Technical Report: Report on Lustre use across an experimental 100Gb network spanning 2,175 mi

Robert Henschel
Stephen Simms
David Hancock
Scott Michael
Tom Johnson
Nathan Heald
Thomas William
Donald Berry
Matt Allen
Richard Knepper
Matthew Davy
Matthew Link
Craig A. Stewart

Indiana University

PTI Technical Report PTI-TR12-002

17 February 2012

Citation:

Henschel, R., S. Simms, D. Hancock, S. Michael, T. Johnson, M. Davy, M. Link and C.A. Stewart. "Technical Report: Report on Lustre use across an experimental 100Gb network spanning 2,175 mi," Indiana University, Bloomington, IN. PTI Technical Report PTI-TR12-002, Jan 2011. http://hdl.handle.net/2022/14137

PERVASIVE TECHNOLOGY INSTITUTE
INDIANA UNIVERSITY

RESEARCH TECHNOLOGIES
INDIANA UNIVERSITY
University Information Technology Services
Pervasive Technology Institute

This material is based upon work supported in part by the National Science Foundation under Grant No. CNS-0521433 to Indiana University for "MRI: Acquisition of a High-Speed, High Capacity Storage System to Support Scientific Computing: The Data Capacitor." Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

More information is available at: http://pti.iu.edu/dc

Table of Contents

Table of Tables

Table of Figures

1. Introduction

The SCinet Research Sandbox (SRS) at the Supercomputing 2011 (SC11) conference encouraged institutions to showcase new and innovative technologies in the area of networking. For the demonstration SCinet, in collaboration with ESnet and Internet2, provided SRS participants with a 100 Gbps network connection from the SC11 show floor to the Internet2 backbone. The 100 Gbps link provided an end-to-end connection from the IU booth on the SC11 show floor to the IU Data Center in Indianapolis, Indiana.

Figure 1 shows a network map of the link with the major routing points. SCinet governed access to the network. Each participant was given time slots for exclusive use of the network. The slots were evenly distributed from Saturday, November 12th to Thursday, November 17th. In total, IU was provided nine test slots for a combined 16 hours. All testing that required access to the network links had to be performed during those times, from setting up the actual end-to-end network connectivity to performing file system and application tests. In addition, we were provided five demonstration slots for a total of four hours. Those time slots were used to showcase the capabilities of the system in the IU booth. All results described in this technical report were obtained in the 20 hours of demonstration and test time.

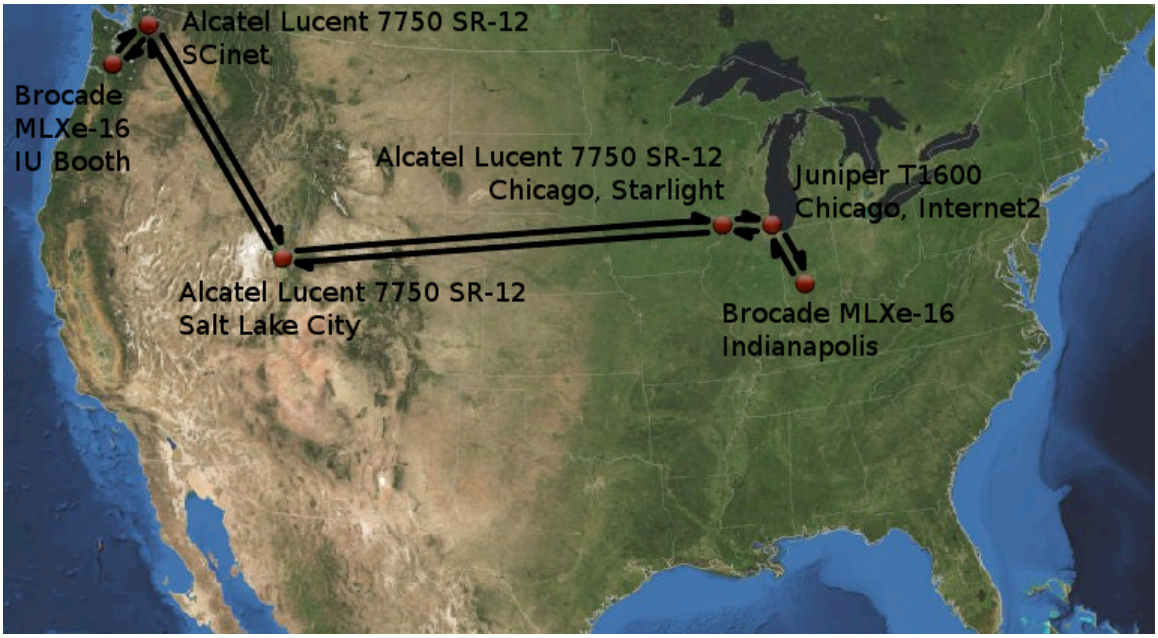

Figure 1. Network map using IU GlobalNOC World View.

