The Lustre File System and 100 Gigabit Wide Area Networking: An Example Case from SC11

Richard Knepper*, Scott Michael†, William Johnson‡, Robert Henschel§, Matthew Link¶
*Pervasive Technology Institute
Indiana University, Bloomington, IN, 47405
Email: rknepper@iu.edu
†Pervasive Technology Institute
Indiana University, Bloomington, IN, 47405
Email: scamicha@iu.edu
‡Global Network Operations Center
Indiana University, Bloomington, IN, 47405
Email: wtjohnso@grnoc.iu.edu
§Pervasive Technology Institute
Indiana University, Bloomington, IN, 47405
Email: henschel@iu.edu
¶Pervasive Technology Institute
Indiana University, Bloomington, IN, 47405
Email: mrlink@iu.edu

***Abstract*—As part of the SCinet Research Sandbox at the IEEE/ACM International Conference for High Performance Computing, Networking, Storage and Analysis (SC11), Indiana University utilized a dedicated 100 Gbps wide area network (WAN) link spanning more than 3,500 km (2,175 mi) to demonstrate the capabilities of the Lustre high performance parallel file system in a high bandwidth, high latency WAN environment. This demonstration functioned as a proof of concept and provided an opportunity to study Lustre's performance over a 100 Gbps WAN. To characterize the performance of the network and file system a series of benchmarks and tests were undertaken. These included low level iperf network tests, Lustre networking tests, file system tests with the IOR benchmark, and a suite of real-world applications reading and writing to the file system. All of the tests and benchmarks were run over a the WAN link with a latency of 50.5 ms. In this article we describe the configuration and constraints of the demonstration and focus on the key findings regarding the networking layer for this extremely high bandwidth and high latency connection. Of particular interest are the challenges presented by link aggregation for a relatively small number of high bandwidth connections, and the specifics of virtual local area network routing for 100 Gbps routing elements.**

I. Introduction

In November 2011 Indiana University (IU) participated in the SCinet Research Sandbox (SRS) at the International Conference for High Performance Computing, Networking, Storage and Analysis (SC11). The purpose of the SRS was to encourage institutions to showcase new and innovative technologies in the area of networking, particularly focusing on OpenFlow technologies. In addition to an OpenFlow test bed, SCinet provided SRS participants with a 100 Gbps network connection from the SC11 show floor to the Internet2 backbone. Working in conjunction SCinet, Internet2, and ESnet were able to provide an end-to-end 100 Gbps connection from the IU booth at SC11 in Seattle, Washington to the IU data center in Indianapolis, Indiana.

The overarching theme of the IU entry into the SRS was to demonstrate how a wide area Lustre file system could be used to empower geographically distributed scientific workflows by utilizing the 100 Gbps wide area network (WAN). The network we used for the SRS demonstration spanned a distance in excess of 3,500 km (2,175 mi) from Seattle, Washington to Indianapolis, Indiana, one of the longest production deployments of a 100 Gbps network. A diagram of the network and routing points is given in figure 1. Such a proof of concept use case is of obvious intrest for geographically distributed workflows; for example, when data sources, such as remote sensors are in locations very distant from the computational resources used to analyze the data [1]. The overall demonstration was successful, with IU achieving the highest-ever reported throughput (6.2 GB/s) with I/O intensive scientific applications reading and writing files using a WAN-based Lustre file system. The applications and IOR benchmarks are described elsewhere [2], as are the lessons learned in regards to the Lustre networking layer (LNET) [3]. The focus of this article is the performance of and lessons learned about the underlying 100 Gbps networking layer.

At the time of the SRS demonstration, production 100 Gbps networks had just begun to be deployed by Internet2 and ESnet. Though Indiana University had previously had great success with Lustre-WAN using cross-country 10 Gbps networks [4] and low latency 100 Gbps networks [5], we had not had a chance to test Lustre-WAN capabilities on 100 Gbps networks over large distances with high latencies. This

