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

Hari Subramoni, Matthew Koop and Dhabaleswar K. Panda
Department of Computer Science and Engineering, The Ohio State University
{subramon, koop, panda}@cse.ohio-state.edu

Abstract

Clusters based on commodity components continue to be very popular for high-performance computing (HPC). These clusters must be careful to balance both computational as well as I/O requirements of applications. This I/O requirement is generally fulfilled by a high-speed interconnect such as InfiniBand. The balance of computational and I/O performance is often changing, with the latest change being made by the Intel "Nehalem" architecture that can dramatically increase computing power.

In this paper we explore how this balance has changed and how different speeds of InfiniBand interconnects including Double Data Rate (DDR) and Quad Data Rate (QDR) InfiniBand HCAs. We explore microbenchmarks, the "communication balance" ratio of intra-node to inter-node performance as well as end application performance. We show up to 10% improvement when using a QDR interconnect for Nehalem systems versus a DDR interconnect on the NAS Parallel Benchmarks. We also see up to 25% performance gain with the HPCC randomly ordered ring bandwidth benchmark.

I. Introduction

High-Performance Computing (HPC) in both scientific and corporate environments are highly dependent on maintaining a healthy balance of computing power and I/O interconnection capabilities. For many HPC environments, commodity clusters are used to give performance and reduced costs over proprietary systems. These clusters generally include commodity processors and a high-speed interconnect such as 10GigE or InfiniBand [1]. The performance of the interconnect can make a significant difference in overall application performance.

Over the past few years many changes have taken place in computing architectures that have changed this balance of I/O and computational power. First, multi-core processors have become common and the computing power per node has increased significantly. Also, recently PCIe 2.0 [2] has doubled the maximum transfer rate per lane to 5 Giga-Transfers/sec (GT/s) from 2.5 in PCIe 1.1.

Also, in response to this increased computing power per node, InfiniBand Host Channel Adapters (HCAs) have also been offered with increased data rates. Many current adapters generally run at 4X Single Data Rate (SDR) (10Gb/sec signaling rate) or Double Data Rate (DDR) (20Gb/sec). Using PCIe 1.1, however, InfiniBand DDR was limited in performance by the PCI interconnect. When PCIe 2.0 was released Quad Data Rate (QDR) (40Gb/sec) was also made available for InfiniBand.

In our previous work [3], we showed that PCIe 2.0 was able to increase the performance of some applications, but QDR showed little improvement. Since this study, the new Intel "Nehalem" or "i7" architecture has been released, which adds an on-chip memory controller and a NUMA architecture for higher memory bandwidth.

In this paper we explore how these continued changes in the computational power and I/O bandwidth interact with each other. In particular, we explore the performance of the new Nehalem systems with both DDR and QDR data rates on InfiniBand. We also examine the computational power differences between prior Intel architectures and the Nehalem architecture to show how much the balance has changed.

We first evaluate each configuration across a number of microbenchmarks. This allows us to quantify the base-level of performance that is available on each platform. We then explore the balance of these machines. To define "communication balance" we look at the ratio of performance between intra-node and inter-node performance. The closer they match the more balanced the system

This research is supported in part by DOE grants #DE-FC02-06ER25749 and #DE-FC02-06ER25755; NSF grants #CNS-0403342, #CCF-0702675 and #CCF-0833169; Equipment donations from Intel, Mellanox, AMD, Advanced Clustering, Appro and Sun Microsystems.

becomes. We take this knowledge and evaluate applications with these configurations. We show that using a QDR interconnect instead of a DDR interconnect on a Nehalem-based cluster can increase performance by up to 10% on the NAS Parallel Benchmarks (NPB). We also see up to 25% performance gain with the HPCC randomly ordered ring bandwidth benchmark.

The remaining part of the paper is organized as follows: Background for the paper is given in Section II. Section III gives an overview of our methodology and platforms used in our evaluation. We present a performance evaluation of various InfiniBand data rates on various computing platforms for various microbenchmarks in Section IV. We present the concept of a communication balance ratio in Section V. We further investigate performance with scientific applications in VI. In Section VII we discuss related work in this area. Finally, we conclude the paper in Section VIII.

II. Background

In this section we present a brief overview of InfiniBand, MPI and modern High Performance systems like Nehalem [4] and High Performance interconnects like the Quad Data Rate (QDR) ConnectX [5] InfiniBand network interconnect from Mellanox [6].

A. InfiniBand Architecture

InfiniBand Architecture (IB) [7] is an industry standard for low latency, high bandwidth, System Area Networks (SAN). An increasing number of InfiniBand network clusters are being deployed in high performance computing (HPC) systems as well as in E-Commerce-oriented data centers. IB supports two types of communication models: Channel Semantics and Memory Semantics. Channel Semantics involve discrete send and receive commands. Memory Semantics involve Remote Direct Memory Access (RDMA) [8] operations. RDMA allows processes to read or write the memory of processes on a remote computer without interrupting that computer's CPU. Within these two communication semantics, various transport services are available that combine reliable/unreliable, connected/unconnected, and/or datagram mechanisms.

