

# Designing Next Generation Clusters: Evaluation of InfiniBand DDR/QDR on Intel Computing Platforms

**Hari Subramoni**, Matthew Koop and Dhableswar.  
K. Panda

Computer Science & Engineering Department  
The Ohio State University

# Outline

- Introduction and Motivation
- Background
- Experimental Setup
- Microbenchmark Level Evaluation
- Communication Balance
- Application Level Evaluation
- Conclusions and Future Work

# Motivation

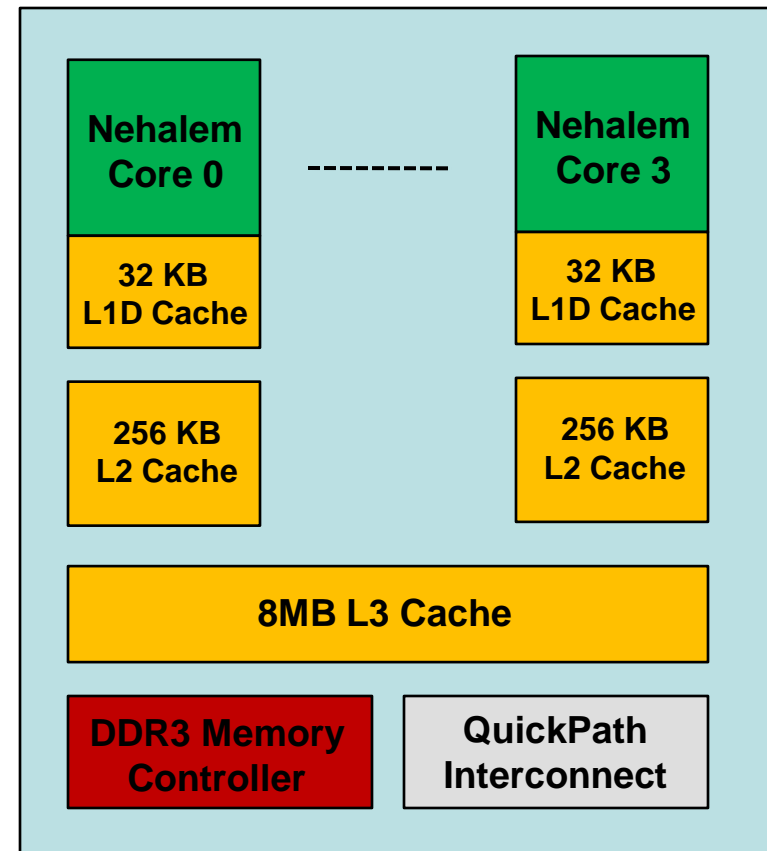
- Commodity clusters becoming more popular
- Balance needs to be maintained between
  - Computational and I/O capabilities
  - Intra-Node and Inter-Node communication performance
- This balance keeps changing with advent of new processor and interconnect technologies
- With the advent of Intel's latest Nehalem series of processors the balance has changed again
- Need exists to qualitatively and quantitatively explore how the new Nehalem processor, along with high speed InfiniBand interconnects, affects communication balance

# Outline

- Introduction and Motivation
- Background
- Experimental Setup
- Microbenchmark Level Evaluation
- Communication Balance
- Application Level Evaluation
- Conclusions and Future Work

# Background

- The Intel Nehalem Processor
  - First true quad core processor with L2 cache sharing
  - 45 nm manufacturing process
  - Integrated memory controller supporting multiple memory channels gives very high memory bandwidth
  - Uses QuickPath Interconnect Technology
  - HyperThreading allows execution of multiple threads per core in a seamless manner
  - Turbo boost technology allows automatic over clocking of processors



# Background (Cont)

- InfiniBand Architecture
  - An industry standard for low latency, high bandwidth, System Area Networks
  - Two communication types
    - Channel Semantics
    - Memory Semantics
  - High data rates
    - Quad Data Rate (QDR) - 40 Gbps
    - Dual Data Rate (DDR) - 20 Gbps
  - Multiple virtual lanes
  - Capable of end-to-end QoS

# Outline

- Introduction and Motivation
- Background
- Experimental Setup
- Microbenchmark Level Evaluation
- Communication Balance
- Application Level Evaluation
- Conclusions and Future Work

# Experimental Testbed

- Experimental Platforms
  - Intel Clovertown
    - Intel Xeon E5345 Dual quad-core processors operating at 2.33 GHz
    - 6GB RAM, 4MB cache
    - PCIe 1.1 interface
  - Intel Harpertown
    - Dual quad-core processors operating at 2.83 GHz
    - 8GB RAM
    - PCIe 2.0 interface
  - Intel Nehalem
    - Intel Xeon E5530 Dual quad-core processors operating at 2.40 GHz
    - 12GB RAM, 8MB cache
    - PCIe 2.0 interface

# Experimental Testbed (Cont)

- InfiniBand Host Channel Adapters
  - Dual port ConnectX DDR adapter
  - Dual port ConnectX QDR adapter
- InfiniBand Switches
  - Flextronics 144 port DDR switch
  - Mellanox 24 port QDR switch
- Open Fabrics Enterprise Distribution (OFED) 1.4.1 drivers
- Red Hat Enterprise Linux 4U4
- MPI Stack used – MVAPICH2-1.2p1
- Legends
  - NH-QDR – Intel Nehalem machines using ConnectX QDR HCA's
  - NH-DDR – Intel Nehalem machines using ConnectX DDR HCA's
  - HT-QDR – Intel Harpertown machines using ConnectX QDR HCA's
  - HT-DDR – Intel Harpertown machines using ConnectX DDR HCA's
  - CT-DDR – Intel Clovertown machines using ConnectX DDR HCA's

# MVAPICH / MVAPICH2

## Software

- High Performance MPI Library for IB and 10GE
  - MVAPICH (MPI-1) and MVAPICH2 (MPI-2)
  - Used by more than 960 organizations in 51 countries
  - More than 31,000 downloads from OSU site directly
  - Empowering many TOP500 clusters
    - 8<sup>th</sup> ranked 62,976-core cluster (Ranger) at TACC
  - Available with software stacks of many IB, 10GE and server vendors including Open Fabrics Enterprise Distribution (OFED)
  - Also supports uDAPL device to work with any network supporting uDAPL
  - <http://mvapich.cse.ohio-state.edu/>

