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

Hari Subramoni, Matthew Koop and Dhabaleswar. K. Panda

Computer Science & Engineering Department
The Ohio State University

HotI '09

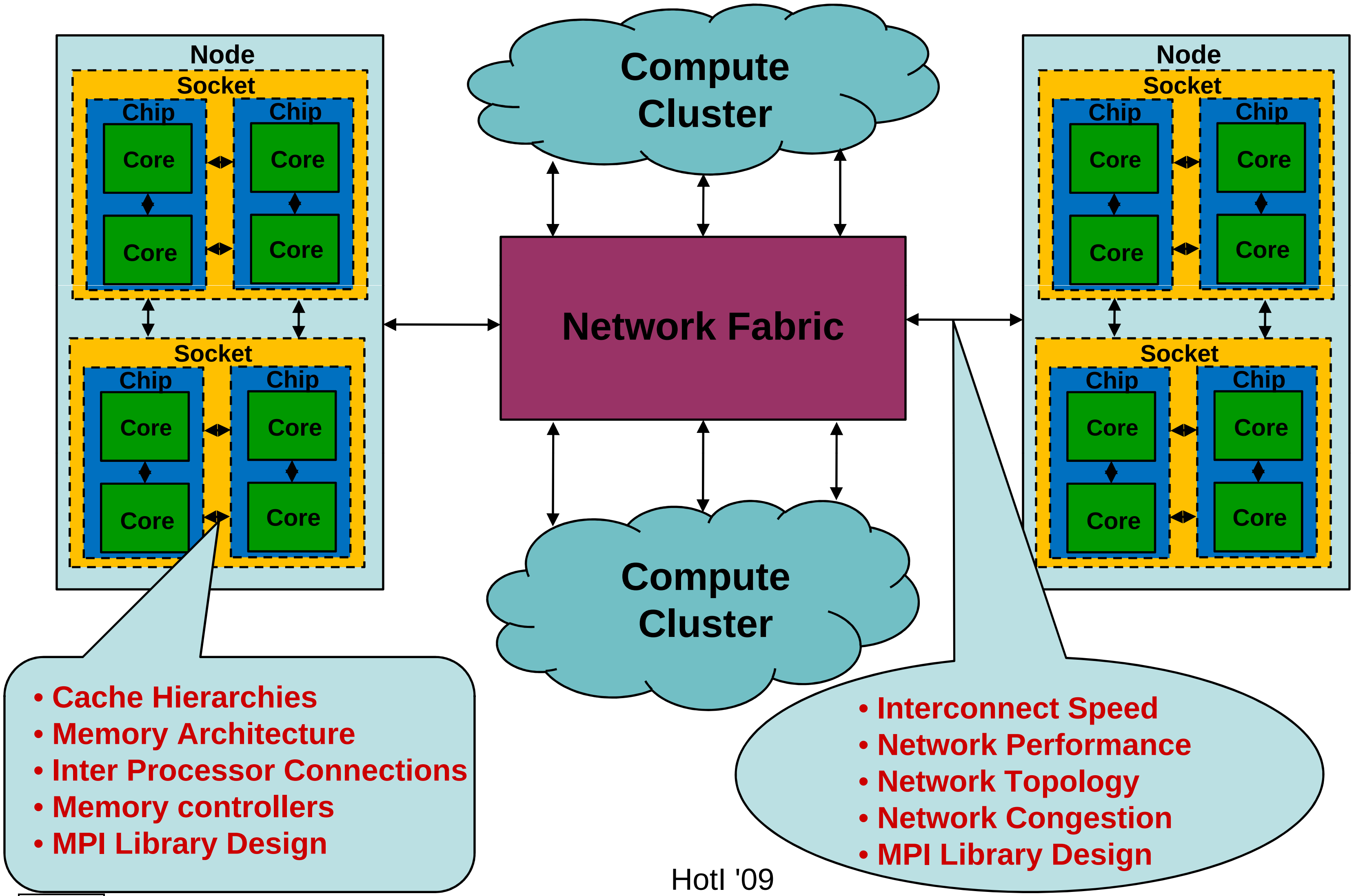
Outline

- Introduction
- Problem Statement
- Approach
- Performance Evaluation and Results
- Conclusions and Future Work

Introduction

- Commodity clusters are becoming more popular for High Performance Computing (HPC) Systems
- Modern clusters are being designed with multi-core processors
- Introduces multi-level communication
  - Intra-node (intra-socket, inter-socket)
  - Inter-node

Factors Affecting the Communication Performance

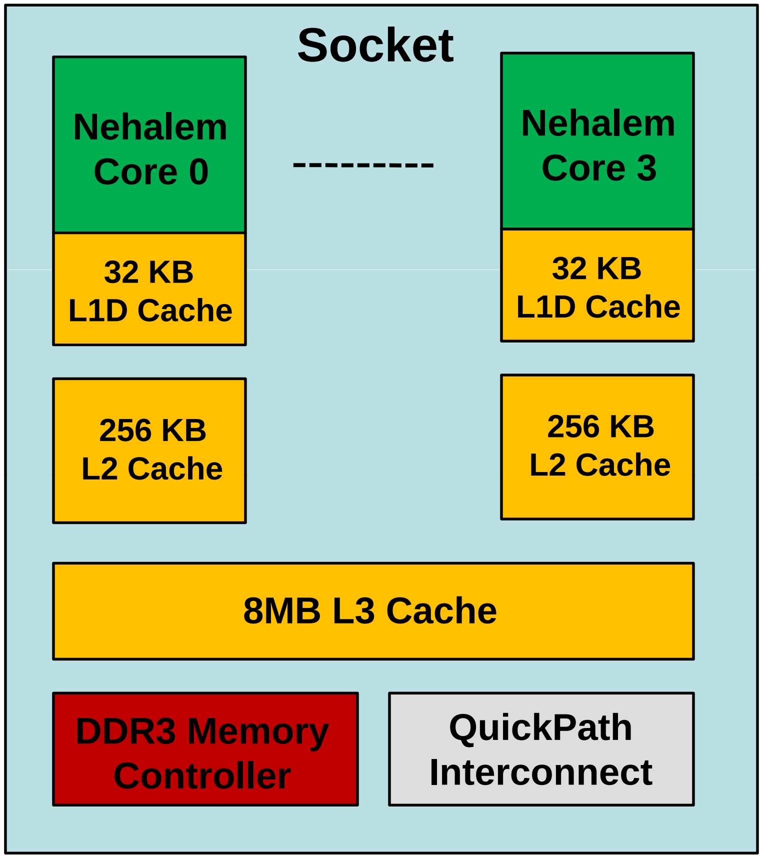


Additional Factors Affecting the Overall Application Performance

- Communication characteristics
- Message distribution
- Mapping of processes into cores/nodes
  - Block vs. cyclic
- Traditionally, intra-node communication performance has been better than inter-node communication performance
  - Most applications use `block' distributions
- Are such practices still valid on modern clusters?

Intel Nehalem Processor

- First true quad core processor with L3 cache sharing
- 45 nm manufacturing process
- 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
- Integrated memory controller supporting multiple memory channels gives very high memory bandwidth
- Has impact on Intra-node Communication Performance



InfiniBand Architecture

- An industry standard for low latency, high bandwidth, System Area Networks
- Multiple features
  - Two communication types
    - Channel Semantics
    - Memory Semantics (RDMA mechanism)
  - Multiple virtual lanes
  - Quality of Service (QoS) support
- Double Data Rate (DDR) with 20 Gbps bandwidth has been there
- Quad Data Rate (QDR) with 40 Gbps bandwidth is available recently
- Has impact on Inter-node communication performance

Outline

- Introduction
- Problem Statement
- Approach
- Performance Evaluation and Results
- Conclusions and Future Work

Problem Statement

- What are the intra-node and inter-node communication performance of Nehalem-based clusters with InfiniBand DDR and QDR?
- How do these communication performance compare with previous generation Intel processors (Clovertown and Harpertown) with similar InfiniBand DDR and QDR?
- With rapid advances in processor and networking technologies, are the relative performance between intra-node and inter-node changing?
- How such changes can be characterized?
- Can such characterization be used to analyze application performance across different systems?

