Evaluation of InfiniBand Range Extension offered by Obsidian

Hussein N. El-Harake, Chris Gamboni, Stefano Gorini and Thomas Schoenemeyer
Swiss National Supercomputing Centre (CSCS), Manno, Switzerland
hussein@cscs.ch, gamboni@cscs.ch, gorini@cscs.ch, schoenemeyer@cscs.ch

Abstract – We evaluated the Obsidian Longbow InfiniBand Range Extender with the overall goal to ensure continuous availability of GPFS through the complete CSCS relocation period by running one single GPFS file system over both sites.

The geographical distance between the current and the future location is about 3 km, the measured distance of dark fiber is 10km.

The evaluation results for the range extender are encouraging and are in line with our expectations and requirements.

1. Introduction

InfiniBand Fabrics are designed to serve the needs of HPC computing and HPC storage. The increase in distribution of computational workloads to sites in different locations and also the ongoing trend for infrastructure consolidation at large research sites creates the necessity to investigate in alternatives to Fiber Channel or iSCSI over Ethernet in order to realize high performance network connections over wide distances.

Three application areas should benefit from this solution: Storage Area Networks by enabling inter-building SANs, Cluster Aggregation and remote visualization.

Various solutions exist in the market for High-Performance Connectivity with InfiniBand over the WAN. As an example, InfiniBand Range Extensions are supplied by Obsidian, Bay Microsystem and Net.com. Net.com offers NX5010, where performance is limited to the use of TCP/IP protocols. [1]

Bay Microsystems provides the IBEx G40/M40 InfiniBand Extension Switch which utilizes proprietary packet and transport-processing technology and allows the IBEx G40 device to seamlessly transport native InfiniBand over most wide-area network technologies. [2]

Obsidian proposes an interesting technology that uses solely optical fiber and has been evaluated by Xiangyong Ouyang et al. [3]. Since proprietary optical encoding is deployed, the devices must be deployed in pairs, connected by direct light paths [4, 5]. When installed, the Longbows appear and behave like native InfiniBand switches.

The motivation for the investigation of this technique at CSCS is based on the center's requirements during its relocation in early 2012. This technology is a potential solution to keep our central parallel file system (GPFS) up and running throughout the relocation period from March 2012 to June 2012, when CSCS will move from its current location in Manno to Lugano.

2. Technique and Evaluation Procedure

Obsidian has been providing the Longbow router to support IB-WAN routing since 2007. This technology allows an InfiniBand fabric to be extended via optical fiber over varying distances, so that two distant InfiniBand clusters can be aggregated. This

InfiniBand router also provides hardware encryption mechanisms to ensure data transfer security.

Since the distance between the old and the new datacenter is less than 10km, we decided to implement the Obsidian Longbows C103 with two **C**oarse **W**avelength **D**ivision **M**ultiplex (CWDM) devices. This equipment is capable of bi-directionally interfacing up to 8 CWDM wavelengths (plus a standard 1310nm channel) with a single fiber pair, suitable for driving up to 90Gbits/s of full-duplex InfiniBand traffic over tens of kilometers.

Our approach is to run one single GPFS parallel file system over two locations with a distance of a few kilometers using a 10 km InfiniBand link.

- *Rivera3, rivera4, rivera5 and rivera6* : I/O dual-socket nodes with Intel Xeon E5649 2.53GHz (Westmere), 48GB DDD3 memory 1333MHz and ConnectX 2 dual port QDR HCA
- *Rivera1 and rivera2*: I/O dual-socket nodes with AMD Opteron 8-core 2.0 Ghz (Magny-Cours), 16 GB DDR2 memory, ConnectX2 dual port QDR
- Two Mellanox QDR switches, one with 36 ports and one with 8 Ports.
- Obsidian C103 solution, consists of 8 SDR boxes and 2 X CWDMs
- 10km of fiber cable in one medium sized box.
- NetApp E5400 with 60 SAS NL disks with 2 TB each (in "Lugano")
- IBM DS5300 with 60 SATA disks with 2 TB each (in "Manno")

In order to obtain reference results for a local InfiniBand fabric, we connected four I/O servers to the 36-port Mellanox QDR switch. After all performance numbers in the local fabric had been measured, we built the test setup as shown in Figure 1. We connected two I/O servers to each InfiniBand Switch, both switches are connected through the Obsidian devices and the 10km fiber spool. Each CWDM box connects to four SDR-devices to get an aggregated bandwidth of 40Gb/s.