978-0-7695-4722-0/12 $26.00 © 2012 IEEE
DOI 10.1109/NAS.2012.36

IEEE computer society

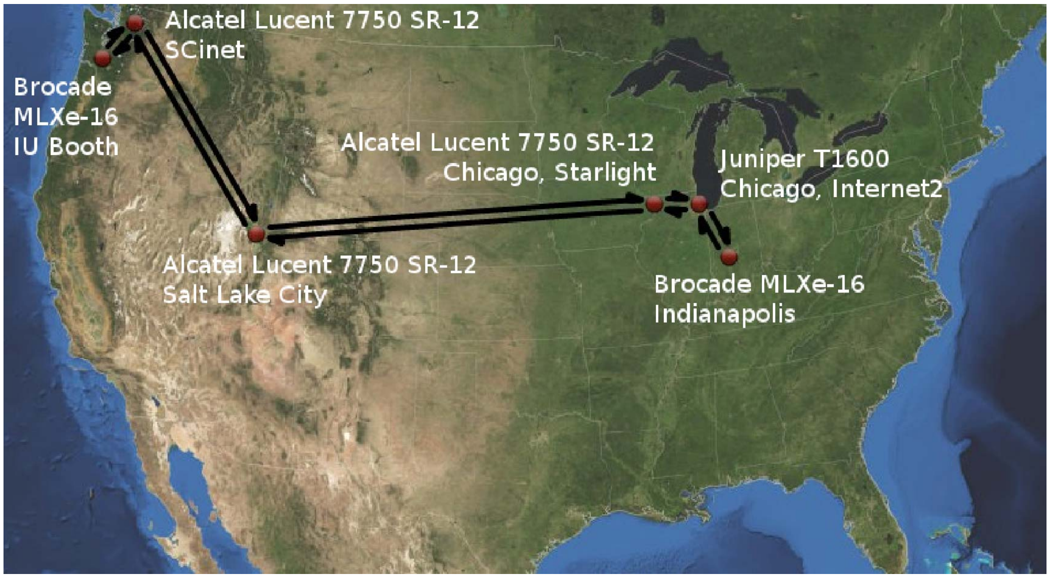

Fig. 1. Networking diagram for SRS demonstration. Routing points are labeled in the diagram, and include a hop in Indianapolis, two in Chicago (to transition from Internet2 to ESnet), a hop in Salt Lake City, and finally two hops in Seattle (to transition from ESnet to SCinet).

demonstration was the first of its kind utilizing Lustre-WAN in such a high bandwidth, high latency environment.

Access to the 100 Gbps WAN was facilitated and coordinated by SCinet. 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 was performed during these times, from enabling 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. These time slots were used to showcase the capabilities of the system in the IU booth. All results described in this article were obtained in the 20 hours of demonstration and test time.

The fact that this study was performed in such a limited timeframe influenced the measurements we were able to take. In addition to the fact that the time allotted to IU was fairly short, the initial connection of the 100 Gbps link required some fine tuning of the routing elements between Seattle and Indianapolis to achieve an eventual uni-directional peak throughput of 96 Gbps on the TCP layer [2]. Although some of the network issues were eventually resolved, the network was quite variable throughout the SC11 conference and, as discussed in section III, we were never able to achieve peak rates with bi-directional throughput. Also, the tuning of the network to achieve this throughput consumed nearly half of the time allocated to IU, leaving approximately 10 hours for demonstrations, gathering network data, gathering data using file system benchmarks, and determining the performance of scientific applications.

Although our time was limited, we were able to make several interesting observations while trying to achieve peak performance from the network. For our relatively small cluster of 30 nodes connected at 10 Gbps each, we found that having a link aggregation layer at each of the endpoints negatively impacted our performance. In addition, we found that the routing of traffic across the two 50 Gbps VLANs that make up the backplane of the Internet2 and ESnet 100 Gbps routers can have an effect on performance.

This paper is structured as follows: Section II describes in detail the SRS use case, including the network configuration and hardware setup that was used for the SRS demonstration. Section III details the changes required to the network configuration to achieve peak performance, and the results we were able to obtain from the networking layer following the tuning. In section IV, we offer some hypotheses explaining the results, and in section V we offer concluding remarks.

II. The SC11 Use Case

As outlined in section I, to showcase the capabilities of Lustre across a wide area network, we set up a Lustre file system and mounted it on the 100 Gbps cross-country WAN. To perform the demonstration we set up a compute cluster on the show floor in Seattle and one in the IU data center in Indianapolis. We also deployed a file system at each location. In addition, we installed and configured networking components to connect to the Internet2 and ESnet endpoints. The rest of this section details the hardware, network, and software set up and configuration that was applied for the demonstration.

A. Networking

Before shipping equipment to the show floor in Seattle the demonstration configuration was assembled in the IU data center for testing. There we performed tests with two Brocade MLXe routers connected back-to-back with the 100 Gbps connection spanning a few meters. Local network tests across this connection showed a latency of 0.24 ms and maximum stable performance of 98 Gbps using TCP iperf [6].

