HP Fabric Clustering System for InfiniBand™ Interconnect Performance on HP-UX 11iv2

White paper

Executive Summary

This document presents the performance characteristics and capabilities of the recently announced HP Fabric Clustering System for InfiniBand cluster interconnect product on HP-UX 11iv2 platforms.

The HP-UX Fabric Clustering System product consists of:

- AB286A - HP PCI-X 2-port 4X Fabric (HPC) Host Channel Adapter (HCA)
- AB291A – HP 12-port 4X Fabric Copper Switch
- 5m, 7m, and 10m 4x Fabric Copper Cables
- HP-UX Fabric Clustering System software stack

HP also lists a reference 96-port 4X IB Switch, the Topspin TS170.

Performance Overview

The following three attributes that affect the scalability of a cluster interconnect solution form the heart and soul of performance characterization of any cluster interconnect solution. The HP-UX Fabric Clustering System interconnect solution shows industry leading numbers on all the three fronts:

- Latency
  HP-UX Fabric Clustering System product offers 1-way latency of 4.6usec for 8-byte messages in RDMA programming model, and sub-10usec latencies for short messages of length 512 or less in both Send/Receive and RDMA programming models.
- Bandwidth
  HP-UX Fabric Clustering System product offers near link rate receive side bandwidth (an aggregate of 924MB/s for 3 parallel streams) limited only by the PCI-X point of attachment.
- Service Demand
  HP-UX Fabric Clustering System product completely leverages the offload and OS-bypass features offered by the InfiniBand technology which yields lower CPU utilization while transferring large messages. The service demand is 9% for 64KB messages and is under 5% for 128KB or larger messages.

This is a living document. The document will be periodically updated with the latest performance data on newer platforms and currently addresses any performance issues found thus far. The reader is encouraged to get the latest version available at the HP website http://docs.hp.com.

Document Structure

This document is organized as follows:

Section 1 contains the executive summary and organization of this paper.
Section 2 outlines the test methodology, tools, and performance considerations as used in generating the performance results listed in this paper.
Section 3 presents the HP-UX Fabric Clustering System interconnect solution performance results in point-to-point configurations.
Section 4 presents the HP-UX Fabric Clustering System interconnect solution performance results in switched configurations.
Section 5 lists the configurations on HP Integrity servers not suited for a high performance Fabric Clustering System solution.
Section 6 summarizes the results.

Methodology

An HP-developed application, *itperf,* was used for driving the data on the cluster. *itperf* was used to measure latency, bandwidth and bi-directional bandwidth across the tested topologies. Service demand generated on the servers was also measured while running the tests. HP Integrity servers rx2600 and rx4640 are used as the data source and sink for performance measurements. Various topologies are used for characterizing the cluster interconnect performance; these include both point-to-point and switched configurations.

Performance aspects of non-optimal configurations are listed in the whitepaper to help system administrators avoid configuration mistakes.

itperf

itperf is a user-space program developed over the ICSC defined IT-API v1.0. *itperf* provides two types of tests and a brief description of each of these types is listed below:

- ***Latency tests***
  itperf supports two kinds of latency tests:
    - Send/Receive programming model
    - RDMA programming model

  The following gives a high-level flow of Send/Receive programming model:
  - Transmit/receive buffers share the same physical memory.
  - Posts the same buffer repeatedly for the whole test.
  - Requests Send Work Request completion notification once every SQ size.
  - Blocks for send and receive completions for 0 seconds.

  In the RDMA programming model, *itperf* latency tests use RDMA writes where there will no be receive side completion. The application polls on the data buffer instead.

- ***Bandwidth tests***
  - Transmit/receive buffers share the same physical memory.
  - Maintains a window of 16 buffers
  - Registers the buffers during the test setup.
  - Requests Send Work Queue completion notification once every half the window size.
  - Waits for send and receive completion notifications.
  - Replenishes RQ on receipt of every message.

HP Integrity Servers

The following configurations of HP Integrity servers are used for conducting the performance tests.

rx2600:
- 2 CPUs (IPF 1.5 GHz)
- 4 GB RAM
- HP-UX 11iv2

rx4640:
- 4 CPUs (IPF 1.5 GHz)/ 3 CPUs (IPF 1.5 GHz)
- 2 GB RAM
- HP-UX 11iv2

Performance Considerations

PCI-X is the primary I/O interface for rx2600 and rx4640 HP Integrity servers. PCI-X slot 4 on the rx2600 and PCI-X slots 7 & 8 on the rx4640 are dual rope slots.

NOTE: A rope is defined as a high-speed, point-to-point data bus.

To achieve the best performance using HP-UX Fabric Clustering System product, it is recommended that the *AB286A - 2-port 4X Host Channel Adapter (HCA)* is plugged into one of the available dual rope slots available on an HP Integrity servers. Unless mentioned otherwise, all the results listed in this whitepaper refer to that configuration.

The HP-UX Fabric Clustering System software stack and the HCA provide better performance when large physical pages are used for the buffers used in data transfers. The *itperf* program as used in obtaining the performance results from bandwidth tests is set to request large physical pages from the operating system.

