US LHCNet: Transatlantic Networking for the LHC and the U.S. HEP Community

Grant Number: DE-FG02-08ER41559

Final Report

April 5, 2013

Abstract

This is the final report for the US LHCNet project, covering the period 2009-2012.

Principal Investigator: Harvey B. Newman, Professor of Physics
California Institute of Technology
Mail Code 256-48, Pasadena, CA 91125
626-395-6656; newman@hep.caltech.edu

Project Summary

US LHCNet provides the transatlantic connectivity between the Tier1 computing facilities at the Fermilab and Brookhaven National Labs and the Tier0 and Tier1 facilities at CERN, as well as Tier1s elsewhere in Europe and Asia. Together with ESnet, Internet2, and other R&E Networks participating in the LHCONE[1] initiative, US LHCNet also supports transatlantic connections between the Tier2 centers (where most of the data analysis is taking place) and the Tier1s as needed. Given the key roles of the US and European Tier1 centers as well as Tier2 centers on both continents, the largest data flows are across the Atlantic, where US LHCNet has the major role.

US LHCNet manages and operates the transatlantic network infrastructure including four Points of Presence (PoPs) and currently six transatlantic OC-192 (10Gbps) leased links. Operating at the optical layer, the network provides a highly resilient fabric for data movement, with a target service availability level in excess of 99.95%. This level of resilience and seamless operation is achieved through careful design including path diversity on both submarine and terrestrial segments, use of carrier-grade equipment with built-in high-availability and redundancy features, deployment of robust failover mechanisms based on SONET protection schemes, as well as the design of facility-diverse paths between the LHC computing sites.

The US LHCNet network provides services at Layer 1(optical), Layer 2 (Ethernet) and Layer 3 (IPv4 and IPv6).

The flexible design of the network, including modular equipment, a talented and agile team, and flexible circuit lease management, allows US LHCNet to react quickly to changing requirements form the LHC community. Network capacity is provisioned just-in-time to meet the needs, as demonstrated in the past years during the changing LHC start-up plans.

[1] LHC Open Networking Environment

Table of Contents

1 Introduction and US LHCNet Mission

Wide area networking is mission-critical for HEP, and the dependence of our field on high performance networks continues to increase rapidly. The expanding network needs of the major physics collaborations have resulted in a rapidly advancing scale of network bandwidth requirements for the field, as fully documented in a series of workshops and studies led by ESnet[2], and the studies and annual reports of the ICFA Standing Committee on Inter-regional Connectivity[3].

This trend has been accelerated over the last few years by the adoption of grids spanning several world regions by the major HEP experiments, and rapid advances in network technologies making the use of multiple "10 Gbps" links[4] over national and transoceanic distances increasingly affordable, and cost-effective. The exponential growth in network use that our field has experienced over the last 15-20 years is expected to continue over the next few years, driven by the convergence of three major factors: (1) the LHC program's ongoing increases in energy and luminosity and the experiments' data taking rates, (2) the continued trend towards computing system components, architectures and software geared for higher bandwidth data transfers, and (3) an ongoing worldwide shift to optical and wireless networks with an order of magnitude greater capacity, supporting more bandwidth-intensive real-time applications.

The longer term outlook is that this exponential growth in use and the corresponding requirements will continue for at least the next ten years, driven by the accumulation of LHC data from the hundreds-of-petabytes range now, to the exabyte range; by continued advances and generational changes in computing, networking and communications technologies; and by specific developments and the rapid emergence standards that respond to the worldwide explosion in the demand for bandwidth.

The effectiveness of US participation in the LHC experimental program is particularly dependent on the speed and reliability of our national and international networks. The achievements of the LHC experimental Collaborations throughout the recently completed three year run at 7 and 8 TeV, including the groundbreaking discovery of a new Higgs-like particle near 126 GeV, are rooted in the ability of physicists at CERN in the US and elsewhere overseas to reliably move many petabytes of data, to access computing and data storage resources, and to collaborate in real time from multiple remote locations. The entire LHC program, and US involvement in it, thus depended, and continues to depend, on unprecedented levels of network performance and reliability.

US LHCNet has been designed to meet these needs by providing a high performance network aiming at 99.95+% service availability, through the use of multiple links across the Atlantic, network equipment that provides robust fallback at the optical layer in case of link failure, and automatic re-direction of network traffic using redundant network equipment at each of the US LHCNet points of presence (PoPs). In order to support its mission, and to enable the US LHC community to make best use of its network resources, the US LHCNet team works to

[2] See Science Requirements for ESnet Networking, at http://www.es.net/hypertext/requirements.html

[3] See http://cern.ch/icfa-scic.

[4] What is often referred colloquially to as "10Gbps link" offers in fact a bandwidth depending on the technology. SONET OC-192, used in WAN connections offers a capacity of 9.4 Gbps to the OSI link layer.

progressively adopt the most cost-effective new network technologies as needed; it continually deploys and integrates state of the art high-throughput methods and tools, working in partnership with the HEP labs, ESnet, the major research and education networks in the US, Europe and Asia, and advanced networking projects funded by DOE/MICS and NSF. Notable examples (discussed later in this report) since 2009 include the UltraLight, PLaNetS, DYNES and ANSE projects funded by NSF, the OliMPS project funded by DOE/OASCR and the LHCONE project originated by Caltech and CERN and now under development and being deployed by all the major R&E network providers and advanced network projects supporting HEP, most notably US LHCNet.

The Network Requirements Workshop[5] organized by DOE/SC HEP and ESnet in August 2009 identified several important requirements, one of them being the US Tier1 sites' networking needs for HEP related traffic. The transatlantic network capacity requirements have been quantified in the bandwidth roadmap (that of US LHCNet) specified in the RFI sent out by OHEP and OASCR in August 2010, where the application throughput to be supported was projected to grow from 48 Gbps in 2010 to 320 Gbps in 2015, assuming well-tuned applications capable of using more than 80% of the theoretical link capacity.

The Workshop on Transatlantic Networking for LHC Experiments[6], jointly organized by CERN and US LHCNet in June 2010 revealed an ongoing evolution in the experiments' computing models, giving greater roles to the Tier2 and Tier3 facilities located at sites throughout the world. As reported at that workshop, the CMS data traffic between Tier2s and Tier1s as well as among Tier2s themselves increased sharply, by factors of 3.5 and 6.7 respectively, at the onset of LHC operation in March 2010. Robust data transfers, with data set transfers complete in a useful time frame, such as 4 hours latency to move a dataset to a Tier2 site as expressed by ATLAS, are important for efficient operation of the experiments, and show both the bandwidth requirements as well as the need for high network service availability.

These exponential traffic growth trends continued throughout the recently completed LHC run, where the data transported by USLHCNet on behalf of ATLAS, CMS and the other LHC experiments reached several tens of petabytes in 2012 alone.

1.1 US LHCNet Design

US LHCNet has been designed to meet these needs, by providing a high performance network with a service availability target of 99.95+%. In order to achieve this unprecedented availability level in a transoceanic network, US LHCNet has deployed multiple links across the Atlantic and cross-links in the US and in Europe, network equipment that provides seamless fallback at the optical layer in case of link failure, and automatic re-direction of network traffic using redundant network equipment at each of the US LHCNet points of presence (PoPs). In order to enable the US LHC community to make best use of its network resources, US LHCNet also works to progressively adopt the most cost-effective new network technologies as needed.

[5] http://workshops.es.net/2009/hep-net-req/wiki/bin/view/HEPNetReq/WebHome

[6] http://indico.cern.ch/conferenceDisplay.py?confId=88883

US LHCNet's robust architecture, together with its equally robust underlying monitoring and management services infrastructure, have allowed US LHCNet to successfully meet the requirements. In terms of availability of at least one OC-192 payload bandwidth, US LHCNet is able to meet the 99.95% availability target. As described in the project status section of this report, this has been made possible by (1) a physically diverse set of cost-effective transatlantic links and continental cross-links, (2) routing and switching as well as optical multiplexing equipment with sufficient redundancy, (3) extremely robust autonomous monitoring and other services supporting the infrastructure, that have been shown to provide seamless, non-stop production operations in the presence of individual or multiple link outages, and (4) a close-knit engineering team with a complete set of skills developed during years of successful transatlantic network field operations.

The ability to deliver the required level of reliability, and seamless operations spanning single and sometimes multiple link outages, is due to the expertise, skills, and years of training of the team, as well as the underlying services architecture that has been developed, deployed and proven in years of successful field-operations by the Caltech team together with CERN and their partners.

The US LHCNet network architecture continues to evolve, making best use of the most cost-effective current-generation equipment and state of the art service technologies. This allows the team to continue to provide the best suitable services in support of the US LHC program, including bandwidth on demand services using dynamic circuits, IPv6 Layer 3 services and soon higher capacity next generation 40 Gbps (40G) and 100G optical as well as Ethernet services. Within the next three years, US LHCNet will replace its present core equipment (much of which has been deployed starting in 2007), and transition to the current generation Optical Transport Network (OTN) ITU standard network paradigm and protocols, including support of higher capacity transatlantic circuits. The first 100G production services across the Atlantic are expected to appear in 2013, following demonstrations already scheduled this June in which the US LHCNet team will take an active part. The transition to the next generation of production services in US LHCNet, including the migration to the next generation optical switching platform, have been carefully planned, and the first design and planning steps in this direction have already been taken in 2011-12.

Based on years of experience in the procurement and deployment of transoceanic and continental links, the construction and implementation of services ensuring seamless operations in the presence of link outages, and the development of inter-domain dynamic circuits and high throughput applications in collaboration with CERN and our HEP and major network partners in the US and Europe, US LHCNet delivers the most cost effective and robust solutions to meet the requirements, and will continue to do so over the next five years, with the next-generation US LHCNet implementation.

1.2 Role of the Caltech Group

The Caltech group first proposed the use of international networks for high energy and nuclear physics (HEP) research in 1982, and has had a pivotal role in transatlantic networks for our field since then. Our group was funded by DOE to provide transatlantic networking for L3 and the other LEP experiments ("LEP3NET") starting in 1986, based on earlier experience and incremental funding for packet networks between the US and DESY (1982-1986). From 1989 onward, the group has been charged by DOE with providing US-CERN networking for the HEP community, and mission-oriented transatlantic bandwidth for many of HEP's major programs.

In December 1995, Caltech and CERN formed the "USLIC" US Line Consortium to fund a dedicated CERN-US line. For the last 18 years, the network has been co-managed and co-operated by the CERN and Caltech network engineering teams. Since November 2006, Caltech and CERN have shared management and operations responsibility for the "US LHCNet" consortium, with Caltech having the primary responsibility for management and operations of the transatlantic and intra-U.S. links among the points of presence in New York, Chicago, CERN and Amsterdam, and for the equipment and its maintenance at the three points of presence outside of CERN.

Starting in the Spring of 2006, Caltech also took over responsibility for the US LHCNet Requests for Proposals issued annually, which are intended to minimize the costs of the network, following a multi-year staged implementation plan that foresees a substantial increase in bandwidth each year at a small to moderate increase in cost. Wherever possible, the plan exploits favorable long-term trends in market pricing per unit bandwidth, especially along the highest capacity transatlantic routes.

1.3 Team Activities

The main activities of the Caltech team are:

- ***Operations and Support:*** Our primary focus is the operation and management of a reliable, high-performance network service 24x7x365. This activity includes equipment configuration, configuring and maintaining the routes and peerings with all of the major research and education networks of interest to high energy physics, as well as monitoring, troubleshooting, and periodic upgrades as needed. The interaction with the physics computing groups and the strict monitoring of the performance of network transfers, as well as solving various requests and trouble tickets from the users and integration with the LHC OPN, are also a part of this activity.
- ***Pre-production Development and Deployment:*** We maintain a "pre-production" infrastructure available for network and grid developments. To keep up with the rapid ongoing emergence of new, more cost effective network technologies, we continually prepare each year for the production network of the following one or two years, by testing new equipment and evaluating new technologies and moving them into production. This ongoing process includes (1) demonstrating and in some cases optimizing the reliability and performance of new architectures and software in field tests for short-term developments, and then (2) completing longer-lasting production-readiness tests prior to release of the new technologies as part of the next-round production service. We typically use major events such as the annual Supercomputing conferences (including SC04-SC12) for large scale demonstrations associated with longer-term planning and developments, where we have also benefited from large-scale vendor support and major R&E network support to minimize costs.
- ***Technical Coordination and Administration:*** Technical coordination includes day-to-day oversight of the team's Operations and Development activities, and technical responsibility for project milestones and deliverables. The number of partners and the variety of network-intensive activities by the LHC experiments and associated Grid

projects that use our transatlantic network also require an excellent degree of coordination. The Caltech and CERN network teams also have a central role in the planning, evaluation and development of new transatlantic network solutions in cooperation with our partner teams at the DOE labs, Internet2, ESnet, National Lambda Rail, and leading universities (Michigan, Nebraska, Vanderbilt, University of Florida, FIU, MIT and many others), as well as the research and education network teams of Canada (CANARIE), the Netherlands (SURFnet), GEANT and other international partners.

The administration activity also includes negotiating contracts with telecom providers and hardware manufacturers, formulating and managing the annual Requests for Proposals for the US LHCNet circuits, and maintaining and periodically renewing the contracts for equipment maintenance, network interconnections and peerings (where these entail charges) and collocation of our equipment at our New York, Chicago and Amsterdam PoPs.

Until the Spring of 2006, CERN was responsible for RFPs and contract negotiations for the transatlantic circuits. Since DOE is the major contributor to US LHCNet, Caltech has taken over this responsibility and negotiated directly with the telecom operators.

- ***Management, Planning and Architectural Design:*** This covers (1) overall management of the team, its year-to-year evolution, training and professional development (2) developing and implementing the strategy and planning for US LHCNet operations and technical development in consultation with our partners, (3) examining technology options, making design choices, and developing the architecture and site-designs (at Starlight, MANLAN, Amsterdam and CERN) for the next upgrade of the network (4) tracking and evaluating current requirements for transatlantic networking, and preparing roadmaps projecting future requirements, with input from the HEP user and network communities, while also taking current and emerging technology trends into account, (5) preparing funding proposals and reviews, reviewing and updating technical coordination plans including the major milestones, (6) developing relationships and joint R&D programs with partner projects, as well as leading network equipment and circuit vendors as appropriate, to maximize the overall benefit to the U.S. and international HEP community within a given funding envelope.

1.4 External Collaborations

During the 3 year grant period, US LHCNet was actively involved in several collaborations, notably the DICE group formed by ESnet, Internet2, GEANT, CANARIE and US LHCNet. The DICE collaboration focused on operational aspects relating to transatlantic networking. A common services portfolio, including IP, dynamic circuits and performance monitoring has been worked out. In 2010, the Caltech team together with CERN originated and helped form the LHC Open Networking Environment, LHCONE, which provides a new form of collaboration between networks as well as the (from network perspective) user community represented by the LHC experiments, and the LHC computing sites.

GLIF, the Global Lambda Integrated Facility, in which US LHCNet participates, has as its goal the promotion of the lambda-networking paradigm for data intensive science. The GLIF resources include GLIF Open Lightpath Exchanges (GOLEs), open lambda exchange points, where R&E networks can interconnect. US LHCNet is collocated with 4 GOLEs, and uses them as exchange points to interconnect with some of its partners: MANLAN in New York, Starlight in Chicago, NetherLight in Amsterdam, and CERNLight in Geneva.

The use of these exchanges provides the only scalable model for large collaborations, with the LHC networking community being a prime example, as it eliminates the need for a full-mesh connectivity among the peering partners. Interconnectivity between R&E networks through lightpath exchanges is today the de-facto mainstream direction, with many of the European Cross-Border Fibers terminating on an exchange point. Together with dynamic bandwidth allocation as promoted and worked on in the GLIF community and the Network Services Interface (NSI) group of the Open Grid Forum (OGF), this model fits best the requirements regarding traffic LHC data flows. It is a strategic direction in US LHCNet to interconnect with its partner networks at GLIF GOLEs. This is implemented in most cases already, and is becoming an increasingly widely adopted direction worldwide as LHCONE continues to advance.

US LHCNet works closely together with ESnet to provision the capacity needed between CERN and the US Tier1s. This includes the matching of capacity, as well as common engineering decisions where the end-to-end paths involve both domains. ESnet provides the last-mile connectivity between the US LHCNet PoPs at Starlight and MANLAN and the US Tier1s (FNAL and BNL, respectively). The US LHCNet and ESnet teams have in collaboration engineered the explicitly path-diverse and PoP-diverse backup circuits to each of the US Tier1 centers, with US LHCNet providing the transatlantic segment from CERN, and ESnet the terrestrial segment in the US. This collaboration at all levels is important for the support of the US LHC program, and has proven to be a very effective operational arrangement in delivering an overall network service with the required capacity and reliability.

Internet2 is the principal partner in the US for Tier2 and Tier3 connectivity, as it provides the most extensive national backbone for Research and Education networking in the US, together with National Lambda Rail. Many of the LHC Tier2 and Tier3 centers in the US are connected to the Internet2 backbone through Regional networks. Caltech and US LHCNet in particular have a long-standing track record of collaboration on R&D projects, through Internet2’s support of the Ultralight network in the past, and currently the direct collaboration in the NSF funded DYNES project where Internet2 and Caltech have the leading roles. The US LHCNet management team also participates in Internet2's policy and direct-setting processes, through its Network Policy, Operations and Architecture Group (Newman) and its Network Technical Advisory Council (Newman and Barczyk).