To evaluate the characteristics, we measured at three different levels:

- RDMA bandwidth, latency and message rates, using the OFED and OSU benchmark suite.
- TCP bandwidth with ***Iperf***
- Bandwidth and latencies within GPFS and GPFS NSDs migration

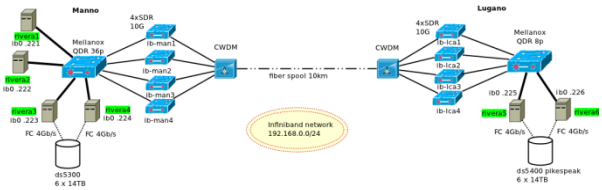

Figure 1: Test Setup to simulate a GPFS over WAN environment. The left-hand side represents "Manno" and the right-hand side represents "Lugano". "Lugano" is not yet physically existent, just simulated by a 10km fiber spool.

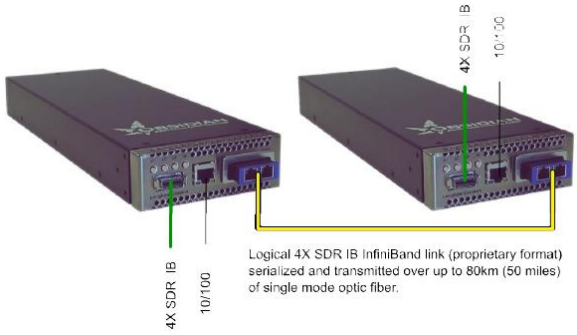

Figure 2: Obsidian C103 series

3. Results

a. RDMA performance with OFED

Figures 3 and 4 show the results for the RDMA tests, ***ib_write_bw*** and ***ib_read_bw*** for the reference setup measured on a local fabric (figure 3) and the test setup (figure 4). The bandwidth within the local fabric

yields results as expected for the QDR network of nearly 3.3GB/s. The test setup provides close to 950MB/s which is the peak of one SDR link. That is consistent with our expectations since we are running one to one tests where only one of the 4 SDR links is used.

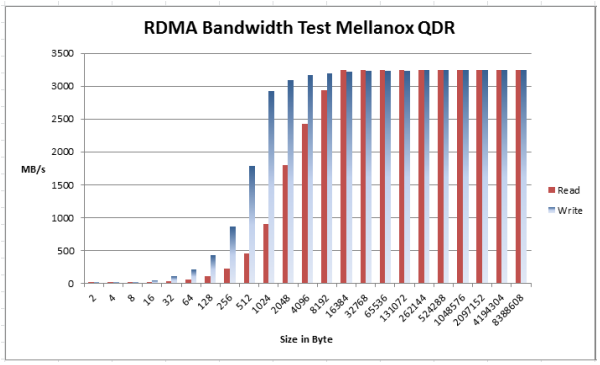

Figure 3: Reference results for RDMA Bandwidth – local IB

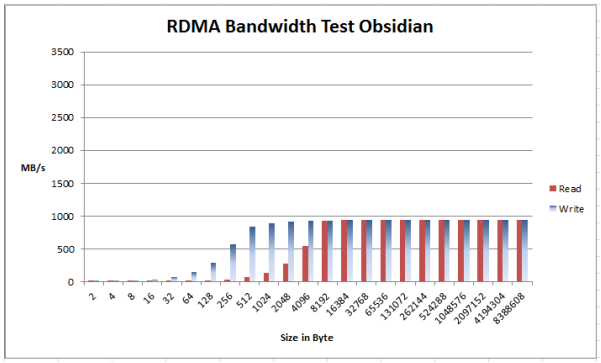

Figure 4: Test setup results for RDMA Bandwidth – over 10km Fiber

With figures 5 and 6, the latencies for the local and the remote setup can be compared. Calculating the minimum latency for the 10km fiber gives a lower boundary of 33µs for that distance. For small messages between 2 Byte and 32Kbyte we observed a latency of 50 to 100µs.

