FERMILAB-CONF-10-709-CD

The US-CMS Tier1 Center Network Evolving toward 100Gbps

A. Bobyshev, P. DeMar
Computing Division, Fermilab, Batavia, IL 605010, U.S.A.

E-mail: {bobyshev, demar}@fnal.gov

Abstract. Fermilab hosts the US Tier-1 Center for the LHC's Compact Muon Collider (CMS) experiment. The Tier-1s are the central points for the processing and movement of LHC data. They sink raw data from the Tier-0 at CERN, process and store it locally, and then distribute the processed data to Tier-2s for simulation studies and analysis. The Fermilab Tier-1 Center is the largest of the CMS Tier-1s, accounting for roughly 35% of the experiment's Tier-1 computing and storage capacity. Providing capacious, resilient network services, both in terms of local network infrastructure and off-site data movement capabilities, presents significant challenges. This article will describe the current architecture, status, and near term plans for network support of the US-CMS Tier-1 facility.

1. Architecture of US-CMS Tier-1 Network

The architecture of US-CMS Tier-1 network is driven by the experiment's needs. CMS applications depend on very intensive wide-area data movement. Data center networks traditionally were based on a multi-tier architecture with access, distribution and core layers. At the access layer, connectivity for end systems is normally provided with 1Gigabit Ethernet connections, although 10 Gigabit Ethernet connections are beginning to appear. The distribution layer provides aggregation for access layer devices, and the core layer interconnects distribution layer devices. The devices at different layers are normally interconnected by one or several 10GbE links, typically with relatively low traffic utilization [3]. That architecture serves well for traditional client-server applications. Major traffic flows of these applications go vertically, from clients at the access layer to servers on other access layer devices. Interaction between clients or between servers is minimal. Oversubscription is not

Operated by Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy.

considered as an issue in such environment. The access layer could be oversubscribed as much as 12:1, the distribution layer up to 8:1 and the core up to 4:1. However, modern applications such as cloud computing and high-impact science data movement applications, utilize network bandwidth much more aggressively, and frequently can run up to 100% utilization over a long period of time [3]. Traffic flows may be fully meshed with any-to-any patterns going not only vertically between layers but also horizontally within same layer. Oversubscription becomes an issue in such environment. There is a great deal of analysis in this emerging trend, with new design solutions are being offered by the major network equipment vendors [3], [4], [5]. In brief, these trends could be summarized as consolidating multiple layers of legacy multitier infrastructure into fewer layers, and correspondingly reducing oversubscription at each layer as much as possible, ideally down to 1:1.

In the CMS computing model, analysis depends on intensive data movement between applications. Since its initial deployment in 2005, the network architecture of US CMS Tier1 has been based on the two layer approach depicted in Figure 1. The access layer provides Layer2 connectivity to end systems, typically connected by 1GE or 2 x1GE. The percentage of 10GE host connections is still very low, but we anticipate sharp growth starting in 2011. The aggregation layer connectivity is non-blocking to the aggregation hubs within data center or to external subnets. This architecture of the US-CMS Tier-1 network has been in place since the Tier-1's initial deployment, although the network platforms in use have evolved over last five years.

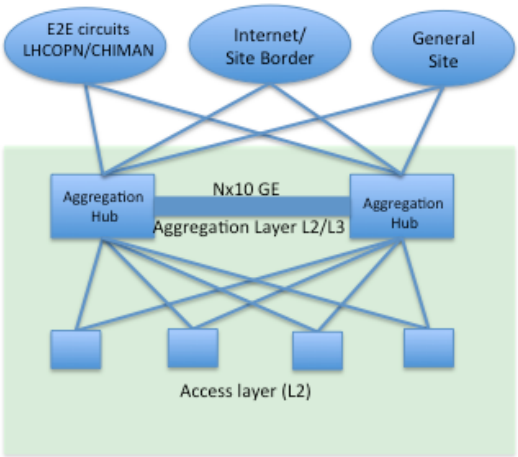

Figure 1: Architecture of US CMS Tier1 Network

2. Status of US-CMS Tier-1 network in 2010-11

In 2009, the aggregation hubs of US-CMS network were upgraded from the Cisco Catalyst 6509E platform to the Cisco Nexus 7000. The upgrade was completed via several intermittent steps, using temporary connections, to avoid disruption of production LHC activities. In 2010, the main focus was on providing a stable operation of LHC data movement. This objective was achieved using network protocols, such as Virtual Port channel (vPC) and Gateway Load Balancing Protocol (GLBP), to support redundancy of connectivity and load balancing traffic across multiple links. All servers and critical end systems are now connected by 2x 1GE links to two different access switches, eliminating single points of failure. Required network maintenance can be conducted without disruption, at the high data rates of production traffic.