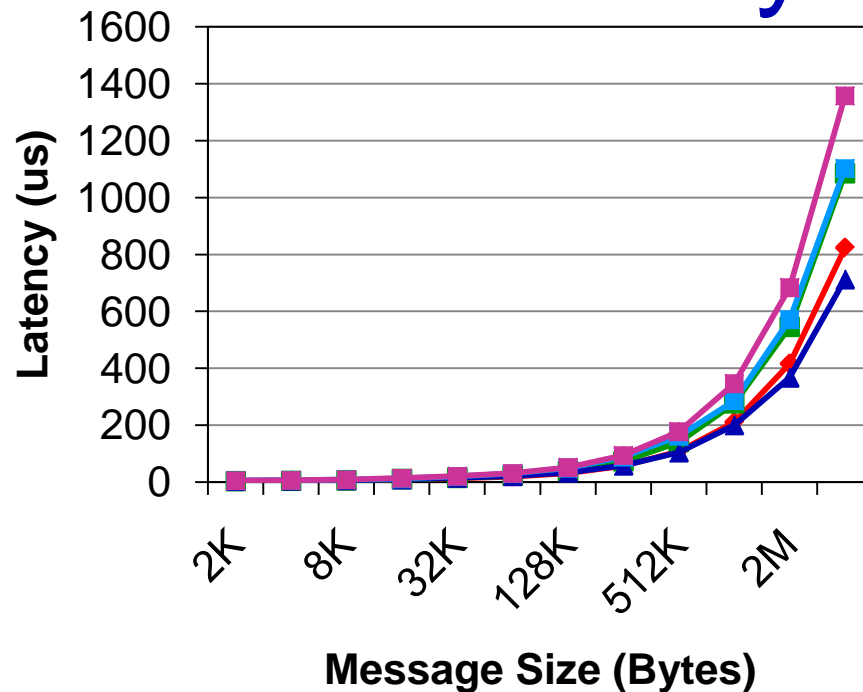
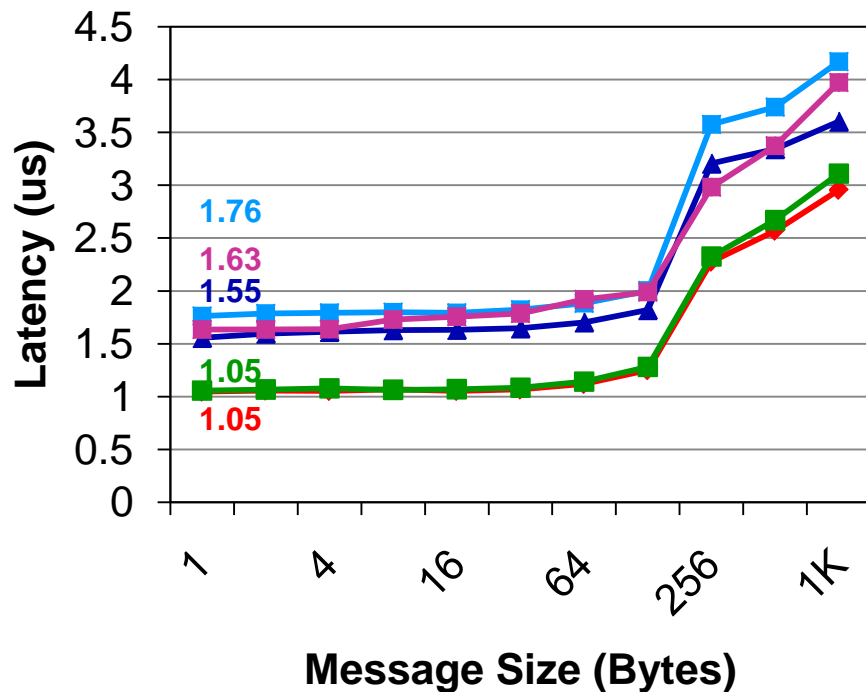
# List of Benchmarks

- OSU Microbenchmarks (OMB)
  - Version 3.1.1
  - <http://mvapich.cse.ohio-state.edu/benchmarks/>
- Intel Collective Microbenchmarks (IMB)
  - Version 3.2
  - <http://software.intel.com/en-us/articles/intel-mpi-benchmarks/>
- HPC Challenge Benchmark (HPCC)
  - Version 1.3.1
  - <http://icl.cs.utk.edu/hpcc/>
- NAS Parallel Benchmarks (NPB)
  - Version 3.3
  - <http://www.nas.nasa.gov/>

# Outline

- Introduction and Motivation
- Background
- Experimental Setup
- Microbenchmark Level Evaluation
- Communication Balance
- Application Level Evaluation
- Conclusions and Future Work

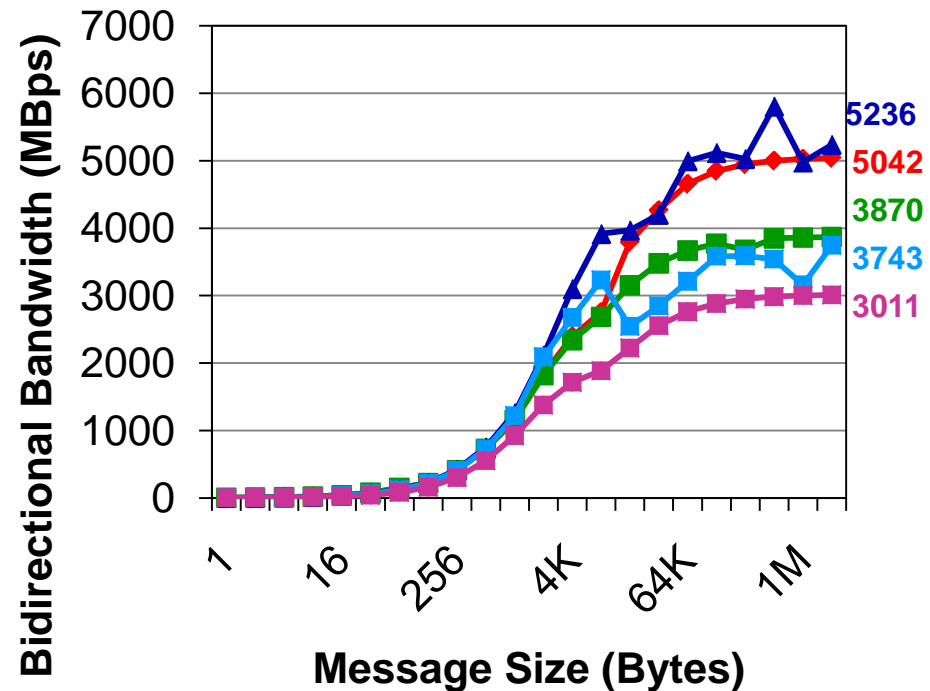
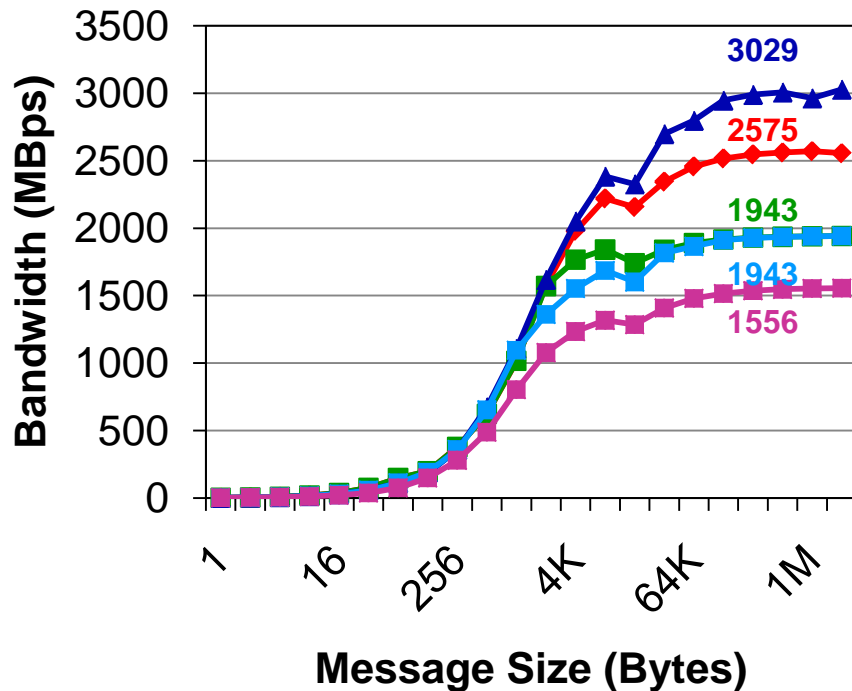
# Microbenchmark Level Evaluation – Inter-Node Latency



