High Performance MPI on IBM 12x InfiniBand Architecture

Abhinav Vishnu,
Brad Benton[1] and
Dhabaleswar K. Panda

{vishnu, panda} @ cse.ohio-state.edu
{brad.benton}@us.ibm.com[1]

Presentation Road-Map

- Introduction and Motivation
- Background
- Enhanced MPI design for IBM 12x Architecture
- Performance Evaluation
- Conclusions and Future Work

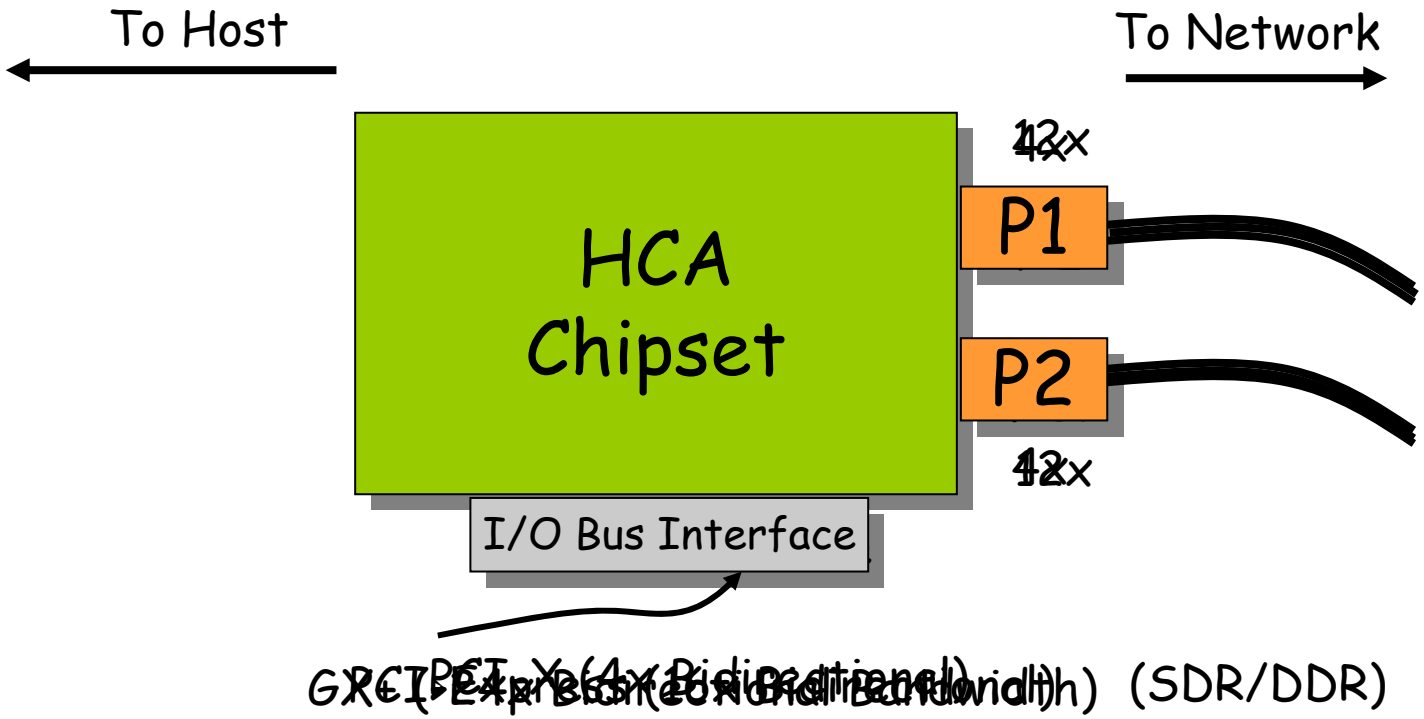
Introduction and Motivation

- Demand for more compute power is driven by Parallel Applications
  - Molecular Dynamics (NAMD), Car Crash Simulations (LS-DYNA) , ... , ...
- Cluster sizes have been increasing forever to meet these demands
  - 9K proc. (Sandia Thunderbird, ASCI Q)
  - Larger scale clusters are planned using upcoming multi-core architectures
- MPI is used as the primary programming model for writing these applications

Emergence of InfiniBand

- Interconnects with very low latency and very high throughput have become available
  - InfiniBand, Myrinet, Quadrics …
- InfiniBand
  - High Performance and Open Standard
  - Advanced Features
- PCI-Express Based InfiniBand Adapters are becoming popular
  - 8X (1X ~ 2.5 Gbps) with Double Data Rate (DDR) support
  - MPI Designs for these Adapters are emerging
- Compared to PCI-Express, GX+ I/O Bus Based Adapters are also emerging
  - 4X and 12X link support

InfiniBand Adapters



MPI for PCI-Express based are coming up

IBM 12x InfiniBand Adapters on GX+ are coming up

Problem Statement

- How do we design an MPI with low overhead for IBM 12x InfiniBand Architecture?