InfiniBand supports multiple transport mechanisms. *Reliable Connected* (RC) transport provides a connected mode of transport with complete reliability. It supports communication using both channel and memory semantics and can transfer messages of sizes up to 4GB. On the other hand, *Unreliable Datagram* (UD) is a basic transport mechanism that can communicate over unconnected modes without reliability and can only send messages of up to the IB MTU size only. Further, this mode of communication does not support RDMA operations.

B. MPI over InfiniBand

Message Passing Interface (MPI) [9] is one of the most popular programming models for writing parallel applications in cluster computing area. MPI libraries provide basic communication support for a parallel computing job. In particular, several convenient point-to-point and collective communication operations are provided. High performance MPI implementations are closely tied to the underlying network dynamics and try to leverage the best communication performance on the given interconnect. In this paper we utilize MVAPICH [10] for our evaluations. However, our observations in this context are quite general and they should be applicable to other high performance MPI libraries as well.

C. Nehalem

Nehalem is the code name for the latest in the series of multi-core processors by Intel. This is Intel's true Quad Core processor with L2 cache sharing and utilizing the revolutionary Quick Path Interconnect (QPI) [11] architecture. The Nehalem also feature Intel's proprietary Hyper-Threading [12] technology which enables the processor to give applications the impression more number of processing units than is actually available on the system. It is also capable of the Turbo-Boost feature whereby it can increase the operating frequency of one of the cores in the system.

D. ConnectX InfiniBand Interconnect

ConnectX [5] is the fourth generation InfiniBand Host Channel Adapter (HCA) from Mellanox Technologies. It has two ports, with 8 virtual lanes for each. It provides fine-grained end-to-end QoS and congestion control with the latest drivers. Each port can be independently configured to be used either as 4X InfiniBand or 10 Gigabit Ethernet. These cards are capable of delivering 10Gb/sec - Single Data Rate (SDR), 20Gb/sec - Double Data Rate (DDR) and 40Gb/sec - Quad Data Rate (QDR). For this evaluation, we use the DDR and QDR versions of the cards.

III. Methodology

In this section we describe the experimental setup and the combinations we evaluate.

A. Experimental Setup

We use three generations of multi-core architectures from Intel for our evaluation:

- *Clovertown* (CT): Intel Clovertown series of processors using Xeon Dual quad-core processor nodes operating at 2.33GHz with 6GB RAM and a PCIe 1.1 interface
- *Harpertown* (HT): Intel Harpertown series of processors using Dual quad-core processor nodes operating at 2.83GHz with 8GB RAM and a PCIe 2.0 interface
- *Nehalem* (NH): Intel Nehalem series of processors with Dual quad-core processor nodes operating at 2.40GHz with 12 GB RAM and a PCIe 2.0 interface

The older PCIe 1.1 bus on the Clovertown machines restricts the bandwidth that can be achieved. The Harpertown and Nehalem machines have significantly higher bandwidth due to the presence of the PCIe 2.0 interface.

We also have three different InfiniBand HCAs available for comparison: DDR PCIe 1.1, DDR PCIe 2.0 and QDR PCIe 2.0. Even DDR is constrained by PCIe 1.1, so QDR is not available for PCIe 1.1.

For all microbenchmark level experiments, the nodes were connected in a back to back manner. For application level experiments, requiring more than 16 cores, we used a switch to interconnect the different machines. A Flextronics 144-port DDR switch is used to connect all the nodes Clovertown nodes, while a Mellanox 24-port QDR switch was used to connect the Nehalem nodes.

B. Configurations

Given the architectures and InfiniBand HCAs, we evaluate the following five configurations:

- *Clovertown-DDR (CT-DDR)*: This uses the older Clovertown architecture with the PCIe 1.1 bus and an InfiniBand DDR card.
- *Harpertown-DDR (HT-DDR)*: This uses the Harpertown architecture with the PCIe 2.0 bus and an InfiniBand DDR card.
- *Harpertown-QDR (HT-QDR)*: Same as HT-DDR, but with a QDR InfiniBand interconnect.
- *Nehalem-DDR (NH-DDR)*: This configuration uses the new Nehalem architecture and the DDR interconnect.
- *Nehalem-QDR (NH-QDR)*: Same as NH-DDR, but with a QDR InfiniBand interconnect.

IV. Microbenchmark Performance Evaluation

In this section we evaluate performance of each configuration on various MPI-level microbenchmarks. We first show performance of the point-to-point benchmarks followed by MPI collective benchmarks.

A. Point-to-Point

Figures 1 and 2 show the inter-node performance of the various systems under consideration. As we can clearly see, the Nehalem machines with the latest QDR rate cards provide the best possible performance as far as bandwidth is concerned. The Nehalem machines delivers an MPI level uni-directional bandwidth of up to 3,029 MBps and bi-directional bandwidth of up to 5,730 MBps.