- ◆ HT-QDR
- HT-DDR
- ▲ NH-QDR
- ◆ HT-QDR
- HT-DDR
- ▲ NH-QDR
- NH-DDR
- CT-DDR
- NH-DDR
- CT-DDR

- Higher small message latency for Nehalem systems due to slower clock rate of machines in our cluster
- Medium to large message shows true capacity of Nehalem systems

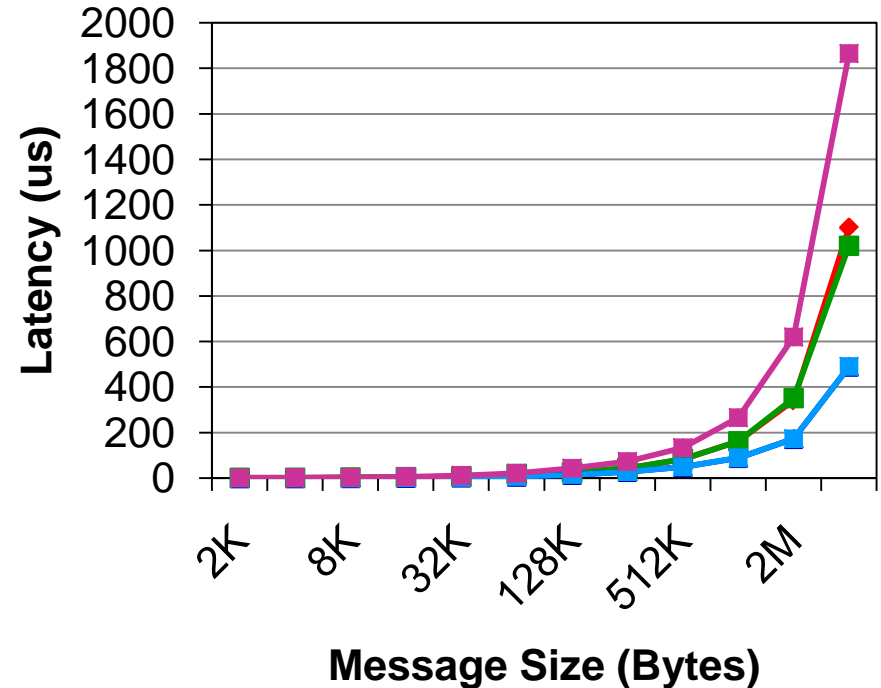
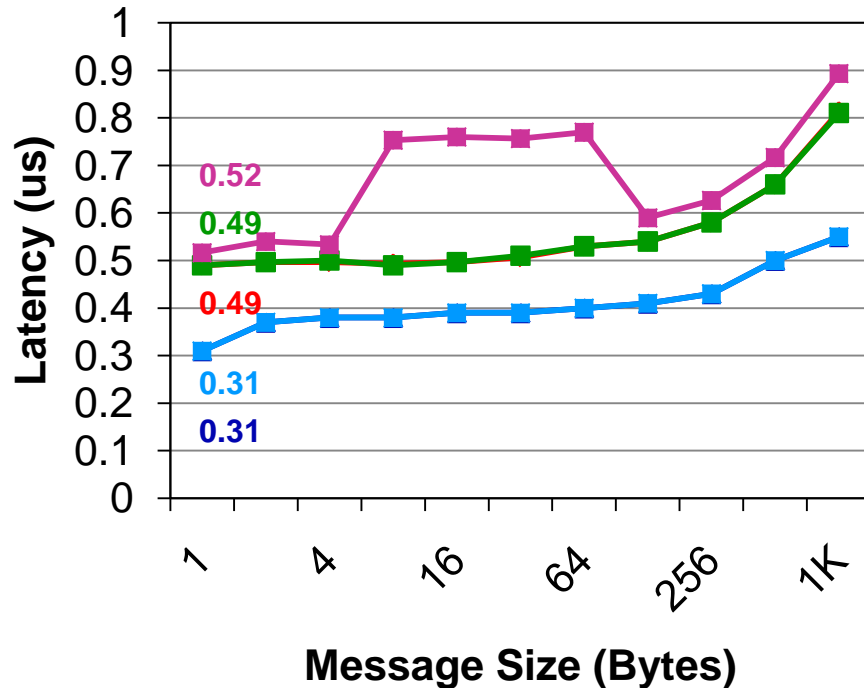
# Inter-Node Bandwidth