US LHCNet was also one of the first users of Internet2’s dynamic circuit network (DCN) services, implemented as the Internet2 On-demand Network (ION). US LHCNet was one of the first networks deploying and testing the DCN Software Suite (DCNSS), and continues this joint testing and deployment program. Internet2 is now moving its ION service to a new "Advanced Layer 2 Services" (AL2S) dynamic circuit infrastructure with much higher capacity. The US LHCNet team will continue to be one of the principal partners in the evolution, co-development and use of the new service, in support of DYNES and other projects supporting the LHC physics program.

1.5 Research and Development Activities

The US LHCNet team members are actively involved in R&D activities with the medium and long-term goal of integrating new and emerging network technologies and services with the LHC experiments' workflows. A current focus of these activities is to develop the next round of services and tools needed by the LHC experiments, in time for the LHC restart at 13 TeV in 2014-15.

LHCONE

The LHC Open Network Environment[7] (LHCONE) is a global-scale collaboration between the R&E networks and the LHC community. The US LHCNet team has been playing a crucial role in LHCONE since even before its conception. The Workshop on Transatlantic Connectivity for HEP in June 2010, co-organized by Caltech and CERN in order to address the requirements to meet the needs for transatlantic networking for the LHC experiments, led to a collaboration which subsequently became the LHCONE. The initial stage of LHCONE operation, based on a "Virtual Routing and Forwarding" (VRF) paradigm that allows LHC-related traffic to flow across the main R&E networks in Europe, the US and Asia via the afore-mentioned open exchange points, is already underway and is heavily used. Today, US LHCNet is actively participating in the development of the next generation services to be used in LHCONE: point-to-point dynamic circuits based on the NSI protocol as standardized by OGF, the development and use of OpenFlow based SDN networks for use in the LHC experiments, and Multipath TCP networking capability.

DYNES

The DYNES[8] project is an NSF funded activity to construct a nation-wide cyber-instrument, interconnecting up to 50 US campuses, many of which are US CMS or US ATLAS Tier2 and Tier-3 sites, though dynamic circuit infrastructure. The DYNES infrastructure has largely been deployed, and will be used in the pilot implementation of the LHCONE point-to-point services, followed by production LHCONE services by the time of the LHC restart.

OLiMPS

The OLiMPS[9] project, under a grant from the DOE OASCR program, is developing an OpenFlow based switched fabric allowing the interconnection of network nodes or end-points via multiple paths. Traditional switched or routed networks in use today are built on the principle of a single forwarding entry, and will forward packets always on the same path between source and destination. Some techniques available today are either not flexible enough (e.g BGP ECMP), complex and lacking scalability (e.g. BGP MEDs), or developed with the data center environment in mind (e.g. IEFT TRILL, IEEE SPB). The approach taken in the OLiMPS project is quite compatible with the directions taken in several NRENs in their deployment of Openflow-based services, most notably the Internet2 implementation of a Software Defined Network

[7] http://lhcone.net
[8] DYnamic NEtwork Service, see http://www.internet2.edu/ion/dynes.html
[9] Openflow Link-layer MultiPath Switching

(SDN)[10]. With this, the results and products of the OLiMPS project are expected to be directly applicable to LHCONE.

ANSE

The most recent project our team is directly involved in is Advanced Network Services for Experiments (ANSE)[11], funded by the NSF CC-NIE program in 2013-14. The goal of this project is the integration of advanced network services, such as dynamic circuit services[12] and real-time network monitoring and measurement[13] with the software stacks and the data and workflow management of the CMS and ATLAS experiments, to raise the operational efficiency of data distribution and analysis in both experiments. Several of the engineers and physicists in both CMS and ATLAS who lead the central data distribution and job processing operations and development, are now directly involved in ANSE, working closely with the US LHCNet team and other members of the project at Caltech, Michigan and Vanderbilt.

[10] Internet2 Open Science, Scholarship and Services Exchange (OSSE), built using OpenFlow technology, see http://www.internet2.edu/ion/dynes.html

[11] Advanced Network Services for Experiments (Caltech, Vanderbilt, Michigan and UT Arlington)

[12] Such as e.g. ESnet's OSCARS, Internet2's ION, DYNES, AutoBAHN and other implementations

[13] E.g. using MonALISA or PerfSONAR

2 Project Status

Caltech together with CERN have been operating transatlantic networks for the HEP community since 1984, and in particular the US LHCNet network funded by DOE OHEP and CERN since its inception as the U.S. Link Consortium (USLIC[14]) in 1995. Based on this long experience, spanning several generations of networks and network technologies, complemented by a decade of operational experience with large-scale distributed system infrastructures, US LHCNet has been designed to deliver the highest levels of service availability, and has met its goals as mentioned above.

Resilience has been engineered into the US LHCNet architecture to ensure nonstop operation of the service, based on:

- Adequate path diversity
- Equipment redundancy
- PoP diversity (two strategically located PoPs on each side of the Atlantic)
- Explicit path and PoP diverse backup paths for each of the US Tier1 centers

The current US LHCNet high-level topology is shown in Figure 1. Partial mesh connectivity at Layer 1 (the optical layer) is used for cost efficiency reasons. The topology shown is designed to match the largest data flows, and it guarantees that each pair of end-points has multiple physical paths available, for resilient operation.

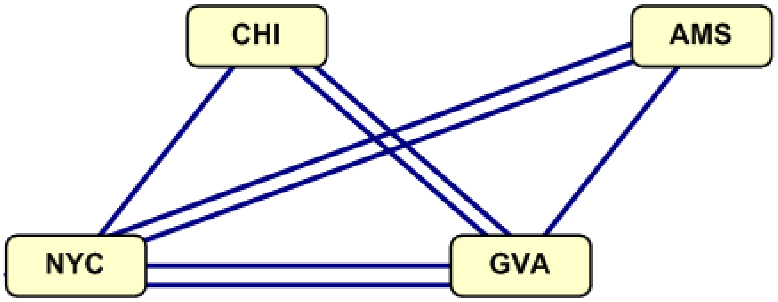


Figure 1: The US LHCNet topology since 2009.

2.1 Technical Status

The RFP carried out in the first year of the grant, resulted in contracting three transatlantic circuits each to Level(3) and T-Systems. Strong focus was given on achieving a good level of path diversity, which resulted in a total of 5 diverse submarine cable systems being used. The contracts were extended on a yearly basis for a total of three years.

[14] USLIC included Caltech, CERN, IN2P3, the WHO (Geneva) and the UN International Computer Center.

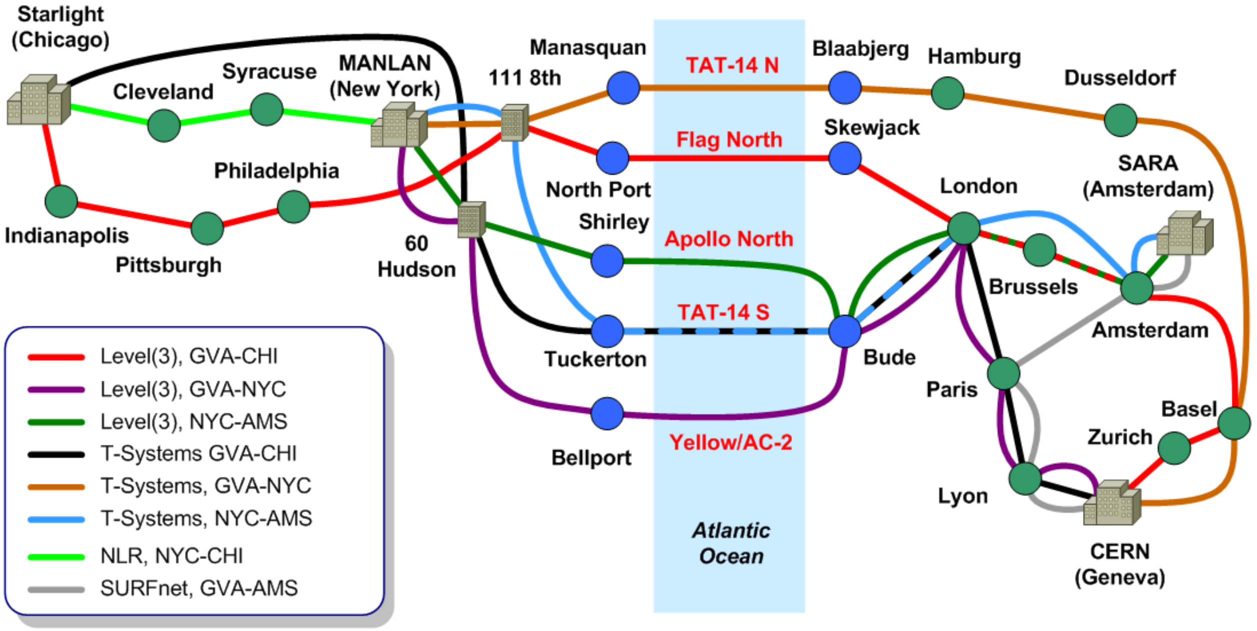


Figure 2: End-to-end routes of the US LHCNet circuits during 2009-2012.

Working together with ESnet, our engineers have prepared and commissioned explicit backup circuits to Fermilab and BNL. These backup circuits are designed to be completely path diverse, and in particular provide PoP diversity between CERN and the US Tier1 centers. The routing of the primary, secondary and backup virtual circuits through US LHCNet and ESnet is shown in Figure 3. The backup paths are only used in the very unlikely case of a simultaneous failure of both the primary and secondary circuits. They have therefore been engineered as the highest priority flow in a shared channel in the US LHCNet network. The flow prioritization has been thoroughly tested by our engineers, and it is guaranteed that the Fermilab and BNL flows in the shared channel will take precedence over any other general purpose network (GPN) traffic. This configuration allows us not to leave any unused capacity reserved, even for such serious failure scenarios where the backup circuits are needed by the Tier1s, while at the same time guaranteeing the Tier1s the agreed-upon bandwidth.

The use of the VCAT[15] standard extension to SONET in the US LHCNet network allows us to provide virtual connections at any capacity with a granularity of STS-3c, i.e. 155 Mbps (150 Mbps payload capacity). The over-all bandwidth allocation in US LHCNet is shown in Table 1. The unallocated bandwidth, which amounts to a fraction of an OC-192 link (0.7x9.4 Gbps), is reserved for protection purposes, and is a key element in providing the required 99.95+% service availability to FNAL and BNL.

[15] Virtual conCATenation protocol, allows for a flexible concatenation of SONET frames to build a sub-rate channel. See e.g. http://en.wikipedia.org/wiki/Virtual_concatenation and http://www.lightreading.com/document.asp?doc_id=30194&page_number=5

Purpose		Endpoint A	Endpoint B	Allocated Bandwidth [OC-192 links]	Allocated Bandwidth [Mbps]
Tier0-Tier1 Tier1-Tier1 (primary, secondary)	CERN-FNAL	Geneva	Chicago	2×0.9	2×8’567
	CERN-BNL	Geneva	New York	2×0.9	2×8’567
	FNAL-FZK	Chicago	Amsterdam	0.2	2’100
Tier1-Tier2	ESnet-GEANT2 peering	New York	Amsterdam	0.5	4’810
	Tier 1/2/3	New York	Amsterdam	0.5	4’810
Tier1 backup, GPN and other peerings	GPN / FNAL backup	Geneva	New York	0.4	4’208
	GPN / BNL backup / FNAL-TIFR	Geneva	Chicago	0.4	4’208
Total allocation				**5.6**	**54’404**

Table 1: Transatlantic bandwidth allocation in US LHCNet between April 2010 and October 2012. Unallocated bandwidth is used for protection of the primary services against link failures.

The primary and backup circuits are mesh-protected, i.e. will be automatically re-routed at the SONET layer within ~100ms in case of link failure. All Tier1 circuits have the highest priority for restoration, i.e. will take precedence over virtual circuits for GPN, ESnet-GEANT peering and other traffic.

It is notable that the dedicated backup circuits protect against major failures at device or facility (PoP) level. They also offer resilience against very particular failure scenarios, such as the failure of all the transatlantic circuits to one of the US LHCNet PoPs.

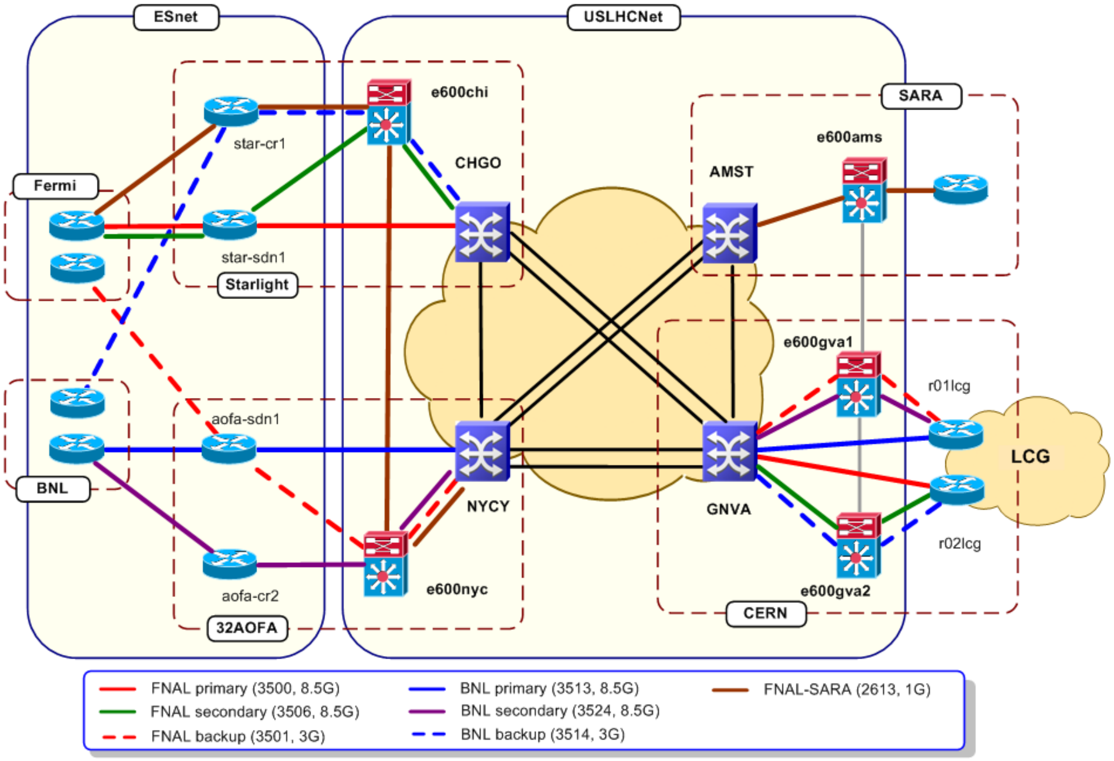


Figure 3: US LHCNet network map showing the Virtual Circuit configuration for both US Tier1 centres.

As part of its mission and to conform with its AUP, US LHCNet provides support for Tier1-Tier2 traffic across the Atlantic. For most of the grant period, U.S. Tier1 – EU Tier2 traffic has been supported in US LHCNet by a Layer 2 connection between ESnet and GEANT, used for IP peering between these two networks.

With the onset of LHCONE, we have consolidated the allocation for ESnet-GEANT and the reserved bandwidth for Internet2-GEANT peering (for US Tier2 – EU Tier1 traffic), into one single channel of 10Gbps, now used by the LHCONE VRF service. It carries part of the Tier2-related traffic between US and Europe. The current capacity allocation is shown in Table 2.

Purpose		**Endpoint A**	**Endpoint B**	**Allocated Bandwidth [OC-192 links]**	**Allocated Bandwidth [Mbps]**
Tier0-Tier1 Tier1-Tier1 (primary, secondary)	CERN-FNAL	Geneva	Chicago	2×0.9	2×8’567
	CERN-BNL	Geneva	New York	2×0.9	2×8’567
	FNAL-FZK	Chicago	Amsterdam	0.2	2’100
Tier1-Tier2	LHCONE	New York	Amsterdam	1.0	9’620
Tier1 backup, GPN and other peerings	GPN / FNAL backup	Geneva	New York	0.4	4’208
	GPN / BNL backup / FNAL-TIFR	Geneva	Chicago	0.4	4’208
Total allocation				**5.6**	**54’404**

Table 2: Current capacity allocation in USLHCNet since October 2012.

Further supporting the DOE science mission, in December 2009, US LHCNet responded to a request from Fermilab, CERN and TIFR, and extended a Layer 2 connection on the existing CERN-TIFR 1 Gbps link to Fermilab. In US LHCNet, this VLAN is carried inside a shared virtual circuit between CERN and Starlight. The installation of this VLAN helps the CMS operations group in its data movement to India, and has been welcomed by both FNAL and TIFR. The layout of this connection is shown in Figure 4.

