FERMILAB-CONF-12-285-CD

Status and Trends in Networking at LHC Tier1 Facilities

A Bobyshev[1], P DeMar[1], V Grigaliunas[1], J Bigrow[2], B Hoeft[3] and A Reymund[3]

[1] Computing Division, Fermilab, Batavia, IL 60510, U.S.A.

[2] Information Technology Division, Brookhaven National Laboratory, Upton, NY 11973, U.S.A.

[3] Steinbuch Centre for Computing, Karlsruhe Institute of Technology, Hermann-von-Helmoltz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

E-mail: bobyshev@fnal.gov, demar@fnal.gov, vyto@fnal.gov, big@bnl.gov,
Bruno.Hoeft@kit.edu, Aurelie.Reymund@kit.edu

Abstract. The LHC is entering its fourth year of production operation. Most Tier1 facilities have been in operation for almost a decade, when development and ramp-up efforts are included. LHC's distributed computing model is based on the availability of high capacity, high performance network facilities for both the WAN and LAN data movement, particularly within the Tier1 centers. As a result, the Tier1 centers tend to be on the leading edge of data center networking technology. In this paper, we analyze past and current developments in Tier1 LAN networking, as well as extrapolating where we anticipate networking technology is heading. Our analysis will include examination into the following areas:

- Evolution of Tier1 centers to their current state
- Evolving data center networking models and how they apply to Tier-1 centers
- Impact of emerging network technologies (e.g. 10GE-connected hosts, 40GE/100GE links, IPv6) on Tier-1 centers
- Trends in WAN data movement and emergence of software-defined WAN network capabilities
- Network virtualization

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

1. Introduction

This paper has been jointly written by networking staff at three Tier1 facilities, the US-CMS Tier-1 Center at Fermilab, the US-Atlas Tier-1 Center at Brookhaven National Laboratory and the multi-LHC experiment Tier-1 Center at Karlsruhe Institute of Technology.

The LHC distributed data processing and analysis computing system is arguably the largest and most complex such system ever created for large-scale science. It has been a resounding success. The LHC Tier1 Centers, with their capacious local and wide area network facilities, are a cornerstone for that success. The Tier1 centers can now count almost three years of high-quality production service. There is an extremely high level of cooperation between the Tier1s on Wide-Area networking issues. That cooperation includes intensive activities and communication on the LHC Optical Private Network (LHCOPN) by the Tier1s, US-LHCNet (the provider of transatlantic connections for LHC community in the U.S.), ESnet (the U.S. Department of Energy Science Network), Internet2 (an advanced networking consortium for the U.S. research and education community), GEANT (the pan European data network dedicated for the research and education) and national and regional R&E network providers throughout the world. However, there is no comparable coordination among Tier1s on Local Area Networking issues. Yet Tier1 LANs should be considered as a sort of "last mile problem" in the challenging task of moving LHC data around the Globe and being made available for analysis. Each Tier1 may have its own specific issues, but we all face similar challenges, tasks, and requirements. This paper is an attempt to look into current status of LAN networking at several Tier1s, trying to examine commonalities or differences.

Our initial intent was to involve a larger number of Tier1 facilities, but we had to reduce our ambitions and focus on a smaller number first. The authors of this paper do not claim that any our findings and observations are applicable for all Tier1 facilities. When we use the term "LHC Tier1 facilities" we assume that it should be clear to readers that is meant to be applicable to the three sites participating in this analysis.

We identify several topics related to Local Area Networking that we intend to review from prospective of current status and trends:

- Network architectures
- Access solutions
- Networking for 10G end systems, 40/100G inter-switch aggregation
- Unified fabrics, Ethernet fabrics, new architectures enabled with new data center technologies
- Network virtualization
- Software-Defined networks
- IPv6

2. Status

Network architecture is the foundation that determines the overall design of any network. Over the past decade, computing has gone through significant transformations while networking has gone through only significant capacity increases, without any corresponding architectural changes. However, that appears to be changing now. All major players in the network industry realize the need for new architectures to bring networking in sync with new technologies. Yet, there is no clear understanding what this new architecture should look like [5],[6],[7],[8],[10],[11].

At a very basic level, network architecture is often described in terms of layers, with a different network function assign to each layer. The classic network architecture is based on a 3-layer design with access, distribution and core layers (figure 1). We are now observing a shift in design of data center and campus networks toward a two-layer model, with consolidation of the core and distribution layers [9]. This transition is largely driven by application needs for greater bandwidth and lower latency, as well as increased agility in configuration and reconfiguration of data center clusters and systems.

