

SAND2005-1555P
Unlimited Release
March 2005

A Quick Look at InfiniBand: November 2004

Douglas Doerfler
Sandia National Laboratories

Introduction

The InfiniBand technology (IB) development community, e.g. Mellanox, Voltaire, Topspin, InfiniCon, etc. are deploying their next generation NICs and switches [1, 2, 3, 4, 5]. The NICs are based on the PCI Express bus and the switches employ a 24 port switch chip developed by Mellanox. This new switch chip has led to the development of high port count (e.g. 144 and 288 ports) standalone switches. Sandia has been acquiring equipment using this technology for its next generation Linux clusters. The purpose of this white paper is to take a quick look at the latest generation InfiniBand technology and contrast it with experience obtained with some other HPC technology. This is an initial "Quick Look" based on the recent acquisition of a small development cluster, named Escher, which contains the latest generation IB components.

This study looks at unidirectional and bidirectional characteristics between two nodes. Granted this is not representative of what an application may see, especially in a large job consisting of several nodes and several layers of switching. These characteristics will be examined in future studies. As the title implies, this is a quick look. This study also takes a look at cost of the components in systems of varying size.

PCI Express

PCI-Express is the latest in a long generation of internal interconnect I/O buses for printed circuit motherboards. It seems that PCI Express is going to be the technology that replaces the PCI-X bus for high-speed intra-node peripheral communications on computational processing boards. PCI and PCI-X will most likely remain to support legacy low performance peripherals, but all of the major HPC interconnect vendors will or will be providing PCI Express based NICs. The IB community is the first to deploy PCI Express based NICs. Quadrics and Myricom will most likely follow in the near future.

PCI Express abandons the traditional bus architecture for I/O buses, where individual peripheral chips sit on the same set of wires (the bus) and pull data when they decode that a transaction is meant for them. True buses have

performance limitations in that each device sitting on the bus provides some level of capacitance and hence limits the frequency at which signals can be driven. PCI Express is really not a bus in this sense. It uses two, one for each direction, point-to-point signals between chips. By limiting the transmitted signal to a single point-to-point connection, and by utilizing low voltage differential (LVDS) signaling, the frequency at which signals can be driven is much greater than that of a traditional bus. Multiple point-to-point signal pairs are used to increase the bandwidth. Multiple peripherals communicate with each other by utilizing a PCI Express switch.

A PCI Express Link is described by its width: x1, x2, x4, ... and x32. A Lane consists of point-to-point connections in each direction for a total of 2 differential pairs or 4 wires. An x1 Link consists of 1 Lane, an x2 Link has 2 Lanes, etc. The signals are nominally driven at 2.5 GHz for 2.5 Gbits/sec per Lane per direction, or 5.0 Gbits/sec aggregate. Note that one cycle of the signal gives you 1 bit, as opposed to a single cycle providing 2 bits by utilizing double data rate (DDR) clock sampling techniques. Each lane uses 8b/10b encoding/decoding logic. Hence, it takes 10 bits per byte of data. PCI Express peak bandwidth for various Link widths is provided in Table 1. It's beyond the scope of this paper to discuss all the attributes of PCI Express. A reference such as [6] should be consulted for further details.

Table 1: PCI Express Peak Aggregate Bandwidth as a Function of Link Width

Link Width	x1	x2	x4	x8	x12	x16	x32
Aggregate Bandwidth (GBytes/sec)	0.5	1	2	4	6	8	16

In a PCI Express Link, there are more pins per device than just those used for signal lines. The Link also requires power and ground pins. For example, an x8 Link has 40 pins per chip, which includes 32 signal pins (8 pairs per direction) and 8 pins for power and ground. A clock signal is not used. The data signal symbols always have a signal level transitions which allows a PLL at the receive end to regenerate the clock.