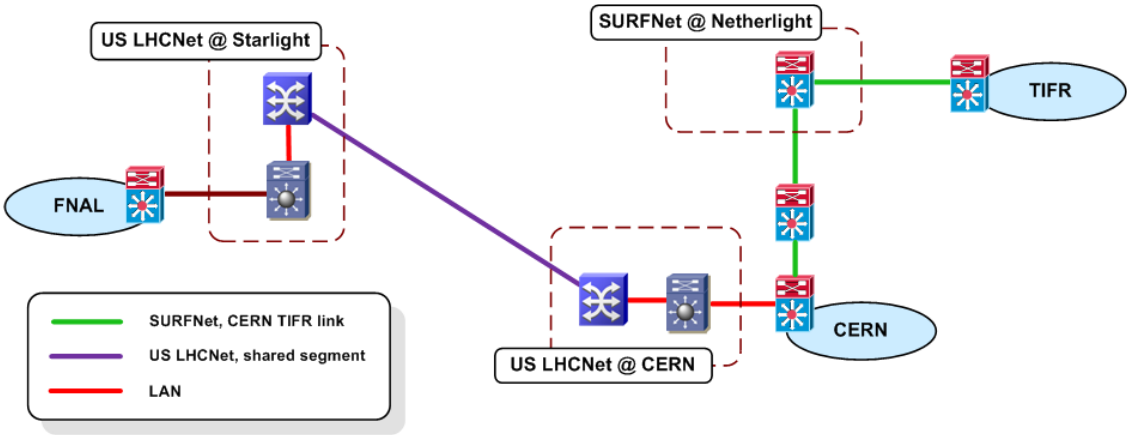


Figure 4: The design of the combined CERN and US LHCNet circuit between Fermilab and TIFR.

US LHCNet, an early adopter of the dynamic circuit paradigm, deployed the Internet2/ESnet Dynamic Circuit Network Software Suite (DCNSS[16]), based on the OSCARS and DRAGON software, already in 2008. US LHCNet is capable of providing dynamic circuits at all of its PoPs, including the European PoPs in Geneva and Amsterdam. The DCNSS is compatible with the Network Services Interface specification as standardized by the Open Grid Forum (OGF), and is expected to work seamlessly with the European AutoBAHN service, once that service is deployed.

[16] See https://wiki.internet2.edu/confluence/display/**DCNSS**

In 2012, the circuit providers were changed, with all 6 circuits contracted to Level(3). This change was done in response to Level(3)'s very cost effective proposal, which also maintains the required level of path diversity across the Atlantic. The current US LHCNet network map is shown in Figure 5.

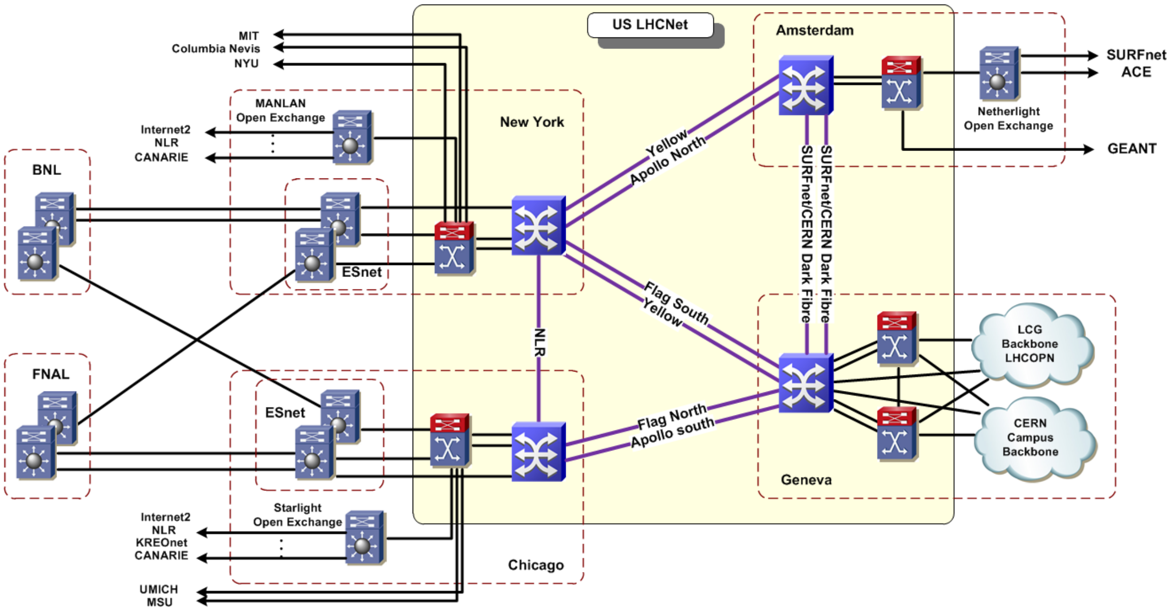


Figure 5: US LHCNet network as of March 2013.

Figure 6 shows the evolution of the network utilization during the grant period. The growth of average and peak network throughput during LHC operation in 2011 and 2012 is clearly visible. High network utilization has continued through the first quarter of 2013, up to the present. Reasons for this include the LHC Heavy Ion run which continued through the end of February, and the Moriond Electroweak and Moriond QCD conferences along with other conferences, where ATLAS and CMS are continuing to present many updated physics results from the three year LHC run, including those bearing on the nature and properties of the new Higgs-like boson.

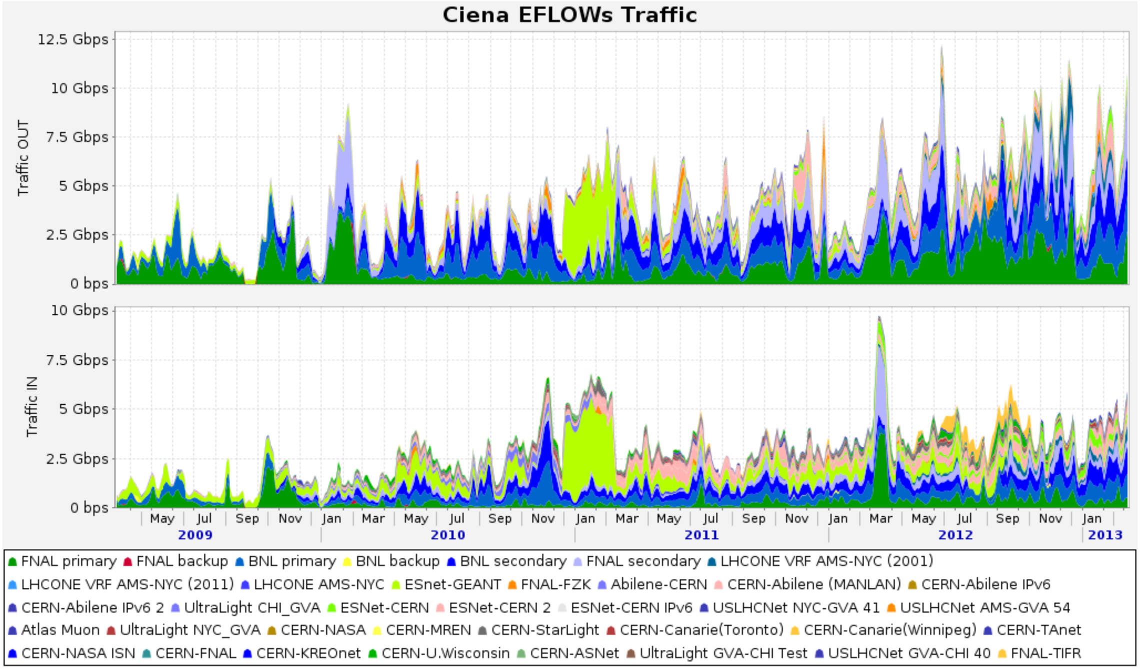


Figure 6: Evolution of the network utilization between 2009 and 2013.

In total, close to 90 PB have been transferred through US LHCNet during the last grant period, as shown in Figure 7.

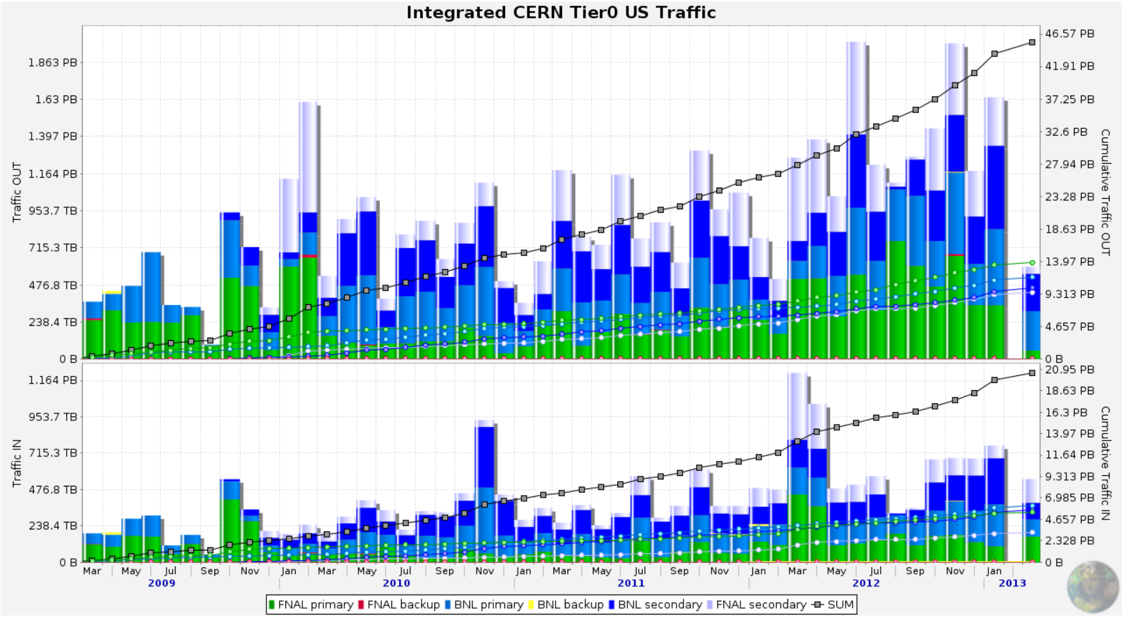


Figure 7: Integrated data traffic through US LHCNet in the period between 2009 and early 2013.

More details on the MonALISA monitoring system are given in Appendix A.

In Spring 2012, our team took the initiative to optimize the operational costs of the transatlantic network. Following a proposal to and discussion with the US CMS and US ATLAS Operations

Management, the Caltech team has followed up on a previous unsolicited sole-source proposal from Level(3), and negotiated new contracts with Level(3) to provide all 6 transatlantic circuits for use in US LHCNet at a very favorable bundle price. The new contract was signed on August 8 2012, for a duration of two years. The changes to the network have been kept to a minimum: the three circuits already in production use remain as is (at a reduced price), while the three circuits previously contracted to T-Systems have been replaced by new Level(3) circuits (also at a favorable price). A high degree of path diversity is maintained, with 5 different cable systems being used for the 6 circuits, as indicated in Figure 5. The main requirement has been that no single event can bring down two circuits interconnecting the same pair of end-points.

US LHCNet completed its transition to the new transatlantic circuit services last Fall, as planned.

The three circuits previously contracted to T-Systems were terminated with the effective date of November 15, 2012. By that date, Level(3) was able to provide two of the three new circuits. US LHCNet continued to operate with five transatlantic circuits over the Christmas/New Year holiday period, until January 15, 2013. Since this affected mainly the protection capacity, operationally, the impact was capacity reduction in case of outages. With the commissioning of the sixth transatlantic circuit in January, US LHCNet resumed operating at full capacity in time for the heavy ion run and the preparations for the winter 2013 conferences.

Thanks to this new contract, based on the volume and in particular longer contract terms, the savings to the US LHCNet program over the two year period of 2013-14 are estimated to amount to approximately $ 600k[17]. These savings will be used to cover a substantial part of the required hardware upgrade. This upgrade is necessary soon in order to replace ageing equipment (the Force10 routers are in service for 7 years now, and the Ciena CoreDirector multiservice switches for 5 years), and to continue to provide the required services (with state of the art uptime, as summarized above) to the LHC experiments.

2.2 Service Availability

In order to provide high availability for its primary mission services, US LHCNet uses Layer 1 protection mechanisms available on its deployed SONET platform. This section describes the protection scheme used, as well as the availability figures which are reached.

2.2.1 Protection for Target 99.95+% Service Availability

The target 99.95% service availability is addressed through resilient network design, partial mesh topology and implementation of circuit protection with an adequate level of bandwidth reserved for this purpose. In order to minimize the capacity that is held in reserve and used only during link failure scenarios, US LHCNet uses CIENA’s mesh protection mechanism. This mechanism is based on a link-state routing protocol implemented at the optical layer (Optical Signaling and Routing Protocol, OSRP). In case of an outage on any segment of the network, the new topology is quickly calculated, and protected services are rerouted on active segments where capacity is available.

[17] The exact amount will depend on the precise date when the transition is finalized.

The advantage of mesh protection over more traditional SONET/SDH linear protection[18] mechanisms is that any free capacity on any segment can be used to protect any circuit, as long as a route exists between the ingress and egress point of the network[19]. As such, CIENA's mesh protection offers the robustness of Layer 1 TDM operation with the flexibility of rerouting usually available only at Layer 3. US LHCNet successfully combines this mechanism with the choice of topology, implementing terrestrial segments between Geneva-Amsterdam and New York-Chicago and providing only a small amount of transatlantic capacity reserved for protection. The advantage of mesh protection over linear protection is depicted in Figure 8: in the case of linear protection (left side of the figure), several additional expensive transatlantic links would have to be provided, while in the case of mesh protection (right side of the figure), two (cheaper) terrestrial segments are added, with only one transatlantic protection link. The latter can be used to protect all other transatlantic circuits – GVA-CHI, GVA-NYC as well as GVA-AMS.

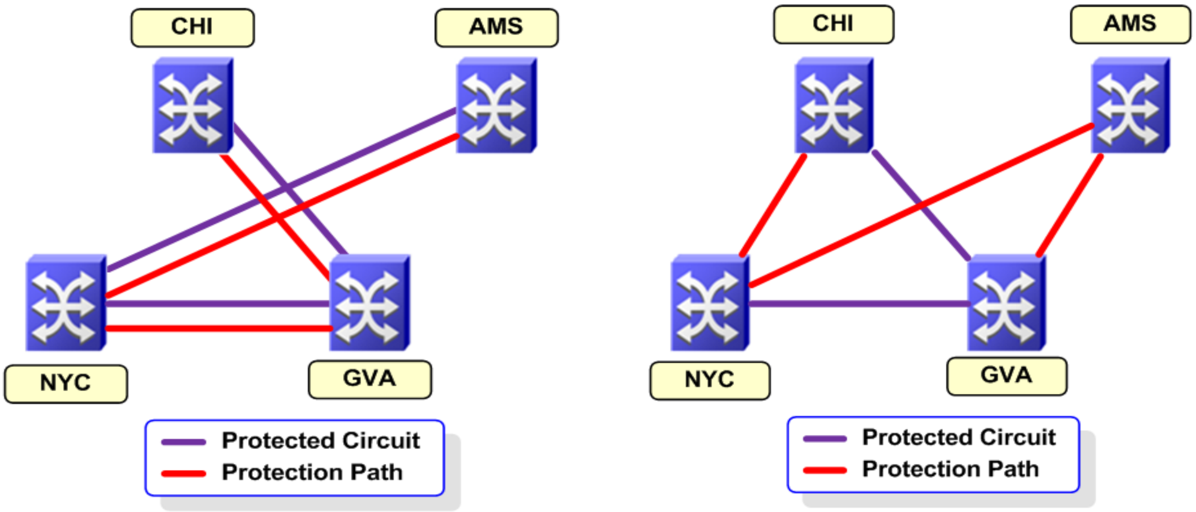


Figure 8: Examples of linear protection (left) and mesh protection (right). Only a subset of the links is shown for clarity of illustration. Note that in the mesh protection scheme, both GVA-CHI and GVA-NYC can use the protection capacity (red) as needed.

The second cornerstone of reaching the high availability target figure is the use of VCAT and LCAS protocol extensions to SONET/SDH (and also OTN). VCAT allows a flexible virtual concatenation scheme, i.e. splitting an end-to-end virtual circuit over multiple physical links using Sub-Network Connections (SNCs). Using LCAS, any of the SNCs can be added or removed dynamically to/from a virtual circuit, thus increasing or decreasing the virtual circuit's capacity. In praxis, this means that an outage on one of the links carrying some of the SNCs will reduce the capacity of a virtual circuit only by this amount – all other SNCs remain operational. For protected virtual circuits, the failed SNCs will be automatically restored on the available

[18] SONET's ring protection is not applicable to US LHCNet due to the very long spans across the Atlantic – a ring protected circuit could result in traffic crossing the Atlantic three times.