- What are the performance benefits of enhanced design over the existing designs?
  - Point-to-point communication
  - Collective communication
  - MPI Applications

Presentation Road-Map

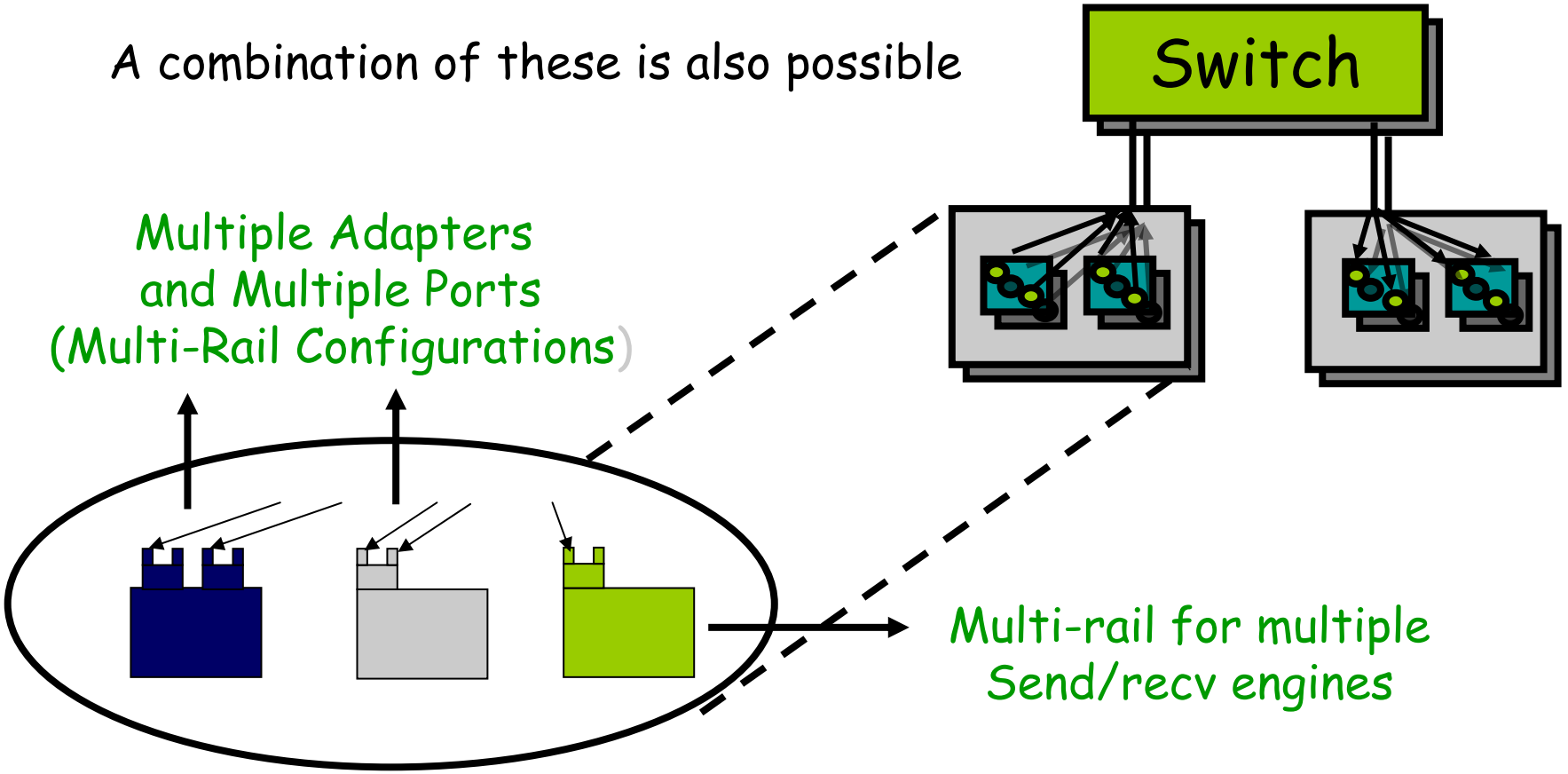
- Introduction and Motivation
- Background
- Enhanced MPI design for IBM 12x Architecture
- Performance Evaluation
- Conclusions and Future Work

Overview of InfiniBand

- An interconnect technology to connect I/O nodes and processing nodes
- InfiniBand provides multiple transport semantics
  - Reliable Connection
    - Supports reliable notification and Remote Direct Memory Access (RDMA) ✓
  - Unreliable Datagram
    - Data delivery is not reliable, send/recv is supported
  - Reliable Datagram
    - Currently not implemented by Vendors
  - Unreliable Connection
    - Notification is not supported
- InfiniBand uses a queue pair (QP) model for data transfer
  - Send queue (for send operations)
  - Receive queue (not involved in RDMA kind of operations)