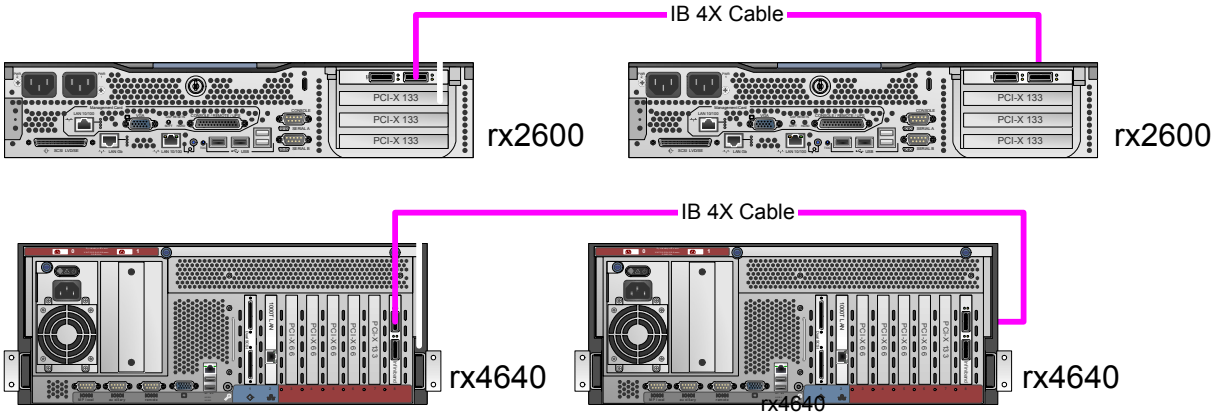
Use *chatr +pd L <application name>* to request the largest physical page available on the host at the time of running the test.

NOTE: Refer to chatr(1M) man page for additional details.

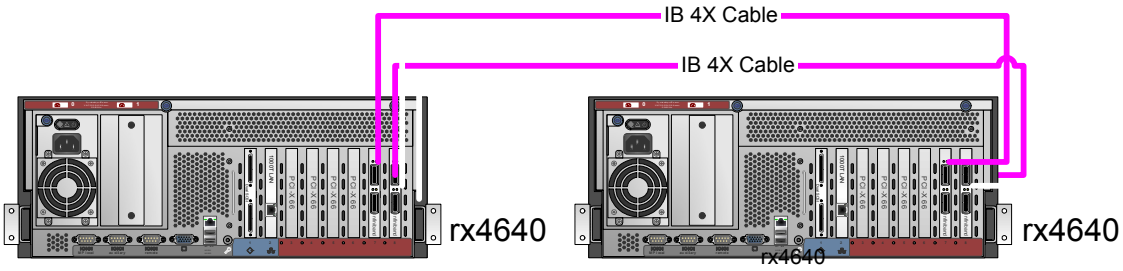
Latency sensitive applications can use a RDMA programming model that eliminates receive side completion processing. An application may also choose to use polling on the EVD for data instead of blocking for completions.

Cluster Solution Performance – Point-to-Point Configuration

A pair of rx2600/rx4640 Integrity servers are connected via a single HCA on each server. One port on each HCA is connected to the other using a 4X cable.

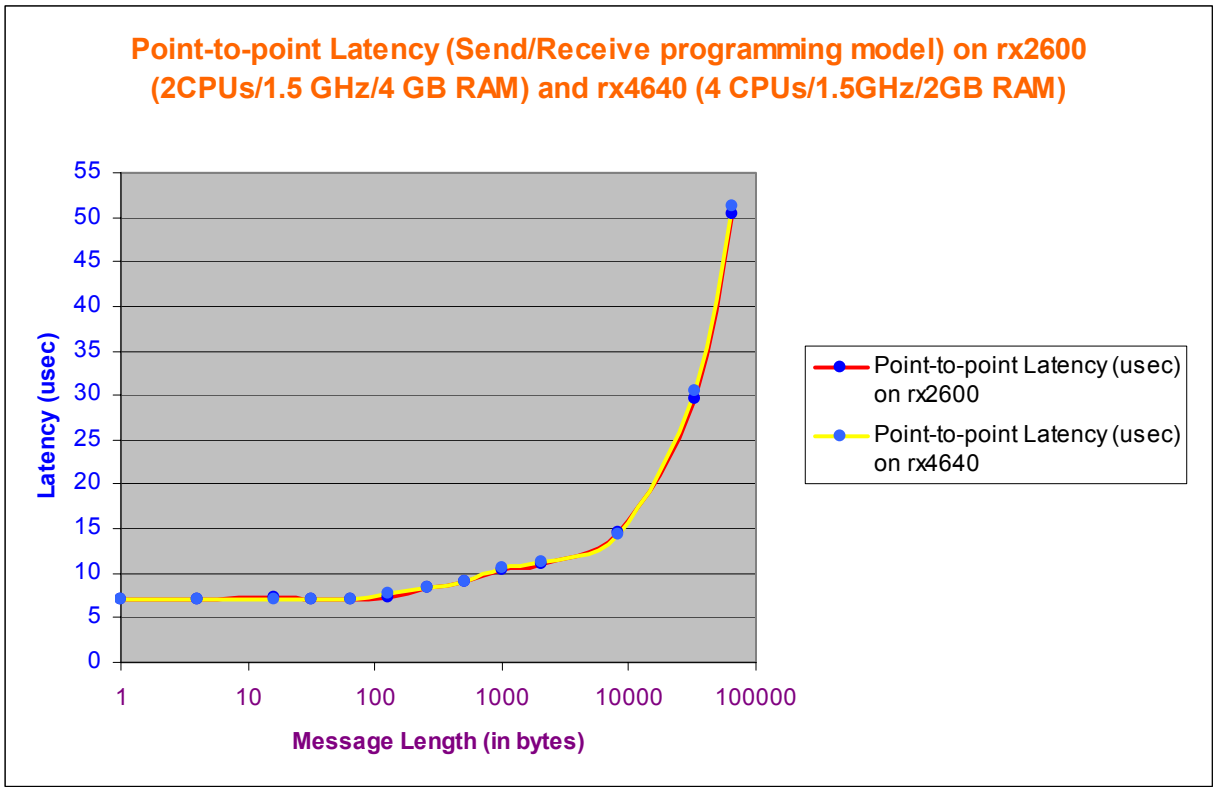


For the HCA scalability tests, 2 HCAs on rx4640 are connected in point-to-point configuration to two HCAs on another rx4640.