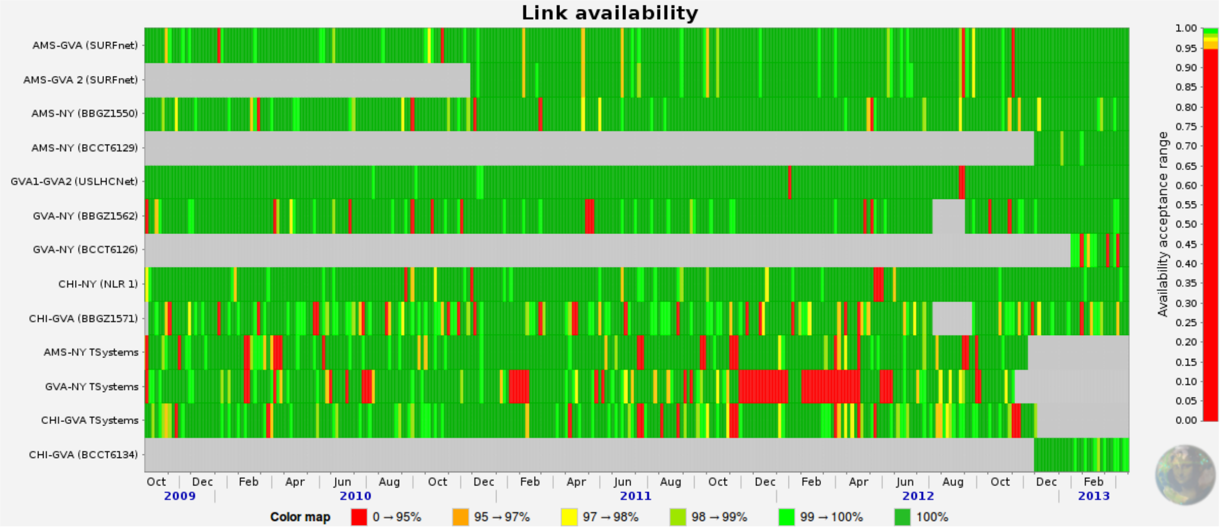
[19] In linear protection, a reserved protection circuit must be provisioned in parallel to the working circuit, on the whole path between the ingress and egress nodes.

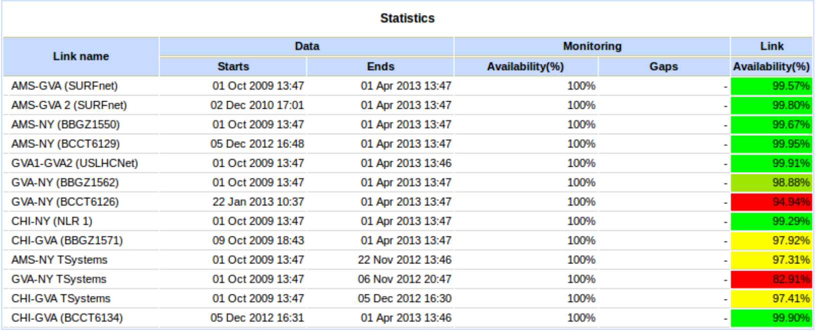
protection path(s). Therefore, a protected end-to-end connection will experience only minor packet loss, but not an interruption. In addition, in case not enough protection capacity is available (e.g. in case of multiple simultaneous outages), the net effect will be a reduction in provisioned bandwidth of the virtual circuit, not a loss of connectivity.

2.2.2 Service Availability Figures

The availability of all the US LHCNet installed links since October 2009 is shown in Figure 9. Grey areas indicate where a given circuit did not exist: either before the link was taken in service, or after de-commissioning. The statistics includes all unscheduled outages as well as planned maintenance periods. The first major effect visible in the plot is the frequent short interruptions impacting each circuit, scaling roughly with the physical length of the links. Transatlantic circuits are more complex than purely terrestrial links, e.g. each of our transatlantic circuits is composed of two submarine segments (Atlantic and English Channel), and several terrestrial segments. Longer spans also mean more fiber, making the link more vulnerable to outages due to road-works etc., as well as more carrier equipment such as repeaters, regenerators and add-drop multiplexers which represent possible points of failure.

The second effect visible in this figure is the several submarine fiber cuts, visible as extended link outages marked in red. The duration varies depending on the cause and location of the fault. As these examples show, such extended outages are to be expected and cannot be avoided, but rather have to be taken into consideration when planning for the service levels to be delivered to the end-sites. US LHCNet does this very efficiently by dedicating only a small fraction of the bandwidth for protection against outages, and the use of advanced features such as mesh protection, LCAS and VCAT protocols.





Statistics

Link name	Data		Monitoring		Link
	Starts	Ends	Availability(%)	Gaps	Availability(%)
AMS-GVA (SURFnet)	01 Oct 2009 13:47	01 Apr 2013 13:47	100%	-	99.57%
AMS-GVA 2 (SURFnet)	02 Dec 2010 17:01	01 Apr 2013 13:47	100%	-	99.80%
AMS-NY (BBGZ1550)	01 Oct 2009 13:47	01 Apr 2013 13:47	100%	-	99.67%
AMS-NY (BCCT6129)	05 Dec 2012 16:48	01 Apr 2013 13:47	100%	-	99.95%
GVA1-GVA2 (USLHCNet)	01 Oct 2009 13:47	01 Apr 2013 13:46	100%	-	99.91%
GVA-NY (BBGZ1562)	01 Oct 2009 13:47	01 Apr 2013 13:47	100%	-	98.88%
GVA-NY (BCCT6126)	22 Jan 2013 10:37	01 Apr 2013 13:47	100%	-	94.94%
CHI-NY (NLR 1)	01 Oct 2009 13:47	01 Apr 2013 13:47	100%	-	99.29%
CHI-GVA (BBGZ1571)	09 Oct 2009 18:43	01 Apr 2013 13:47	100%	-	97.92%
AMS-NY TSystems	01 Oct 2009 13:47	22 Nov 2012 13:46	100%	-	97.31%
GVA-NY TSystems	01 Oct 2009 13:47	06 Nov 2012 20:47	100%	-	82.91%
CHI-GVA TSystems	01 Oct 2009 13:47	05 Dec 2012 16:30	100%	-	97.41%
CHI-GVA (BCCT6134)	05 Dec 2012 16:31	01 Apr 2013 13:46	100%	-	99.90%

Figure 9: Link availability for all US LHCNet links since October 2009.
Table under the plot gives the numerical values, averaged over the entire period.

The table in Figure 9 shows the availability numbers during the past 4 years, and includes both the circuits contracted in 2009, as well as the new circuits in service since late 2012/early 2013.

Looking at only the circuits in operation since 2009, we observe an average link availability of 97.0%. This low value is caused by one particular circuit, a statistical outlier: The T-Systems link between Geneva and New York had several very long outages, in particular one in late 2011 due to a submarine shunt fault, with bad weather preventing a repair for over 2 months. The second long outage, following shortly in early 2012 was due to a repeater replacement on the entire TAT-14 cable. During both outages, T-Systems was not able to provide backup capacity, despite repeated requests from the US LHCNet side[20].

The average link availability without the bad circuit amounts to 98.2% between the 5 submarine circuits only, or 98.7% if averaging over all (8) other circuits, including the terrestrial links between Geneva and Amsterdam, and New York and Chicago. This is consistent with our long-term experience including previous contracts, where we were quoting 99% circuit availability.

[20] Paying due diligence, we have decided to not extend the T-Systems contract, and instead opted for substituting them with Level(3) circuits in late 2012.

The emphasis on resilience through path diversity[21], coupled with efficient protection mechanisms implemented within the US LHCNet network, has paid off, and led to only minimal impact to service levels provided. For the calculation of service availability during the entire grant period, we have included all contracted circuits.

The results are summarized in Table 3. Single circuit outages do not impact the primary US LHCNet services thanks to the automatic protection and restoration that is an integral part of US LHCNet's architecture. Multiple simultaneous outages however reduce the service levels. Due to partial overlap on several segments, as shown in Figure 2, we rely on a careful analysis of past failures in order to estimate the probability of multiple outages.

The results indicate an average of less than 17 days per year with two circuits unavailable. In this condition, US LHCNet can still guarantee bandwidth availability to each of the US Tier1s at a lower capacity, but still higher than a single OC-192 circuit, i.e. between 9.4 and 16.8 Gbps (8-14.3 Gbps of usable bandwidth).

The design of the US LHCNet network has put an emphasis on avoiding more than two simultaneous outages as much as possible within the allowed budget envelope. The architecture takes the small overlaps into account, making sure that a single outage cannot impact two circuits terminating at the same Tier1 site. Independent outages affecting two such circuits at the same time cannot be however completely excluded, as the example of hurricane Irene in 2011 shows. The storm caused flooding in places 200 miles apart, resulting in two of the main US LHCNet links between Geneva and Chicago both being cut for over 24 hours. USLHCNet continued to provide 4.5 Gbps bandwidth for the LHCOPN connection between CERN and Fermilab, through the backup path Geneva-New York-Chicago.

Three or more simultaneous failures are statistically possible, but the probability is low, and estimated[22] at less than 5 hours per year.

With 6 transatlantic links			
# transatlantic links failed simultaneously	**Effect on US Tier1 services (primary and secondary)**	**Effect on Tier2 and other unprotected services**	**Maximum expected duration within one year**
1 link	No impact, service protected, 16.8 Gbps operational per Tier1	Degraded, operational	95 days/year
2 links	Degraded, available bandwidth between 9.4 and 16.8 Gbps	Degraded or not operational	7 days/year
3 links	Degraded, but at least 8.4 Gbps bandwidth available	Degraded or not operational	5 hours/year

Table 3: Impact and projected duration of transatlantic circuit failures on the current US LHCNet services, using six transatlantic links, based on operational experience.

[21] Some overlap on terrestrial segments has been taken into account, and allowed for cost efficiency reasons.

[22] The approximation uses observed availability per circuit, but assumes independent outages, and neglects the outage probability on the terrestrial links used on backup paths. Strictly speaking, the figures given for the maximum duration of single or multiple link outages shown in the table, while expected to be accurate, thus represent lower bounds.

Over the past four years, the availability of capacity equivalent or higher than one OC-192 circuit between CERN and each of the US Tier1 sites is calculated to be above 99.94%. The availability of full provisioned bandwidth (both primary and secondary FNAL and BNL circuits available) is estimated at 98.1%. The service availability, where the Tier 1 sites were reachable through US LHCNet is estimated to be 99.99%.

2.3 Fair-sharing during degraded operation

To guarantee fair sharing of the network resources during degraded operation, US LHCNet uses the standard VCAT[23] and LCAS[24] protocol extensions to SONET as well as Layer 1 protection mechanisms. VCAT allows splitting end-to-end Virtual Circuits (VCs) into multiple Sub-Network Connections (SNCs), with the SNCs taking different physical paths through the US LHCNet network. Each physical link segment has SNCs belonging to different VCs allocated. In case of an outage, a proportional fraction of each VC is affected. For protected VCs, the failed SNCs are rapidly restored, by using the protection path. Unprotected VCs are reduced in bandwidth using LCAS.

Figure 10 and Figure 11 demonstrate the principle of VCAT/LCAS operation in case of a link outage. Two virtual circuits spanning multiple physical links are shown: one with protected circuit behavior (in green), and one without protection (in blue). In this example both virtual circuits continue to operate during the link outage: the protected circuit continues at the full rate, while the unprotected one continues at two-thirds of the initially provisioned bandwidth.

[23] Virtual conCATenation
[24] Link Capacity Adjustment Scheme

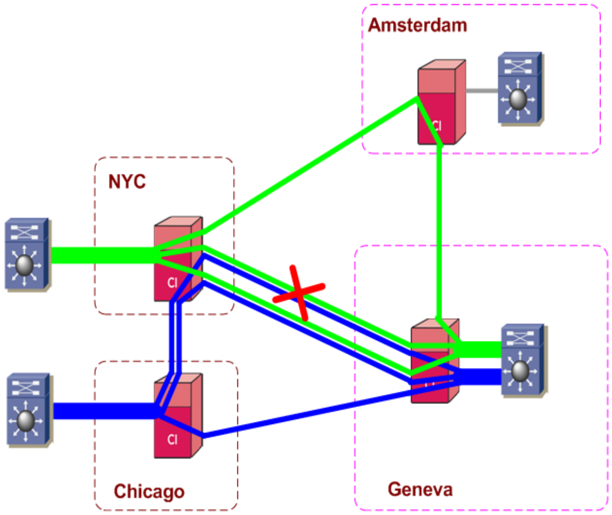


Figure 10: VCAT/LCAS operational principle: Initial multi-path distribution of Sub-Network Connections, concatenated using VCAT to form one protected (green) and one unprotected (blue) Virtual Circuit. The red cross indicates an assumed fiber cut disabling the link.

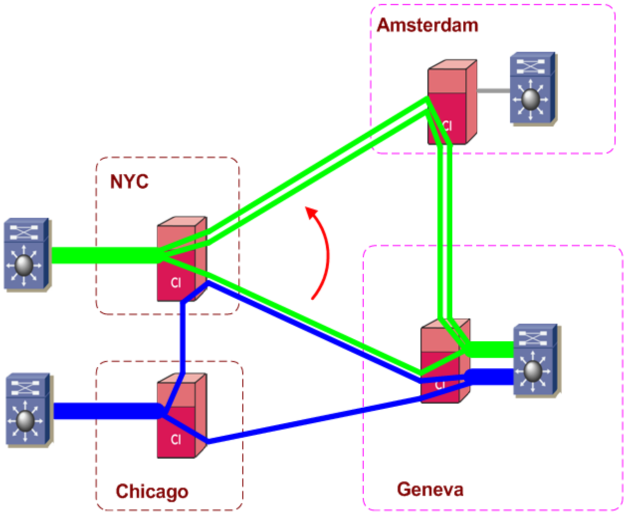


Figure 11: VCAT/LCAS operational principle: the Sub-Network Connections distribution after restoration. The green (protected) virtual circuit continues at the full provisioned bandwidth, while the blue (un-protected) one is reduced in bandwidth, but continues to pass traffic.

Note that at the moment of failure, both circuits are affected to the same degree. If both circuits are not protected, both will continue to operate at lower capacity, while the equal distribution of the SNCs guarantees fairness.

2.4 Monitoring

US LHCNet uses the MonALISA monitoring framework as well as perfSONAR for operational purposes. We use the perfSONAR protocol standard to export operational values to external clients, in particular the E2EMON used in the LHCOPN.

MonALISA forms the basis of our remote operations infrastructure, which includes continuous, fully consistent monitoring of all of our links in real-time, as well as automated operations in case of link outages. Based on in-the-field experience, the MBTF of this pervasive infrastructure is measured in years.

MonALISA, with its high performance underlying real-time messaging infrastructure, provides the unique capability of deriving the precise uptime on each link, thereby helping to detect and measure the time of each outage and help enforce our SLAs. It also includes open source end-system client software agents that can profile the hardware, operating system and kernel settings, network interfaces and IO subsystems, as well as the CPU and IO load on the data senders and receivers, which has proven very useful in diagnosing and resolving end-to-end throughput problems. MonALISA, along with other robust monitoring tools such as Spectrum, also provides monitoring information to PerfSONAR where needed.

MonALISA gathers and exports via the perfSONAR protocol the following information:

- Bandwidth utilization for each data flow
- Port status

- Device status
- Alarms on each device used

MonALISA also monitors the flows across the dynamic circuits created by the Internet2's dynamic circuit system (DCNSS), and rapidly recognizes their creation and destruction, to provide precise information on the data volumes sent and received over each circuit, and to flag any performance problems when needed.

US LHCNet will continue to use the MonALISA monitoring framework, as it offers functionality currently beyond what is available in the perfSONAR toolset. These features are vital for operation of the US LHCNet network. Basic functions provided by MonALISA (which are also provided by alternative commercial software packages such as Spectrum) are:

- Real time alarm notification via e-mail and SMS
- Availability reports
- Multi-platform support (from routers, switches down to optical multiplexers)
- Multi-protocol support (SNMP, TL1, CLI)

The US LHCNet team is aware of, and actively collaborating in the efforts of the perfSONAR development teams to provide the above features in the future. Our collaborations are carried out through various working groups in this area, including the Internet2 Performance Working Group and the DICE monitoring working group. As part of these efforts, MonALISA is and will remain a perfSONAR compliant monitoring system.

2.5 Operations

The US LHCNet operations team uses a "unified" Level 1/2/3 support model, based on the flexibility and versatility of each team member. In addition, as part of the 24x7 operational support, US LHCNet uses the CERN Computer Center staff when needed for Level 1 support. The CERN Computer Center operators are present 24x7x365. US LHCNet exports monitoring data to the SPECTRUM system used by CERN, followed in real-time by the operators.

The monitoring and automated operations services infrastructure deployed in US LHCNet, are themselves highly resilient and diverse. This enables the engineering team to provide a highly successful converged Level 1+2 response around the clock, where all engineers on the US LHCNet Caltech team, as well as the CERN operators, are notified immediately of any link or equipment failure. For completeness, messages also reach the PI who is on the NOC notification list, and who may participate directly in case of the need for strategic decisions.

This approach has been shown to lead to the rapid application of the team's expertise to minimize downtime, and also the proven ability to deal rapidly with major outages involving multiple links, or other complex problems, escalating these problems to Level 3, where the entire team and in some cases the PI will be involved, on the rare occasions when they arise.

This mode of operation requires expert team members with a wide variety of skills, from network operations, to switch and router configuration and testing, to Layer 1 optical multiplexer configuration and testing, to end-system and network interface administration, testing and optimization. This also extends to a general "distributed system service", that includes debugging in partnership with the experiments, as well as sister teams in Caltech HEP and at collaborating institutions, to resolve end-to-end throughput problems.

