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

Abstract

Part I of this proposal presents the status and progress in 2006-7, and the technical and financial plans for 2008-2010 for the US LHCNet transatlantic network supporting U.S. participation in the LHC physics program. US LHCNet provides transatlantic connections of the Tier1 computing facilities at Fermilab and Brookhaven with the Tier0 and Tier1 facilities at CERN as well as Tier1s elsewhere in Europe and Asia. Together with ESnet, Internet2, the GEANT pan-European network, and NSF's UltraLight project, US LHCNet also supports connections between the Tier2 centers (where most of the analysis of the data will take place, starting this year) and the Tier1s as needed.

The plans presented for the next three years, which include an upgrade from the present New York-Chicago-Amsterdam-CERN network that includes three transatlantic 10 Gbps links to eight 10 Gbps links across the Atlantic by 2010 when the LHC program is expected to be in full swing, are a part of the multi-year plan to meet the requirements of the LHC experiments along with other major U.S. HEP programs that require networking between the U.S. and CERN in the most cost-effective manner. These plans, which include annual Requests for Proposals to obtain the best link pricing for a set of fully diverse network paths ensuring uninterrupted 24 X 7 network operation, have been developed and evolved in consultation with CERN, ESNet, Fermilab, BNL and the LHC experiments, and presented to DOE periodically since 2001. In 2006-7 the status, progress and plans have been discussed frequently by these partners, as well as with DOE, in the U.S. LHC Network Working Group formed for this purpose.

The LHC experiments and other major DOE-funded HEP programs face unprecedented engineering and organizational challenges due to the volumes and complexity of the data, and the need for scientists located at sites around the world, remote from the experiment, to work collaboratively on data analysis. LHC physicists in the U.S. face exceptional challenges as they are separated from the experimental site by 6-9 time zones.

It is now well established that national and international networks of sufficient (and rapidly increasing) bandwidth and end-to-end performance are the key to meeting many of these challenges. The Computing Models of the major HEP experiments are based on a grid of Tier0, Tier1 and Tier2 centers interconnected with high speed networks that support multi-Terabyte data transfers (a concept developed at Caltech in 1999 whose network aspects have been refined in US LHCNet as well as in the NSF DISUN and UltraLight projects since 2004).

The US LHCNet transatlantic network is a lynchpin in the global ensemble of networks used by the HEP community today, and an essential resource for US participation in the LHC. The current US LHCNet program and plan, led by Caltech, has evolved from DOE-funded support and management of international networking between the US and CERN dating back to 1985, as well as a US-DESY network in the early 1980's. US LHCNet today consists of a set of 10 Gbps links interconnecting CERN, MANLAN[1] in New York and Starlight[2] in Chicago. The network has been architected to ensure efficient and reliable use of the 10 Gbps bandwidth of each link, up to relatively high occupancy levels, to cover a

[1] The MANLAN exchange point is designed to facilitate peering among US and international research and education networks in New-York. See http://networks.internet2.edu/manlan/

[2] StarLight is an international peering point for research and education networks in Chicago. See http://www.startap.net/starlight

wide variety of network tasks, including: large file transfers, grid applications, data analysis sessions involving client-server software as well as simple remote login, network and grid R&D-related traffic, videoconferencing, and general Internet connectivity.

Caltech shares with CERN the responsibility for the implementation, operation and management of the US-CERN network, as well as the peerings with ESnet and the major US academic network backbones (in particular Abilene and more recently National Lambda Rail). Funding for the network bandwidth, the routing and switching equipment and other infrastructure required for network operation, is shared by Caltech (under a DOE/HEP grant) and CERN.

The effectiveness of the LHC ensemble of networked computing and storage facilities is being driven, and the appropriate scale of networking to enable the physics program, have seen set by a convergence of several related factors: (1) the emergence of affordable network technologies supporting gigabit/sec (Gbps) and 10 Gbps links in both local and wide-area networks, notably gigabit and 10 gigabit Ethernet, (2) dense wavelength division multiplexing (DWDM) that supports multiple high-bandwidth “wavelengths” on an optical fiber-pair over long distances, and (3) advances by members of the HEP community working with network engineers and computer scientists, exploiting advances in network protocols, computing systems and network interfaces, and the Linux kernel to achieve multi-Gigabit per second throughput over long distances.

These developments are accelerating HEP’s large scale use and dependence on long-range networks. This is particularly apparent in the series of “service challenges” in recent years, which have given way in 2006-7 to “computing, storage and analysis challenges” involving increasingly large data transfers coordinated with distributed processing and storage in parallel with data access and analysis by many hundreds of physicists. These challenges mark the ramp-up of operations of the Tiered centers to the scale required to support data-taking at the LHC starting in 2008.

The latest developments that will continue these trends, and drive the continued evolution of efficient network usage and management for the LHC and other DOE-funded programs include: (1) the integration of the high-throughput data transfer methods mentioned above with the major storage systems, starting with the widely-used dCache system developed by Fermilab and DESY, and (2) the implementation of a new networking infrastructure on US LHCNet and its partner networks (ESnet and Internet2) that support virtual circuits with bandwidth guarantees to support the largest and highest priority data transfer tasks. In US LHCNet these virtual circuits will be built around a core of advanced CIENA optical multiplexers and the use of emerging network standards developed for this purpose. The construction of the services required to form and manage these virtual circuits links are being developed by the combined efforts of network teams supported by DOE-funded and NSF-funded projects, working in concert[3]: ESNet (OSCARS), Internet2 (DRAGON), Fermilab and Caltech (LambdaStation), Brookhaven along with Michigan and SLAC (Terapaths), UltraLight and US LHCNet.

[3] These efforts also include the DICE consortium composed of the GEANT2 pan-European network, the national research and education networks in Europe, ESnet and Internet2.

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

Table of contents

Part I: US LHCNet: Transatlantic Networking for the LHC................................1
Table of contents..3
1 Introduction and US LHCNet Mission...4
1.1 Role of the Caltech Group...4
1.2 Transatlantic Network Needs for LHC and HEP..6
1.3 Meeting the Needs: Dynamic Virtual Circuits with Bandwidth Guarantees...10
1.4 US LHCNet Status, Current Activities and Future plans 11
1.5 Team Organization and Activities..13
1.6 Manpower Profile and Funding Plan..15
2 Project status..17
2.1 Working Methodology: Production and "High-Impact" Networking 17
2.2 Technical Status (July 2007)...18
2.3 Recent Milestones in 2006-2007, and Planned Milestones for 2007-2008.....21
2.4 Caltech-CERN Team Organization and Breakdown of Team Activities........25
3 US LHCNet Bandwidth and Funding Plan Through 2010......................................28
4 External Collaborations 32
5 Conclusion 35

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. Referring to the global HEP collaborations of 500 to 2000 physicists each from 10-40 countries and 50-180 institutions, a well known physicist[4] summed it up by saying:

"Collaborations on this scale would never have been attempted, if they could not rely on excellent networks."

This trend has been accelerated by the recent 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 over national and transoceanic distances increasingly affordable, and cost-effective.

The effectiveness of US participation in the LHC experimental program is particularly dependent on the speed and reliability of our national and international networks. As we approach the startup of the LHC, the ability of scientists to move large amounts of data, access computing and data resources, and collaborate in real time from multiple remote locations require unprecedented network performance and reliability.

In addition to the application of state of the art high-throughput methods and tools, US LHCNet has been designed to meet these needs by providing a high performance network with 99.9+% 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).

1.1 Role of the Caltech Group

The Caltech group first proposed the use of international networks for high energy and nuclear physics (HENP) 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 ("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 more than 10 years, the network has been co-managed and co-operated by the CERN and Caltech network engineering teams. Since November 2006, Caltech and CERN share management and operations responsibility for the "US LHCNet" consortium, with Caltech having the primary responsibility for management and operations of the the transatlantic and intra-U.S. links among the points of presence in New York, Chicago, CERN and Amsterdam, and for the three points of presence outside of CERN.

[4] Larry Price, Argonne National Laboratory, in the TransAtlantic Network (TAN) Working Group Report, October 2001; see http://gate.hep.anl.gov/lprice/TAN

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 moderate increase in cost, by exploiting favorable long-term trends in market pricing per unit bandwidth, especially along the highest capacity transatlantic routes.

The timing of this plan is designed to be as late as possible, to reduce costs while meeting the needs of the LHC experiments, while also foreseeing the time required to deploy, test, and make production-ready the new optical multiplexers and network services required to meet the LHC experiments' needs at each stage of development. Ongoing development is foreseen to effectively utilize and manage the available bandwidth on a scale that increases as the volume of the data and the corresponding data transport needs increase, as the number of affordable links in US LHCNet increases, as the LHC experiments' distributed computing systems mature, and as the state of network and server technologies continue to progress in capability and cost-effectiveness.

Since 2001, US LHCNet has become a leading developer in the use of TCP-based data transfers over long distance networks. We are also engaged, together with leading partner groups in the network and computer science communities[5], in advanced R&D programs that are developing a new set of end-to-end managed network services integrated with grid systems, to ensure that the available bandwidth across the U.S. and the Atlantic, on the increasing scale required in the LHC era, can be utilized and managed effectively. These programs exploit new optical network technologies, developments of new protocol stacks and streaming-dataflow tools, and a new generation of circuit-oriented network services with bandwidth guarantees, in combination with the major storage systems of the LHC experiments, to ensure predictable, sustained high throughput for transporting multi-Terabyte datasets over long distances.

Following a Request for Proposals issued in May 2006, Caltech negotiated new cost-effective contracts for 2007 for three transatlantic links, under Caltech's responsibility, thereby upgrading the transatlantic US LHCNet bandwidth from 20 to 30 Gbps, according to plan. In order to provide a redundant, uninterruptible service, we have implemented three physically diverse paths across the Atlantic (the diversity including the end station equipment in each case). The US LHCNet transatlantic circuits currently in service are shown in Figure 1.

[5] Including SLAC, Fermilab, Brookhaven, ESnet, Internet2, National Lambda Rail, StarLight, MANLAN, and Netherlight, Michigan, Florida, Caltech's Network Laboratory led by Steven Low, the NSF-funded Ultralight, FAST-TCP, PLaNetS and DRAGON projects, and the DOE-funded LambdaStation, Terapaths and OSCARS projects.

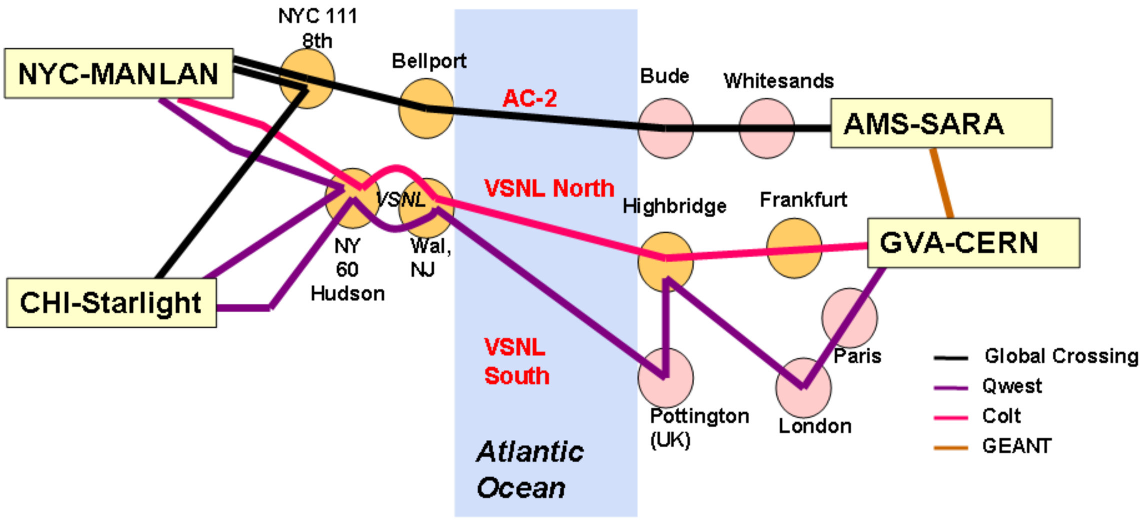


Figure 1 US LHCNet Transatlantic Circuits

In July 2007, Caltech issued another Request for Proposals for the US LHCNet links in FY2008. Following the responses to the RFP (due in late August), and following the US LHCNet plan, we expect to upgrade the network to include four transatlantic links along at least three fully diverse physical network paths. This year we are also opening up the option of concatenating transatlantic links terminating at the US LHCNet PoP in Amsterdam, or GEANT2 PoPs in London or Paris, with links provided at relatively low cost by GEANT2 between these PoPs in Europe and CERN. This enhanced strategy is being adopted because of its potential to deliver full diversity of the network paths including the intra-European segments coming into CERN, and/or lower overall cost.

1.2 Transatlantic Network Needs for LHC and HEP

The baseline bandwidth needs of the LHC experiments are periodically reviewed and discussed extensively within our community, and updated as needed. The conclusions up to now are summarized in Table 1, as presented and adopted by the recently created US LHC Network Working Group[6] which includes members from CERN, ESnet, Caltech, FNAL, BNL, ATLAS, CMS, DOE and Internet2. These requirements are put in place through our close collaboration with the US Tier1s (BNL and Fermilab) as well as ESNET, GEANT2 and the other members of the LHC OPN[7] and are consistent with the current funding plan and the current circuit costs. The provisioning of the circuits and the commissioning of services over the US LHCNet network will be just in time for the upcoming startup of the LHC in July 2008.

The roadmap summarized in the table updates earlier estimates of the baseline network

[6] The US LHC Network Working Group met for the first time at CERN on July 7th 2005. Its members will continue to meet regularly, as needed, to evaluate network needs and coordinate efforts between CERN, the US Tier1s and Tier2s, ESnet, US LHCNet, and other transatlantic networks including CANARIE (Canada), the IRNC links funded by NSF, the GEANT2 pan-European network, etc.

[7] LHC Optical Private Network – the sum of all links designed to carry Tier0 to Tier1 traffic

requirements for HEP, originally determined by the ICFA Network Task Force in 1997-1999, the Transatlantic Network Working Group in 2000-2001, and by the ICFA Standing Committee on International Connectivity (SCIC[8]) in 2002-2006. In 2005-6 the bandwidth profile in the plan was stretched out by more than a year, to match the LHC schedule, and to provide the necessary level of just-in-time provisioning. As discussed above, the plan represented in the table includes just-sufficient time for deployment, commissioning and preparations for full-scale production operations year by year. The upper half of the table gives the projected bandwidth requirements for CMS and ATLAS between the US and CERN, along with other requirements for transatlantic networking. The corresponding projections of the transatlantic bandwidth to be installed that could be used to meet HEP's needs, including US LHCNet and other links, are given in the lower half of the table.

Table 1. Transatlantic Network Requirements Estimates and Bandwidth Provisioning Plan, from the T0/T1 networking group, in Gbps

Year	2006	2007	2008	2009	2010
CERN-BNL (ATLAS)	5	15	20	30	40
CERN-FNAL (CMS)	15	20	20	30	40
Other (ALICE, LHCb, LARP, Tier1-Tier2, Development, Inter-Regional Traffic, etc.)	10	10	10-15	20	20-30
TOTAL US-CERN BW	30	45	50-55	80	100-110
US LHCNet Bandwidth	20	30	40	60	80
Other bandwidth (GEANT2, IRNC, SURFnet, WHREN, CANARIE, etc.)	10	10	10-20	20	20-30

It is important to note that: (1) the US ATLAS and US CMS contingents will share network access to the CERN laboratory with ALICE, LHCb and the non-LHC programs at CERN, (2) traffic between US-Tier1 and EU-Tier1 centers may transit via CERN, consuming a significant part of the US LHCNet bandwidth, and (3) some of the bandwidth listed in the third line of the table (GEANT2, SURFnet, WHREN, CANARIE, etc.) is to support "inter-regional" traffic from the Asia Pacific region, Russia, China, and Latin America, that transits the US or Canada before crossing the Atlantic.

The rising bandwidth requirements shown in the table above are confirmed by the evolution of traffic in ESNET in the past four months. Figure 2 shows the evolution of the traffic accepted by ESnet per month (2000-2007). The accepted traffic was 2.4 petabytes/month in June 2007. Looking at the growth since mid-2003, the trend is still in

[8] See http://cern.ch/icfa-scic. The 2007 ICFA SCIC documents and presentations are available at http://monalisa.caltech.edu:8080/Slides/ICFASCIC/SCICReports2007

line with the long term growth trend or a factor of ten every 47 months, although the growth spurts associated with the (still reduced-scale) "data challenges" of the LHC experiments are evident in the plot. The “pre-LHC” scale of the CMS data traffic alone, for example, already amounts to eight petabytes over four months (an average of two petabytes per month) as shown in Figure 3.

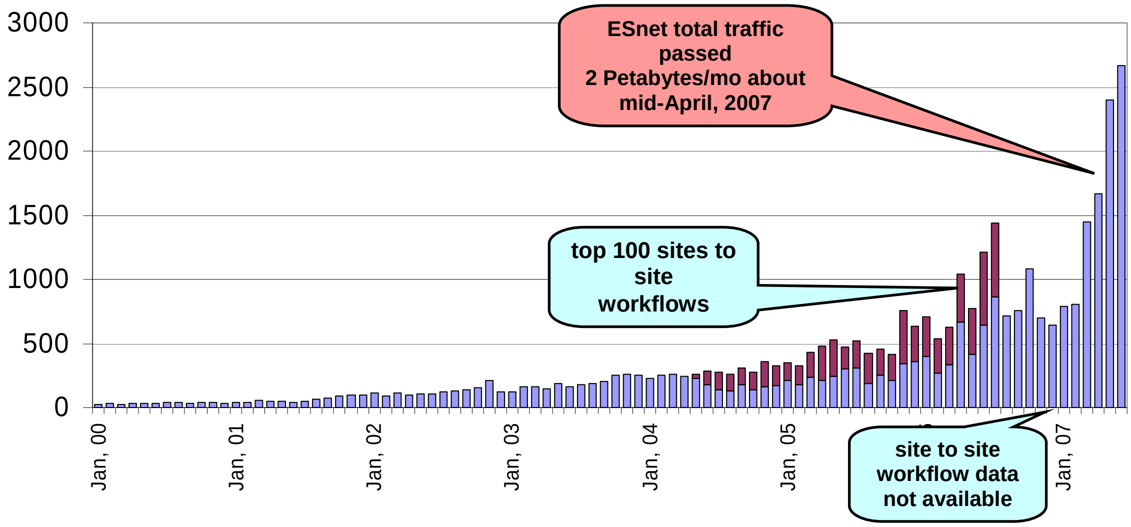


Figure 2 ESnet monthly accepted traffic January, 2000 - May 2007; more than 50% of the traffic is now generated by the top 100 sites *(courtesy of ESnet)*

We should also note that the view of HEP network requirements continues to change year-to-year as the “data challenges” of the LHC experiments have progressed. Applications and system concepts are becoming increasingly demanding in terms of network capacity and the need for reliable high performance end-to-end, notably through the adoption of grids with increasing scope and “data intensiveness”.

The ability to use long-distance 10 Gbps links effectively using low cost data servers has been amply demonstrated by Caltech, CERN, SLAC, Fermilab and others over the last five years. Single streams and flows of multiple streams reaching 10 Gbps server-to-server are now possible, and are becoming increasingly routine where needed. This means that peak bandwidth demands could potentially exceed the numbers in the table, and the bandwidth usage will need to be carefully managed.

With the current transatlantic capacity of 30 Gbps, the US CMS and US ATLAS Tier1 centers will each have the equivalent of at most one 10 Gbps dedicated link that will be used for Tier0-Tier1 as well as Tier1-Tier1 traffic. Some designated Tier1-Tier2 traffic also will be carried by US LHCNet, where part of the New York – Amsterdam link is to be dedicated to support traffic between Tier2 centers in Europe and the U.S. Tier1s, using a peering set up between ESnet and GEANT2. US LHCNet will also be used to support ALICE, LHCb and the LHC Accelerator Research Project (LARP). Part of the bandwidth also will be used as required for development, especially during periods when new versions of the network services are released, deployed and tested, with increased functionality and scale following each development cycle (as presented in detail in Annexes B and F).

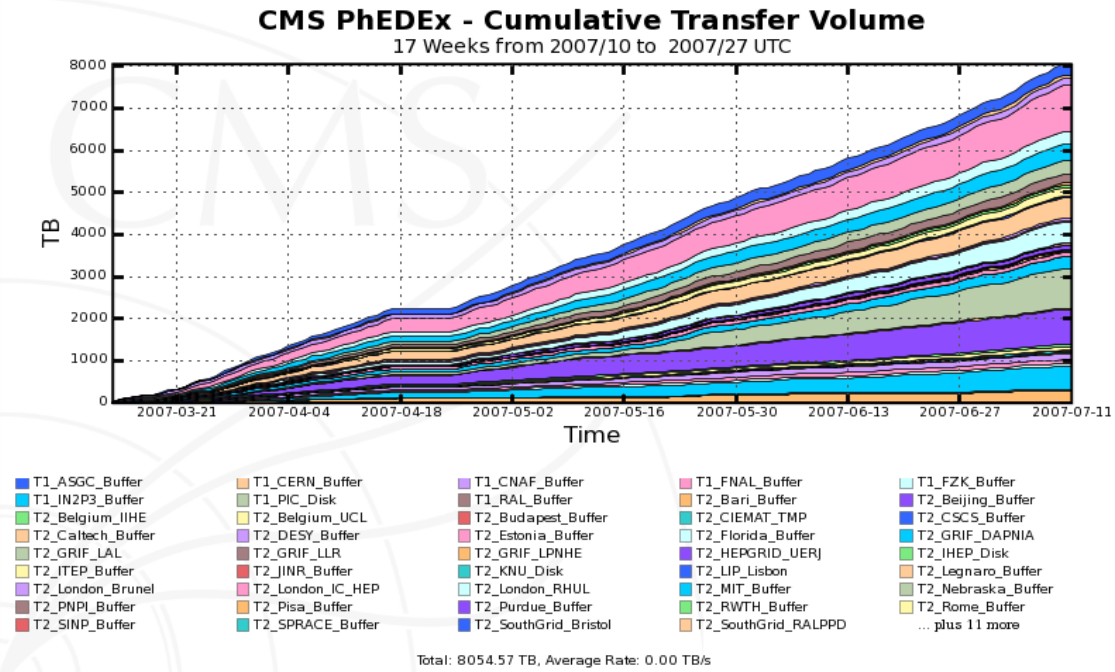


Figure 3 Accumulated data (Terabytes) received by CMS Data Centers ("Tier1" sites) and many analysis centers ("Tier2" sites) during the past four months (8 petabytes of data) [LHC/CMS]; this sets the scale of the LHC distributed data analysis problem

The implementation of the 4th transatlantic link at end of 2007 will therefore be crucial in providing provide some effective redundancy, and some additional bandwidth required to develop the necessary production-ready circuit services prior to the start of LHC operations. Once the LHC starts, the bandwidth used for production-operations and that used for development of circuit-provisioning and scheduling services, and the associated monitoring and management services, will have to be carefully managed, especially as the just-in-time plan for provisioning bandwidth over the next three years (summarized in the table above) provides relatively limited bandwidth and thus limited flexibility for development and testing at-scale. We expect that the competition for resources will be especially constraining during data challenges in 2007-8, and from the start of the LHC onwards, up to the stage where there are at least six transatlantic links in 2009.