MultiPathing Configurations

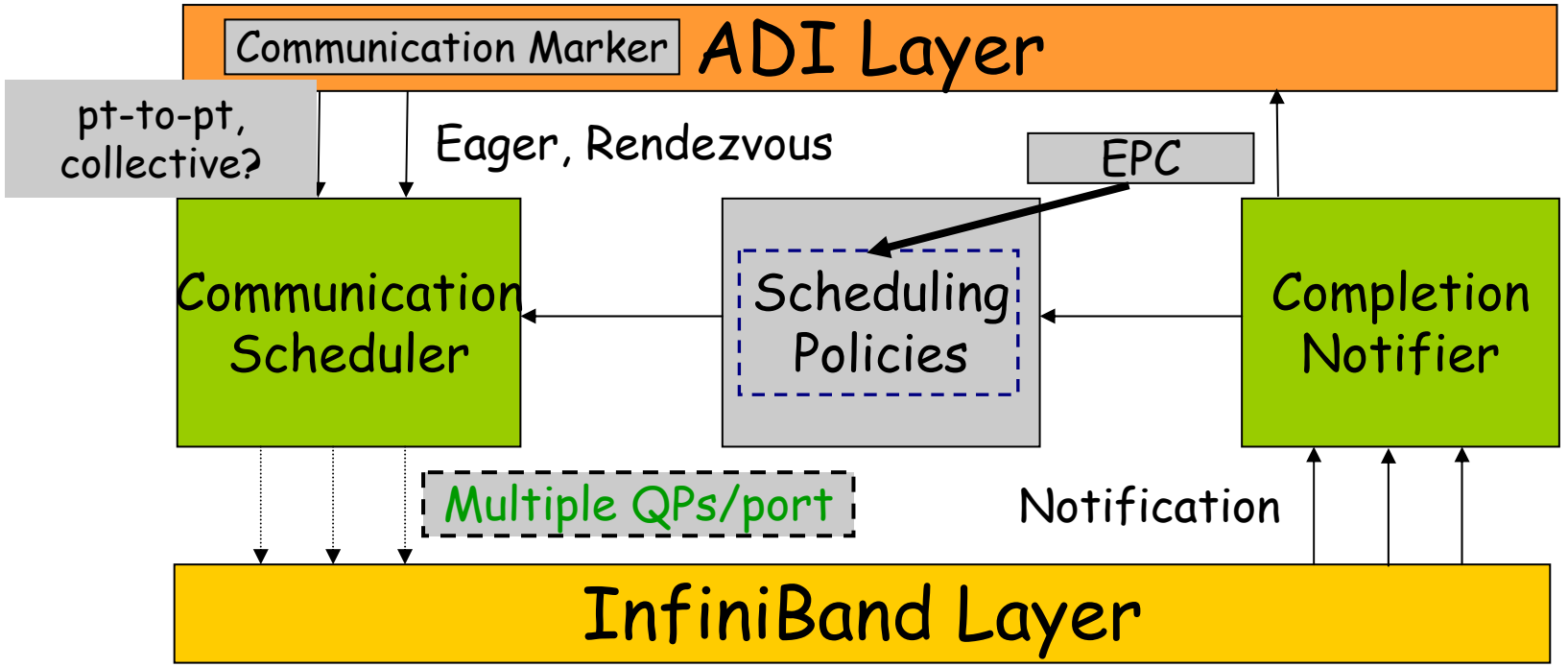
A combination of these is also possible



Presentation Road-Map

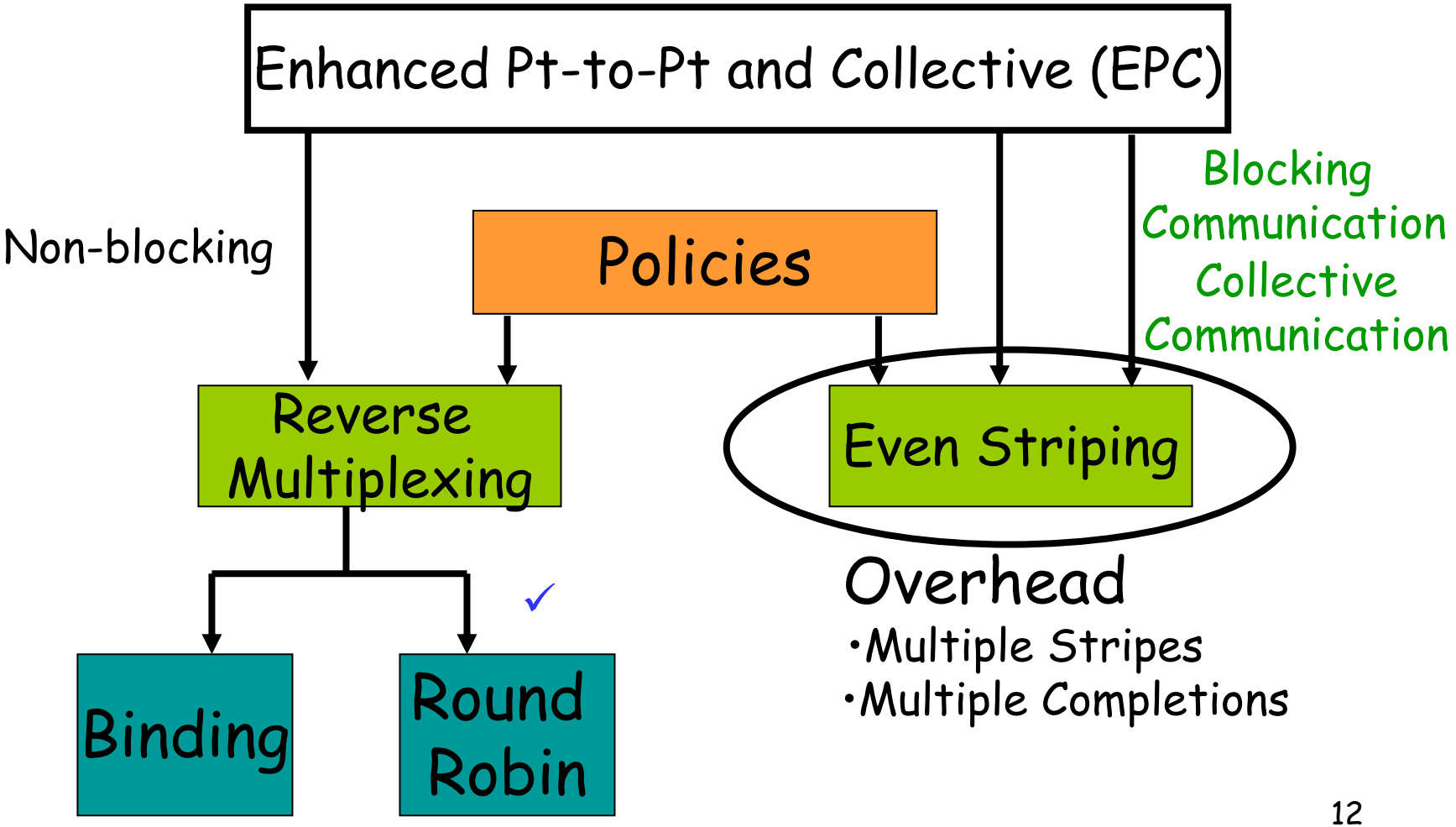
- Introduction and Motivation
- Background
- Enhanced MPI design for IBM 12x Architecture
- Performance Evaluation
- Conclusions and Future Work

MPI Design for 12x Architecture



*Jiuxing Liu, **Abhinav Vishnu** and Dhabaleswar K. Panda. ,"*
Building Multi-rail InfiniBand Clusters:
MPI-level Design and Performance Evaluation,". SuperComputing 2004

Discussion on Scheduling Policies



EPC Characteristics

pt-2-pt	blocking	striping
	non-blocking	round-robin
collective		striping

- For small messages, round robin policy is used
  - Striping leads to overhead for small messages

MVAPICH/MVAPICH2

- We have used MVAPICH as our MPI framework for the enhanced design
- MVAPICH/MVAPICH2
  - High Performance MPI-1/MPI-2 implementation over InfiniBand and iWARP
  - Has powered many supercomputers in TOP500 supercomputing rankings
  - Currently being used by more than 450 organizations (academia and industry worldwide)
  - http://nowlab.cse.ohio-state.edu/projects/mpi-iba
- The enhanced design is available with MVAPICH
  - Will become available with MVAPICH2 in the upcoming releases