The apparent loss in latency performance in Figure 1(b) is due to the relatively slower clock speeds of the Nehalem machines used in our experiments as opposed to the Harpertown (HT) and Clovertown (CT) systems as mentioned in Section III. The latency performance of small messages depends on the ability of the processor to send out packets at a very high rate. The slower clock speeds prevent the Nehalem systems used in our experiments from sending the same number of packets per second as the the other systems can, resulting in slightly lowered performance for small message sizes. In this context it is to be noted that there exists other more expensive Nehalem systems which are capable of higher clock rates than the ones we have at our facility. The multi-pair and bi-directional bandwidth tests also show the same performance trends seen with basic bandwidth and latency.

Figures 3 and 4 shows the intra-node performance of the systems. The intra-node tests clearly show the high memory bandwidth provided by the Nehalem systems as opposed to the other machines. The results of the multi-pair bandwidth tests on the other hand seems to go against the trend. This is due to the higher number of cache conflicts that occur with the Nehalem machines as opposed to the other systems. In the normal micro benchmarks, the same send and receive buffers are used each time while performing either a send or a receive operation, causing the buffer to be moved into the processors cache after the first access itself. Nehalem, being a true quad-core processor, shares the cache with more number of cores than the older machines resulting in a greater number of cache conflicts. To alleviate the effect of caching and measure the true memory bandwidth of the systems as well as to make the benchmark more realistic, we modified our benchmarks to use different sets of send and receive buffers. We declare a set of 64 memory regions and use the regions in a round robin fashion for the send and the receive operations. This kind of buffer usage reduces the possibility of the presence of the buffer being used for send and receive in the processors cache. But as we can see from Figure 4 (b), even this optimization only helps when the message size is large enough to cause all the messages not to fit in the cache. Thus the true memory bandwidth of the Nehalem systems can be seen for the large messages ($>$ 64 KB).

Since these are intra-node tests, there are no differences

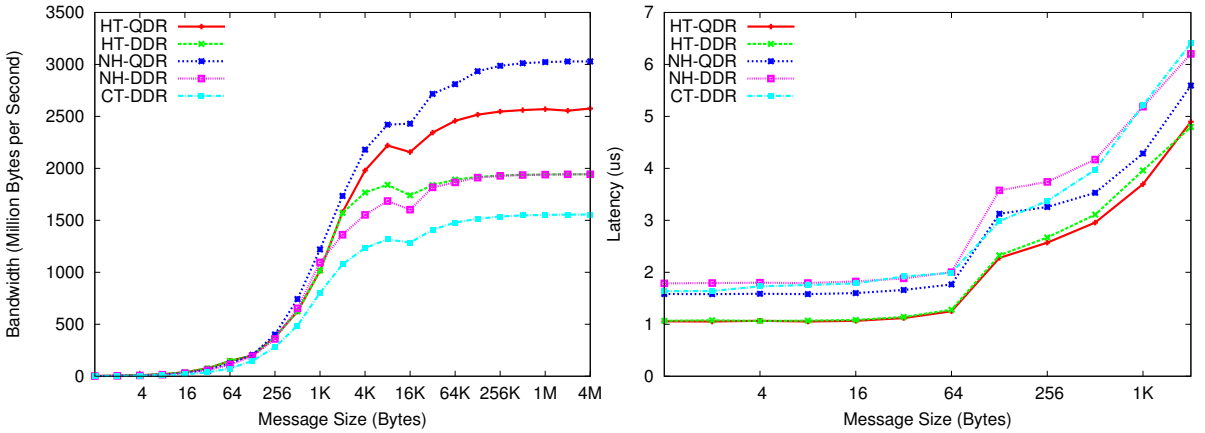

Fig. 1. Inter Node Performance: (a) Bandwidth and (b) Small Message Latency

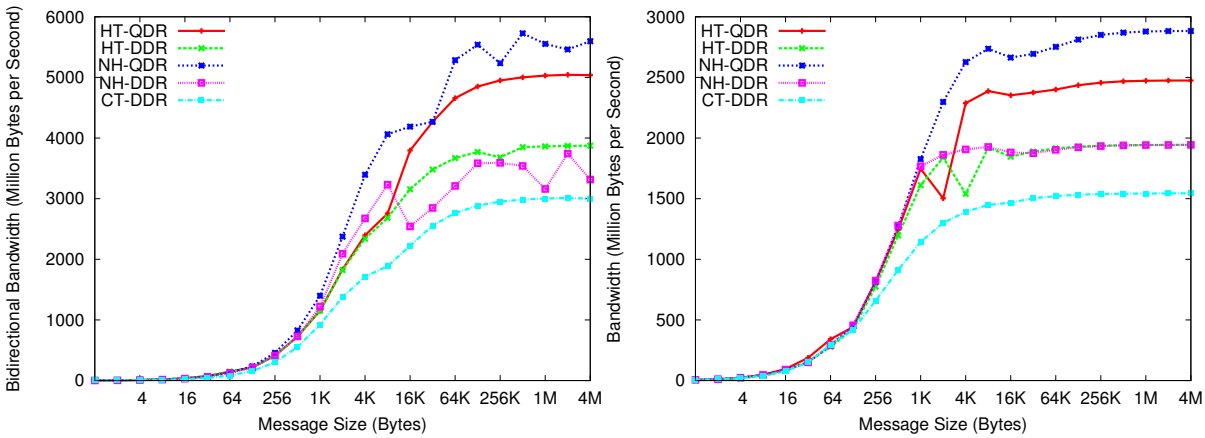

Fig. 2. Inter Node Performance: (a) Bi-directional Bandwidth and (b) Multi-pair Bandwidth