InfiniBand

InfiniBand at the physical layer is similar to PCI Express, but differs in significantly in its implementation. Of course, the upper level protocols (hard and soft) are very different. But the terminology used at the physical layer is similar and it does tend to cause confusion. Like PCI Express, IB uses differential signal pairs, one in each direction, to create links. The IB links also use LVDS signaling. However, IB uses a 1.25 GHz signal and generates two bits per cycle using DDR edge triggering. That is, the data signals are sampled on the rising and falling edge of the clock. Like PCI Express, the receive side clock is regenerated from

the transmitted signal using a PLL. Data sent over IB links are encoded using 8b/10b encoding. IB links have other signals, which are used for management and control of the link.

The width of an IB link is described by the number of “lanes” it has: 1x, 4x and 12x. Each lane consists of transmit and receive of wire pairs. Note that IB uses the “x” as a suffix and PCI Express uses the “x” as a prefix! The peak bandwidth for various IB link widths is provided in Table 2. It’s beyond the scope of this paper to discuss all the attributes of InfiniBand. A reference such as [7] should be consulted for further details.

Table 2: InfiniBand Peak Aggregate Bandwidth as a Function of Link Width

Link Width	1x	4x	12x
Aggregate Bandwidth (GBytes/sec)	0.5	2	6

Note that the peak aggregate bandwidths for PCI Express and InfiniBand are matched, based on the “x” terminology. This is convenient when looking at PCI Express/InfiniBand NICs. This will be illustrated in the description of the test platform provided below.

Test Platforms

The InfiniBand platform used for this study, code named Escher, is a 12 node cluster acquired from Verari Systems. Nodes are based on the Supermicro X6DAE-G2 motherboard, which utilizes dual Intel Xeon EM64T (Nocona) processors and the Intel E7525 (Tumwater) chipset [8]. The Tumwater chipset has an x16 PCI Express interface, which is brought out by the Supermicro motherboard to an expansion slot. The nodes run SuSE Linux Professional, version 9.1, and version 2.6.4-52-smp of the Linux kernel. Benchmark codes were compiled using the SuSE bundled Gnu compiler suite, gcc version 3.3.3.

The InfiniBand hardware is from Voltaire. The NIC is the HCA 400 with an x8 PCI Express interface, which is plugged into the x16 slot on the motherboard. This NIC provides two 4x IB ports. In theory, the x8 PCI Express interface is balanced with the two 4x IB ports. That is, the 4 GBytes/sec peak at the host I/O interface and $2 \times 2 = 4$ GBytes/sec peak at the IB side of the NIC provides balance between PCI Express and IB data rates. However, at the time of this study the Voltaire system software stack is unable to utilize one of the 4x IB ports. Hence the peak IB bandwidth achieved with the configuration is 1.0 GBytes/sec per direction, or 2.0 GBytes/sec aggregate.

The twelve nodes are InfiniBand interconnected via a Voltaire ISR 9024 switch. This switch has 24, 4x IB ports. The test platform uses 12 of the ports, one 4x IB connection per NIC and one NIC per node.

The results of the Escher platform are contrasted with the results obtained from the Red Squall and Vplant/Callisto platforms [9, 10]. Their key attributes relative to this study are summarized in Table 3.

Table 3: Test Platform Comparison

	Escher (4x InfiniBand w/PCI-Express)	Red Squall (Elan4 w/PCI-X)	Callisto (4x InfiniBand w/PCI-X)
Link Peak BW (Aggregate)	2 GB/sec	2.133 GB/sec	2 GB/sec
Host Interface Peak BW	4 GB/sec	1.064 GB/sec	1.064 GB/sec
Host Processor	dual 3.4 GHz Xeon EM64T	dual 2.0 GHz Opteron	dual 3.06 GHz Xeon
Memory Subsystem	dual DDR2-400 (2 x 3.2 GB/sec)	dual DDR-333 (2 x 2.67 GB/sec)	dual DDR-266 (2 x 2.13 GB/sec)

Pallas MPI Benchmark Suite (PMB)

Pallas, before being purchased by Intel, provided a very nice MPI benchmark suite. It “was” freely available for download. This benchmark suite has been used in past studies and it provides a nice suite of routines to measure the performance of several point-to-point and collective MPI routines. It is used here to maintain consistency with previous studies. The PMB suite (version 2.2.1) reports bandwidth results as 1 MB = 2²⁰ bytes. For this study, the bandwidth results were converted to 1 MB = 10⁶ bytes, which is more consistent with communication theory practice.