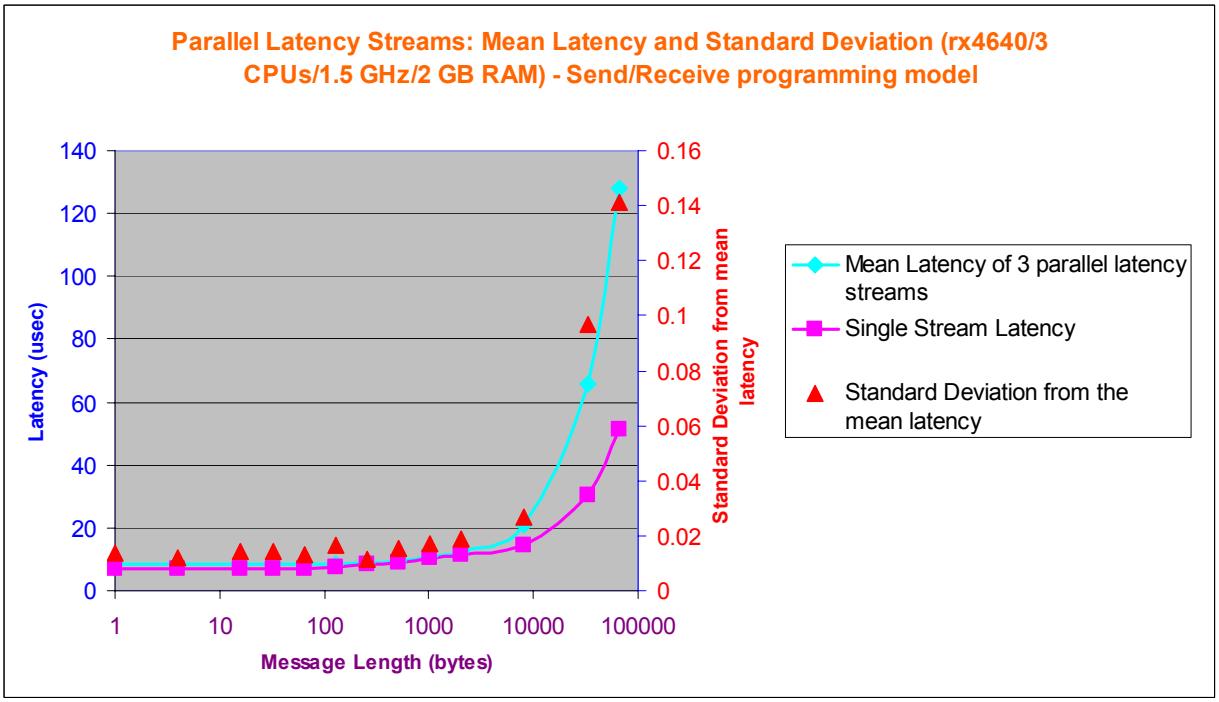
Latency

On rx2600 servers in point-to-point configuration, HP-UX cluster delivers 1-way latency of 4.3usec for 8-byte messages using the RDMA programming model and 7.0usec for 1-byte messages using the Send/Receive programming model. The latency remains low for larger messages. On rx2600 servers in point-to-point configuration, a 512 byte message gets a 1-way latency of 7.1usec using the RDMA programming model and 9.0usec using the Send/Receive programming model. Cumulative service demand on the available processors is 100% as the *itperf* application polls for completion events.



The latency values remain virtually unaffected when the results are obtained from two independent pairs of HCAs plugged into rx4640 servers.

As the number of parallel latency streams that share the same HCA increases, latency suffers as the message size increases. The standard deviation of the latencies for various streams from the mean latency depends on the number of processors on the server. Given below is an illustration of the impacts of 3 parallel streams on mean latency and standard deviation.