Outline

- Introduction
- Problem Statement
- Approach
- Performance Evaluation and Results
- Conclusions and Future Work

Two Angles

- Absolute Performance of Intra-node and Inter-node communication
  - Different combinations of Intel processor platforms and InfiniBand (DDR and QDR)
- Characterization of Relative Performance between Intra-node and Inter-node communication
  - Use such characterization to analyze application-level performance

Communication Balance Ratio

- Applications have different communication characteristics
  - Latency sensitive
  - Bandwidth (uni-directional) sensitive
  - Bandwidth (bi-directional) sensitive
- Introduce a set of metrics Communication Balance Ratio (CBR)
  - CBR-Latency = Latency_Intra / Latency_Inter
  - CBR-Bandwidth = Bandwidth_Intra / Bandwidth_Inter
  - CBR-Bi-BW = Bi-BW_Intra / Bi_BW_Inter
  - CBR-Multi-BW = Multi-BW_Intra / Multi-BW_Inter
- CBR-x=1 => Cluster is Balanced wrt metric x
  - Applications sensitive to metric x can be mapped anywhere in the cluster without any significant impact on overall performance

Outline

- Introduction
- Problem Statement
- Approach
- Performance Evaluation and Results
- Conclusions and Future Work

Experimental Testbed

- Three different compute 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, 6MB cache
    - 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)

- Two different InfiniBand Host Channel Adapters
  - Dual port ConnectX DDR adapter
  - Dual port ConnectX QDR adapter
- Two different InfiniBand Switches
  - Flextronics 144 port DDR switch
  - Mellanox 24 port QDR switch
- Five different platform-interconnect combinations
  - 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
- Open Fabrics Enterprise Distribution (OFED) 1.4.1 drivers
- Red Hat Enterprise Linux 4U4
- MPI Stack used – MVAPICH2-1.2p1

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 32,000 downloads from OSU site directly
  - Empowering many TOP500 clusters
    - 8th 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/

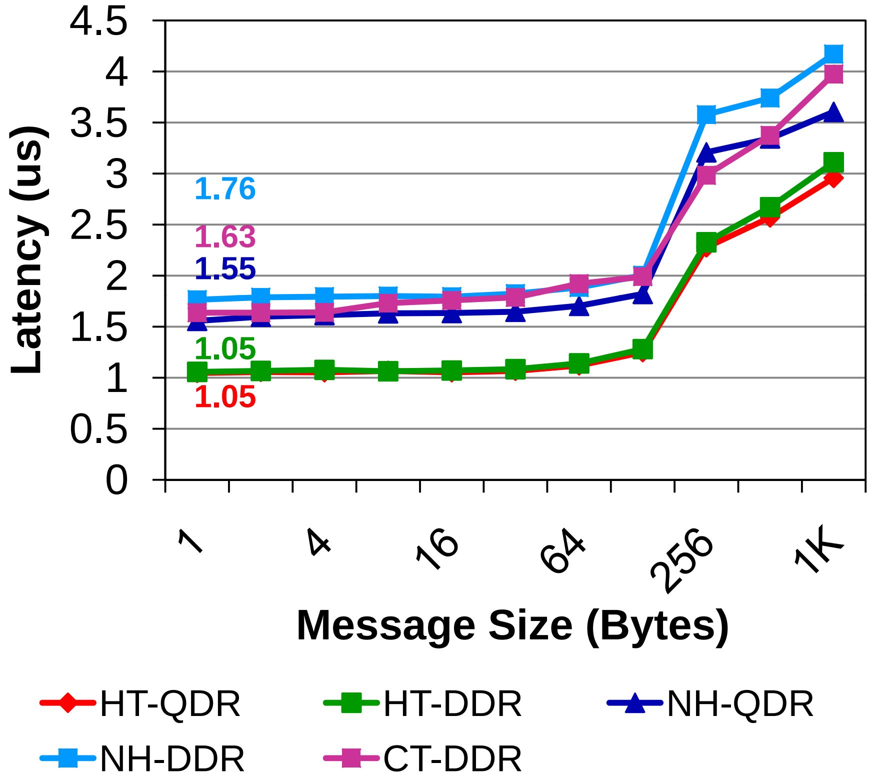
List of Benchmarks

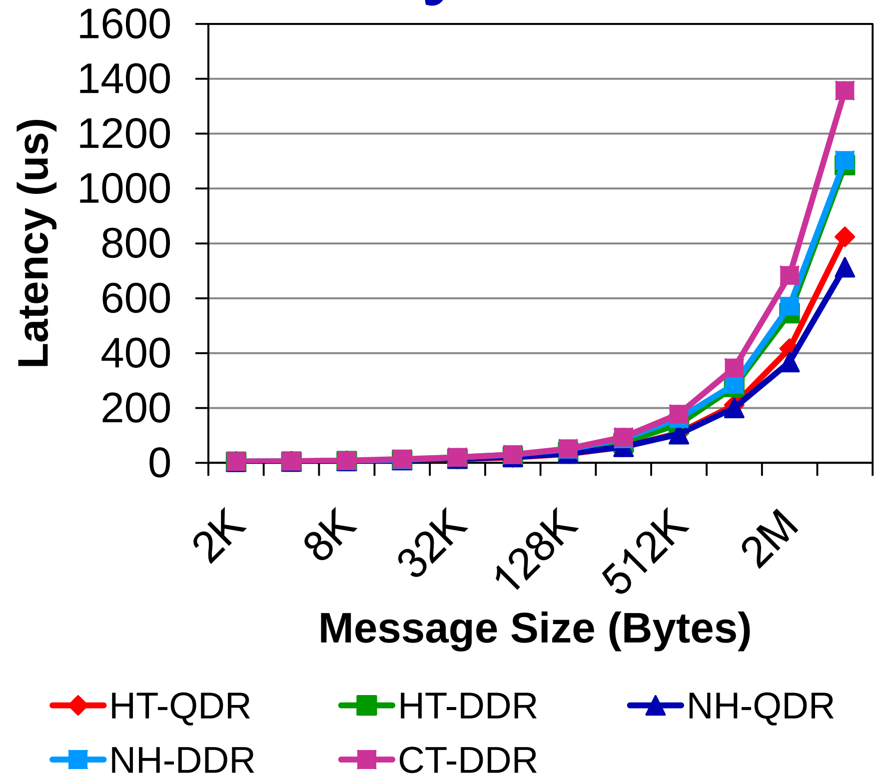
- OSU Microbenchmarks (OMB)
  - Version 3.1.1
  - http://mvapich.cse.ohio-state.edu/benchmarks/
- Intel Collective Microbenchmarks (IMB)
  - Version 3.2
  - http://software.intel.com/en-us/articles/intel-mpi-benchmarks/
- HPC Challenge Benchmark (HPCC)
  - Version 1.3.1
  - http://icl.cs.utk.edu/hpcc/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

Performance Results

- Absolute Performance
  - Inter-node latency and bandwidth
  - Intra-node latency and bandwidth
  - Collective All-to-all
  - HPCC
  - NAS
- Communication Balance Ratio
  - CBR-Latency
  - CBR-Bandwidth (uni-directional)
  - CBR-Bandwidth (bi-directional)
  - CBR-Bandwidth (multi-pair)
- Impact of CBR on Application Performance

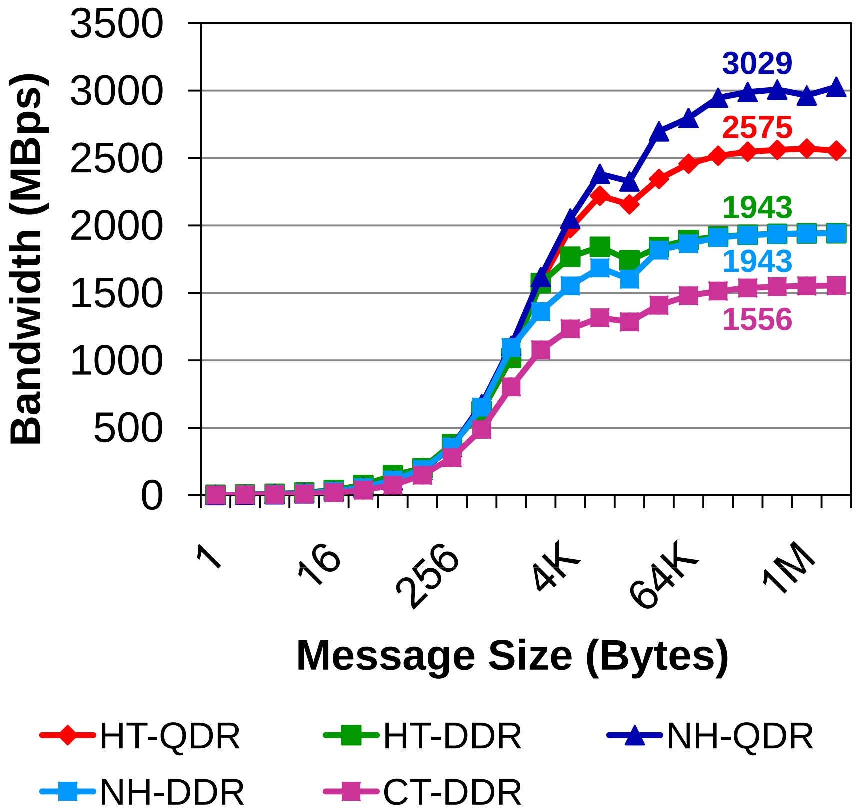
Microbenchmark Level Evaluation – Inter-Node Latency