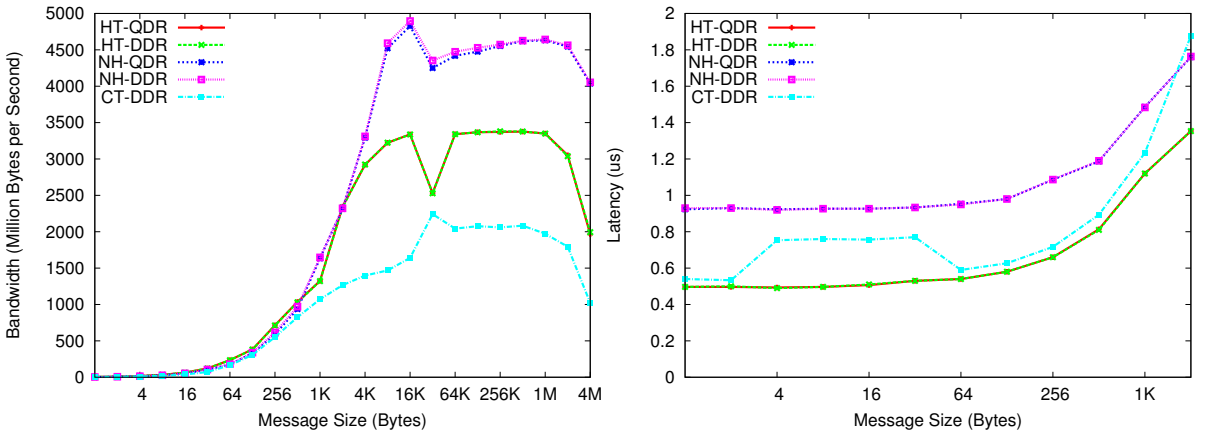

Fig. 3. Intra Node Performance: (a) Bandwidth and (b) Small Message Latency

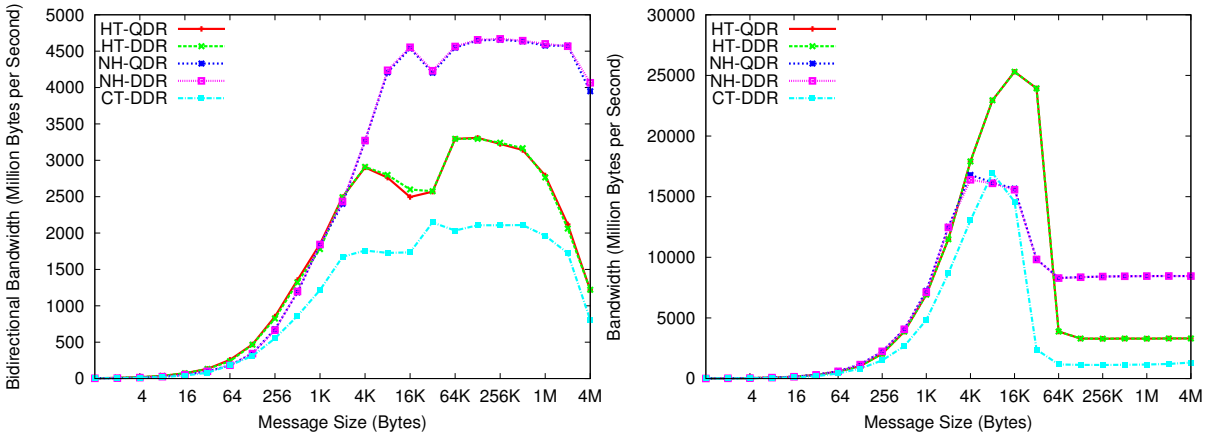

Fig. 4. Intra Node Performance: (a) Bi-directional Bandwidth and (b) Multi-pair Bandwidth

between the tests using the DDR and QDR version of the interconnect.

B. Collectives

In this section we show the results of our evaluation of the various collective algorithms on NH-QDR, NH-DDR and CT-DDR systems. Experiments were not performed on HT-DDR or HT-QDR since at the time of writing only two nodes (of each type) were available. Our experiments were performed over four systems each having eight cores giving a total of 32 computational cores. Our experimental evaluation shows uniform improvement in performance for all message sizes for some of the commonly used collectives such as Alltoall, Scatter and Broadcast.

Figures 5 (a) and (b) show the performance of the *Alltoall* algorithm for small and large messages, respectively. As we can see, utilizing the latest QDR interconnects with the Nehalem systems yields the best performance. We are able to achieve more than 100% improvement in the performance of all message sizes.