The network link that was used for the SRS demonstration provided 100 Gbps connectivity from the IU booth on the SC11 show floor to the IU data center in Indianapolis. In the months preceding the SC11 conference, Indiana University worked with Internet2 to upgrade the existing link from Chicago to Indianapolis to 100 Gbps. For the SRS demonstration Internet2 and ESnet provided access to the 100 Gbps link from Chicago to Salt Lake City and on to Seattle. SCinet was responsible for the network link from the Interent2/ESnet route point to the show floor. The connection from Salt Lake City to Seattle had been established at 100 Gbps only two weeks prior to SC11 and had not undergone performance testing.

Once the link was established, we measured a 50.5 ms round-trip time (RTT) between the clusters in Seattle and Indianapolis. After addressing some initial routing difficulties and establishing connectivity, TCP iperf tests showed good performance for two parallel streams, each at 10 Gbps. We expected to be able to continue to add streams and to achieve near peak performance. However, when adding more iperf streams to the link, performance dropped and throughput became unstable. Working with Internet2, ESnet, and SCinet, we were able to diagnose the problem as being caused by dynamic virtual local area network (VLAN) routing. By adjusting the routing parameters on several of the routers in the network, we were able to achieve a stable 80 Gbps throughput with TCP iperf. The root cause of this issue and parameters we changed are further detailed in section IV-B. Following all of our modifications, we were able to put a stable 80 Gbps of TCP iperf traffic on the link by using 8 servers and clients. The routing behavior between Seattle and Indianapolis was such that the traffic was split across two 50 Gbps VLANs on the ESnet and Internet2 routers. Due to this setup, adding another single 10 Gbps stream to either of the connections resulted in congestion on one of the VLANs and degraded performance. However, when oversubscribing the link by using all 30 compute nodes on each end, we were able to achieve a peak throughput of 96 Gbps. Another unfortunate outcome was that the tuning steps yielded a stable connection only in one direction, from Seattle to Indianapolis. We were never able to achieve similar results in the other direction, and due to time constraints, we focused our efforts on just one direction. These issues are discussed in more detail in section III.

Listing 1. Tuning parameters for the network

```
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
```

Listing 1 shows the network tuning parameters that were set on all nodes. We increased the maximum TCP buffer to 167 MB and increased the sending and receiving queues to 10000 and 30000. In addtion, we enabled MTU 9000 and configured all nodes to use the `tcp_bic` network stack. Flow control was disabled on the network adapters and the routers.

B. Hardware

Figure 2 shows the final hardware configuration that was used for the SRS demonstration. Each site was equipped with 31 IBM servers that functioned as compute nodes as well as 16 storage servers that were attached to DataDirect Networks (DDN) storage devices. Brocade and Ciena provided the network 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 central 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. This core router also provided a 10 Gbps link to an IBM BNT G8264 OpenFlow enabled switch at each endpoint. The SRS demonstration comprised a OpenFlow component in addition to the main Lustre 100 Gbps demonstration. However, we will not discuss the OpenFlow component in this article (see [7] for details on the OpenFlow component of the SRS demonstration). In the final configuration, 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), which was populated with 300 2 TB SATA hard 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 throughput of the DDN SFA10000 allowed us to use a single storage system for the Lustre OSS nodes at each site.

III. Results

Due to the limited deployment of active cross-country 100 Gbps network paths all SRS participants shared a single link. Entities who participated in the SRS were allocated time slots throughout the day to utilize the single 100 Gbps link, all