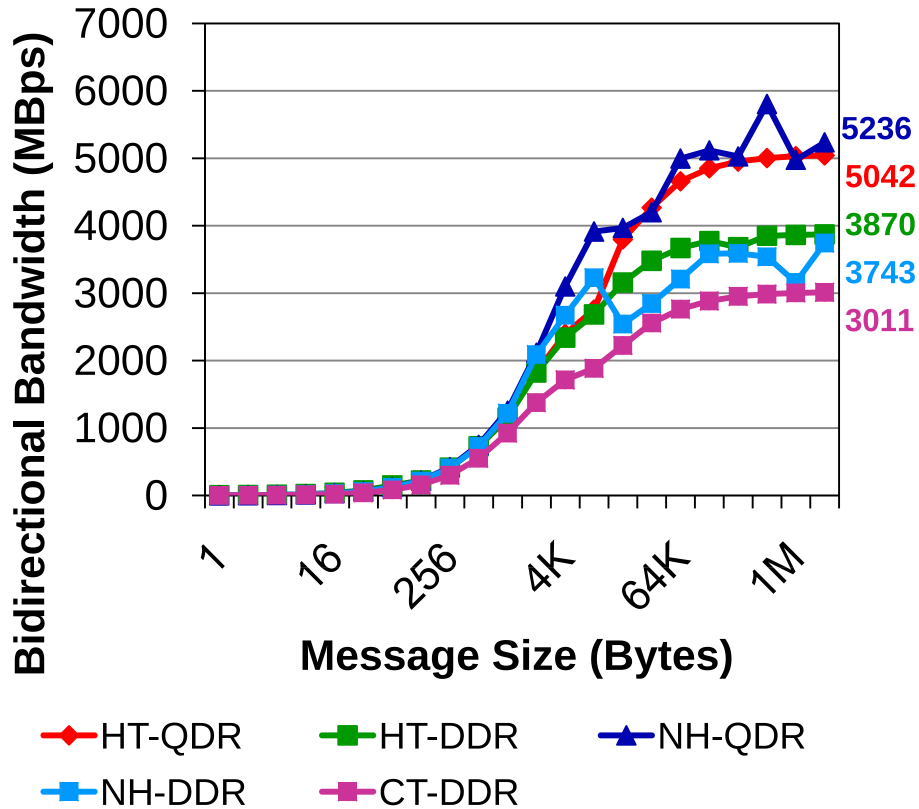




- **Harpertown systems deliver best small message latency**
- **Up to 10% improvement in large message latency for NH-QDR over HT-QDR**

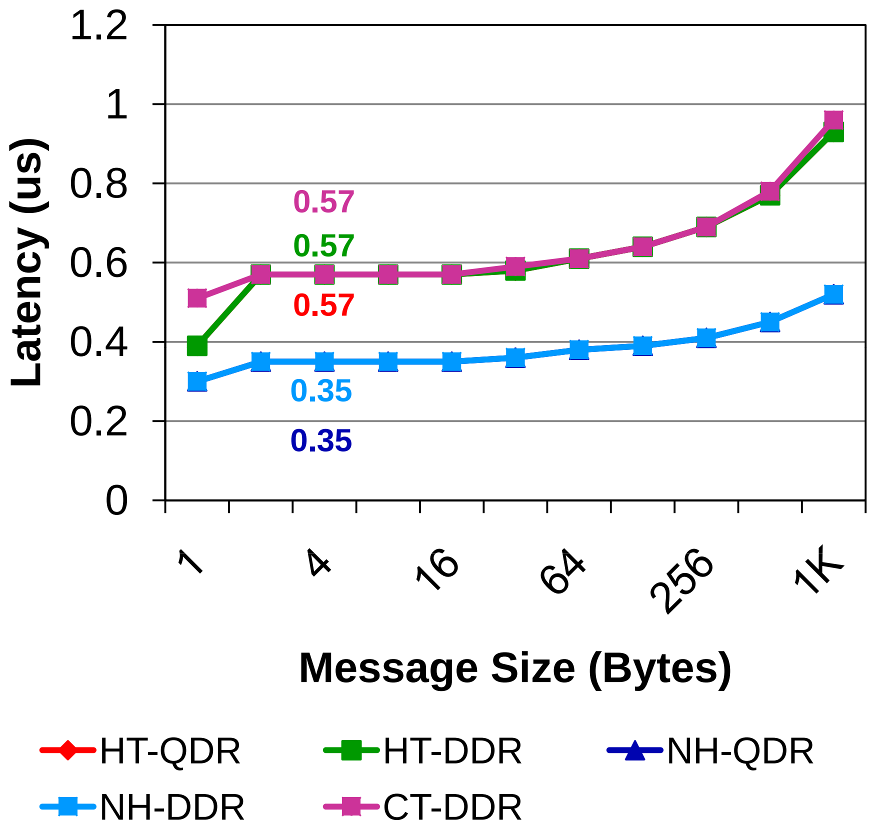
Inter-Node Bandwidth

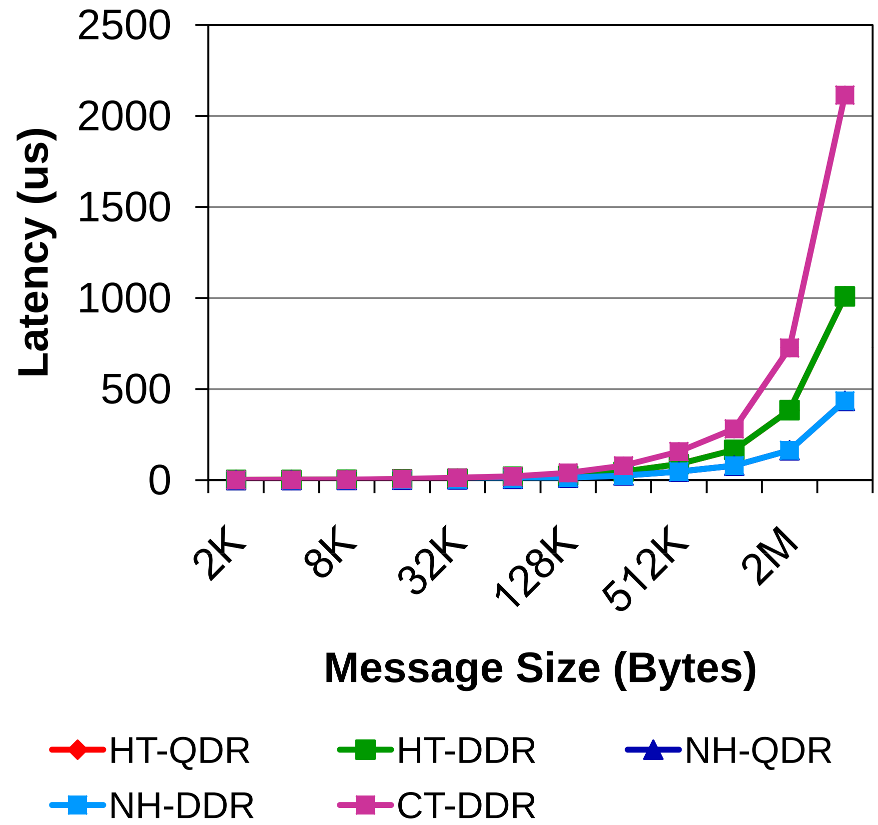




- **Nehalem systems offer a peak uni-directional bandwidth of 3029 *MBps* and bi-directional bandwidth of 5236 *MBps***
- **NH-QDR gives up to 18% improvement in uni-directional bandwidth over HT-QDR**