Given the full range of network requirements for production and development, the ability to dedicate an entire, physical 10 Gbps link to a designated class of flows for long periods (e.g. Tier0-Tier1 traffic to Brookhaven or Fermilab) is not expected to occur until early in 2009 (when six transatlantic links are planned; during the 2008-9 LHC shutdown), and again in the first half of 2010 (when eight transatlantic links are planned; during the LHC 2009-2010 shutdown).

If network usage continues to expand, as strongly indicated by historical trends in HEP network usage, further network upgrades will be required in the years beyond 2010. As noted by ESnet in its long range planning, the transition to 40 or 100 Gbps wavelengths on optical fibers is expected to occur sometime in the 2012 timeframe, which may allow the historical trend of exponentially expanding network usage by the HEP community to continue, at an acceptable cost.

1.3 Meeting the Needs: Dynamic Virtual Circuits with Bandwidth Guarantees

The solution now being implemented to meet the requirements, following the widely adopted direction set at the DOE High Performance Network Planning Workshop in 2002[9], is the use of dynamically provisioned, switched virtual circuits with bandwidth guarantees, together with network services that allow these circuits to be automatically constructed, adjusted in real-time, and torn down as needed. The US LHCNet technical plans follow those being implemented by ESnet, Internet2, US LHCNet and GEANT2, and the development of the necessary provisioning and management services are coordinated with several DOE- and NSF-funded projects, including OSCARS (ESnet), DRAGON (Maryland and Internet2), Fermilab LambdaStation (Fermilab and Caltech), Terapaths (Brookhaven, Michigan and SLAC), and UltraLight (Caltech, Florida, Michigan, Fermilab, SLAC CERN and many other partners). Following the direction taken in Internet2, the implementation of the circuit-oriented paradigm in US LHCNet in a highly reliable way involves interconnecting the US LHCNet links using cost-effective CIENA SONET multiplexers with robust circuit-fallback and granular bandwidth allocation and management capabilities at the optical layer. The use of SONET equipment and dynamically provisioned and sized channels allows for bandwidth guarantees for long duration flows, and significant cost savings compared to traditional alternatives.

To further increase the reliability of US LHCNet, we will use an interconnection topology among the multiplexers and the US LHCNet Force10 switch-routers that ensures non-stop operation of the network even in case of a link failure, a multiplexer failure, or a switch-router failure. These plans, which are now being implemented, are further detailed in Sections 2.2 and 2.3 and Annexes A and B .

The use of these circuits avoids much of the cost of traditional “carrier class” routers that implement “Quality of Service” (QOS) algorithms, with very expensive interfaces. By separating a designated set of flows from other network traffic in a set of separate “channels”, many of the problems inherent in traditional provisioning methods can be avoided. One key problem with traditional methods, where TCP-based flows with different round trip times[10] (RTT), or congestion algorithms or parameter settings share a single link, is that some of the flows can compete “unfairly”, or interfere with each other in a way that degrades and destabilizes all the flows, making the time to complete a given transfer quite unpredictable. The use of dynamic channels also opens up the possibility of non-traditional protocols that can provide stable high throughput without impeding other flows.

The dynamic component of the circuits is achieved by close monitoring of the end-to-end path including the end host, the router interfaces and the application and is designed to compensate for the shortcomings of circuit provisioning, namely underutilization or bandwidth starvation. By using the CIENA multiplexers which employ the standard

[9] See http://www.doecollaboratory.org/meetings/hpnpw

[10] A particularly relevant example to US LHCNet is the fact that among competing standard TCP flows, those with shorter round trip times tend to have higher throughput. So transatlantic flows would, in this scenario, tend to have lower performance than flows within continental Europe.

VCAT[11] and LCAS[12] protocols to form and adjust the bandwidth channels in real-time, together with our monitoring systems, we are able to dynamically adjust (up or down) the bandwidth for any particular circuit in 51 Mbps increments. This gives us the ability to protect high priority traffic, reduce the bandwidth allocated to underperforming transfers, add bandwidth within a quota to allow high-performance traffic flows to complete in a short time, and so on. When working in a bandwidth constrained situation, as expected once the LHC is in operation, we also plan to deploy queues for sets of flows with similar characteristics and/or priority levels, in a fashion analogous to the batch queues of major computing facilities, and to coordinate the use of network bandwidth with computing and storage resources in order to better manage the overall workflow. Further details on the US LHCNet dynamic services architecture and its mode of operation are given in Annex F.

1.4 US LHCNet Status, Current Activities and Future plans

The Caltech engineering team operates and manages, in collaboration with CERN, a high performance transatlantic network infrastructure to meet the HEP community's needs. This facility is developed as needed to address HEP's rapidly advancing requirements, while taking advantage of the equally-rapid evolution of (and occasional revolutions in) network technologies, in order to provide the most cost-effective solutions with adequate performance, year-by-year. The operation, management and development of the current network service and its ongoing development build on more than a decade of leadership by the joint Caltech-CERN team in transatlantic networking.

The new US LHCNet backbone shown in Figure 4 has been operational since November 1st, 2006. The backbone includes three OC-192 transatlantic circuits (Geneva-NewYork, Geneva-Chicago and Amsterdam-Chicago) on three separate transatlantic cables, and another three continental OC-192 circuits (two New York-Chicago circuits and one Amsterdam-Geneva circuit)

The network connections between the US Tier1 centers and US LHCNet's points of presence in New York and Chicago are operated and managed by ESnet. These connections involve the use of metropolitan area rings, and also a dark fiber leased by Fermilab between its campus and the Starlight point of presence. ESnet's LIMAN (Long Island Metropolitan Area Network) provides dedicated 10 Gbps optical links (or "lambdas") to BNL since May 2006. In Chicago, ESNet has taken over the operation of the currently-used 10 Gbps lambdas which connect the FNAL campus to Starlight. ESnet and Caltech have co-designed an "optical private network" (OPN) that directly connects FNAL and BNL to CERN via virtual circuits, and that functions as an integral part of the LHC OPN that connects CERN to all the LHC Tier1 centers. The complete US LHCNet and ESnet setup, which includes backup paths to ensure uninterrupted operation, is detailed in Annex B.

[11] Virtual Concatenation protocol. See http://en.wikipedia.org/wiki/Virtual_concatenation and http://www.lightreading.com/document.asp?doc_id=30194&page_number=5

[12] The Link Capacity Adjustment Scheme specified in ITU-T G.7042.
See for example http://en.wikipedia.org/wiki/LCAS

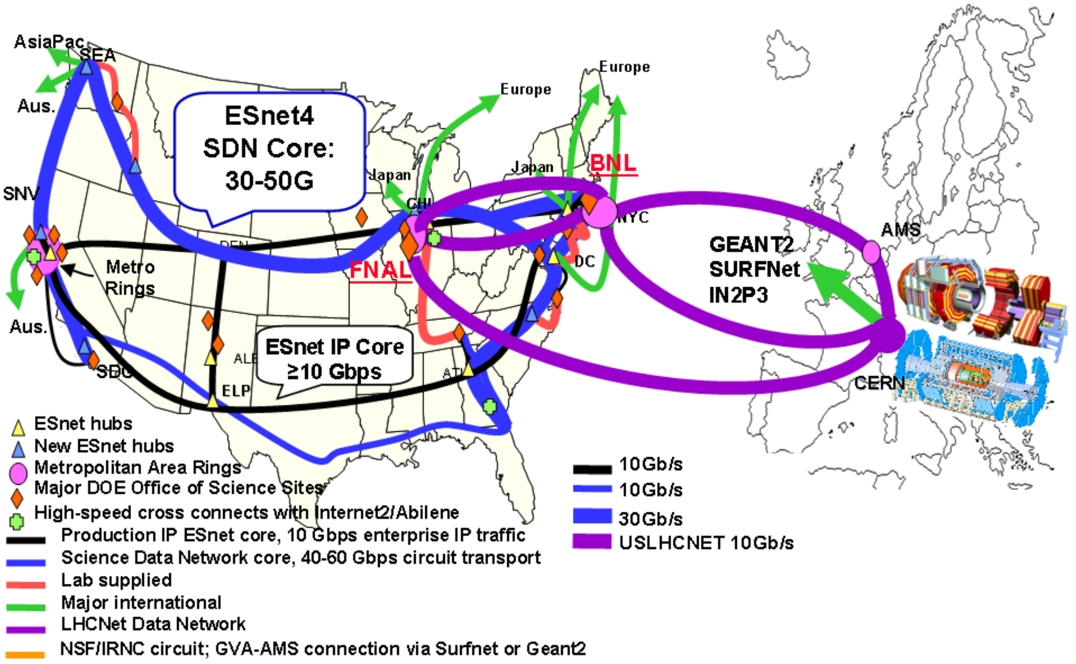


Figure 4 US LHCNet Quadrangle with connectivity to ESnet and the DoE HEP labs (since Nov 2006)

In 2007 and 2008, the US LHCNet bandwidth will be increased by 10 and then 20 Gbps respectively, as discussed further in Section 2.3 and Annex A[13]. In order to reduce costs associated with the transatlantic circuits, we have deployed a new PoP in Amsterdam, as Amsterdam is a strategic place where interconnections to other transatlantic networks (IRNC, Gloriad, Surfnet, etc.) are simple and relatively inexpensive.

Continued upgrades are planned during 2008-2010, to reach the bandwidths required as the LHC becomes fully operational, and approaches design luminosity.

It is now well-established that our current shared network infrastructure which emulates point-to-point connections by applying traditional "QoS" attributes to individual IP flows, will not by itself meet the HEP community's requirements for guaranteed data throughput for its highest-priority data transport tasks. As a result, there is a clear demand within our community for circuit-based services to supplement the usual switched and routed network services. The Caltech and CERN engineering teams have carefully selected[14] and deployed CIENA Core Director SONET multiplexers implementing the VCAT/LCAS protocols[15]. The new infrastructure will offer circuit-oriented services, by provisioning dedicated Ethernet circuits dynamically, with a

[13] Annexes A – I may be found at http://mgmt.uslhcnet.org/DOE/2007/USLHCNetAnnexes_2007.doc

[14] The selection criteria included the production readiness of the VCAT/LCAS and "virtual private line" software suite, the favorable experience of UltraScience Net, and took into account the in-depth evaluation of the Core Directors, their features and functionality relative to potential competitors performed by Internet2 and the DRAGON team. Our own evaluation completed at the beginning of 2007 showed that potential competitors were two to three years behind in the features and functions most relevant to the use of dynamic virtual circuits at the optical layer.

[15] http://www.cisco.com/en/US/products/hw/optical/ps2001/products_white_paper0900aecd802c8630.shtml

selectable bandwidth from 51 Mbps up to 10 Gbps. In addition, the new infrastructure is OC-768 (40 Gbps) capable, to accommodate the new transatlantic circuit capacities foreseen to appear within the next five years.

1.5 Team Organization and Activities

Over the last 15 years, the Caltech engineering team has acquired a unique level of experience and expertise in the operation, management, planning and development of a transatlantic network on behalf of the HEP community. In parallel, we have developed a strong and efficient collaboration with the CERN network engineering team. H. Newman shares the direction and strategic development of the network with D. Foster, head of the CERN Communications Group. Our technical engineering lead is now Dan Nae, taking over from Sylvain Ravot in November 2006. He was joined in February 2007 by former CERN staff member and chief engineer of the LHCb online data network Artur Barczyk (based at CERN), and half-time assistant engineer Ramiro Voicu. We are also replacing engineer Yang Xia at Caltech[16] in April by Tony Cheng. Cheng comes to our group with nearly 20 years of professional experience in network operations, architecture and the design and installation of large scale production networks for several Fortune 100 companies[17]. Cheng will be the principal engineer in the US, working with Nae, Barczyk and Voicu at our points of presence as needed. In particular our present team (including Yang until March and Cheng starting in April) has been working together, with help from the CERN/IT network team at CERN, to being up the four CIENA CoreDirectors on our network, moving the Force10 routers behind them, in a carefully staged transition that will avoid any interruption in the US LHCNet service[18].

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

[16] Taking up a network engineering position for the UCLA library system, to reduce his daily commute between home and work.

[17] Cheng's experience includes work with AT&T, Charles Schwab, Bank of America, Citicorp, Mitsubishi, GTE Bell Atlantic, Unisys, and Martin Marietta.

[18] The steps in the transition, and the configuration at each stage, are detailed in ***Section 2.3.***

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 iGrid or the annual Supercomputing conferences (including SC04-SC07) for large scale demonstrations associated with longer-term planning and developments[19], where we have also benefited from large-scale vendor support to minimize the costs of these developments.

- ***Technical Coordination and Administration:*** The 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-CERN network team also has 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 (University of Florida, FIU, Michigan and many others), as well as the research and education network teams of Canada (CANARIE), the Netherlands (SURFnet), GEANT2[20] and other international partners.

  The administration activity also includes negotiating contracts with telecom providers and hardware manufacturers, the formulating and managing calls for tender for bandwidth upgrades, 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 and Chicago PoPs.

  Until the Spring of 2006, CERN was responsible for Requests for Proposals 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. The latest RFP, for the US LHCNet circuits in FY2008, has been issued in mid-July 2007. Responses to the RFP are due on August 29.

- ***Management, Planning and Architectural Design:*** This covers (1) overall management of the team and its year-to-year evolution, (2) developing and implementing the strategy and planning for US LHCNet Operations and Development in consultation with our partners, (3) examining technology options, making design choices, and developing the architecture and site-designs

[19] Especially for the demonstration of networks beyond the scale of current production networks. These exercises, including the SC03, SC04, SC05, and SC06 efforts led by Caltech, have proven to be very valuable for medium and long term planning and development. An SC07 demonstration showing storage to storage networking, integration with the dCache system, and data distribution and analysis combined with dynamic circuit provisioning across the Atlantic is planned for this November.

[20] GÉANT2 is the pan-European research and education network http://www.geant2.net/

(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.

1.6 Manpower Profile and Funding Plan

The current and planned activity distribution of the Caltech team, split among the 4 major categories outlined above, is shown in Figure 5. As shown in the figure, the team's operations have migrated towards an increased focus on deployment, operations and support of the production network, and to the rapid preparation of production-ready services in the case of our development activities. Tables describing the distribution of CERN's activities, and detailing activities of each member of our Caltech network team, are provided in Annex C.

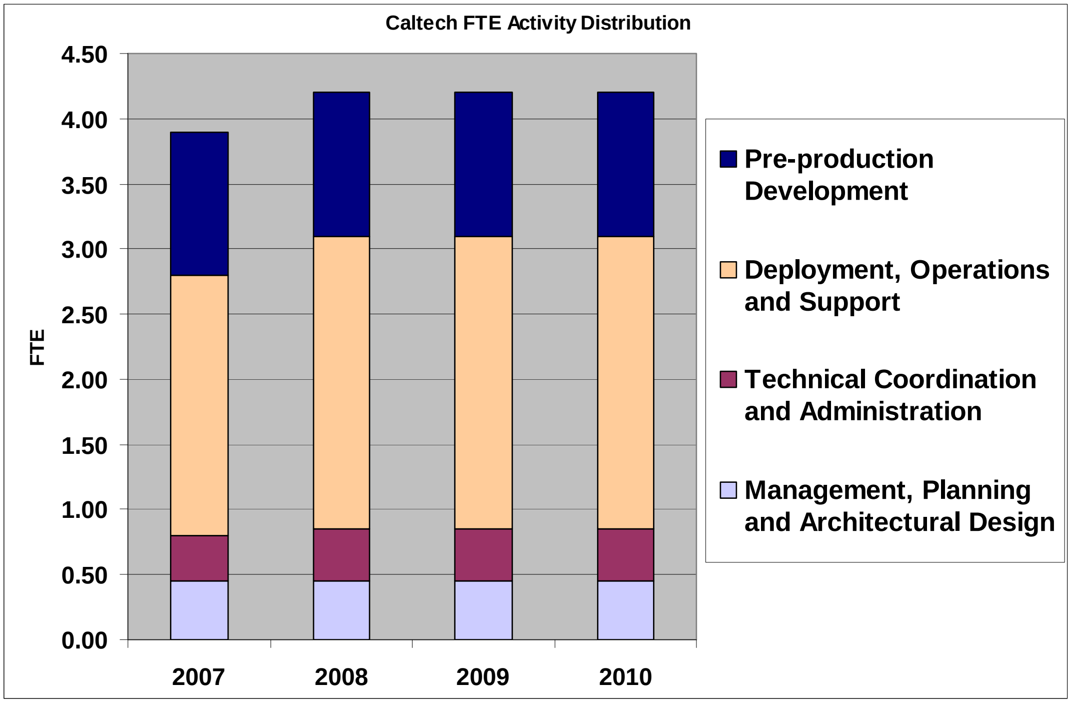


Figure 5 The actual (2007) and planned (2008-2010) effort distribution of US US LHCNet engineering activities (FTEs)[21] by the Caltech team, year-by-year.

The US LHCNet manpower level currently funded by DOE in 2007 through a grant to

[21] The management and strategic development activity includes 0.2 FTE of effort by H. Newman.

Caltech[22] level in 2007 (3.5 full-time Caltech engineers) is just sufficient to operate the network and to support the ramp-up of networking activities supporting Fermilab and Brookhaven, their connections to other Tier1 and Tier2 centers as needed. This year, with the provisioning of additional transatlantic circuits, the development and deployment of circuit-oriented services, the increasing need for integrating a new generation of network management services with Grid services, and the general increase in network-related activities as the LHC experiments start up, we are now making a transition from 3 to 4 full-time Caltech engineers, to strengthen the overall operations and support effort of the team as required[23]. The Caltech US LHCNet team would then remain at the constant manpower level of 4 full time engineers[24] from now on. Section 2.4 gives a detailed description of the manpower requirements.

The actual (2005-2007) and proposed (FY2008-2010) funding profile, which represents the current best estimate of the DOE funding required (and just sufficient) to meet HEP's network needs, is summarized in Figure 6. Apart from the charges for leasing the transatlantic link, there are significant charges for “Infrastructure” which includes the required network hardware (routers, switches, optical fibers, and interfaces), rental costs for placing and maintaining racks at the point of presences in New York, Chicago and Amsterdam, connections to the general purpose Internet, the salaries of the network engineers, a modest amount for test and server equipment placed at StarLight, MANLAN and SARA and maintenance (24x7x365). In addition to the totals shown, CERN pays for the Force10 and CIENA equipment located at CERN and contributes a total of 350k Swiss Francs (approximately $ 280k) to cover part of the overall link and equipment costs.

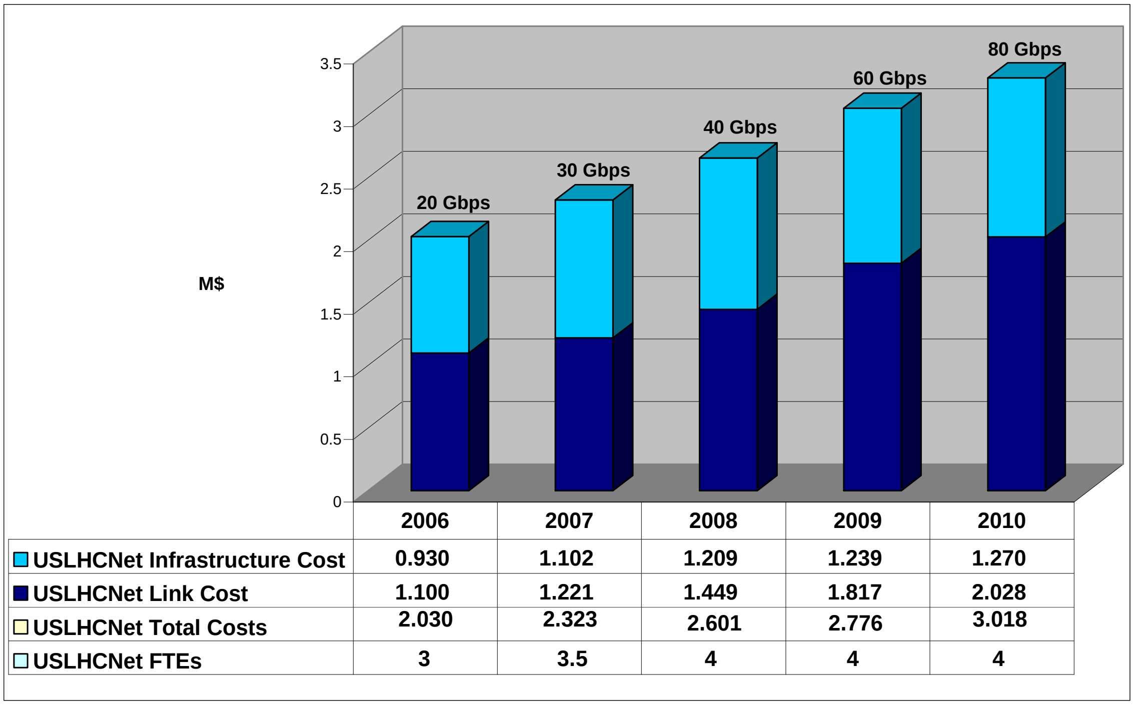
A particular issue that arose in FY2007 is a funding shortfall of $ 223k relative to the costs for the equipment already purchased, and the circuits already contracted for, for this fiscal year. The initial funding provided by DOE (under a continuing resolution) was necessarily just part of the total need, and then the supplemental request submitted in March, that reflected a precise accounting of the remaining costs this year, was only partly funded.

In order to resolve the problem temporarily, CERN has agreed to advance its contribution for FY2008 to this Fall, which will allow US LHCNet to continue to operate uninterrupted as required. The funding requests for FY2008 – FY2010 take this advance from CERN into account.

[22] Our overall funding for 2007 inancial plan also includes a strategy to deal with the current funding shortfall of $ 223k in 2007, as discussed further in this section.

[23] This corresponds to 3.5 FTEs of engineering effort in FY2007, as shown in the proposed budget tables.

[24] There is additional expert network engineering manpower at Caltech dedicated to advanced network R&D, education and outreach, and the development of networking to Latin America. These various efforts have been funded by the DOE/MICS’ LambdaStation project, and are currently funded by NSF’s UltraLight, CHEPREO, and PLaNetS (Physics Lambda Network System) projects.