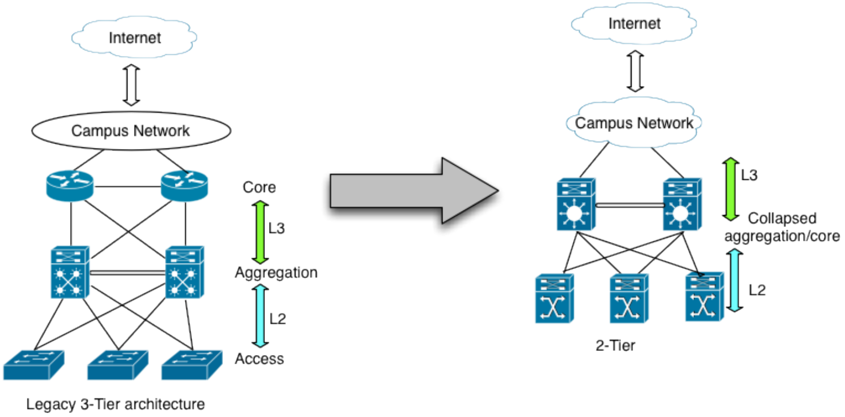


Figure 1. Transition of network architectures

2.1. Network architectures

At LHC Tier1 centers, 2-layer designs has been commonly implemented from very beginning. Such designs provide optimal performance for LHC applications, with their very high LAN data rates. A typical architecture of an LHC Data Center is depicted in figure 2.

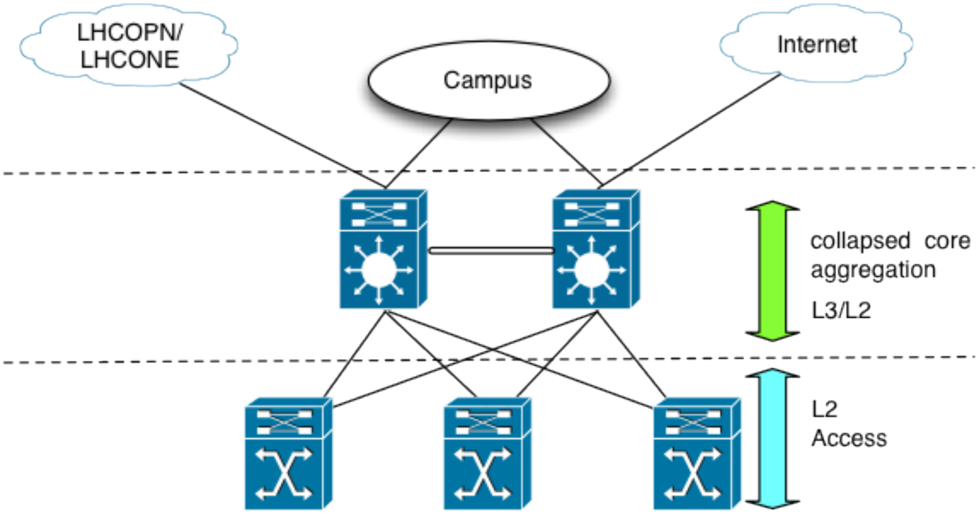


Figure 2. Typical architecture of LHC Tier1 center

Different sites may label the layers differently. In this paper we will refer to them as aggregation and access layers. The aggregation layer provides both Layer2 forwarding of traffic within the same subnet and the Layer3 function of routing between different subnets. The aggregation layer is composed of aggregation switches and the links interconnecting them. The number of aggregation switches is determined by several factors, including distributed locations with computing resources, the number of access devices that need to be connected to the aggregation layer, and redundancy requirements. The aggregation layer is normally designed to be non-blocking. At an LHC Tier1 center, the aggregation layer usually has direct connection to private wide-area network paths, such as the LHCOPN, as well as a connection to the site core. If security policies allow, some sites also support some direct connections to site perimeter, bypassing the site's core network.

The access layer provides network connectivity (layer 2) for end systems. A vast majority of the end systems at LHC centers are connected at 1GE. At some sites, server nodes, typically dCache or other cache file service systems, are connected at 2x1GE, and the number of 10G-connected servers is growing rapidly.