Figures 6 (a) and (b) show the performance of the *Scatter* collective communication algorithm for small and large message respectively. The trends seen with the Alltoall algorithm are continued here, with NH-QDR outperforming the NH-DDR by a significant margin. We are able to gain a performance benefit of more than 100% by utilizing the QDR HCAs over the DDR HCAs.

V. Communication Balance Ratio

In this section we explore the concept of a "balanced" system by looking at ratio of intra-node versus inter-node communication performance.

A. Metric

This balance metric looks at intra-node communication performance as compared to the inter-node communication performance. If these ratios are equal to 1 it means that the inter-node communication performance that can be obtained can be similar to that of the intra-node. The closer the ratio is to 1, the more balanced the system will be, in the sense that the physical location of the process, i.e, in the same node or on a different node would not matter as far as the communication performance of the application goes.

We look at four different sets of parameters: Unidirectional Bandwidth: BW_{intra}/BW_{inter}, Latency: L_{intra}/L_{inter}, Bidirectional Bandwidth: $BiBW_{intra}/BiBW_{inter}$ and Multi-pair Bandwidth: $MP\_BW_{intra}/MP\_BW_{inter}$.

B. Evaluation

We take an evaluation of each of the combinations in Section III and plot the ratio of intra-node versus inter-node communication performance.

Figure 7(a) shows the communication balance ratio for bandwidth. We observe that the CT-DDR and NH-DDR systems are the most unbalanced. In the case of CT-DDR, the PCIe 1.1 prevents DDR from achieving a balanced ratio since small messages are also effected. For NH-DDR, the improved memory bandwidth shows that the intra-node bandwidth is much higher than the inter-node bandwidth. This imbalance is corrected when QDR is used.

Intra-node latency is much more balanced in general and is shown in Figure 7(b). The NH-QDR and NH-DDR are generally the most balanced.

Bidirectional and multi-pair bandwidth ratios show the most striking differences between the configurations. From

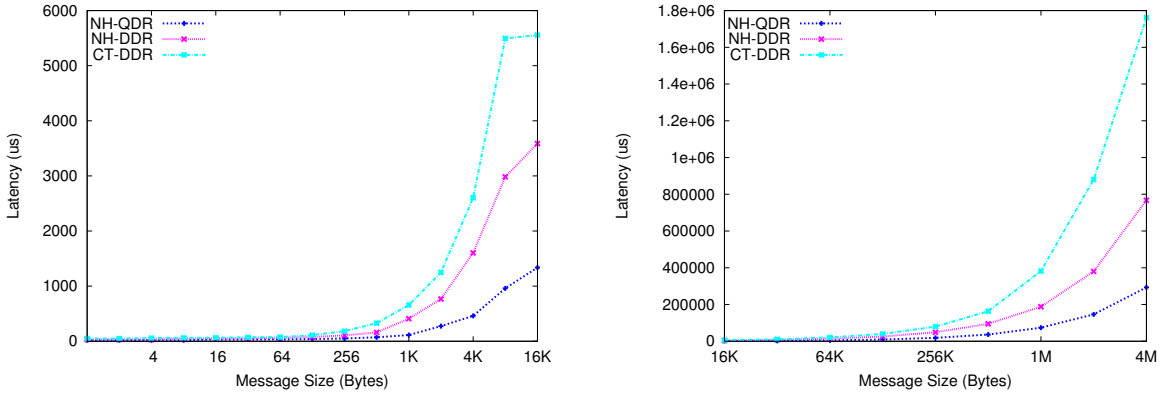

Fig. 5. Performance of Alltoall Collective over 32 cores for (a) Small Messages and (b) Large Messages

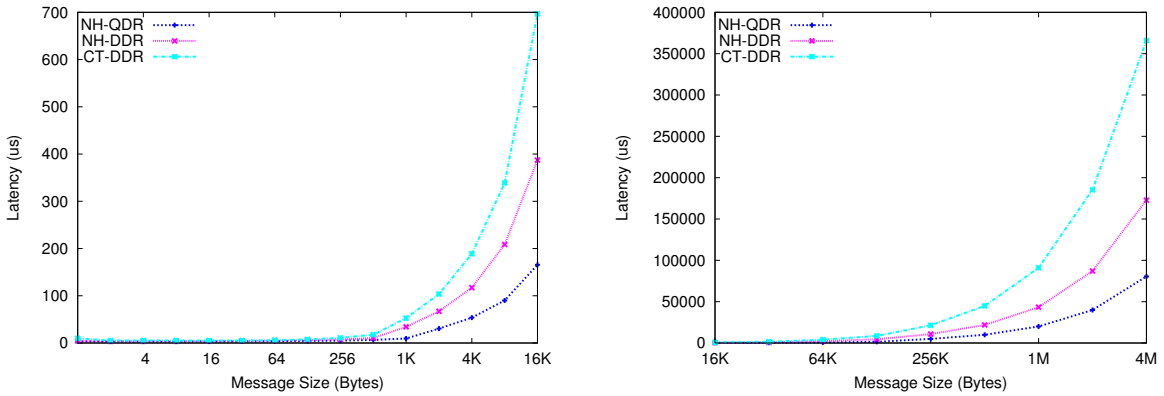

Fig. 6. Performance of Scatter Collective over 32 cores for (a) Small Messages and (b) Large Messages