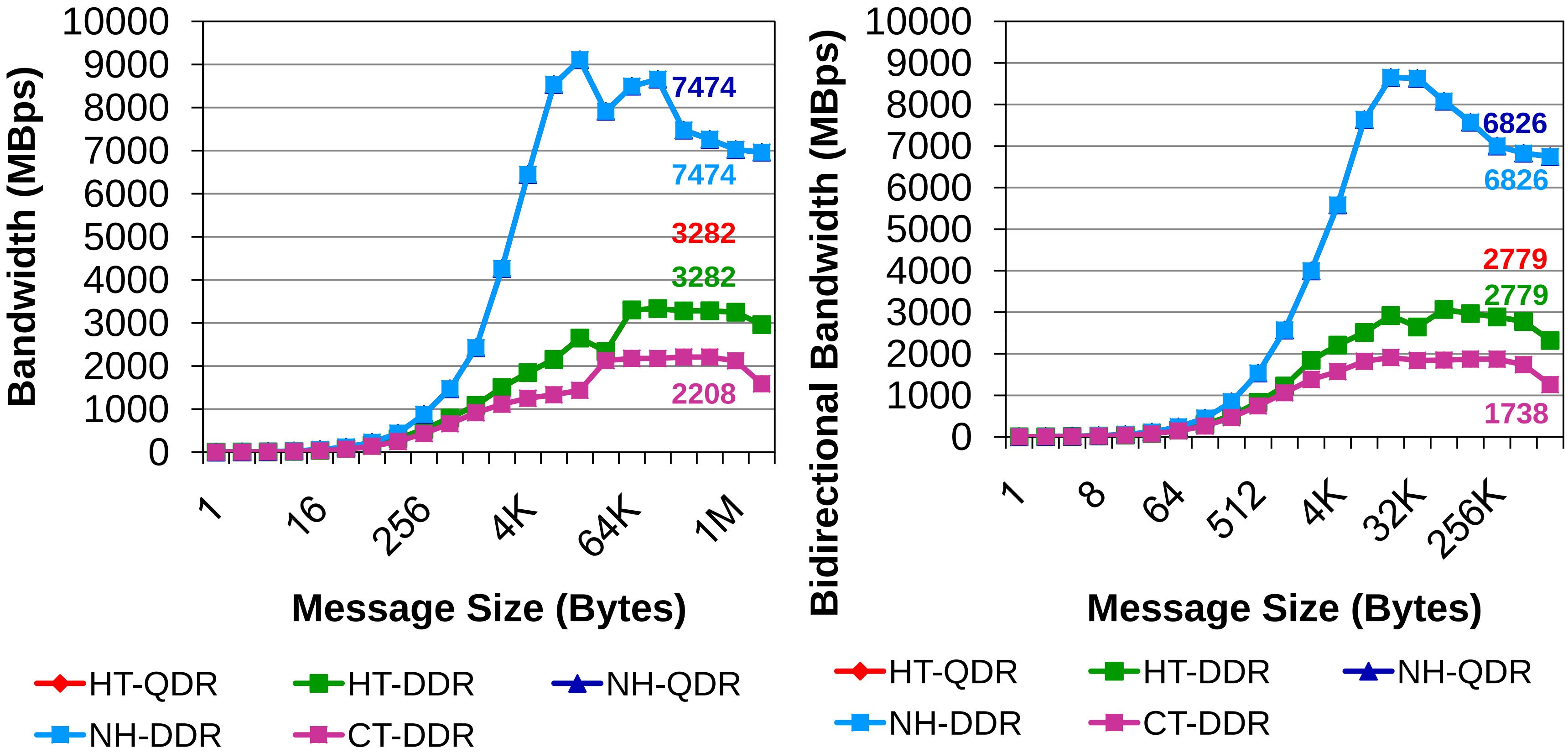
Intra-Node Latency





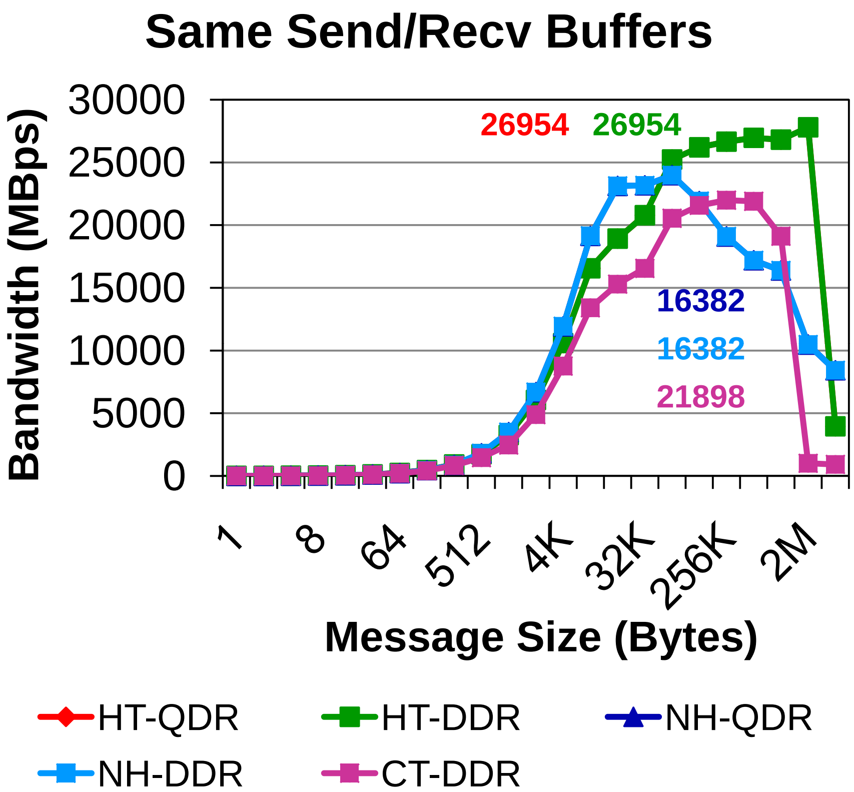
- **Intra-Socket small message latency of 0.35 *us***
- **Nehalem systems give up to 40% improvement in Intra-Node latency for various message sizes**

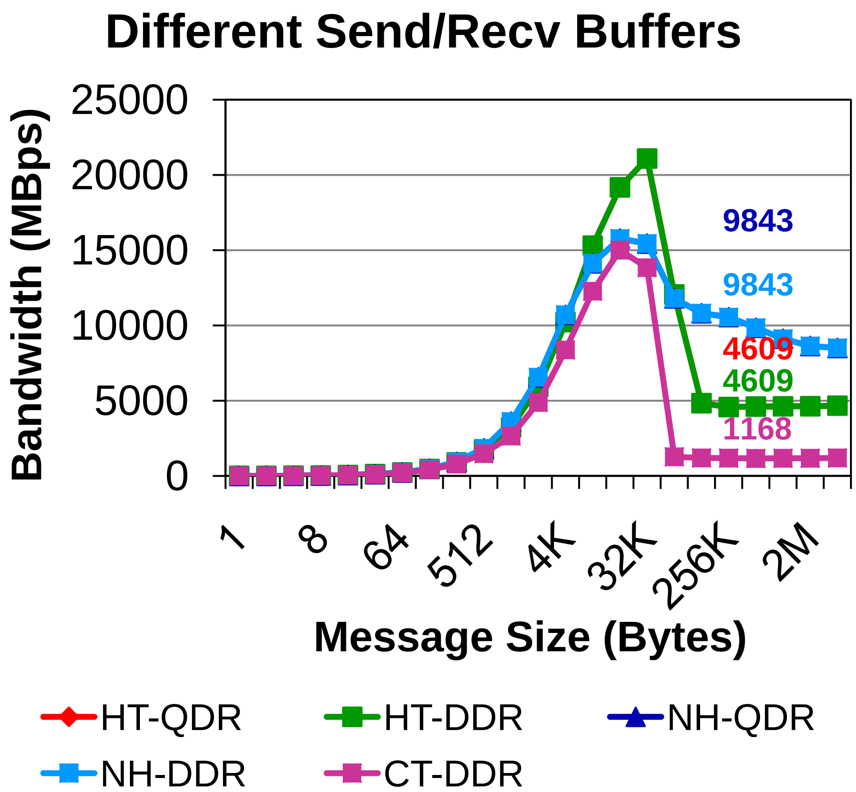
Intra-Node Bandwidth



- **Intra-Socket bandwidth (7474 *MBps*) and bidirectional bandwidth (6826 *MBps*) show the high memory bandwidth of Nehalem systems**
- **Drop in performance at large message size due to cache collisions**