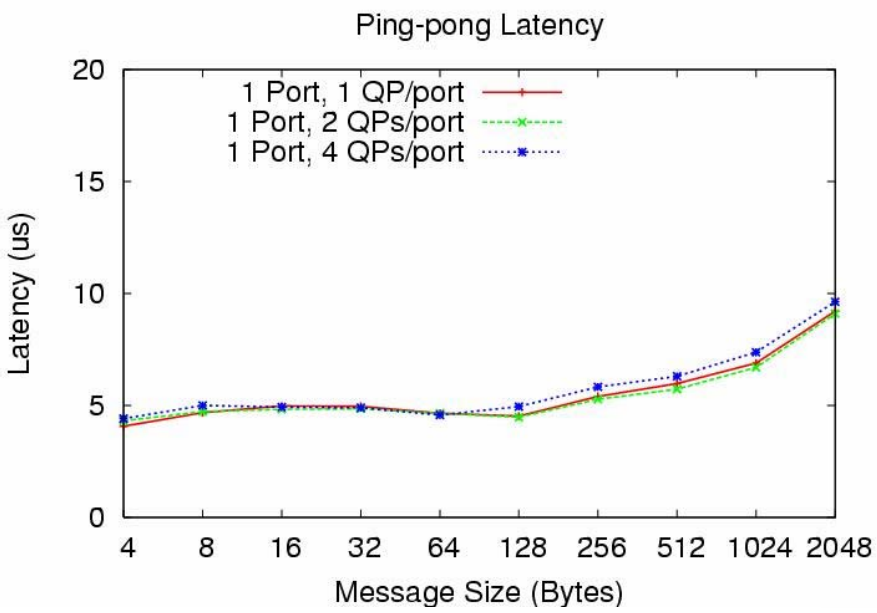
Presentation Road-Map

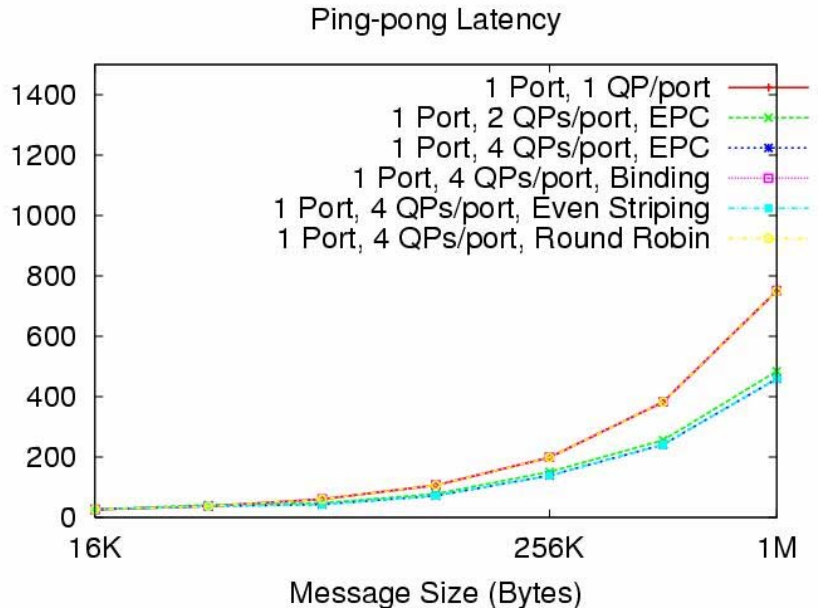
- Introduction and Motivation
- Background
- Enhanced MPI design for IBM 12x Architecture
- Performance Evaluation
- Conclusions and Future Work

Experimental TestBed

- The Experimental Test-Bed consists of:
  - Power5 based systems with SLES9 SP2
  - GX+ at 950 MHz clock speed
  - 2.6.9 Kernel Version
  - 2.8 GHz Processor with 8 GB of Memory
  - TS120 switch for connecting the adapters
- One port per adapter and one adapter is used for communication
  - The objective is to see the benefit with using only one physical port

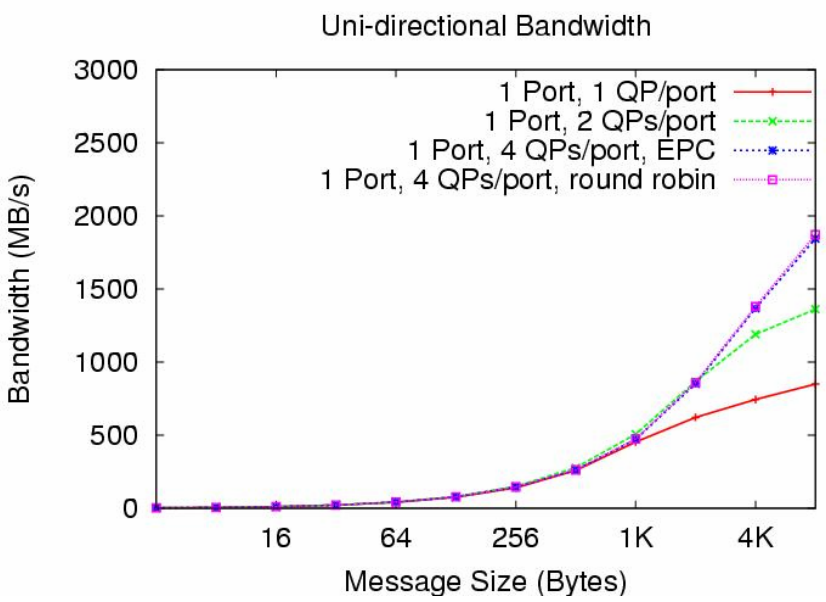
Ping-Pong Latency Test

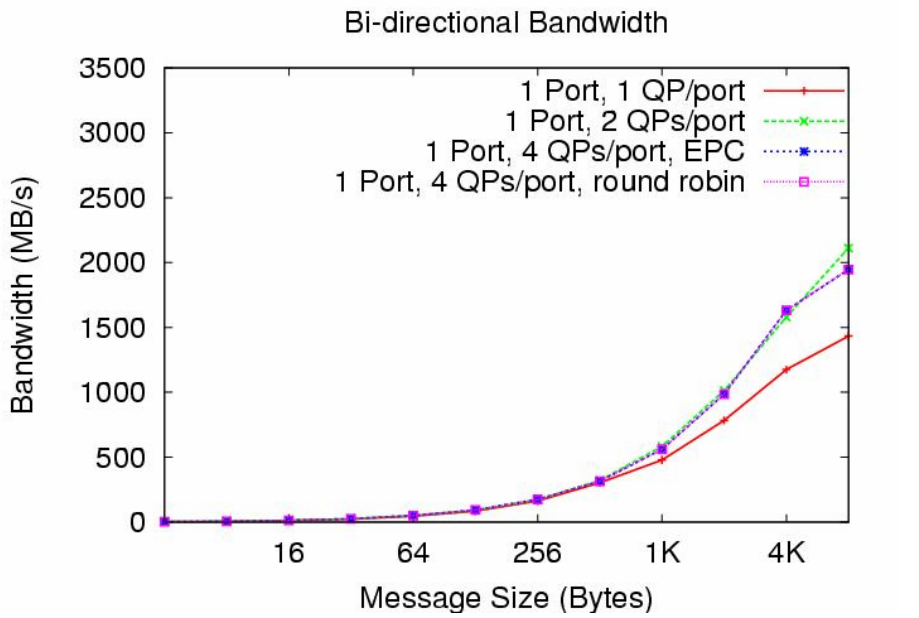




- EPC adds insignificant overhead to the small message latency
- Large Message latency reduces by 41% using EPC with IBM 12x architecture