	2006	2007	2008	2009	2010
🟦 USLHCNet Infrastructure Cost	0.930	1.102	1.209	1.239	1.270
■ USLHCNet Link Cost	1.100	1.221	1.449	1.817	2.028
□ USLHCNet Total Costs	2.030	2.323	2.601	2.776	3.018
□ USLHCNet FTEs	3	3.5	4	4	4

Figure 6 Annual actual and projected bandwidth and costs for the US LHCNet US-CERN network. The costs in 2006 and 2007 are actual, committed amounts. The costs projected for FY2008 to FY2010, calculated as explained in Section 3 are equal to the funding requests for those years. As explained in the text, the advance of the FY2008 CERN contribution to US LHCNet in FY2007 is accounted for in the table.

2 Project status

2.1 Working Methodology: Production and "High-Impact" Networking

The US LHCNet backbone is architected and operated to guarantee 24x7x365 network availability and full performance, supporting both large data transfers and real-time traffic such as that from VRVS/EVO. Our team works closely with the CMS and ATLAS software and computing projects, to make the network and its mode of use evolve according to the needs of the LHC experiments, and to consistently meet the particular needs of the U.S. physics groups. We keep the US LHCNet bandwidth and technology in line with the ESnet backbone, thereby providing U.S. researchers with adequate networking, and potentially a competitive advantage for their research.

Our working methodology, which has been very successful over the last 12 and especially the last 6 years, is illustrated in Figure 7. While our primary focus is the operation of the production network, with the rapid advance of network technologies (and the associated requirements-evolution) year-by-year there is a necessary continuing process of "Pre-Production" and some "Experimental" network development, where the production networks of any given year are prepared in the previous one to two years. This is driven by the fact that developing and maintaining a reliable, bandwidth-efficient network service brings with it an ongoing need to (a) develop the expertise and experience to work with higher performance, and often newer and more cost-effective models of network routers, switches, optical multiplexers, servers and server-interfaces, (b) develop new protocols and/or optimized protocol and interface parameter settings, to achieve new levels of throughput over long distance networks, (c) develop new modes of monitoring, dynamic provisioning and managing networks end-to-end, while incorporating these capabilities (on increasing scales and with increasingly functionality) into integrated grid systems. This parallels the DOE Network Roadmap[25], where the "Production" network is accompanied by a "High Impact" network in which the next-round production capabilities are developed[26].

2.2 Technical Status (July 2007)

US LHCNet has been architected to ensure efficient and reliable use of the multi-10 Gbps backbone up to relatively high occupancy levels for each of a wide variety of network tasks. On the CERN side, the network has redundant connections to the CERN backbone and the LCG (LHC Computing Grid) farms. On the U.S. side, the bandwidth to research networks and DOE laboratories is continually being increased in partnership with ESnet, Internet2 and National Lambda Rail, as well as through regional and university-funded network initiatives. As described below and shown in Figure 9, eleven of our partners already have a 10 Gbps connection to our equipment either via a dedicated fiber or via the StarLight switching exchange infrastructure. MIT, NYU, and SUNY Buffalo also are in the process of upgrading their connections to MANLAN. MIT in particular has purchased a dark fiber ring between Cambridge and MANLAN, and NYSERNet (who manages the MANLAN facility) is beginning a developmental program of "Lambda" networking and intends to partner with US LHCNet and UltraLight in these developments.

The US LHCNet network has evolved significantly in 2006-2007. In particular, new circuits were deployed as a result of our call for tender last year and we added the four SONET optical multiplexers to our backbone. Our requirements for the next generation US LHCNet optical equipment call for circuit-oriented services with bandwidth guarantees for high-priority network traffic flows. The allocation of bandwidth channels will be dynamic, and policy-driven, in order to optimally match the allocation of the available bandwidth across the Atlantic to the experiments' needs. The protocols used to match the transatlantic links to 10 Gigabit Ethernet interfaces at the ends, to form logical channels with bandwidth guarantees, and to adjust these channels dynamically in response to shifting demands are called respectively GFP, VCAT and LCAS.

These technical capabilities will allow us to proceed with our vision of dynamically provisioned circuits and optical paths at the application level, and dynamically adjustable bandwidth provisioning based on the application needs and capabilities which

[25] The DOE Science Networking Challenge: Roadmap to 2008 report is at http://www.osti.gov/bridge/product.biblio.jsp?osti_id=815539

[26] Also see the report of the DOE High-Performance Network Planning Workshop at http://www.doecollaboratory.org/meetings/hpnpw/finalreport/

are required for the next generation US LHCNet, and for efficient use of its network resources.

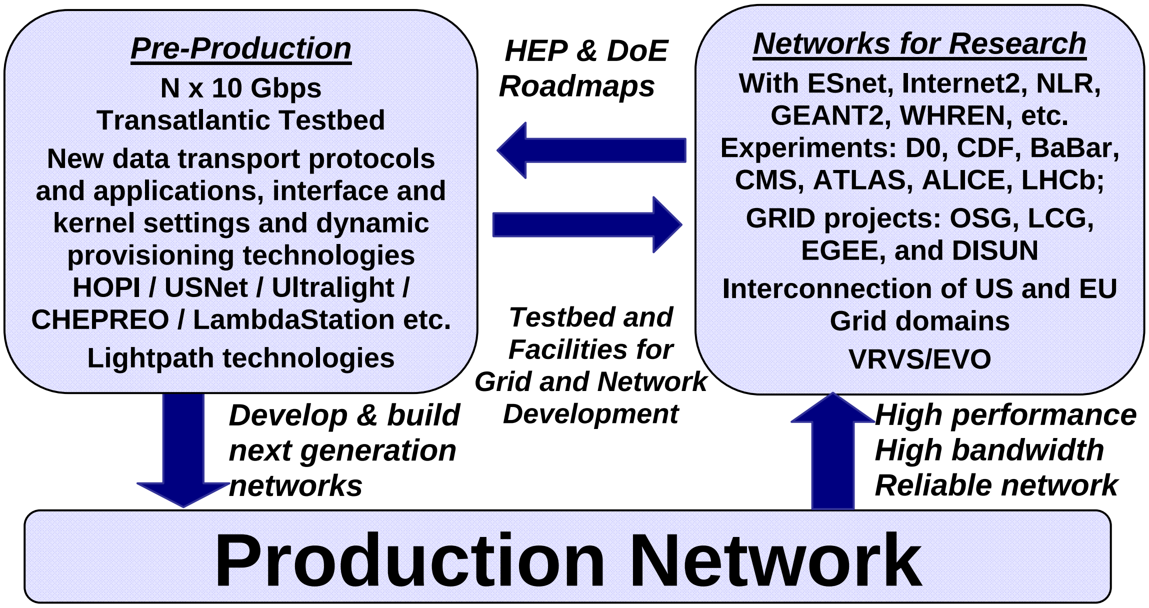

Pre-Production
N x 10 Gbps
Transatlantic Testbed
New data transport protocols and applications, interface and kernel settings and dynamic provisioning technologies
HOPI / USNet / Ultralight / CHEPREO / LambdaStation etc.
Lightpath technologies
HEP & DoE Roadmaps
Testbed and Facilities for Grid and Network Development
Networks for Research
With ESnet, Internet2, NLR, GEANT2, WHREN, etc.
Experiments: D0, CDF, BaBar, CMS, ATLAS, ALICE, LHCb;
GRID projects: OSG, LCG, EGEE, and DISUN
Interconnection of US and EU Grid domains
VRVS/EVO
Develop & build next generation networks
High performance
High bandwidth
Reliable network
Production Network

Figure 7: US LHCNet Working Methods. The prime focus is on the provisioning of a reliable high performance transatlantic production network in support of DOE's major HEP programs. This is achieved, on increasing scales, and with effective exploitation of the latest technologies, by a tight planning cycle, strong synergy with and leadership in several network R&D projects, and strong partnerships with leading network and grid projects as well as vendors.

Based on these requirements, we selected the Ciena CoreDirector/CI because it's the only production-ready equipment available on the market today that is fully capable of supporting the GFP/VCAT/LCAS protocol set. The CIENA CD/CI's stability and reliability has been proven by Internet2 which has installed more than 25 CoreDirectors at key sites throughout the U.S. on their next-generation network backbone, by the DOE-funded UltraScience Net project, and by Qwest who have chosen CoreDirectors for their new nationwide network.

- Between November 2006 and March 2007 the new circuits tendered last year became operational. A timeline of the events is shown in Figure 8. We have increased our transatlantic capacity from 20 to 30 Gbps and we added two new continental circuits between New York and Chicago so Fermilab and Brookhaven each can have access to all of the transatlantic circuits, and also maintain the majority of their connectivity with CERN during link outages.
- In February 2007 a new US LHCNet PoP was deployed in Amsterdam, located in the SARA Facility. A new Force10 router was installed to connect the Geneva - Amsterdam and the Amsterdam – New York circuits

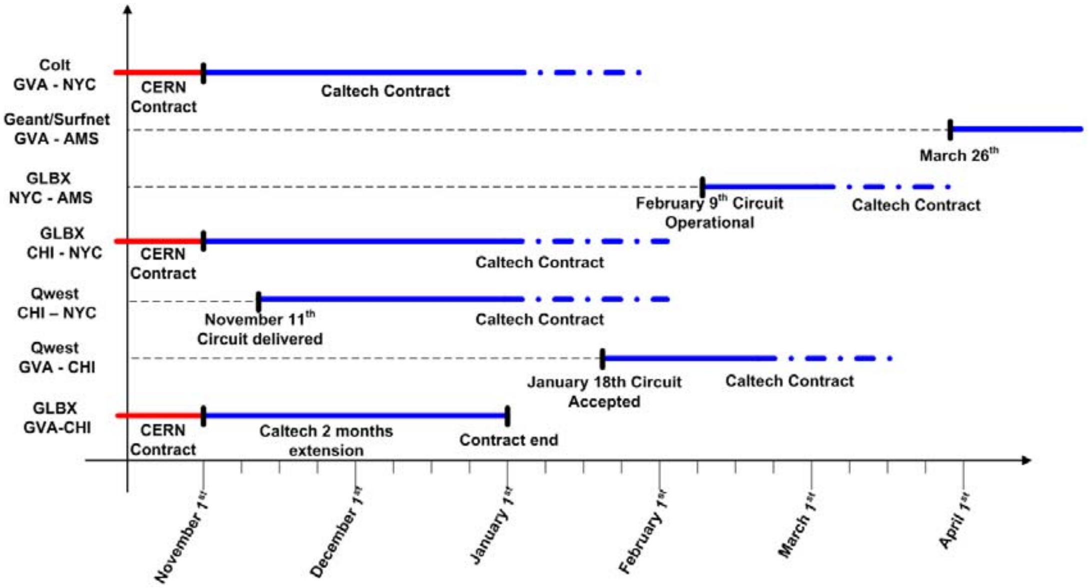


Figure 8 US LHCNet Circuit Deployment Timeline

- Since January 2007 US LHCNet exports monitoring data to the perfSONAR monitoring service providing end-to-end monitoring for the LHC Optical Private Network (OPN)
- The new Ciena CoreDirector/CI's were installed in our PoPs in New York City and Chicago in March
- In April the Ciena CD/CI was commissioned in Geneva

The current topology of US LHCNet network is shown in Figure 9. At the time of this writing, the OC-192 transatlantic circuits still terminate on the Force10 E600s, with the CIENA CD/CIs situated behind them, so that we can exercise the CIENA functions but also ensure production operations in the mode used in 2005-6, based on 10 GbE WAN-PHY in the Force10s. The configuration shown in the figure provides a variety of services running across the Atlantic to support both production and "pre-production" needs. In addition to standard IP services, we provide "Layer 2"[27] point-to-point connections between CERN and the US-Tier1 centers, extensive Quality of Service (QoS) configuration and policy-based routing (PBR).

We are now gradually moving the OC-192 circuits to the new CIENA multiplexers, in a step-by-step procedure described in the following section.

[27] Layer 2 is the Data Link Layer in the ISO standard seven-layered network model is the Data Link Layer that describes the logical organization of data bits transmitted. For example, this layer defines the framing, addressing and checksumming of Ethernet packets. See http://www.freesoft.org/CIE/Topics/15.htm

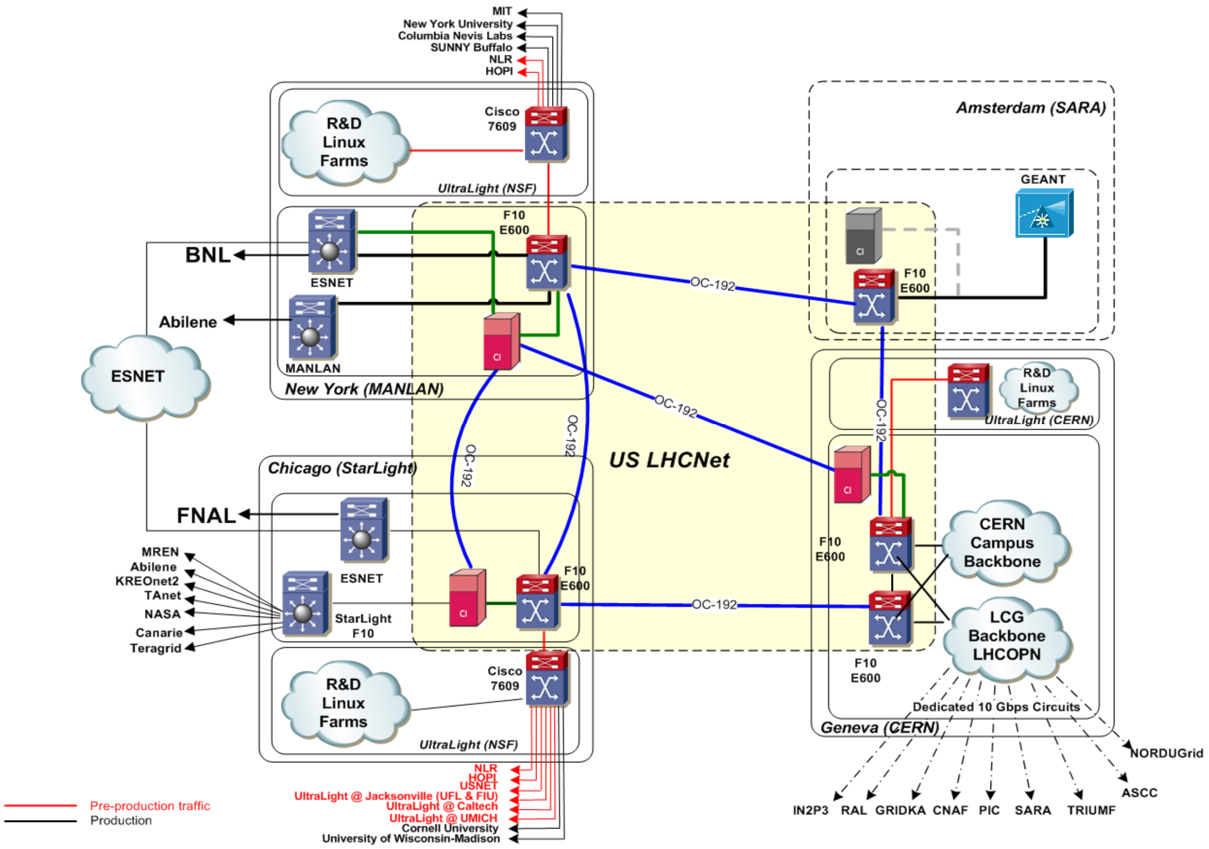


Figure 9 The Current US LHCNet topology (July 2007)

2.3 Recent Milestones in 2006-2007, and Planned Milestones for 2007-2008

The Caltech engineering team continues to develop and provide, in partnership with CERN, and in collaboration with ESnet, Fermilab, BNL, UltraLight, Internet2, National Lambda Rail and GEANT2, the appropriate cost-effective transatlantic network infrastructure required to meet the HEP community's needs. This includes a highly reliable, high-performance production network available 24x7, and a "pre-production" infrastructure for networking (and associated grid) developments. We have designed, commissioned and deployed a scalable network which can support additional transatlantic circuits, and accommodate any new partners that wish to directly or indirectly connect to US LHCNet, while carefully managing the available network resources.

From now until the start of operation of the LHC, the majority of the network bandwidth will remain available to an ongoing series of grid-based data processing, distribution and analysis "challenges", which aim to prototype the data movement services that will be used at full scale once the LHC experiments begin data-taking. This is allowing us to acquire an understanding of how the entire system performs when exposed to the level of usage we expect during LHC running. Based on this experience, combined with our ongoing experience in several network and grid development projects, we will continually update our operational procedures as needed, and evolve the setup and develop the necessary tools to administer and monitor the network end-to-end. As mentioned above, we will also work with the major grid projects, such as OSG and LCG, to ensure consistency between the managed network and Grid operations.

The milestones described below are in addition to the daily production-network support.

- **May 2006:** Caltech issued a Request for Proposals for the supply of multiple 10Gbps transatlantic circuits. The RFP documents are available in Annex I.
- **August 2006:** Telecom provider(s) Qwest, Global Crossing and Colt were selected from among those responding to the RFP.
- **October 2006 – April 2007:** Provisioning of new transatlantic circuits. This proceeded on schedule, as shown in Figure 8. We replaced our existing (CHI-GVA) transatlantic circuit from Global Crossing by one from Qwest, kept the existing transatlantic circuit (NYC-GVA) from Colt, and added a transatlantic circuit (NYC-AMS) from Global Crossing. We also added continental links from Global Crossing (NYC-CHI) and GEANT2 (AMS-GVA). Overall this was the most cost-effective plan, taking path-diversity and reliability requirements into account.

  This transition brought US LHCNet capacity to include three transatlantic circuits, as planned. As described above, this is the minimum bandwidth required to support the network needs of Fermilab and BNL, in addition to other transatlantic network needs, in the period up to and including LHC startup this year.
- **US LHCNet-Tier2 Center connectivity (2007):** We have experienced a strong ramp-up of network-related activities by Caltech, Nebraska, UCSD, Florida, Michigan, Wisconsin, MIT, Argonne, LBNL, Boston University and other U.S. Tier2 centers both in LHC data analysis challenges and in development projects such as NSF's DISUN project[28], where the DISUN sites and others have one or more 10 Gbps connections to Starlight. Other large U.S. computing centers including for example Cornell and Vanderbilt[29] in US CMS[30], which are not designated as Tier2s, will continue to increase their activities. Switch ports supporting connections to these partners are being funded by UltraLight and other NSF projects, or by the partners themselves.
- **July 2007:** US LHCNet will provide transit for a ESNET – GEANT2 peering over our NYC-Amsterdam link. The peering is designed to provide support for European Tier2s accessing US Tier1s. Discussions with IRNC PI Tom deFanti to use the IRNC links for traffic between the US Tier2s and the Tier1s in Europe have also begun, and use of the IRNC links has been agreed. The idea to use up

[28] The Data Intensive Sciences University Network. This NSF-funded project, based on a concept and work in support of grid-based data analysis developed by Caltech, is led by former U.S. CMS Deputy Program Manager and NSF PI Bob Cousins (UCLA), and involves the four Tier2 sites mentioned in the text.

[29] See http://www.vampire.vanderbilt.edu/about/index.php about Vanderbilt's Advanced Computing Center for Research and Education (ACCRE). Director P. Sheldon is working towards making ACCRE a major data storage site for US CMS.

[30] Lawrence Livermore National Lab also joined US CMS early in 2006.

to half of these links for such traffic has been agreed to in principle by NSF program manager Kevin Thompson. Future calls with GEANT2, who operates the general purpose IRNC link between Amsterdam and New York, will include discussions of the use of part of this link to support some of the traffic between US Tier2s and European Tier1s.

Further discussions with Internet2, to carry US Tier2 traffic in the US to the international points of presence in the US, and with the goal of setting up a peering with GEANT2 in Europe this year, started in March between Internet2 CEO Van Houweling and H. Newman, and have been carried forward by Internet2 engineers R. Summerhill and R. Carlson. These discussions, and a series of workshops organized by Carlson at several US CMS and US ATLAS sites, have also covered Internet2 carrying the traffic required between US Tier2s and Tier3s.

- **August 2007:** Ciena CD/CI installation in Amsterdam and migration of the Amsterdam – Geneva and Amsterdam – New York circuits to the new equipment; The Ciena CD/CI will provide the possibility of having circuit oriented services as well as equipment redundancy at all four US LHCNet sites for the transatlantic connectivity. A network map is provided in Figure 10.

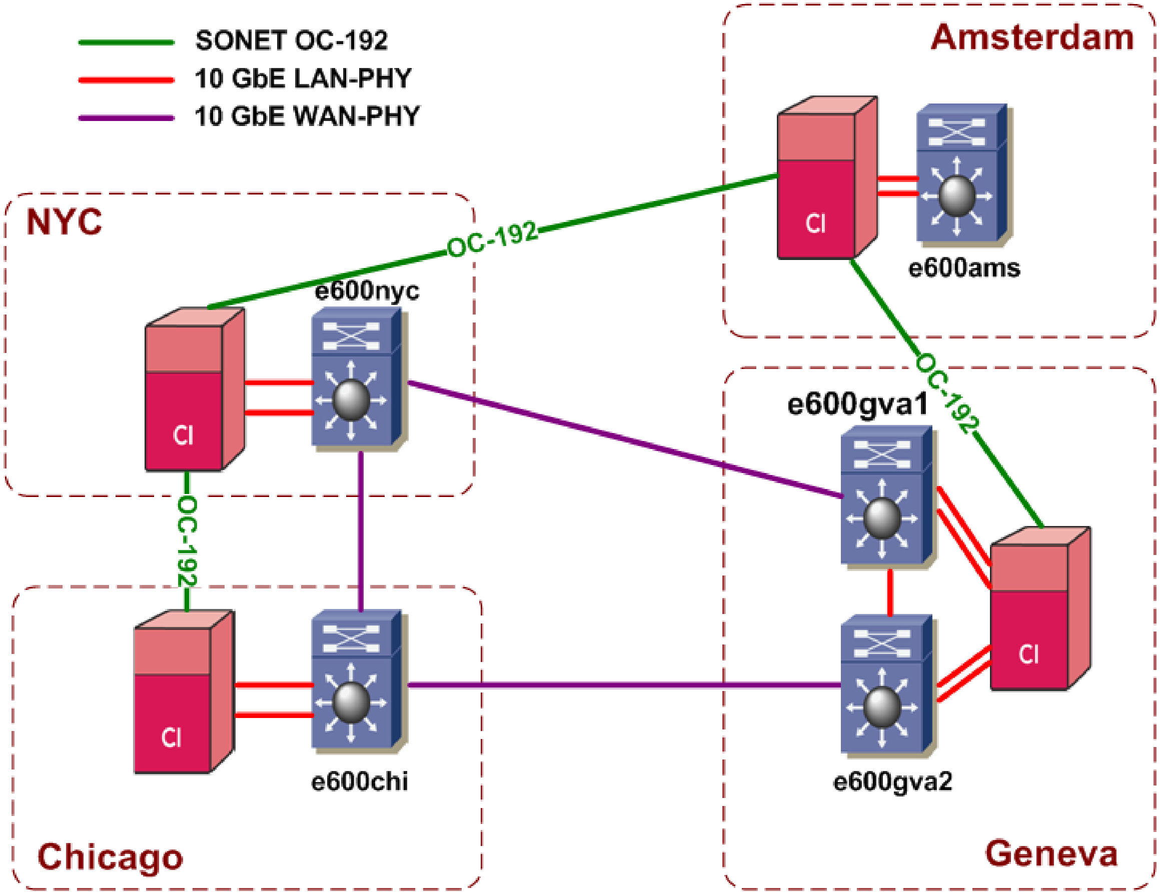


Figure 10 US LHCNet Transitional Topology Map (August 2007)

- **“Pre-production”:** The new infrastructure initially deployed in the Spring and Summer of 2007 will offer circuit-based services intended to provide redundant paths and on-demand, high bandwidth end-to-end dedicated circuits. Circuit-switched services will be used to directly interconnect the DOE laboratories to CERN and will be available on demand to policy-driven, data-intensive applications, managed by MonALISA services. These new services, now under development in the Caltech-led UltraLight project, the Fermilab+Caltech LambdaStation project, BNL and Michigan’s Terapaths project, and ESnet’s OSCARS projerct, will be interfaced to the future ESnet lambda-based infrastructure.
- **October 2007:** initial deployment of our circuit oriented network services on US LHCNet; simple scheduler with fixed bandwidth circuits for site to site on-demand data set transfers.
- **Spring 2008:** interaction with the data transfer application of the experiments, as well as with other intra-domain and inter-domain (LambdaStation, TeraPaths, DRAGON, Oscars) control plane services in order to provide an end-to-end path reservation.
- ***LHC Startup:*** July 2008: We will begin to exercise the network and services with real data, in close cooperation with the LHC experiments. The planned US LHCNet configuration by this time is shown in Figure 11.