The USCMS Tier 1 facility maintains high capacity links for off-site connectivity and to the EnStore tape robotic complex. We have established two 20GE channels to the general campus network, where the EnStore facility is connected. While the CMS data model [1][2] specifies a minimum of 3Gbp/s of available bandwidth for moving data into/out of Enstore, this requirement is based on the anticipated sustained (24x7) rate of pre-processed data from the Tier-0. The aggregate bandwidth capacity of the CMS Enstore mover nodes is approximately 20Gbp/s today. Combined with the 20Gbp/s of general off-site network bandwidth capacity, we believe that overall 40GE of bandwidth provisioned to the general campus should provide enough capacity to avoid any potential bottlenecks for next 2 - 3 years. In addition, another 20GE channel extends to the Laboratory perimeter router that supports the CMS virtual data circuits to CERN and a number of other CMS Tier-1/Tier-2 centers. Figure 2 depicts the current USCMS Tier 1 network, including a few in-progress bandwidth upgrades. By beginning of 2011 all uplinks from the access switches to Nexus 7000 aggregation hubs were upgraded to 8x10GE. The deployment of the second generation Nexus 10GE modules that provide 32 10GE ports and 230Gbps switching fabric connections per slot enabled us to implement these upgrades. The central hardware in the aggregation hubs is fully redundant, with inter-switch channels distributed across multiple 10GE modules that allow maintenance of modules without disruption of service.

When we migrated to the Nexus 7000 platform, we adopted an asymmetric scheme for provisioning bandwidth to access switches in proportion 7:1 or 3:1. In other words, for an (n)x10GE channel of an access switch, (n-1) 10GE links are connected to primary aggregation hub and the remaining one to secondary aggregation hub. This configuration allows use of the primary Nexus 7000's non-blocking switching fabric as the core network fabric for the entire Tier-1. The vPC technology enables use of all 10GE links in this configuration, instead of keeping some links in standby mode.

