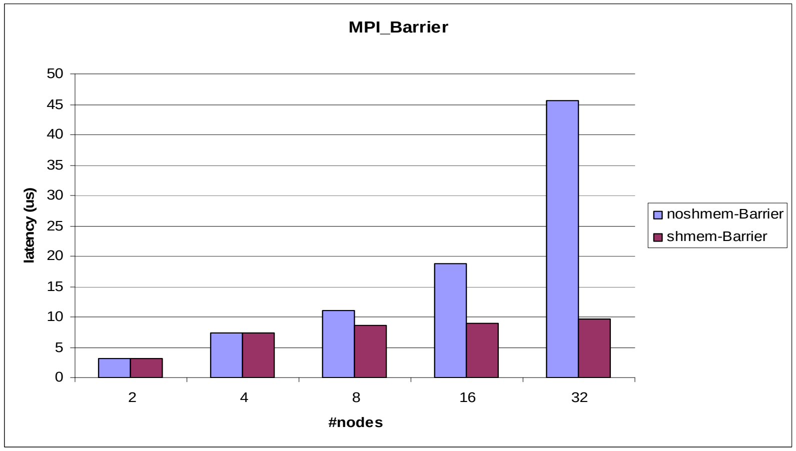
- Intel Bensley system with Clovertown processors
  - Quad-core, 2.33GHz, shared 4MB L2 cache
- MVAPICH-0.9.9 improves application performance by up to 5%



L. Chai, Q. Gao and D. K. Panda, Understanding the Impact of Multi-Core Architecture in Cluster Computing: A Case Study with Intel Dual-Core System, The 7th IEEE International Symposium on Cluster Computing and the Grid (CCGrid 2007), May 2007.

Multi-core Aware Shared Memory-based Collectives





- Four Intel Quad-core systems with dual sockets (32 processors)
- Performance improvement
  - up to 3.2 times for small message and 33% for large message (MPI_Allreduce)
  - up to 4.5 times for (MPI_Barrier)

Amith R. Mamidala, Debraj De, Abhinav Vishnu, Sundeep Narravula and D. K. Panda, Scalable Collective Communication for next generation Multicore clusters with InfiniBand,, under review

Efficient Shared Memory and RDMA based design for MPI_Allgather over InfiniBand, Amith R. Mamidala, Abhinav Vishnu and D. K. Panda, EuroPVM/MPI, September 2006

Fault Tolerance

- Component failures are the norm in large-scale clusters
- Imposes need on reliability and fault tolerance
- Working along the following three angles
  - End-to-end Reliability with memory-to-memory CRC
    - Available in MVAPICH (since MVAPICH 0.9.7)
  - Application transparent Process Fault Tolerance with Efficient Checkpoint and Restart
    - Available in MVAPICH2 0.9.8
  - Reliable Networking with Automatic Path Migration (APM) utilizing Redundant Communication Paths
    - Will be available soon
  - uDAPL-based Network Fault-Tolerance
    - Will be available soon

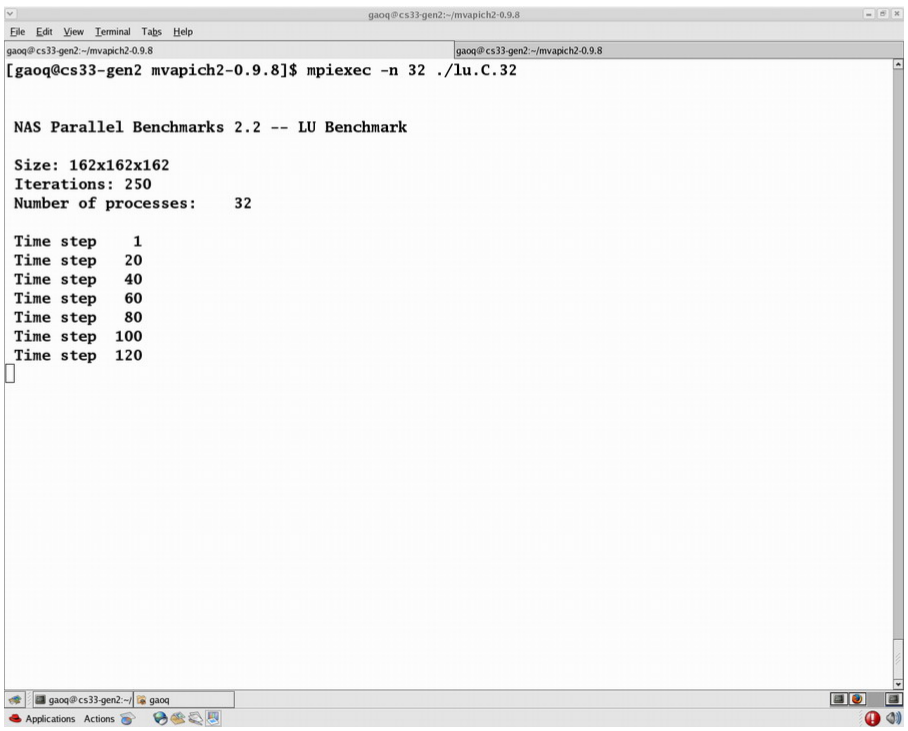
Checkpoint/Restart Support in MVAPICH2 0.9.8

- Process-level Fault Tolerance
  - User-transparent, system-level checkpointing
  - Based on BLCR from LBNL to take coordinated checkpoints of entire program, including front end and individual processes
  - Designed novel schemes to
    - Coordinate all MPI processes to drain all in flight messages in IB connections
    - Store communication state and buffers, etc. while taking checkpoint
    - Restarting from the checkpoint
  - Tested with NFS, PVFS2, Ext3 (local disk)

A Running Example (Cont.)