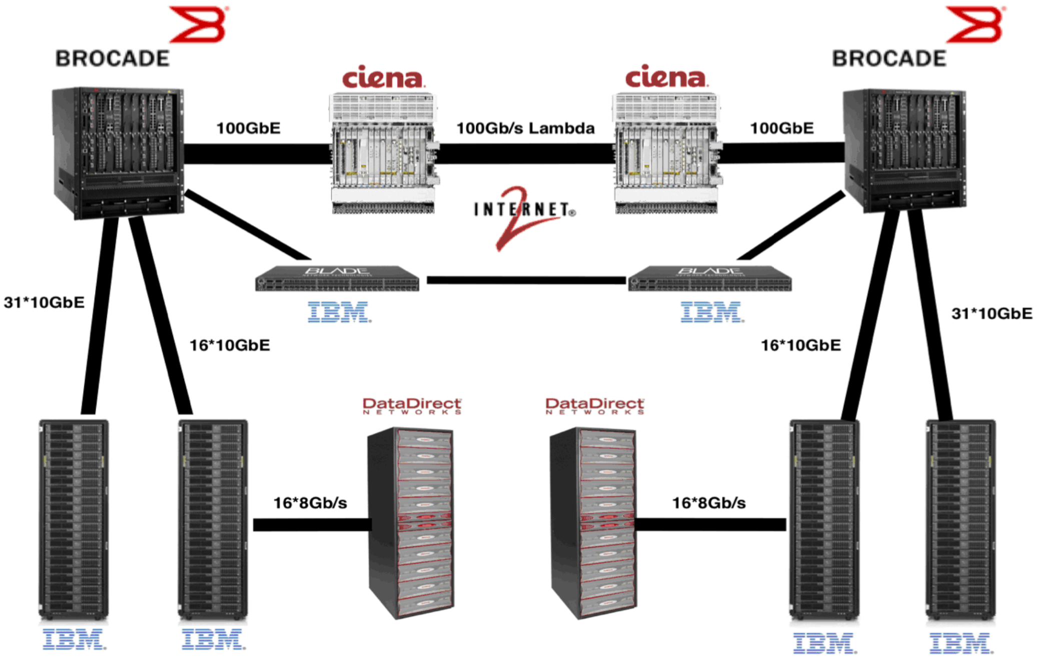

Fig. 2. Shown here is the hardware configuration used for the SRS demonstration. An IBM iDataPlex cluster and Lustre file system were connected to a Brocade 100 Gpbs MLXe-16 router at each endpoint, one in Seattle and one in Indianapolis. The IBM routing equipment connected to the Brocade routers was used for the OpenFlow component of the demonstration and is not discussed in this article.

with goals of demonstrating real world experiments. During allocated time, the link was 100% dedicated to the scheduled participant, and at all other times, bandwidth was restricted. All organizations participating in the SRS demonstration at SC11 were allocated a similar number of time blocks.

For our initial deployment, IU selected a Brocade MLXe-16 chassis, accompanying high speed fabric modules, management modules, three 8-port 10GbE modules and a 2-port 100GbE module. To provide connectivity to the 30 compute nodes and 16 storage nodes, we also selected a Brocade VDX 6720-60. The VDX provided 60 ports of 10GbE connectivity. The compute and storage nodes also leveraged Brocade 1860 Fabric Adapters which are 2-port small form-factor pluggable plus (SFPP) modular PCI express adapters capable of any combination of line rate 10GbE or 8Gb FC. The compute nodes utilized one port of 10GbE using twinax (or direct attach copper) cables to the VDX while the Lustre OSS nodes used 10GbE twinax and multi-mode fiber (MMF) for FC (see section II-B for a description of the file system configuration). With 16 OSS, this configuration provided an aggregate bandwidth of 160 Gbps on the Ethernet side and 128 Gbps on the Fibre Channel side. Either of these aggregate bandwidth numbers exceeded the I/O capability of the SFA10000 due to the number of drives it contained, and the configuration of those drives into the OSTs. To ensure that the Brocade 1860 Fabric Adapters could perform at peak on both the 10GbE and FC interfaces simultaneously, we ran several tests through individual OSS nodes and measured the throughput on both the interfaces. We found that the hardware was able to deal with the simultaneous load and gave very good performance on both the 10GbE and FC.

The entire system consisted of 48 nodes (storage, compute, and management) connected to the VDX with 12 parallel link aggregation connections between the VDX and the MLXe. These 12 connections all used the same length twinax cabling, and formed the link aggregation group (LAG). The 12 LAG connections plugged into the MLXe, which connected to the 100GbE network. We used the remaining 12 10GbE ports for additional experiments, management connections and external connectivity. The initial configuration of the MLXe came with 24 10GbE ports, meaning that we lacked the necessary number of ports in the MLXe to connect all 48 storage and compute nodes directly to the MLXe. Figure 3 shows how the VDX and link aggregation fit into the overall networking of this initial configuration.

Prior to the demonstration, we staged both of the storage and compute systems in the Indianapolis data center and directly connected the two MLXe devices. In this test configuration, the 100GbE link was a ten meter jumper between the two mirrored systems. Latency across the link from host to host was 0.24 ms, and we were able to achieve a stable 98 Gbps unidirectional transfer. This result required utilizing 30 compute nodes transmitting to thirty 30 compute nodes receiving.