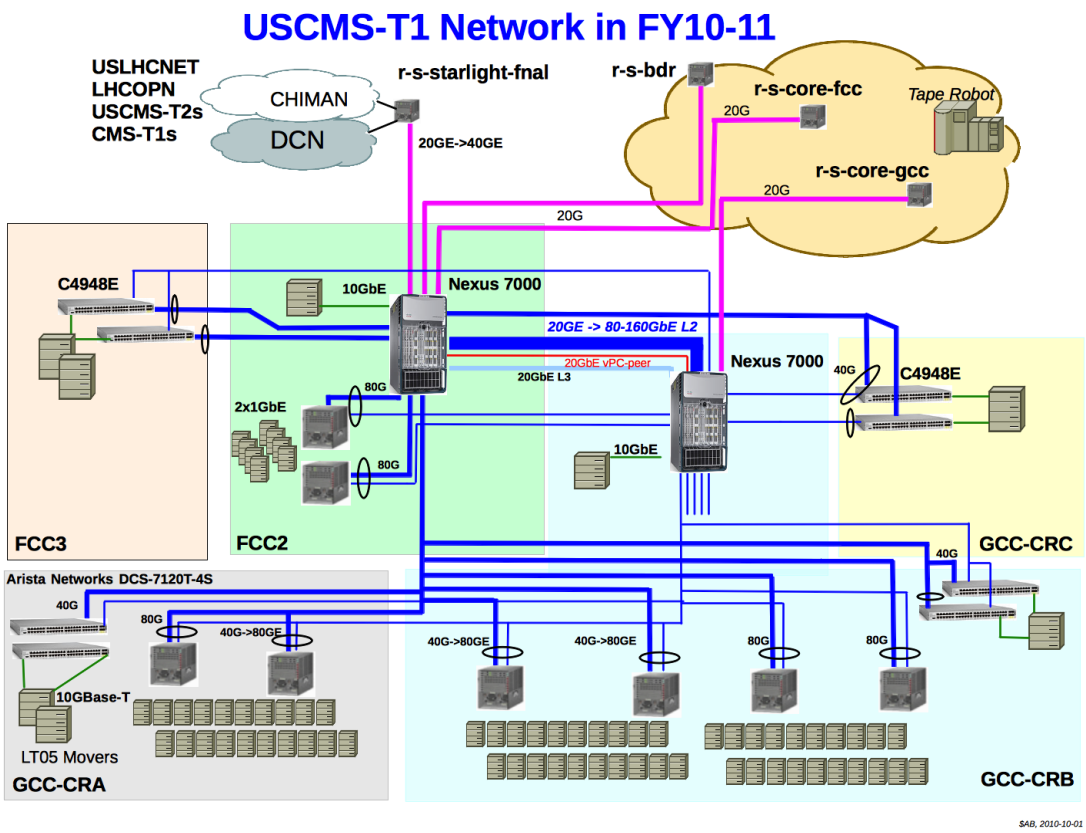

Figure 2: USCMS-Tier1 Network in FY10-11

Due to CMS's increasing data processing requirements, demand for bandwidth within the local area network continues to grow. Currently, there are about 1200 worker nodes within the Tier-1, connected to network across seven Cisco C6509 switches. In 2010, we initiated upgrade of that infrastructure to 80GE per Cisco 6509E chassis, leaving 288 worker nodes per switch. This decreased the level of oversubscription from 8:1 down to 3:1. Due to continual changes within the network and to CMS computing at the Tier1 facility, it would be difficult to provide an accurate quantitative number for improved applications' performance from decreased oversubscription levels. Deployment of new servers, retirement of obsolete equipment, application modifications, and changes in job processing, have all impacted the traffic patterns and loads within the Tier-1 LAN. However, two graphs (figures 3 and 4) should give an indication of the level of performance improvements.

The first plot (figure 3) displays utilization loads that we typically observed during data movement runs at the access layer with 40GE provisioned bandwidth per a switch and 8:1 oversubscription. The network is shown operate with an almost fully saturated channel for an

extended period of time. This level of saturated LAN would have definitely a negative impact on CMS data movement and analysis times.

The second plot (figure 4) shows a typical utilization pattern after the uplinks for the Tier-1 access switches were upgraded by 8x10GE, decreasing the oversubscription level to 3:1. This graph demonstrates better LAN conditions for distributed application performance.