Terminal A:

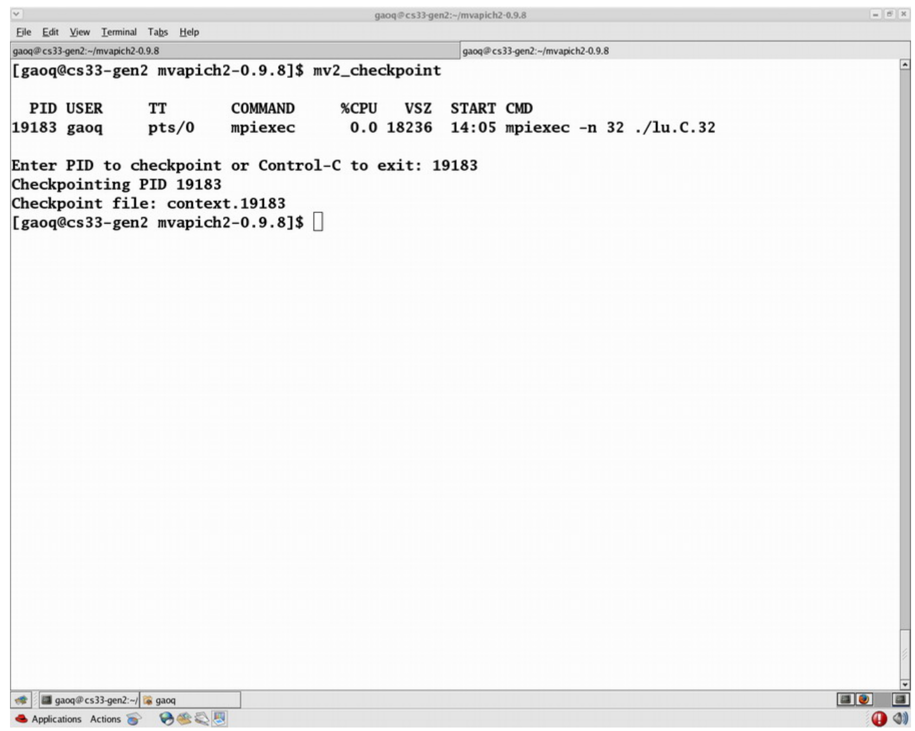
LU is running



1

Terminal B:

Now, Take checkpoint



2

A Running Example (Cont.)

Terminal A:

LU is not affected.

Stop it using CTRL-C

```
[gaoq@cs33-gen2 mvapich2-0.9.8]$ mpiexec -n 32 ./lu.C.32


 NAS Parallel Benchmarks 2.2 -- LU Benchmark

 Size: 162x162x162
 Iterations: 250
 Number of processes:     32

 Time step    1
 Time step   20
 Time step   40
 Time step   60
 Time step   80
 Time step  100
 Time step  120
 Time step  140
CTRL+C Caught... exiting
[gaoq@cs33-gen2 mvapich2-0.9.8]$
```

3

Terminal B:

Then, restart from the checkpoint

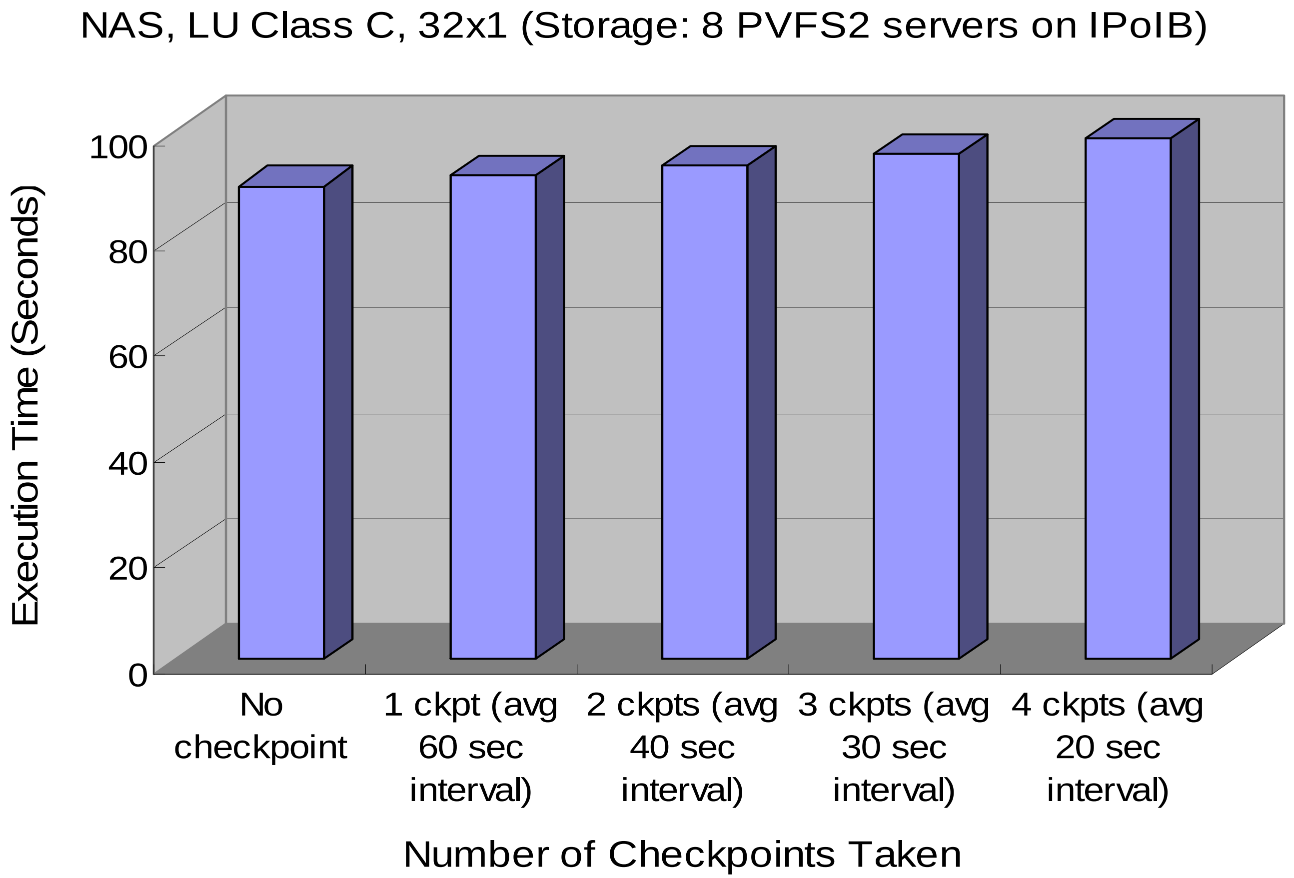
```
[gaoq@cs33-gen2 mvapich2-0.9.8]$ mv2_checkpoint

  PID USER      TT       COMMAND      %CPU    VSZ  START CMD
19183 gaoq      pts/0    mpiexec       0.0 18236  14:05 mpiexec -n 32 ./lu.C.32

Enter PID to checkpoint or Control-C to exit: 19183
Checkpointing PID 19183
Checkpoint file: context.19183
[gaoq@cs33-gen2 mvapich2-0.9.8]$ cr_restart context.19183
mpiexec_cs33-gen2 (mpiexec 334): mpiexec: Restarting
 Time step  140
 Time step  160
 Time step  180
 Time step  200
 Time step  220
```

4

Checkpoint/Restart Performance with PVFS2



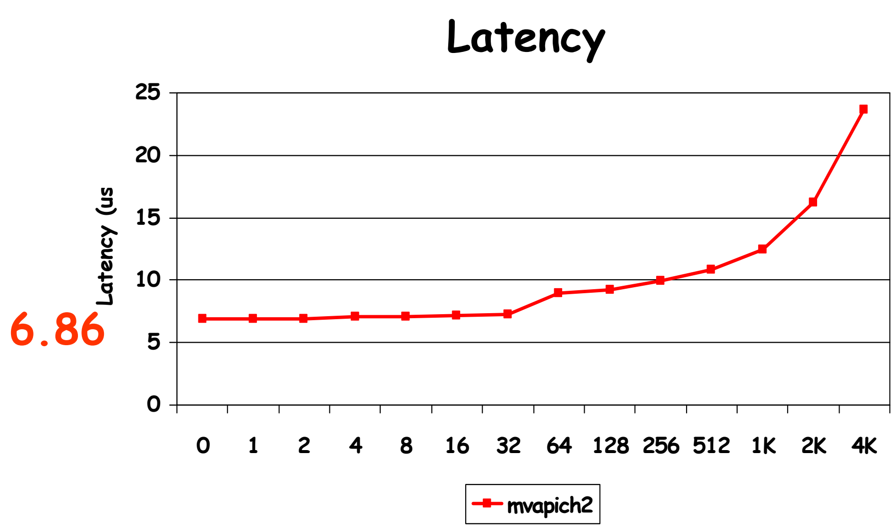
RDMA CM and iWARP Support

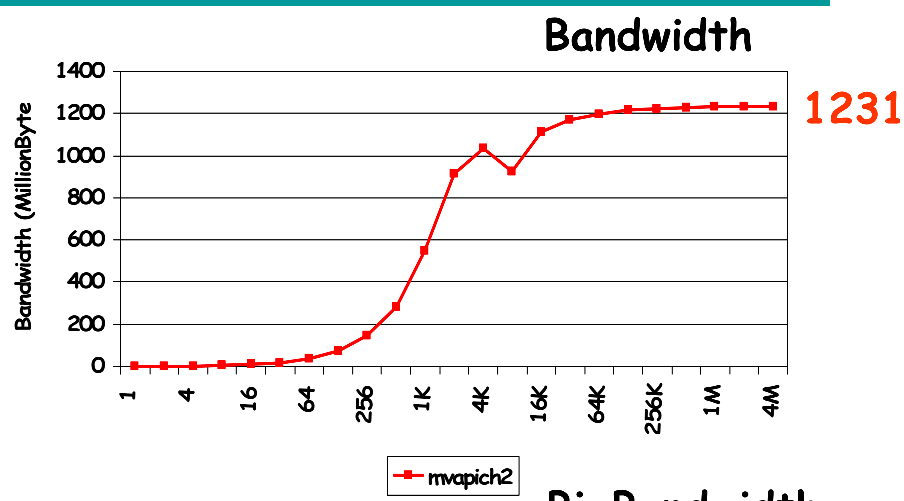
- Available in MVAPICH2 0.9.8
- RDMA CM is supported for both
  - IB
  - iWARP
- iWARP support is tested with
  - Chelsio adapter