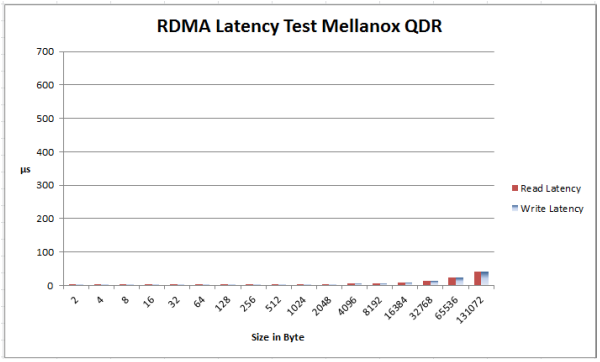

Figure 5: Reference Results for a local IB-Fabric for Latency

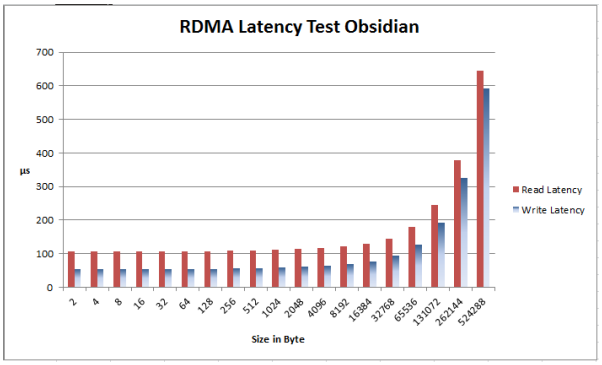

Figure 6: Latency results for the Test Setup over 10km fiber

We repeated the bandwidth and latency tests with the OSU benchmark. Figures 7 to 10 summarize all the results. Performance numbers and characteristics are very similar to those achieved with the OFED benchmark test.