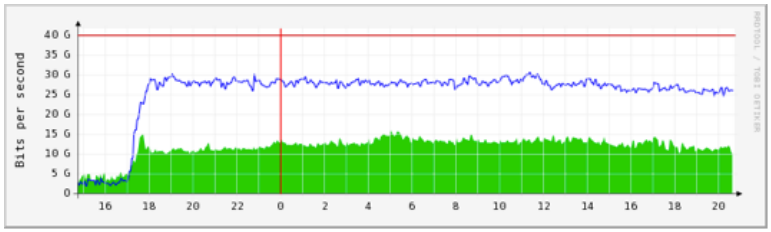

Figure 3 Utilization at the access layer before upgrades

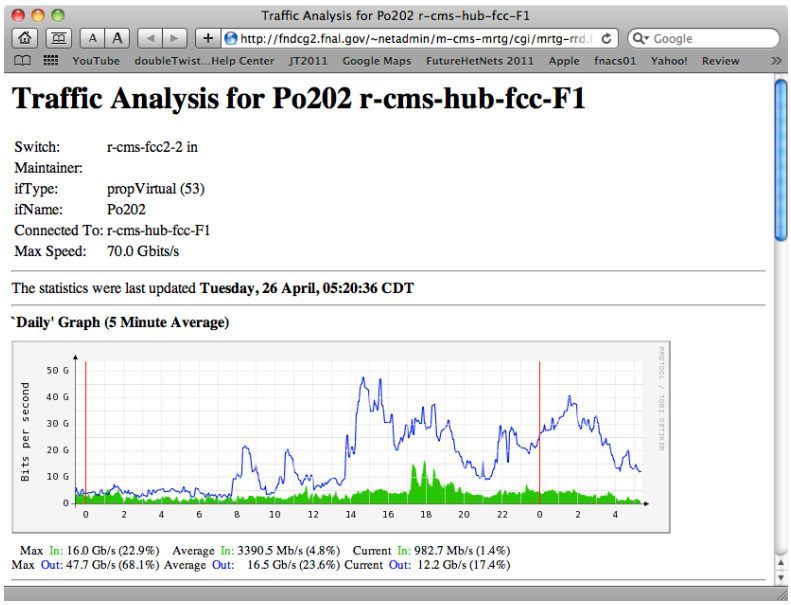

Figure 4 Utilization at the access layer after upgrade

3. QoS at US-CMS Tier1 Network

Data movement between file server nodes and job execution (worker) nodes is capable of saturating any link within the Tier-1 network. Sustained network congestion inevitably leads to packet loss. While the impact of packet loss on the job submissions is merely lengthened job processing times, the impact on other types of Tier-1 network traffic is far more severe. Examples of performance sensitive types of traffic include data being read from or written to tape, middleware control and monitoring traffic, and interactive traffic. To address these concerns, we have deployed QoS within the Tier-1 LAN. All CMS traffic is divided into several classes, based on source IP address/destination IP address tuple. Each class is guaranteed a portion of a total bandwidth. The USCMS QoS model is depicted in figure 5. The critical traffic has source/destination address pairs from Tier0/CERN and to/from a tape robot system. It will use about same portion of total LAN bandwidth as is provisioned for offsite connectivity.

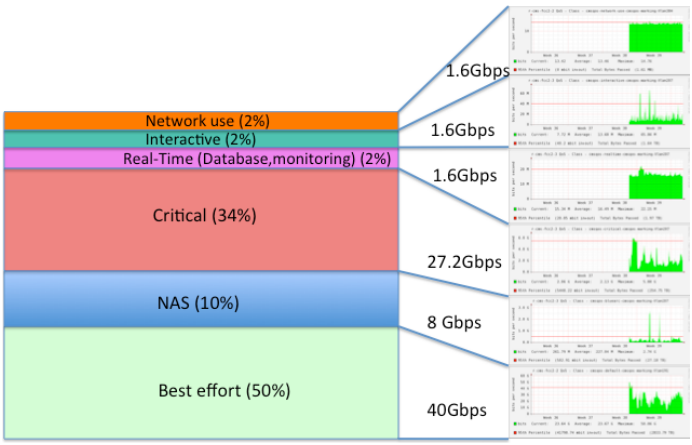

Figure 5: QoS Classes at USCMS-T1 Network

4. Work in progress

4.1. End-to-End dynamic circuits

At the US-CMS Tier1, we have deployed end-to-end circuit technology to setup virtual data circuits to the Tier-0 Centre (CERN), as well as a number of Tier-1 and Tier-2 sites. However, majority of these circuits are static or created on a long-term basis. It is anticipated that Tier-3 sites will play significant role in analysis of the LHC data. Tier-3s are typically physics groups at universities. It may be too difficult or expensive for these groups to establish and operate the network infrastructure to establish virtual data circuits, particularly if the circuits require static bandwidth allocation. Dynamic circuits may be a solution for data movement by Tier3s. Establishing dynamic circuits on demand of applications is a very challenging task for end sites because it requires modification of network configurations on-the-fly. Dynamic circuits could be requested for the duration needed to move CMS experiment data without impacting the university's general network traffic. Several R&D projects, such as TeraPaths [13], Lambda Station [12], Phoebus [14] have explored this area. The completed

Lambda Station project resulted in a deployment evaluation of dynamic circuits at US CMS Tier1, and was used for move production data to the Tier2 facility at the University of Nebraska.

4.2. Data centre monitoring approaches

In complex networks, such as the USCMS-Tier1 Facility, detecting connectivity and performance problems is a very challenging task. Modern network technologies, such as Layer2/Layer3 multipath and vPC (virtual port channel), support redundancy and load sharing, as well as bring the benefits of efficiency and network resilience. Such benefits come at the expense of network complexity. This complexity makes it difficult to detect and troubleshoot problems when the network does not perform as expected. For several recent years, we have observed pair-wise behavior with connectivity and performance problems. Performance or even connectivity can be normal between certain nodes and not acceptable between adjacent nodes that use the same network path. Such problems are often intermittent, making them even more difficult to troubleshoot and fix. To ensure the USCMS Tier1 network supports the required network performance of its applications, there is the need to deploy new monitoring approaches, techniques and infrastructure that will enable any production node to be a monitoring element. The primary objective of our on-going investigation is to design techniques, develop monitoring systems, and build basic infrastructure to detect, troubleshoot and resolve pair-wise connectivity and performance problems. The system we are investigating would allow any production host to become an active measurement element, capable of self-detecting and alarming on connectivity and performance problems. The proposed system would combine passive analysis of netflow data to detect performance anomaly in order to trigger, justify and initiate active measurements. If a problem is confirmed by active measurements, the system would alert these issues to a central management station. In an ideal case, the system would maintain any-to-any connectivity and performance matrix. In practice, scope will probably need to be limited to provide performance matrix between various network clouds or areas according to logical description of the system's main application data flows. In addition, such a system would need to be aware of network and end system load. Preliminary investigations are proceeding to identify a rough timeline, effort required, and available resources needed to build such a system. We are considering a simplified version of such a system that could be built on existing tools within the Tier1 network. Initial efforts are estimated at about 1 Full Time Equivalent (FTE), with 20 percent to design effort and 80 percent to software development for a period of one year.