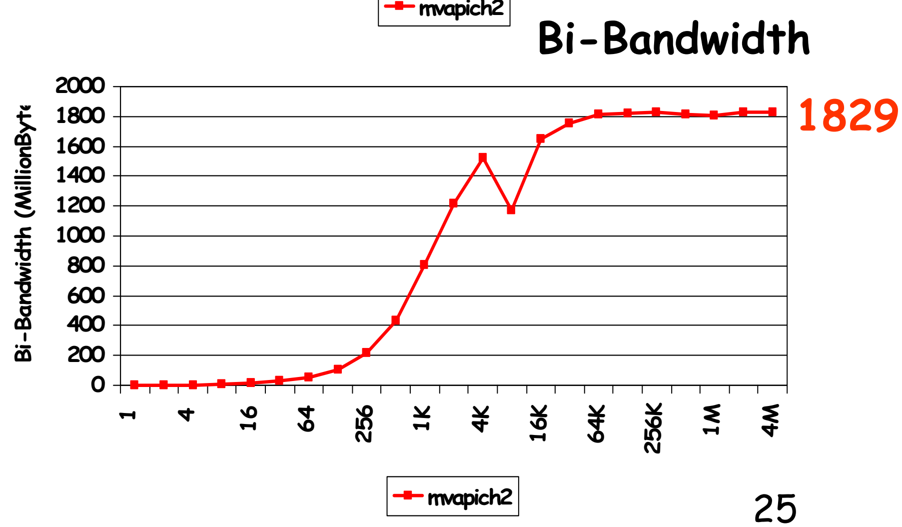
MVAPICH2-iWARP Performance

- Hardware:
  - 2.33 GHz Intel Xeon quad dual-core systems
  - 4 GB memory
  - Chelsio T3B2 Adapters (FW 4.0)
  - Fulcrum 10GigE evaluation switch
- Software
  - Redhat 4 u4 (2.6.20.9)
  - OFED 1.2
  - MVAPICH2 0.9.8

