Transition to Trinity: Preparing a Next-Generation Network

Kathryn Protin, Susan Coulter, Jesse Martinez, Alex Montaño, and Charles Wilder
Los Alamos National Laboratory, High Performance Computing Systems Group
{ksprotin, skc, jmartinez, afm, wilder}@lanl.gov

Introduction

The High Performance Computing Systems (HPC-3) group at Los Alamos National Laboratory (LANL) currently manages about a dozen supercomputers dedicated to modeling and simulation projects in the areas of weapons, biology, climatology, and more. 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 an 80-petabyte parallel file system. Trinity, as well as its future peers, will require an advanced infrastructure that can handle the performance requirements brought on by increased memory and a multi-core architecture. The current network backbone is insufficient to handle this performance demand, so in the months leading up to Trinity's arrival, our team completely overhauled LANL's HPC network infrastructure.

Parallel Scalable Backbone: 2005-2014

Trinity is installed in the largest of our server rooms, along with four other supercomputers and several file systems committed to classified computing. The previous network backbone, installed in 2005 and upgraded in 2008, consisted of twelve lane switches, all connected to a single top-level switch, as well as several smaller service switches. This backbone was known as the Parallel Scalable Backbone, or PaScalBB (Figure 1), and was introduced to support the arrival of the Redtail supercomputer. This backbone used Force10 chassis switches that supported 10 Gigabits per second (Gbps) bandwidth.

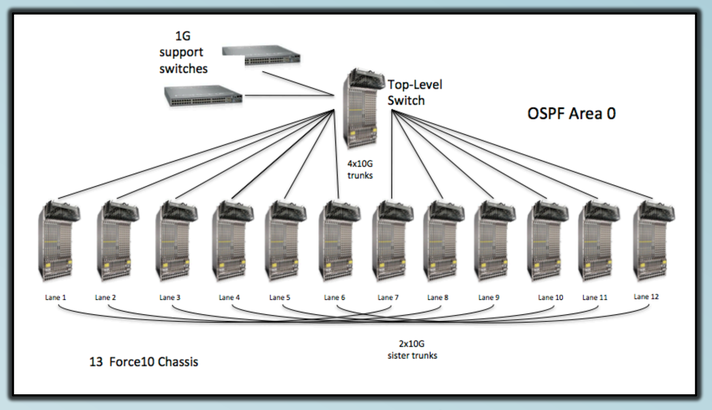

Figure 1: PaScalBB

Next Generation Backbone: 2015-2025 (est.)

The new backbone design is called the Next-Generation Backbone (NGBB), divided into NGBB-Ethernet (Figure 2) and NGBB-InfiniBand (see future work). It uses Arista chassis switches with 100 Gbps bandwidth, with 40 and 10 Gbps breakouts based on network needs. Because of the increased bandwidth, we were able to reduce the twelve lane switches into just two top-level switches. Previously, the network had many VLANs, split up based on cluster membership. In the new design, VLAN membership is based upon both function—admin/monitoring, user, services and configuration management– and system, with each Ethenet cluster's IO nodes in their own VLAN.

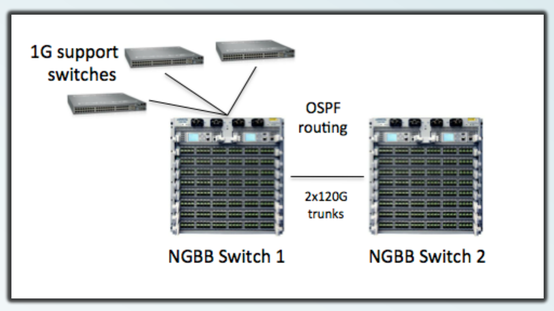

Figure 2: NGBB-Ethernet

Transition Phase

Each of LANL's HPC systems has a monthly Dedicated System Time (DST) in which it is made unavailable to users for routine maintenance. Our goal was to completely replace our network backbone—which was connected to every secure supercomputer and file system—without causing any additional outages. We completed the transition in several phases, moving the section of the backbone that corresponded to each system during its regularly scheduled DST. 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 cables between each lane switch and one of the new Arista switches. We then connected the two Arista switches together and depended on OSPF (open shortest-path first) routing for communication between the two backbones.