The same partnership, on other occasions, participates in the deployment of pre-production tools and subsystems using new technologies and/or state-of-the-art services for high throughput as needed, to move them systematically into production so that the network deployed at each stage of the roadmap can be utilized by the experiments to full advantage.

2.5.1 Network Monitoring

Network monitoring is vital to ensure proper network operation over time, and is tightly integrated with all the procedures implemented by the US LHCNet NOC, as described in Annex F. Besides the central role in the performance of pro-active interventions in the case of network failure, network monitoring helps establish long term trends in terms of network utilization, and provides global quality figures about the network health as well as a database of historical events that should help speed up incident resolution in the future.

US LHCNet uses several monitoring tools in a coordinated fashion, each with its clearly defined role:

- **MonALISA** monitors the CIENA CD/CIs by means of a custom TL1 module, and sends alarm notifications via email to the US LHCNet NOC. It provides true end-to-end monitoring as well as global trends, and precise real-time as well as historical information on link availability and utilization.
- **PerfSONAR** is used to provide monitoring values to E2ECU[25].
- **E2ECU** monitors the PerfSONAR link status and contacts the US LHCNet NOC. The E2ECU acts as a central repository for network performance data for the entire LHCOPN, as well as a dissemination channel for scheduled or unscheduled network events.
- **Spectrum** monitors most US LHCNet devices and can also receive CIENA alarms through the experimental TL1 gateway installed at CERN. Spectrum alarms are monitored by the CERN Operators which function as the Level 1 NOC for US LHCNet. There are a set of procedures in place for the operators to follow in order to determine the urgency level of an alarm and the action to take (currently send an email or phone the US LHCNet engineer on-call).

There are also other tools in place to insure configuration change tracking (RANCID) and centralization of network equipment logs (syslog-ng).

2.5.2 Network Service

US LHCNet is deeply involved in the design and operation, and is an integral part of the LHC “Optical Private Network” (OPN), which is the primary network interconnecting the Tier0 at CERN with the Tier1 sites as well as connections among the Tier1s themselves.

At present, the LHCOPN is built as a set of Layer 2 connections between CERN and the Tier 1 sites, and as of recently includes also Tier1-Tier1 Layer 2 circuits primarily used for LHC data traffic. Originally the design concentrated on the Tier0-Tier1 connectivity, but has moved on to

[25] The “End-to-End Coordination Unit”, a NOC like entity monitoring the status of the LHCOPN links, and coordinating problem resolution.

include Tier1-Tier1 data movement as scavenger (i.e. lowest priority) traffic. Today, Tier1-Tier1 data movement exceeds Tier0-Tier1 traffic.

The LHC OPN is an example of a truly federated network, made up of the connecting Tier1 end-sites and the NRENs together with US LHCNet, ESnet, GEANT, and Internet2 as network service providers. Together with regular conference calls, quarterly face-to-face meetings are used for coordination between the end-sites and the networks. US LHCNet participates in all of these activities.

The LHC OPN has developed its operational model based on the interaction between users, operators (Tier0/Tier1 sites and Layer 3 operators) and infrastructure providers (Layer 1/Layer 2 network operators). It uses the Global Grid User Support System (GGUS) for trouble ticket handling and problem resolution at IP level. The End-to-End Coordination Unit (E2ECU), operated by DANTE, has as its role, as the name implies, the coordination of problem resolution in the inherently multi-domain environment of the LHC OPN. The E2EMON system is used for multi-domain monitoring of end-to-end circuits.

US LHCNet is fully involved in all operational aspects of the LHC OPN: it responds to GGUS ticket processing where US LHCNet is involved, provides information to and works on problem resolution with the E2ECU, as well as provides monitoring information to the E2EMON system through the installed perfSONAR monitoring server.

US LHCNet’s operational procedures include cooperation within the LHCOPN operational model, through information exchange including advance notification of planned maintenance, problem reporting and repair status updates. US LHCNet’s contact information is available to GGUS as well as E2ECU.

US LHCNet also engaged and continues to engage with the LHC OPN community and beyond, including Tier2s and Tier3s in several other ways, including:

- Operation of a Tier2 and Tier3 by sister teams at Caltech supported by DOE through the US CMS program as well as direct grants to Caltech. This enables the US LHCNet team to keep current with all associated network-related plans and issues, and to understand the relation of the network services to the LHC experiments’ computing and software needs and issues, in depth.
- Collaborative work in the LambaStation[26], Terapaths[27], UltraLight[28], PLaNetS projects, and now also the DYNES program funded by NSF (which was conceived and initiated by Caltech) in partnership with Internet2.
- Large scale pre-production deployment and testing of new technologies “at scale” in cooperation with major network equipment vendors as well as telecommunications providers, also in cooperation with the Tier1 teams at BNL and Fermilab, the teams at Caltech and many other US Tier2 sites, as well as CERN OpenLab and the Tier2s in Brazil and Korea. These exercises, including Tier2-scale and larger annual demonstrations of leading edge network, network interface and data server technologies

[26] http://www.lambdastation.org/
[27] https://www.racf.bnl.gov/terapaths/
[28] https://www.ultralight.org/web-site/ultralight/html/index.html

at the annual Supercomputing conferences, receive massive support from some of the world’s leading network providers, the major R&E networks (NLR, Internet2, ESnet, RNP and SURFNet in particular), and the major optical (CIENA), switch/router network vendors (Cisco, Force10) and network interface (Intel, Neterion, Chelsio, and more recently Myricom and Mellanox) manufacturers.

- The US LHCNet PI has been the originator and initial developer of the LHC Computing Model and the associated network roadmaps (including the latest roadmap developed by the PI and Lead Engineer, which is identical to that in the recent DOE RFI, and which also was presented in US LHCNet’s May 2010 annual Status Report).

  The US LHCNet team, and the PI in particular, devotes continuous attention and contributes to the evolution of the LHC Computing Models by the experiments. This includes participation in workshops and meetings on data and network operations, synthesizing the needs and outlook for requirements including the relationship of network needs to other needs such as the storage volumes, and reporting the needs to partner organizations such as Internet2, ESnet, the U.S. funding agencies, and ICFA.

- The US LHCNet PI also has a pivotal leading role in many aspects of the contacts between the networking and high energy physics communities, ranging from triggering the a series of Internet2 Tier2 and Tier3 workshops in 2008-10 (led by R. Carlson), to co-leading the Internet2 HEP Special Interest Group, to initiating the DYNES project led by Internet2, to representing HEP and the scientific community as a whole in Internet2's Network Policy, Architecture and Operations Advisory Council, as well as its former Applications Advisory Council. Other ties to the community include the PIs former role as the US CMS Computing Liaison and the longstanding chair (since 2002) of ICFA's Standing Committee on Inter-regional Connectivity, where he reports to ICFA semi-annually on the state and requirements of HEP networking and computing, and their future directions.

  A major recent example of this community role was in the proposal stage of the DYNES project, where the PI had the central role in explaining and developing support for the project among more than 40 campuses and 14 regional network Connectors to Internet2, as well as the Open Science Grid.

2.5.3 Software Support

Several software packages are being used by US LHCNet today. This section summarizes the development and maintenance aspect of each of them.

DCNSS (OSCARS/DRAGON): the Dynamic Circuit Network Software Suite is maintained by Internet2 and in the case of OSCARS, by ESnet. US LHCNet has deployed the DCNSS on its Inter-Domain Controller in order to provide transatlantic Dynamic Circuit operation. Internet2 is committed to support DCNSS, and in particular has agreed to continue the development and support relating to the CIENA platform. Extending these dynamic circuit services to campuses with Tier2s and Tier3s, via regional and state network connectors to Internet2 across the U.S. will be done by the NSF-funded DYNES project starting in the first quarter of 2011. DYNES and the DICE collaboration will work together to extend these services across the Atlantic, in particular to and from the Tier2 and Tier3 sites in Europe.

The DYNES services and software used will be maintained by the DYNES collaborating institutes, and in particular the DCN software and ION service will be maintained by Internet2, who provides the PI (E. Boyd) and administrative staff for the DYNES project. In Europe, the main European project addressing dynamic bandwidth allocation is AutoBAHN, an activity in the GN3 project expected to reach service deployment in 2011. The support for these development efforts comes from GEANT's managing institution DANTE, and from the NRENs participating in the GEANT project.

The PerfSONAR monitoring toolkit: is an open-source community development led by ESnet, Internet2 and GEANT. Development and maintenance is performed by the community. Distinguishing between the perfSONAR protocol and the implementations, there are currently two of the latter: perfSONAR-PS, developed and maintained mainly in the US, and perfSONAR-MDM, which is a European development in the GEANT3 project.

MonALISA: is the monitoring framework, as well as a monitoring system, both of them designed, developed and maintained by Caltech. US LHCNet has direct communication channels with the MonALISA developers, guaranteeing quick response for reported bugs as well as development and customization requests.

2.6 Future Planning

After the end of the DOE grant DE-FG02-08ER41559, the US LHCNet project is continuing with funding from the US CMS and US ATLAS Operations Program.

The US LHCNet team is currently evaluating the best suited technology and architecture for the next generation US LHCNet services. We are working on multiple tracks, including hardware platform evaluation, such as Carrier-Ethernet based solutions, as well as investigating OpenFlow based solutions as part of the important global trend towards software-defined networks (SDN). Targeting the LHC startup in late 2014, we plan to submit the proposal including a detailed architectural design to the US CMS and US ATLAS Operations Management in Spring 2013, and start the implementation soon thereafter, to be ready in fully operational state and integrated in the LHCONE next generation services by Fall 2014.

More detailed plans are laid out in Appendix B.

3 Conclusions

The US LHCNet network has provided high availability, high-capacity network services across the Atlantic to the US LHC community in support of the LHC physics program. Through state-of the art, resilient, cost effective design, we have achieved service availability numbers of 99.99% and above. The US LHCNet network has been a key component of the LHCOPN network system, linking CERN with all Tier 1 sites world-wide. The US LHCNet team has played an important role in the conception, design, deployment and ongoing development of the LHC Open Network Environment (LHCONE). With its expertise in Layer 1 through Layer 3 networking, as well as new, revolutionary mainstream concepts such as Software Defined Networking and OpenFlow, the team continues to be a key player in the construction of the advanced services in LHCONE.

After signing new circuit contracts until end of 2014, the US LHCNet network is currently preparing for the implementation of the new architecture, matching robustness as needed for successful continuation of the LHC program, with modern techniques and services tailored to the new concepts used in the LHC experiments' data processing and analysis workflows.

Appendix A The MonALISA Framework

MonALISA (Monitoring Agents in A Large Integrated Services Architecture) (http://monalisa.caltech.edu) is a globally scalable framework of services developed by Caltech to monitor and help manage and optimize the operational performance of grids, networks and running applications in real-time. MonALISA is currently used in several large scale HEP communities and grid systems including ALICE, CMS, ATLAS and the Open Science Grid (OSG). It actively used to monitor all network devices and links in US LHCNet. MonALISA also is used to monitor, control and administer all of the EVO[29] reflectors, and to help manage and optimize their interconnections.

As of this writing, more than 360 MonALISA services are running throughout the world. These services monitor more than 40,000 compute servers, and thousands of concurrent jobs. More than 4 million persistent parameters are currently monitored in near-real time with an aggregate update rate of approximately 25,000 parameters per second.

This information also is used in a variety of higher-level services that provide optimized grid job-scheduling services, dynamically optimized connectivity among the EVO reflectors, and the best available end-to-end network path for large file transfers. Global MonALISA repositories are used by many communities to aggregate information from many sites, to properly organize them for the users and to keep long term histories. During the last year, the repository system served more than 8 million user-requests.

A.1 MonALISA System Design

The MonALISA system is designed as an ensemble of autonomous self-describing agent-based subsystems which are registered as dynamic services. These services are able to collaborate and cooperate in performing a wide range of distributed information-gathering and processing tasks.

An agent-based architecture of this kind is well-adapted to the operation and management of large scale grids, by providing global optimization services capable of orchestrating computing, storage and network resources to support complex workflows. By monitoring the state of the grid-sites and their network connections end-to-end in real time, the MonALISA services are able to rapidly detect, help diagnose and in many cases mitigate problem conditions, thereby increasing the overall reliability and manageability of the grid.

The MonALISA architecture, presented in Figure 12, is based on four layers of global services. The network of Lookup Discovery Services (LUS) provides dynamic registration and discovery for all other services and agents. Each MonALISA service executes many monitoring tasks in parallel through the use of a multithreaded execution engine, and uses a variety of loosely coupled agents to analyze the collected information in real time.

[29] See http://evo.caltech.edu

The secure layer of Proxy services, shown in the figure, provides an intelligent multiplexing of the information requested by clients or other services. It can also be used as an Access Control Enforcement layer.

As has been demonstrated in round-the-clock operation over the last six years, the system integrates easily with a wide variety of existing monitoring tools and procedures, and is able to provide this information in a customized, self-describing way to any other set of services or clients.

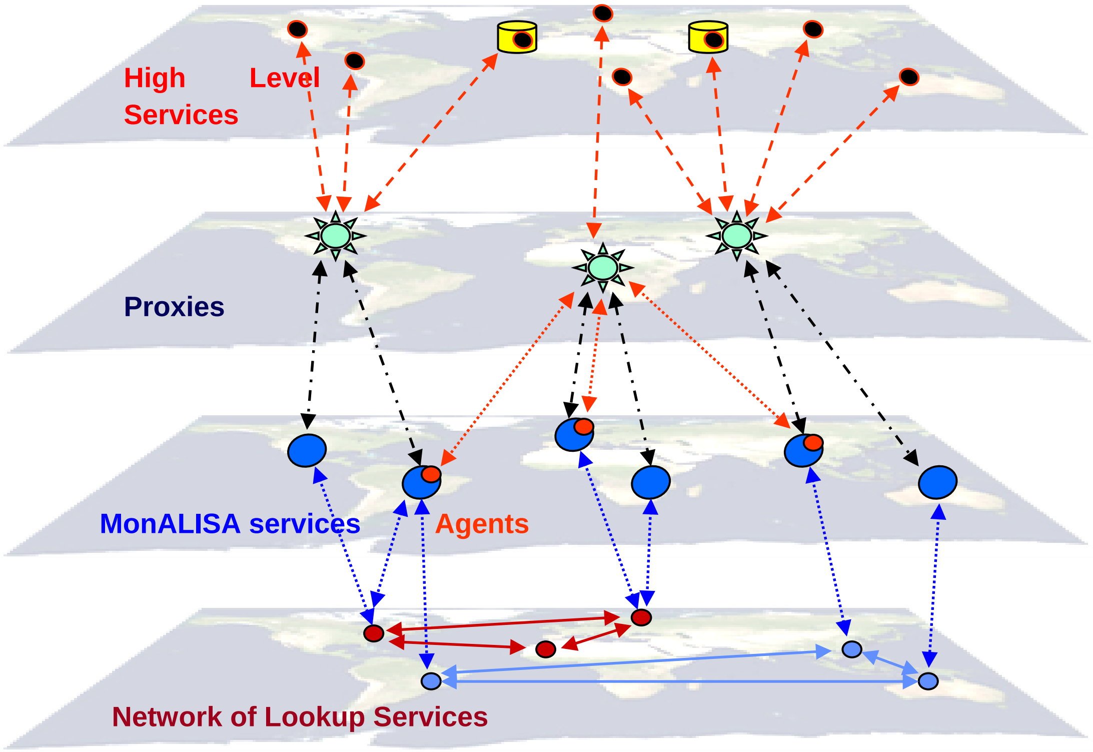


Figure 12: The four layers, main services and components of the MonALISA framework.

MonALISA Deployment in Grids

The MonALISA services currently deployed are used by the HEP community to monitor computing resources, running jobs and applications, different Grid services and network traffic.

MonALISA and its APIs are currently used by a wide range of grid applications in the High Energy Physics community:

For CMS it is used by the ARDA project for the CMS dashboard, and by all the job submission tools for analysis jobs (CRAB), production jobs (ProdAgent) and the Tier0 submission application for the main production activities at CERN. The system monitors detailed information on how the jobs are submitted to different systems, the resources consumed, and how the execution is progressing in real-time. It also records errors or component failures during this entire process.

In ALICE MonALISA is used to provide complete monitoring for their entire offline system, which is based on the “ALIEN” software. Here MonALISA is used to monitor jobs, facilities, experiment-specific services and all the data transfers. It also provides accounting of the resources used. Analysis elements, such as the XROOT servers and clients are instrumented with MonALISA APIs, and this near real-time information is used for load balancing during parallel interactive analysis. ALICE extensively uses MonALISA's ability to react to alarm conditions and rapidly take appropriate action, specifically to restart services which do not work correctly, and to control the overall submission of production jobs.

CMS and ATLAS are using MonALISA to monitor the traffic and usage for all the xrootd data servers.

For network monitoring the system allows one to collect, display and analyze a complete set of measurements and to correlate these measurements from different sites to present global pictures of WAN topology, delay in each segment, and an accurate measure of the available bandwidth between any two sites. As described in the previous Annexes and the following section, these particular functions will be extensively used in US LHCNet’s circuit services.

MonALISA Network Monitoring and Management

In order to build a coherent set of network management services it is very important to collect in near real-time information about the network traffic volume and its quality, and analyze the major flows and the topology of connectivity. Access to both real-time and historical data, as provided by MonALISA, also is important for developing services able to predict the usage pattern, to aid in efficiently allocating resources “globally” across a set of network links.