MPI Two-Sided Performance

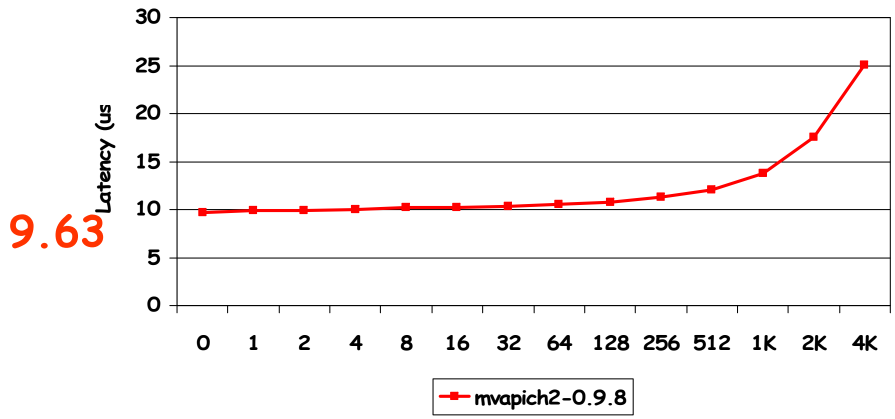




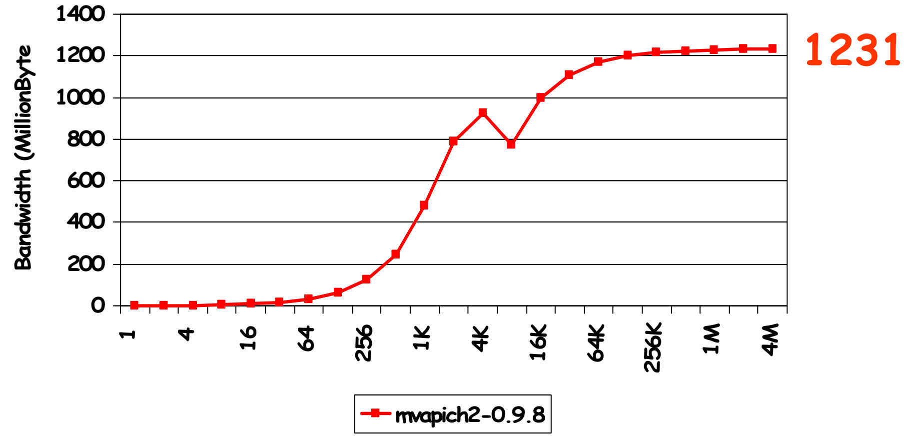


MPI Put Performance

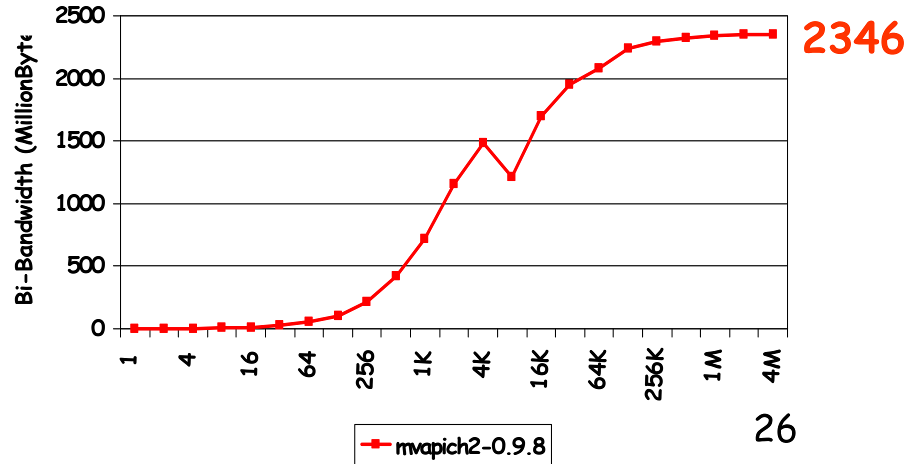
Latency



Bandwidth



Bi-Bandwidth



OSU Benchmarks

- Expanding on OSU benchmarks
  - Multi-core platforms
- Added a new Multiple-pair Bandwidth and Message Rate benchmark with MVAPICH 0.9.9 release
- Working on other benchmarks
  - Multi-pair latency test
  - Multi-pair bidirectional bandwidth test
  - Overlap test
  - Extended one-sided tests
- Being done as a part of the following project
  - Center for Performance Evaluation of Cluster Networking and I/O Technologies (PECNIT)
- Funded through the AVETEC/DICE Program

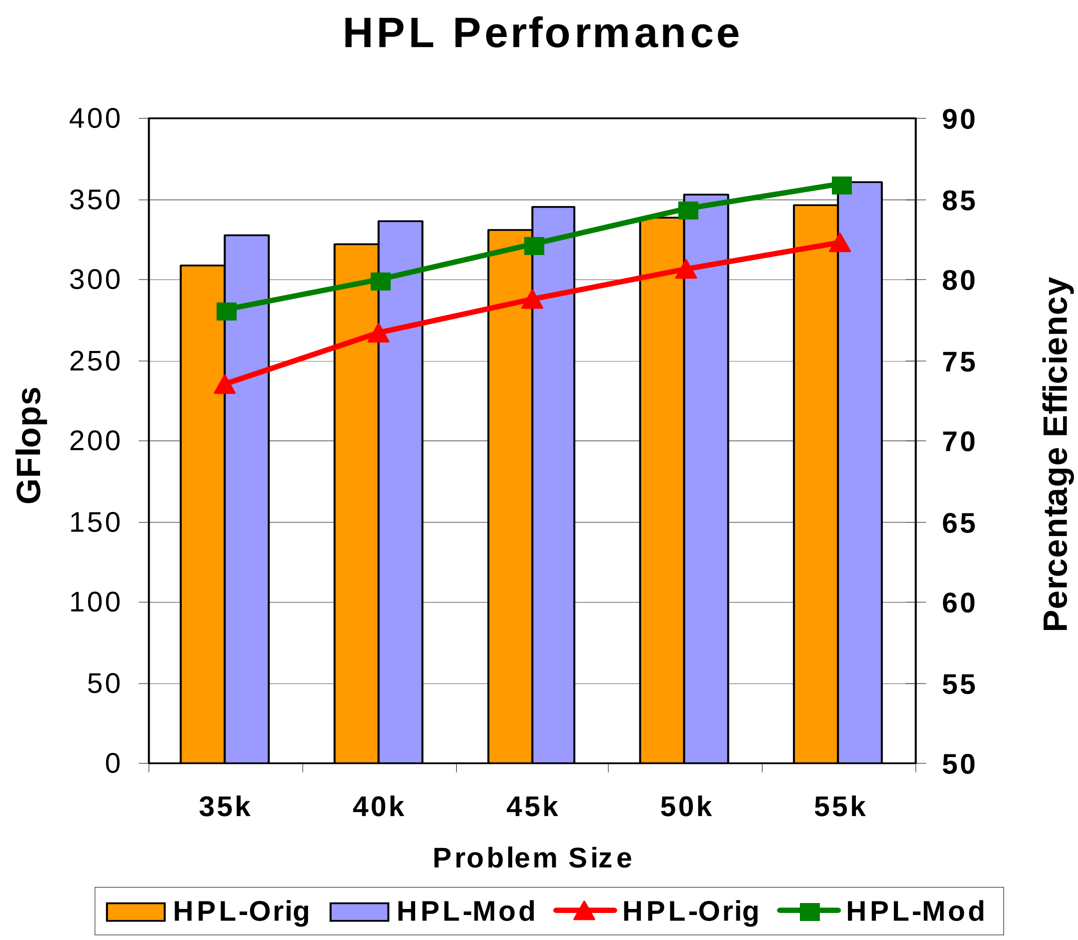
  www.avetec.org/dice/index.htm

Presentation Overview

- Overview of MVAPICH/MVAPICH2 Project
- Selected Features of the latest releases
  - Message Coalescing and Memory Scalability
  - ConnectX Performance
  - Congestion Avoidance with Multi-Pathing
  - Multi-core-aware Point-to-point
  - Multi-core-aware Optimized Collectives
  - Checkpoint/Restart
  - RDMA CM and iWARP
  - OSU Benchmarks
- Upcoming Features and Issues
  - Overlap of Computation and Communication
  - Automatic Path Migration (APM)
  - UD-based Design
  - Multi-Network Support using uDAPL
  - MVAPICH-PSM (QLogic) Implementation and Performance
- Conclusions