PMB PingPong

PingPong measures the time it takes for a message to travel to a remote node and return back to the sender. It uses the MPI_Send() and MPI_Recv() calls. The reported time is round trip time divided by two. The bandwidth is based on the amount of data sent per transfer, the unidirectional bandwidth. The test was executed on two nodes.

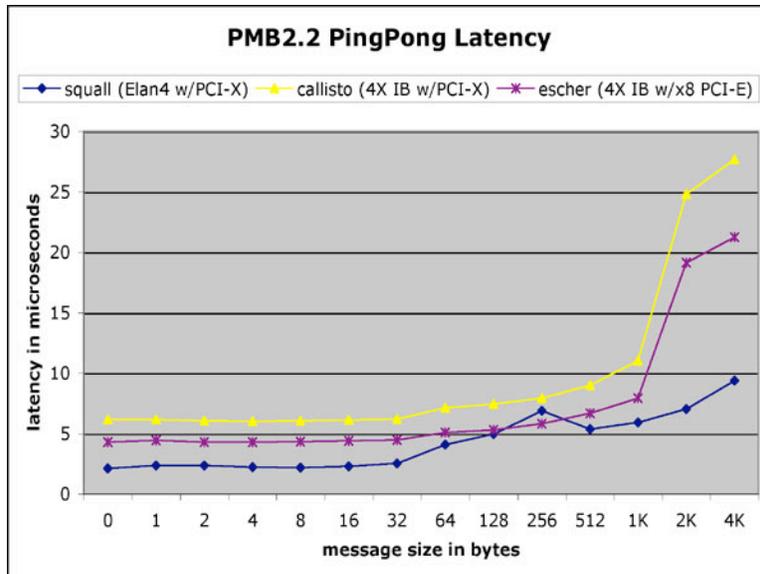


Figure 1: PingPong Latency

Table 4: PingPong Latency Results (time in microseconds)

Msg size	Escher	Red Squall	Callisto
0	4.29	2.1	6.17
1	4.42	2.36	6.15
2	4.3	2.36	6.07
4	4.3	2.2	6.02
8	4.31	2.19	6.07
16	4.39	2.28	6.11
32	4.45	2.51	6.21
64	5.09	4.1	7.14
128	5.3	4.96	7.43
256	5.83	6.89	7.94
512	6.67	5.36	8.99
1K	7.93	5.9	11.04
2K	19.14	7.04	24.8
4K	21.26	9.37	27.68

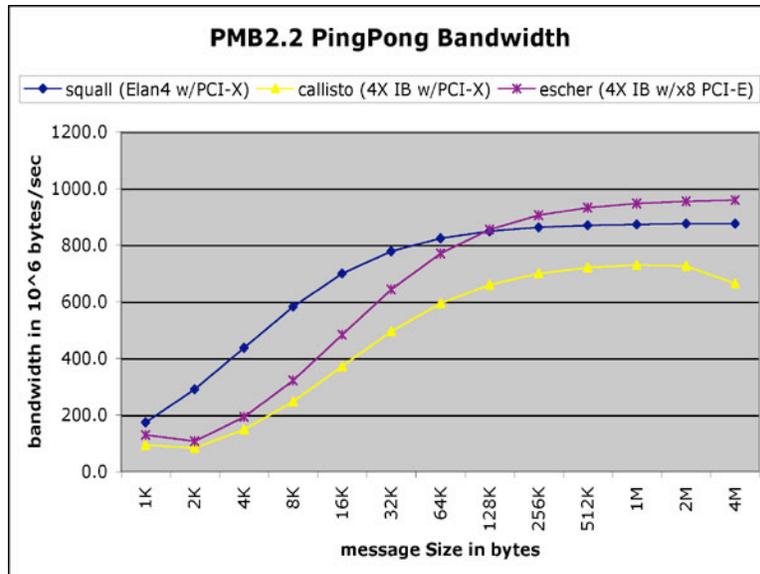


Figure 2: PingPong Bandwidth

Table 5: PingPong Bandwidth Results (10^6 bytes per second)