A large set of MonALISA monitoring modules has been developed to collect specific network information or to interface it with existing monitoring tools, including:

- SNMP modules for passive traffic measurements
- Active network measurements using simple ping-like measurements
- Tracepath-like measurements to generate the global topology of a wide area network
- Interfaces with the well-known monitoring tools MRTG, RRD, IPBM, PIPEs
- Data Transfer Applications such as GridFTP, xrootd, FDT
- Modules to collect dynamic NetFlow / Sflow information
- Available Bandwidth measurements using tools like pathload
- Dedicated modules for TL1 interfaces with CIENA’s CD/CIs, optical switches (GlimmerGlass and Calient) and GMPLS controllers (Calient)

These modules have been field-proven to function with a very high level of reliability over the last few years.

The way in which MonALISA is able to construct the overall topology of a complex wide area network, based on the delay on each network segment determined by tracepath-like measurements from each site to all other sites, is illustrated in Figure 13. The combined information from all the sites allows one to detect asymmetric routing, route instability or links with performance problems. For global applications, such as distributing large data files to many grid sites, this information is used to define the set of optimized replication paths.

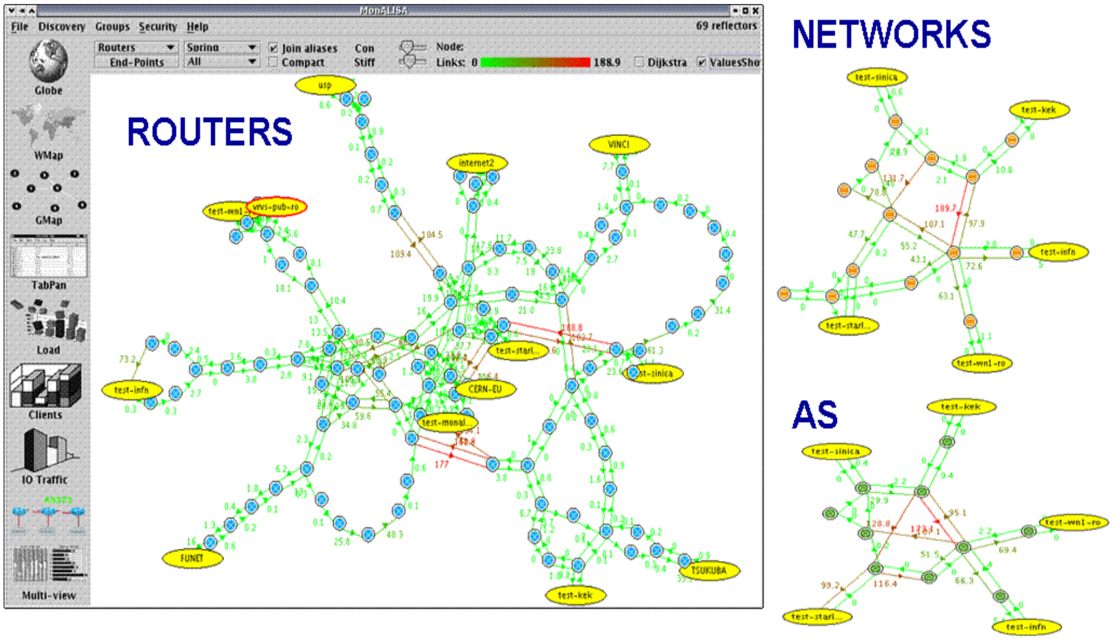
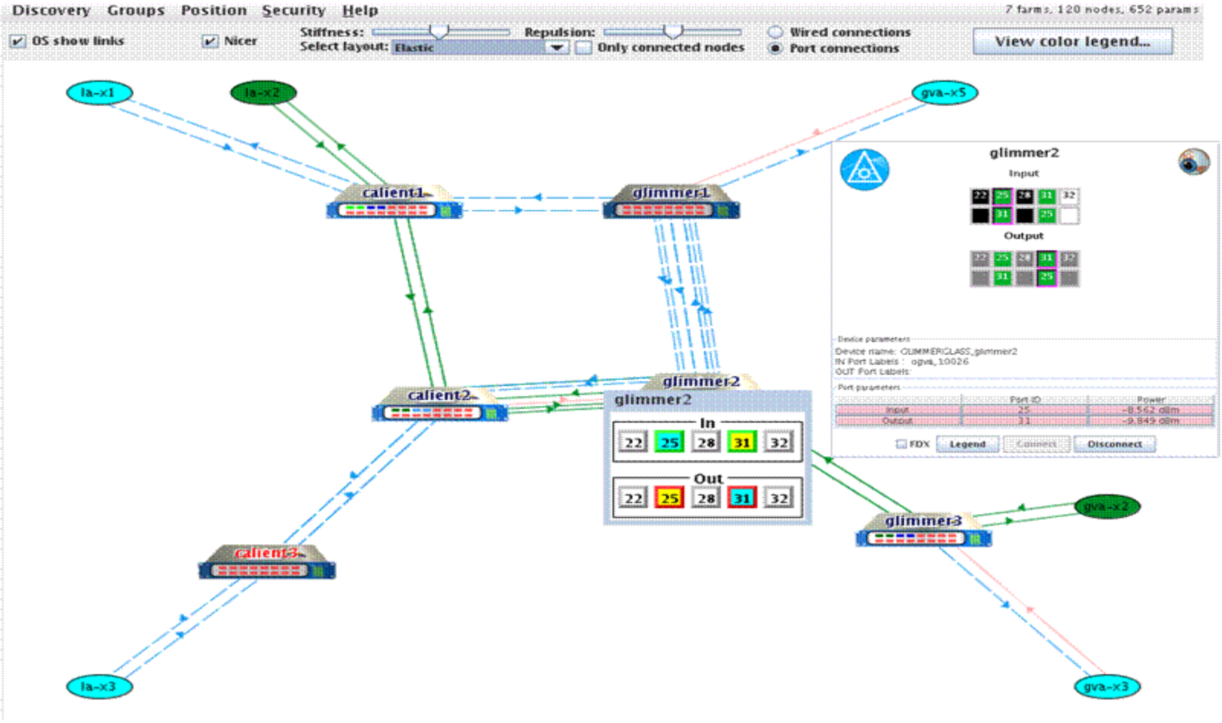


Figure 13: MonALISA real time view of the topology of WANs used by HEP. A view of all the routers, or just the network or "autonomous system" identifiers can be shown.

Specialized TL1 modules are used to monitor the power on Optical Switches and to present the topology. The MonALISA framework allows one to securely configure many such devices from a single GUI, to see the state of each link in real time, and to have historical plots for the state and activity on each link. It is also easy to manually create a path using the GUI. In Figure 14 we show the MonALISA GUI that is used to monitor the topology on Layer 0/1 connections and the state and optical power of the links.



Device parameters
Device name: GLIMMERGLASS_glimmer2
IN Port Labels : ogva_10026
OUT Port Labels:

Port parameters

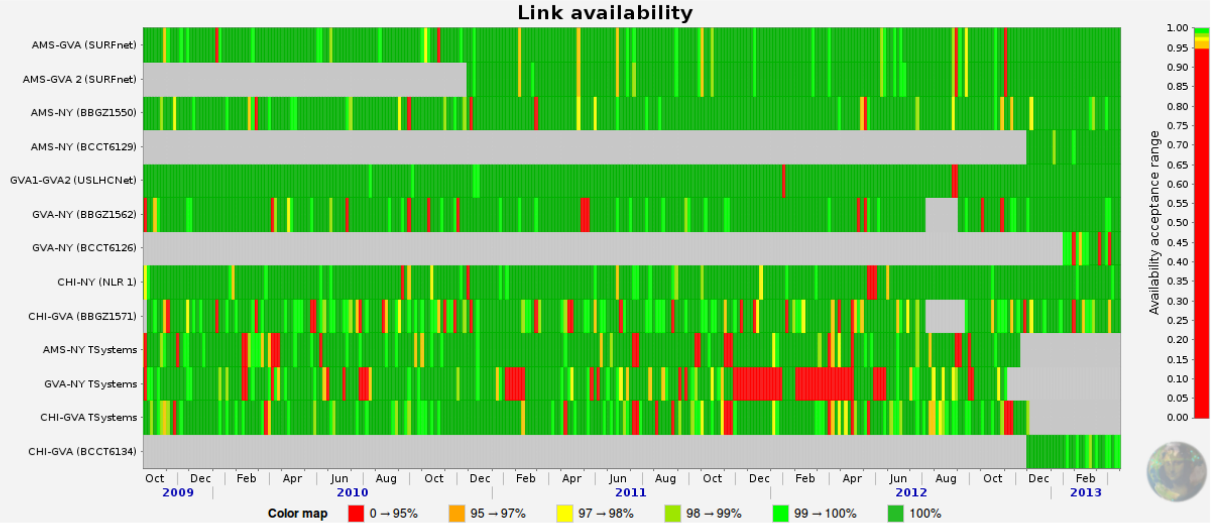
	Port ID	Power
Input	25	-8.562 dBm
Output	31	-9.849 dBm

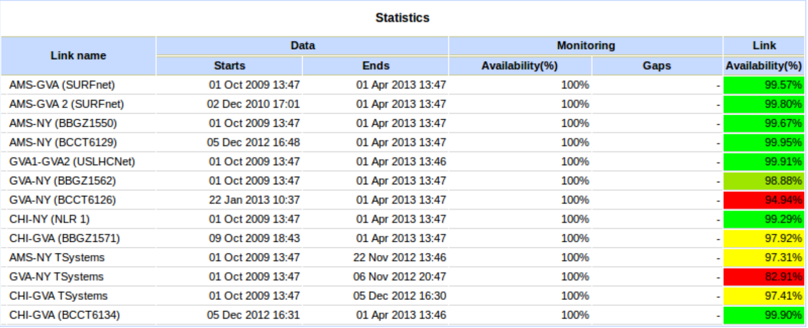
Figure 14: Monitoring and autonomous control for optical switches and optical links.

Monitoring USLHCNet

MonALISA is used to provide reliable, real-time monitoring of the USLHCNet infrastructure. In each point of presence (GVA, AMS, CHI, NYC) we run a MonALISA service to monitor the links, the network equipment and the peering with other networks. Each major link is monitored at both ends from two independent MonALISA services (the local one and one from a remote site). MonALISA services keep locally the history of all the measurements and a global aggregation, for long term history, is kept in a MonALISA repository. Dedicated TL1 modules for the Ciena CD/CI were developed to collect specific information on topology, dynamic circuits and operational status.

Link Status. We monitor the status for all WAN links and peering connections. For the Force10 switches we use SNMP and for the Ciena CD/CI the TL1 interface. The repository analyzes the status information from all the distributed measurements, for each segment, to generate reliable status information. Measurements are done every ~30s and the full history is kept in the repository database. The system allows one to transparently change the way a WAN is operated (via Force10 or Ciena CD/CI) and keeps consistent history. Figure 15 shows the panel that allows one to analyze the links' availability for any time interval.





Statistics

Link name	Data		Monitoring		Link
	Starts	Ends	Availability(%)	Gaps	Availability(%)
AMS-GVA (SURFnet)	01 Oct 2009 13:47	01 Apr 2013 13:47	100%	-	99.57%
AMS-GVA 2 (SURFnet)	02 Dec 2010 17:01	01 Apr 2013 13:47	100%	-	99.80%
AMS-NY (BBGZ1550)	01 Oct 2009 13:47	01 Apr 2013 13:47	100%	-	99.67%
AMS-NY (BCCT6129)	05 Dec 2012 16:48	01 Apr 2013 13:47	100%	-	99.95%
GVA1-GVA2 (USLHCNet)	01 Oct 2009 13:47	01 Apr 2013 13:46	100%	-	99.91%
GVA-NY (BBGZ1562)	01 Oct 2009 13:47	01 Apr 2013 13:47	100%	-	98.88%
GVA-NY (BCCT6126)	22 Jan 2013 10:37	01 Apr 2013 13:47	100%	-	94.94%
CHI-NY (NLR 1)	01 Oct 2009 13:47	01 Apr 2013 13:47	100%	-	99.29%
CHI-GVA (BBGZ1571)	09 Oct 2009 18:43	01 Apr 2013 13:47	100%	-	97.92%
AMS-NY TSystems	01 Oct 2009 13:47	22 Nov 2012 13:46	100%	-	97.31%
GVA-NY TSystems	01 Oct 2009 13:47	06 Nov 2012 20:47	100%	-	82.91%
CHI-GVA TSystems	01 Oct 2009 13:47	05 Dec 2012 16:30	100%	-	97.41%
CHI-GVA (BCCT6134)	05 Dec 2012 16:31	01 Apr 2013 13:46	100%	-	99.90%

Figure 15: Monitoring the status of major links.

The link availability monitoring in Figure 4 shows a full history over the last four years of the circuits we used in USLHCNet the overall percentage of uptime. It is important to note the monitoring availability for all these circuits was 100% during this period. This is a result of the MonALISA architecture which offers several layers of redundancy and is a really stable distributed service system.

Traffic Monitoring. We monitor the total traffic on all the Force 10 ports and on the Ethernet ports on the CIENA CD/CIs. The traffic on Ciena virtual circuits is also monitored by dedicated modules in MonALISA. Different aggregated views are presented, such as the total traffic on all the US LHCNet circuits, as well as integrated traffic over any time interval, as were shown in Figure 6 and Figure 7. Figure 16, shows the traffic through US LHCNet circuits indicating short-term bursts on each circuit. This view reveals the presence of such bursts close to 9 Gbps, when a transfer reaches the maximum capacity of an OC-192 link. While most monitoring systems do not register such peaks, usually reporting longer term averages, the MonALISA monitoring service clearly shows that such peaks saturating link capacity do occur.

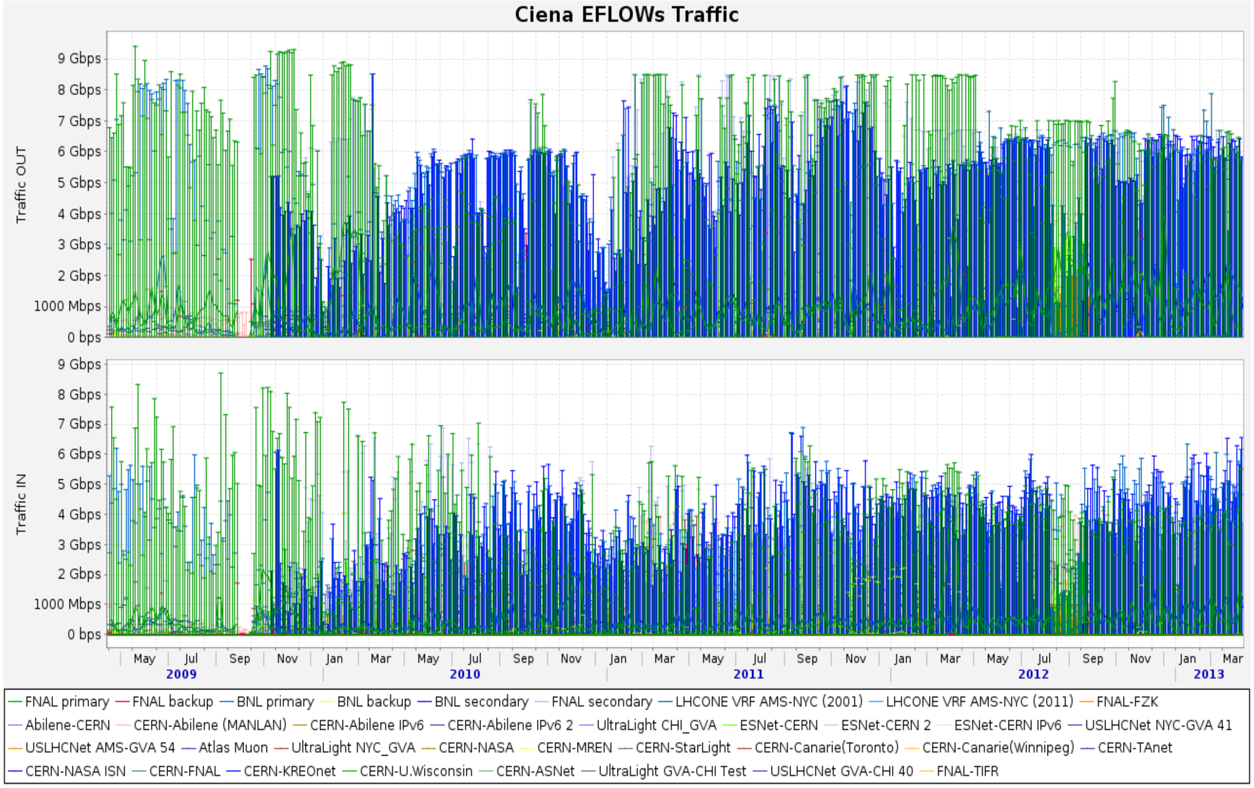
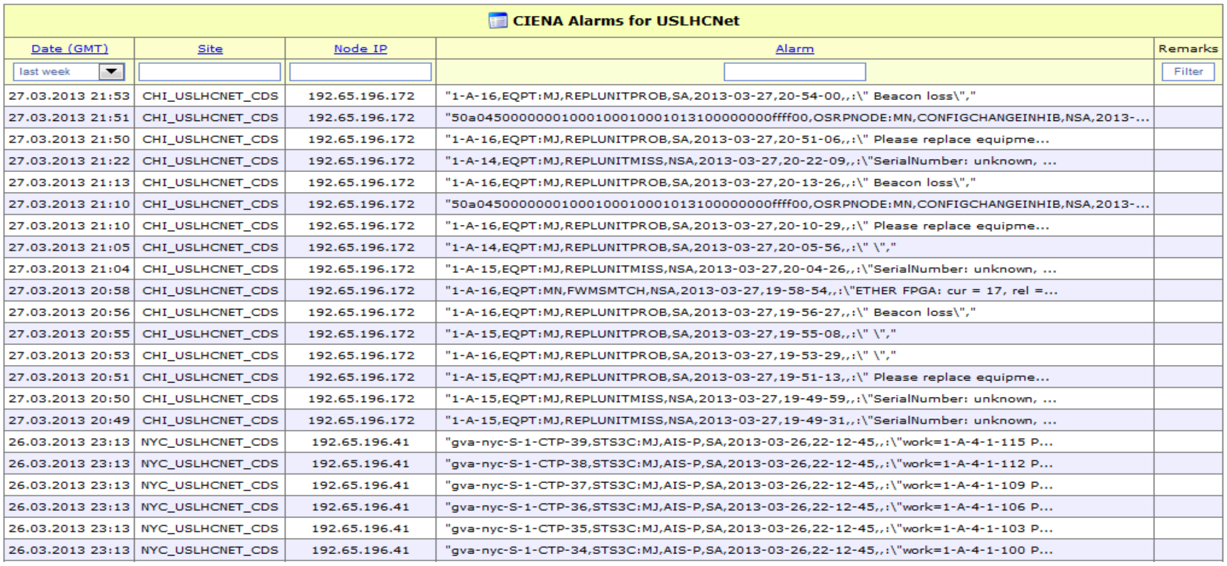


Figure 16: Traffic history for all the circuits in USLHCNet. A large number of peaks of 7-9 Gbps is observed.