During the conference, latency across the 100GbE shared infrastructure increased to a consistent 50.5 ms. This increased

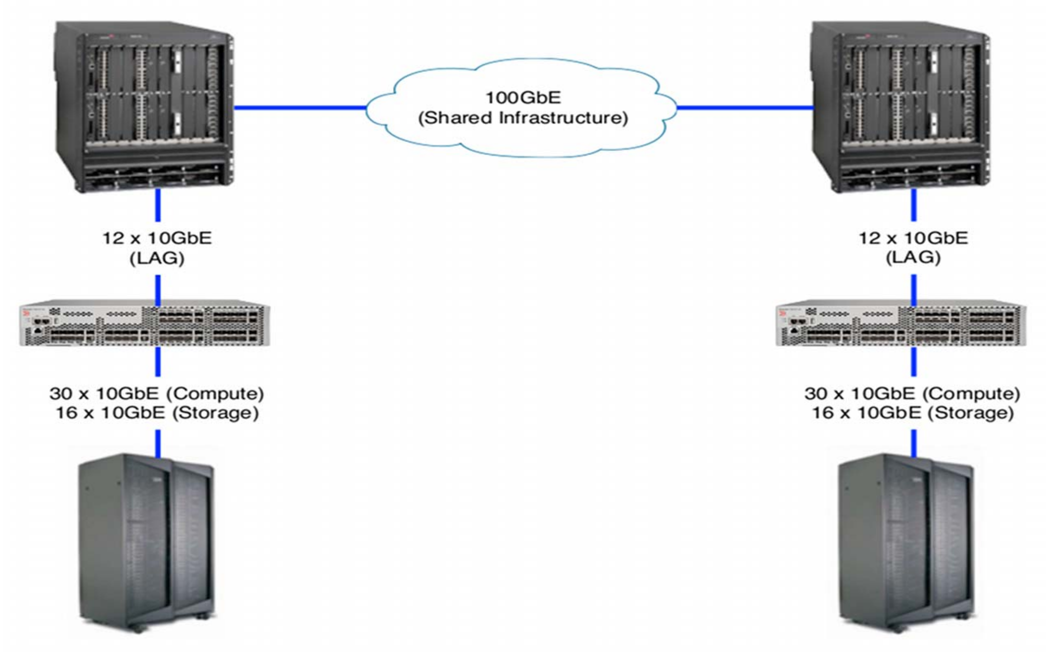

Fig. 3. The initial setup of the 100 Gbps testbed including the VDX devices. Each VDX connected the 48 compute and storage nodes to the 12 link LAG, which connected into the MLXe-16.

latency was primarily due to the fact that the distance the link spanned had increased from 10 meters to over 3500 kilometers. In addition, the two systems were no longer on a single subnet, but now had several routing elements between them that were not present in the staging setup. We experienced a greater challenge utilizing link aggregation in situ. With 30 iperf clients communicating to 30 iperf servers through the LAG and across the 100 Gpbs link, congestion on the constituent connections of the LAG was inevitable. For 30 clients transmitting 10 Gbps traffic through a bottleneck of twelve 10 Gbps links, TCP congestion was unavoidable, but it did not seem to negatively affect our results when both systems were tested in the Indianapolis data center. However, the added latency amplified the impact of TCP retransmission resulting in all clients dramatically suffering from the added congestion. The result was that we could only sustain 30 Gbps of traffic from thirty 10 Gpbs clients.

As described in section IV, we had the VDX configured with the hashing algorithm that provided maximum throughput in the Indianapolis data center. Due to time constraints we were unable to test other hash algorithms at the conference. In order to insure success during the limited time allotted to us for the SRS demonstration, we worked with Brocade to remove the VDX and local link aggregation, and install hardware to provide additional 10GbE ports directly into the MLXe-16.

An additional technical challenge we encountered was that not all 100GbE network elements were capable of forwarding 100 Gbps of traffic contiguously. For two of the routing elements, (Juniper T1600 and Alcatel-Lucent 7750 SR-12) provided by Internet2 and ESnet, respectively, traffic had to be separated into two virtual local area networks (VLANs), each parallel across the physical infrastructure. That is, each side of the routing element, the inbound port and the outbound port, were both capable of 100 Gbps of traffic, but the internal routing fabric broke each stream into 50 Gbps VLANs. In figure 4 this is represented by the orange and green paths. Our solution was to implement two VLAN ID's, with an equally divided number of compute and storage nodes statically routed across each of the parallel paths.