Msg size	Escher	Red Squall	Callisto
1K	129.0	173.5	92.7
2K	107.0	291.0	82.6
4K	192.7	437.2	148.0
8K	321.9	582.5	247.7
16K	482.6	698.6	371.1
32K	642.7	778.5	494.7
64K	770.5	823.6	594.1
128K	855.4	848.3	659.9
256K	905.1	862.1	698.8
512K	932.4	869.4	720.1
1M	946.9	873.2	728.9
2M	954.1	875.0	725.6
4M	958.5	875.9	663.4

PMB SendRecv

The SendRecv benchmark measures bidirectional bandwidth between nodes. It uses the MPI_Sendrecv() call. The time reported is the average time for all iterations of the test. The bandwidth is the amount of data transferred times two, once for each direction. The results are for a job of size two nodes.

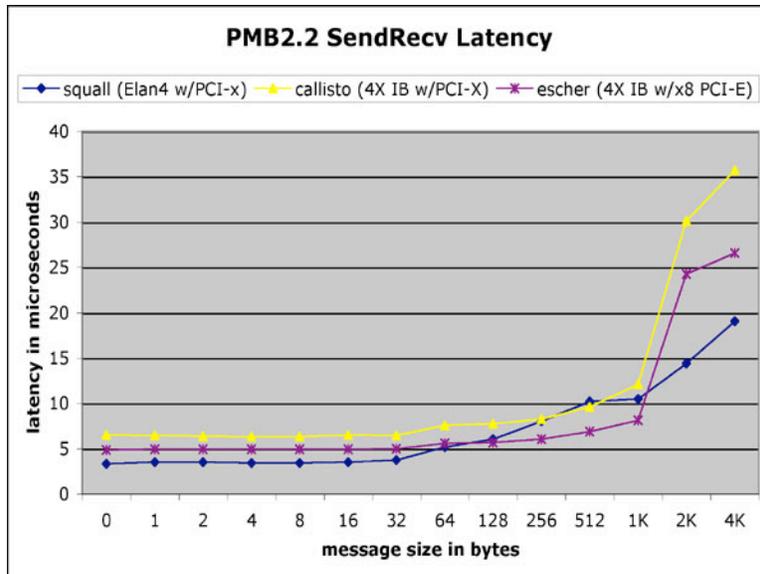


Figure 3: SendRecv Latency

Table 6: SendRecv Latency Results (time in microseconds)

Msg size	Escher	Red Squall	Callisto
0	4.86	3.31	6.49
1	4.92	3.52	6.46
2	4.93	3.52	6.35
4	4.93	3.42	6.31
8	4.92	3.4	6.33
16	4.94	3.5	6.52
32	4.96	3.72	6.45
64	5.57	5.19	7.55
128	5.69	6.05	7.75
256	6.05	8.04	8.25
512	6.87	10.18	9.59
1K	8.14	10.46	12.08
2K	24.26	14.38	30.14
4K	26.58	19.05	35.7