2. System Description

2.1. Hardware

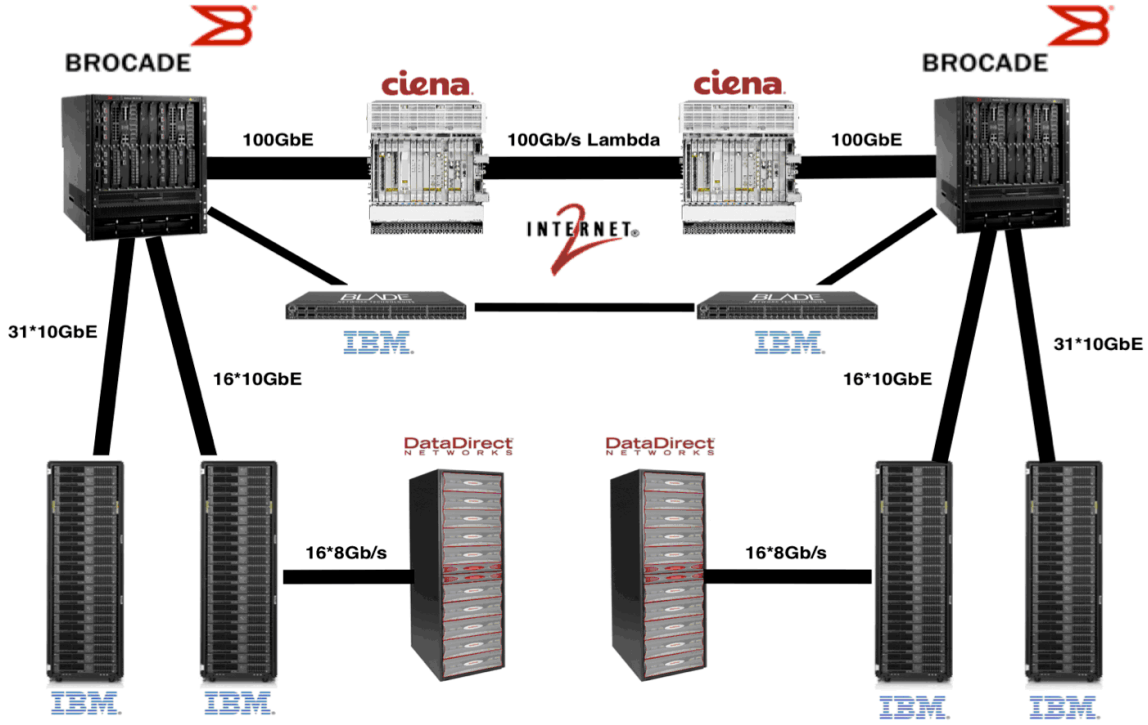

Figure 2. Hardware setup.

Figure 2 shows the final configuration that was used between Seattle and Indianapolis. IBM provided 31 servers that functioned as compute nodes as well as the 16 storage servers that were attached to the DataDirect Networks (DDN) storage devices. Brocade and Ciena provided the routing equipment that enabled the network link from the show floor to Indianapolis.

The configurations of the compute cluster, storage, and networking components were identical in both Indianapolis and Seattle. The core networking component at each endpoint was a Brocade MLXe-16 router that provided a 100 Gbps Ethernet connection to the Ciena optical terminal managed by Internet2. The core router also provided a 10 Gbps link to an IBM BNT G8264 OpenFlow enabled switch at each endpoint. These switches were connected at 10 Gbps over a separate Internet2 connection. The 31 compute servers and 16 Lustre storage servers were attached directly to the Brocade core router at 10 Gbps using Twinax cables and Brocade 1860 dual-port adapters.

