Transition to Trinity: Preparing a Next-Generation Network

Kathryn Protin, Susan Coulter,
Jesse Martinez, Alex Montaño,
and Charles Wilder
Los Alamos National Laboratory
Los Alamos, NM, USA
{ksprotin, skc, jmartinez, afm,
cwilder}@lanl.gov

ABSTRACT
This poster will discuss the work done by the network team in the High Performance Computing Systems (HPC-3) group at Los Alamos National Laboratory (LANL) to prepare a network infrastructure suitable for the Trinity supercomputer. The team transitioned from our previous backbone, Parallel Scalable Backbone (PaScalBB), to the Next Generation Backbone (NGBB) upon Trinity's arrival in June 2015. This poster outlines the need for an improved network infrastructure, the past and future network backbones, and the planning and execution of the transition to the new backbone. In conclusion, we review the challenges faced and lessons learned throughout the transition process.

Categories and Subject Descriptors
C.2.2 [**Computer-Communication Networks**]: Network Architecture and Design – *distributed networks, network topology.*

General Terms
Management, Design

Keywords
network, backbone, infrastructure, file system, supercomputer, high performance computing, Ethernet, InifiniBand, VLAN, switch

1. INTRODUCTION
The High Performance Computing (HPC) group at Los Alamos National Laboratory (LANL) is committed to providing supercomputing support while advancing the lab's nuclear security and weapon science missions. LANL is consistently seen as one of the world leaders in supercomputing, with several machines making the LINPACK TOP500 list, including Roadrunner at number one in 2008 and Cielo in sixth place in 2011 [4]. Beginning in June 2015, LANL deployed its next big project, the Trinity supercomputer. Trinity is scheduled to top the TOP500 list with unprecedented speed—forty or more petaflops— and comes with an 80-petabyte parallel file system [3]. The near future of supercomputing, as evidenced by Trinity, will require an advanced infrastructure that can handle the performance requirements brought on by increased memory and a multi-core architecture. Since LANL's previous network infrastructure was built to support systems one-tenth the speed of Trinity, there was a need to replace this infrastructure with one that had higher bandwidth yet was simpler to maintain.

2. BACKBONES

2.1 Parallel Scalable Backbone (PaScalBB): 2005-2014
The previous network backbone, installed in 2005 and upgraded in 2008, consisted of twelve lane switches, all of which are connected to a single top-level switch. This backbone was known as the Parallel Scalable Backbone, or PasScalBB, and was composed of Force10 chassis switches that supported 10 Gigabits per second (Gbps) bandwidth (Figure 1). Although PaScalBB allowed for enough bandwidth and redundancy to meet past needs, it was not sufficient for Trinity-era computing. Because of the age and number of switches involved in PaScalBB, it was also becoming difficult to maintain.

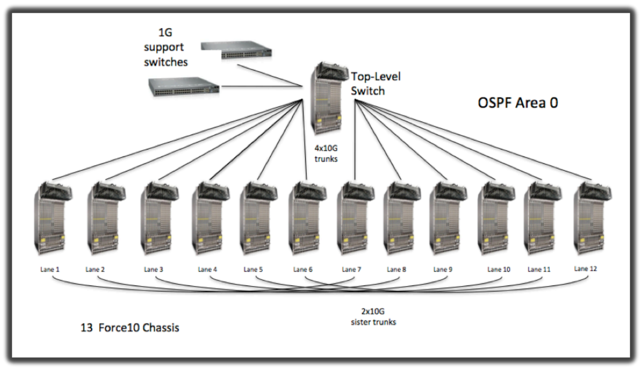

Figure 1. PaScalBB

2.2 Next Generation Backbone (NGBB): 2014-Present
The new backbone design is called the Next-Generation Backbone (NGBB), split into NGBB-Ethernet and NGBB-IB (for InfiniBand). In order to accommodate the Ethernet needs of Trinity-era systems, we used Arista chassis switches with 40 Gbps capability (Figure 2). Previously, the network had many VLANs, which can be considered logical subnets, split based upon cluster membership. The new backbone reorganizes these VLANs into functional groups, divided into admin/monitoring, user, services, configuration management and I/O. This makes the network infrastructure easier to maintain, since similar components are grouped together, and improves network security by having fewer subnets and therefore a simpler configuration.

