Disk-to-Disk network transfers at 100 Gb/s

Artur Barczyk\textsuperscript{a}, Ian Gable\textsuperscript{b}, Marilyn Hay\textsuperscript{c}, Colin Leavett-Brown\textsuperscript{b}, Iosif Legrand\textsuperscript{a}, Kim Lewall\textsuperscript{b}, Shawn McKe\textsuperscript{d}, Donald McWilliam\textsuperscript{c}, Azher Mughal\textsuperscript{a}, Harvey Newman\textsuperscript{a}, Sandor Rozsa\textsuperscript{a}, Yvan Savard\textsuperscript{d}, Randall J. Sobie\textsuperscript{b}, Thomas Tam\textsuperscript{e}, Ramiro Voicu\textsuperscript{a}

\textsuperscript{a}: California Institute of Technology, Pasadena, CA, USA
\textsuperscript{b}: University of Victoria, BC, Canada
\textsuperscript{c}: BCNET, Vancouver, BC, Canada
\textsuperscript{d}: University of Michigan, Ann Arbor, MI, USA
\textsuperscript{e}: CANARIE Inc, Ottawa, ON, Canada

E-mail: newman@hep.caltech.edu
E-mail: rsobie@UVic.ca

\textbf{Abstract.} A 100 Gbps network was established between the California Institute of Technology conference booth at the Super Computing 2011 conference in Seattle, Washington and the computing center at the University of Victoria in Canada. A circuit was established over the BCNET, CANARIE and Super Computing (SCInet) networks using dedicated equipment. The small set of servers at the endpoints used a combination of 10GE and 40GE technologies, and SSD drives for data storage. The configuration of the network and the server configuration are discussed. We will show that the system was able to achieve disk-to-disk transfer rates of 60 Gbps and memory-to-memory rates in excess of 180 Gbps across the WAN. We will discuss the transfer tools, disk configurations, and monitoring tools used in the demonstration.

1. Introduction

The ATLAS [1] and CMS [2] experiments located at the LHC [3] have accumulated in excess of 100 Petabytes of data since 2010. The analysis of the data from these experiments follows the LHC Computing Model that was initially based on rigid hierarchical structure where Tier 2 centres exchange traffic primarily with their regional Tier 1 centre. Today the LHC Computing model is evolving to an agile peer-to-peer model which makes efficient use of compute and storage resources by exploiting high bandwidth networks [4]. In this model, the Tier 1-2 centers can directly access data from any other center. This new paradigm is being enabled by 100 Gbps networking technology that is being deployed in the cores of major research and education network providers such as ESnet, GEANT, CANARIE and Internet2.

The LHC experiments need to be ready to exploit the current generation of 100 Gbps networks and the coming era of Terabit/s network transfers. In this paper we show how large data sets can be rapidly transferred using high-speed network technologies. In particular, we present the results of a demonstration staged during the Super Computing 2011 (SC11) Conference in Seattle Washington. A 100 Gbps network was established between the University of Victoria (UVic) computing centre and the California Institute of Technology (Caltech) booth at SC11 in Seattle. A bi-directional throughput of 186 Gbps memory-to-memory and a single direction...
throughput of 60 Gbps disk-to-disk (using Solid State Disk (SSD) technology) was achieved during the conference. This throughput was achieved with roughly half a standard 42U 19-inch rack of Linux servers at each location. We describe the network and the server systems used at the Caltech booth and the UVic computing centre. The results presented in this work were obtained in the one week period during the SC11 exhibition.

2. Network Design
A point-to-point 100 Gbps circuit was established between the UVic Data Centre and the Seattle Convention Centre over a total distance of 212 km using production segments of the BCNet and CANARIE networks. Figure 1 shows a schematic of the network. A Brocade MLXe-4 with LR4 100G optic was located in the UVic Data Centre connecting via BCNET to a Ciena OME 6500 located in the Victoria Transit Exchange (VicTX) in downtown Victoria. From there the circuit is carried across CANARIE and BCNET via an OTU4 link to a second OME 6500 located in SCinet network in the Seattle Convention Centre, and then to an MLXe-4 (also with LR4 optic) located in the Caltech conference booth. Each MLXe-4 was equipped with two 8 port 10GE line cards.