Once both of these challenges had been addressed, we were able to achieve a stable unidirectional TCP throughput of 80 Gpbs iperf traffic using eight clients and servers on each side of the link. When using 30 iperf clients and servers we were able to achieve a peak unidirectional TCP throughput of 96 Gbps. We were only ever able to achieve good TCP results in one direction on the link, for traffic traveling from Seattle to Indianapolis. Though the link from Indianpolis to Seattle did not appear to have loss, we were able to saturate the link with UDP traffic and little loss, it seems that packet reordering caused a great deal of TCP retransmits resulting in overall poor TCP performance. We made every effort to work with SCinet, Internet2, and ESnet to address the packet reordering issue, but ultimately, since our time was limited, we were unable to resolve the issue.

IV. Discussion

In our experience with the SRS demonstration we faced challenges in achieving the throughput we expected to see from a 100 Gbps cross country link. These challenges were mainly caused by TCP congestion due to TCP flows not

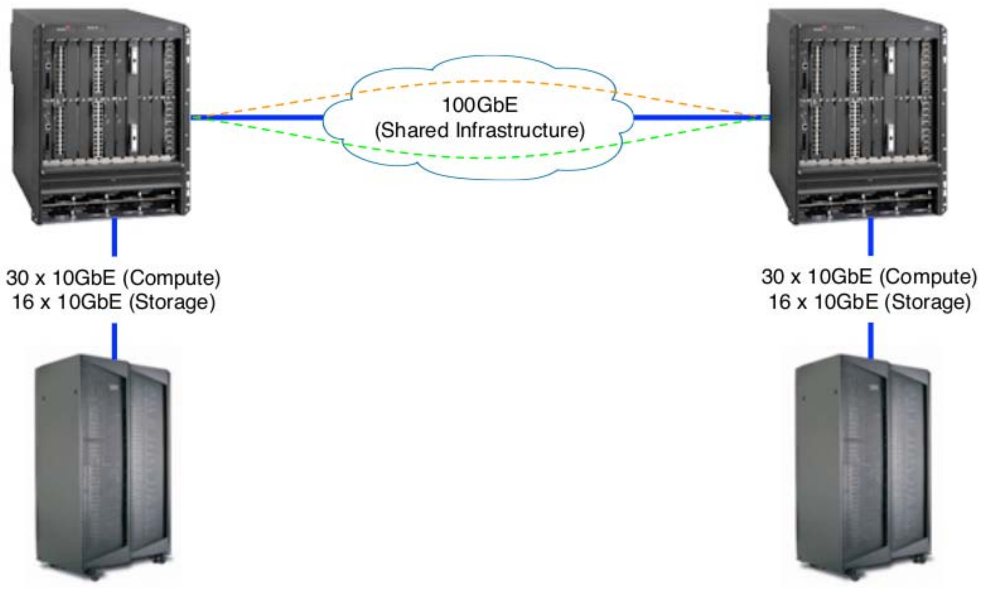

Fig. 4. Diagram of 100 Gbps configuration after removing the VDX. The two separate 50 Gbps VLAN paths are represented by the green and orange dashed lines.

being optimally routed. Both link aggregation and the dynamic VLAN routing within the 100 Gbps router fabrics presented some initial challenges. Although the default settings may be appropriate in many situations, they did not deliver optimal performance for our SRS demonstration.

A. Link Aggregation

We found when the two systems were side-by-side in the Indianapolis data center, we were able to achieve good results with the LAG in the network path. This configuration is represented in figure 3. However, using the same configuration across the WAN our performance was significantly degraded. There were several factors that could have affected the overall throughput for the LAG, including the number of parallel flows and the hashing algorithm used to evenly distribute traffic across the link aggregation. A flow consists of a TCP exchange between a pair of hosts. Any pair of hosts can have multiple TCP flows if they operate on different ports. However, in our network testing we sought to simulate the traffic we would be experiencing in the demonstration of the Lustre-WAN file system, and Lustre uses a relatively small number of ports for its communication. So, in all of our iperf testing we instantiated a single flow on a single port per host pair. However, each host pair received a *different* port, thus increasing the hash parameter space, as explained below. In the Lustre-WAN demonstration nearly all of the communication across the 100 Gpbs link was due to I/O. That is, between one compute cluster and the storage nodes on the other end of the 100 Gbps link. Each new host either writing to a Lustre OSS or an OSS serving a file constitutes a new flow. Using 30 compute nodes on one side of the link and 16 OSS on the other allows for a minimum of 16 flows, in a 1:1 configuration, and a maximum of 480 flows, in a 1:all configuration. To test the setup we used the 30 compute nodes on either side of the link as iperf clients and servers, and minimized the number of flows to in order to represent a “worst case scenario” for the distribution of traffic across the constituent connections of the link aggregation.