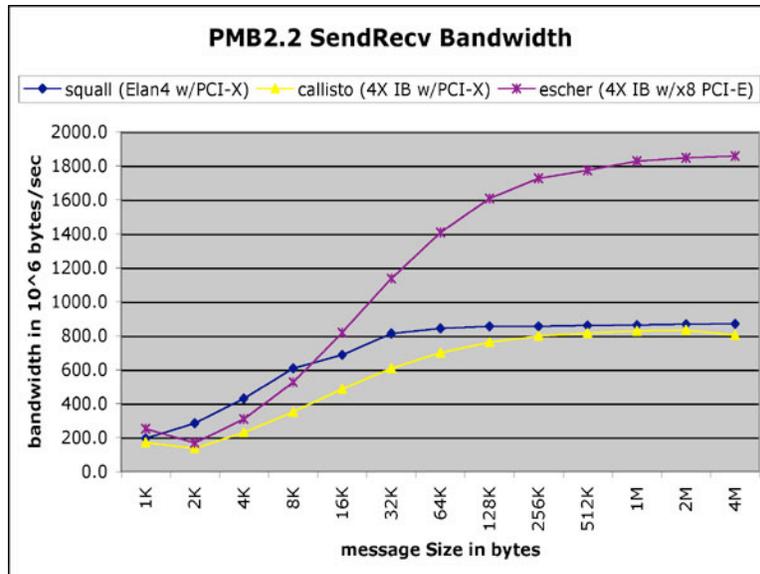


Figure 4: SendRecv Bandwidth

Table 7: SendRecv Bandwidth Results (10^6 bytes per second)

Msg size	Escher	Red Squall	Callisto
1K	251.5	195.8	169.4
2K	168.8	284.9	135.9
4K	308.2	430.0	229.5
8K	527.1	607.6	351.7
16K	818.3	686.9	485.4
32K	1134.9	812.4	608.4
64K	1408.1	841.5	699.7
128K	1606.5	853.3	761.1
256K	1726.5	855.5	798.7
512K	1774.1	859.8	816.0
1M	1827.8	860.9	827.8
2M	1846.0	865.8	831.0
4M	1857.3	868.4	803.7

PMB Conclusions

The PingPong latency of the IB network is 4.29 microseconds, which is almost 2 microseconds less than the previous generation IB network and it is over 2 microseconds more than the latency obtained by the Elan4 network. The excellent performance of the Elan4 based Red Squall cluster is partly attributed to the excellent memory subsystem performance of the Opteron. The discontinuities illustrated in the plots at 512 bytes for Elan4 and 2K bytes for IB is due to the transition from a short message protocol to a large message protocol in the software stacks.

The PingPong bandwidth and the SendRecv bandwidth results of Escher's IB interconnect are impressive. Unidirectional bandwidth is 958.5 MB/sec, or 95.8%

of the link's peak. Bidirectional bandwidth is 1,857.3 MB/sec, or 92.9% of the links peak. Callisto's IB results are 728.9 MB/sec, or 72.9% of peak, and 831 MB/sec, or 41.6% of peak, for unidirectional and bidirectional bandwidth respectively. Red Squall's results are 875.9 MB/sec, or 82.1% of peak, and 868.4 MB/sec, or 40.7% of peak, respectively.

Table 8: Test Platform Comparison

	Escher (4x InfiniBand w/PCI-Express)	Red Squall (Elan4 w/PCI-X)	Callisto (4x InfiniBand w/PCI-X)
Maximum Unidirectional BW / % of peak	958.5 / 95.8%	875.9 / 82.1%	728.9 / 72.9%
Maximum Bidirectional BW / % of peak	1,857.3 / 92.9%	868.4 / 40.7%	831 / 41.6%

Escher's IB results demonstrate the potential improvements that are available by replacing PCI-X with PCI Express as the host interface.

It should be noted that the Elan4 network's bandwidth curves "ramp up" quicker than the IB networks. So if an application is moving messages in that message size range, e.g. ~1K bytes to ~128K bytes for unidirectional messaging, the Elan4 network may provide better performance. Even though its "peak" bandwidth is lower. For a SendRecv operation, the range is much narrower, ~1K bytes to ~8K bytes.

Pricing Analysis

Using recently published [11, 12] list pricing schedules, the list price per port has been calculated. This includes the cost of the NICs, all levels of switching, and cables. It does not include items such as racks, power distribution units, blanking panels, maintenance, etc. All of which can become significant costs when looking at total cost of ownership.

Although looking at list pricing is helpful, it does not really reflect what the actual cost would be in an acquisition. For a large acquisition, discounts would be applied and depending on the vendor and timing the discounts could be significant.

For the pricing analysis, the IB network was configured using only 24 port switches for small systems, and 24 port switches at the node level with 288 port switches at the upper level for large systems. All network configurations, including Quadrics, utilize a "half-bandwidth" connection architecture to the upper level of switching in a fat-tree network. This is appropriate for most "capacity"

type deployments. Capability deployments would require a full-bandwidth connection architecture to the upper layers of the network.