3. End Site Systems
The focus of the UVic end system was to provide a stable platform for testing the more experimental end system installed on the exhibition floor at SC. The UVic cluster consisted of 10 identically configured Dell R710 servers each with Intel X520 DA network cards and six
240 GB OCZ SSDs. The six SSD drives were configured in a RAID-0 configuration\(^1\) using a hardware RAID controller with the XFS file system. Each RAID controller was configured with 1 MB stripe size (max available) and write-through algorithm. Scientific Linux 6.1 [6] was configured on the machines with common kernel TCP optimizations [7]: increased TCP buffers, txqueueien increase (10000 from 1000) and htcp congestion control algorithm. The Linux kernel disk IO scheduler was configured to ‘noop’ from the typical ‘deadline’ scheduler. This configuration was found to give optimum performance between servers connected directly to the MLXe-4. The transfers were made using the high performance data transfer tool FDT [8] developed by Caltech and Polytechnica University (Bucharest). Each host pair was able to achieve a throughput that ranges from 9.49 to 9.54 Gbps.

The system deployed to the Caltech booth consisted of a mix of PCIe Gen 2 servers with 10 GE NICs and PCIe Gen 3 servers with 40 GE NICs supplied by Dell and Supermicro as shown in Figure 2. Most of the equipment making up the Caltech booth was available only days before the start of the conference, and some delivered after the start of the conference exhibition. Therefore, limited time was available to benchmark the systems performance before deploying the systems as part of the demonstration. Three Supermicro systems were configured with 40 GE Mellanox CX3 NICs and 16 OCZ 120 GB SSDs (sc32 - sc34 in Figure 2). Four Supermicro storage servers were used with Areca RAID controllers installed and connected with

\(^1\) RAID-0 would not typically be used in a production environment, but given the limited hardware available this was a good choice for maximum performance.
Figure 3. Total traffic at SC week. Traffic In is traffic from UVic to SC and traffic out is traffic from SC to UVic. Total integrated traffic for the week was 4.4 PB. Each colored band represents the contribution from a single machine.

Figure 3. Total traffic at SC week. Traffic In is traffic from UVic to SC and traffic out is traffic from SC to UVic. Total integrated traffic for the week was 4.4 PB. Each colored band represents the contribution from a single machine.

4. Results
The 100G circuit was established from the conference to the UVic data centre on November 13th with first data flowing on the afternoon of November 14th. The program for the exhibition week proceeded over three distinct phases. First, maximum uni-directional memory-to-memory traffic, then bi-directional memory-to-memory traffic and finally disk-to-disk throughput from UVic to SC. The evolution of the data transfers can be seen in Figure 3. Notable features include the start of bi-directional memory-to-memory flow (morning Nov 15th) and the large disk-to-disk flow starting the evening of Nov 16th.

Once the circuit was established we were able to quickly achieve over 98 Gbps sustained uni-directionally over the circuit with no packet drops. All network switching equipment was remarkably stable and presented no problems in configuration. After attempting bi-directional maximum throughput we observed a decrease in throughput to roughly 60 Gbps in (UVic to SC) and 80 Gbps out (SC to UVic). This decrease was eliminated by changing the load balancing algorithm of the 12 port-channel Link Aggregation Group (LAG) between the Brocade MLXe-4 and Dell-Force10 Z9000 to round robin rather than hash based. After this fix we were able to achieve a stable throughput of 98 Gbps in and 88 Gbps out for a combined total of 186 Gbps.

The next focus of experimentation was to use as few as possible 40GE PCIe Gen 3 machines to receive 100 Gbps. As can be seen in Figure 4, data flows were gradually consolidated such that two Supermicro PCIe Gen 3 machines were receiving 30 Gbps each and 2 PCIe Gen 2...
Figure 4. The traffic in section of the plot shows an ever smaller number of machines receiving memory-to-memory transfers. Near 15:30 we see 100 Gbps being received by only 4 machines on the show floor. The two machines each receiving 30 Gbps (top right) are PCI Gen 3 machines with 40 GE Melanox NICs.

Figure 5. Disk-to-disk transfers occurring from UVic to the SC show floor Nov 17-18. Total transfer speed peaked above 60 Gbps around 22:00 Nov 16th. The sharp drops occurring Nov 16 were the results of Linux Kernel panics. The large drop at 4:30 on Nov 17 was the result of the failure of a 16 disk RAID array on PCI Generation 3 machine.