Alarms and Notification. The operational status for the Force10 ports and all the Ciena CD/CI alarms are recorded by the MonALISA services. The alarms are analyzed and SMS/email notifications are generated based on different error conditions. We also "monitor" the services used to collect monitoring information. A global repository for all these alarms is available on the MonALISA servers, which allows one to select and sort the alarms based on different conditions. Figure 17 presents the panel which allows one to analyze the alarms from the entire system.

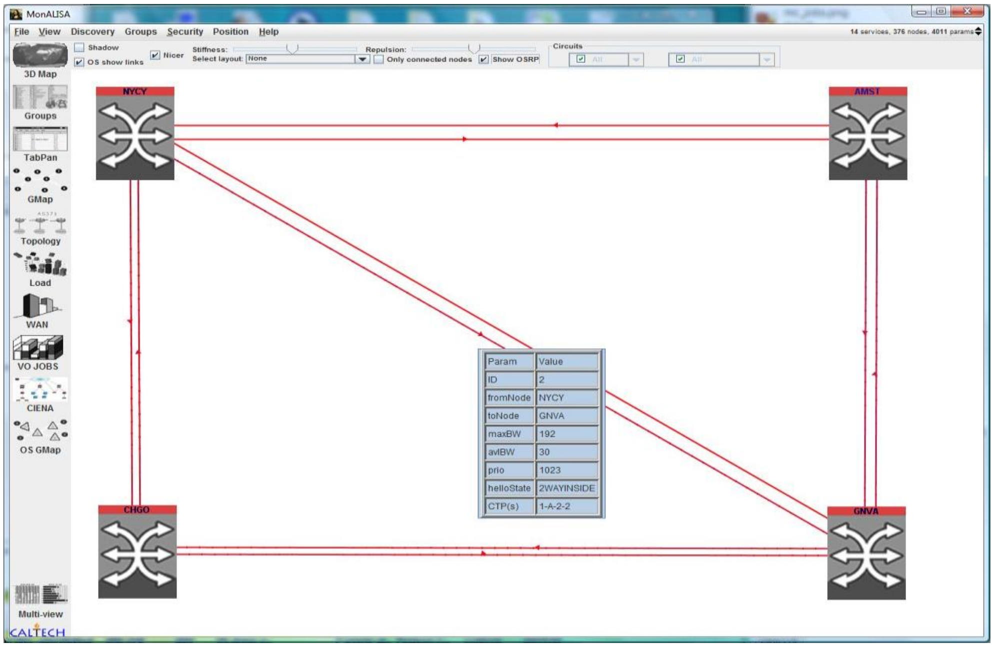


CIENA Alarms for USLHCNet

Date (GMT)	Site	Node IP	Alarm	Remarks
last week ▾				Filter
27.03.2013 21:53	CHI_USLHCNET_CDS	192.65.196.172	"1-A-16,EQPT:MJ,REPLUNITPROB,SA,2013-03-27,20-54-00,,:\" Beacon loss\","	
27.03.2013 21:51	CHI_USLHCNET_CDS	192.65.196.172	"50a0450000000100010001000101310000000ffff00,OSRPNODE:MN,CONFIGCHANGEINHIB,NSA,2013-...	
27.03.2013 21:50	CHI_USLHCNET_CDS	192.65.196.172	"1-A-16,EQPT:MJ,REPLUNITPROB,SA,2013-03-27,20-51-06,,:\" Please replace equipme...	
27.03.2013 21:22	CHI_USLHCNET_CDS	192.65.196.172	"1-A-14,EQPT:MJ,REPLUNITMISS,NSA,2013-03-27,20-22-09,,:\"SerialNumber: unknown, ...	
27.03.2013 21:13	CHI_USLHCNET_CDS	192.65.196.172	"1-A-16,EQPT:MJ,REPLUNITPROB,SA,2013-03-27,20-13-26,,:\" Beacon loss\","	
27.03.2013 21:10	CHI_USLHCNET_CDS	192.65.196.172	"50a0450000000100010001000101310000000ffff00,OSRPNODE:MN,CONFIGCHANGEINHIB,NSA,2013-...	
27.03.2013 21:10	CHI_USLHCNET_CDS	192.65.196.172	"1-A-16,EQPT:MJ,REPLUNITPROB,SA,2013-03-27,20-10-29,,:\" Please replace equipme...	
27.03.2013 21:05	CHI_USLHCNET_CDS	192.65.196.172	"1-A-14,EQPT:MJ,REPLUNITPROB,SA,2013-03-27,20-05-56,,:\" \","	
27.03.2013 21:04	CHI_USLHCNET_CDS	192.65.196.172	"1-A-15,EQPT:MJ,REPLUNITMISS,NSA,2013-03-27,20-04-26,,:\"SerialNumber: unknown, ...	
27.03.2013 20:58	CHI_USLHCNET_CDS	192.65.196.172	"1-A-16,EQPT:MN,FWMSMTCH,NSA,2013-03-27,19-58-54,,:\"ETHER FPGA: cur = 17, rel =...	
27.03.2013 20:56	CHI_USLHCNET_CDS	192.65.196.172	"1-A-16,EQPT:MJ,REPLUNITPROB,SA,2013-03-27,19-56-27,,:\" Beacon loss\","	
27.03.2013 20:55	CHI_USLHCNET_CDS	192.65.196.172	"1-A-15,EQPT:MJ,REPLUNITPROB,SA,2013-03-27,19-55-08,,:\" \","	
27.03.2013 20:53	CHI_USLHCNET_CDS	192.65.196.172	"1-A-16,EQPT:MJ,REPLUNITPROB,SA,2013-03-27,19-53-29,,:\" \","	
27.03.2013 20:51	CHI_USLHCNET_CDS	192.65.196.172	"1-A-15,EQPT:MJ,REPLUNITPROB,SA,2013-03-27,19-51-13,,:\" Please replace equipme...	
27.03.2013 20:50	CHI_USLHCNET_CDS	192.65.196.172	"1-A-15,EQPT:MJ,REPLUNITMISS,NSA,2013-03-27,19-49-59,,:\"SerialNumber: unknown, ...	
27.03.2013 20:49	CHI_USLHCNET_CDS	192.65.196.172	"1-A-15,EQPT:MJ,REPLUNITMISS,NSA,2013-03-27,19-49-31,,:\"SerialNumber: unknown, ...	
26.03.2013 23:13	NYC_USLHCNET_CDS	192.65.196.41	"gva-nyc-S-1-CTP-39,STS3C:MJ,AIS-P,SA,2013-03-26,22-12-45,,:\"work=1-A-4-1-115 P...	
26.03.2013 23:13	NYC_USLHCNET_CDS	192.65.196.41	"gva-nyc-S-1-CTP-38,STS3C:MJ,AIS-P,SA,2013-03-26,22-12-45,,:\"work=1-A-4-1-112 P...	
26.03.2013 23:13	NYC_USLHCNET_CDS	192.65.196.41	"gva-nyc-S-1-CTP-37,STS3C:MJ,AIS-P,SA,2013-03-26,22-12-45,,:\"work=1-A-4-1-109 P...	
26.03.2013 23:13	NYC_USLHCNET_CDS	192.65.196.41	"gva-nyc-S-1-CTP-36,STS3C:MJ,AIS-P,SA,2013-03-26,22-12-45,,:\"work=1-A-4-1-106 P...	
26.03.2013 23:13	NYC_USLHCNET_CDS	192.65.196.41	"gva-nyc-S-1-CTP-35,STS3C:MJ,AIS-P,SA,2013-03-26,22-12-45,,:\"work=1-A-4-1-103 P...	
26.03.2013 23:13	NYC_USLHCNET_CDS	192.65.196.41	"gva-nyc-S-1-CTP-34,STS3C:MJ,AIS-P,SA,2013-03-26,22-12-45,,:\"work=1-A-4-1-100 P...	

Figure 17: Global repository for the Ciena CD/CI alarms. MonALISA provides a user friendly interface to sort and analyze them as needed.

Network Topology. For the Ciena CD/CI nodes, MonALISA provides real-time information for the OSRP connections with all the attributes for the SONET links (illustrated in Figure 18).



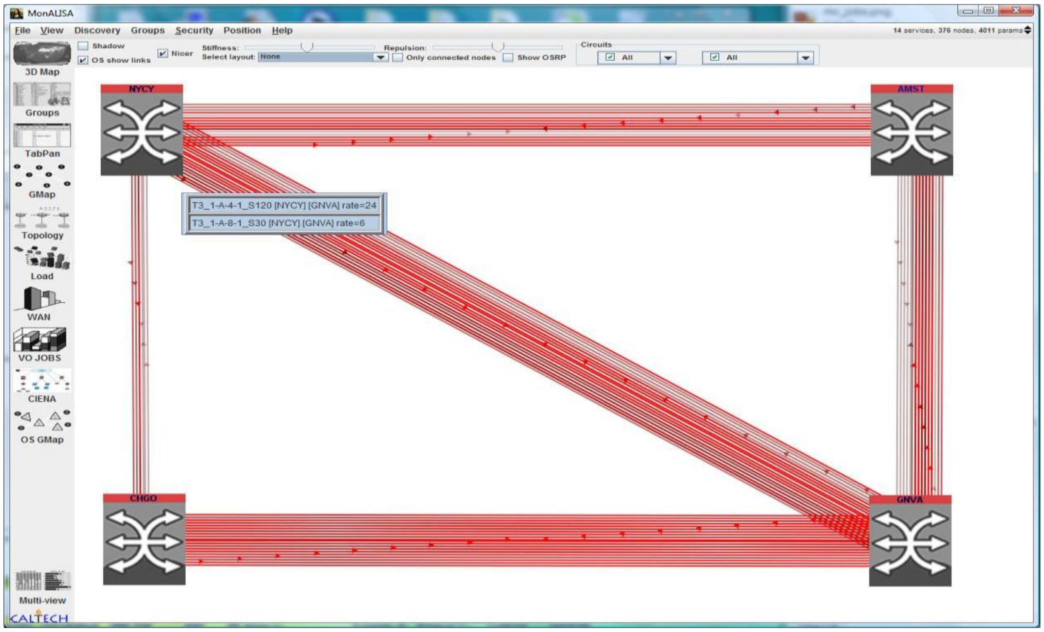
14 services, 376 nodes, 4011 params

☐ Shadow
☑ OS show links
☑ Nicer
Select layout: None
☐ Only connected nodes
☑ Show OSRP

Param	Value
ID	2
fromNode	NYCY
toNode	GNVA
maxBW	192
avlBW	30
prio	1023
helloState	2WAYINSIDE
CTP(s)	1-A-2-2

Figure 18: The Physical topology for the Ciena SONET links

Topology of dynamic circuits. The topology of all the circuits created in the entire network is presented in real time in the MonALISA interactive client. This panel allows one to select any set of circuits, and it presents how they are mapped onto the physical network, with all their attributes (as shown in Figure 19).



T3_1-A-4-1_S120 [NYCY] [GNVA] rate=24

T3_1-A-8-1_S30 [NYCY] [GNVA] rate=6

Figure 19: The topology for the dynamic circuits in the entire network and their attributes, presented in compact graphical form.

Development of an automated network management system

US LHCNet is continuously striving to provide best possible performance, including shortening the time to repair in case of outages. In some contingency planning scenarios, we foresee the migration of circuits between CIENA and Force10 devices, which today requires manual intervention and remote hands support in remote PoPs. A possible improvement in this respect, significantly shortening re-provisioning time in major outage scenarios, is an automated system capable of reconfiguring interconnects at the optical level between the US LHCNet devices and long-haul links.

The MonALISA team has started the development of an automated network management system using the Telescent Light switch optical patch panel[30]. A prototype Telescent optical switch was deployed at Caltech and CERN for a two weeks period and was integrated into a test environment for the USLHCnet infrastructure (Figure 20).

[30] www.telescent.com

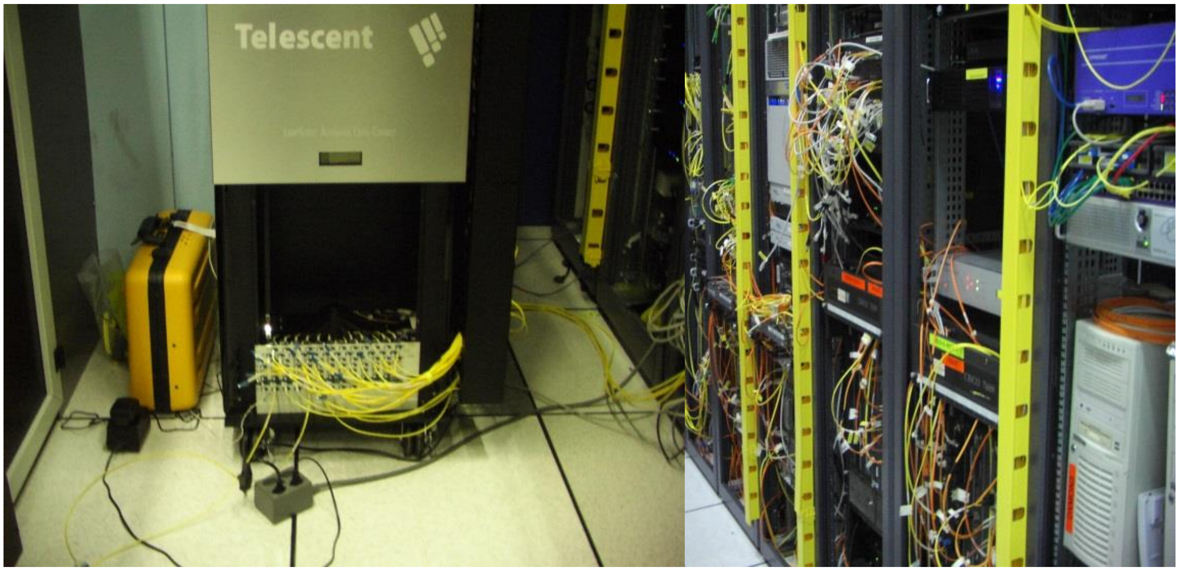
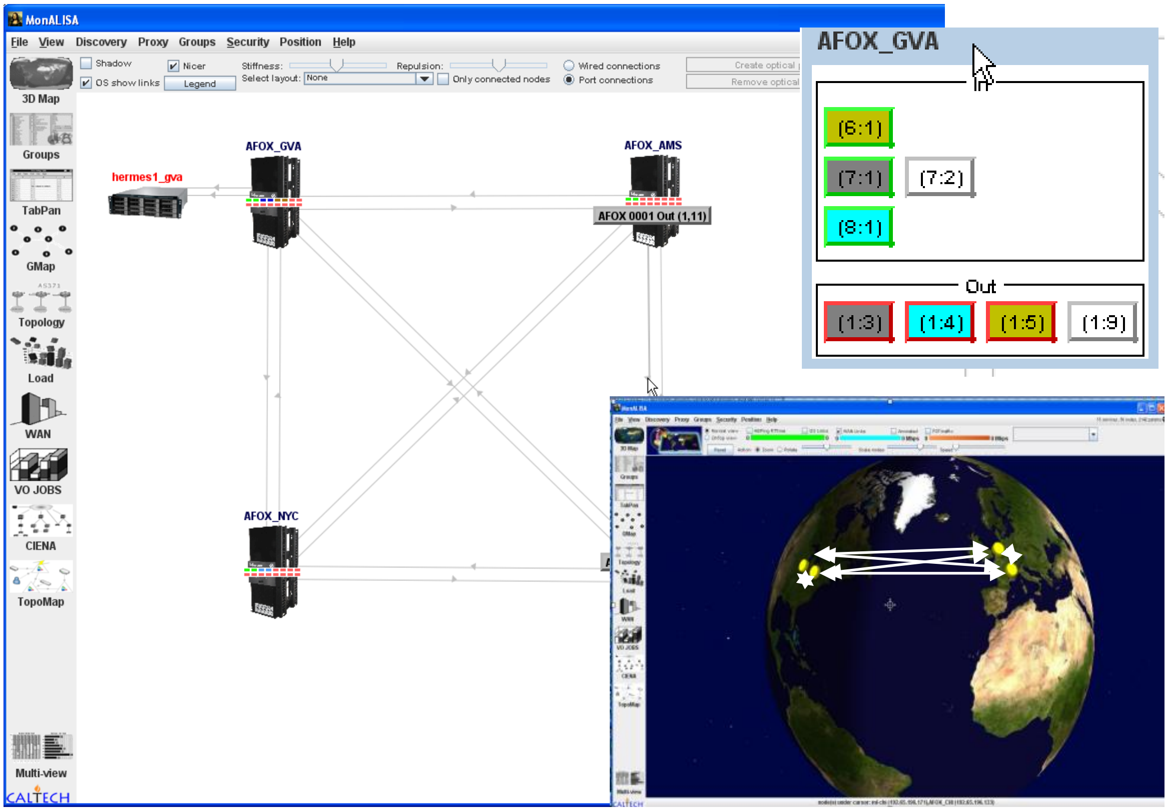


Figure 20: The Telescent optical switch was integrated into a test environment in the USLHCnet infrastructure at CERN.