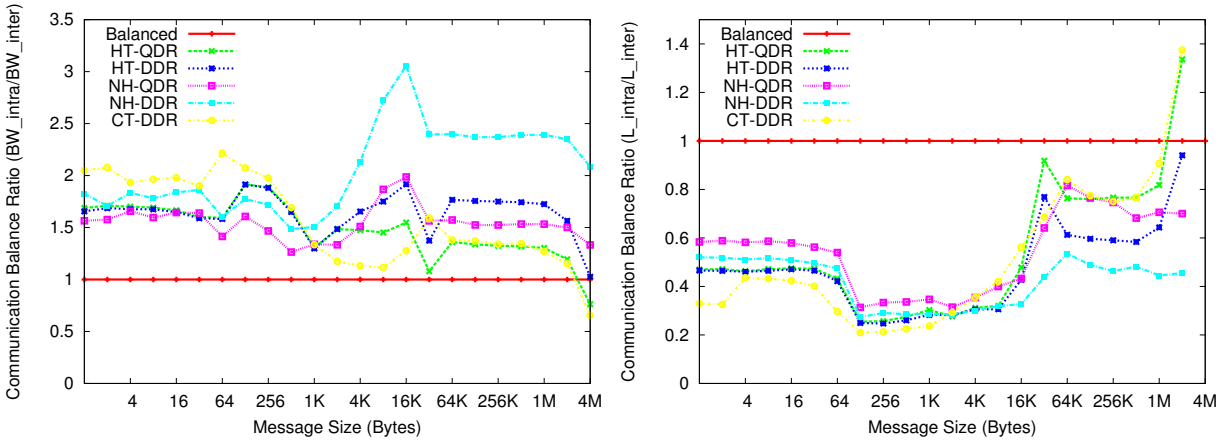

Fig. 7. Communication Balance of Intra Node to Inter Node Performance for (a) Bandwidth and (b) Latency

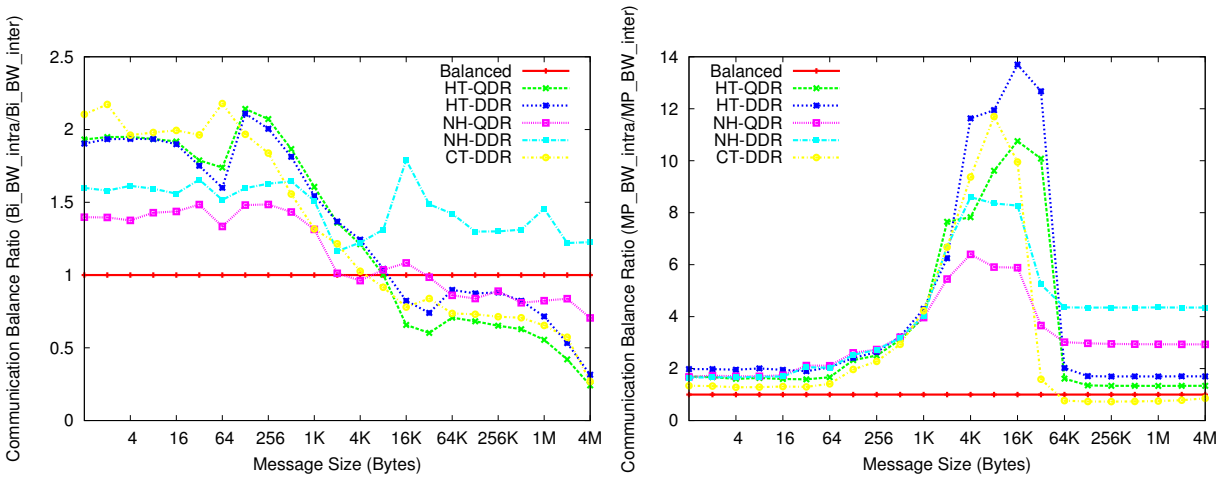

Fig. 8. Communication Balance of Intra Node to Inter Node Performance for (a) Bi-directional Bandwidth and (b) Multi-pair Bandwidth

Figure 8(a), for smaller messages the intra-node bidirectional bandwidth is 50% higher than that of the inter-node, however, as message sizes increase they become roughly the same. NH-QDR shows the most balanced system with having higher inter-node bandwidths to match with the increases in memory bandwidth that Nehalem brings.

The multi-pair bandwidth ratios are shown in Figure 8(b). For large message sizes NH-QDR shows a clear advantage over NH-DDR. The other configurations show a closer ratio, however, this is due to the lower intra-node communication performance that can be achieved on those systems.