2.2. Access Solutions

Data center top-of-rack (ToR) access solutions have been widely deployed over the past several years. In brief, a small access switch with 24 – 48 1GE connections and up to 4x10G fiber uplinks is installed in each rack. If the density of systems in each rack is not too high, one access switch can service several racks. All connections within the racks are patched by UTP cables. Access switches are connected then to a consolidation switch at end of each row of racks. An access solution based on ToR switches is commonly used at LHC data center as well. However, the deployment scenario differs from the one normally recommended by vendors. At LHC centers, ToR switches are directly connected to the aggregation layer, without an intermediate consolidation switch (figure 3). The standard ToR deployment recommended by vendors raises several points of concern for LHC data centers. A consolidation switch at the end of row is a potential oversubscription point. A consolidation switch will be typically connected to the aggregation layer at 'n' x 10GE, where 'n' is normally between '1' and '8'. This limits the total bandwidth of the uplink to 80Gb/s. At LHC data centers there might be up to 30 servers installed in one rack. Thus, as few as 3-ToR switches could seriously oversubscribe the consolidation switch's uplink. In the network-intensive environment of LHC-computing, this could well be a common occurrence. Directly connecting the ToR switches to the aggregation layer at 'n' x 10GE ('n' being between '1' and '4') avoids the consolidation switch bottleneck. Under these conditions, ToR is an efficient solution for LHC Tier1 Centers, even when connecting dCache nodes via 2x1G.

The largest component of computing resources at LHC centers are the analysis farms systems. For connecting analysis nodes, a large chassis switch, such as the Cisco 6509, Dell Force10, Brocade BigIron and MLX, are most often used (figure 4). A big switch can provide enough 1G ports to connect up to 300 servers, with 80G uplinks to the aggregation layer. Big chassis swiches provide a higher density of 1G ports and support more efficient data movement because of their extremely capacious switch fabrics. Analysis systems have burst traffic characteristics that permit some level of oversubscription. Typically, oversubscription levels of 3:1 or even 4:1 are acceptable.

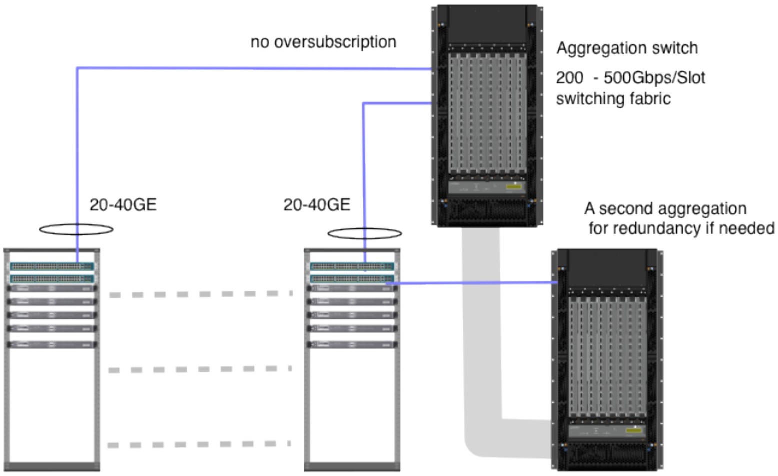


Figure 3. Connecting ToR access switches at LHC Tier1s

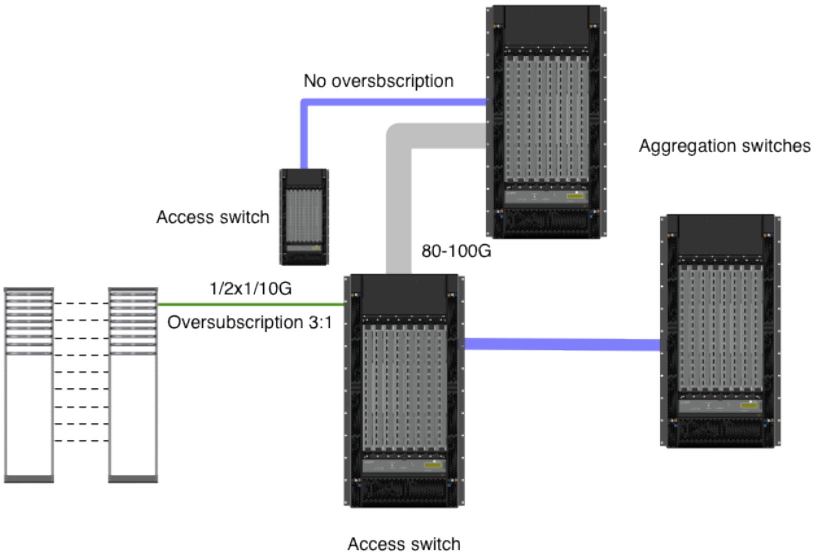


Figure 4. Connecting server farms to a large chassis switches

2.3. Bandwidth provisioned at LHC data centres

LHC applications are very dependent on highly efficient data movement over the local network. From a networking prospective, we identify four major entities for LHC data movement model. These entities are:

- long term storage (tape)
- disk caches to stage data for moving to other entities
- analysis farms
- WAN.

A typical bandwidth provisioned at LHC data center is depicted in figure 5. At an LHC data center, there is a goal to reduce oversubscription to the very minimum, 1:1 in the ideal case. Oversubscription levels 3:1 or 4:1 may be considered acceptable for data movement into and out of analysis servers. Dcache nodes, on the other hand, are connected via 2x 1GE or 10G, and generally without any oversubscription. In practice, it is not always possible to follow the structure depicted in figure. For example, it might be necessary to co-locate dcache nodes with analysis servers. In such cases, additional measures can be taken to provide the desired characteristics for designated traffic.

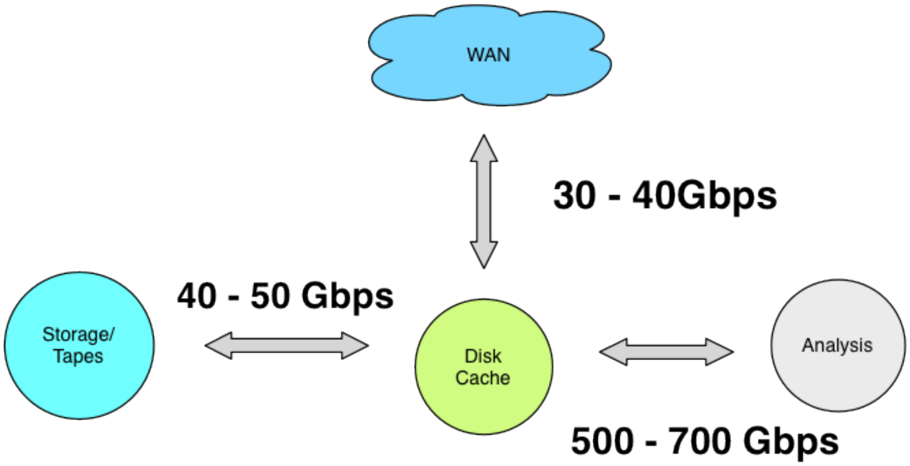


Figure 5. Typical bandwidth provisioned at LHC Tier1 LAN

These measures include QoS and logical partitioning of devices with independent uplinks.

Back in 2008 – 2009, when the LHC was just starting up, the capacity of the switches available at that time was often fully utilized. The graph below (figure 6) shows a typical utilization of a LAN channel during data movement. Utilization levels were frequently at 100%, for extended periods of time. High utilization of LAN channels often affected the efficiency of data analysis.