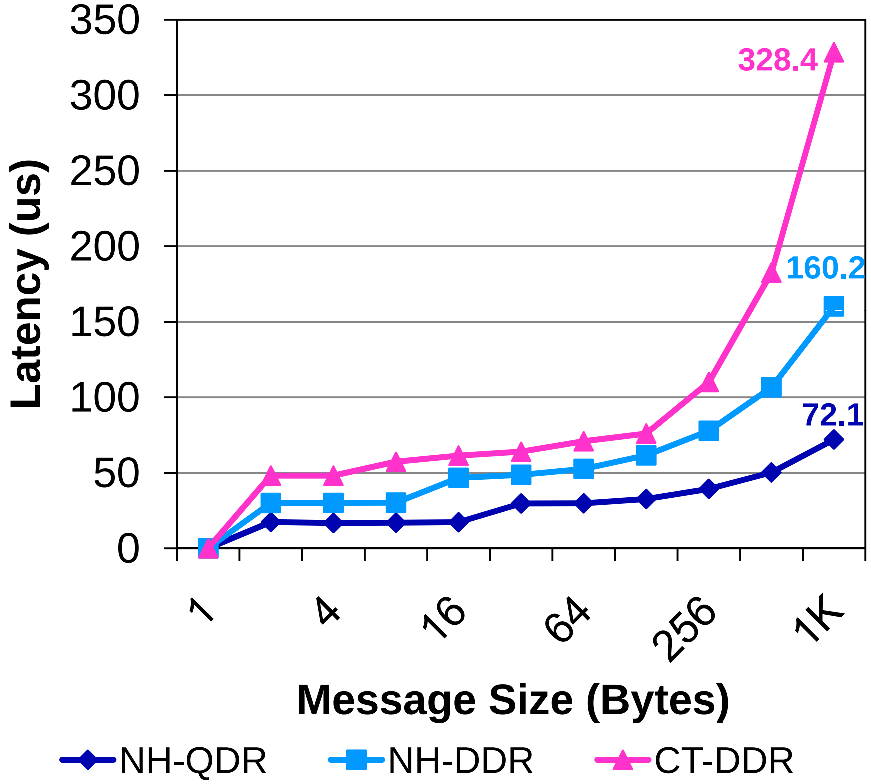
Intra-Node MultiPair Bandwidth

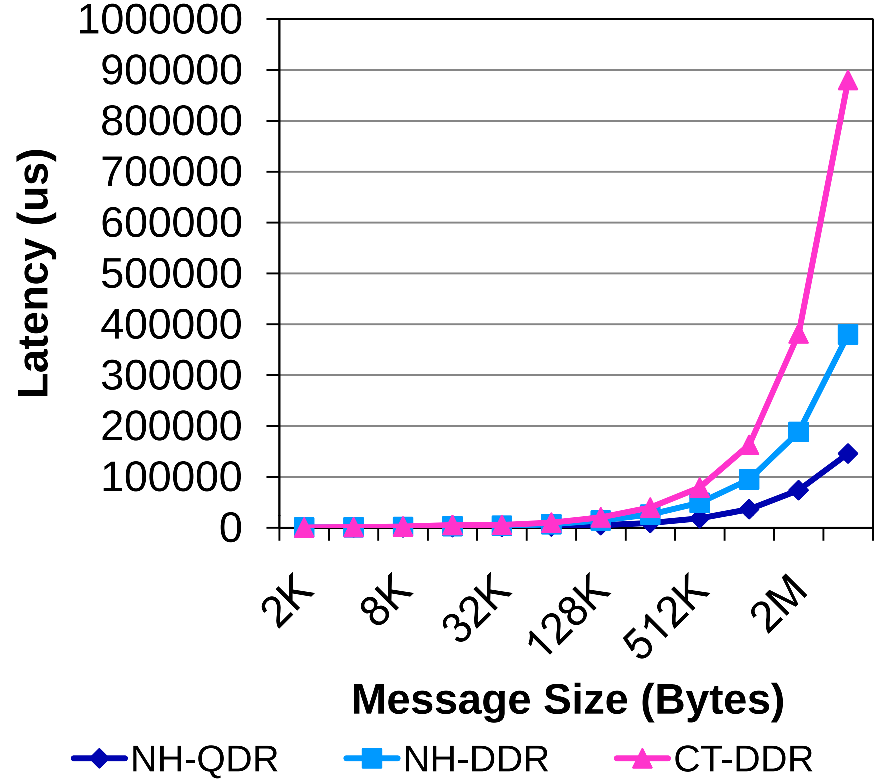




- **Different send/recv buffers are used to negate the caching effect**
- **Nehalem systems show superior memory bandwidth with different send/recv buffers**

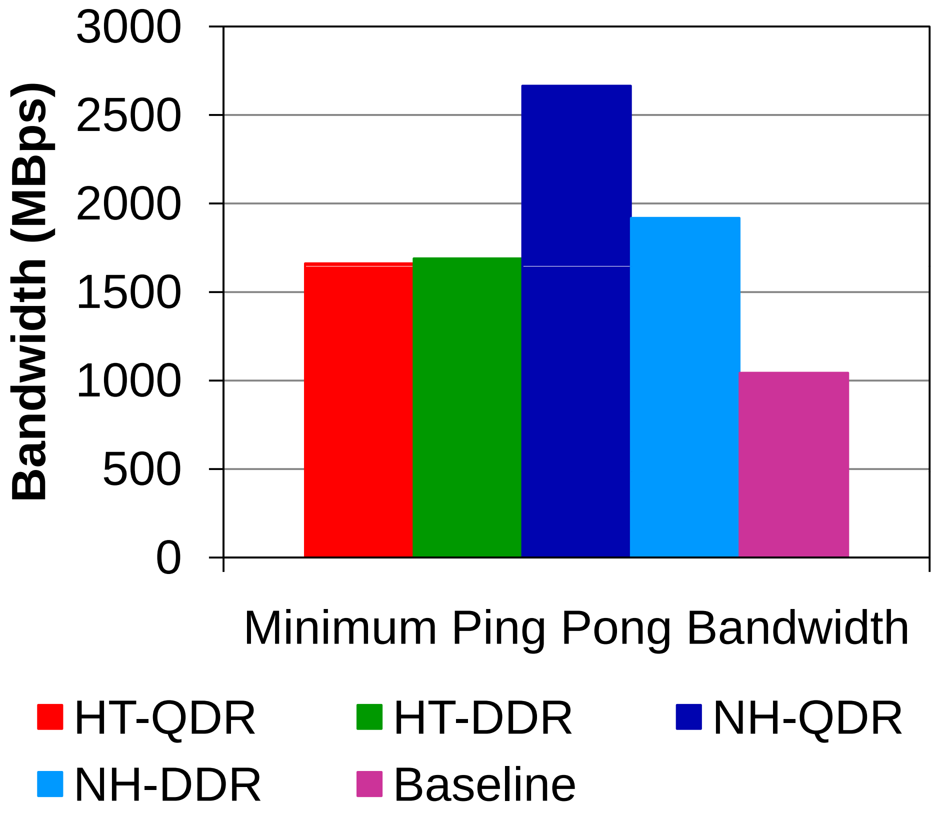
Collective Performance
Alltoall Latency (32-cores)





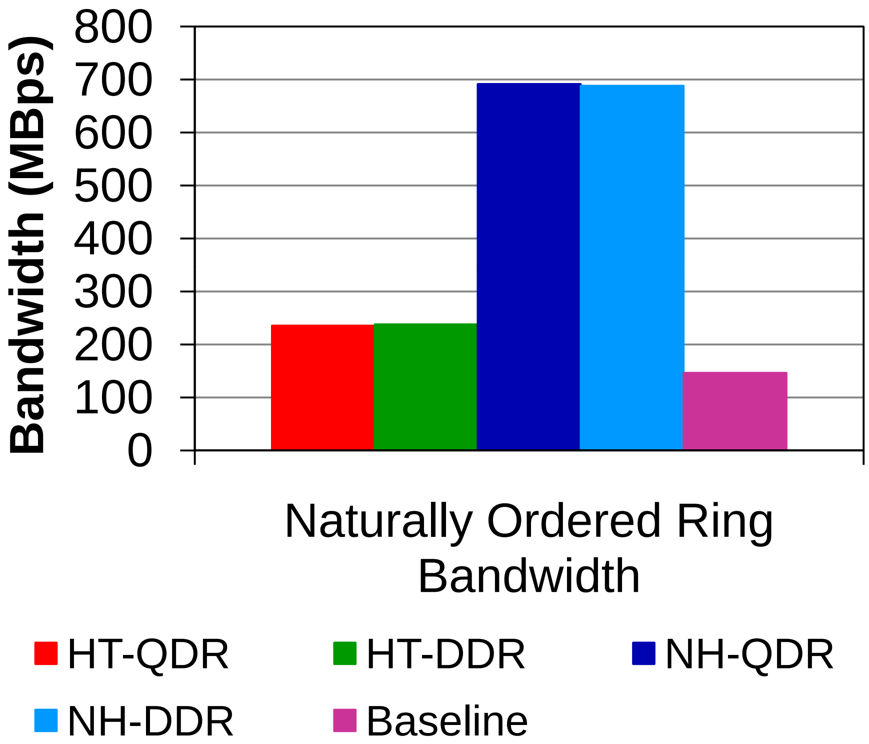
- **A 43% to 55% improvement by using QDR HCA over a DDR HCA**
- **Harpertown numbers not shown due to unavailability of more number of nodes**