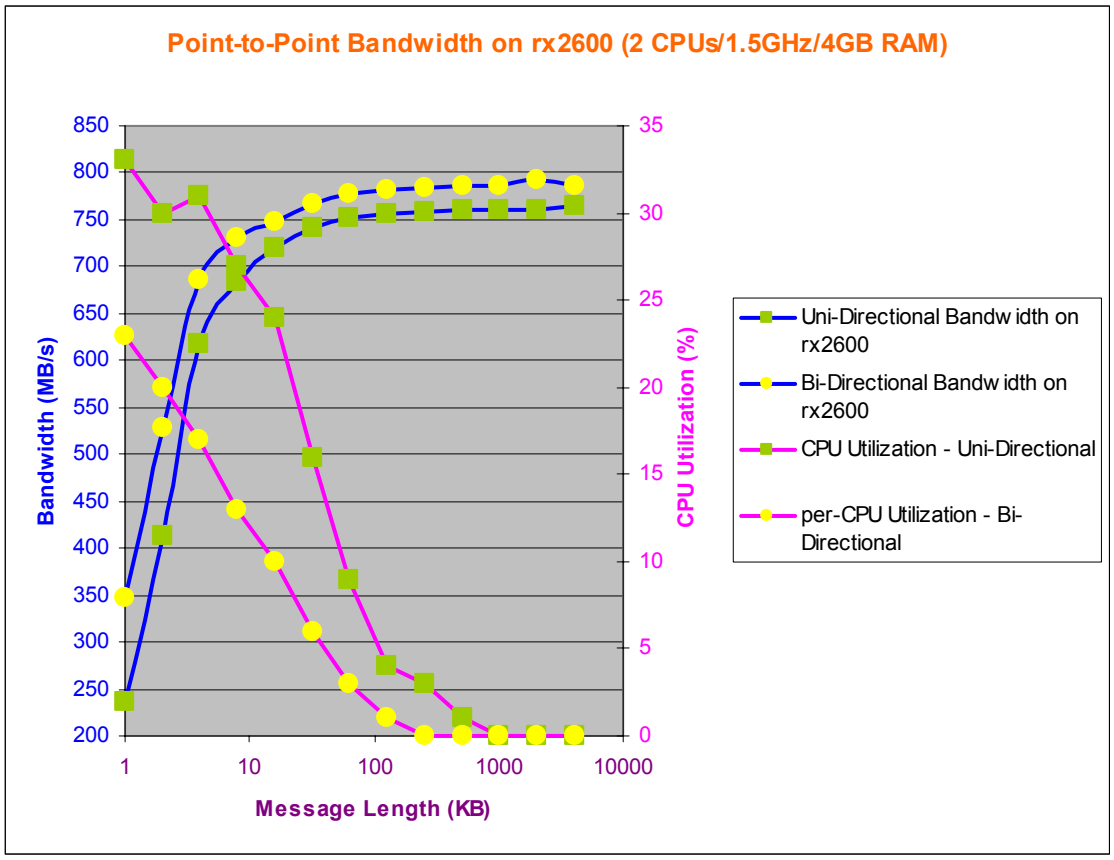


As the number of parallel latency streams, running on the same HCA on a server, start exceeding the number of available processors on the server, process scheduling issues will have an adverse impact on the mean latency and standard deviation values. Process scheduling issues also impact parallel latency streams running on different HCAs, when the number of such streams start exceeding the number of processors on the server.

Bi-directional 1-way latencies on a rx2600 in point-to-point configuration are 5.6usec on each stream in the RDMA programming model for 8-byte messages, and 7.1usec on each stream in the Send/Receive programming model for 1-byte messages. The bi-directional latencies for short messages, up to 512 bytes, remain pretty close to the respective uni-directional latencies. On a rx2600 in point-to-point configuration, the bi-directional 1-way latencies for 512 byte messages are 7.6usec on each stream in RDMA programming model, and 9.1usec on each stream in the Send/Receive programming model.

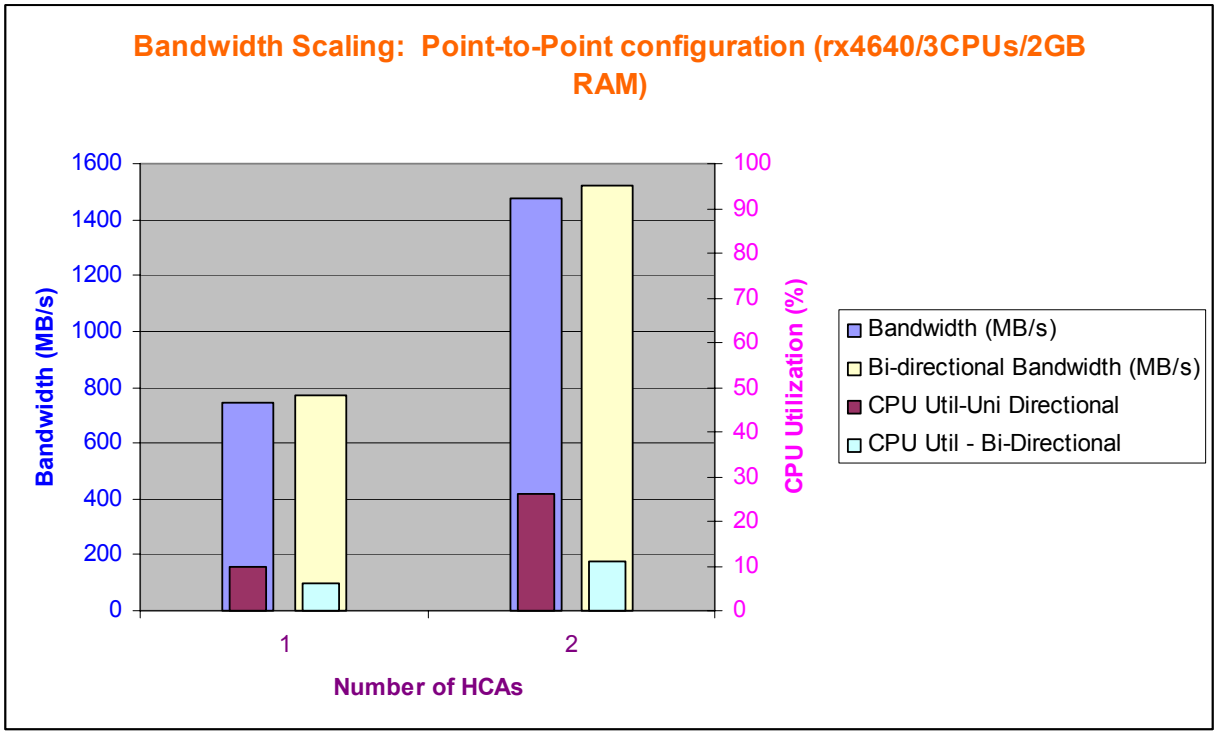
Bandwidth

The HP-UX Fabric Clustering System interconnect solution offers high bandwidth rates in the order of 700MB/s right from messages of length 16KB and more. Service demand requirements on the server falls exponentially as the message size increases providing the applications with a high bandwidth at lower CPU utilization rates.