Enhancing Overlap Capabilities in HPL

- MVAPICH has RDMA Read support
- RDMA Read with Interrupt can provide
  - Asynchronous progress
  - Overlap of computation and communication
- Have enhanced HPL to add overlapping at the sender side
- Results on 32 dual dual-core nodes with IB DDR
- MPI overlap increase the overall application efficiency by 5-6%
- Improvement rate consistent with increasing problem size



Network-Level Fault Tolerance with APM

- Designed a solution using InfiniBand Automatic Path Migration (APM) Hardware mechanism
  - Utilizes Redundant Communication Paths
    - Multiple Ports
    - LMC
- OFED 1.2 is providing APM support
- Will be available with MVAPICH2-1.0

A. Vishnu, A. Mamidala, S. Narravula and D. K. Panda, Automatic Path Migration over InfiniBand: Early Experiences, Third International Workshop on System Management Techniques, Processes, and Services, to be held in conjunction with IPDPS '07, March 2007.

UD-based Design

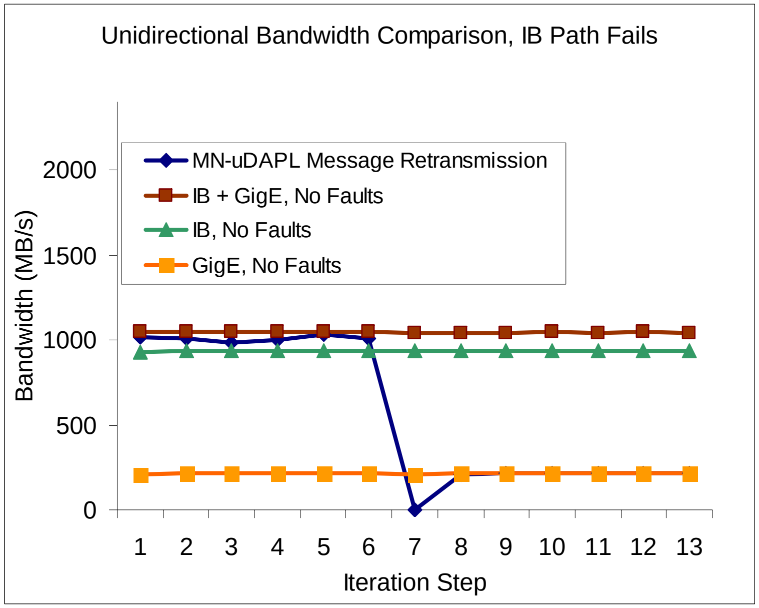
- For very large clusters when the number of connections is increased, memory usage can still be significant
- Designed a high-performance MPI over the Unreliable Datagram (UD) transport
- Has been tested up to 4K processes on LLNL Atlas system with applications
- Benefits:
  - Near-constant memory usage per process as number of peers increase
    - Only 38 MB for 4K processes
    - 43 MB for 16K processes (estimated)
  - Increased performance when communicating with many peers (due to better HCA cache utilization)
- Will be available in the upcoming MVAPICH 1.0 release

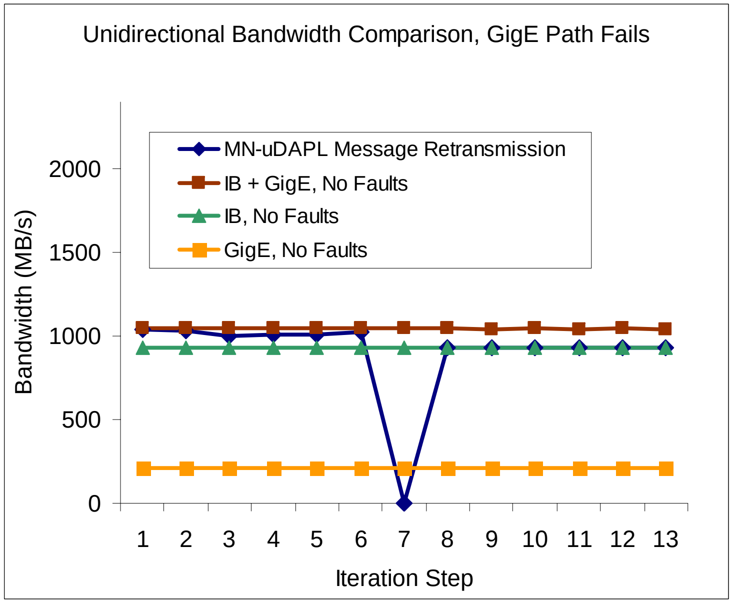
M. Koop, S. Sur, Q. Gao and D. K. Panda, High Performance MPI Design Using Unreliable Datagram for Ultra-Scale InfiniBand Clusters, 21st ACM International Conference on Supercomputing (ICS07), June 2007.