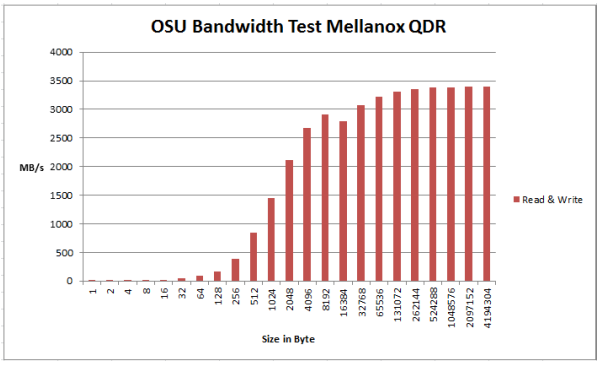

Figure 7: Reference Results for OSU – local InfiniBand

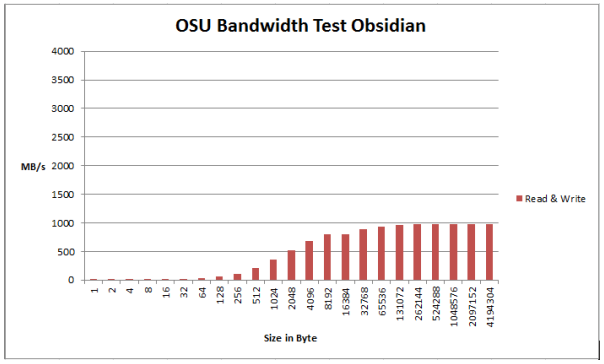

Figure 8: Test Setup Results – with 10 km fiber

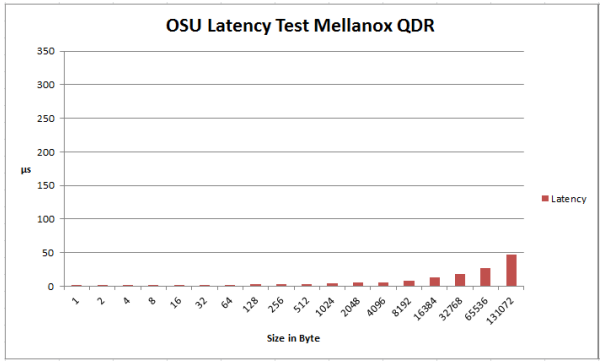

Figure 9: Reference Results for OSU latency – local InfiniBand

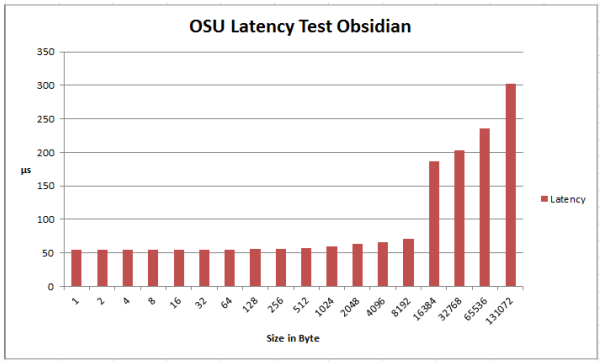

Figure 10: Results for Test Setup – with 10 km fiber

In addition, we looked at the message rates which, in the reference setup achieve up to 52MMr/s for block sizes between 1Byte and 2K Byte (figure 11). For the test setup we measured up to 3.5MMr/s (figure 12), when using one link. If 4 SDR links are used simultaneously, we expect to obtain 14MMr/s.

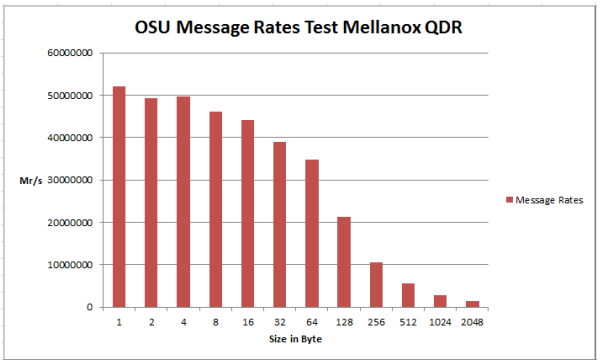

Figure 11: Reference Results for Message Rates, small block sizes – local InfiniBand

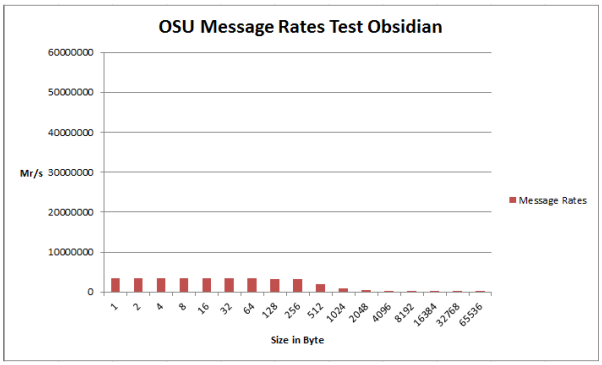

Figure 12: Test setup results, small block sizes – with 10 km fiber

b. TCP Performance

We tested the TCP throughput using the tool *iperf* Version 2.0.4 with the following parameter:

```
iperf -c 192.168.0.224 -P4

-c 192.168.0.224 = ip address of the client
-P4 = 4 threads in parallel
-s = server mode
```

iperf between *rivera3* and *rivera5* connected on the same IB switch:

```
rivera3:~ # iperf  -c 192.168.0.225 -P4
------------------------------------
Client connecting to 192.168.0.225, TCP port 5001
TCP window size:  193 KByte (default)
------------------------------------
[SUM]  0.0-10.0 sec  27.2 GBytes  23.3 Gbits/sec

rivera5:~ # ./iperf -c 192.168.0.223 -P4
------------------------------------Client
connecting to 192.168.0.223, TCP port 5001
TCP window size:  193 KByte (default)
------------------------------------------
[SUM]  0.0-10.0 sec  27.6 GBytes  23.7 Gbits/sec
```