4.3. 10GBase-T to end systems is on rise

Starting in 2011, we anticipate significant demand for 10GE connections to end systems. In order to be ready for such demands, we have been evaluating 10Gbase-T Ethernet technology. Two different solutions were considered and investigated. One involved hosts with a 10GE connection directly to

Nexus 7000 aggregation switch. The second solution targeting end systems located too far away for direct 10GE connection to a Nexus. In this instance, we evaluated a small top-of-rack switch, Arista Networks' DCS-7120T-4S with 20 x 10GBase-T ports and 4x 10GE SFP+ uplinks. On the host side, we deployed the Intel E10G41AT2 10Gigabit AT2 Server adapter.

Evaluations of products described above have demonstrated that our physical and network infrastructure is capable of handling 10GE systems directly connected to a Nexus. A critical issue with a top-of-rack solution for 10GE hosts is oversubscription concerns. A typical oversubscription of 5:1 (4x10GE uplinks for twenty 10GE host connections) is sufficiently high to raise performance-impacting congestion concerns. In 2011, we expect to continue to evaluate remote top-of-rack solutions for 10GE-connected hosts, while proceeding with direct 10GE connections to the Nexus 7000 aggregation switches.

Acknowledgments

Many people at Fermilab are involved into support of US-CMS Tier1 networking. The list of people who made significant contributions is very long. We would like to mention at least some people who are involved on a daily basis. Vyto Grigaliunas provides support for off-site connectivity and end-to-end circuits. Wenji Wu provides data analysis and wide-area troubleshooting of performance problems. Maxim Grigoriev provides support for the monitoring infrastructure. Orlando Colon leads the effort to provide the physical infrastructure of the US CMS Tier 1 network.

References

[1] Challenges for the CMS Computing model in the First Year, I.Fisk, 17th International Conference on Computing in High Energy and Nuclear Physics(CHEP09), IOP Publishing, Journal of Physics, Conference Series 219(2010)072004

[2] Documenting BW requirements for use cases from the Computing Model at CMS-Tier1s, M.C.Sawley, ETH Zurich, 18 June 09, http://indico.cern.ch/getFile.py/access?contribId=4&sessionId=3&resId=0&materialId=slides&confId=45475

[3] Architecting Low Latency Cloud Networks, Arista Whitepaper http://www.aristanetworks.com/media/system/pdf/CloudNetworkLatency.pdf

[4] Cloud Networking: A Novel Network Approach for Cloud Computing Models, Arista Whitepaper http://www.aristanetworks.com/media/system/pdf/CloudNetworkingWP.pdf, http://www.aristanetworks.com/en/blogs/?p=311

[5] Next-Generation Federal Data Center Architecture, Cisco Systems Whitepaper, http://www.cisco.com/en/US/solutions/collateral/ns340/ns517/ns224/net_implementation_white_paper0900aecd805fbdfd.html

[6] Data Center Architecture Overview, Cisco Systems design guide, http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_Infra2_5/DCInfra_1.html

[7] Juniper Data Center LAN Connectivity, http://www.juniper.net/us/en/solutions/enterprise/data-

center/simplify/#literature
[8] Data Center LAN Migration Guide Juniper Network Design toward two tier acrchitecture, http://www.juniper.net/us/en/solutions/enterprise/data-center/simplify/#literature
[9] An overview and implementation of New Data Center Technology. Steven Carter, Cisco Systems, NLIT summit 2010, https://indico.bnl.gov/conferenceDisplay.py?confId=258
[10] End Site Control Plane Subsystem, https://plone3.fnal.gov/P0/ESCPS/
[11] P.Demar, A.Bobyshev, M.Crawford, V.Grigaliunas, Use of alternate path WAN circuits at Fermilab, CHEP07, Victoria BC, Canada, 2-7 September 2007
[12] Lambda Station: Production Applications exploiting advanced networks in data intensive High Energy physics, CHEP06, TIFR, India, 13-17 February 2006
[13] Terapaths: Configuring End-to-End Virtual Paths with QoS Guarantees, https://www.racf.bnl.gov/terapaths
[14] Phoebus: High-performance Data transfer for Dynamic Circuit Networks, http://e2epi.internet2.edu/phoebus.html