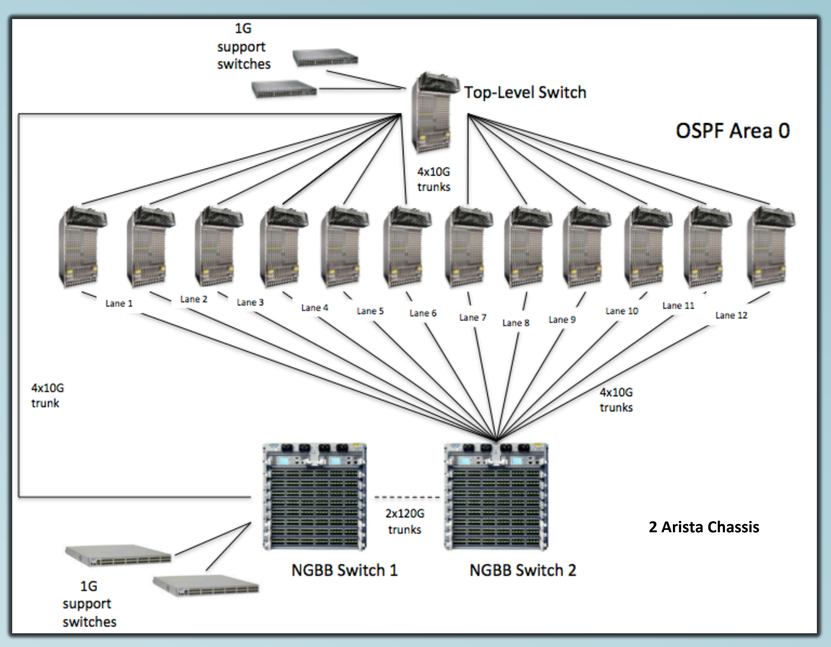

Figure 3: Transition Backbone

During this transition, we utilized our existing patch panels, which connected clusters to switches, and ran new cables to the Arista switches as we removed all connections to the old Force10 switches. During each DST, a member of the network team coordinated with the system administrators to make changes. The network team moved cables and organized the logical structure of the network (Figure 4) as the network information on each system component was updated. Once all changes were complete, the teams worked together to test network connectivity and functionality.

Lessons Learned

The biggest challenge the network team faced during this transition was migrating production supercomputers to a new backbone without causing excessive system downtime. We also had to manage a short deadline because of facilities work in preparation for Trinity, with power to the PaScalBB switches decommissioned only five months after the transition began. Finally, we had to coordinate with other teams within HPC-3, as the backbone transition affected the systems they managed. For example, we had to make sure that the system administrators were aware of IP address and subnet changes (Figure 4) well in advance of each DST.

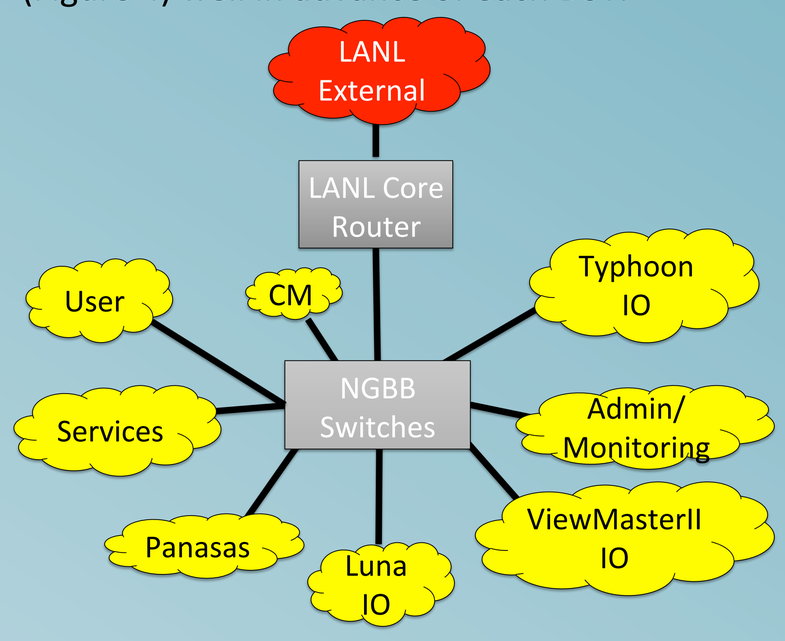

Figure 4: Logical Network Structure

Conclusion and Future Work

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. The increased bandwidth of the Arista switches provides high-speed data transfer between Trinity and its file system in order to minimize data loss. Additionally, our simplified logical network structure improves both management and security as hosts are arranged by both system and functionality. Work on NGBB-InfiniBand is scheduled to begin in Fall 2015, as we replace the current system-specific IB fabrics with a network-wide IB backbone.

Acknowledgements

Thank you to our HPC-3 group leaders, Jeff Johnson, Cory Lueninghoener and Carolyn Connor, and division leaders, Gary Grider and Randal Rheinheimer.

LA-UR-15-25199