Multi-Network Support using uDAPL

- Network-independent interfaces like uDAPL are being available
- Can we design a unified MPI framework, with low overhead, flexibility, and adaptivity to support following
  - Network Heterogeneity
  - Network Failover
  - Asynchronous recovery of previously failed paths

Network Failover





- The peak bandwidth achieved after failover is same as achievable in no-faults case

A. Vishnu, P. Gupta, A. Mamidala and D. K. Panda, A Software Based Approach for Providing Network Fault Tolerance in Clusters Using the uDAPL Interface: MPI Level Design and Performance Evaluation, SC '06, November 2006

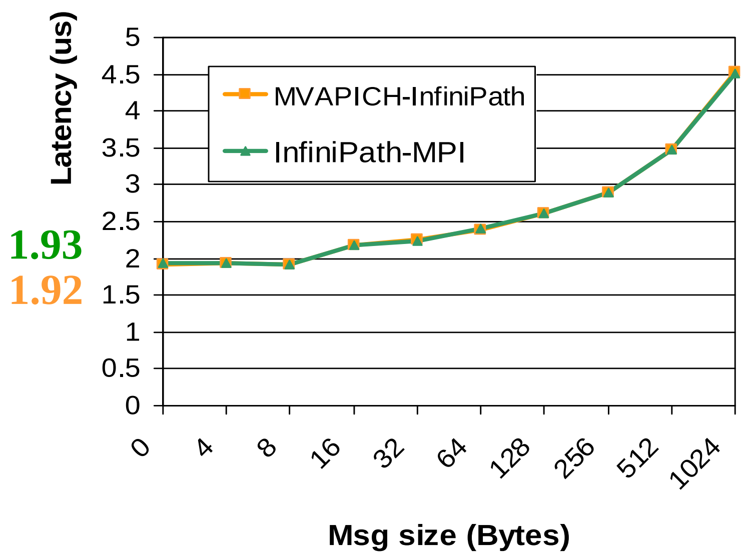
MVAPICH-PSM Design

- Designed support for MVAPICH-PSM
- Performance matches very close to InfiniPath-MPI (QLogic's proprietary stack)
- Will be available with MVAPICH 1.0

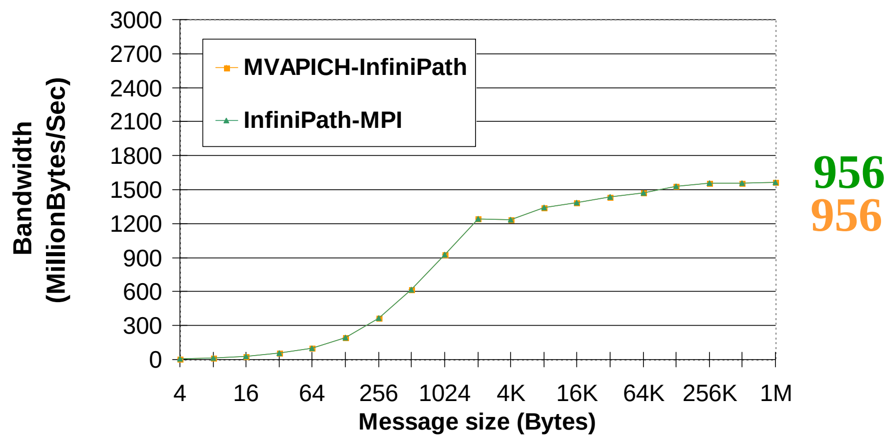
MVAPICH-PSM Performance

- Two different setups
- Setup1 (Intel Quad-core with PCI-Express)
  - Intel(R) Xeon(R) CPU E5345 @ 2.33GHz
  - FB DIMM 533MHz
  - InfiniPath Release 2.0
  - gcc (GCC) 3.4.6
  - RH AS4u4
- Setup2 (AMD Opteron with HT)
  - AMD Opteron(tm) Processor 254 @ 2.8Ghz
  - InfiniPath Release 2.0
  - gcc (GCC) 3.4.6
  - RH AS4u4
- Numbers are taken with the switch being present in both setups

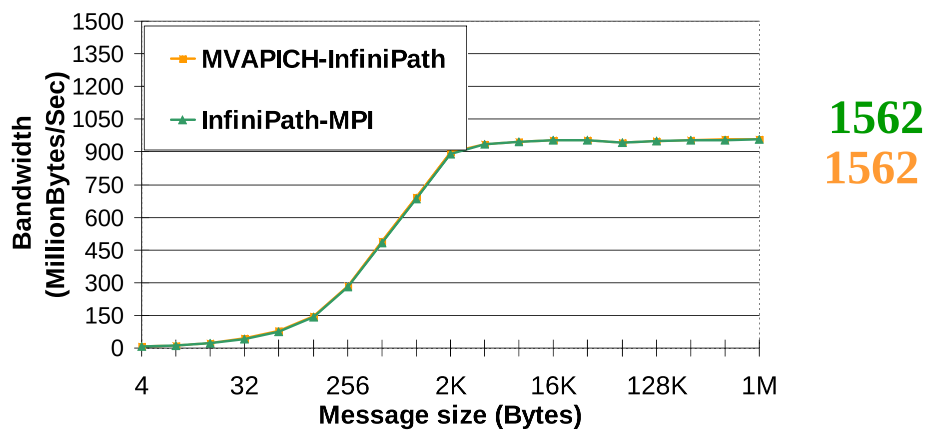
MVAPICH-PSM Performance: Intel Quad-core