- ◆ HT-QDR
- HT-DDR
- ▲ NH-QDR
- ◆ HT-QDR
- HT-DDR
- ▲ NH-QDR
- NH-DDR
- CT-DDR
- NH-DDR
- CT-DDR

•NH-QDR gives up to **18%** improvement in uni-directional bandwidth over HT-QDR

# Intra-Node Latency

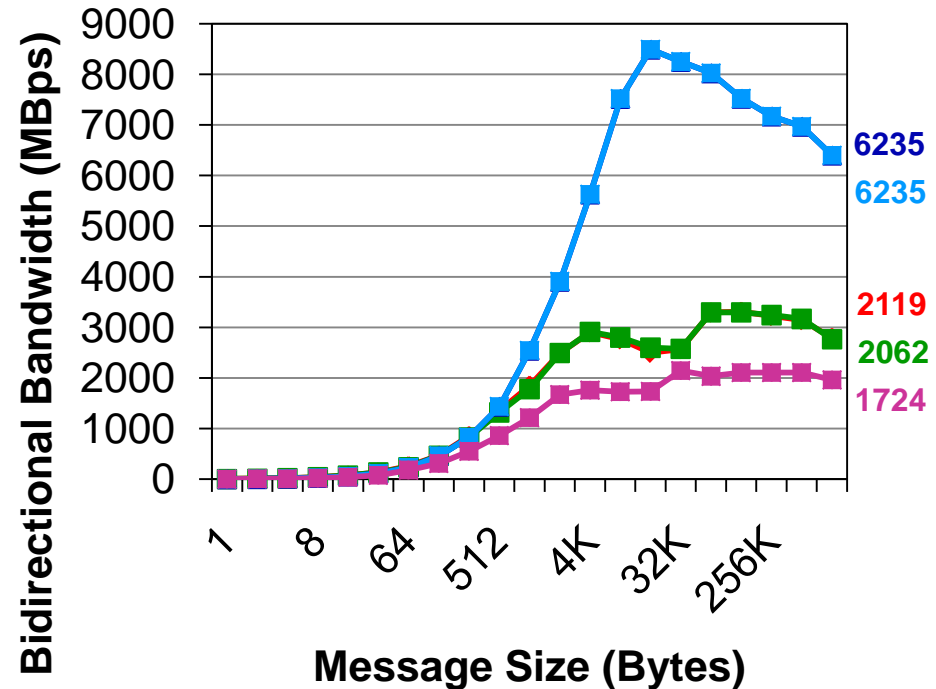
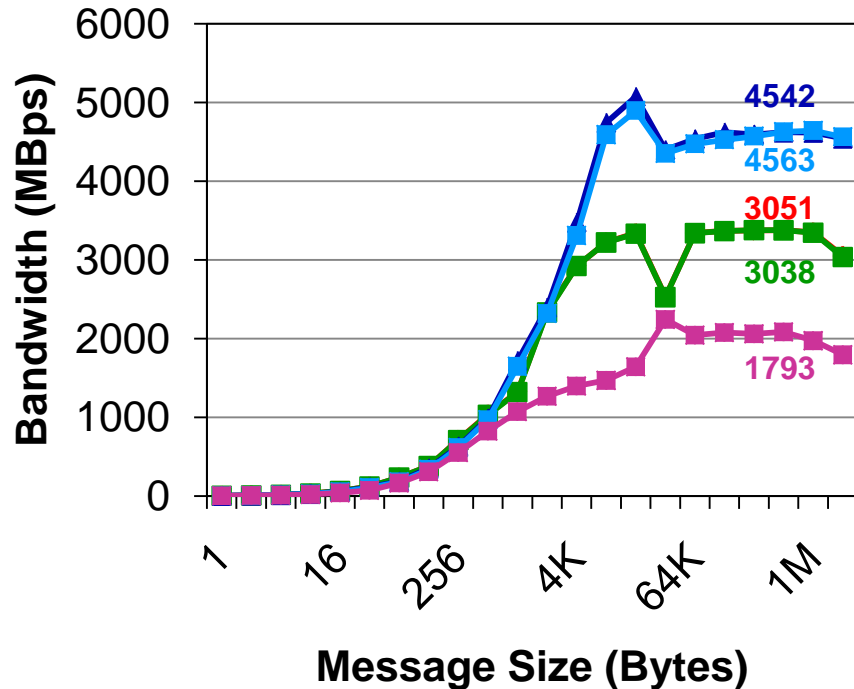


◆ HT-QDR    ■ HT-DDR    ▲ NH-QDR  
■ NH-DDR    ■ CT-DDR

◆ HT-QDR    ■ HT-DDR    ▲ NH-QDR  
■ NH-DDR    ■ CT-DDR

•Nehalem systems give up to **45%** improvement in Intra-Node latency

# Intra-Node Bandwidth



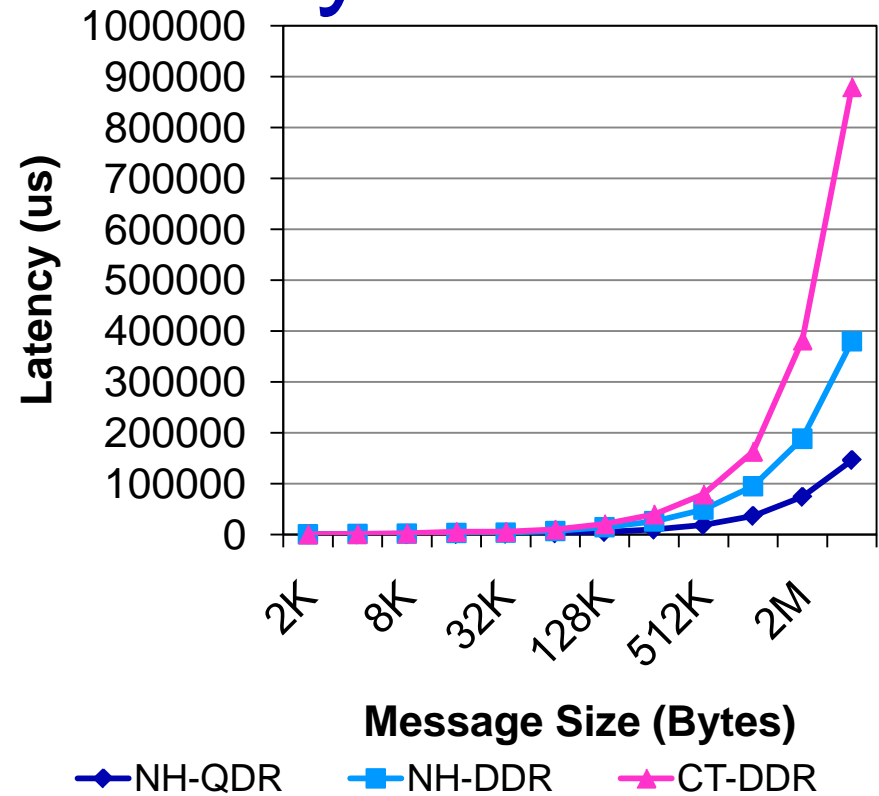
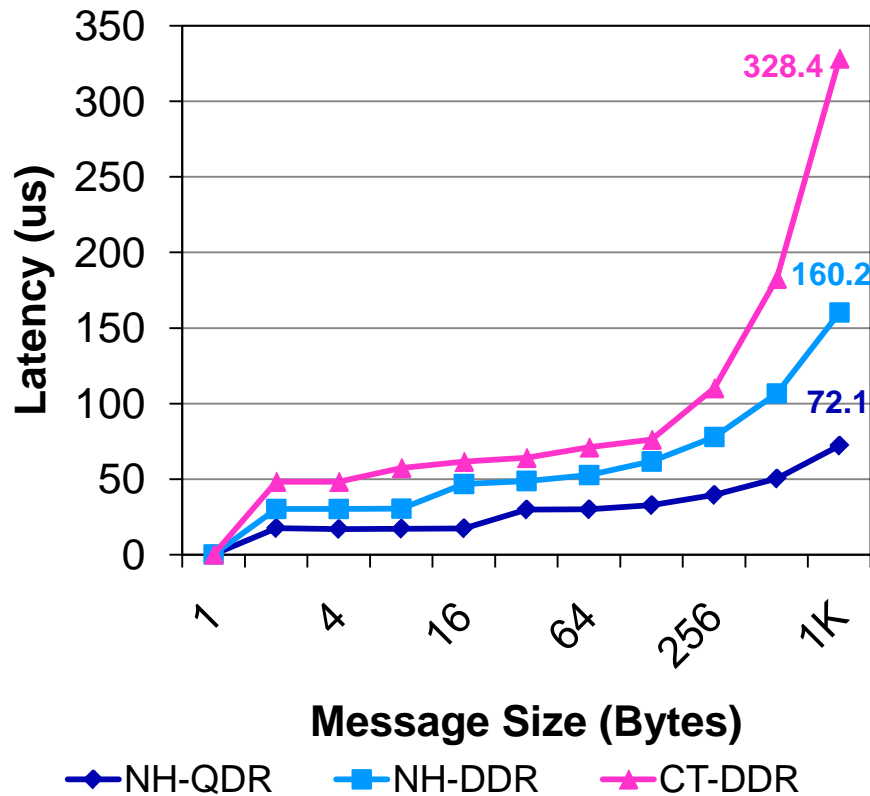
◆ HT-QDR    ■ HT-DDR    ▲ NH-QDR  
■ NH-DDR    ■ CT-DDR