When moving *rivera5* and *rivera6* to the

other side of the WAN, we get the following results:

```
rivera3:~ # iperf -c 192.168.0.225 -P4
------------------------------------------Client
connecting to 192.168.0.225, TCP port 5001
TCP window size:  193 KByte (default)
------------------------------------------------
[SUM]  0.0-10.0 sec  9.21 GBytes  7.89 Gbits/sec
rivera5:~ # ./iperf -c 192.168.0.223 -P4
------------------------------------------Client
connecting to 192.168.0.223, TCP port 5001
TCP window size:  193 KByte (default)
------------------------------------------

[SUM]  0.0-10.0 sec  9.20 GBytes  7.89 Gbits/sec
```

The performance of nearly 8Gbit/s is close to the peak rate for a single SDR x4 link with a maximal signaling rate of 10Gbit/s.

c. **GPFS Tests**

For the final GPFS test we simulated an environment with 2 I/O nodes in Manno and 4 I/O nodes in Lugano, the nodes are connected to the associated storage controller with 60 SATA disks. The I/O nodes are connected through the 10km fiber. The file system was configured with 6 Network Storage Devices (NSD) assigned to 6 LUNs in Manno.

The intention was to accommodate GPFS cluster over two sites (10KM), migrate NSDs online, analyze the bandwidth and latency characteristics when running GPFS over the remote fabric.

Test1: Comparison of bandwidth performance within the local fabric and with remote fabric.

Our main interest here was the impact of the 10km fiber connection on the GPFS bandwidth. We run the ***gpfsperf*** tool on *rivera2* by creating a 30GB file with 4MB blocks using 1 process and 6 Threads on all 6 local NSDs in "Manno". We measured a rate of 624 MB/s, with a thread utilization of 0.98.

The same test was repeated, but on *rivera6* in "Lugano" over the 10km fiber writing to the 6 NSDs in "Manno". We measured 700MB/s, which is even slightly higher. This can be explained by the fact that *rivera6* is equipped with higher main memory than *rivera2*.

We conclude, that in our case the bandwidth is independent of the distance of the NSDs to the IO nodes.

Test2: Migrate 6 NSDs (6 LUNs) from "Manno" to "Lugano"

Creating and removing NSDs is the procedure to follow for migrating data from Manno to Lugano. The purpose of this test is to guarantee the functionality, measure the latency effects on GPFS and time to completion for the migration of all NSDs from one site to the other with ***mmdelnsd***. The command was launched on *rivera6* in "Lugano".

For this test we filled the NSDs in "Manno" with 2 TB of data and 5050 files. The data were evenly distributed over the 6 NSDs. This migration took 90 minutes. Analysis of the data distribution on the NDSs in "Lugano" after the migration showed an almost identical distribution.

The test was repeated migrating only 2 out of 6 NSDs, however the total migration time did not change.

Test3: Latency

In order to get the full I/O statistics we used the ***mmpmon*** tool based on 75K files. The command was launched from *rivera3* ("Manno") to *rivera5/6* over the 10km link and also from *rivera6* to *rivera5/6* in order to compare the local latency.

In the distribution picture for the local GPFS (figure 13), we see a maximum between 20 and 50 ms, for the remote GPFS the curve is slightly shifted to higher latencies between 50 and 100ms (figure 14).

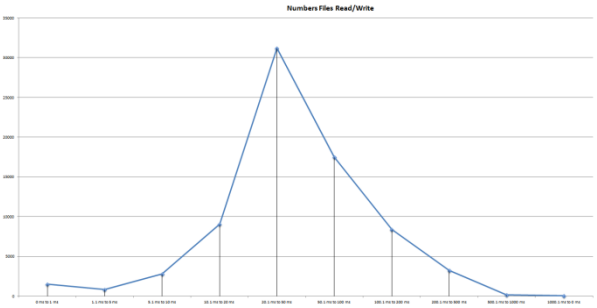

Figure 13: Distribution of latency for "local" GPFS, unit is ms.