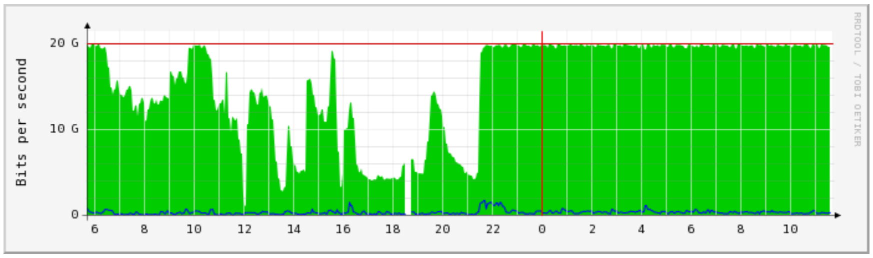
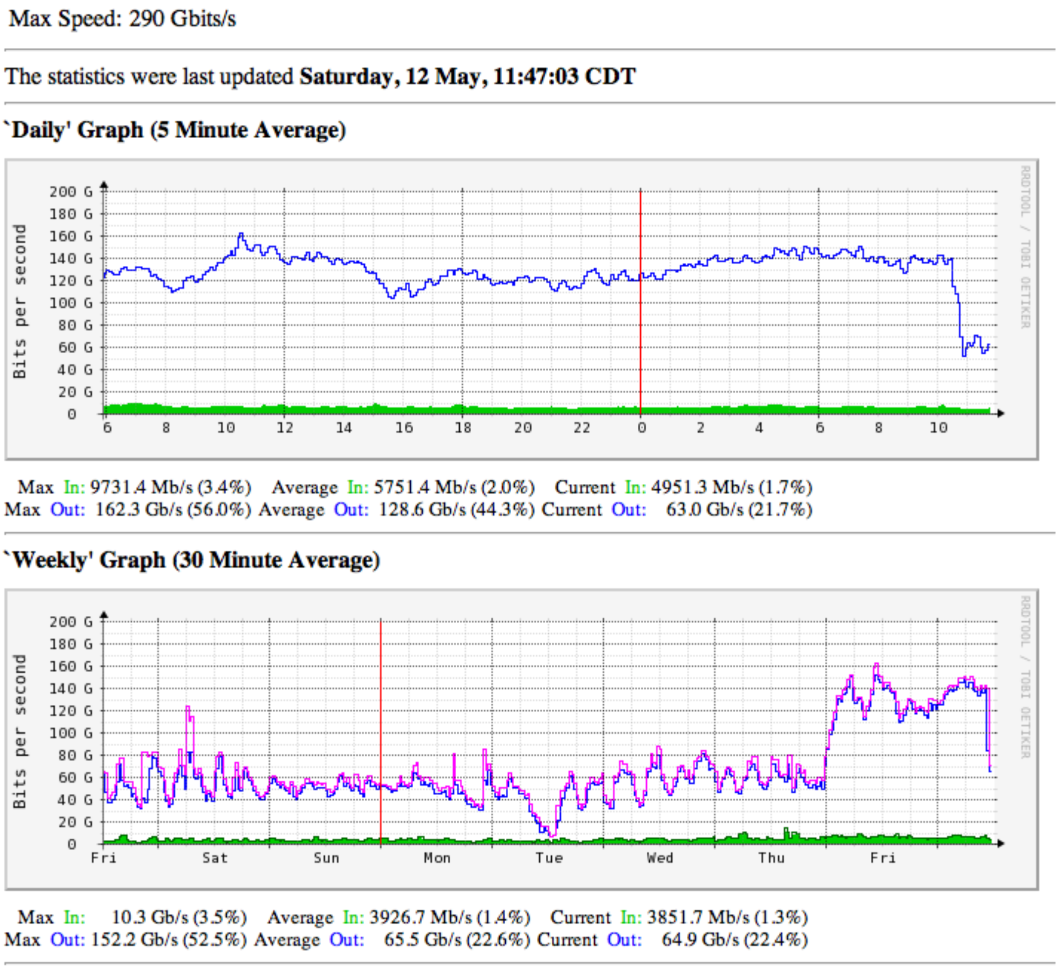


Figure 6. Utilization of a typical LAN channel in 2008 - 09

In the same time period, the next generation switching platforms were becoming available on the market. These switches provided higher capacity, with 200 – 500Gbps/slot bandwidth per line card. Upgrades to these platforms allowed provisioning of adequate bandwidth for LHC local data movement. The graphs in figure 7 show typical utilization in 2011-12. As we can see, local network bandwidth capacity moved well ahead of application needs. On the graphs below, utilization is at 40 – 50% of 300Gbps of provisioned bandwidth



Max Speed: 290 Gbits/s

The statistics were last updated **Saturday, 12 May, 11:47:03 CDT**

`Daily' Graph (5 Minute Average)

Max In: 9731.4 Mb/s (3.4%) Average In: 5751.4 Mb/s (2.0%) Current In: 4951.3 Mb/s (1.7%)
Max Out: 162.3 Gb/s (56.0%) Average Out: 128.6 Gb/s (44.3%) Current Out: 63.0 Gb/s (21.7%)

`Weekly' Graph (30 Minute Average)

Max In: 10.3 Gb/s (3.5%) Average In: 3926.7 Mb/s (1.4%) Current In: 3851.7 Mb/s (1.3%)
Max Out: 152.2 Gb/s (52.5%) Average Out: 65.5 Gb/s (22.6%) Current Out: 64.9 Gb/s (22.4%)

Figure 7. Utilization of a typical channel in 2011-12

The current situation with provisioning bandwidth well ahead of utilization levels might not last long. 10G Ethernet is become available on server motherboards at very affordable cost. This means that future servers will likely have 10G Ethernet ports by default. To accommodate this emerging demand, several solutions for connecting of 10G servers are under consideration. For very small deployments, ToR switches could be a viable options, since small switch with n x 40GE uplinks are becoming available on the market.