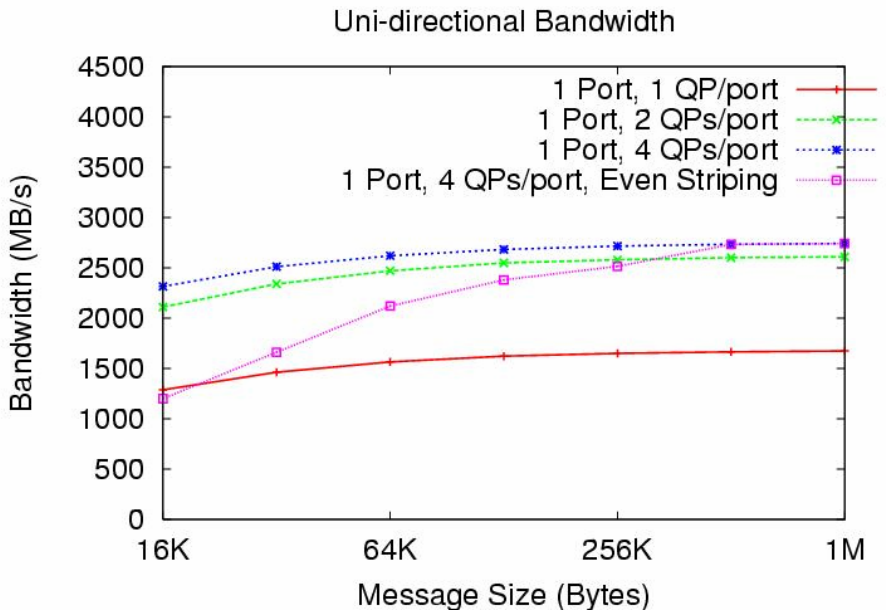
Small Messages Throughput

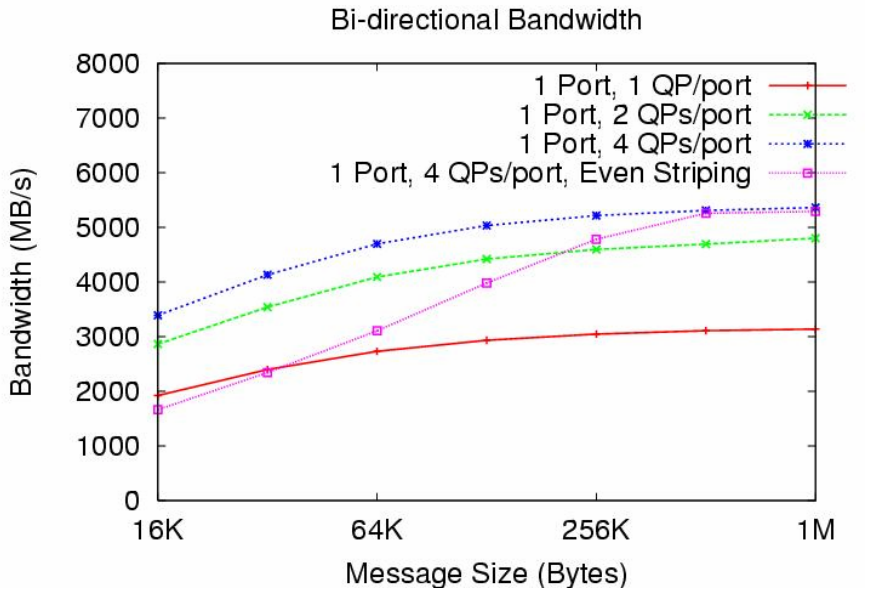




- Unidirectional bandwidth doubles for small messages using EPC
- Bidirectional bandwidth does not improve with increasing number of QPs due to the copy bandwidth limitation