VI. Application Level Performance Results

In this section we show the results of the application level performance evaluation we conducted using the HPCC [13] and the NAS [14] benchmarks.

The HPCC benchmark suite is the standard benchmark used to evaluate the ranking of the top supercomputers in the world. The suite consists of seven different benchmarks designed to test various aspects of the supercomputing system.

Figures 9 (a), (b) and (c) show the performance of the HPCC Communication bandwidth and latency benchmark on 16 core systems. As expected, the Nehalem system with QDR interconnects outperforms all other systems. From Figure 9 (a), we see that the QDR enabled systems get up to 28% more Ping Pong bandwidth than the ones without the QDR interconnects. We see similar performance gains with the randomly ordered ring bandwidth, where the QDR enabled systems get up 25% better performance than systems with DDR.

In order to see how the QDR performance affects larger scale applications we ran the NAS Parallel Benchmarks (NPB) on four nodes with eight cores each. NPB is developed at NASA and contains a set of benchmarks which are derived from the computing kernels common on Computational Fluid Dynamics (CFD) applications. Figure 10 shows the normalized performance of the various NAS Parallel Benchmarks run on 32 processes. We see performance improvement of up to 10% for the NAS - Integer Sort (IS) benchmark. Similar performance improvement is seen for the CG and FT benchmarks. The actual performance numbers for the various benchmarks are given in Table I and II.

VII. Related Work

Liu, et al. [15], provided a performance evaluation of PCI-X based InfiniBand adapters versus the first-generation PCI-Express InfiniBand adapters. The authors in [3] showed the benefits QDR can offer over PCI-Express 2.0 gen2 interface. Other authors have done similar evaluations of high performance computing systems and interconnects in the past. Previous evaluation by Sur, et al., compared ConnectX and InfiniHost III HCAs from Mellanox and showed performance improvements for multi-core systems [16].

Authors in [17] and [18] have evaluated the InfiniPath series of InfiniBand interconnects from QLogic. In [19], the authors have demonstrated the improvement in performance that can be obtained with the then state-of-the-art Bensely platform from Intel. Our work explores how we can combine advanced computing platforms like Nehalem and the latest QDR InfiniBand interconnects to create a communication-balanced high performance computing system for next generation supercomputing clusters.

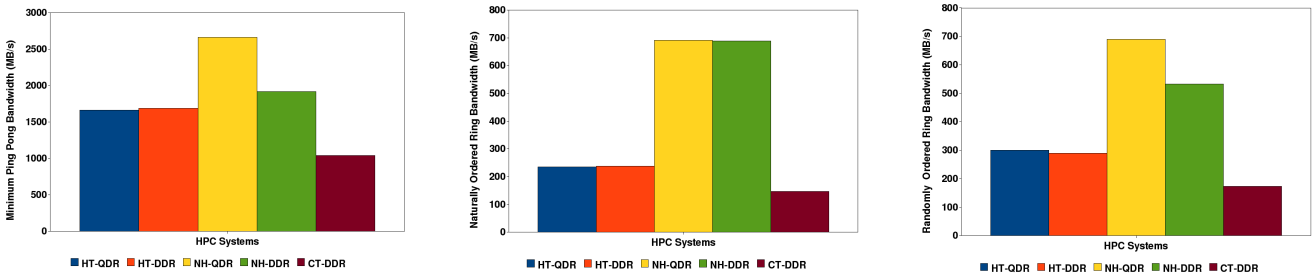

Fig. 9. Performance of HPCC Benchmark: (a) Ping Pong Bandwidth, (b) Naturally Ordered Ring Bandwidth and (c) Randomly Ordered Ring Bandwidth

TABLE I. Performance (in seconds) of NAS Parallel Benchmark Suite for Class C on 32 Processes (Four nodes with eight cores each)

Benchmark	CT-DDR	NH-DDR	NH QDR
CG	64.5	12.13	11.4
FT	55	23.6	22.2
IS	4.4	1.75	1.6
LU	140	86.8	86.8
MG	31	8.1	8

TABLE II. Performance (in seconds) of NAS Parallel Benchmark Suite for Class B on 32 Processes (Four nodes with eight cores each)

Benchmark	CT-DDR	NH-DDR	NH QDR
CG	26	4.8	4.5
FT	13.7	5.5	5.1
IS	0.92	0.41	0.37
LU	35	22.1	22.1
MG	4.8	1.1	1.1