In case of very wide deployment of 10GE-connected systems, big data center switches, such as the Cisco Nexus 7000, or Brocade MLX, would be more suitable solutions.

3. Trends

3.1. Converged or unified fabrics.

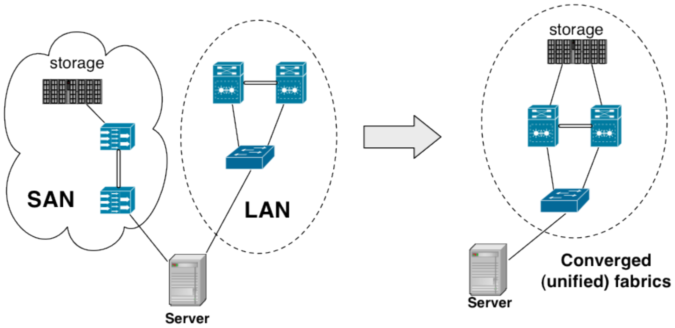


Figure 8. SAN/LAN converging fabrics

Disk storage is an important element for LHC data analysis. Until recently two independent infrastructures, SAN and LAN, were usually created for separate storage and Ethernet networks. Data center network infrastructure now support lossless Ethernet. This opens up the possibility for a single infrastructure to support both LAN and SAN. This is a popular trend in generic data center networking.

However, LHC Tier1s do not seem very interested in this capability at present. Storage at LHC data centers is created as a federated file system, based on individual LAN servers with disks attached via fibre channel (FCS). Each disk system can support up to 200TB using current technology (figure 9).

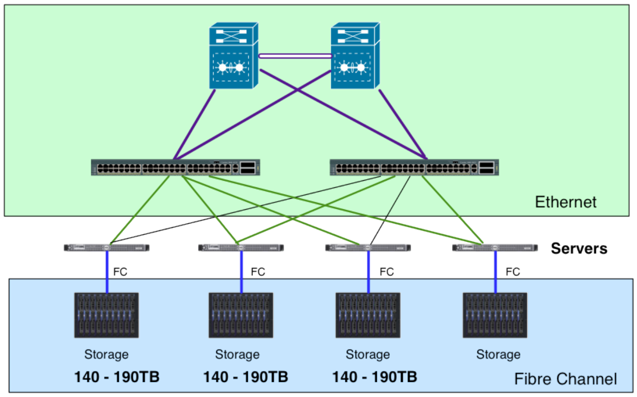


Figure 9. Architecture of disk storage at LHC Tier1

3.2. Connecting 10G servers

The majority of servers within LHC Tier1 centers are 1G or 2x1G-attached systems. The cost of a 10G connection for both 10GBase-T and 10GBase-SX is dropping rapidly. We believe that we have reached a point where the cost of 10G-connected servers is affordable and economically efficient for small (10 – 50) and medium (50 – 120) servers deployments. One limiting factor is the absence of suitable switching solutions with 10GBaseT ports for servers. Fortunately, the marketplace is rapidly changing, and we anticipate a much broader spectrum of such products on the market by the end of this year or beginning of the next year. At that point, we expect 10GE-connected servers to become the norm at LHC Tier1 centers.

3.3. Network Virtualization

Almost everything nowadays can be virtualized. Until recently, virtualization was mainly associated with virtualized computing, where a single system runs multiple instances of logically-independent operating systems in parallel, with each one providing a separate computing environment. In networking, VLAN technology is considered a standard LAN feature, rather than a virtualization technique. Similarly, VPNs are simply considered a solution for remote access to LAN. In actual fact, both are really network virtualization technologies.