Under bi-directional traffic loads, an acid test for any system performance, HP-UX Fabric Clustering System interconnect solution offers excellent performance results. The bi-directional bandwidth on a rx2600 crosses 731MB/s at a short message length of 8KB with a service demand of only 13% per-CPU utilization. The bi-directional bandwidth is almost equally split across each individual stream for all message sizes.

HP-UX Fabric Clustering System interconnect solution scales almost linearly (approximately 2X for 2 HCAs), in both uni-directional as well as bi-directional bandwidth rates on a rx4640 with 2 HCAs.

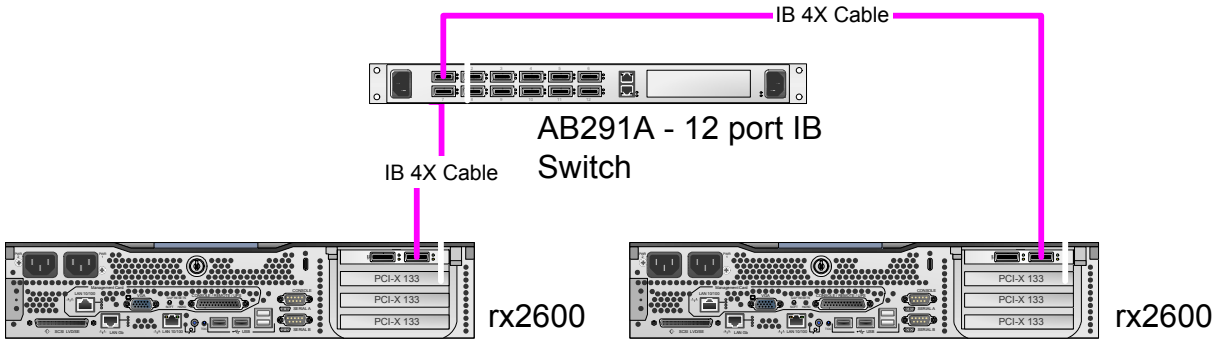


The HP-UX Fabric Clustering System interconnect solution scales without adversely affecting the service demand requirements on the server. Because of lack of dual rope slots on rx4640, HCA scaling was tested only up to 2 HCAs.

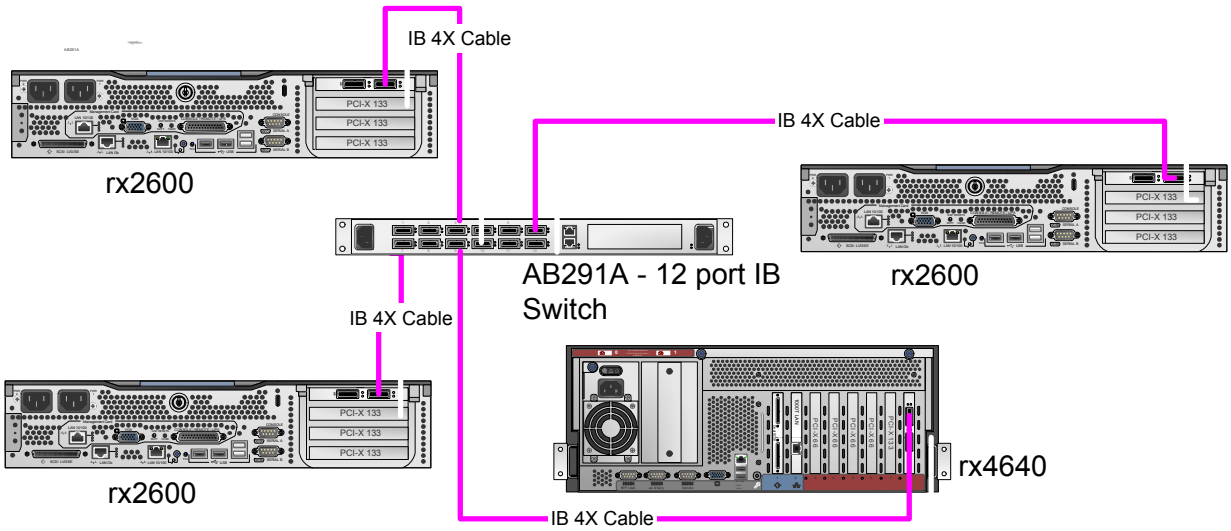
HP Fabric Clustering System – Switch Fabric

HP-UX Fabric Clustering System Interconnect solution can be used to build multi-node clusters using these switches. The AB291A – 12-port 4X Switch is used in testing the switch configurations. The following switch configurations were used for testing the performance characteristics:

- **2-node switch configuration**

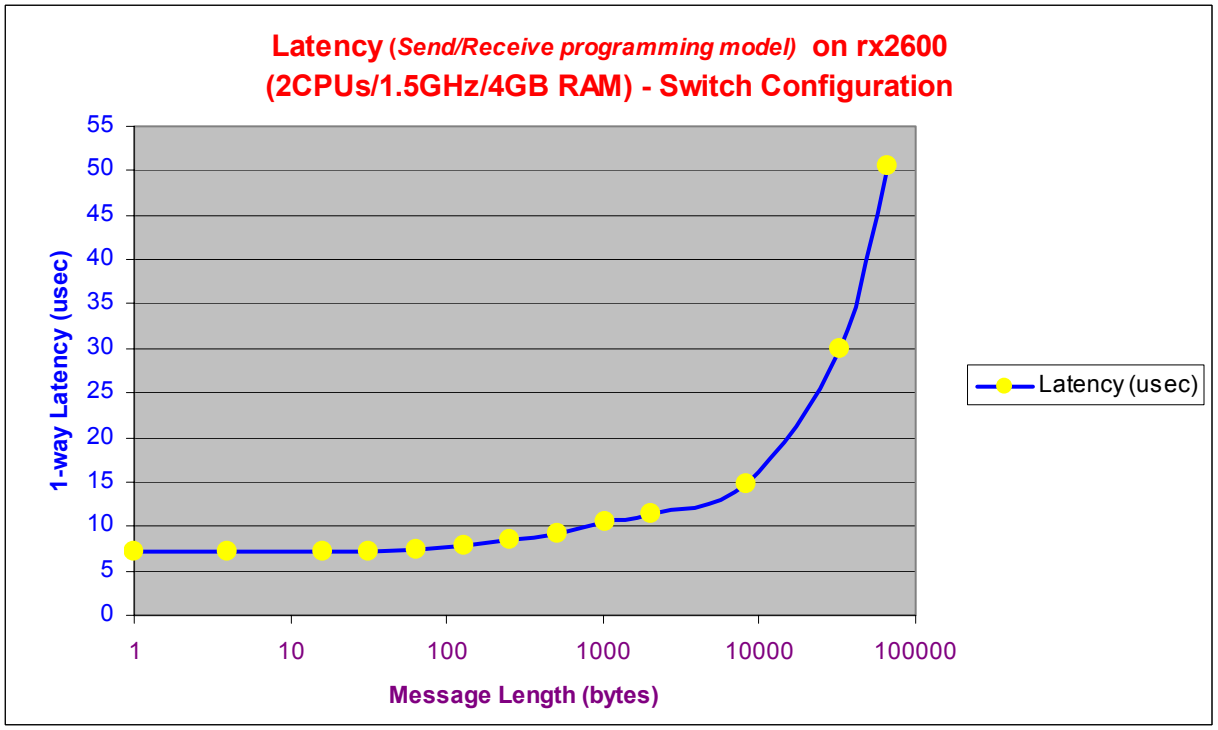


- **4-node (3 clients & 1 server) switch configuration**



Latency

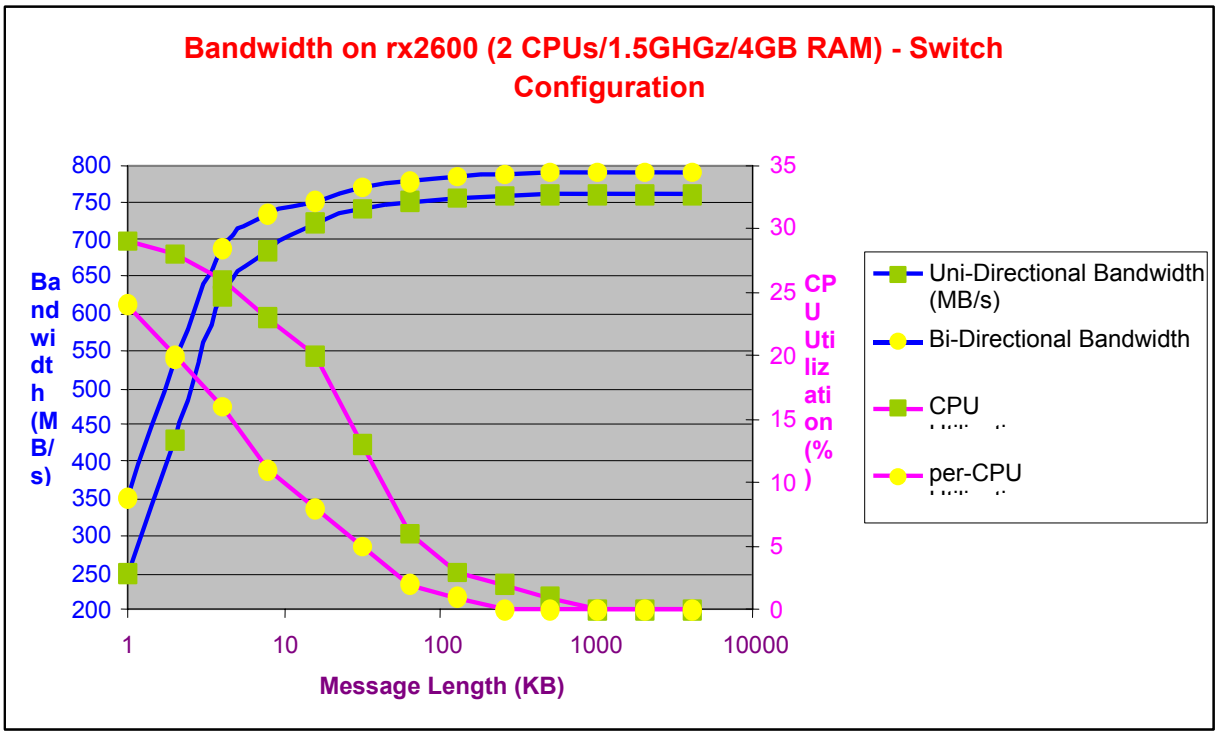
In switched configurations, the HP-UX solution offers a 1-way latency of 4.6usec for 8-byte messages using the RDMA programming model and 7.2usec for 1-byte messages using the Send/Receive programming model, on rx2600 node clusters. In switch configurations, the latency remains sub-10usec for messages of length 512 bytes.