The compute servers were IBM System x iDataPlex dx360 M3 systems, each configured with dual Intel Xeon E5645 6-core 2.40 GHz processors, 24 GB of DDR3 RAM, a Brocade 1860 adapter, and a 250 GB SATA hard drive. The object storage servers (OSS) were IBM System x iDataPlex dx360 M3 servers each configured with an Intel Xeon E5645 6-core 2.40 GHz processor, 48 GB of DDR3 RAM, a Brocade 1860 adapter, and a 1 TB SATA hard drive. The OSS nodes at each site were connected directly to a DDN SFA10000 via 8 Gb Fibre Channel (FC). The SFA10000 drove five 60-slot storage enclosures configured with 2 TB SATA disk drives. The metadata server was identical to the compute servers, except it had 96 GB of RAM and was directly connected to a DDN EF3015 RAID system that contained twelve 300 GB 15K RPM SAS disk drives for Lustre metadata.

Due to space constraints on the show floor, server density was important. The dual-port Brocade 1860 adapters were able to saturate the 8 Gb FC links as well as the 10 Gbps Ethernet links simultaneously,

allowing the use of rack-dense IBM iDataPlex servers with only one PCI-Express slot available. The throughput of the DDN SFA10000 allowed us to use a single storage system for the Lustre OSS nodes at each site. In Seattle and Indianapolis, 14 of the 16 Object storage servers had 2 OSTs, while two had a single OST.

2.2. Software

Both compute and storage servers were installed with Red Hat Enterprise Linux Version 5.7 and version 1.8.6-wc of Lustre. During testing we made no significant changes to the software itself, though some software parameters were altered as follows:

For all testing, Linux TCP tuning parameters were altered to accommodate the 50.5 ms of latency between Seattle and Indianapolis as follows:

net.ipv4.tcp rmem=4096 65536 167772160
net.ipv4.tcp wmem=4096 65536 167772160
net.core.rmem max=167772160
net.core.wmem max=167772160
net.core.netdev_max_backlog=30000
eth2 txqueuelen 10000
eth2 mtu 9000
FlowControl off

3. Performance Testing

In the course of our studies across the 100 Gbps network the following Lustre parameters were varied for the IOR and LNET tests:

rpcs_in_flight between 8 and 128
peer_credits between 8 and 64
credits between 64 and 2048
max_dirty_mb between 32 and 256

The results below represent the best results that we achieved.

	Compute nodes to local storage over 100 Gbps	**From Seattle computes nodes to Indianapolis storage**
Latency	0.24 ms	50.5 ms
TCP iperf, one stream	9.8 Gbps	9.8 Gbps
TCP iperf, 30 streams	98 Gbps	96 Gbps
IOR, 1 client, 16 servers	1.2 GB/s	1.2 GB/s
IOR, 30 clients, 16 servers	9.6 GB/s	6.5 GB/s
LNET, 1 client, 8 servers	1.2 GB/s	1.1 GB/s
LNET, 10 clients, 10 servers	11.9 GB/s	9.5 GB/s
8 applications (peak)	8.8 GB/s	6.2 GB/s
8 aplications (sustained)	7.9 GB/s	5.6 GB/s

Table 1. Key results.

Table 2 lists the applications that were used during testing.

Application / Workflow	Domain	Number of Nodes
Heat3D	Heat Diffusion	8
VampirTrace	Application Performance Analysis	7
Enzo	Astronomy	6
NCGAS	Genomics	3
OLAM	Weather	2
CMES	Computational Neuroscience	2
Gromacs	Molecular Dynamics	1
ODI-PPA	Astronomy	1

Table 2. Applications.

4. Conclusion

Our SRS entry was as much an experiment at the network layer as it was at the file system or applications layer. This is substantiated by the fact that across the SRS we observed 100 Gbps equipment from three different vendors interoperating for the first time across thousands of miles. We assume that as both 100 Gbps technology and national deployments mature, issues like those we encountered at the SRS demonstration will be addressed and overcome. It should also be noted that even though we were unable to fully saturate the SRS 100 Gbps link, real world applications were still able to be run over the WAN and, on average, each achieve several Gbps throughput. Indeed, from the end user's point of view, the true breakthrough of Lustre-WAN is not necessarily the ability to achieve 90%+ network efficiency, though it has been shown to be possible, but the streamlining of their workflow by eliminating the need for user-directed file transfers. Overall, we see the future of Lustre-WAN and 100 Gbps networking to be very promising, albeit with a period of adjustment in the near term future.