



Cluster used for Evaluation



Jharrod LaFon wires InfiniBand connections.

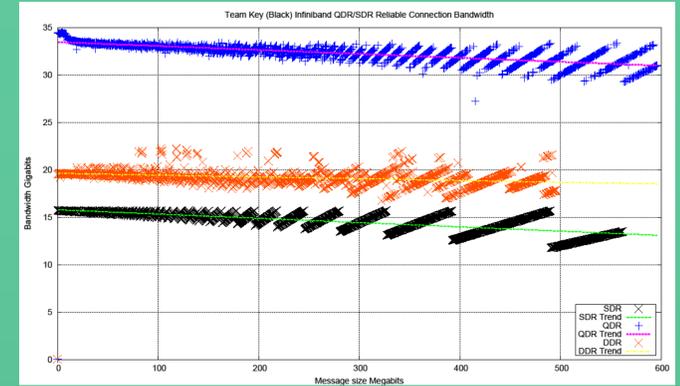


Troubleshooting kernel modules.

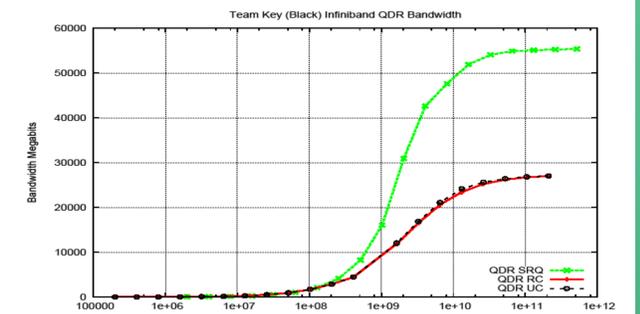
Summary

In this study, we examined the various abilities and performance metrics of the new Mellanox Connect X QDR InfiniBand Interconnect cards. These are designed to increase the speed at which nodes can talk to each other in a High Performance Computing environment. We evaluated a variety of aspects, including the performance as compared to previous generations of InfiniBand interconnects. This allows us to make a logical decision on the cost effectiveness, and capabilities of the next generation of high speed cluster interconnects.

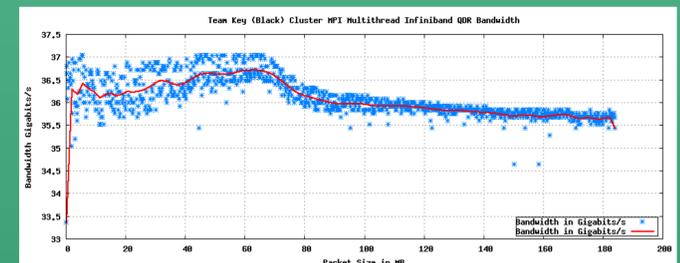
QDR/DDR/SDR Comparison



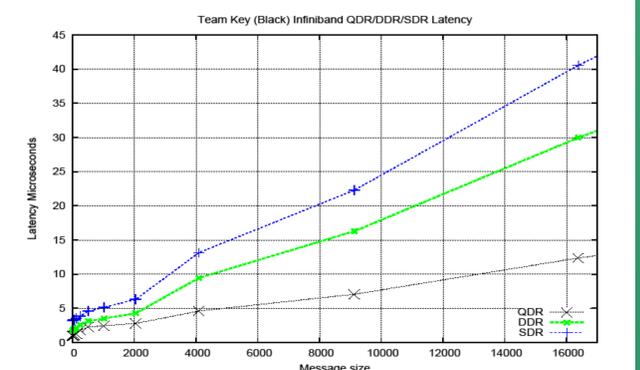
QDR Protocol Performance



Bandwidth and Message Size



QDR/DDR/SDR Latency



IPOIB vs. Verbs API

- Bandwidth Maximums (Bidirectional)
 - IPOIB \approx 32 Gigabit/s
 - VAPI \approx 55 Gigabits/s
- Latency Minimums
 - IPOIB 50.0 μ s (optical)
 - VAPI 1.5 μ s (optical)
- Ease of use
 - VAPI compilation
- Internet protocol over InfiniBand (IPOIB)
 - Uses the default kernel and existing TCP/IP standards
 - Requires more overhead
- Verbs API
 - Bypass CPU
 - Uses Kernel Module
 - Allow RDMA (Remote Direct Memory Access)