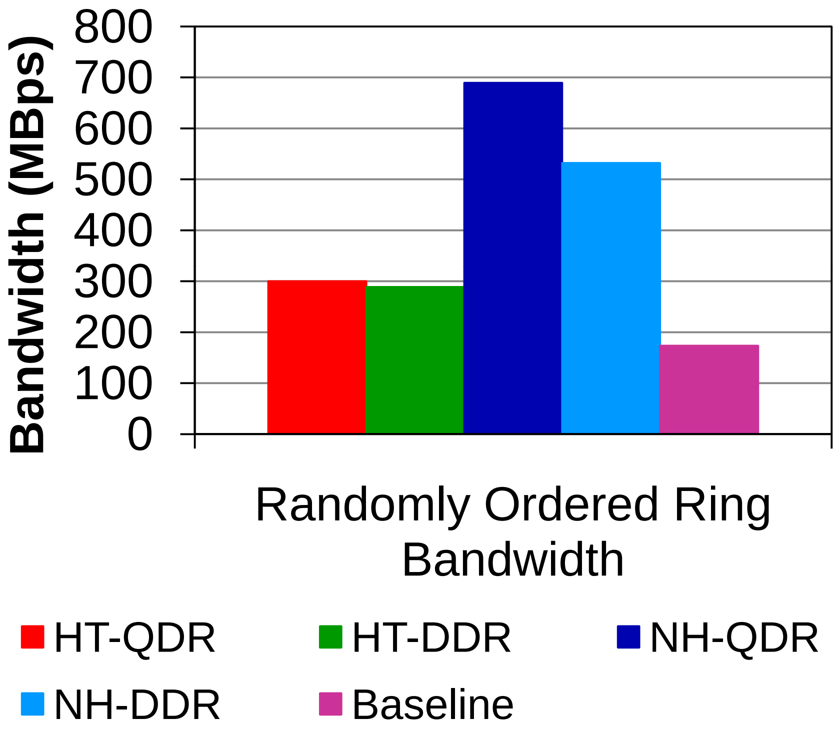
Application Level Evaluation – HPCC



- **Baseline numbers are taken on CT-DDR**
- **NH-DDR shows 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)

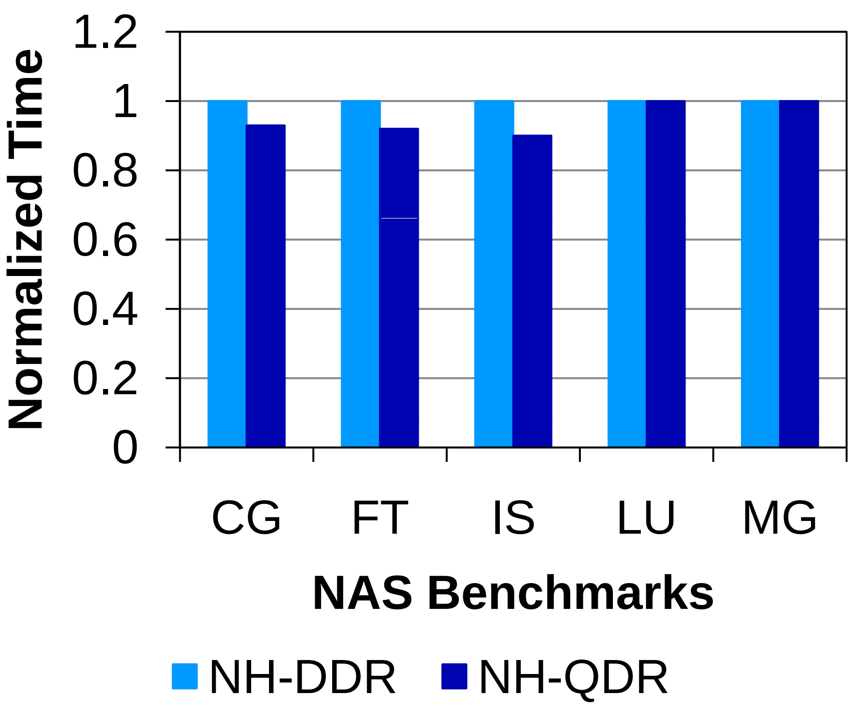




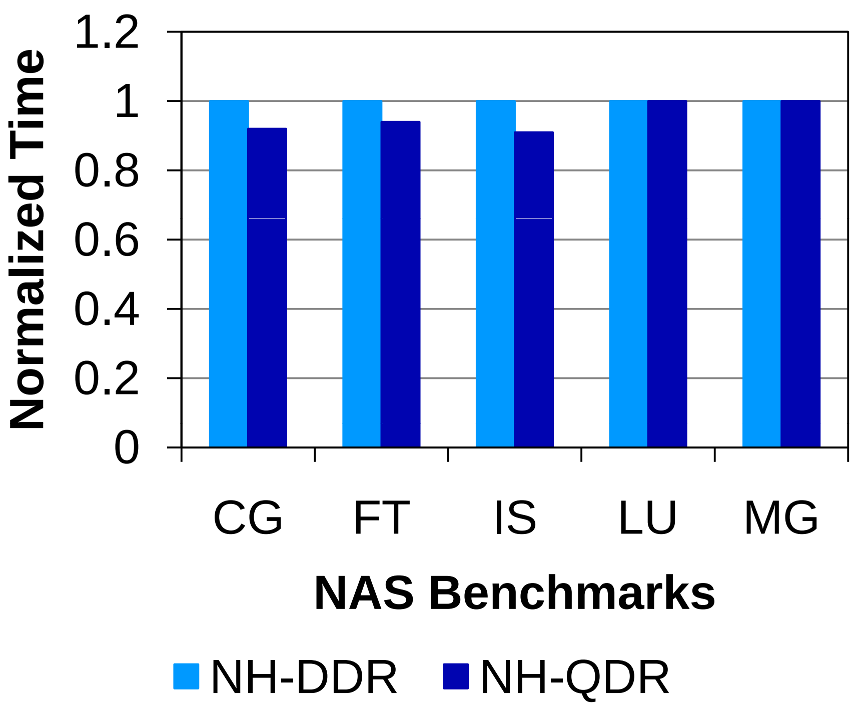
- **Up to 190% improvement in Naturally Ordered Ring bandwidth for NH-QDR**
- **Up to 130% improvement in Randomly Ordered Ring bandwidth for NH-QDR**