Small Message Latency

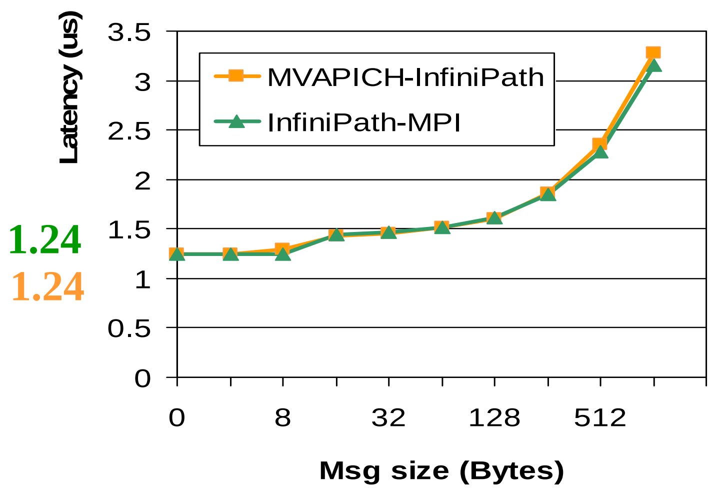


Uni-Directional Bandwidth

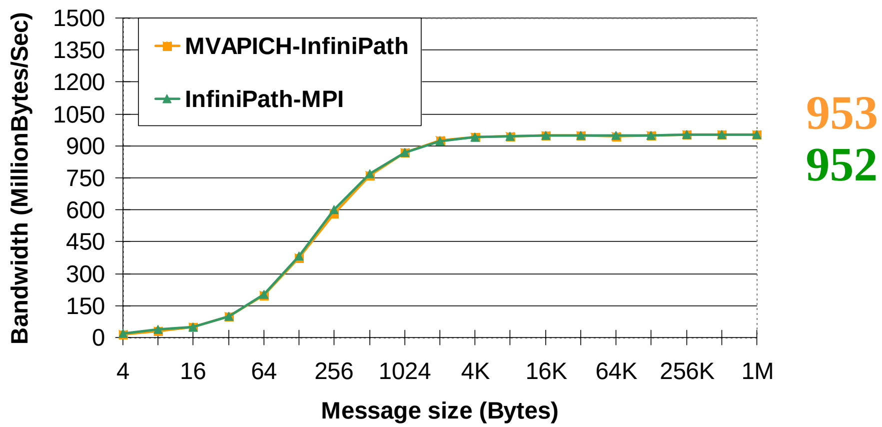


Bi-Directional Bandwidth

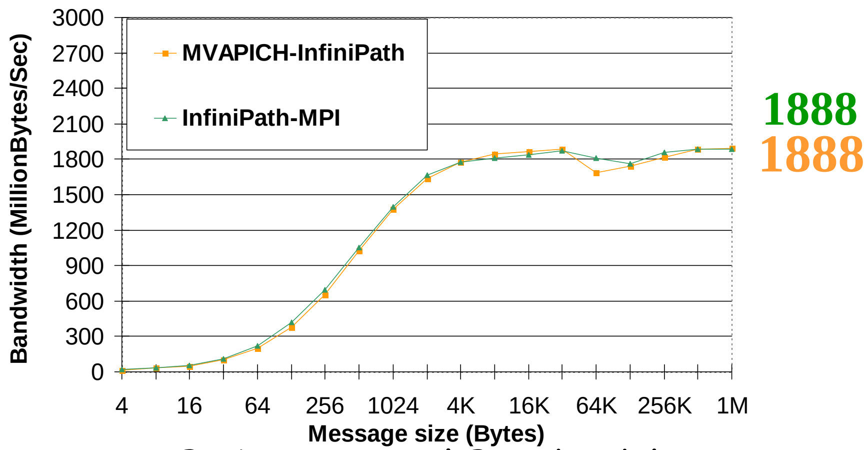
MVAPICH-PSM Performance: AMD Opteron



Small Message Latency



Uni-Directional Bandwidth



Bi-Directional Bandwidth

Presentation Overview

- Overview of MVAPICH/MVAPICH2 Project
- Selected Features of the latest releases
  - Message Coalescing and Memory Scalability
  - ConnectX Performance
  - Congestion Avoidance with Multi-Pathing
  - Multi-core-aware Point-to-point
  - Multi-core-aware Optimized Collectives
  - Checkpoint/Restart
  - RDMA CM and iWARP
  - OSU Benchmarks
- Upcoming Features and Issues
  - Overlap of Computation and Communication
  - Automatic Path Migration (APM)
  - UD-based Design
  - Multi-Network Support using uDAPL
  - MVAPICH-PSM (QLogic) Implementation and Performance
- Conclusions

Conclusions

- MVAPICH and MVAPICH2 are being widely used in stable production IB clusters delivering best performance and scalability
- Also enabling clusters with iWARP support
- The user base stands at more than 495 organizations
- New features for scalability, high performance and fault tolerance support are aimed to deploy large-scale clusters (20K-50K) nodes in the near future

Acknowledgements

- Current Students
  - Lei Chai (PhD)
  - Qi Gao (PhD)
  - Wei Huang (PhD)
  - Matthew Koop (PhD)
  - Amith Mamidala (PhD)
  - Sundeep Narravula (PhD)
  - Ranjit Noronha (PhD)
  - G. Santhanaraman (PhD)
  - Sayantan Sur (PhD)
  - K. Vaidyanathan (PhD)
  - Abhinav Vishnu (PhD)
- Past Students
  - P. Balaji (PhD)
  - D. Buntinas (PhD)
  - Sitha Bhagvat (MS)
  - B. Chandrasekharan (MS)
  - Weihang Jiang (MS)
  - Sushmita Kini (MS)
  - S. Krishnamoorthy (MS)
  - Jiuxing Liu (PhD)
  - Jiesheng Wu (PhD)
  - Weikuan Yu (PhD)
- Current Programmers
  - Shaun Rowland
  - Jonathan Perkins

Web Pointers

MVAPICH

MVAPICH Web Page
http://mvapich.cse.ohio-state.edu/

E-mail: panda@cse.ohio-state.edu