The situation is now changing. Next generation products from networking vendors are introducing new capabilities for LAN virtualization. For example, Virtual Device Context from Cisco Systems Inc., consolidates multiple physical devices in a single logical. It can also partition a single physical device into multiple logical devices, each with its allocation of required resources. Cisco FabricPath (Cisco's implementation of Ethernet fabric) enables creation of multiple Layer2 networks on the same physical infrastructure, each customized for its specific tasks and provisioned with adequate bandwidth. LAN virtualization now can mean many things beyond the well-known VLAN technology such as:

- Multi-tenant capability enables a single physical network to support multiple logical networks with each logical network appearing to be its own physical network to its customers
- Converged fabrics – using a single physical infrastructure that is based on lossless Ethernet to support both SAN (Fibre Channel) topology and LAN-based Ethernet. Converged fabrics must provide guaranteed delivery because SANs are extremely loss-sensitive.
- Virtualized access layers provide a foundation for Virtualized Computing, storage and service mobility
- All virtualized services are provided adequate bandwidth provisioning for any-to-any connections.

LHC Tier1 facilities are investigating these new virtualized technologies. Below are a few ways that network virtualization could benefit Tier1 centers:

- Application of LHC-centric security, routing or other policies per a logical network basis rather than on per interface or devices basis
- Better segmentation and management of networking resources
- Facilitates different levels of services
- Provides mechanisms to isolate traffic per application, user groups or services
- Computing resources can be distributed across multiple computing rooms, while still treated as being co-located in same one.

3.4. Software Defined Networks

Interest in Software Defined Networks (SDN) has increased in the past several years [19]. The term Software Defined Networks means a capability to programmatically change the network behavior on-demand of the applications or management software. The changes are intended to be dynamic, covering a specified period of time or while certain networking conditions exist. The spectrum of potential uses for SDN is very broad. At LHC Tiers1, a primarily interest in SDN is to establish dynamic end-to-end virtual circuits between sites and control the site's traffic on to and off of these circuits.

The dedicated Layer2 end-to-end circuits between Tier1s and Tier0 and between Tier1s play a significant role in efficient data movement of LHC data across the Globe. The LHCOPN is the cornerstone for LHC data movement, and represents the prototype for this type of statically configured virtual private network. However, the LHCOPN approach does not scale to the hundreds of LHC Tier2 and Tier3 facilities that exist. Short-lived, dynamic network connections would be necessary for virtual private network-based data movement to those facilities. The capability for such services is now emerging. In the US, circuit service is provided to LHC community by ESNet (OSCARS)[15] and Internet2 ION [16]. In Europe, an automated bandwidth on-demand service called AutoBAHN is offered by GEANT[17]. This service is currently available in the US and in Europe.

Today, end-to-end (E2E) data circuits usually terminate at a site's perimeter. These circuits can be requested via a Web-GUI for a short or long period of time. In practice,Tier1 facilities typically create circuits and leave them permanently established because of the difficulties in setting them up and tearing them down. To have E2E services be truly dynamic, several essential features are still missing. The major issues are

- Lookup service
- Network model
- Service negotiation between sites
- Circuit monitoring

When sites initially connect to an SDN service provider, the first questions that network administrators will ask is what are other sites can be reached through their SDN service. Currently, this is manually-acquired knowledge. The availability of a lookup service, with an API capability, is still not in place. LHC Tier1s today are restricted to following the manual knowledge path in order to participate in E2E circuit services

If network infrastructure is going to be modified dynamically, we will need an abstract representation or a model of the local network that can be easily understood. The main challenge here is not just abstract representation or open API for a single networking device, such as OpenFlow configuration of a router or a switch, but an abstract representation of network as a complex system.

End-to-End data movement between two sites that are linked by that circuit requires coordinated configuration, as well as synchronized setup and tear down of the circuit by each site. If one site fails to create its circuit segment with proper characteristics, or at the proper time, the connection will perform badly, or not at all. Last but not least, circuits need to be constantly monitored for status and performance. To date, monitoring capabilities do not provide a quick enough response time to reroute traffic without performance degradation.

From very beginning, LHC Tier1 facilities have been involved in early deployment of SDN services, and participated in circuit-related R&D projects. In 2008, Fermilab in collaboration with California Institute of Technology completed the LambdaStation project [13]. A prototype system has been deployed at the US CMS Tier1 facility and used to recover production data to the Tier2 center at University of Nebraska Lincoln.. Terapaths is a project headed by Brookhaven National Laboratory to explore End-to-End circuits creation across WAN with QoS guarantee[14]. In Europe, DE-KIT is working on AutoKNF project[18], which is based on AutoBAHN tool offered by GEANT to implement a bandwidth-on-demand service from end customer prospective.