Clearly, if there are exactly the same number of TCP flows as links in the LAG there should be no TCP congestion, if distributed properly each TCP flow will get its own link. Also, if the number of TCP flows is much greater than the number of links in the LAG, then the throughput of TCP traffic should stabilize near the peak throughput for the LAG. That is, with a sufficiently large number of flows, the scaling back of a few flows due to congestion should not dramatically impact the aggregate throughput. However, in our setup we typically had a number of flows slightly greater than the number of links in the LAG, up to a few times the number of links in the LAG.

The ratio of the number of flows to the number of LAG links affects the performance of the system due to the fact that link aggregation uses a hashing algorithm to attempt to dynamically distribute traffic evenly. The parameters used to determine the hash are source IP address, destination IP address, port number, MAC address and VLAN ID. Since all hosts were part of the same broadcast domain, being part of a single cluster at each respective side, the VLAN ID was consistent, and therefore did not change from one TCP flow to another. Likewise with the MAC address, the destination MAC address of all the clients on either end of the 100 Gbps link was that of the virtual MAC on the MLXe, so it did not vary from one TCP flow to another. Of the three remaining variables, source IP, destination IP, and port number, we found that if we removed the port number from the hashing algorithm performance suffered dramatically.

The best throughput is achieved when the TCP flows are evenly distributed across the members of the LAG. If this is not the case, some LAG links will have more TCP flows, resulting in TCP congestion and scaling back of the flow, while others will have fewer TCP flows, resulting in the underutilization of the link. By using source IP, destination IP, *and* port number, we increased the parameter space of the hashing algorithm. This, in turn, increased the probability that flows will be distributed evenly since there are more potential combinations. However, even when using the settings that produced the optimal measurements in the IU data center, we were unable to see similar performance over the 50.5 ms of latency. This is probably due to the fact that the number of TCP flows we were generating was at most a few times larger than the number of links in the LAG. So, when flows scaled back because there was TCP congestion on a link, it produced a dramatic effect. Perhaps, if we had many more flows with a greater number of source and destination IPs and port numbers, we would have seen better results from the LAG.

B. VLAN Routing

Once the LAG had been removed from the network, and the compute and storage nodes were directly connected to the MLXe-16 routers, we had a configuration as shown in figure 4. However, initial tests in this configuration still gave poor performance. Again, it seemed to be due to the way that TCP flows were assigned to different routes. For several of the routing elements between Seattle and Indianapolis, specifically one Juniper T1600 and three Alcatel-Lucent 7750 SR-12s, the internal routing fabric split the single 100 Gbps path into two 50 Gbps VLANs. Each of the 50 Gbps paths is represented by a different colored dashed line in figure 4. Different providers determined how TCP flows would be assigned to the the split VLAN solution in different ways. One vendor required odd and even VLAN ID's (decimal value represents the bits value in the transmitted Ethernet frame), we selected 265 and 266. We then assigned half of the compute nodes and half of the storage nodes to the odd VLAN ID and the other half to the even VLAN ID. As long as the amount of traffic coming from each half of the cluster was roughly equivalent, the two 50 Gbps VLANs should have been evenly loaded. Another provider implemented two Virtual Private LAN Service (VPLS) services, each with a unique VLAN ID. The default behavior of the VPLS service was to use an unknown hashing algorithm to distribute traffic evenly between the two 50 Gbps VLANs. However, the dynamic nature of the hashing algorithm seemed to interfere with the TCP congestion control algorithms, resulting in poor performance. Upon further testing, participants in the conference discovered that distribution using a per-packet method was preferred as it evenly balanced the load (i.e. putting subsequent packets on alternating VLANs) across the parallel 50 Gpbs connections.