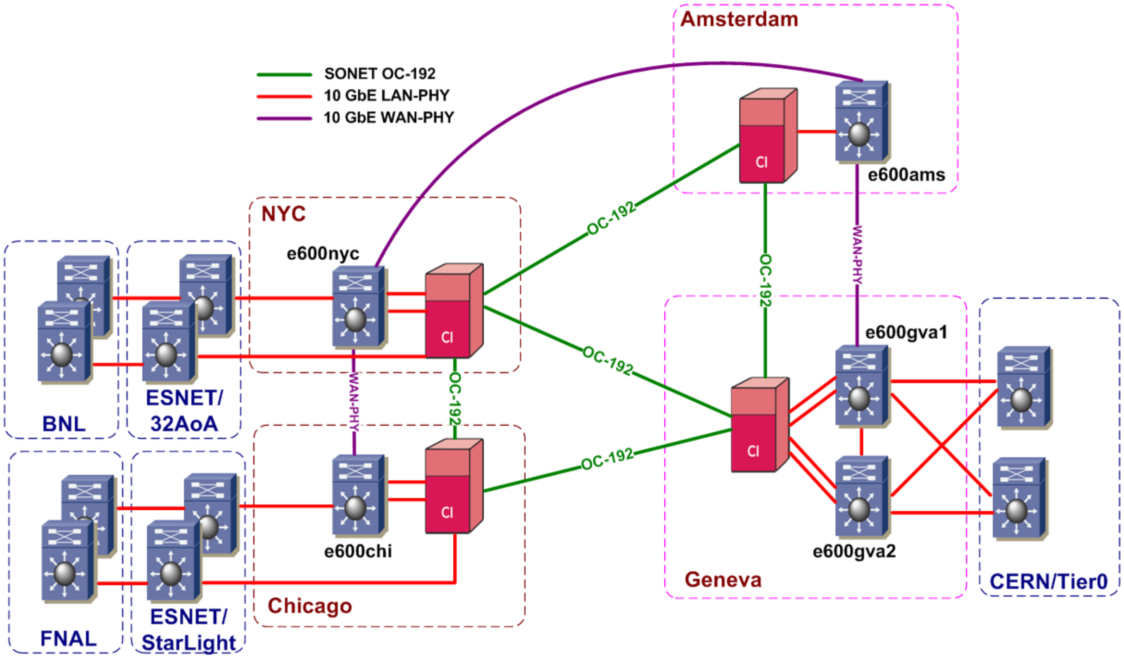


Figure 11 Planned US LHCNet Network Map for LHC Startup in 2008

- **International connectivity (2007-2010):** Maintain excellent connectivity to international partners. US LHCNet will be connected at 10 Gbps with most or all of its international partner networks, and we will set up connections via multiple lambdas with some of them, when and where appropriate. (The switch ports supporting connections to these partners are being funded by UltraLight and other NSF projects, or by the partners themselves, together with donations of part of the equipment from Cisco Systems.) International collaborations include those described in Section 2.2, and connections via AMPATH[31] to the CHEPREO/WHREN link between Miami and Sao Paulo, to enable high speed communications between the US and South America (starting with the Tier2 centers in Rio de Janeiro (UERJ) and Sao Paulo (UNESP) that are now partners working closely with the US Tier2 team, and members of OSG.

We also maintain collaborations with KISTI in Korea, RoEduNet in Romania, SANET in Slovakia, and PERN2 in Pakistan, as well as the Chinese research network CSTnet[32] that supports IHEP Beijing, and GLORIAD[33] – a global ring network around the earth.

- **Circuit oriented services (2008):** As described in earlier sections and Annex F, we will implement dynamic circuit provisioning. Dynamic circuit adjustments and queuing mechanisms for multiple bandwidth request will be progressively automated where needed (through the use of MonALISA services, for example) and made available to (authenticated, authorized) users in the context of the new network-management services infrastructure which is currently under development.
- **The network as a Grid resource (2009 —):** Add the “network” dimension to Grid-based systems by extending advanced planning and optimization into the networking and data-access layers. Provide interfaces and functionalities for physics applications to effectively interact with the networking resources.

Further details of US LHCNet bandwidth planning for 2008-2010, coordinated with ESnet's plans, are given in Annexes B and E.

2.4 Caltech-CERN Team Organization and Breakdown of Team Activities

The management and the operation of the link are based on a strong collaboration between the CERN and Caltech network engineering teams, dating back to 1995. The current organization of the US LHCNet project is shown in Figure 12

The US LHC Network Working Group includes members from US LHCNet, the DOE HEP laboratories involved in the LHC, ESnet, key university representatives with leading roles in the NSF UltraLight and DISUN projects, and the U.S. funding agencies. The group met for the first time at CERN in July 2005, and holds conference calls and meets annually to discuss strategic development and coordinate US-CERN networking activities.

[31] http://www.ampath.fiu.edu/

[32] A 2.5 Gbps link is planned to start October 1, 2005.

[33] http://www.gloriad.org/gloriad/index.html

Harvey Newman is the Principal Investigator of US LHCNet and is responsible for direction of the project from the U.S. side. He takes a lead role in developing the medium and long-term plans, and contributes to the corresponding bandwidth roadmap, oversees the Caltech network team and its activities, and allocates the funding to US LHCNet circuits and network equipment under the DOE grant to Caltech. He shares the direction of network operations and development with D. Foster, head of CERN's Communications Group.

Dan Nae is the senior network engineer who is responsible for the technical coordination of all networking activities between CERN and the U.S. across US LHCNet. He is a member of the Caltech HEP networking team based at CERN since March 2003, and he has been in charge of the design, implementation and operation of US LHCNet since November 2006, including the installation of the Ciena CD/CI multiplexers, the design of the network topology, and the migration of the circuits. He leads the Caltech network team, supervises the day-to-day operation of the network, and coordinates as well as contributing to pre-production network development activities. Dan is also responsible for the routing and peering policies with other networks. He is a member of the ICFA Standing Committee on Inter-regional Connectivity.

Artur Barczyk joined the team in February 2007, coming from his previous position as a CERN staff member and head of the CERN LHCb online data network. He is contributing to the day-to-day network operations and support including network element configuration and handling trouble tickets and outages, and manages our Service Level Agreements (SLA) with the circuit vendors. He is responsible for the deployment of dynamic circuit-oriented services on the US LHCNet network, as well as for testing and evaluation new equipment and technologies. Artur leads the preparation of the US LHCNet Requests for Proposal documents for renewing and upgrading our circuits, works with Newman and Nae to oversee the RFP process and its results, and acts as a liaison with the various LHC computing groups.

Tony Cheng joined the team in April 2007. He is based at Caltech, and can relatively easily travel to our US PoPs to work on new hardware installation and maintenance. He has over 20 years of professional experience in network operations, architecture and the design and installation of large scale production networks for several Fortune 100 companies. Tony is now the primary network engineer in the US and contributes to overall US LHCNet operations and support, especially during Caltech working hours. He will be working on bandwidth management techniques as well as pre-production activities (equipment and technologies evaluation).

Ramiro Voicu joined the team in September, 2006. He is a software engineer working mainly on developing monitoring software for US LHCNet. His main tasks are integration of the US LHCNet networking hardware with the MonALISA monitoring platform as well as the other monitoring systems, developing network services software for automated provisioning and management of the CIENA Core Directors and the Layer 1 and 2 hybrid circuits to be formed by US LHCNet together with ESnet, Internet2 and the HEP labs. He also manages the various servers at CERN, and helps in the day to day operations of the network, and he travels with other network engineers to remote PoPs when needed.

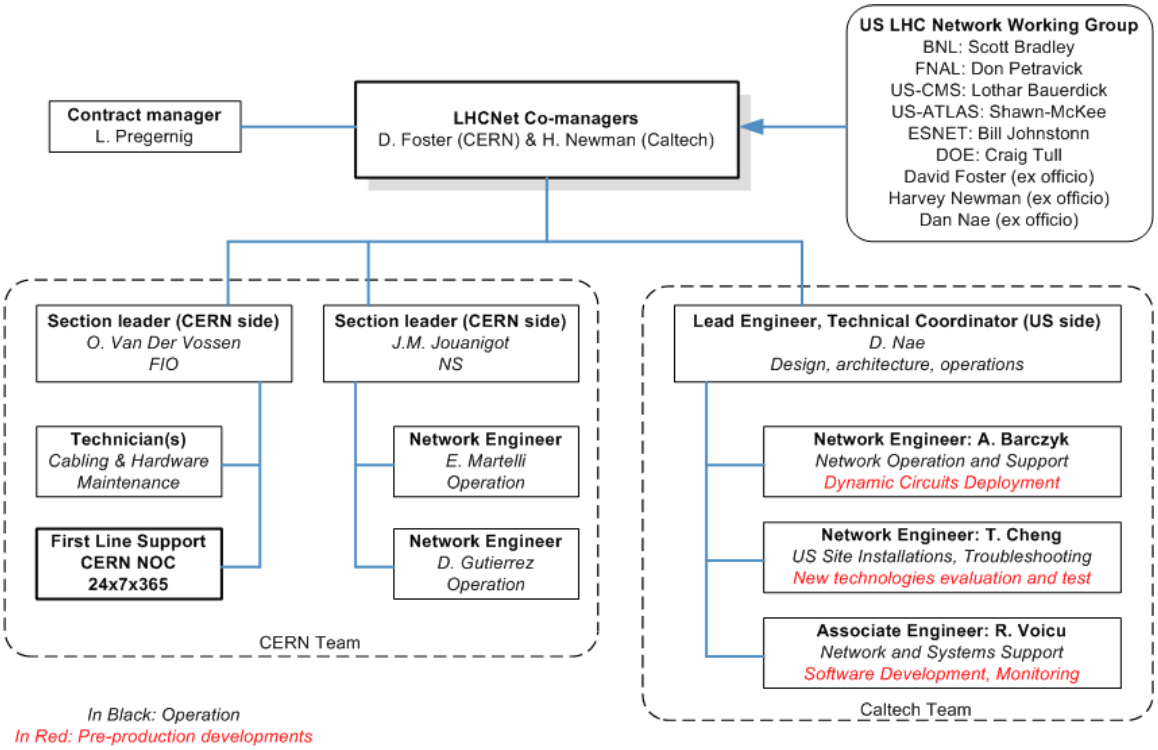

Contract manager
L. Pregernig
LHCNet Co-managers
D. Foster (CERN) & H. Newman (Caltech)
US LHC Network Working Group
BNL: Scott Bradley
FNAL: Don Petravick
US-CMS: Lothar Bauerdick
US-ATLAS: Shawn-McKee
ESNET: Bill Johnstonn
DOE: Craig Tull
David Foster (ex officio)
Harvey Newman (ex officio)
Dan Nae (ex officio)
Section leader (CERN side)
O. Van Der Vossen
FIO
Section leader (CERN side)
J.M. Jouanigot
NS
Technician(s)
Cabling & Hardware Maintenance
First Line Support
CERN NOC
24x7x365
Network Engineer
E. Martelli
Operation
Network Engineer
D. Gutierrez
Operation
CERN Team
Lead Engineer, Technical Coordinator (US side)
D. Nae
Design, architecture, operations
Network Engineer: A. Barczyk
Network Operation and Support
Dynamic Circuits Deployment
Network Engineer: T. Cheng
US Site Installations, Troubleshooting
New technologies evaluation and test
Associate Engineer: R. Voicu
Network and Systems Support
Software Development, Monitoring
Caltech Team
In Black: Operation
In Red: Pre-production developments

Figure 12 US LHCNet Organization

CERN Network Engineers. The CERN engineers (shown in Figure 12) are involved in the design and the operation of the network, in partnership with the Caltech team. They manage the network equipment on the CERN site, the commodity internet access and the peerings with research and education networks.

The CERN Network Operation Center (NOC) delivers the first level support 24 hours a day, 7 days a week. The CERN NOC watches out for alarms and can be reached at any time by any user of the network. If a problem cannot be resolved immediately, it is escalated to the Caltech/CERN network engineering team.

FNAL and BNL Network Engineers. Following the US – CERN networking meeting held at CERN in July 2005, it was decided to include FNAL and BNL engineers in the operation of the links. In particular, these engineers (D. Petravick, P. Demar, M. Crawford, V. Grigaliunas at Fermilab; W. Bradley, J. Bigrow, F. Burstein at Brookhaven) working in coordination with ESnet are responsible for the local connectivity of DOE laboratories in New York and Chicago to the site networks and facilities. Since they are relatively close to the US LHCNet PoPs, they can also provide some local support. Since 2006, they have been involved in circuit-oriented service development, and in focused efforts to increase the data throughput to and from the CERN Tier0, and among the CMS and ATLAS Tier1 (and in some cases Tier2) sites. In 2007 the Fermilab engineers have also worked with the Caltech US CMS and UltraLight teams to integrate Caltech's Fast Data Transport (FDT) with dCache.

The actual (2006-2007) and planned (2008-2010) distribution of engineering manpower (in FTEs) is shown in Figure 13. Note that the CERN contribution significantly decreased in 2005 and 2006 with the end of the former EU-funded DataTAG project and the rampup of many European Tier1 centers that require an increasing share of the attention and manpower available to the CERN network team. In spite of these pressures, the level of effort from CERN increased to 0.3 FTEs in 2007 as we approach LHC startup, and is expect to remain at at least this level in future years. The contributions from Fermilab and Brookhaven to our increased overall effort in US LHCNet also rose to a total of 0.5 FTEs in 2007.

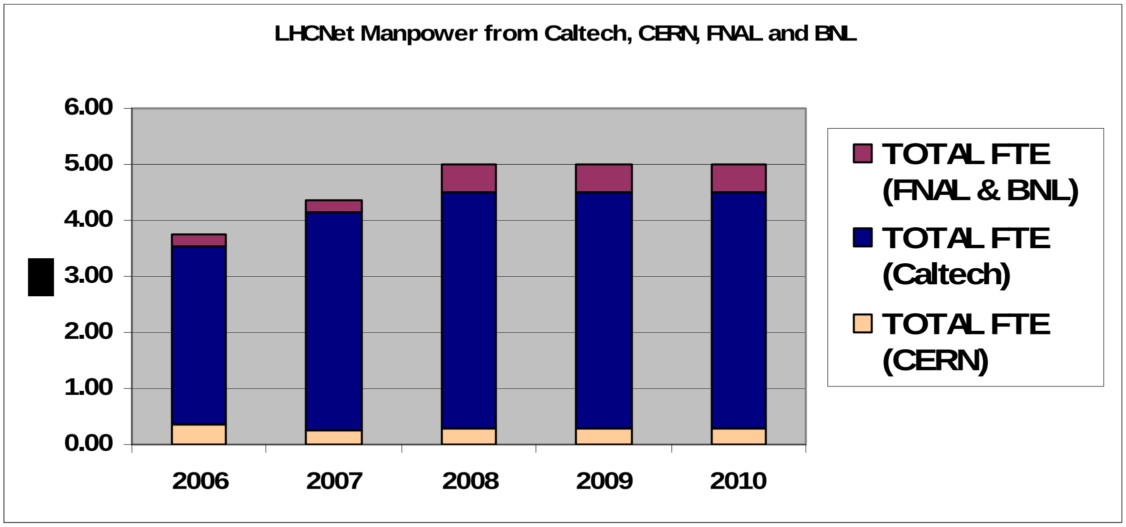


Figure 13 The actual and planned manpower contributions from Caltech, CERN, FNAL and BNL to US LHCNet

3 US LHCNet Bandwidth and Funding Plan Through 2010

The ongoing bandwidth and funding plan described in this section has been designed to meet the transatlantic networking needs for U.S. participation in the LHC program in the most cost effective manner possible. This currently includes three 10 Gbps transatlantic wavelengths on the US LHCNet network, which is (as discussed above) just adequate to meet the needs during the LHC startup period. Plans on a corresponding scale are being developed for some of the principal domestic links, including the links to the Tier1 centers at FNAL and BNL and the Tier A center at SLAC by ESnet, as well as links to the Tier2 centers over Internet2 and regional networks, or using dedicated links.

The plan is cost-optimized to take advantage of the evolution of link prices per unit bandwidth, and the emergence of “wavelength-based networks”. Further updates to optimize this plan and its cost will be made yearly, as the Computing Models of the LHC and the other major experiments are refined, and as our experience with optical network technologies (and pricing) progresses. However, it must also be mentioned that with recent information on link costs, and the fact that we have already defined a cost-optimized plan for the (CIENA) equipment with the required functionality over the next three years, we cannot expect significant cost reductions below the level of the requests presented in this section. We have also agreed with DOE that if the requested funds are provided, and if subsequently costs for the links, for any major equipment, or for the collocation rises, then we will adapt our scope and/or link upgrade timing to remain within the currently requested funding envelope.

The basic US-CERN bandwidth requirements and cost parameters in the plan are summarized as follows:

1. Ramp up the bandwidth from 30 to 40 Gbps by the end of 2007, then to 60 Gbps by 2009 and 80 Gbps by 2010. This will allow us to keep up with grid-based event productions, data distribution for CMS and ATLAS data challenges, and grid-based data analysis up to and during the first years of LHC operation, along with other transatlantic network use between the U.S. and CERN by the U.S. HEP community. This proposed rate of growth in bandwidth is somewhat slower than the longer term trends of HEP network usage over the last 15 years, as derived by the ICFA Network Task Force and ESnet, of a factor of 10 in bandwidth every 3-4 years.
2. Based on current market trends, assume constant transatlantic bandwidth costs, per unit of bandwidth, in 2006 and 2007, and an annualized cost-deflation factor per unit of bandwidth of 30% for each two-year period, starting in 2008 and continuing through at least 2010.
3. Assume CERN's contribution to circuit costs remains constant at a level of 350k CHF (approximately $ 280k).
4. As discussed in Section 1.2, assume that one-quarter to one-third of the bandwidth required is provided by other networks such as GEANT2, SURFnet or the NSF IRNC links.
5. Carry out the necessary network development to prepare each year for the production network of the following year. We began this procedure by installing the former "DataTAG" 2.5 Gbps research wavelength in the Summer of 2002. We (as well as Fermilab, BNL and others) are currently using a non-negligible part of the 10 Gbps wavelength for network. This is a very important part of the plan, since the new SONET optical switching and multiplexing equipment to be used, and the dynamic circuit provisioning and network path management services and software to be deployed, are both new and in many ways different from the traditional IP routers and switches we have used in past years.

   As described earlier in this proposal, we currently have 3.5 FTEs of engineering manpower, having added Artur Barczyk and Tony Cheng to the team replacing Sylvain Ravot and Yang Xia, and added Ramiro Voicu half-time this year. Starting in 2008, as described above, we will require 4 full time engineers to meet the needs as we approach and begin LHC operations. We assume a constant engineering manpower level from 2008 onward.

   Our manpower costs per person are relatively low: an average annual cost-base of $150k per FTE in 2008 (including indirect costs), and an annual salary index of 3.5%. These costs per FTE are low compared to the typical costs of network engineers with the requisite level of expertise on the open market.

6. From 2007 onward, a constant annual funding level for equipment is assumed; this supports the CIENA upgrades detailed in A and the annual addition of ports on switch-routers as needed, to distribute traffic over an increasing number of transatlantic and continental circuits, as summarized in Annex B. and permit the replacement of some old equipment, typically after 5 years in service for network routers and switches.
7. Reach an approximately constant DOE funding level during LHC operations, from 2010 onwards. This assumes that non-DOE sources contribute a significant portion of the overall transatlantic link cost, as is the case now. In 2010 and beyond, it is assumed that the installed bandwidth will continue to grow, but at a reduced annual percentage rate, corresponding to continued annual funding at approximately the 2010 level, in 2010 constant dollars[34].
8. Assume an exchange rate of 1.25 Swiss Francs per dollar. This unfavorable exchange rate, which represents an average over the last two years, has substantially increased the cost-of-living adjustments paid to engineers based at CERN relative to earlier years.

The proposed funding profile, which represents the best current estimate of what is required to meet HEP's network needs, is summarized in **Table 2**[35]. Apart from the charges for leasing the transatlantic link, there are significant charges for “Infrastructure” which includes the required network hardware (routers, switches, optical fibers, and interfaces), rental costs for placing and maintaining racks at the points of presence in New York, Chicago and Amsterdam, connections to the general purpose research and education networks (for example to the MANLAN router in New York to peer with Internet2), salaries of the Caltech network engineers, and maintenance (24 hour/7 day per week/4 hour response time). The breakdown of the infrastructure budget is shown in **Table 3** and Figure 14[36].