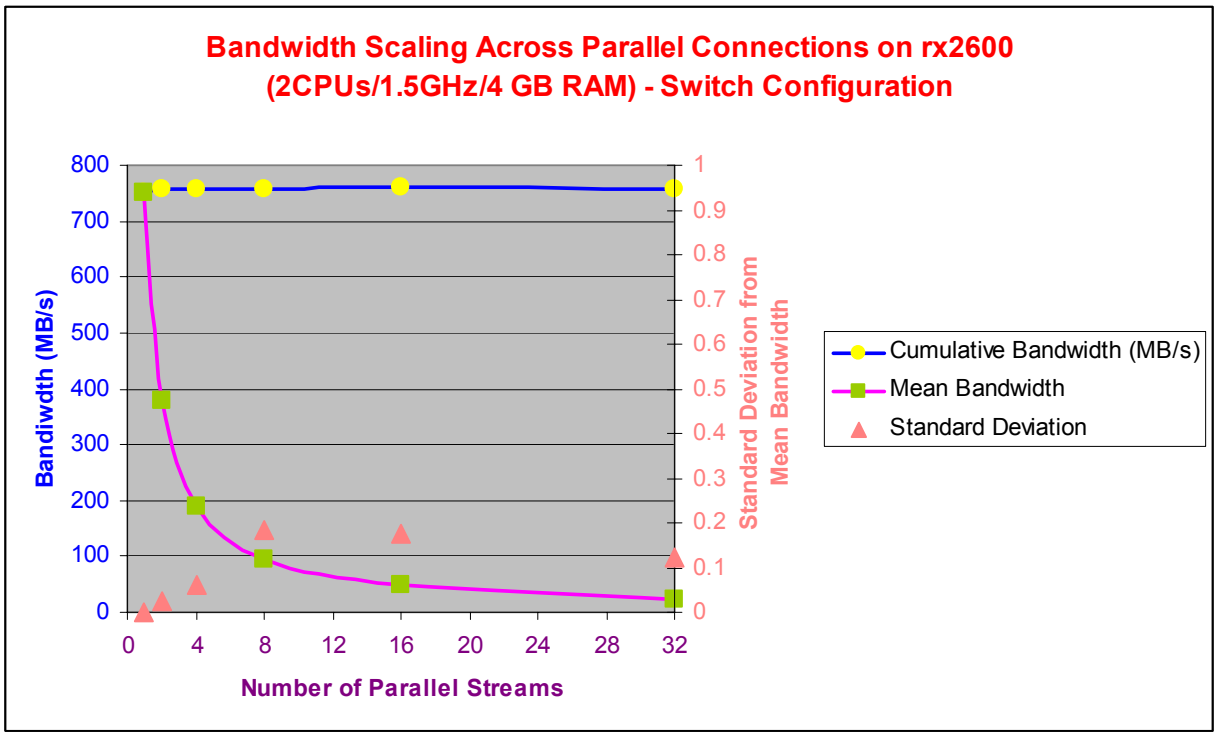
Bi-directional 1-way latencies on a rx2600 (2 CPUs/1.5GHz/4GB RAM) are 5.8usec for each stream using the RDMA programming model, and 7.2usec for each stream using the Send/Receive programming model. The bi-directional latencies for short messages, up to 512 bytes, remain pretty close to the respective uni-directional latencies. On a rx2600 (2 CPUs/1.5GHz/4GB RAM) in switched configuration, bi-directional 1-way latencies for 512-byte messages are 7.7usec for each stream using the RDMA programming model, and 9.30usec for each stream using the Send/Receive programming model.

Bandwidth

Uni-directional bandwidth on a rx2600 node cluster provides an application bandwidth of 761MB/s at 4MB message size with 0% service demand. Bi-directional bandwidth for the same message size is 790MB/s with 0% service demand on the server.