◆ HT-QDR    ■ HT-DDR    ▲ NH-QDR  
■ NH-DDR    ■ CT-DDR

- Intra-Node bandwidth and bidirectional bandwidth shows the high memory bandwidth of Nehalem systems
- Drop in performance for Nehalem systems at large message size due to cache collisions

# Collective Performance

## Alltoall Latency



- The 32 process Alltoall latency numbers shows a **43% to 55%** improvement by using QDR HCA over a DDR HCA
- Harpertown numbers not shown due to insufficient number of nodes

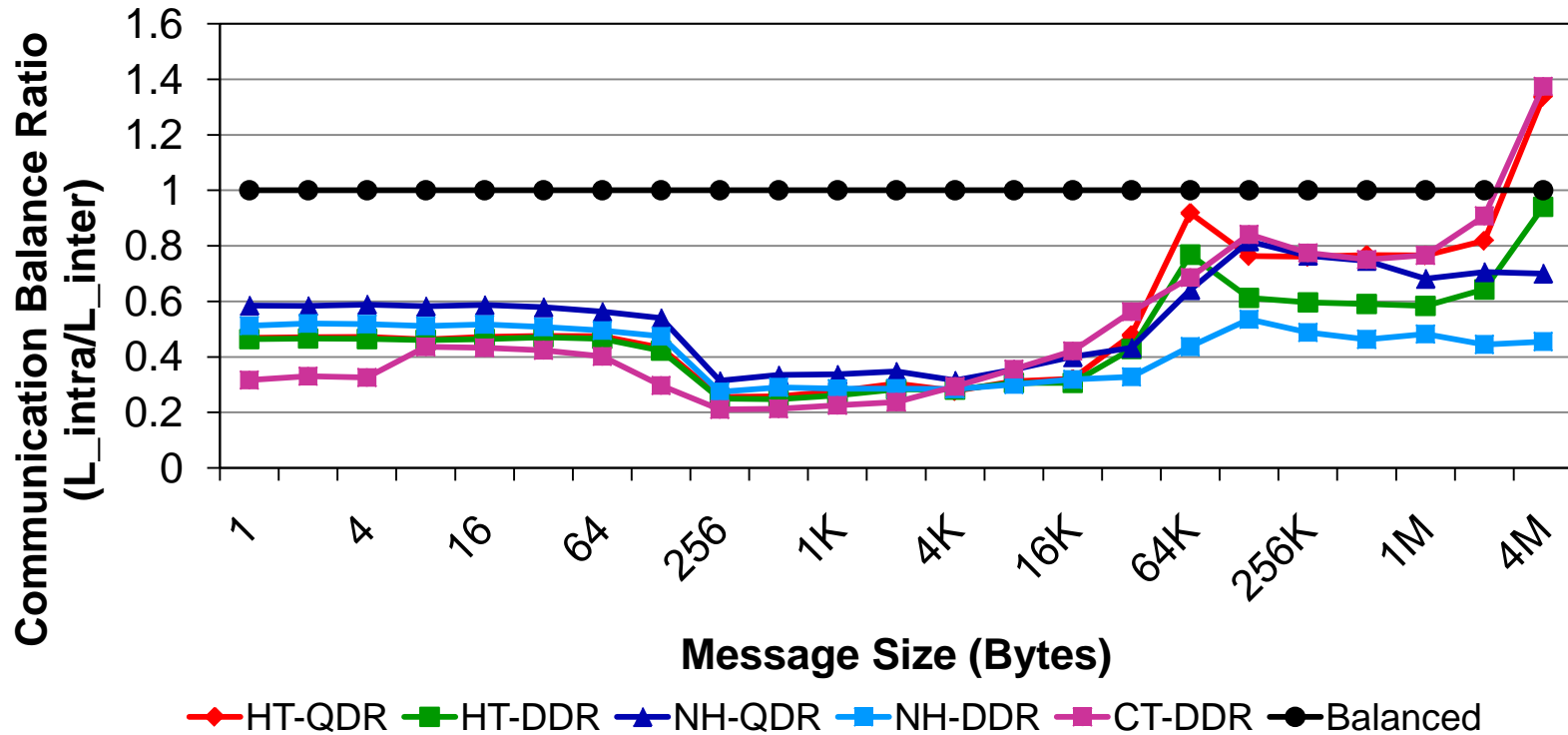
# Outline

- Introduction and Motivation
- Background
- Experimental Setup
- Microbenchmark Level Evaluation
- **Communication Balance**
- Application Level Evaluation
- Conclusions and Future Work