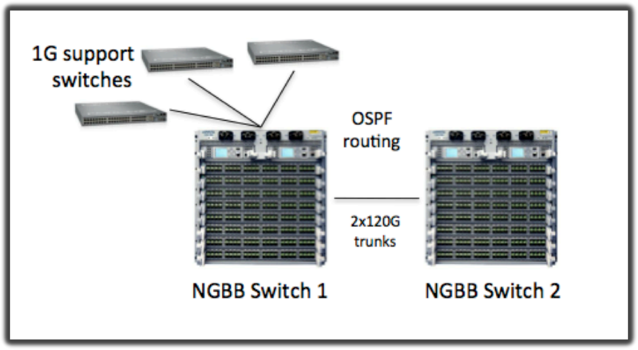

Figure 2. NGBB-Ethernet

The second phase of this project, known as NGBB-IB, depends on InfiniBand, a popular high speed/ low latency interconnect technology. As of November 2014, InfiniBand was the fabric of choice for 225 of the supercomputers in the TOP500 [2]. It facilitates high-speed data transfer and connects the cluster components to file systems including Lustre, the high-speed parallel file system that will be utilized by Trinity [1]. The InfiniBand backbone is currently being installed, with scheduled completion by the end of 2015.

3. PLANNING AND EXECUTION

Planning the network transition began in early 2014, with consideration of how the new backbone should be designed as well as an implementation schedule. The network team needed to accomplish this transition with limited impact on the daily operation of the four secure computing platforms. Each system at LANL has a monthly Dedicated System Time (DST) in which a system is made unavailable to users for routine maintenance. Our goal was to complete this transition without causing any additional outages. As we transitioned between PasScalBB and NGBB, we implemented a temporary solution linking the two backbones (Figure 3). In order for systems on both backbones to communicate, we ran "trunks" composed of four 10 Gpbs fiber optic cables between each lane switch and one of the new Arista switches.

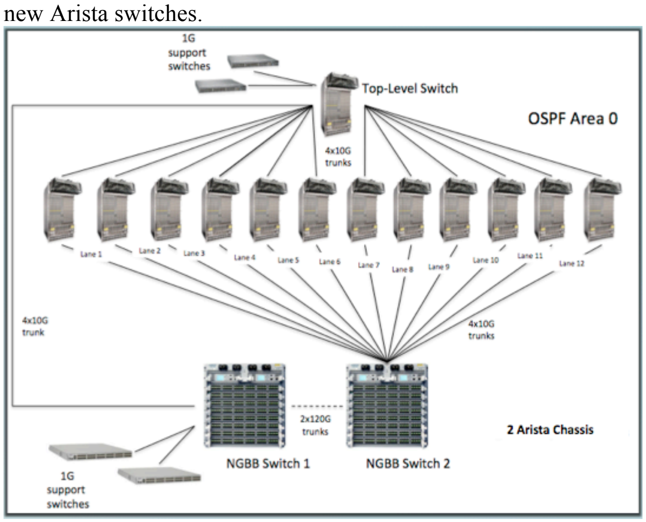

Figure 3. Transition Backbone

In addition to careful planning of the transition backbone, we had to coordinate with other teams within HPC-3, as this work affected the systems they managed. For example, we had to make sure that the system administrators were aware of IP address and subnet changes well in advance of each DST. For each system we moved, a different member of the network team coordinated with cluster and file system administrators, user consultants and facilities managers to plan each outage. We had to work to ensure we weren't causing additional system downtime outside of regularly scheduled DSTs as well as complete each part of the transition before testing the systems and returning to users.

4. CONCLUSION

Throughout the transition, we had several setbacks, including a delay in ordering network supplies and working around the ongoing facilities work. We also learned the importance of communicating with other teams, including system administrators, operators and facilities. We believe that the final configuration of NGBB-Ethernet will meet the needs of Trinity as a next-generation supercomputer with unprecedented speed and memory capacity. Additionally, our simplified logical network structure improves both management and security of HPC resources. The completed NGBB-Ethernet, when paired with the in-process NGBB-IB, will enhance the functionality of Trinity and keep LANL at the top of the high-performance computing world.

5. ACKNOWLEDGMENTS

Thank you to our HPC-3 group leaders, Jeff Johnson, Cory Lueninghoener and Carolyn Connor, and division leaders, Gary Grider and Randal Rheinheimer. This work is approved for release by LA-UR-15-25199.

6. REFERENCES

[1] "About the Lustre File System." Lustre. Seagate Technology LLC. [Online]. Available: <http://lustre.opensfs.org/about/>.

[2] "InfiniBand continues growth among supercomputers on TOP500 List." InfiniBand Trade Association Press Release. InfiniBand, 19 Nov. 2014. [Online]. Available: <http://www.InfiniBandta.org/content/pages.php?pg=press_room_item&rec_id=805>.

[3] N. Hemsoth. "First details emerge from Cray on Trinity supercomputer." HPCWire. Tabor Communications, July 2014. [Online]. Available: <http://www.hpcwire.com/2014/07/10/first-details-emerge-cray-trinitysupercomputer/>.

[4] "Top 500 list," Top 500 Lists. [Online]. Available: <http://www.top500.org/>.