Performance of NAS Benchmarks

Class B – 32 processes

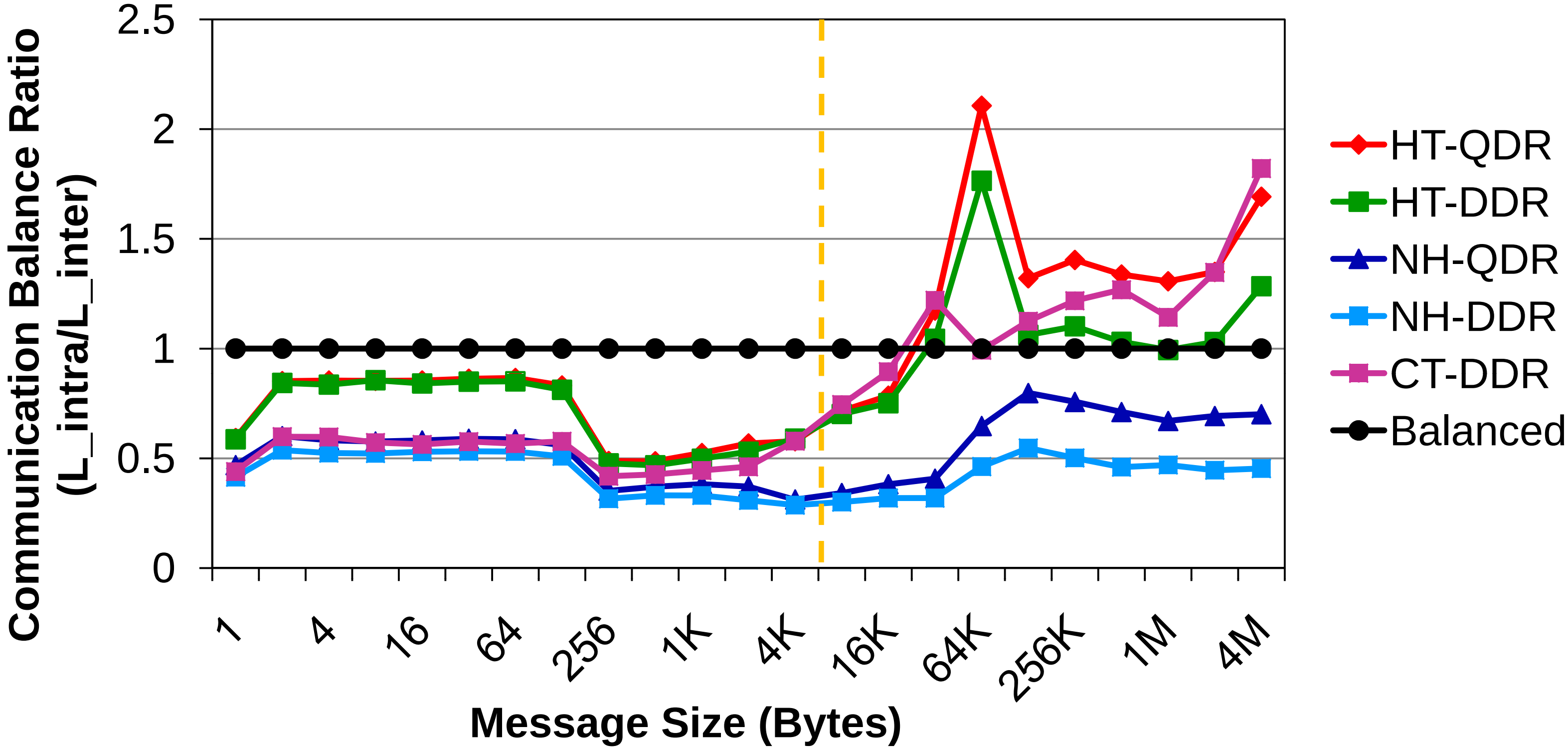


Class C – 32 processes



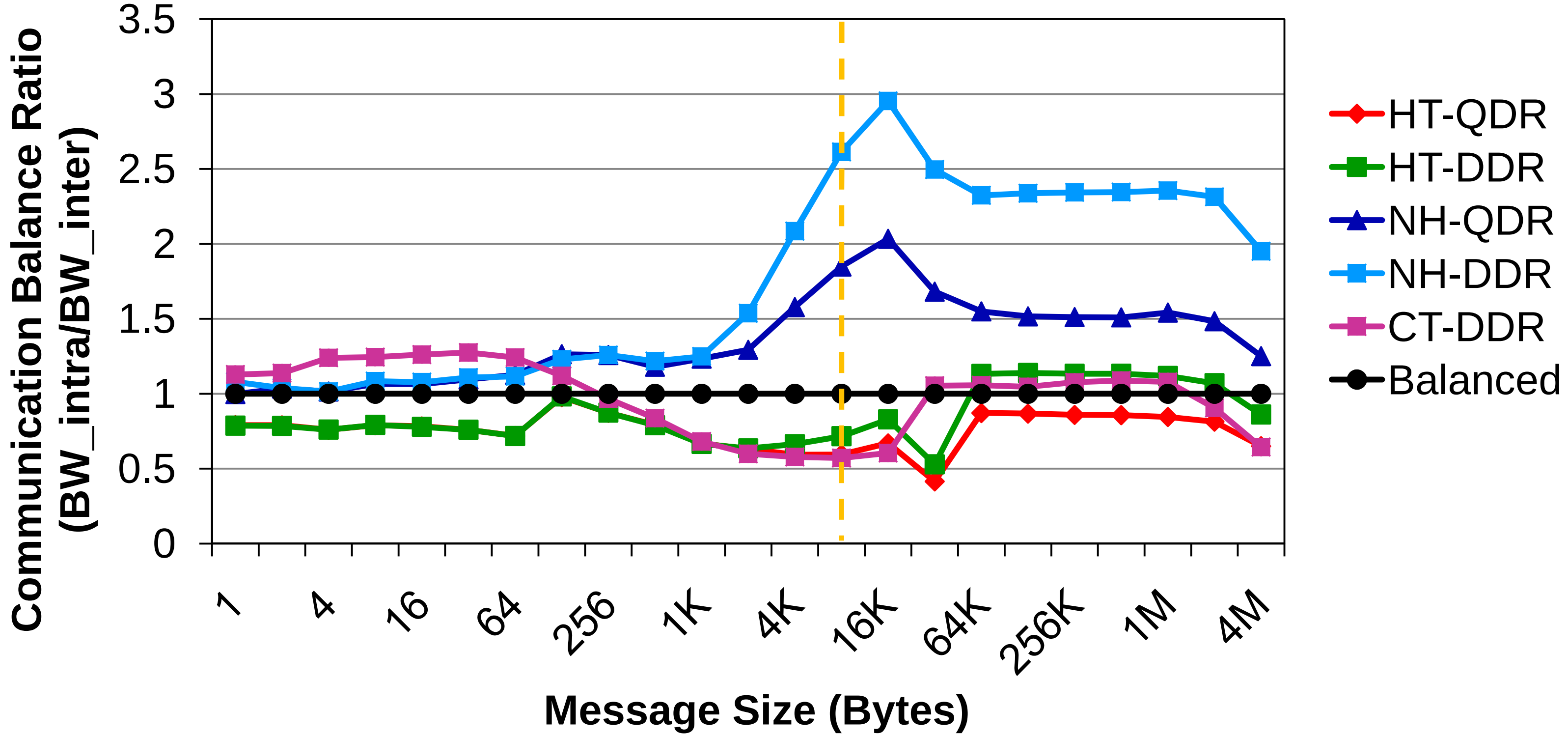
- **Numbers normalized to NH-DDR**
- **NH-QDR shows clear benefits over NH-DDR for multiple applications**

CBR-Latency



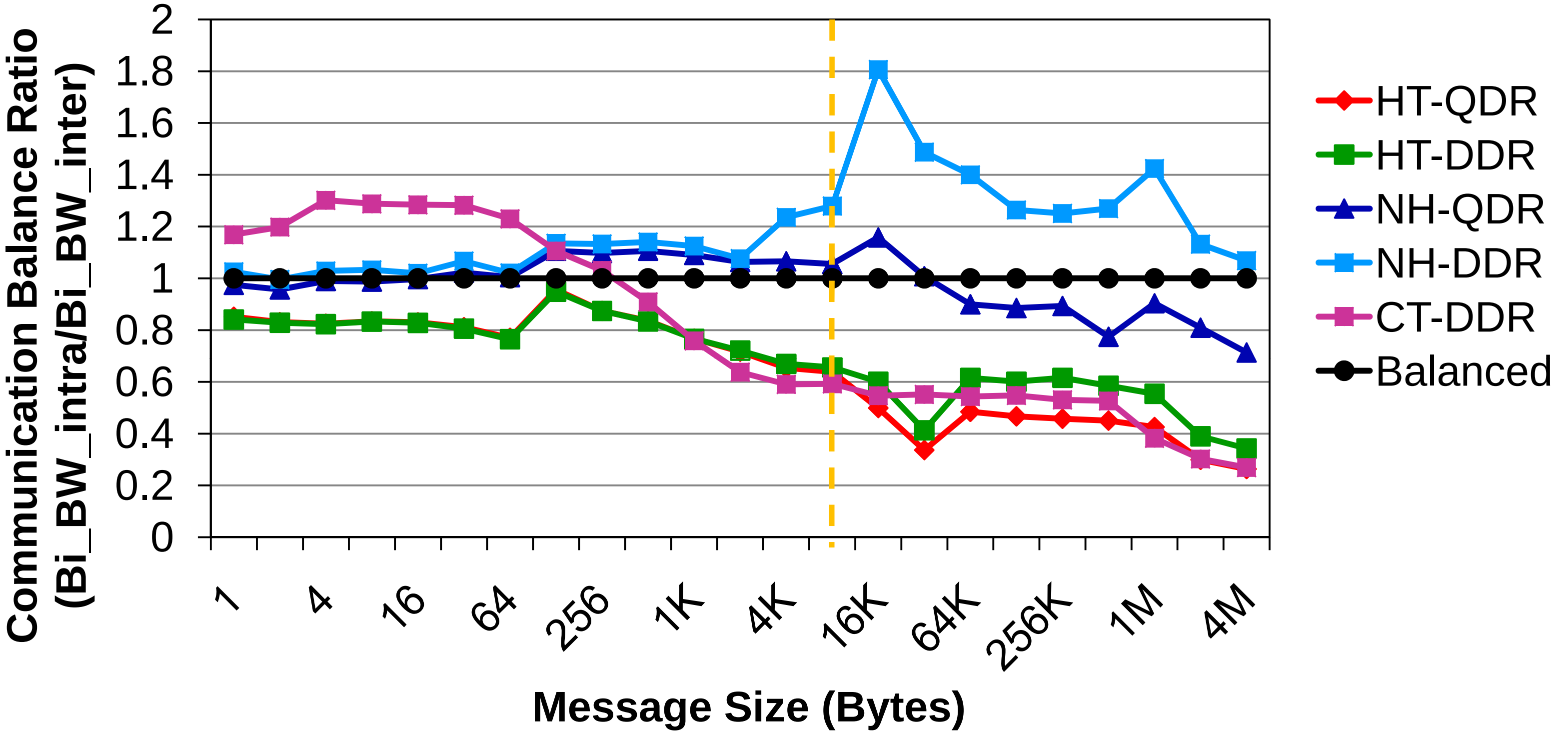
- **Useful for Latency bound applications**
- **Harpertown more balanced for applications using small to medium sized messages**
- **HT-DDR more balanced for applications using large messages followed by NH-QDR**