# Communication Balance Ratio

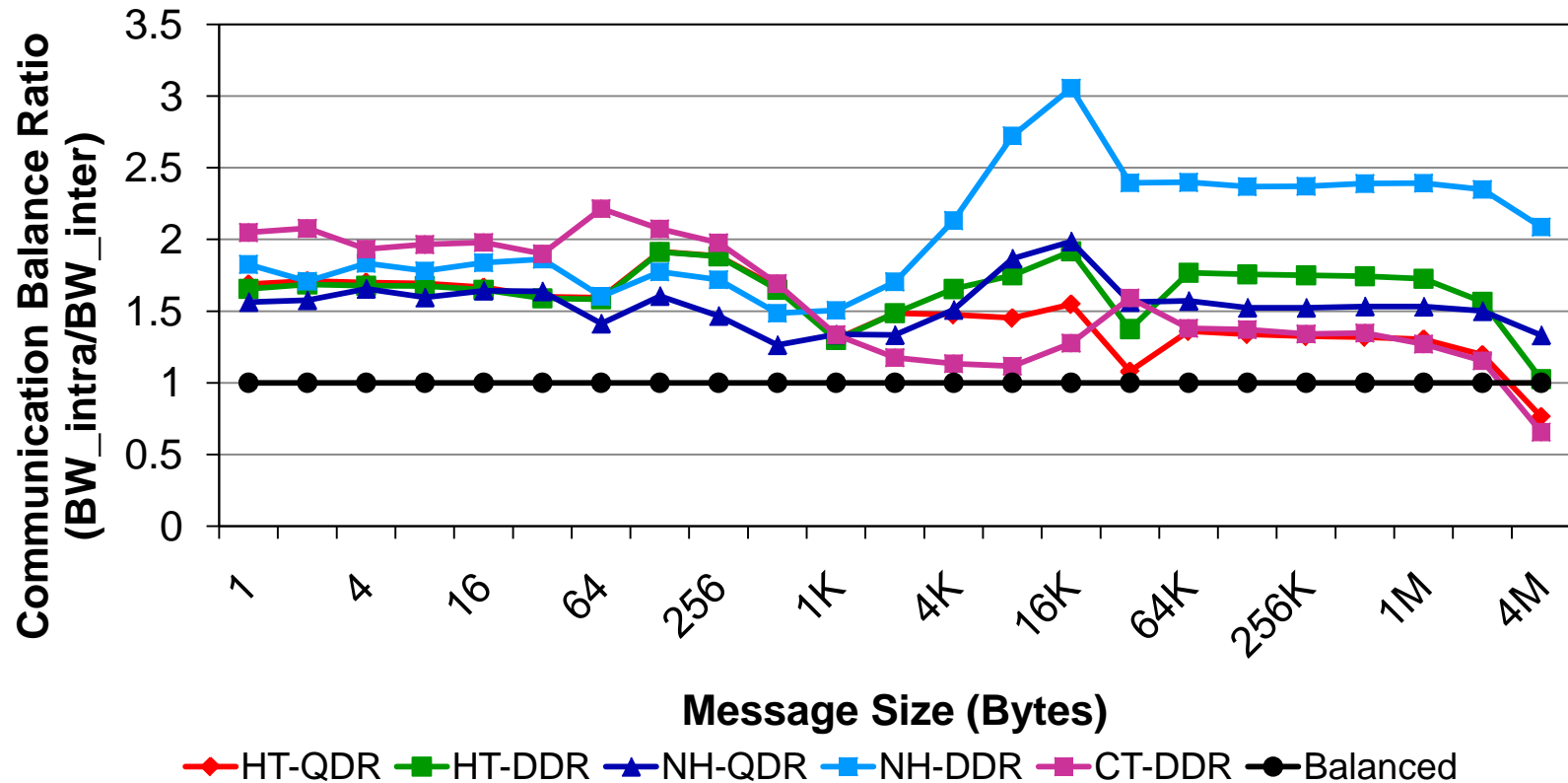
- Intra-Node communication gives better performance than Inter-Node communication for small to medium sized messages
- This make physical location of process very important in parallel computing
- A good super computing cluster is one where Intra-Node performance is same as Inter-Node performance