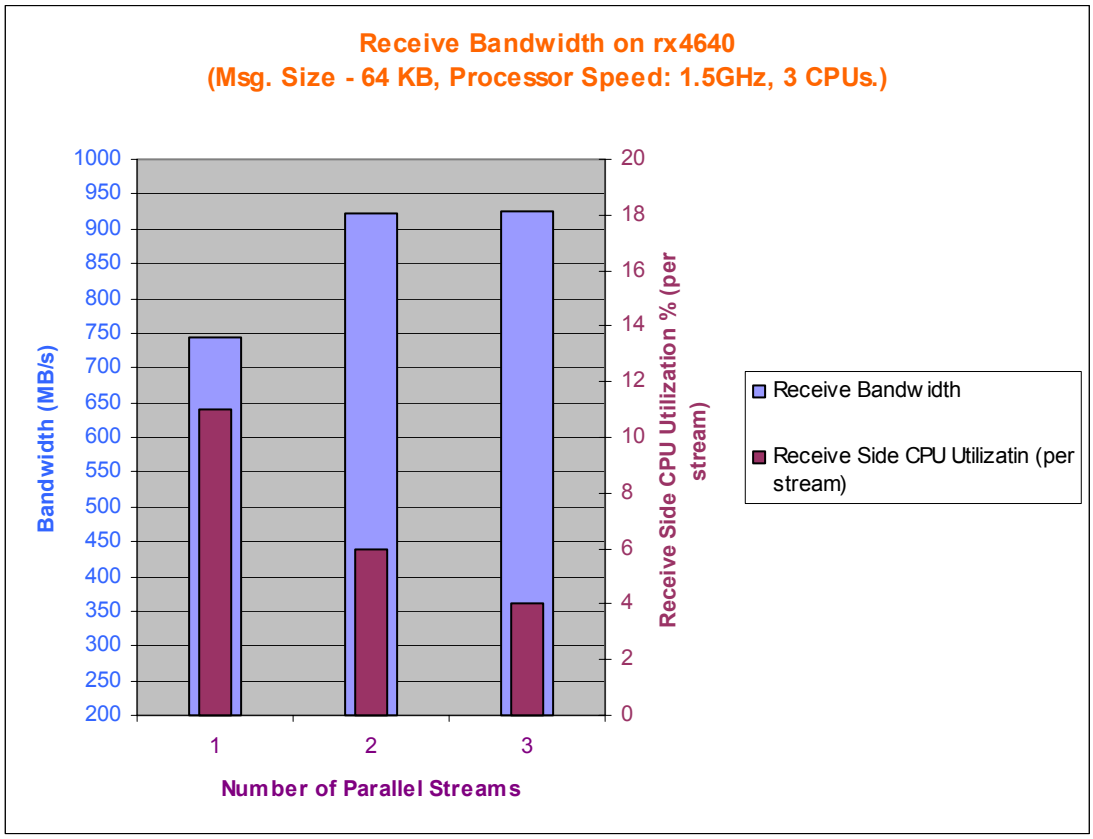


In a typical application environment, the HCA will be shared by more than a single connection. HP-UX cluster interconnect solution scales across multiple connections providing almost an equal share of the available bandwidth for all the parallel connections. On a rx2600 (2 CPUs/1.5Gz/4GB RAM) in a switch configuration, for 32 parallel bandwidth streams, the bandwidth offers a mean bandwidth of 23.67MB/s with a standard deviation of 0.123MB/s, resulting a cumulative bandwidth of 757MB/s.



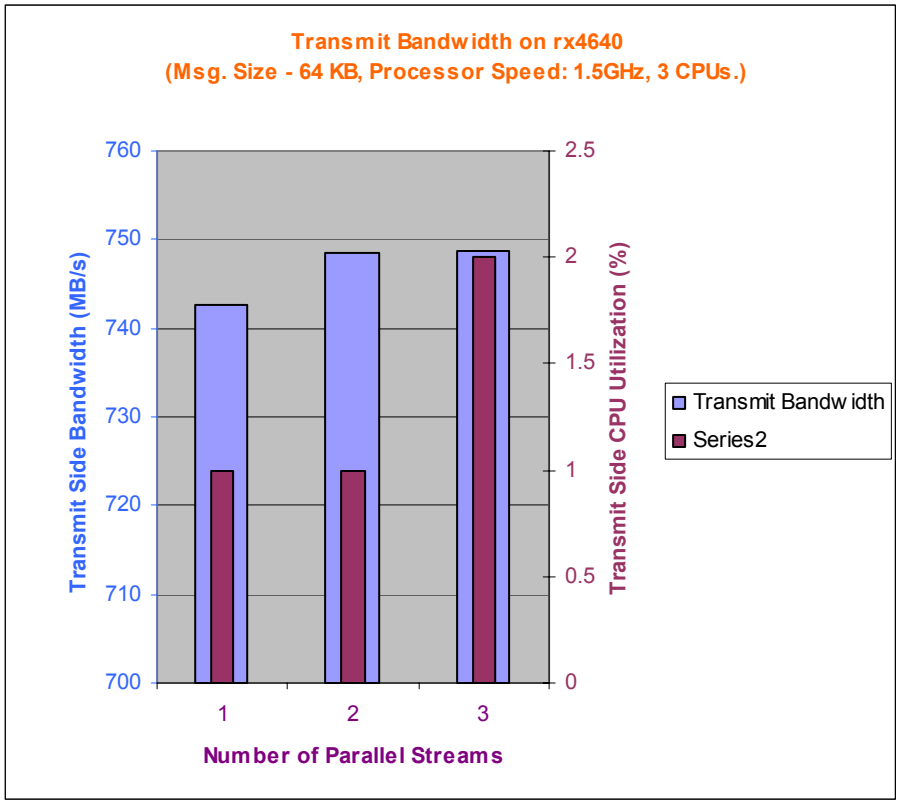
Receive Bandwidth

Three application streams can saturate the receive bandwidth offered by the HP-UX HCA. Cumulative receive bandwidth on a rx4640 (3 CPUs/1.5 GHz/2 GB RAM) for a message size of 64KB is 924.5MB/s; limited only by the limitations of the PCI-X point of attachment.



Transmit Bandwidth

On the transmit side, a single stream can drive the HCA to the limits of the interconnect solution. Using multiple streams will have only insignificant gains.



Non-Optimal Configurations

HCAs when plugged in single rope PCI-X slots on a rx2600 (2 CPUs/1.5 GHz/4GB RAM) in a point-to-point configuration offer 1-way latency of 7.8usec, as against 7.0usec latency on dual rope slots in a similar configuration. HP-UX clustering interconnect solution saturates the single rope PCI-X slot quite fast and can only offer a bandwidth of 452 MB/s for a 4MB message, as against 760MB/s bandwidth on dual ropes slots in a similar configuration.

HP Integrity server rx4640 has shared slots. Shared slots normally operate at 66MHz and can be lowered to 33MHz if a 33MHz card is plugged into the other shared slot on the server. Thus usage of shared slots in high performance oriented HP-UX Fabric Clustering System interconnect solution configurations is not recommended.

Summary

HP-UX Fabric Clustering System interconnect solution offers 4.6usec 1-way latency and a single stream bandwidth of 760MB/s, all using industry standard technologies and off-the-shelf components. The solution scales well across multiple HCAs as well as multiple connections enabling an overall improvement for real world applications. The HP-UX Fabric Clustering System interconnect solution supports both point-to-point and switch configurations without any negative impact on the application performance.

References

Refer to whitepapers on HP Fabric Clustering System found at www.hp.com.
www.hp.com/go/mpi
www.infinibandta.org