Table 2: US LHCNet annual actual (2007) and projected (2008-2010)

[34] We also expect a change in the market when wavelengths of 40 or 100 Gbps replace the presently available 10 Gbps circuits. As in the ESnet plans, this is expected to occur by approximately 2012. Implementation of the higher bandwidth wavelengths on transoceanic cables is expected to come somewhat later than on continental links.

[35] As explained in Section 1.6 above, CERN will advance its contribution for FY2008 to US LHCNet to this Fall, to allow us to meet the total expenditure of $ 2.323M in FY2007. $ 223k of this contribution is already included in the total for 2007 in Table 2. The CERN contribution for FY2008 is thus only $ 57k, ad shown in the table.

[36] The collocation and maintenance costs shown in the table are accurate estimates based on current costs in 2007, including the increased rack space required at each PoP for the CIENA multiplexers, and the projected marginal costs that correspond to maintenance on the increasing value of the CIENA and Force10 linecards each year, corresponding to the plans presented in detail in Annexes A and B.

bandwidth (in Gbps) and costs (in M$)

Year	2007	2008	2009	2010
Bandwidth (Gbps)	30	40	60	80
Transatlantic Circuit Lease Cost (M$)	1.221	1.449	1.817	2.028
US LHCNet Infrastructure Cost (M$)	1.102	1.209	1.239	1.270
Contribution from CERN in US LHCNet		(0.057)	(0.280)	(0.280)
Total US LHCNet Cost (M$)[37]	**2.323**	**2.601**	**2.776**	**3.018**

Table 3: Actual (2007) and projected (2008-2010 US LHCNET infrastructure budget breakdown (in M$) by category

Year	2007	2008	2009	2010
Routing and Switching Equipment	0.305	0.292	0.279	0.266
Salaries for Network Engineers	0.507	0.600	0.621	0.643
Network Perf. Monitoring and Test Systems	0.050	0.050	0.050	0.050
HW and Software Maintenance and Colocation	0.195	0.215	0.235	0.255
Travel Network Engineers	0.045	0.052	0.054	0.056
TOTAL US LHCNet Infrastructure Cost (M$)	**1.102**	**1.209**	**1.239**	**1.270**

[37] As explained above, the total in this table for FY2007 already takes into account part of the advanced FY2008 CERN contribution to US LHCNet, amounting to $ 223k. The remainder of this $ 280k contribution, $ 57k, is shown in the table as the remaining CERN contribution in 2008.

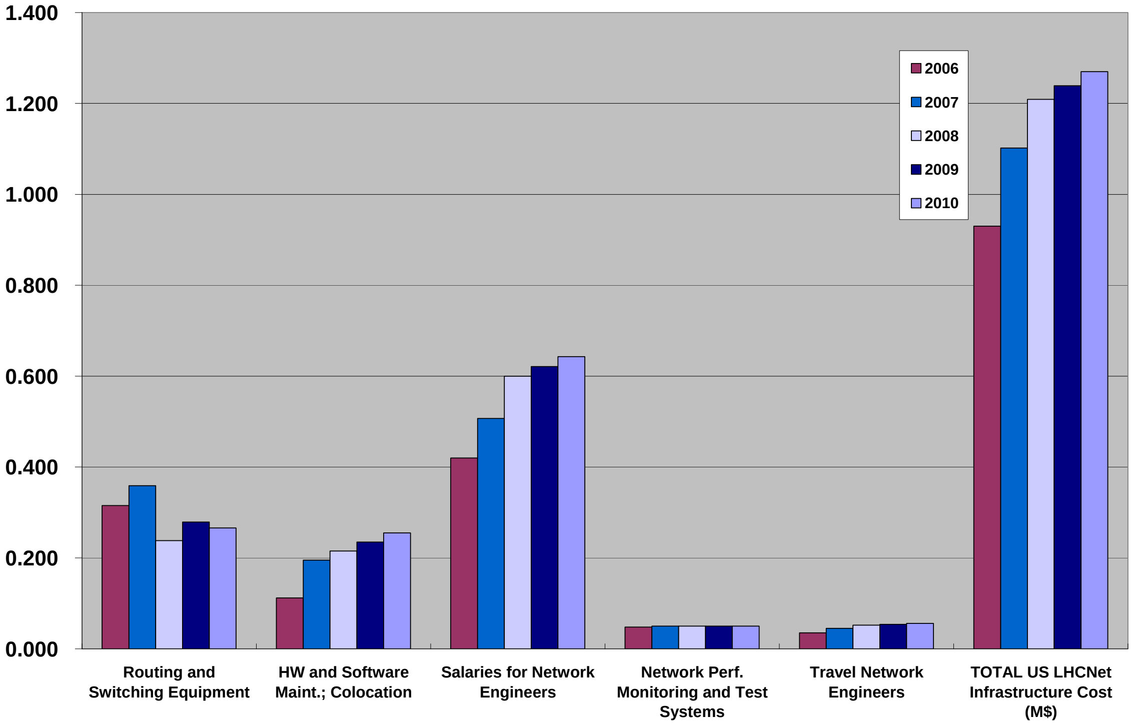


Figure 14 Breakdown of annual infrastructure costs for the US-CERN network (in M$)

4 External Collaborations

As mentioned in earlier sections of this proposal, we are working closely with the major mission- and research and education networks, most notably ESnet, Internet2, National Lambda Rail and Internet2, to ensure the efficient operation and management of transatlantic networking for HEP. With the transition to a hybrid network that includes circuit-oriented network provisioning and services, integrated with the grid systems of the major LHC experiments, we are also engaged in collaborations with the leading projects developing these new network and grid technologies, as summarized below, namely: Terapaths, OSCARS, and LambdaStation funded by DOE, and UltraLight, PLaNetS and DRAGON funded by NSF. Work with these projects, the DOE labs and the LHC experiments is an integral part of the overall plan to be able to deliver end-to-end network paths with bandwidth (or rather throughput) guarantees crossing multiple administrative domains, and thus to enable effective use of US LHCNet and its partner networks in support of the LHC program.

Terapaths[38]: This project, led by Brookhaven and Michigan, is investigating the integration and use of differentiated network services based on LAN QoS and MPLS in the ATLAS data intensive distributed computing environment as a way to manage the network as a critical resource; much as resource scheduler/batch managers currently

[38] http://www.atlasgrid.bnl.gov/terapaths/index.shtml

manage CPU resources in a multi-user environment.

OSCARS[39]: The focus of the ESnet On-Demand Secure Circuits and Advance Reservation System (OSCARS) is to develop and deploy a prototype service that enables on-demand provisioning of guaranteed bandwidth secure circuits within ESnet.

UltraLight[40]: Caltech is leading a consortium of high energy physicists, computer scientists and network engineers together with all the major optical network projects (HOPI, UltraScience Net and NLR), in the UltraLight project funded by NSF. UltraLight concerns the development of next-generation integrated network-aware Grid systems, and especially the means to use these systems effectively in support of the LHC and several other major DOE- and NSF-supported physics programs. With the support of Cisco systems, Ultralight has deployed a set of Cisco switch-routers that connect US LHCNet to some of the US Tier2 sites, as well as providing facilities for the deployment and testing of circuit-oriented network services.

PLaNetS[41] project: : The Physics Lambda-based Network System (PLaNetS) project funded by NSF for 2007-2008 is closely allied with the work on circuit-oriented services discussed throughout this proposal, PLaNetS supports Terabyte and multi-Terabyte "*data transactions*" that complete in minutes to hours, rather than hours to days, to significantly improve the overall efficiency of use network resources in the context of the LHC experiments' computing models. PLaNetS builds on the work in UltraLight to develop a core suite of high performance end-to-end data transfer tools and applications, enhanced by real-time network and end-system monitoring and management services, components of which have been developed and proven in sustained field-trials over the last five years. The PLaNetS paradigm for network operations and management includes (1) queues for tasks (transfers) of different lengths and levels of priority, coupled to dynamic (real or virtual) path-construction services for the most demanding, high-priority tasks, leveraging the work of the OSCARS, TeraPaths and LambdaStation projects, (2) a task "director" aided by end-system agents to partition the work among foreground, real-time-background and queued transfers, (3) end-to-end monitoring, network path and topology discovery, and path performance estimation and tracking services, based on the MonALISA and the Clarens frameworks, as well as SLAC/IEPM monitoring services, and (4) Policy-based network path-request and utilization services, incorporating the OSG infrastructures for authentication, authorization and accounting.

LambdaStation[42] Caltech has worked together with Fermilab on the LambdaStation project (DE-FG0104ER04-03). This project has enabled the very large, production-use mass storage systems at Fermilab to exploit the advanced research network infrastructures provided by ESnet and US LHCNet, using Fermilab's dark fiber and the ESnet Chicago MAN, between the Fermilab campus and Starlight. Necessary innovations have been implemented in the Fermilab local network and application environments to enable HEP applications (notably for US CMS) to send traffic, on a per-flow basis, across advanced network paths, specifically DOE's UltraScience Net.

[39] http://www.es.net/oscars/

[40] http://ultralight.caltech.edu

[41] http://pcbunn.cacr.caltech.edu/PLaNetS/PLaNetS_PIF_DraftV21.htm

[42] http://www.lambdastation.org/

DRAGON[43] project: The DRAGON project is developing technology and deploying network infrastructure that will allow advanced e-science applications to dynamically acquire dedicated and deterministic network resources. The objective is to link computational clusters, storage arrays, visualization facilities, remote sensors, and other instruments into globally distributed and application specific topologies.

Openlab[44] project: US LHCNet is interconnected at 10 Gbps to the CERN's Openlab testbed which develops data-intensive Grid solutions to be used by the worldwide community of scientists working at the LHC. US LHCNet infrastructure has been used by the project to experiment with high throughput data transfers across the Atlantic, using the latest data servers and network interfaces. To acheive these goals DRAGON has developed a GMPLS based control plane which includes advanced inter-domain service routing techniques and detailed application formalizations [2]. A reference network implementation has also been constructed in the Washington D.C. area [3].

Collaboration with equipment manufacturers and network providers. We have developed strong relationships with equipment manufacturers. We are currently collaborating with Intel to build high performance end-systems equipped with dual ported 10 GbE PCI-Express cards. We are also continuing collaborations (involving substantial network interface card donations) with Neterion[45] and Myricom[46]. For several years, Cisco has been a major partner supporting our R&D efforts by loans, donations, free access to their research waves on National Lambda Rail, and strong technical support. We are also currently discussing a joint R&D program with CIENA to help us in our development of connection-oriented network services across transoceanic and continental networks, managed by our MonALISA system.

As results of those collaborations, we have conducted a series of breakthrough data transfer trials over the last four. We broke the Internet2 land speed record eleven times between 2002 and 2005, increasing TCP performance across long haul networks from 400 Mbps to 7.3 Gbps. Working together with CERN, Fermilab, SLAC, Michigan and many other HEP partners, we captured the Supercomputing 2005 (SC05) Sustained Bandwidth Award for the demonstration of "High Speed TeraByte Transfers for Physics", where we aggregated a peak rate of 151 Gbps to the show floor at Seattle, which, significantly, is of the same order of magnitude as the total traffic expected in the early phase of LHC operation.

In 2006, with the advent of PCI Express bus in computers, we demonstrated stable flows to and from servers at close to 10 Gbps, and then focused on new application developments for higher speed storage-to-storage flows over long distances. At SC06 we demonstrated a new application FDT[47] (Fast Data Transport) developed by Caltech that provides storage-to-storage data flows sustainable for hours over long distance networks, that are limited only by the speed of the disks, controllers, and the file system. At SC06 we used FDT to demonstrate a sustained 17 Gbps storage-to-storage

[43] http://dragon.maxgigapop.net. The control plane is described briefly at: http://dragon.east.isi.edu/data/dragon/documents/dragon-ctrl-plane-overview-v1.0a.pdf

[44] http://proj-openlab-datagrid-public.web.cern.ch/proj-openlab-datagrid-public/

[45] http://www.neterion.com/

[46] http://www.myri.com/

[47] http://monalisa.cern.ch/FDT/

flow using a single 10 Gbps link in both directions.

5 Conclusion

Wide area networking is a fundamental requirement for HEP. U.S. physicists involved in the LHC are especially dependent on the development of reliable transatlantic networks and grid systems of sufficient capacity and capability, if they are to contribute effectively to the LHC physics program and take part in the physics discoveries.

The Caltech group has played an important role in the development of international networks for the HEP community, and has led the planning, development, operation and management of transatlantic networking on behalf of the U.S. for the last 22 years.

US LHCNet, operated and managed by the Caltech network engineering team in partnership with the CERN network team, ESnet and the network teams at Fermilab and BNL, is now an essential mission-critical resource for U.S. participation in the LHC. Building on this experience, we have therefore designed, developed and are now implementing a highly costs-effective four year plan to meet these needs between 2007 and 2010, as presented in this proposal.

The funding request for FY2008-FY2010 required to carry out the plan, which has been presented to the US LHC Network Working Group over the last year, has been accurately estimated based on current cost experience, the plan for periodic upgrades of the links, optical multiplexers and switch-routers (presented in detail in Annexes A and B), and the minimum engineering manpower required to operate and manage the network with high reliability and performance.

Provision of the requested funds for the next three years, enabling the plans now underway to be executed, will be an essential, major step required for the success of U.S. involvement in the LHC program, and of the LHC program as a whole.

US LHCNet Annexes

Table of Contents

Annex A: US LHCNet Technical Plan for 2007-2010 Based on CIENA CD/CI SONET Multiplexers 1

- US LHCNet Today; Transition to an Optical Switch-Fabric 1
- Phase I: Initial deployment (2007) 2
- Phase II (2008) 4
- Phase III (2009) 5
- Phase IV (2010) 6

Annex B: US LHCNet Architecture and Migration Plan 8

- Current Status 8
- US LHCNet Migration Plan 10
- Redundancy and Fallback Scenarios 13
  - RSTP Configuration 13
  - Equipment Diversity 15
  - Additional Capabilities 17

Annex C: US LHCNet Breakdown of Engineering Activities 19

Annex D: The CERN Campus Network and the LHC OPN 21

Annex E: ESNET and US LHCNet Bandwidth Planning for 2008-2010 23

Annex F: Network Services Infrastructure for managed high performance end-to-end data transfers 28

- Circuit Oriented Services 29
- Managed Network Services 29
- Operational Scenarios 36
- Authentication, Authorization, and Accounting (AAA) 41
- Inter-domain provisioning and local connectivity at data centers 41
- Data transfer and management tools 42
- Milestones 43
- Summary 44
- References 45

Annex G: The MonALISA Framework 46

- MonALISA System Design 46
- MonALISA Deployment in Grids 47
- MonALISA Network Monitoring and Management 48

Annex H: VINCI: Virtual Intelligent Networks for Computing Infrastructures 52

- End To End Optical Control Plane 52

Annex I: US LHCNet Request for Proposals 55

Annex A: US LHCNet Technical Plan for 2007-2010 Based on CIENA CD/CI SONET Multiplexers

US LHCNet Today; Transition to an Optical Switch-Fabric

The US LHCNet transatlantic network has evolved from DOE-funded support and management of international networking between the US and CERN dating back to 1985, as well as a US-DESY network in the early 1980's. The first US LHCNet 10 Gbps transatlantic link was installed in the Fall of 2004, connected to Juniper T320 routers at Starlight and CERN. In the Fall of 2005 we created a 10 Gbps "wavelength triangle" among Starlight, MANLAN and CERN, resulting in two 10 Gbps transatlantic links, with enough redundancy to ensure that Fermilab and BNL would each remain connected to CERN, even in the cases where one of the links fails. Because of the very high cost of all Juniper 10 Gbps interfaces (even 1 Gbps interfaces have a high cost) we changed over to more cost-effective Force10 switch-routers that support 10 GbE WAN-PHY at that time.

US LHCNet today consists of a set of multiple 10 Gbps links interconnecting CERN, MANLAN[1] in New York, Starlight[2] in Chicago, and the SARA PoP[3] in Amsterdam. The network has been architected to ensure efficient and reliable use of the 10 Gbps bandwidth of each link, up to relatively high occupancy levels, to cover a wide variety of network tasks, including: large file transfers, grid applications, data analysis sessions involving client-server software as well as simple remote login, network and some grid R&D-related traffic, videoconferencing, and general Internet connectivity.

During 2006-7 we decided to adopt the "circuit-oriented services" paradigm, for the reasons given in the main body of this proposal. We also found, on several occasions, that the Force10 architecture would not allow stable reconfiguration and re-routing of traffic when a link "flaps" (goes up and down repeatedly). We therefore turned to the CIENA Core Director CD/CI multiplexers that provide stable fallback in case of link outages at Layer 1 (the optical layer), and full support for the GFP/VCAT/LCAS protocol suite.

In February 2007 US LHCNet deployed a new Point of Presence in Amsterdam, in the SARA Computing Facility. Having a PoP in Amsterdam allows us to take advantage of the very diverse transatlantic connectivity options available there, along with the continental GEANT2 infrastructure across Europe for Amsterdam – Geneva connectivity. In March 2007 we deployed our new CIENA CD/CIs in Chicago, New York City and later in Geneva. The fourth CIENA device is scheduled to be shipped to Amsterdam on August 15 and installed on August 21.

The stepwise migration to the CIENAs, and the plans to evolve their configuration in 2008-2010 are described in detail in this Annex. The equipment configuration and the

[1] The MANLAN exchange point is designed to facilitate peering among US and international research and education networks in New-York. See http://networks.internet2.edu/manlan/

[2] StarLight is an international peering point for research and education networks in Chicago. See http://www.startap.net/starlight

[3] SARA Computing and Networking Services is an advanced ICT service center that supplies a complete package of high performance computing & visualization, high performance networking and infrastructure services. See http://www.sara.nl

future evolution and upgrades were carefully considered according to the foreseeable budget and the vendor's development plans[4]. Following the discussions, we designed a four year plan meant to accommodate the US LHCNet need for bandwidth, reliability and added functionality. We devised the most cost effective plan possible, working with CIENA to (a) stay with the half-rack CI chasses, until at least 2010, thus avoiding the higher costs of the full rack CIENA CD chasses and the increased collocation costs (b) therefore migrating to the more cost-effective double-density modules and double-capacity switch fabric as they become available, and as space in the chasses at each of our 4 PoPs is needed for additional connections, and (c) returning the older modules to recover part of the original cost.

Phase I: Initial deployment (2007)

Phase I started in March 2007 and included the deployment of four Ciena CD/CI nodes in Chicago, New York, Geneva and Amsterdam. The initial port count accommodates all the existing OC-192 links, and has enough 10 GbE links for intra-PoP connections. The Chicago and New York nodes were installed in March, and the Geneva node was installed in April; the Amsterdam node will be installed at the end of August. The initial deployed configurations are shown in detail in Figure 1.

There are two types of linecards present in our configuration:

- single port linecards with enhanced 10 GbE interfaces (ESLM), each of which can be used as 1x10Gbps or 10x1Gbps Ethernet
- two port linecards with OC-192/STM-64 interfaces

The number of ports of each type at each location, and the PoP-to-PoP interconnections at each phase from August 2007 to the end of 2010, are shown in the following figures.

[4] We also considered potential competing vendors. Only Nortel initially appeared to have the capability to provide the functionality needed, and we investigated their offerings and development schedule in depth. In the end, we determined that there is a 2-3 year CIENA lead in the development of mature GFP/VCAT/LCAS products, and that in addition products with sufficient port density and backplane capacity would not be available in time from Nortel. As a result we chose CIENA as the vendor. A similar conclusion was reached by Internet2 and the DRAGON project. We were also encouraged in our choice by the good experience of UltraScience Net with the CIENA Core Directors.

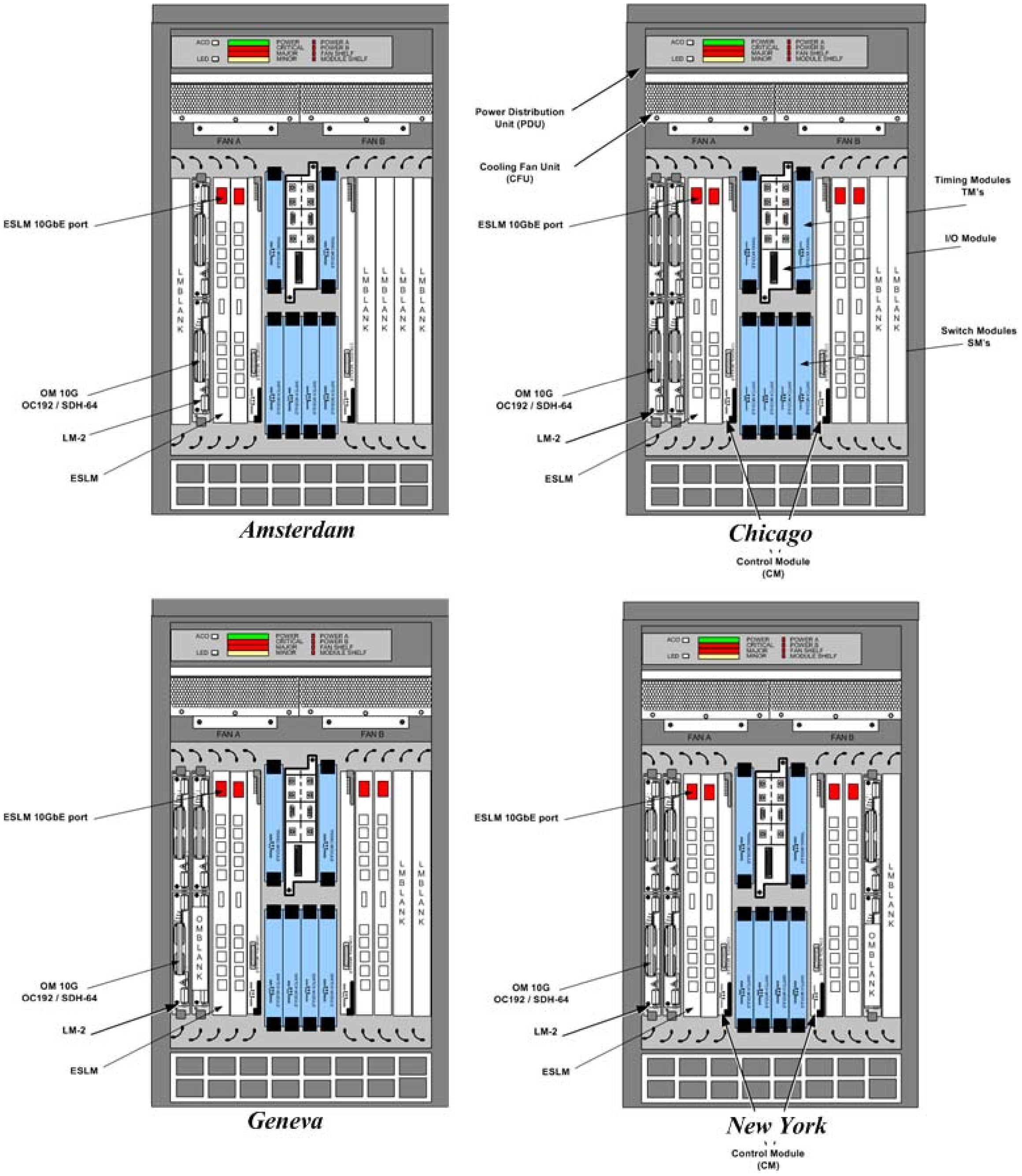
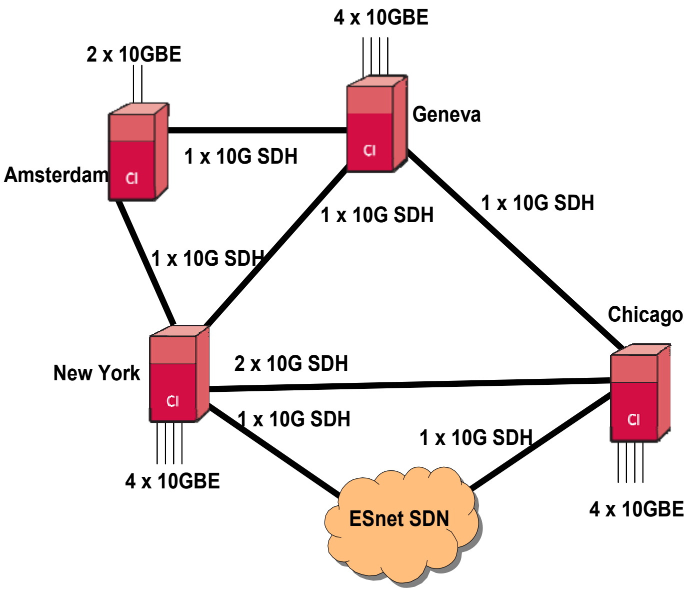


Figure 1 ***Phase I View of CIENA Node Configurations (End-August 2007)***



Ports at Each PoP	10 GbE Ports	OC-192 /STM-64 Ports
AMS	2	2
GVA	4	3
NYC	4	5
CHI	4	4

Figure 2 ***Phase I Circuit Layout and Port Count at Each Site (August 2007)***

Phase II (2008)

In Phase II we simply increase the number of interfaces at each PoP. There is no change in technology, all interface types are the same as in Phase I.