# Communication Balance Ratio Latency



•The black line indicates the ratio displayed by a balanced system

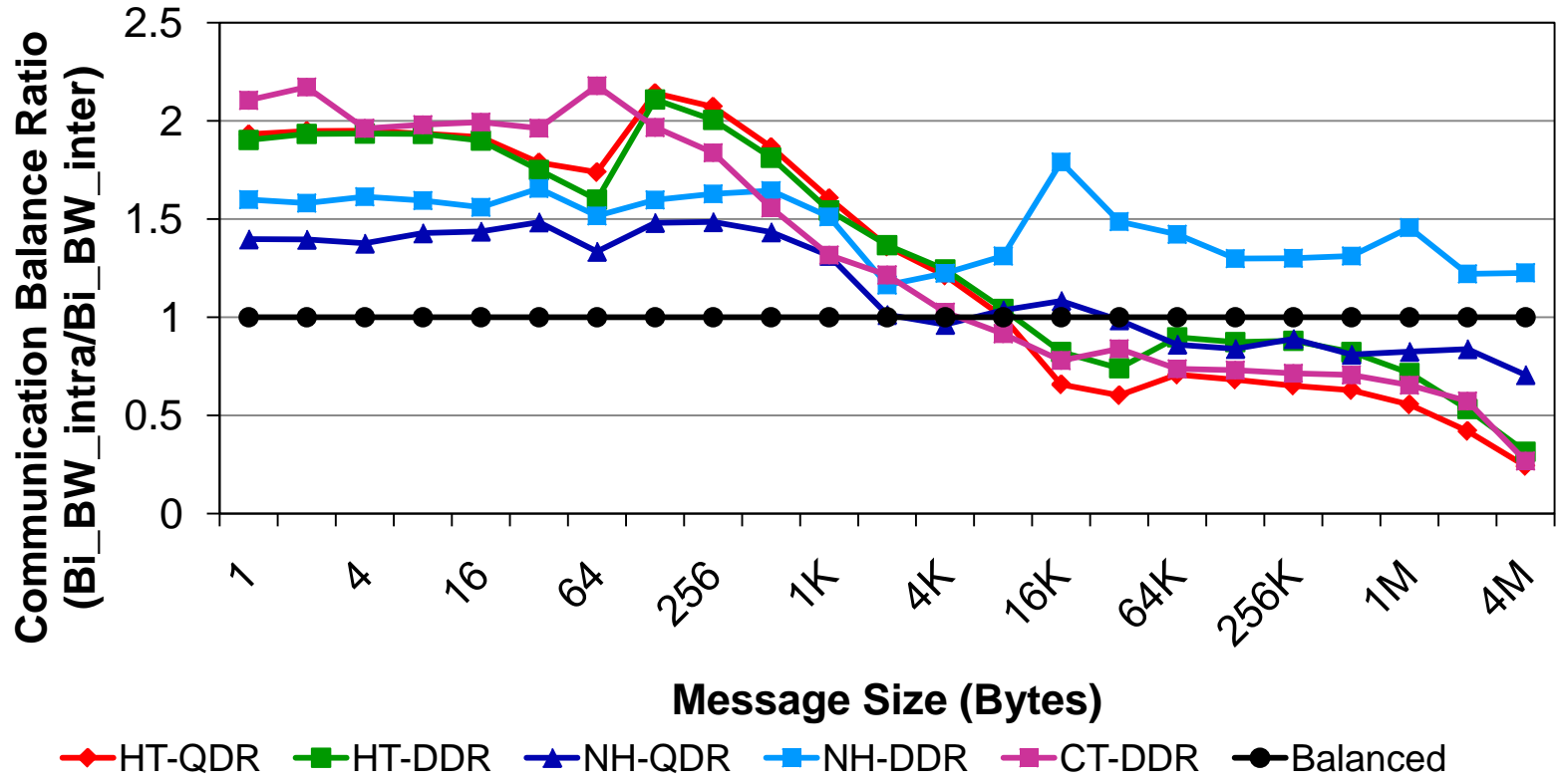
# Communication Balance Ratio Bandwidth



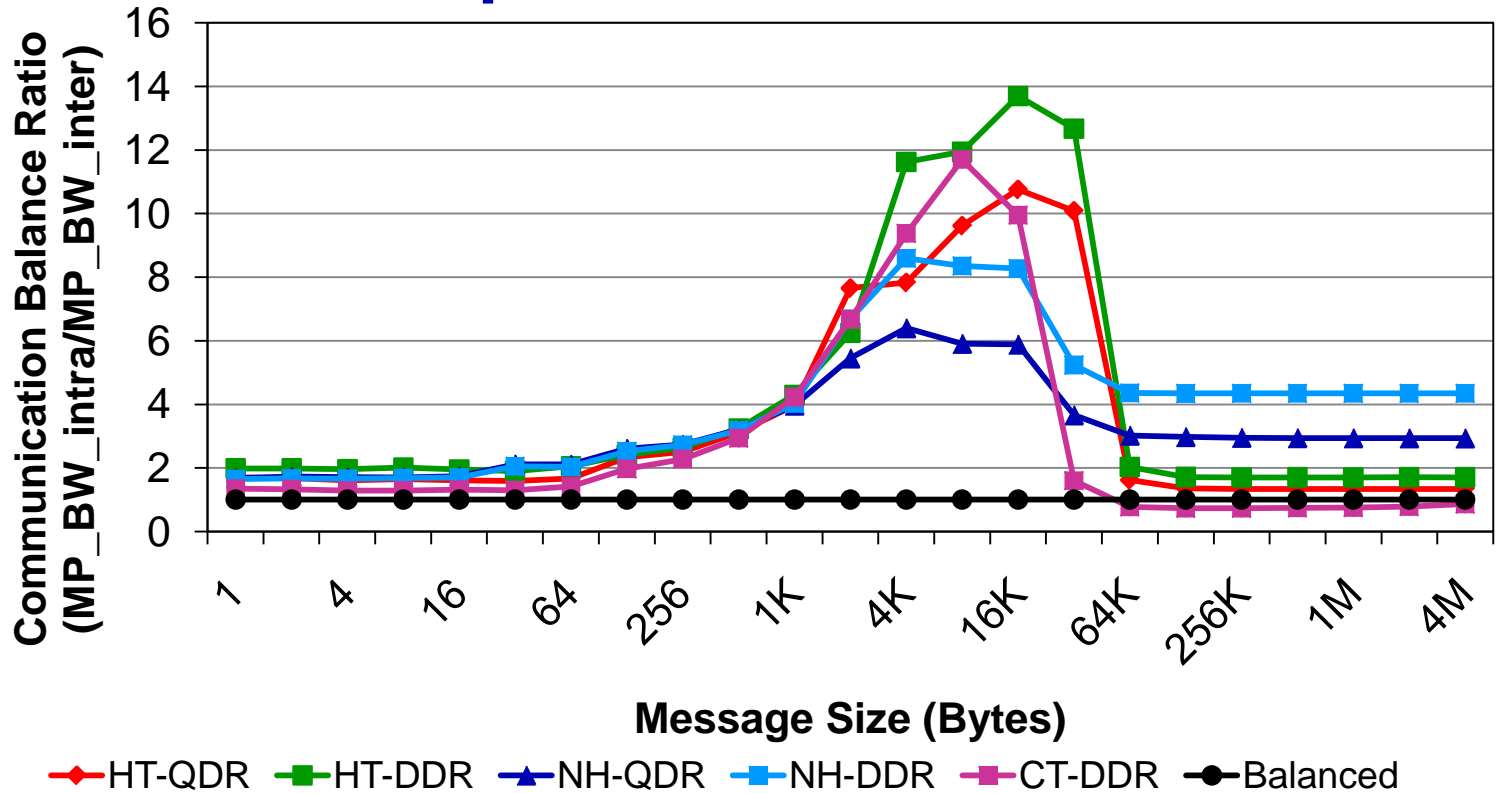
•The black line indicates the ratio displayed by a balanced system

Hotl '09

# Communication Balance Ratio Bidirectional Bandwidth



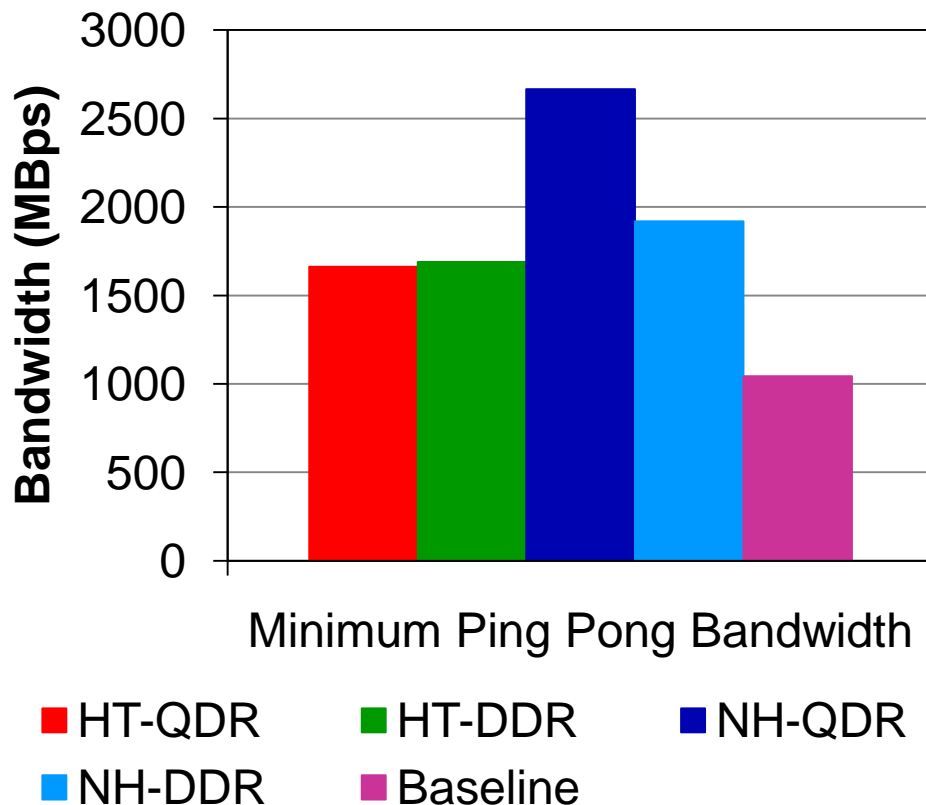
# Communication Balance Ratio Multipair Bandwidth



# Outline

- Introduction and Motivation
- Background
- Experimental Setup
- Microbenchmark Level Evaluation
- Communication Balance
- Application Level Evaluation
- Conclusions and Future Work