While interest in SDNs continues to grow, the technology is still not mature enough for production deployment in local area networks at Tier1 facilities. It will likely take several more years of broader support from network vendors before the technology reaches the point for production deployment in Tier1 LANs.

3.5. IPv6

There are several ongoing IPv6 activities at institutions hosting LHC Tier1 centers. These activities are primarily driven by campus networking needs, not LHC experiments requirements. In the US, the Federal Government has mandated support for native IPv6 on all public-facing services by end of FY2012. Scientific computing, such as LHC computing, is not within the scope of that mandate. As a result, while there is an ongoing IPv6 support effort at the US facilities hosting LHC Tier1s, it does not directly involve the Tier1s themselves. However, as IPv6 becomes more widely deployed within these facilities, it is likely that IPv6 support within the Tier1s will occur naturally, as an extension of generic TCP/IP support.

At DE-KIT, some services are available worldwide over IPv6 (e.g. www.kit.edu) and IPv6 is fully deployed over the campus network in production mode available for every customer who needs an IPv6 presence. Some features are still not available from the vendor, but workarounds where configured and makes it possible.

Especially in the LHCOPN, DE-KIT deployed BGP for IPv6, and exchanges IPv6 prefix information with CERN. For the moment, only one test node speaks IPv6 at the German T1, but the protocol is ready to be used. DE-KIT participates in the HEPIX and the EGI IPv6 working group.

4. Site News

In this section we would like to present very recent news from each Tier1 facility participated in this analysis.

- BNL completed upgrade of its Perimeter Defense Network from 2 6509's to a pair of Nexus 7010 to provide a path for 100GE from the perimeter inward
- DE-KIT has deployed at the Core two Nexus 7010's. A border router upgrade is planned for provisioning of a 100G capability for external links, such as the LHCOPN
- DE-KIT has deployed 50 additional 10G fileservers for matching the experiment expectations of storage accessibility
- Fermilab has reached an agreement with ESnet to establish a dedicated 100G link to the new 100G-based ESnet5 network. This project is planned to be completed in the summer of 2012. Fermilab deployed one hundred ten 10G servers, primarily dCache nodes and tape movers

5. Conclusion

In this paper we presented the summary of analysis LAN networking at LHC Tier1 facilities. This analysis is based on the practical experience operating networks at three LHC centers, the US-CMS, the US-ATLAS in the U.S. and the DE-KIT in Germany. Facing similar challenges we found lots of commonalities in our approaches to design and operate our networks.

We would like to be more proactive in planning network resources in our Tier1 centers. For this, we need to understand where we are at any particular time, what we might need to anticipate on requirements, technology progress, and so on. Exchange of information and ideas between LHC facilities might be useful. If other Tier1/2/3s are interested to join this informal forum, feel free to contact any authors of this paper.

6. References

[1] What does the NEW architecture look like ? Network World Magazin. http://www.networkworld.com/community/blog/what-does-new-architecture-look

[2] 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

[3] 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

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

[5] 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

[6] 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

[7] 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

[8] Juniper Data Center LAN Connectivity, http://www.juniper.net/us/en/solutions/enterprise/data-center/simplify/#literature

[9] Data Center LAN Migration Guide Juniper Network Design toward two tier acrchitecture, http://www.juniper.net/us/en/solutions/enterprise/data-center/simplify/#literature

[10] Dell Virtual Network Architecture, http://en.community.dell.com/techcenter/networking/w/wiki/3489.dell-virtual-network-architecture.aspx

[11] An overview and implementation of New Data Center Technology. Steven Carter, Cisco Systems, NLIT summit 2010, https://indico.bnl.gov/conferenceDisplay.py?confId=258

[12] Brocade Campus LAN Switches: Redefining the Economics of Effortless Campus Networks, White paper, Brocade Communications Systems, Inc. http://www.brocade.com

[13] LambdaStation project http://www.lambdastation.org

[14] Terapaths project http://www.racf.bnl.gov/terapaths

[15] ESnet OSCARS project, http://es.net/oscars

[16] Interoperable on-demand network service, Internet2 ION, http://internet2.edu/ion/

[17] Automated Bandwidth Allocation across Heterogeneous Networks (AutoBAHN) http://www.geant.net/service/autobahn/pages/home.aspx

[18] AutoKNF- “End customer” AutoBAHN deployment, https://tnc2012.terena.org/core/presentation/32

[19] Open Networking Foundation, https://www.opennetworking.org/