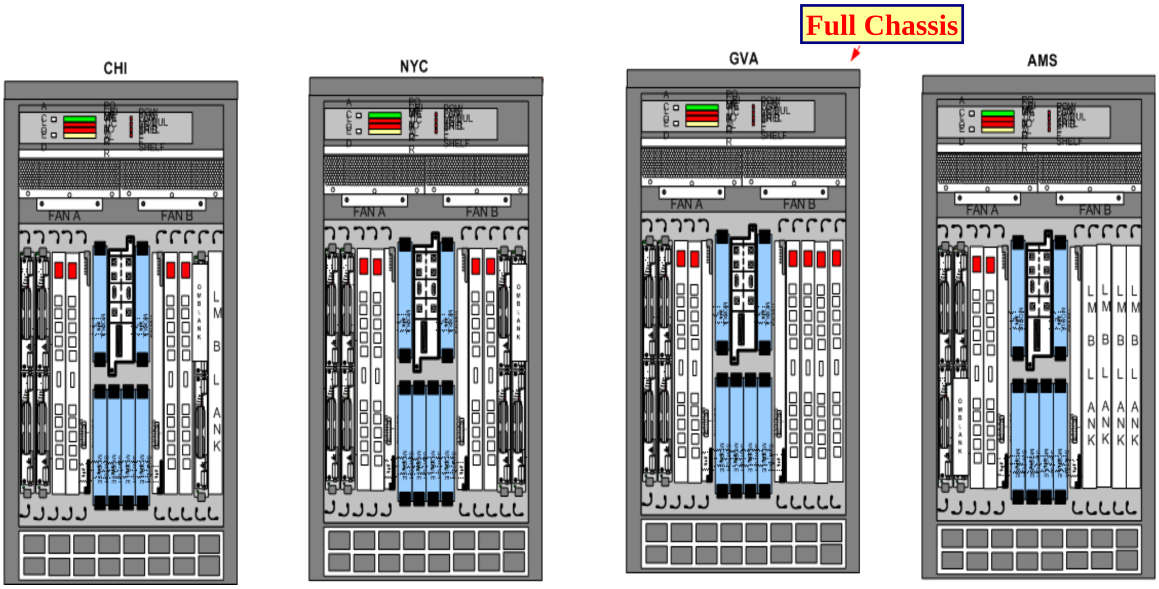
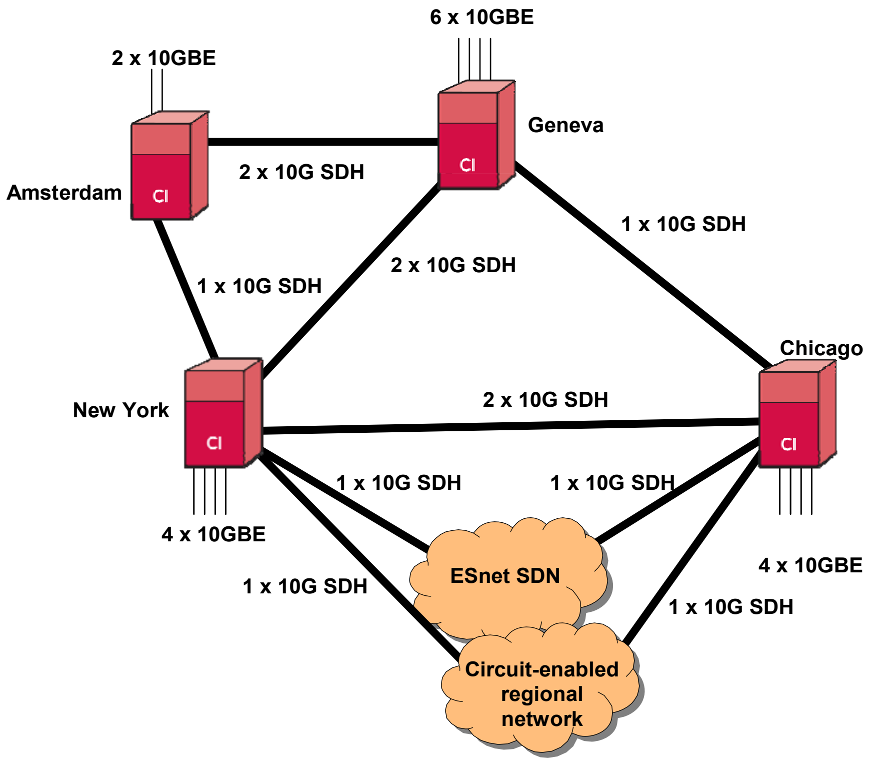

Full Chassis

Figure 3 ***Phase II View of CIENA Node Configurations (2008)***



Ports at Each PoP	10 GbE Ports	OC-192 /STM-64 Ports
AMS	2	3
GVA	6	5
NYC	4	7
CHI	4	5

Figure 4 *Phase II Circuit Layout and Port Count at Each Site (2008)*

Phase III (2009)

In Phase III we will introduce the ESLM2 card (two port 10 GbE Ethernet card). The new linecards will replace the existing one-port cards, to achieve the required port density at all PoPs.

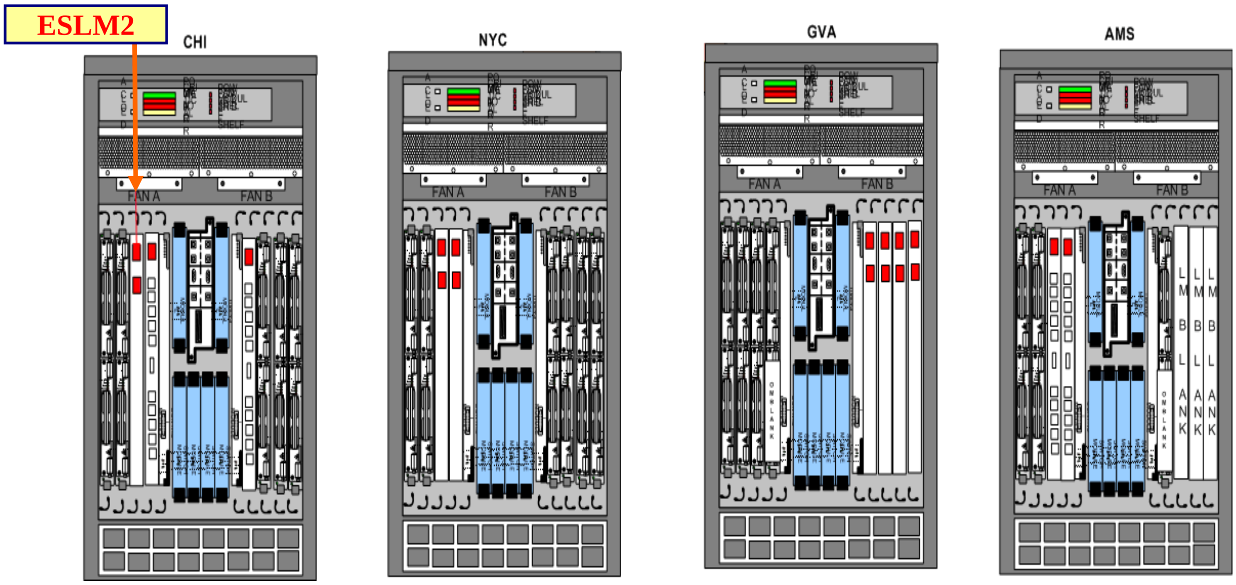
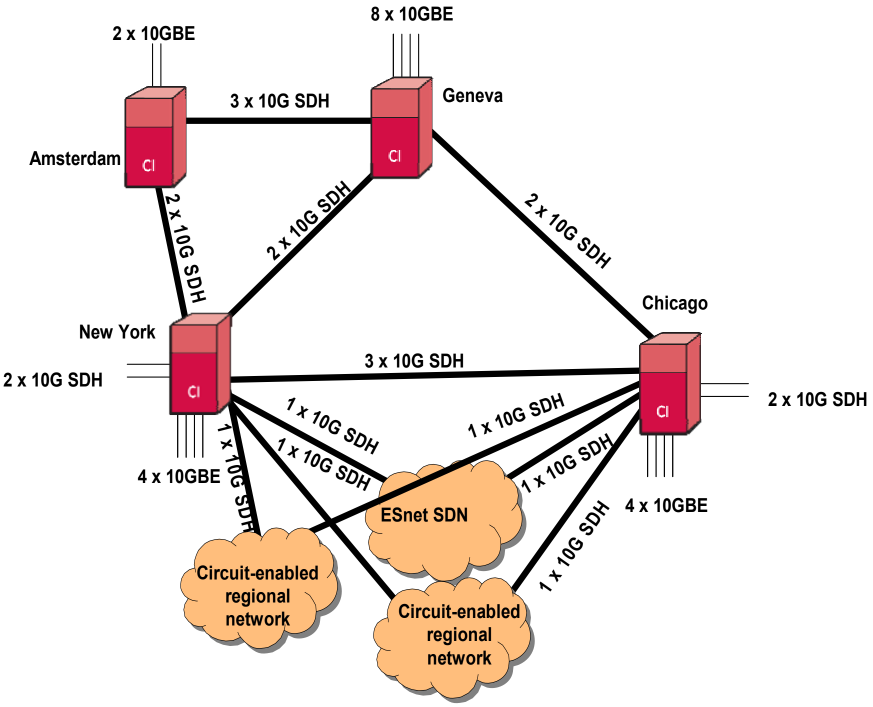


Figure 5 *Phase III View of CIENA Node Configurations (2009)*



Ports at Each PoP	10 GbE Ports	OC-192 /STM-64 Ports
AMS	2	5
GVA	8	7
NYC	4	12
CHI	4	10

Figure 6 ***Phase III Circuit Layout and Port Count at Each Site (2009)***

Phase IV (2010)

Since the previous phase leads to saturation of the currently available TDM switch-fabric capacity of the New York and Geneva CD/CI chasses, in Phase IV we will replace the switching-matrix in these two chasses with one that has twice the capacity. The new matrix also will allow for a greater number of higher capacity linecards. We will replace the existing two port OC-192 cards with then-available higher density 4 port cards where needed, to fit into the CD/CI chasses. As agreed with CIENA, the old cards will be traded in against the more cost-effective ones, recovering part of the original cards' cost.

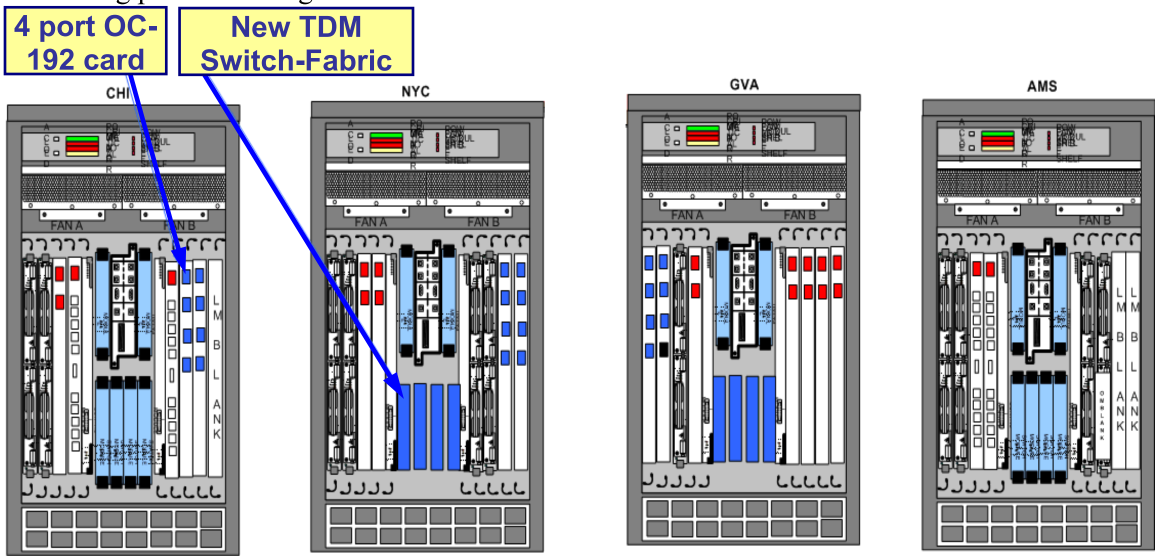
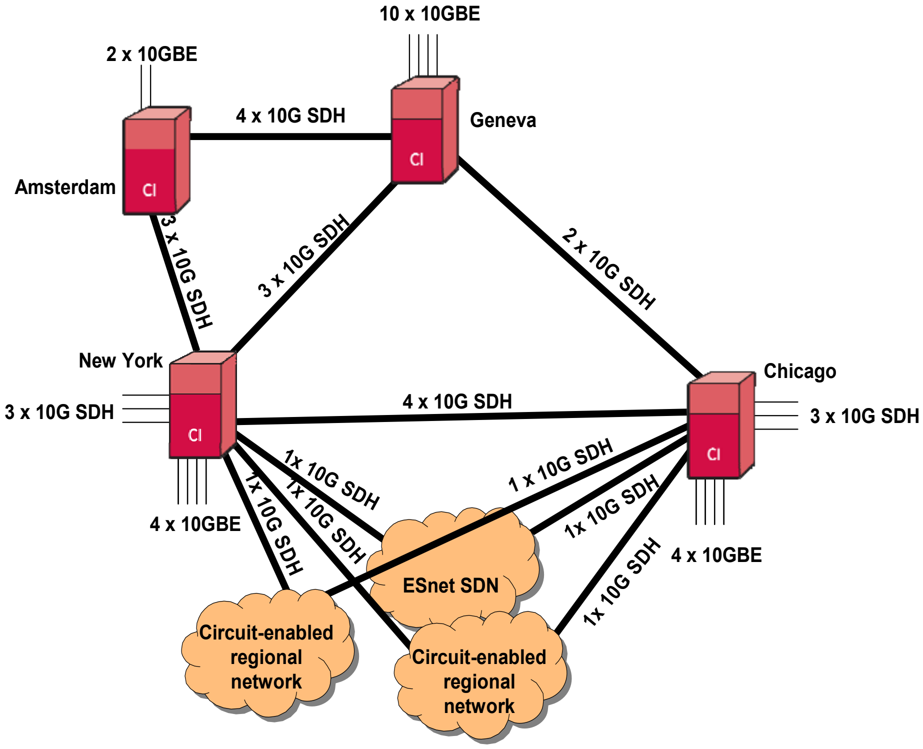


Figure 7 ***Phase IV View of CIENA Node Configurations (2010)***



Ports at Each PoP	10 GbE Ports	OC-192 /STM-64 Ports
AMS	2	7
GVA	10	9
NYC	4	16
CHI	4	12

Figure 8 *Phase IV Circuit Layout and Port Count at Each Site (2010)*

Annex B: US LHCNet Architecture and Migration Plan

Current Status

The US LHCNet network has been designed with redundancy and resiliency in mind in order to provide a highly reliable, uninterrupted service to the US Tier1s, while being very cost effective at the same time. Our core equipment consists of Force10 routers (using WAN-PHY interfaces to terminate SONET OC-192 circuits) and CIENA CD/CI SONET multiplexers for the US LHCNet network and services. The current map of the network is shown in *Figure 1*.

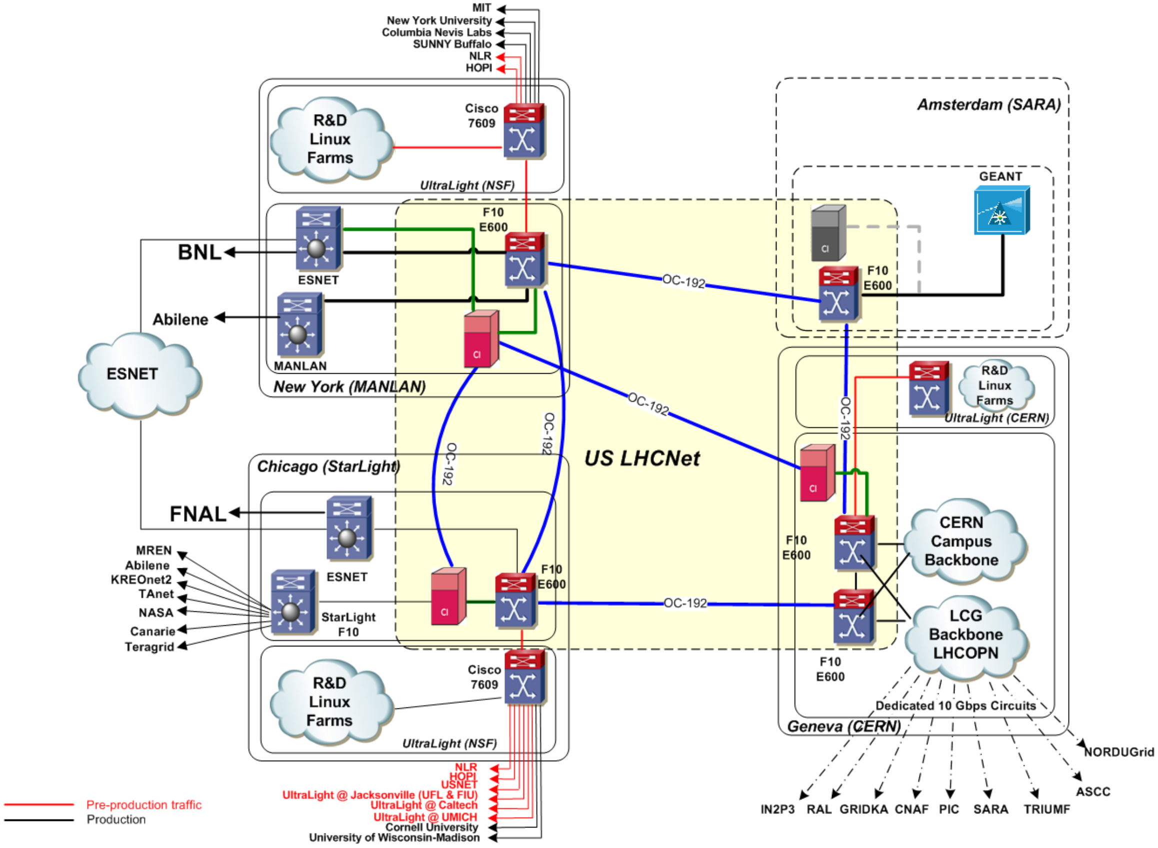


Figure 9 ***US LHCNet Network Map (July 2007)***

Several important factors were taken into consideration for the design and the operation of the network:

- **Transatlantic circuit redundancy:** each of our transatlantic circuits (*Figure 2*) runs over a different submarine cable; a submarine cable cut can take much longer to repair than a continental link (on the order of days to weeks, as opposed to hours to a few days in the case of continental circuits). Therefore it was imperative that one such cable cut would not bring the entire network down for an extended period of time.
- **Continental circuit redundancy:** a recent study by DANTE (www.dante.net) shows that many of the OPN (Tier0-Tier1 network) circuits share a common

segment inside the European continent. While having entirely diverse physical paths is not always possible due to budget constraints, the US LHCNet engineers have been working together with the DANTE engineers to ensure a minimum level of diversity for the underlying fiber paths supporting the LHC OPN (Optical Private Network) circuits.

- **Equipment redundancy:** all of our core network devices come with dual power supplies and dual control/supervisor modules, so that one module failure doesn't bring an entire node down; also, our OC-192 circuits are connected as much as possible to different linecards, to avoid a single card failure bringing down multiple circuits.