# Application Level Evaluation – HPCC

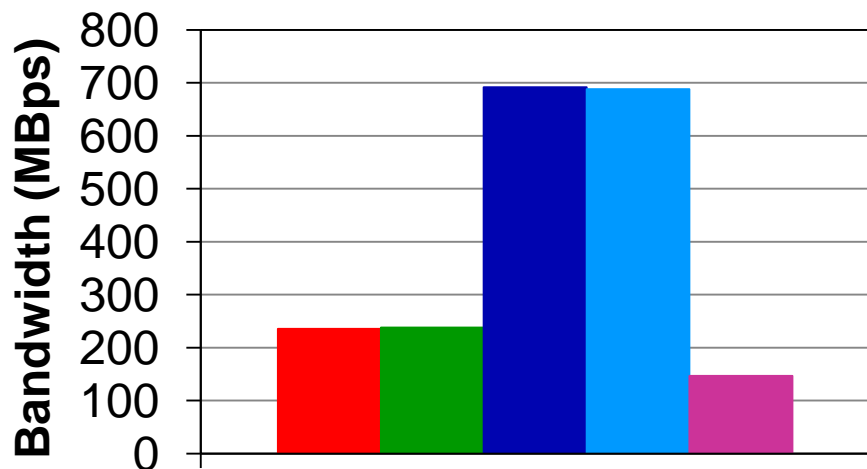


- Baseline numbers are taken on CT-DDR

- NH-DDR show a **13%** improvement in performance over Harpertown and Clovertown systems

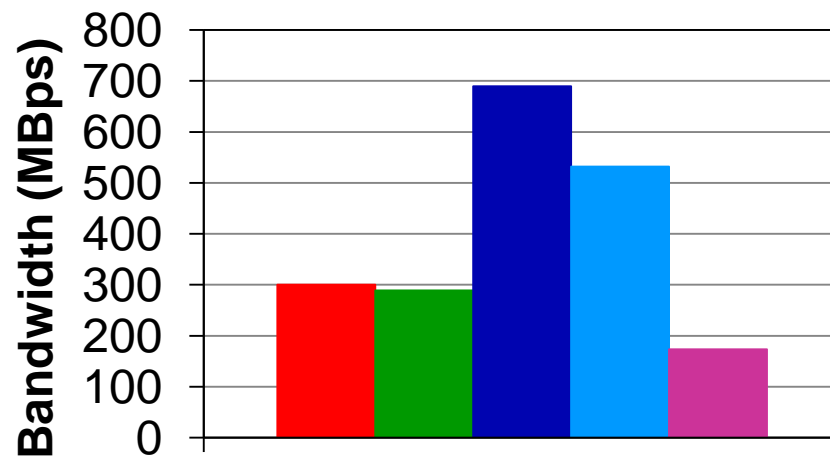
- NH-QDR shows a **38%** improvement in performance over NH-DDR systems

# Application Level Evaluation – HPCC (Cont)



Naturally Ordered Ring  
Bandwidth

■ HT-QDR    ■ HT-DDR    ■ NH-QDR  
■ NH-DDR    ■ Baseline



Randomly Ordered Ring  
Bandwidth

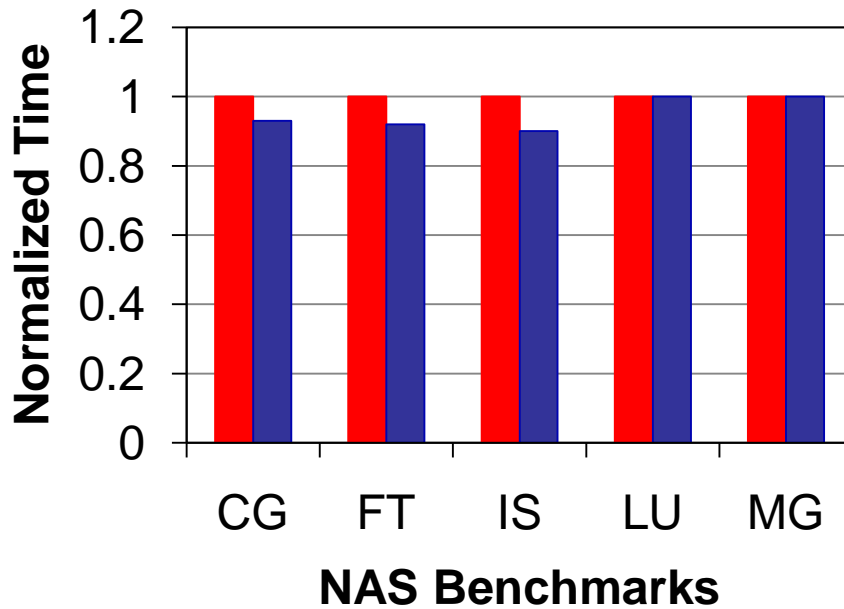
■ HT-QDR    ■ HT-DDR    ■ NH-QDR  
■ NH-DDR    ■ Baseline

•Upto **190%** improvement in Naturally Ordered Ring bandwidth for NH-QDR

•Upto **130%** improvement in Randomly Ordered Ring bandwidth for NH-QDR

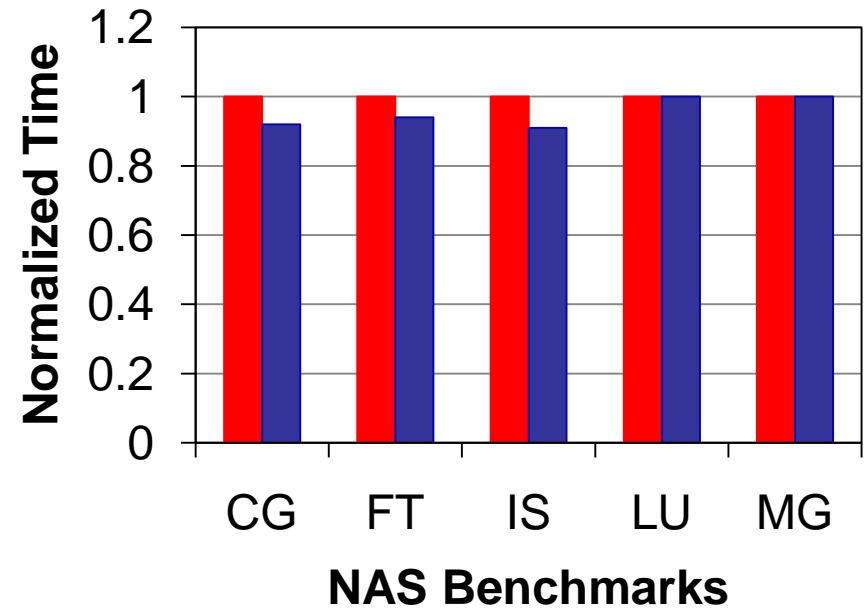
# Performance of NAS Benchmarks

**Class B – 32 processes**



■ NH-DDR ■ NH-QDR

**Class C – 32 processes**



■ NH-DDR ■ NH-QDR

•NH-QDR shows clear benefits over NH-DDR for multiple applications

# Outline

- Introduction and Motivation
- Background
- Experimental Setup
- Microbenchmark Level Evaluation
- Communication Balance
- Application Level Evaluation
- Conclusions and Future Work

# Conclusions & Future Work

- Evaluate multiple computational platforms with multiple InfiniBand HCAs
- Nehalem systems with QDR interconnect gives best performance
- Propose the metric of *communication balance* to find out the best components to design next generation clusters
- Nehalem systems with QDR interconnects offers best communication balance
- We plan to perform larger scale evaluations and study impact of these systems on performance of end applications

# Thank you !

{subramon, koop, panda}@cse.ohio-state.edu



Network-Based Computing Laboratory

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