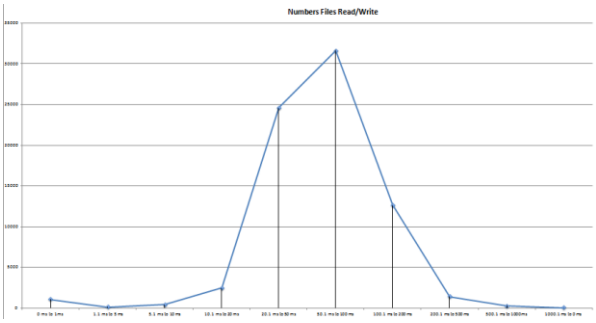

Figure 14: Distribution of latency for "remote" GPFS, unit is ms.

Test 4: Bandwidth over of all four Longbow links

In order to saturate the 4 Longbow links, we changed the setup using *rivera1* and *rivera2* on the Manno side, and *rivera3, rivera4, rivera5 and rivera6* writing to 4 NSDs on the Lugano side. ***GPFSperf create*** launched on *rivera1* and *rivera2* yields an aggregated average write bandwidth of 1558 MB/s. Figure 15 shows the performance which is evenly distributed over all four links. We expect a performance of more than 2GB/s by adding more NSDs in this setup.

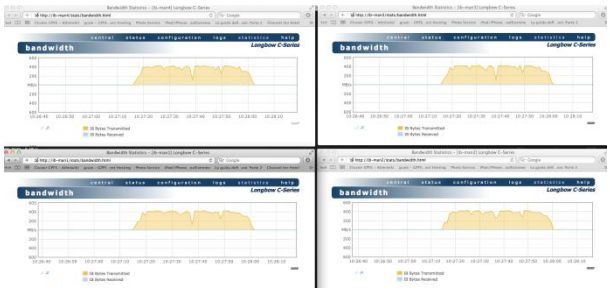

Figure 15: Write Bandwidth for all four Longbow links when running two GPFS clients.

Test 5: Failover Test

Permanent operation of GPFS is one of the key requirements of CSCS. Therefore we were interested in the resiliency features of this solution. The stress test was carried out during the execution of the **gpfsperf** performance test.

We decided to unplug up to three of the 4 cables between the SDR-box and the CWDM device on the "Manno" side. We started by unplugging one cable, followed by unplugging another cable 5 minutes later. In both cases the aggregated bandwidth performance did not change, because the remaining two links absorbed the additional traffic. After another 5 minutes we unplugged the third cable, and the aggregated bandwidth performance dropped by 30% as expected. In terms of functionality, GPFS was not affected, even when running with only one SDR link over the Longbow.

When we reconnected the cables the link was recognized within one minute, and the performance fully recovered.

We finally ran ***mmfsck*** and no error was reported.

4. Conclusion

Based on our results of this evaluation study, CSCS will deploy the Obsidian InfiniBand Range Extender C103 Series for the migration of User data from the current location in Manno to the new center in Lugano-Cornaredo.

The performance will be sufficient for migrating the complete GPFS File system from Manno to Lugano within a few weeks.

The key characteristics of resilience have been analyzed and we are convinced that the Obsidian technology is a very cost-efficient and reliable solution to complete the relocation until June 2012.

5. Literature

[1] Presentation Oak Ridge, ofed-2008-wyu.pdf, Spring 2008.

[2] Bay Microsystems, IBEx G40/M40 Global IB Extension Switch/Router, Product Brief, 2011

[32] Xiangyong Ouyang et al., Filesystem Performance Evaluation of Obsidian Longbow Routers, Ohio State University, Evaluation Report, March 2011.

[4] Obsidian, InfiniBand Range-Extending Switch – Dark Fiber, Long-Bow Datasheet, 2011.

[5] David Southwell, Obsidian Strategies, Presentation, 2009