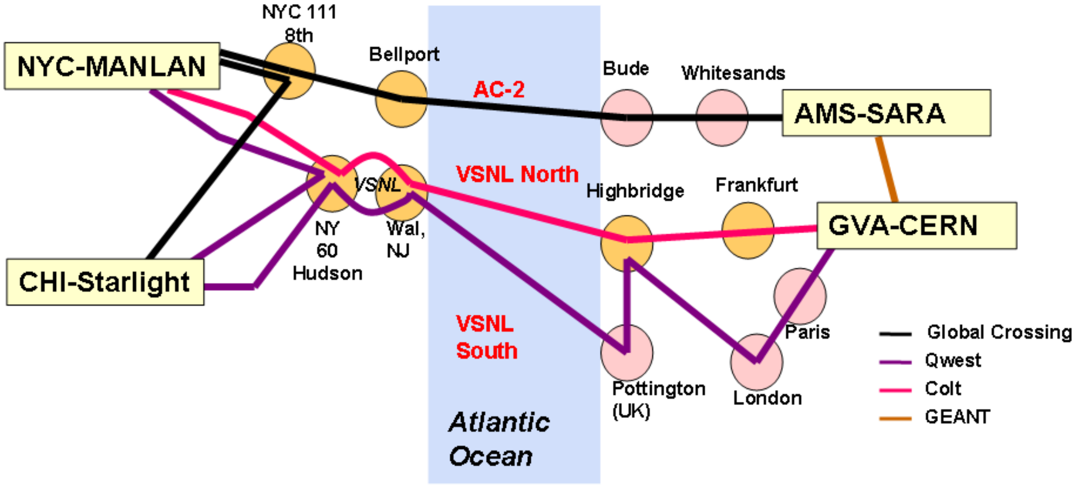


Figure 10 ***US LHCNet Transatlantic Circuits Currently Installed (Mid-2007)***

- **Equipment diversity:** we keep some of our links connected to the Force10 routers to maintain equipment diversity at the New York and Chicago PoPs; in the unlikely scenario that an entire node will go down, the traffic can be re-routed through the secondary equipment. There is a drawback for this setup, having equipment diversity using both the Force10s and the CIENAs doesn't allow for the Layer 1 fallback capabilities of the CD/CIs, and it also limits the use of new services in the normal situation. Therefore we will migrate, as shown below in the following figures, to a topology with sufficient redundancy that has a minimum number of transatlantic links connected directly to the Force10s. The migration will be completed over the next 12 months, by July 2008. This will allow us time to install and commission the fourth link across the Atlantic, and to gain the necessary operational experience with the CIENA CD/CIs and the new services.

  Once the transition is completed, in the unlikely even of a failure of a major piece of equipment terminating one of the transatlantic links, an emergency intervention at the remote PoP would be scheduled, to physically move the circuits from the failed equipment to the corresponding secondary one, within 2 to 4 hours of the failure.

- **Layer 2 fallback:** we are using the RSTP protocol[5] to assure redundancy for the VLANs provisioned across US LHCNet involving the Force10 routers; the switchover of the layer2 circuits is automatic in case of a circuit failure, and takes less than 1 second in most cases.

- **Layer 1 fallback:** While the Layer 2 RSTP-based redundancy functions well in most situations, it has a number of limitations:
    - Sometimes it's not fast enough to keep the existing TCP sessions up (which leads to the flapping BGP sessions and extended unavailability, on the order of minutes)
    - It cannot deal with degrading and especially with flapping circuits
    - The VLAN-based setup doesn't provide strict QoS guarantees

  To address all these limitations we will use the Mesh Restoration feature of the CIENA CoreDirectors. This will provide fast (<50ms) and efficient fallback for Layer 1 circuits across all the existing OC-192 links. The SONET multiplexer monitors and takes into account the circuit quality, and can trigger a switchover in case of a circuit outage or degradation; or it can refrain from switching over in case of rapidly flapping circuits in order to preserve the network stability. The possible impairment scenarios, and the corresponding responses of the US LHCNet network, are described later in this Annex.

- **Layer 3 fallback:** The current LHC Computing Model calls for direct Layer 1 or Layer 2 circuits between all Tier1s and CERN. However, US LHCNet maintains a number of redundant peerings with CERN, ESnet, Internet2, NLR and other major research networks in the U.S. to help maintain the necessary connectivity to the US Tier1 and also Tier2 sites, even under adverse conditions. As a last resort, the current peering arrangements among CERN, US LHCNet, ESnet, GEANT2 and the US Tier1s – BNL and FNAL - would allow traffic to flow over general purpose IP connections or over the public IP infrastructure; however, this is not a desired behavior for Tier0-Tier1 traffic. The alternative now under discussion is additional Tier1-Tier1 connectivity and transit arrangements (the currently preferred scenario in the LHC OPN)

US LHCNet Migration Plan

The ongoing deployment of the Ciena CD/CI devices poses a number of operational challenges, in order to keep the network functional and to simultaneously provide a full service to the US Tier1s while progressively migrating some of the current links from the Force10 devices. We have therefore carefully staged and scheduled a number of migration steps, to ensure we maintain uninterrupted network service while at the same time having enough time to familiarize ourselves and extensively test the new technologies present in the CIENA Core Directors. Due to the heterogeneous nature of

[5] The Rapid Spanning Tree Protocol (IEEE 802.1w). See www.cisco.com/warp/public/473/146.html

the terminating interfaces at each PoP (the Force10 interfaces are WAN-PHY, 1310nm and 1550nm, the CIENA interfaces are SONET I64-2/SR2 1550nm) each circuit migration has to be done simultaneously at both ends.

March, 2007: After the initial deployment of the Chicago and NYC CD/CIs, we started by migrating one of the backup links between Chicago and New York to the new devices. This would allow us to test and validate the newly installed equipment. The resulting configuration is shown in Figure 11.

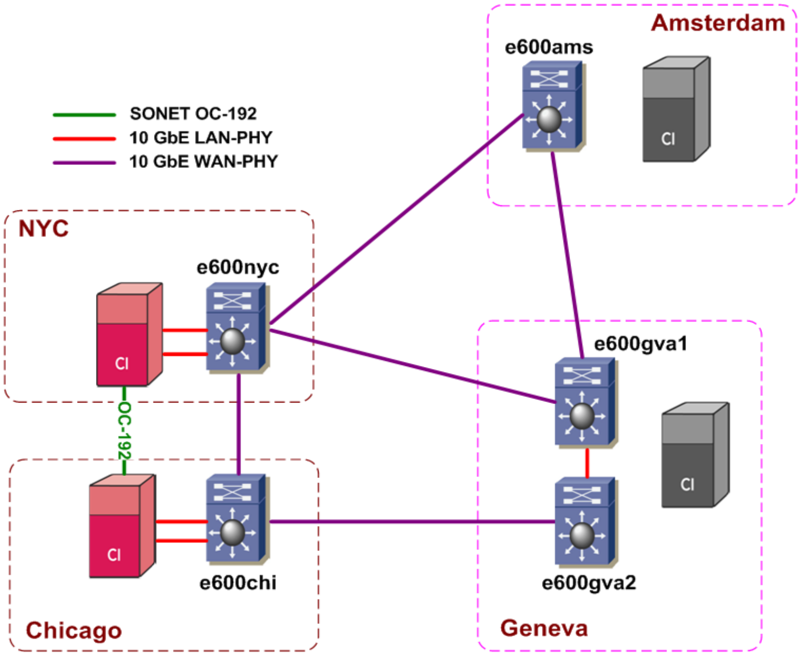


Figure 11 ***Ciena Migration Status, March 2007***

July 2007: After the installation of the Geneva node, we have decided to connect it directly to the network, as shown in Figure 12. Because the terminating interfaces at both ends of a link need to be the same, our preferred solution was to move the Geneva-New York traffic over the Geneva-Amsterdam-New York path and connect the New York – Geneva Colt link to the Core Director (see the figure). This setup effectively creates two parallel networks: one consisting of the Force10 switch-routers, and the other one consisting of the new Ciena CD/CI multiplexers.

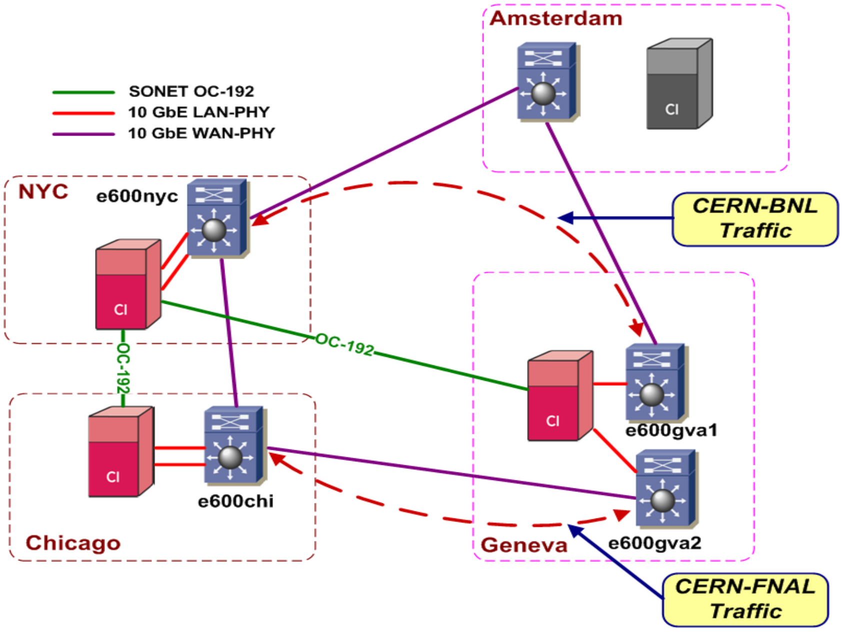


Figure 12 ***Ciena Migration Status Map, July 2007 (current)***

August 2007: After the upcoming installation of the fourth Ciena node in Amsterdam this month, we will reconfigure our network as shown in Figure 13. This will allow us to fully test the capabilities of all four new devices, as well as to develop the required dynamic network services without disrupting the production traffic. This is a milestone configuration, as it provides network stability, equipment diversity and optimal link utilization of the three transatlantic US LHCNet links.

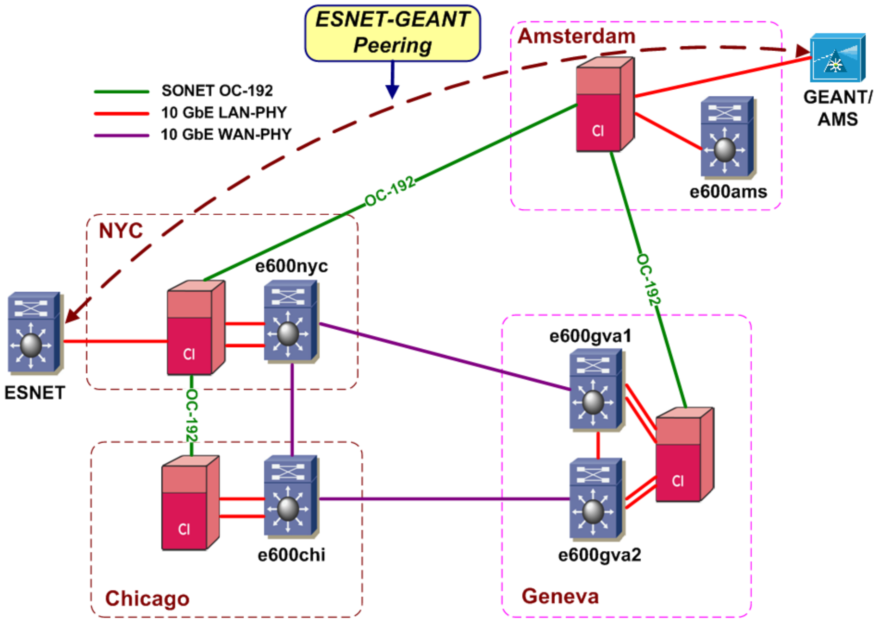


Figure 13 ***Ciena Migration Status, End-August 2007***

In order to better support the HEP traffic between the US and Europe (in particular the traffic between US Tier1s and European Tier2s), US LHCNet and ESNET have signed an MOU to allow ESNET to peer directly with GEANT over the US LHCNet link between NYC and Amsterdam. US LHCNet will provision a static circuit over this link. The speed of this link is under discussion, but will most likely be half the available bandwidth, i.e. 5 Gbps at the outset, and will be adjusted as needed according to experience.

February 2008: After completing the planned upgrade of the US LHCNet circuit topology for 2008 (see Annex A) following the responses to the recently-released RFP and the new link installation at the end of 2007, we will start to gradually move the circuits to the “final” configuration planned for the first LHC physics run. By February 2008 we would have acquired sufficient operational experience and development tools to provide a broad range of circuit-oriented scheduled services across US LHCNet. Figure 14 reflects what we think is the most probable outcome of the RFP, but the final configuration might differ slightly, according to the offers received and their cost.

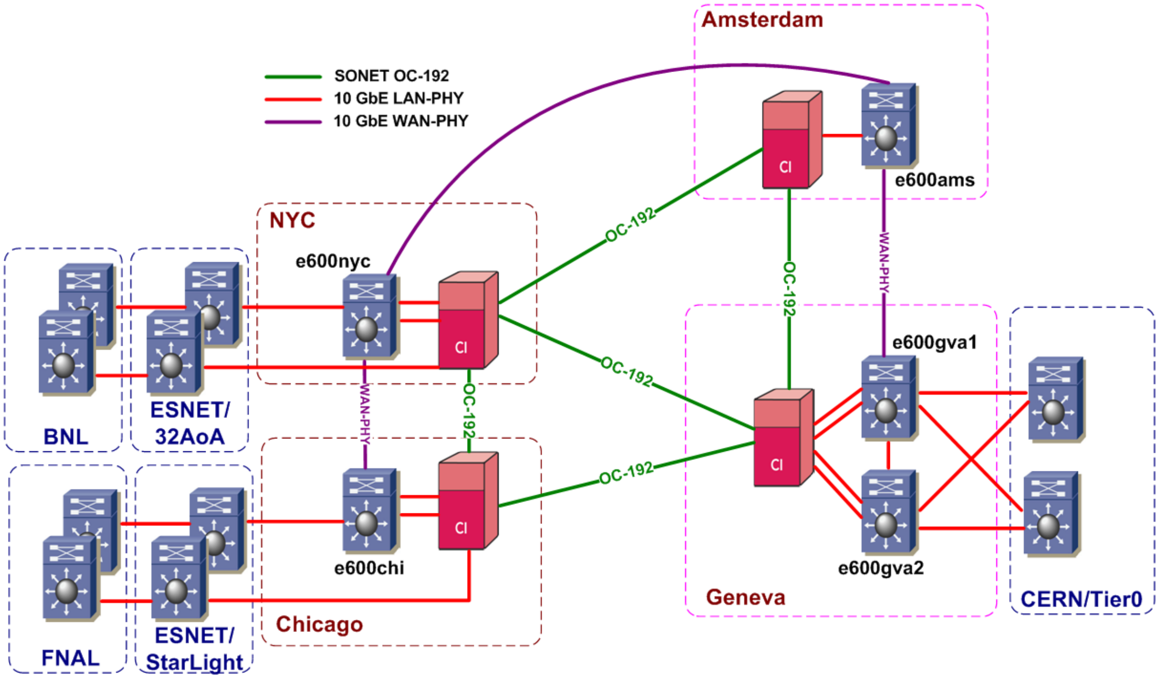


Figure 14 US LHCNet Configuration for the First LHC Physics Run (July 2008)

Redundancy and Fallback Scenarios

The US LHCNet Network was designed with redundancy and resiliency in mind. In the following paragraphs we describe some of the failure modes, and how the network will react in each of these situations to mitigate the failure.

RSTP Configuration

In the current network setup (see Figure 12) we are still relying on RSTP in case of link failure. RSTP builds and maintains a loop-free topology over the existing links.

In case one of the links fails (see Figure 15), the Layer 2 VLANs are automatically re-routed over the remaining links. Under normal situations, this takes less than a second, so it doesn't affect already established TCP connections (such as BGP peerings). When the failed link comes back up traffic is resumed over the original link[6].

We use this setup for the peerings between CERN and various US and international research networks, and for transporting the corresponding traffic flows between CERN, New York and Chicago.

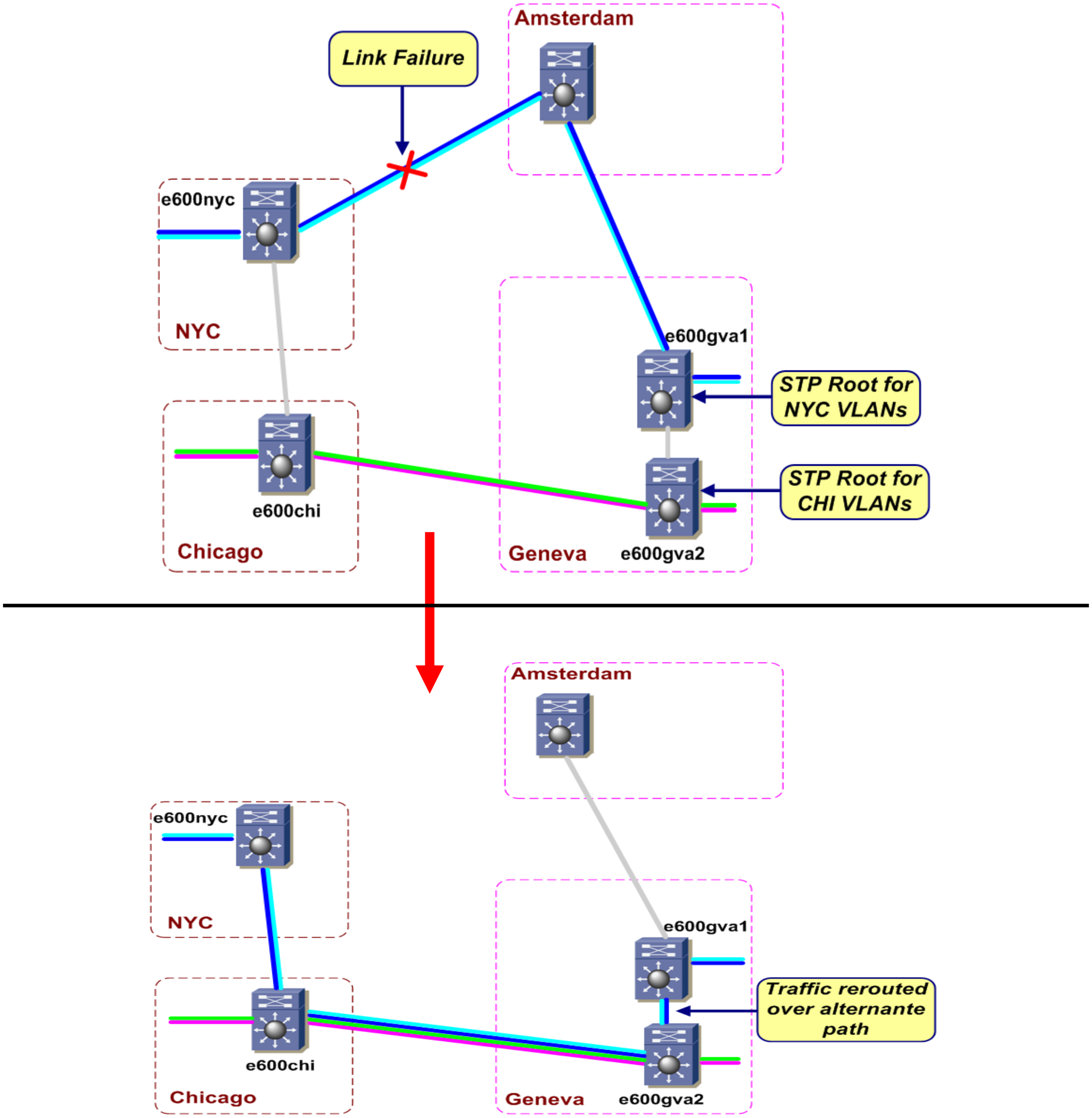


Figure 15 ***Following a link failure, traffic is re-routed over the available paths***

[6] However this can be a drawback in case of a flapping link, as described above.

Equipment Diversity

The reason for having equipment diversity is to be able to provide two alternate paths between CERN and each of the Tier1 centers in the US that will not share a single point of failure (either a link or a major piece of equipment terminating one or more links). The US Tier1s have expressed concerns that the failure of the US LHCNet or ESnet edge equipment facing them would interrupt the stream of data coming from the Tier0 for an extended period of time, irrespective of the available paths across the Atlantic or across the metropolitan areas of Chicago and New York. Having equipment diversity also allows us to perform software maintenance on the US LHCNet equipment *without interrupting the end-to-end traffic flow*. By carefully scheduling the upgrades we can re-route the traffic ahead of time, perform the software upgrade and then move the traffic back.

Due to the deployment of both Force10 and Ciena devices at all PoPs, US LHCNet has already achieved equipment (node) diversity at its edge locations. ESnet has the capability of achieving equipment and path diversity from the US LHCNet demarcation point to the Tier1 edge device(s). BNL and Fermilab are in turn working towards achieving equipment diversity from the ESNet edge to the computing and storage facilities at the respective labs.

Equipment failure cannot be handled in an optimal way without manual intervention (i.e. physically moving the circuits away from the failed equipment) but we can take some automated actions which can keep the traffic flowing to its final destination over a longer path. The actions to take in case of equipment failure at the edge of the network have to be coordinated with all our partner networks (CERN, BNL, Fermilab and ESnet). We present some of the possible scenarios for a network element failure below.