Dell machines were receiving 20 Gbps. The maximum achievable throughput for a PCIe Gen 2 system with 40 GE nic was 24 Gbps (this was also demonstrated by the Caltech team at SC10).

The final and most challenging part of the demonstration was to achieve the maximum
possible disk-to-disk throughput between the two end systems. Figure 5 shows the achieved
disk-to-disk throughput for the 12 hour period starting at 20:30 on November 16th. Each host
at UVic was loaded with 10 GB files to roughly 80% capacity, and a script was established to
copy those same series of files repeatedly to the hosts using FDT client and server with 6 parallel
streams. Total disk-to-disk throughput peaked at 60 Gbps. A number of Linux kernel panics in
the XFS module were encountered early in the transfer. Each machine suffering the panic was
rebooted and the transfers restarted. Caltech booth machines used the 3.0.1 UltraLight kernel [9]
rather than the RedHat provided 2.6.32 series kernels in order to get improved performance with
the latest hardware. The source of the kernel panics was never resolved, but their occurrence
was unsurprising given the relatively untested nature of the kernel and file system combination
at the time. The frequency of the kernel panics was reduced by dropping to two parallel streams
for those systems experiencing the panics. One 16 disk SSD RAID array failed near 04:30
November 17 because of a drive failure. Because the array was in RAID-0 configuration (no
redundancy) the copy operation was unable to continue. One PCIe Gen 3 with 16 SSD was able
to sustain continuously 12.5 Gbps continuous write (orange bar in Figure 5). The performance
of the systems degraded after many repeated writes to the same system.

5. Conclusion
The SC11 demonstration achieved its goal of clearing the way to Terabit/sec data transfers
by showing that a modest set of systems is able to efficiently utilize a 100 Gbps circuit near
100% of its capacity. The latest generation of servers (based on the recently released PCIe Gen 3
standard) equipped with 40GE interface cards and RAID arrays with high-speed SSD disks
allowed the team to reach a stable throughput of 12.5 Gbps from network to disk per 2U server.
A total disk-to-disk throughput between the two sites of 60 Gbps was achieved in addition to 186
Gbps to total bi-directional memory-to-memory throughput. It is important to underline that
pre-production systems with relatively few SSDs were used during this demo, and no in-depth
tuning was performed due to the limited amount of time in the preparation. We therefore expect
these numbers will improve further, and approach the 40GE line rate within the next year.

6. Acknowledgements
The generous support in kind of our industrial partners Ciena, Brocade, Dell, Mellanox,
Supermicro and Color-Chip is acknowledged. We would like to acknowledge the support of the
Natural Sciences and Engineering Research Council of Canada, the National Science Foundation,
and the US Department of Energy.

References
[1] The ATLAS Collaboration et al 2008 The ATLAS Experiment at the CERN Large Hadron Collider JINST
3 S08003 doi:10.1088/1748-0221/3/08/S08003
[2] The CMS Collaboration et al 2008 The CMS experiment at the CERN LHC JINST 3 S08004
doi:10.1088/1748-0221/3/08/S08004
[3] Evans L and Bryant P 2008 LHC Machine JINST 3 S08001 doi:10.1088/1748-0221/3/08/S08001
[4] Bos K and Fisk I 2010 The Bos-Fisk Paper, http://lhcone.web.cern.ch/node/19
[5] Newman H 2011 A New Generation of Networks and Computing Models for High Energy Physics in the
LHC Era J. Phys.: Conf. Ser. 331 012004 doi:10.1088/1742-6596/331/1/012004
[6] The Scientific Linux Distribution www.scientificlinux.org
[7] The ESNet Linux Host Tuning Guide http://fasterdata.es.net/host-tuning/linux/
[8] Maxa Z, Ahmed B, Keira D, Legrand I, Mughal A, Thomas M and Voicu R 2011 Powering physics data
transfers with FDT J. Phys. Conf. Ser. 331 052014 doi:10.1088/1742-6596/331/5/052014
[9] The UltraLight Linux Kernel http://ultralight.caltech.edu/web-site/ultralight/workgroups/network/Kernel/kernel.html