CBR-Bandwidth



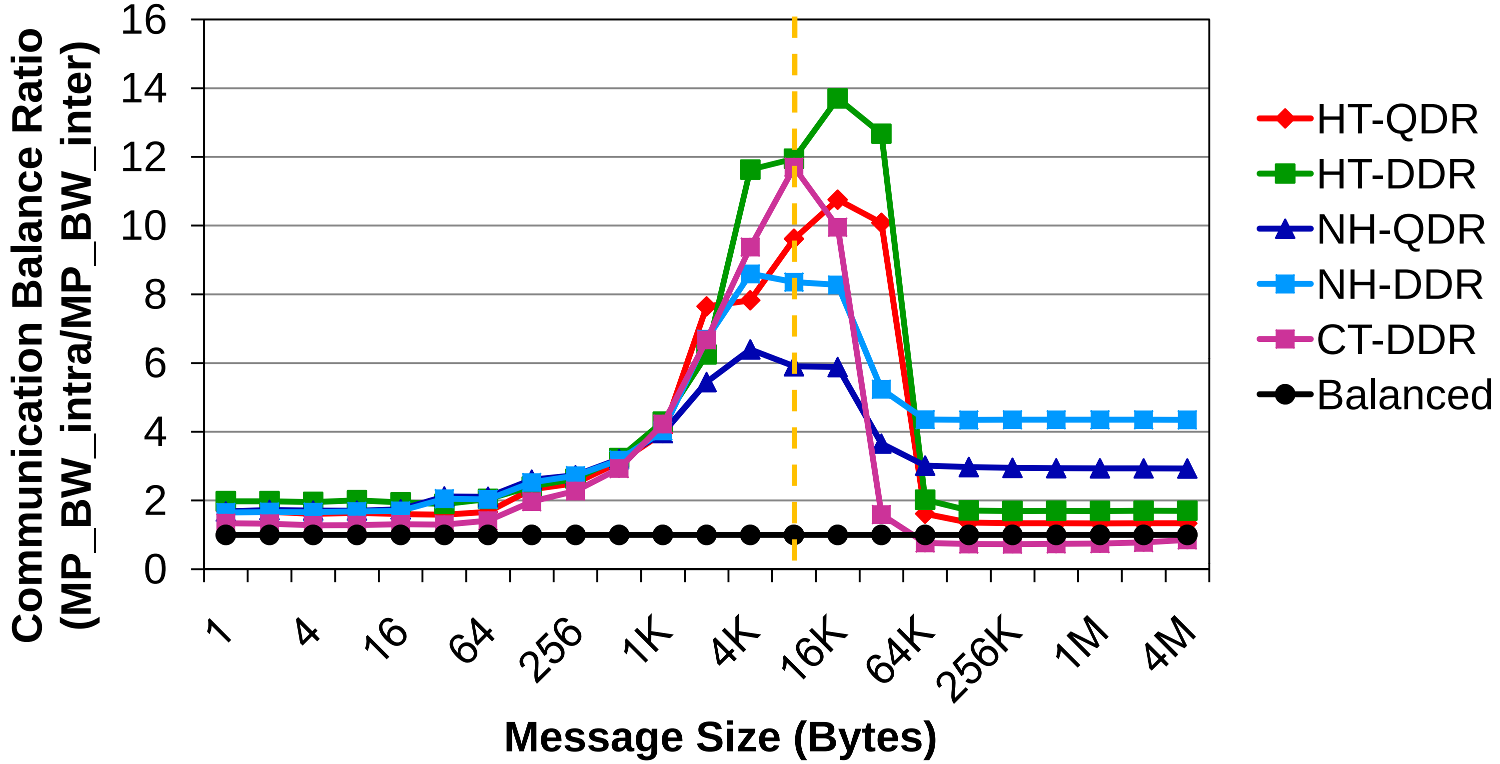
- Useful for Bandwidth bound applications
- Nehalem systems more balanced for applications using small to medium sized messages
- Harpertown systems more balanced for applications using large messages

CBR-Bidirectional Bandwidth



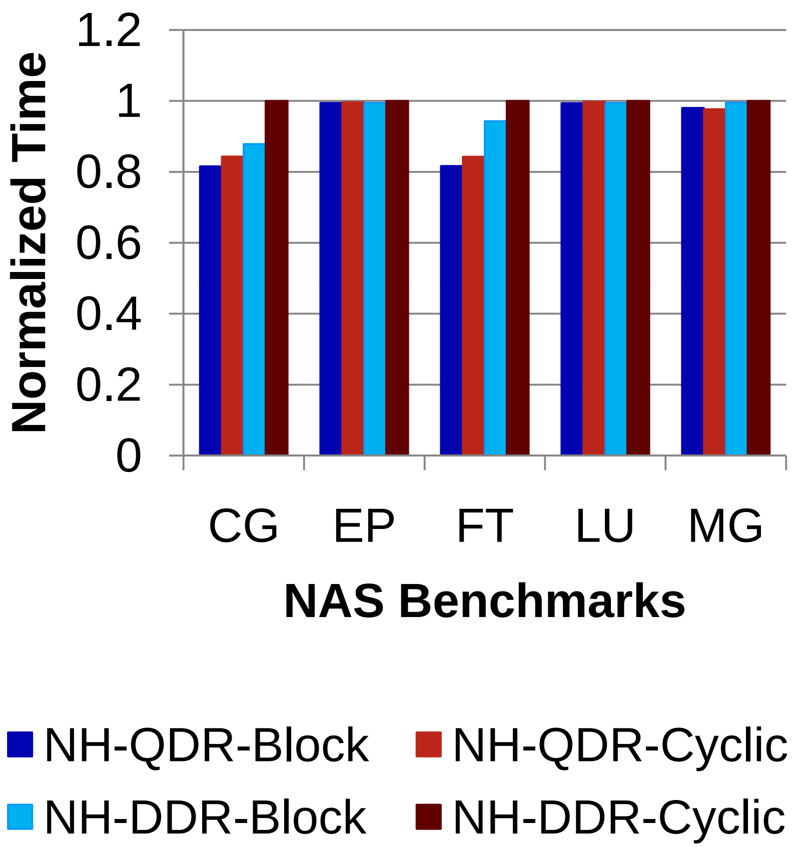
- **Useful for Applications using frequent bidirectional communication pattern**
- **Nehalem systems balanced for applications using small to medium sized messages in bidirectional communication pattern**
- **NH-QDR balanced for all message sizes**

CBR-Multipair Bandwidth



- Useful for Communication intensive applications
- NH-QDR balanced for applications using mainly small to medium sized messages
- Harpertown balanced for applications using mainly large messages

Impact of CBR on Applications (NAS)



- NH-QDR is more balanced than NH-DDR especially for medium to large messages
  - Process mapping should have less impact with NH-QDR than NH-DDR for applications using medium to large messages
- We compare NPB performance for block and cyclic process mapping
- Numbers normalized to NH-DDR-Cyclic
- NH-QDR has very similar performance for both block and cyclic mapping for multiple applications
- CG & FT uses a lot of large messages, hence show difference
- MG is not communication intensive
- LU uses small messages where CBR for NH-QDR and NH-DDR is similar

Outline

- Introduction
- Problem Statement
- Approach
- Performance Evaluation and Results
- Conclusions and Future Work

Conclusions & Future Work

- Studied absolute communication performance of various Intel computing platforms with InfiniBand DDR and QDR
- Proposed a set of metrics related to Communication Balance Ratio (CBR)
- Evaluated these metrics for various computing platforms and InfiniBand DDR and QDR
- Nehalem systems with InfiniBand QDR give the best absolute performance for latency and bandwidth in most cases
- Nehalem based systems alter the CBR metrics
- Nehalem systems with InfiniBand QDR interconnects also offer best communication balance in most cases
- Plan to perform larger scale evaluations and study impact of these systems on the performance of end applications

Thank you !

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

MVAPICH

Network-Based Computing Laboratory

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