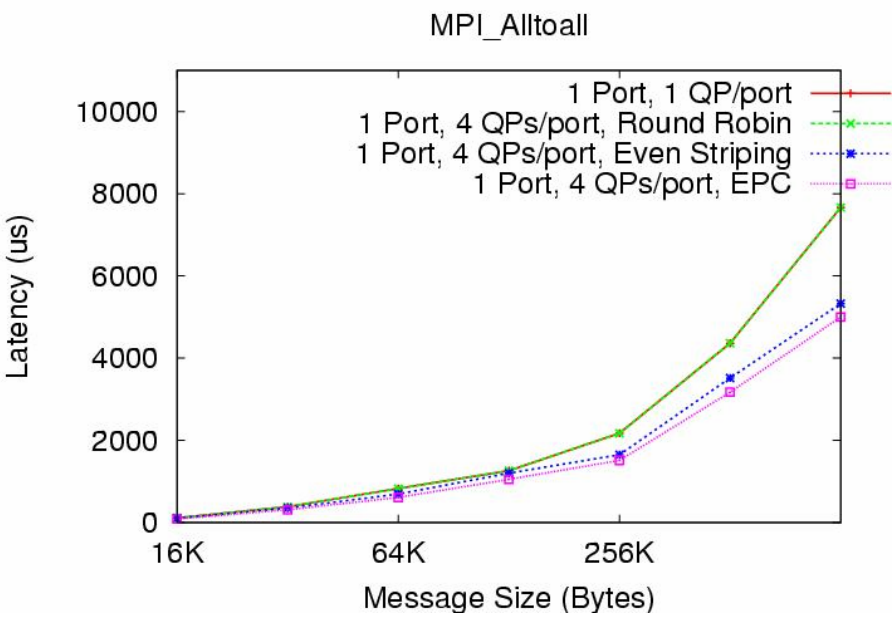
Large Messages Throughput

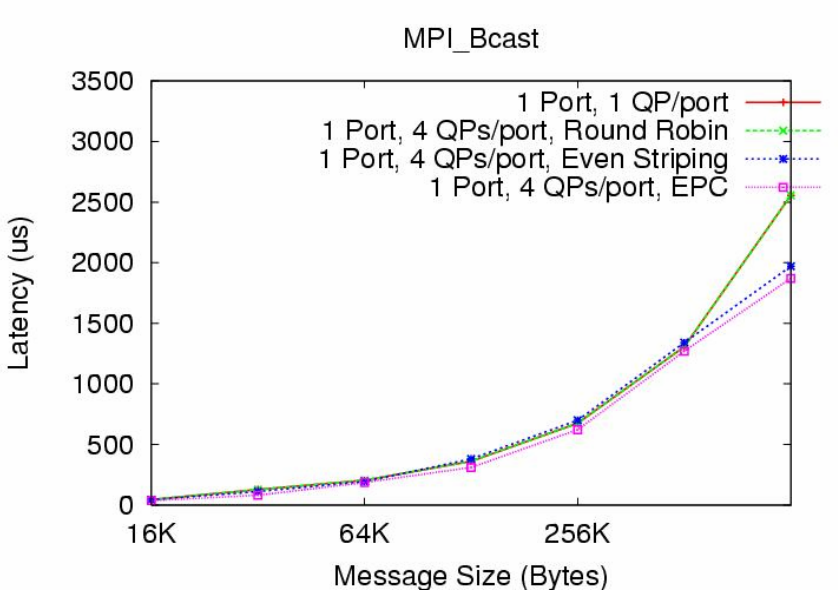




- EPC improves the uni-directional and bi-directional throughput significantly for medium size messages
- We can achieve a peak unidirectional bandwidth of 2731 MB/s and bidirectional bandwidth of 5421 MB/s

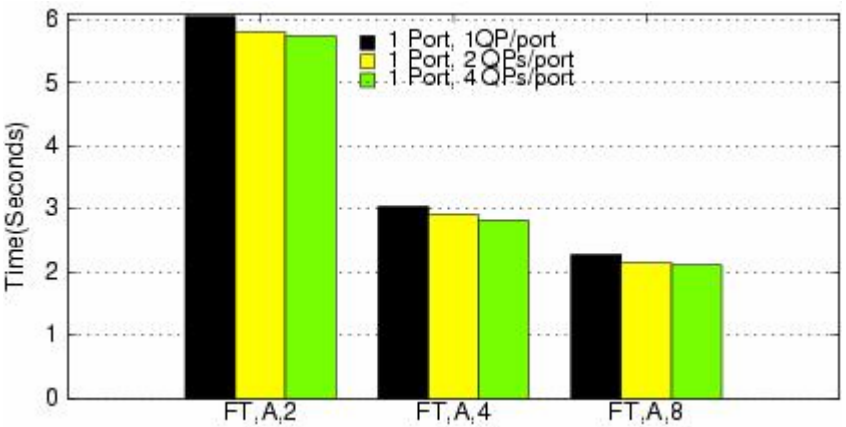
Collective Communication

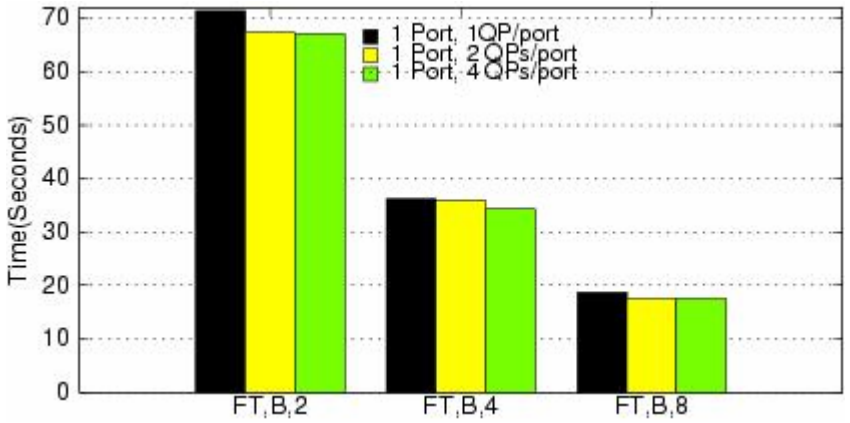




- MPI_Alltoall shows significant benefits for large messages
- MPI_Bcast shows more benefits for very large messages