Switched vs. B2B

- Bandwidth (Bidirectional)
 - Switched \approx 55.34 Gigabits/s (SRQ copper) Max
 - Back to Back \approx 55.42 Gigabits/s (SRQ copper) Max
- Latency
 - Switched: 0.99 μ s (Copper) Min
 - Back to Back: 0.86 μ s (copper) Min
- Reliability/Management
 - The switch requires a management module, which was unavailable during the time that we were testing, therefore we can not evaluate the management features of the MTS3600
 - The switch intelligently chooses the highest signal rate possible for each individual port

QDR/DDR/SDR Comparison

- Bandwidth (Max, Copper, Switched)
 - QDR: 55.34 Gigabits/s
 - DDR: 23.22 Gigabits/s
 - SDR: 15.61 Gigabits/s
- Latency (Min, Copper, Switched)
 - QDR: 0.99 μ s
 - DDR: 1.77 μ s
 - SDR: 3.14 μ s
- All 3 signal rates exhibit step-like functions in bandwidth when message size is varied
- SDR Measurements were the most consistent, DDR and QDR were somewhat sporadic, competing for hardware resources.
- Tests were completed with legacy hardware, meeting or exceeding system requirements.

Copper vs. Fiber

- Bandwidth (Bidirectional, Back to Back)
 - Copper: 55.42 Gigabits/s (SRQ QDR Max)
 - Fiber: 55.41 Gigabits/s (SRQ QDR Max)
- Latency (Min, B2B)
 - Copper: 0.86 μ s
 - Fiber: 0.95 μ s
- Reliability
 - Fiber Cables contain their own firmware, which can effect performance and compatibility with some platforms.
 - Copper has a limited cable length of approx. 10m
 - Fiber converts the signal from electric, to optical, and back for each cable.
 - The lasers and receivers are part of the InfiniBand cable

Procedure

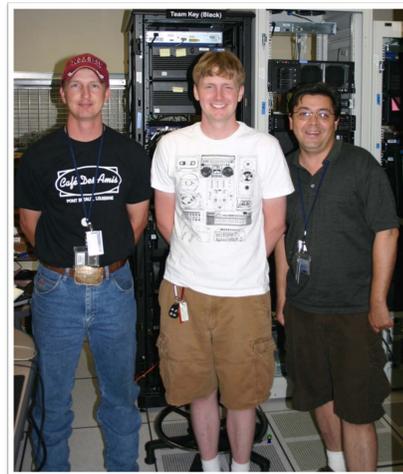
After building a cluster, we installed Mellanox Connect X QDR InfiniBand PCI express 8x 2.0 network interface cards. These were then used in conjunction with the Mellanox MTS3600 36 Port 40Gb/s InfiniBand Switch System to analyze and test the performance of QDR data rate InfiniBand interconnects.

We found that the bandwidth of the interconnect varied greatly with the type of test used, but that the switch, and type of cable made a small, almost inconsequential difference.

Since the cards are cutting-edge, the current kernel for Centos 5.3 was not optimal for the maximum performance. Furthermore the OFED (Open Fabrics Enterprise Distribution) stack that supported the cards was still in alpha testing, and therefore not stable.

A large performance gap exists between Internet Protocol over InfiniBand (IPOIB) and the Remote Direct Memory Access (RDMA) supported by the cards. At the rates of transfer that is expected, even the memory bandwidth may have been limiting the throughput.

Furthermore, the packet size and MTU make a significant impact on the throughput and Latency. For messages much larger than 1MB bandwidth is constant, but under that there is a notable drop when using very small message sizes. The highest bi-directional bandwidth measured was approximately 55Gigibits/second, and the lowest latency measured was about 1 micro-second. The N/2 value was about 7.5 kilobytes.



Team Key (Black) above. From left to right: Jharrod LaFon, Ben Haynes, John Herrera. Photographer: Dane Gardner.

Reliable vs. Unreliable Data Transport

Both IPOIB and Verbs API (using RDMA) have protocols for both reliable and unreliable data transport. The advantage to the unreliable modes, unreliable connection (UC), unreliable datagram (UD) and Unreliable Datagram Protocol (UDP), is that they are capable of sending packets at a much higher rate than the reliable modes. The reliable modes, reliable connection (RC), reliable datagram (RD), and Transport Control Protocol (TCP), protect the data and make sure that the packets get where they need to be.

When using UD communications, we were able achieve a send rate of over 30Gb/s, yet our receive rate was only 25 Gigabits. Alternatively our reliable data range was approximately 22 Gigabits.



Left: Mellanox MTS3600 Switch. Right: Infiniband QDR cable. Photographer: Dane Gardner