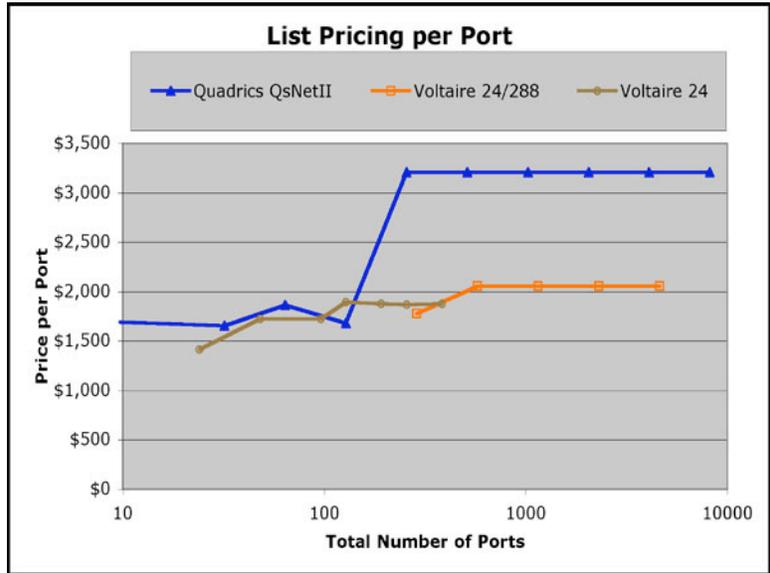


Figure 5: List Pricing as a Function of Port Count

Table 9: Pricing for IB using 288 & 24 Port Switches

Network Size	List Price	Price per Port
4608	\$9,473,840	\$2,056
2304	\$4,736,920	\$2,056
1152	\$2,368,460	\$2,056
576	\$1,184,230	\$2,056
288	\$511,390	\$1,776

Table 10: Pricing for IB Using 24 Port Switches

Network Size	List Price	Price per Port
384	\$720,320	\$1,876
256	\$477,730	\$1,866
192	\$360,160	\$1,876
128	\$242,590	\$1,895
96	\$165,680	\$1,726
48	\$82,840	\$1,726
24	\$33,970	\$1,415

Table 11: Pricing for Elan4

Network Size	List Price	Price per Port
8192	\$26,260,128	\$3,206
4096	\$13,130,064	\$3,206
2048	\$6,565,032	\$3,206

1024	\$3,282,516	\$3,206
512	\$1,641,258	\$3,206
256	\$821,529	\$3,209
128	\$214,851	\$1,679
64	\$119,177	\$1,862
32	\$52,961	\$1,655
8	\$13,582	\$1,698

References

1. <http://www.mellanox.com>
2. <http://www.voltaire.com>
3. <http://www.topspin.com>
4. <http://www.infinicon.com>
5. <http://www.quadrics.com>
6. Ravi Budruk, Don Anderson, Tom Shanley, *PCI Express System Architecture*, MindShare, Inc., Addison-Wesley, 2003.
7. Tom Shanley, *InfiniBand Network Architecture*, MindShare, Inc., Addison-Wesley, 2003.
8. *Super X6DAE-G2 User's Manual*, Super Micro Computer Inc., 2004.
9. Doerfler, Clauser, Laguna, Lawry, Maestas and Simonds, "A Comparison of 4X InfiniBand and Quadrics QsNetII Technologies", Internal Publication, Sandia National Laboratories, Draft 1.1, March 17, 2004.
10. Brightwell, Doerfler and Underwood, "A Comparison of 4X InfiniBand and Quadrics Elan-4 Technologies", Proceedings of the 2004 IEEE International Conference on Cluster Computing, September 20-23, 2004, San Diego, CA.
11. "Voltaire U.S. Price List", Rev 3.3, Sept. 2004, Voltaire, Inc., Bedford, MA.
12. "Quadrics QsNetII – Price List", June 2004, Quadrics, Ltd., Bristol, UK.