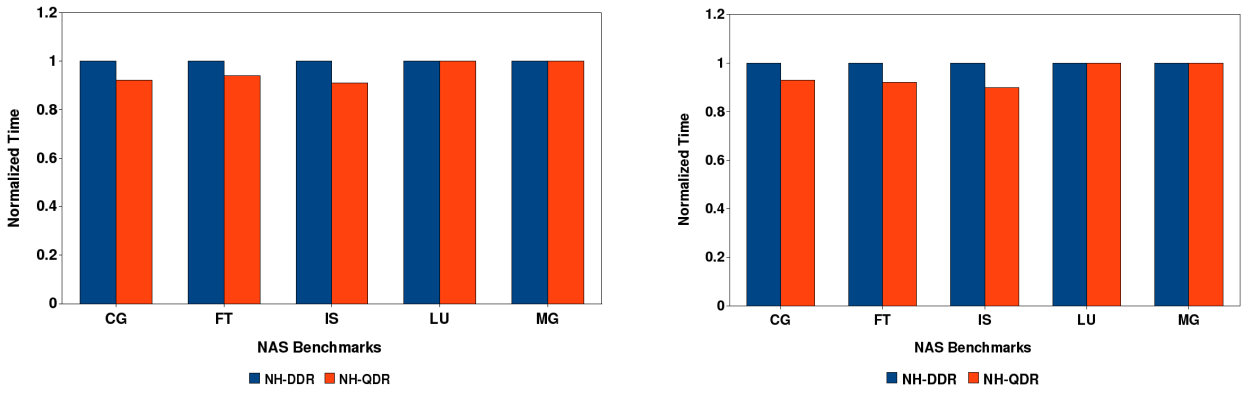

Fig. 10. Normalized Performance of NAS Parallel Benchmark Suite for (a) Class C 8x4 Processes and (b) Class B 8x4 Processes

VIII. Conclusions and Future Work

We evaluate three generations of computing platforms from Intel - Clovertown, Harpertown and Nehalem with Quad Data Rate and Double Data Rate InfiniBand interconnects over a wide variety of parameters. Through our analysis of the microbenchmark level evaluations, we propose the metric of *communication balance ratio* to rate various systems based on their intra-node and inter-node performance. We see that the latest Nehalem systems with InfiniBand QDR head towards designing communication-balanced cluster. We also observe that, when compared to the same platforms with DDR interconnects, such systems also provide good performance benefits as evidenced by the results of the various microbenchmark, collective and application level evaluations that we did.

As part of future work we intend to conduct evaluations on larger test beds and study the impact that these new systems have on various aspects of high performance computing. We also intend to carry out in-depth studies on how the MPI inter-node and intra-node designs can be improved to take advantage of the latest advances in processor and interconnect technologies.

References

[1] "InfiniBand Trade Association, InfiniBand Architecture Specification, Volume 1, Release 1.0," http://www.infinibandta.com.

[2] PCI-SIG, "PCI Express Base 2.0 Specification," http://www.pcisig.com/specifications/pciexpress/base2.

[3] M. Koop, W. Huang, K. Gopalakrishnan, and D. K. Panda, "Performance Analysis and Evaluation of PCIe 2.0 and Quad-Data Rate InfiniBand," in *HotI16*, August 2008.

[4] Intel Corporation, http://www.intel.com/technology/architecture-silicon/next-gen.

[5] Mellanox Technologies, "ConnectX Architecture," http://www.mellanox.com/products/connectx_architecture.php.

[6] "Mellanox Technologies," http://www.mellanox.com.

[7] Infiniband Trade Association, http://www.infinibandta.org.

[8] RDMA Consortium., http://www.rdmaconsortium.org/home/draft-recio-iwarp-rdmap-v1.0.pdf.

[9] MPI Forum, "MPI: A Message Passing Interface," in *Proceedings of Supercomputing*, 1993.

[10] MVAPICH2: High Performance MPI over InfiniBand and iWARP, http://mvapich.cse.ohio-state.edu/.

[11] Intel Corporation, http://www.intel.com/technology/quickpath/.

[12] Intel, http://www.intel.com/technology/turboboost/.

[13] HPC Challenge Benchmarks, http://icl.cs.utk.edu/hpcc/.

[14] D. H. Bailey, E. Barszcz, J. T. Barton, D. S. Browning, R. L. Carter, D. Dagum, R. A. Fatoohi, P. O. Frederickson, T. A. Lasinski, R. S. Schreiber, H. D. Simon, V. Venkatakrishnan, and S. K. Weeratunga, "The NAS parallel benchmarks," vol. 5, no. 3, Fall 1991, pp. 63–73.

[15] J. Liu and A. Mamidala and A. Vishnu and D. K. Panda, "Performance Evaluation of InfiniBand with PCI Express," in *HotI12*, August 2004.

[16] S. Sur, M. Koop, L. Chai, and D. K. Panda, "Performance Analysis and Evaluation of Mellanox ConnectX InfiniBand Architecture with Multi-Core Platforms," in *HotI15*, Palo Alto, CA, August 2007.

[17] L. Dickman, G. Lindahl, D. Olson, J. Rubin, and J. Broughton, "PathScale InfiniPath: A First Look," in *HotI-13*, 2005.

[18] R. Brightwell, D. Doerfler, and K. D. Underwood, "Preliminary Analysis of the InfiniPath and XD1 Network Interfaces," in *CAC*, 2006.

[19] M. Koop, W. Huang, A. Vishnu, and D. Panda, "Memory Scalability Evaluation of the Next-Generation Intel Bensley Platform with InfiniBand," in *HotI-14*, August 2006.