V. Conclusions

As a demonstration for the SCinet Research Sandbox, Indiana University deployed a wide area Lustre file system spanning more than 3,500 kilometers over a 100 Gbps network. Our initial assumptions regarding the networking and local testing within the IU data center required some adjustments once we deployed on the wide area link with 50.5 ms of latency. Namely, we found that for the number of clients and servers we were testing, 30 clients to 30 servers, link aggregation and VLAN routing across the 100 Gbps routers played a key role in our ability to achieve peak throughput. By removing the link aggregation group and adjusting how TCP flows were assigned to the constituent 50 Gbps VLANs we were able to achieve a peak unidirectional TCP throughput of 96 Gbps. The key in both instances was to insure that traffic was distributed evenly across the available paths, and, whenever possible, TCP congestion was avoided. When we were unable to do this our throughput suffered dramatically.

We assume that as 100 Gbps technology matures issues such as even distribution of TCP flows will be more fully addressed and dynamic routing schemes will become more sophisticated. In addition, as 100 Gbps router technology advances we expect that the division of 100 Gbps paths into two 50 Gbps paths will be eliminated. For now, peak performance is achievable with a relatively small number of clients and servers, but in order to get optimal results, some fine tuning may be required.

Acknowledgment

The authors would like to thank the contributions of our collaborators and technical staff. The following staff at IU provided excellent support in the administration of the systems, testing applications, visualization, technical writing, and project management: Edward Balas, Janae Cummings, Kurt Seiffert, Daphne Siefert-Herron, Bill Sherman, Martin Swany, David Hancock, Jenett Tillotson, George Turner, Matt Allen, Nathan Heald and Josh Walgenbach. Internet2 provided network setup and support: Andrew Lee, Chris Robb, and Matthew Zekauskas. ESnet staff also provided setup and troubleshooting assistance: Evangelos Chaniotakis and Patrick Dorn.

The following vendors provided loaner hardware to support this project: Brocade loaned all the server host adapters, two MLXe-16 100 Gbps switches, required Fibre Channel and Ethernet optics, and all the Twinax cables used to connect the servers to the network. Ciena loaned optical equipment to Internet2 to enable the 100 Gbps network from Chicago to Seattle. DataDirect Networks loaned two SFA10000 storage systems for Lustre object storage targets and two EF3015 RAID arrays for Lustre metadata storage. IBM contributed 22 iDataPlex dx360 M3 servers for the project as well as two OpenFlow enabled BNT G8264 switches with required optics. Internet2 provided networking equipment to extend their 100 Gbps network from Chicago to Indianapolis.

References

[1] R. Henschel, S. Michael, and S. Simms, "A distributed workflow for an astrophysical OpenMP application: using the data capacitor over WAN to enhance productivity," in *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, ser. HPDC '10. New York, NY, USA: ACM, 2010, pp. 644–650.

[2] R. Henschel, S. Simms, D. Hancock, S. Michael, T. Johnson, N. Heald, T. William, M. Allen, R. Knepper, M. Davy, M. Link, and C. Stewart, "Demonstrating Lustre over a 100Gbps Wide Area Network of 3500km," in *Proceedings of 2012 International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '12. Submitted: ACM, 2012.

[3] S. Michael, L. Zhen, R. Henschel, S. Simms, E. Barton, and L. Matthew, "A Study of Lustre Networking Over a 100 Gigabit Wide Area Network with 50 milliseconds of Latency," in *Proceedings of the 21th ACM International Symposium on High Performance Distributed Computing*, ser. HPDC '12. Accepted: ACM, 2012.

[4] S. C. Simms, G. G. Pike, S. Teige, B. Hammond, Y. Ma, L. L. Simms, C. Westneat, and D. A. Balog, "Empowering distributed workflow with the data capacitor: maximizing lustre performance across the wide area network," in *SOCP '07: Proceedings of the 2007 workshop on Service-oriented computing performance: aspects, issues, and approaches*. New York, NY, USA: ACM, 2007, pp. 53–58.

[5] M. Kluge, S. Simms, T. Wiliam, R. Henschel, A. Georgi, C. Meyer, M. Mueller, C. Stewart, W. Wuensch, and W. Nagel, "Performance and quality of service of data and video movement over a 100 Gbps testbed," *Future Generation Computer Systems*, vol. Sumbitted, 2012.

[6] iperf Team, "Home page," http://sourceforge.net/projects/iperf/, 2012.

[7] E. Kissel, G. Fernandes, M. Jaffee, M. Swany, and N. Zhang, "Driving Software Defined Networks with XSP," in *Accepted for SDN'12: Workshop on Software Defined Networks*. IEEE, 2012.