A set of MonALISA dedicated prototype modules were developed to monitor and control the Telescent optical patch panel. These modules are based on a set of java based APIs that are used to communicate with the switch firmware. The monitoring modules are currently used to get the connectivity matrix for the switch. The control modules can send reconfiguration commands to the switch.

For the development and testing of a global management system, the Telescent switch was logically divided into four sub-switches and each one was monitored and controlled independently by a MonALISA service. The topology for such distributed setups is currently done using configuration files for each switching unit, but it can be extended to use the RFID information from the Telescent switch as soon as this will be available.

Figure 21 presents the topology GUI in the MonALISA framework for global systems. It presents all the connected links and the interfaces used for each device.



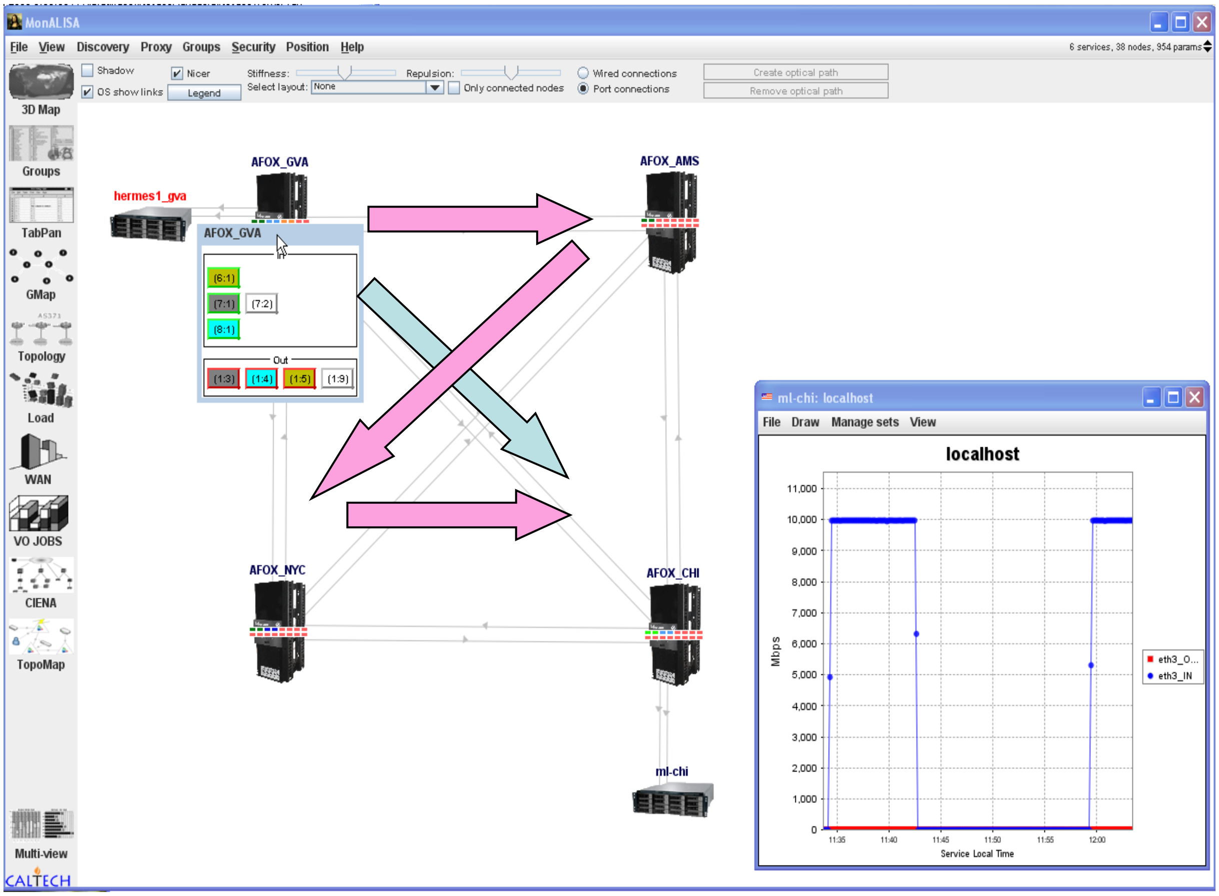
☐ Shadow
☑ OS show links
☑ Nicer
Stiffness:
Repulsion:
Select layout: None
☐ Only connected nodes
○ Wired connections
● Port connections

Figure 21: The MonALISA GUI presenting global topology for four Telescent sub-switches.

We used this setup to test different reconfigurations using four virtual switches. Two servers were connected to the switch (ml-chi and hermes3-gva) and we used a set of cross connects to simulate the connectivity between the virtual switches. The Fast Data Transfer[31] (FDT) application was used to send data between the two systems. In

Figure 22 we present topology reconfiguration and the traffic between the two servers. As soon as the reconfiguration was done, the FDT transfer recovered. The reconfiguration time for this test is quite long because all the operations and done on the same physical switch and it requires two fire connections per link. The switching time per fiber will also improve in the next versions of the Telescent switches.

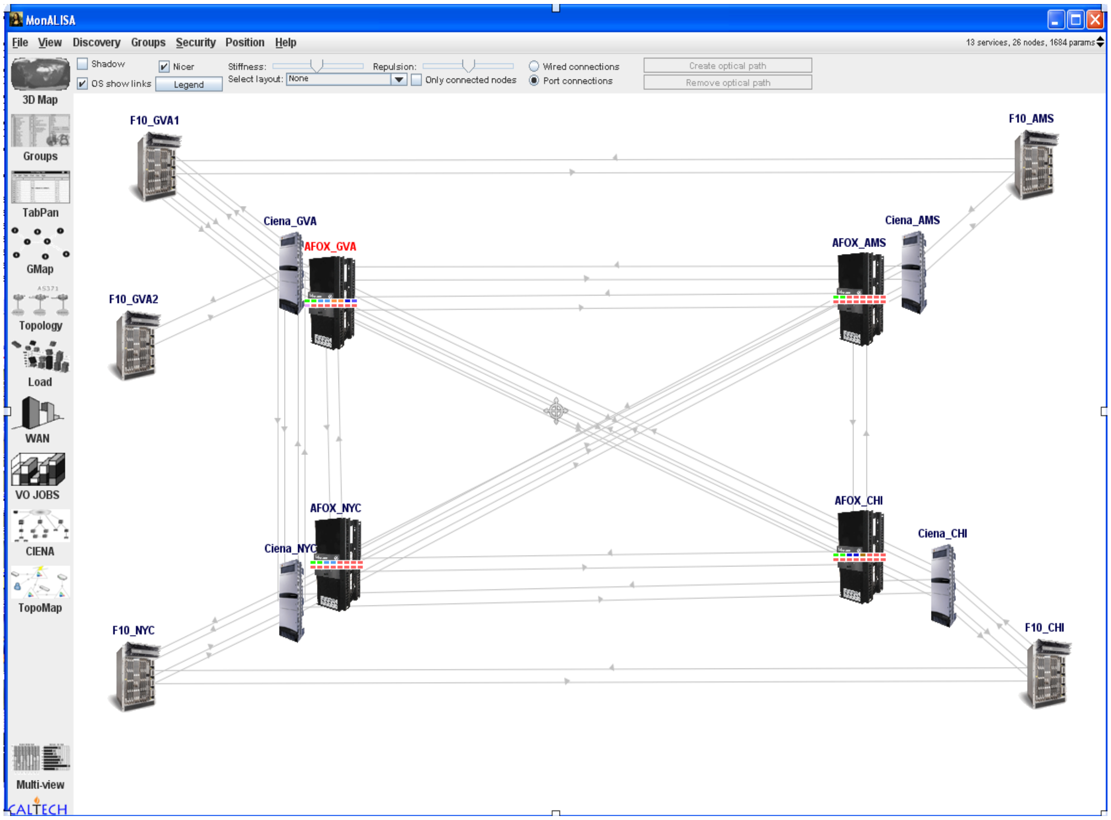
[31] FDT web page : http://monalisa.cern.ch/FDT



□ Shadow
☑ OS show links
☑ Nicer
Stiffness:
Repulsion:
Select layout: None
□ Only connected nodes
○ Wired connections
● Port connections

Figure 22: A global reconfiguration for the four sub-switches. Initially the two systems were connected using the direct link (blue arrow). The network was then reconfigured to connect the two systems using all four sub systems (pink arrows). The data transfer between the two servers recovered once the reconfiguration was done.

To build a global network management system, we need to integrate different types of network devices and complex topologies. Dedicated MonALISA modes are under development to provide status and connectivity information for different types of routers and switches. This information is used together with the Layer 0 connectivity maps from the optical switches. In Figure 23 is presented a simulation of the US LHCNet topology which includes several types of network devices. These global views will be used to develop higher level services capable to take automatic action and generate the reconfigurations maps when we detect failures in connectivity or network equipment.



13 services, 26 nodes, 1684 params

☐ Shadow ☑ Nicer

☑ OS show links

Select layout: None

☐ Only connected nodes

○ Wired connections

● Port connections

Figure 23. A simulation of the global topology for USLHCNet.

Appendix B Future Planning

The LHC Long Shutdown 1 (LS1) period now underway provides an important opportunity to deploy and commission the necessary changes and upgrades to the network infrastructure, in time for the LHC startup. We propose to address this in two stages: in 2013, the team will execute the modifications to Layer 2 and Layer 3 services, and followed by the upgrades to the Layer 1 service in 2014.

3.1.1 Plan for 2013

In 2013, we will replace the aging Force10 e600 Terascale routers. Dell/Force10 has released the End of Life announcement for this platform in 2012.

While in the past US LHCNet has deployed a meshed routed network, for the next generation architecture we will consolidate the routing functionality in only one PoP, at CERN. This is in good agreement with the fact that the larger fraction of the services provided by US LHCNet are at Layer 2.

We have chosen the Brocade MLXe as the routing/switching hardware platform, for several reasons:

- It provides high-density 10GE, 40GE, and 100GE switching and routing capability
- It provides SDN[32] capability through port-based OpenFlow configuration
- It supports Virtual Routing and Forwarding (VRF), enabling
- It supports WANPHY operation, consistent with US LHCNet emergency requirements
- Brocade is one of the leading routing and switching hardware providers
- CERN operates a large Brocade MLXe installation, including CERN as well as the LCG backbones, which through collaboration gives us access to a large know-how pool, and to the potential of on-site immediate spare parts loans in certain urgent cases involving the CERN site equipment.

The technical aspects of the Brocade MLXe, including all the main functions and the ability to operate flawlessly at full performance, have been thoroughly tried by the Caltech team during the very successful Supercomputing 2012 field trials and demonstration in Salt Lake City.

Apart from the issue of ageing hardware, an additional essential consideration guiding the plan for 2013 is that the functions summarized in the first points in the list above are not supported by the current routers used in US LHCNet. In particular the absence of Virtual Routing and Forwarding (VRF) support prevents US LHCNet from full integration in the LHCONE routed services. The replacement of the router at CERN in Geneva, foreseen for mid-2013, will enable us to connect to the LHCONE VRF infrastructure.

The new design is shown in Figure 24. The optical multiservice switches act at Layer 1 and Layer 2, providing switched and in particular virtual circuit connectivity. During 2013, we'll continue this operation using the existing CIENA CoreDirector switches. Routing functions will

[32] Software Defined Networking

be performed at only one PoP, which acts as the central peering point. Virtual circuits (or simple VLANs) are extended through the CoreDirector core to the external peering points at the other PoPs. This view is consistent with the main mission of US LHCNet and the fact that the majority of the traffic through US LHCNet is transported in virtual circuits, i.e. switched, not routed.

The single MLXe16 device, with the necessary number of 10GE ports, will replace the two Force10/Dell e600 routers currently in operation at the US LHCNet CERN PoP. For reasons of resilience, the team also will install a small, 1 RU high, router for backup operation in case of a major failure of the main device, or to take over during software upgrades (not shown in the picture).

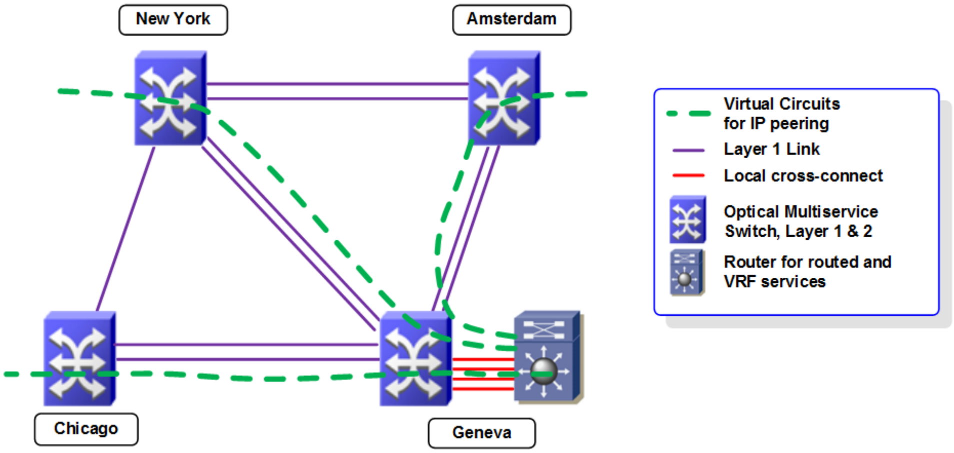

New York
Amsterdam
Chicago
Geneva
Virtual Circuits for IP peering
Layer 1 Link
Local cross-connect
Optical Multiservice Switch, Layer 1 & 2
Router for routed and VRF services

Figure 24: Next generation US LHCNet design with centralized routing.

In the second half of 2013, we intend to deploy the first two Layer 1 switches, initially in a test setup. We have chosen the Geneva (CERN) PoP, and Chicago (Starlight) as locations for the evaluation and integration of the first nodes. The detailed plan for this will be completed in the May-June timeframe, immediately followed by the installation, which is planned to start in July 2013.

The Software Defined Networking (SDN) paradigm targets a separation of control plane and forwarding plane functionality, whereby the controller code is provided by the operator or user. The trend towards SDN has rapidly gained traction, and become a major focus of development in the last two years, both in industry and in many R&E networks including Internet2 and ESnet. OpenFlow[33], a protocol for communication between the controller and network device originally developed at Stanford, that allows the controller code to be executed on an external server, has taken the lead in this trend and has already been implemented by several manufacturers. As there is much interest in OpenFlow in Internet2, ESnet, and other national research and education networks in the US and overseas, and as their plans for upcoming advanced network services

[33] See e.g. https://www.opennetworking.org/ The OpenFlow protocol serves to modify the forwarding database entries on the controlled device.

include the use of OpenFlow, we plan to deploy an OpenFlow network in US LHCNet at the same time as the proposed router platform upgrade. Then Caltech network engineering team is already working with OpenFlow, and one member of the team, M. Bredel who is the lead developer in the OLiMPS project, is developing OpenFlow and contributing the widely used Floodlight controller.

3.1.2 Outlook for 2014

In 2014, we plan to finalize the upgrade of the Layer 1 platform, by installing the remaining two optical nodes in New York (MANLAN) and Amsterdam (SARA). This will complete the new architecture as shown in Figure 24, and give us the possibility to upgrade to higher-capacity connections, probably 100Gbps, at our PoPs, in time for the LHC restart in 2015. We would then be positioned to upgrade to higher speed inter-PoP connections as needed, including higher speed transatlantic circuits as soon as they become cost-effective.