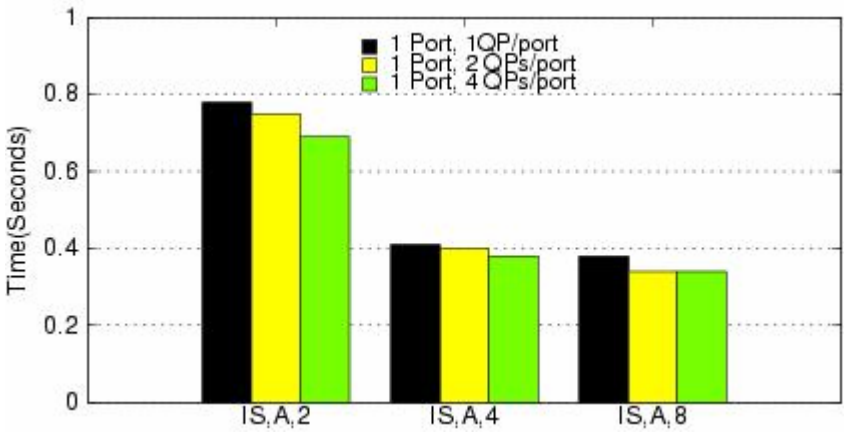
NAS Parallel Benchmarks

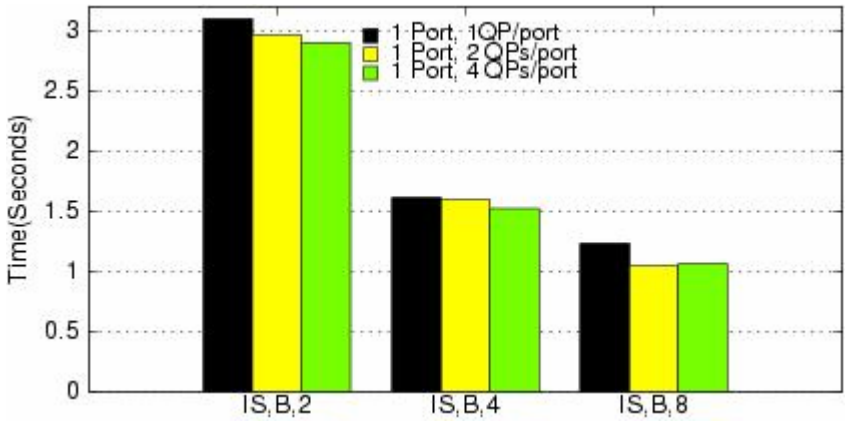




- For class A and class B problem sizes, x1 configuration shows improvement
- There is no degradation for other configurations on Fourier Transform

NAS Parallel Benchmarks





- Integer sort shows 7-11% improvement for x1 configurations
- Other NAS Parallel Benchmarks do not show performance degradation

Presentation Road-Map

- Introduction and Motivation
- Background
- Enhanced MPI design for IBM 12x Architecture
- Performance Evaluation
- Conclusions and Future Work

Conclusions

- We presented an enhanced design for IBM 12x InfiniBand Architecture
  - EPC (Enhanced Point-to-Point and collective communication)
- We have implemented our design and evaluated with Micro-benchmarks, collectives and MPI application kernels
- IBM 12x HCAs can significantly improve communication performance
  - 41% for ping-pong latency test
  - 63-65% for uni-directional and bi-directional bandwidth tests
  - 7-13% improvement in performance for NAS Parallel Benchmarks
  - We can achieve a peak bandwidth of 2731 MB/s and 5421 MB/s unidirectional and bidirectional bandwidth respectively

Future Directions

- We plan to evaluate EPC with multi-rail configurations on upcoming multi-core systems
  - Multi-port configurations
  - Multi-HCA configurations
- Scalability studies of using multiple QPs on large scale clusters
  - Impact of QP caching
  - Network Fault Tolerance

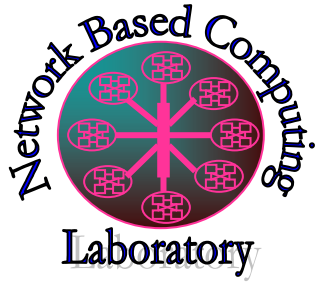
Acknowledgements

Our research is supported by the following organizations

- Current Funding support by

- Current Equipment support by


Web Pointers



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

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

E-mail: {vishnu, panda}@cse.ohio-state.edu,
brad.benton@us.ibm.com