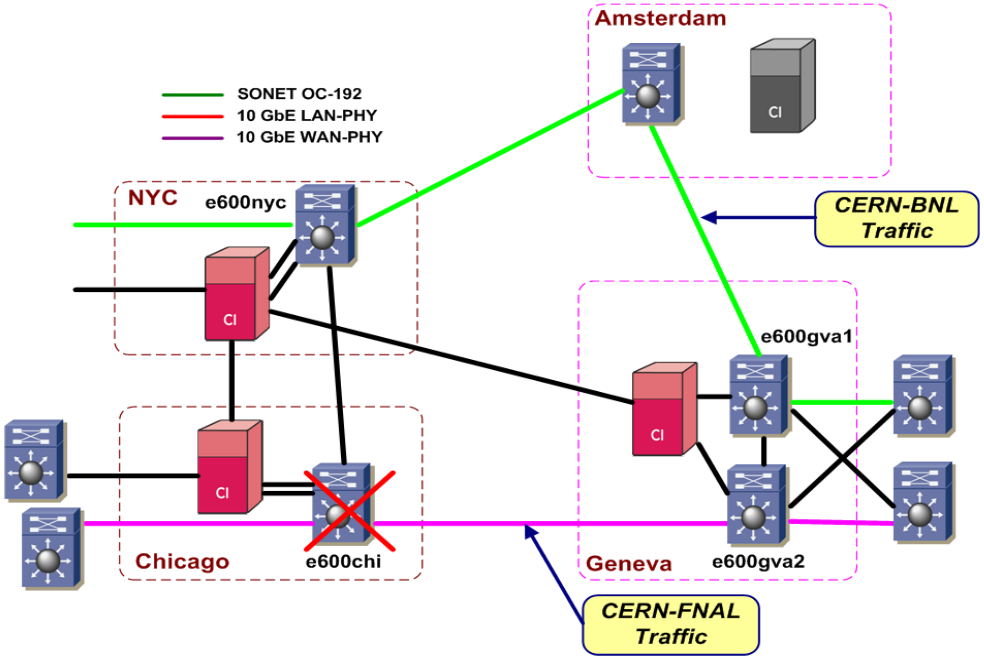


Figure 16 ***A network element failure should not interrupt the traffic flow***

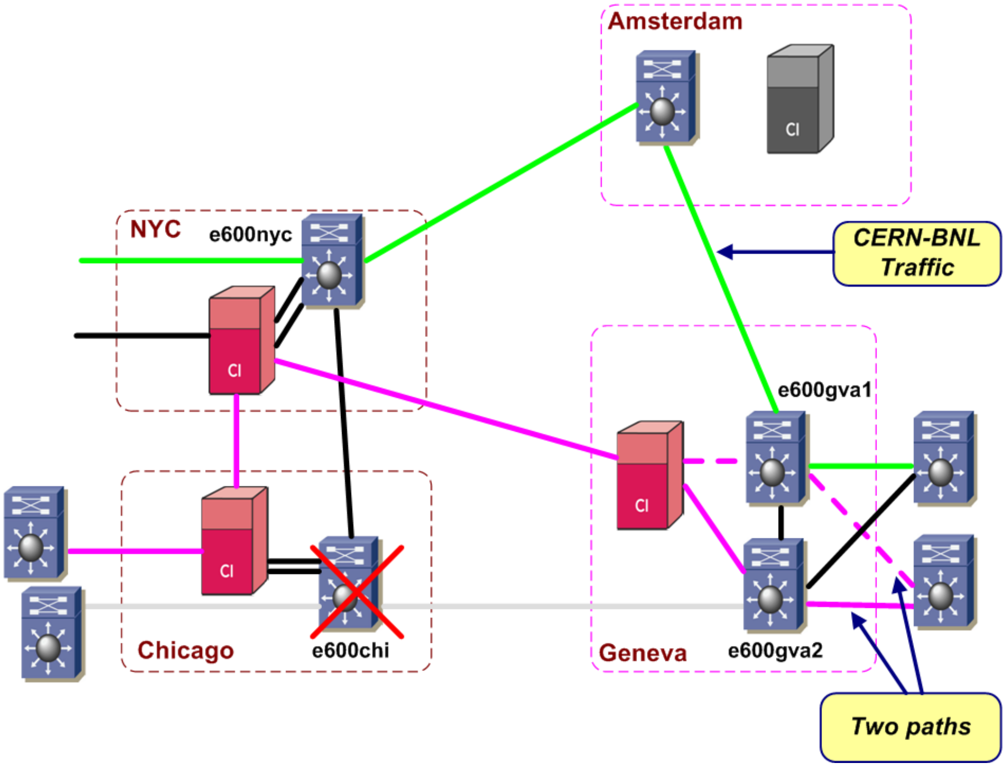


Figure 17: *Failure mode 1 (Automated): Traffic is re-routed over a longer path bypassing the failed equipment*

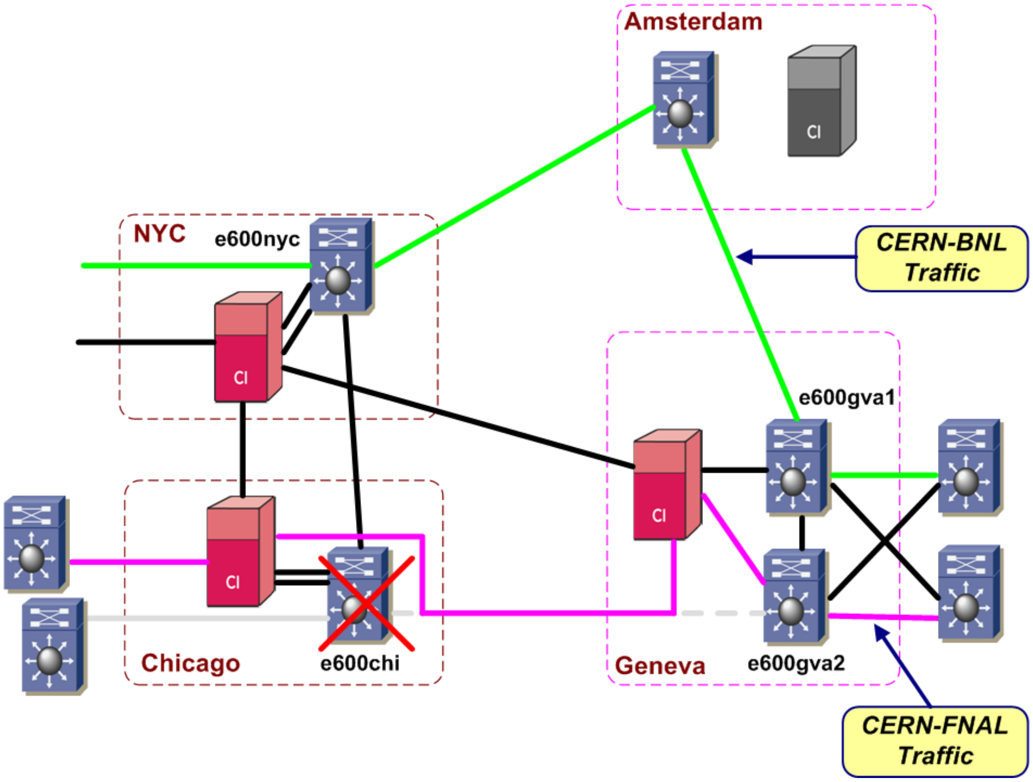


Figure 18: *Failure mode 2 (Manual intervention required): The link is reconnected to another device*

Additional Capabilities

With the deployment of the CIENA devices, US LHCNet has acquired additional technical capabilities which would allow us to cope with complex failure modes in a dynamic way. The CD/CI devices can automatically monitor and react to link degradation (BER rising above a configured threshold) and are more robust when it comes to flapping links (a hold-down timer can be configured which would prevent a circuit from being re-routed rapidly back and forth across the network, disrupting the overall stability. But perhaps the most important ability is to associate priorities to each of several circuits provisioned across a single physical link. In case of need, the circuits are re-routed while continuing to guarantee a designated bandwidth level to each of them (see Figure 19 and Figure 20), taking the circuit-priorities into account.

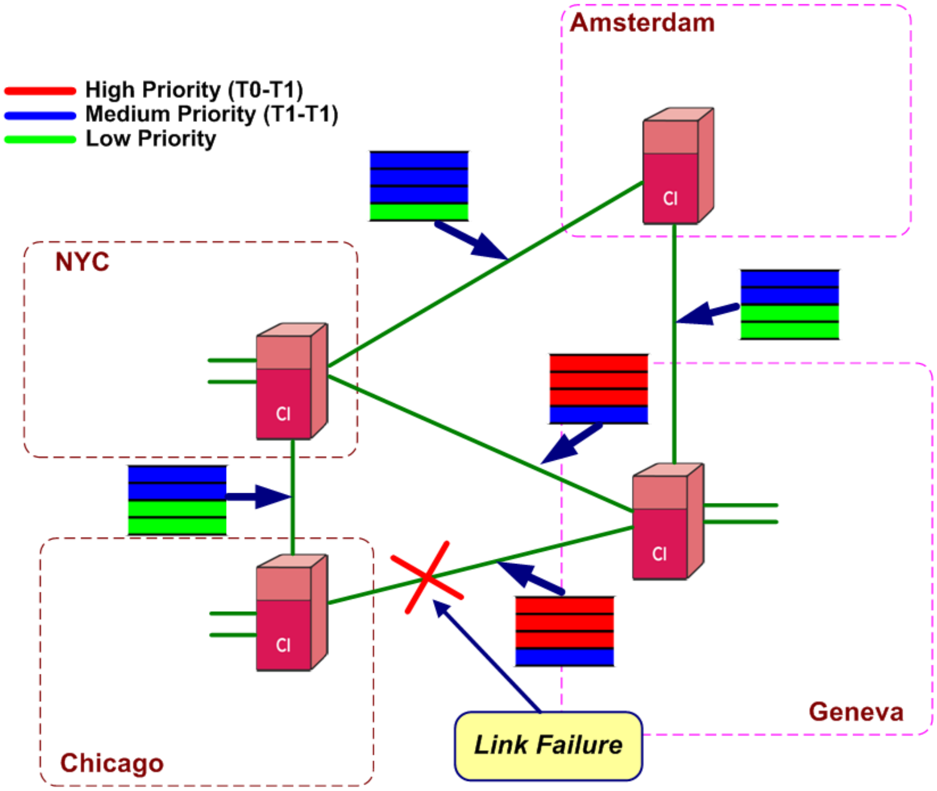


Figure 19 ***Link Failure with Core Directors***

Using a CIENA feature called Mesh Restoration, the provisioned circuits over a failed SONET link can be re-routed according to priorities and can preempt lower priority circuits. The fallback is automatic and very fast (<50ms once the failure is detected)

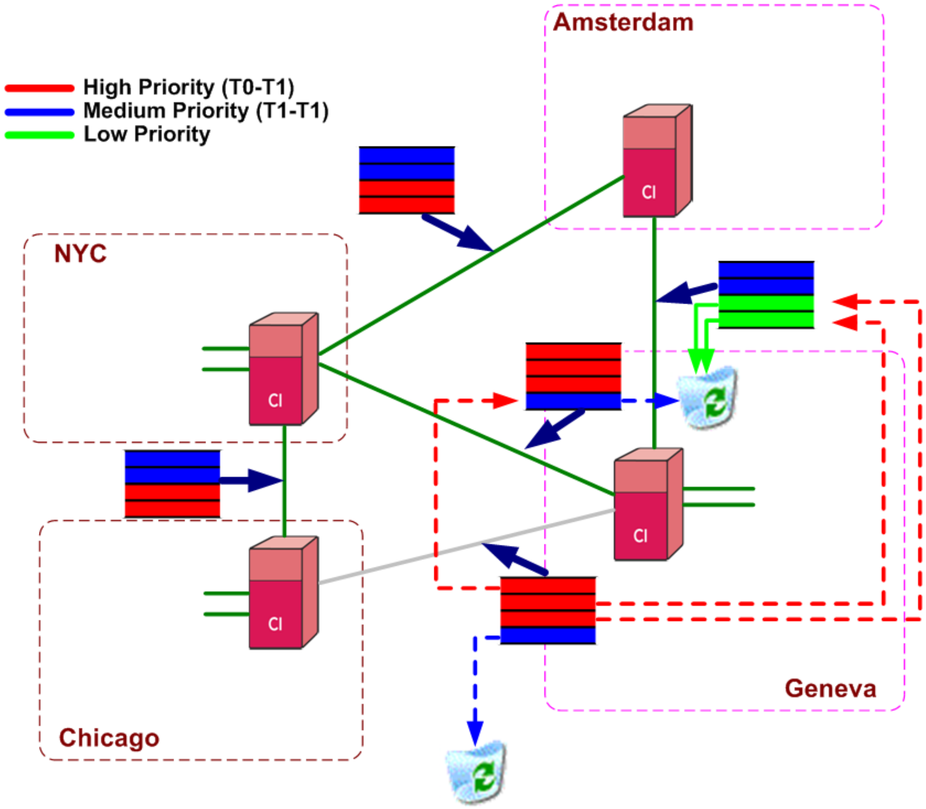
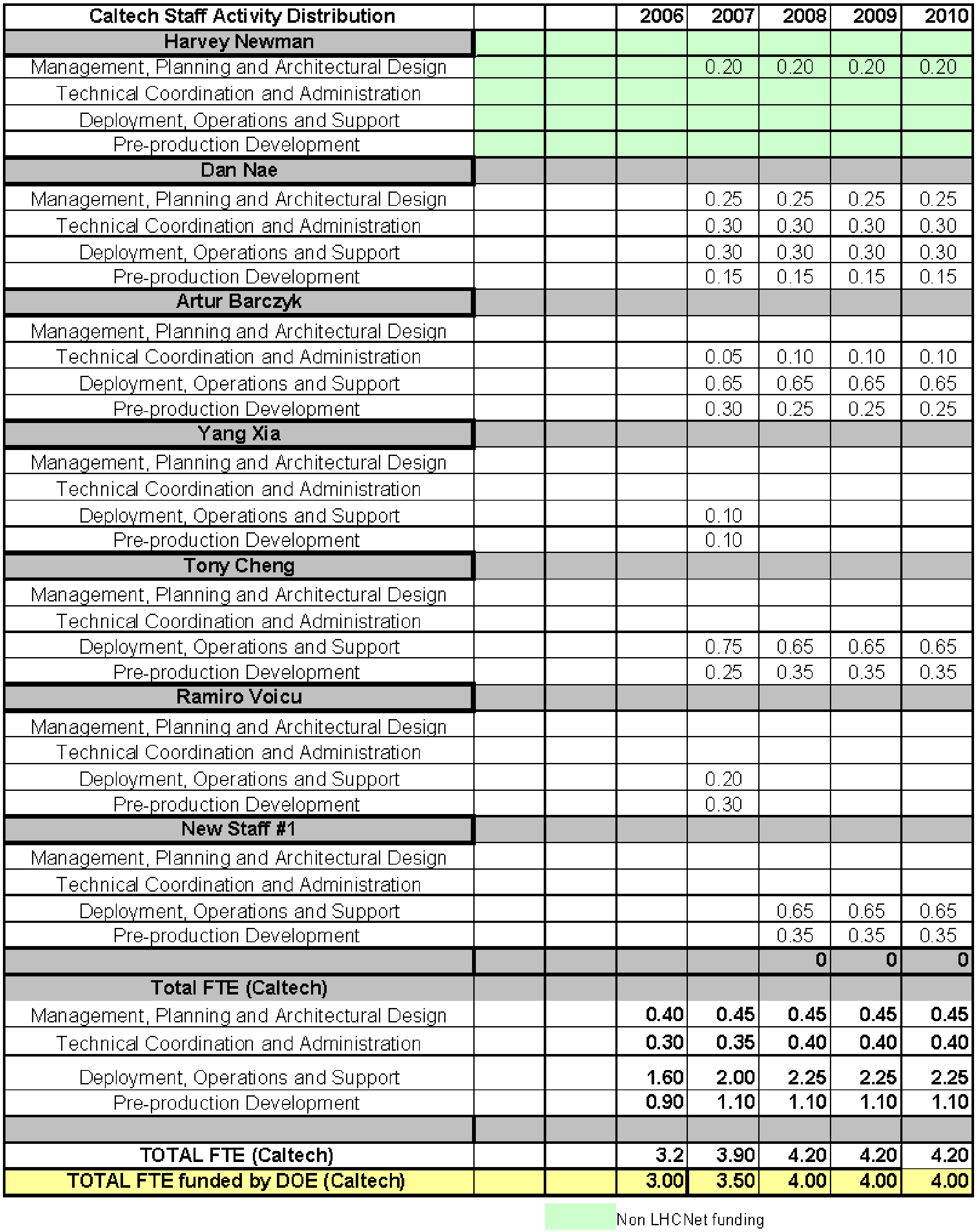


Figure 20 High priority circuits are mesh-restored, and can preempt the lower priority ones which may be discarded

Annex C: US LHCNet Breakdown of Engineering Activities

Table 1: Distribution of activities for each Caltech US LHCNet team member: actual (2007) and planned (2008-2010)



Caltech Staff Activity Distribution			2006	2007	2008	2009	2010
Harvey Newman							
Management, Planning and Architectural Design				0.20	0.20	0.20	0.20
Technical Coordination and Administration							
Deployment, Operations and Support							
Pre-production Development							
Dan Nae							
Management, Planning and Architectural Design				0.25	0.25	0.25	0.25
Technical Coordination and Administration				0.30	0.30	0.30	0.30
Deployment, Operations and Support				0.30	0.30	0.30	0.30
Pre-production Development				0.15	0.15	0.15	0.15
Artur Barczyk							
Management, Planning and Architectural Design							
Technical Coordination and Administration				0.05	0.10	0.10	0.10
Deployment, Operations and Support				0.65	0.65	0.65	0.65
Pre-production Development				0.30	0.25	0.25	0.25
Yang Xia							
Management, Planning and Architectural Design							
Technical Coordination and Administration							
Deployment, Operations and Support				0.10			
Pre-production Development				0.10			
Tony Cheng							
Management, Planning and Architectural Design							
Technical Coordination and Administration							
Deployment, Operations and Support				0.75	0.65	0.65	0.65
Pre-production Development				0.25	0.35	0.35	0.35
Ramiro Voicu							
Management, Planning and Architectural Design							
Technical Coordination and Administration							
Deployment, Operations and Support				0.20			
Pre-production Development				0.30			
New Staff #1							
Management, Planning and Architectural Design							
Technical Coordination and Administration							
Deployment, Operations and Support					0.65	0.65	0.65
Pre-production Development					0.35	0.35	0.35
					0	**0**	**0**
Total FTE (Caltech)							
Management, Planning and Architectural Design			**0.40**	**0.45**	**0.45**	**0.45**	**0.45**
Technical Coordination and Administration			**0.30**	**0.35**	**0.40**	**0.40**	**0.40**
Deployment, Operations and Support			**1.60**	**2.00**	**2.25**	**2.25**	**2.25**
Pre-production Development			**0.90**	**1.10**	**1.10**	**1.10**	**1.10**
TOTAL FTE (Caltech)			**3.2**	**3.90**	**4.20**	**4.20**	**4.20**
TOTAL FTE funded by DOE (Caltech)			**3.00**	**3.50**	**4.00**	**4.00**	**4.00**

🟩 Non LHCNet funding

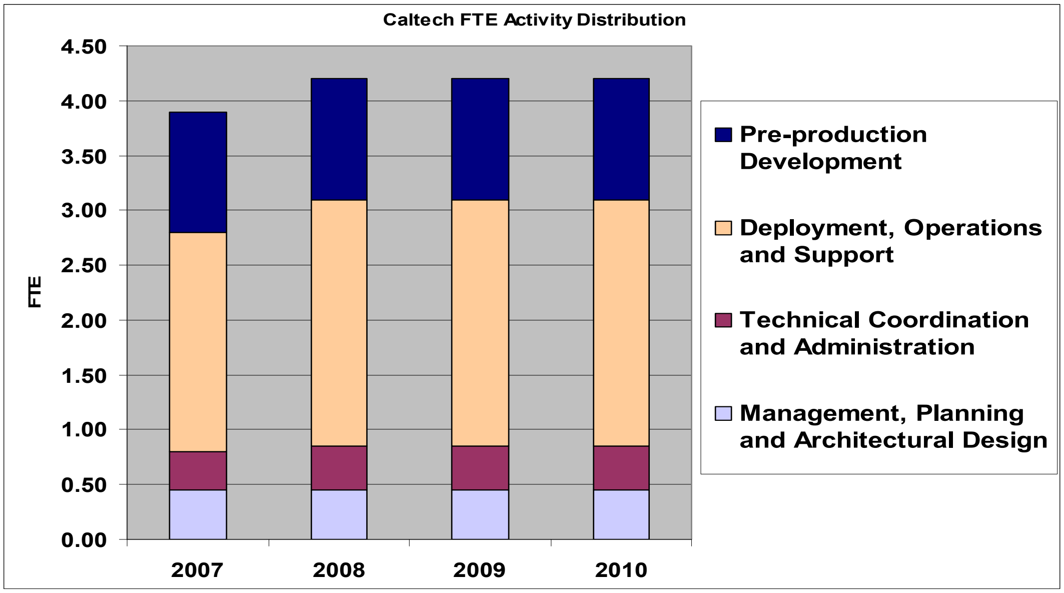


Figure 21: Actual (2007) and planned (2008-2010) Caltech FTE Activity Distribution

Table 2: Actual and planned contribution of CERN to the US LHCNet activities

CERN Staff Activity Distribution	2006	2007	2008	2009	2010
Management, Planning and Architectural Design	**0.10**	**0.05**	**0.05**	**0.05**	**0.05**
Technical Coordination and Administration	**0.10**	**0.05**	**0.05**	**0.05**	**0.05**
Deployment, Operations and Support	**0.10**	**0.10**	**0.15**	**0.15**	**0.15**
Pre-production Development	**0.05**	**0.05**	**0.05**	**0.05**	**0.05**
TOTAL FTE (CERN)	**0.35**	**0.25**	**0.30**	**0.30**	**0.30**

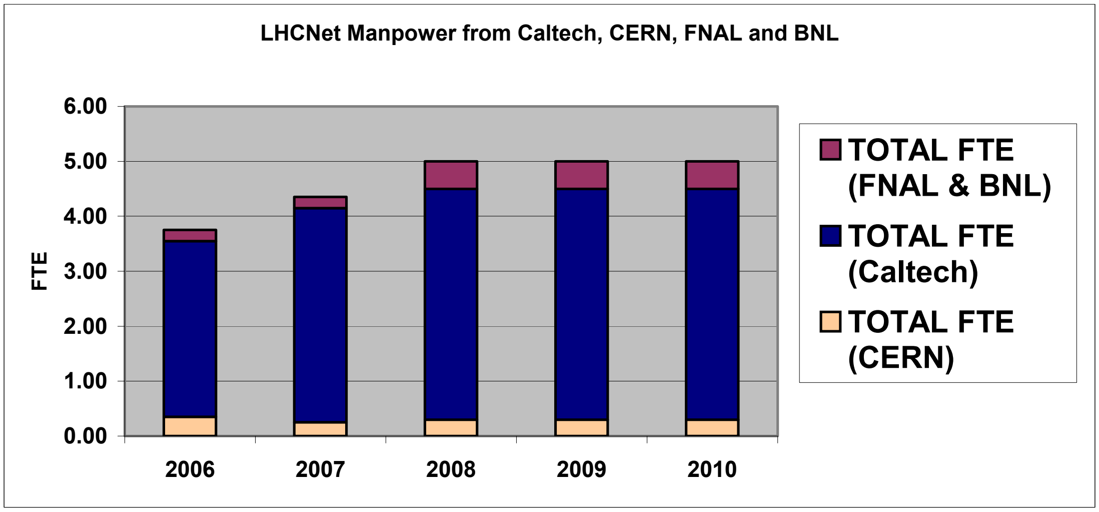


Figure 22: US LHCNet actual (2006-2007) and planned (2008-2010) manpower contributions from Caltech, CERN, FNAL and BNL

Annex D: The CERN Campus Network and the LHC OPN

As shown in Figure 23, the CERN campus network is split into four distinct areas.

- The **General Purpose Internet** area includes the commodity Internet access, the access to GEANT and the CERN Internet exchange point (CIXP) where Internet service providers (ISP) can interconnect.
- The **Security Zone** contains firewalls (a primary firewall and a backup). A High Throughput Access Route (HTAR) service that by-passes firewalls is available to high-bandwidth users.
- The **Campus Backbone** is the backbone used by users at CERN for their daily work. It connects all buildings to the general purpose services.
- The **LCG Backbone** is a very high capacity backbone that is going to connect the experiments' computing nodes, disk servers and tapes to WAN circuits to/from the Tier1 centers.

US LHCNet is directly attached to the CERN LCG backbone via a set of 10 Gbps connections for CMS and ATLAS traffic to/from Fermilab and BNL, as well as the U.S. Tier2 centers as needed. Traffic with CERN's General purpose services will transit via the security zone.

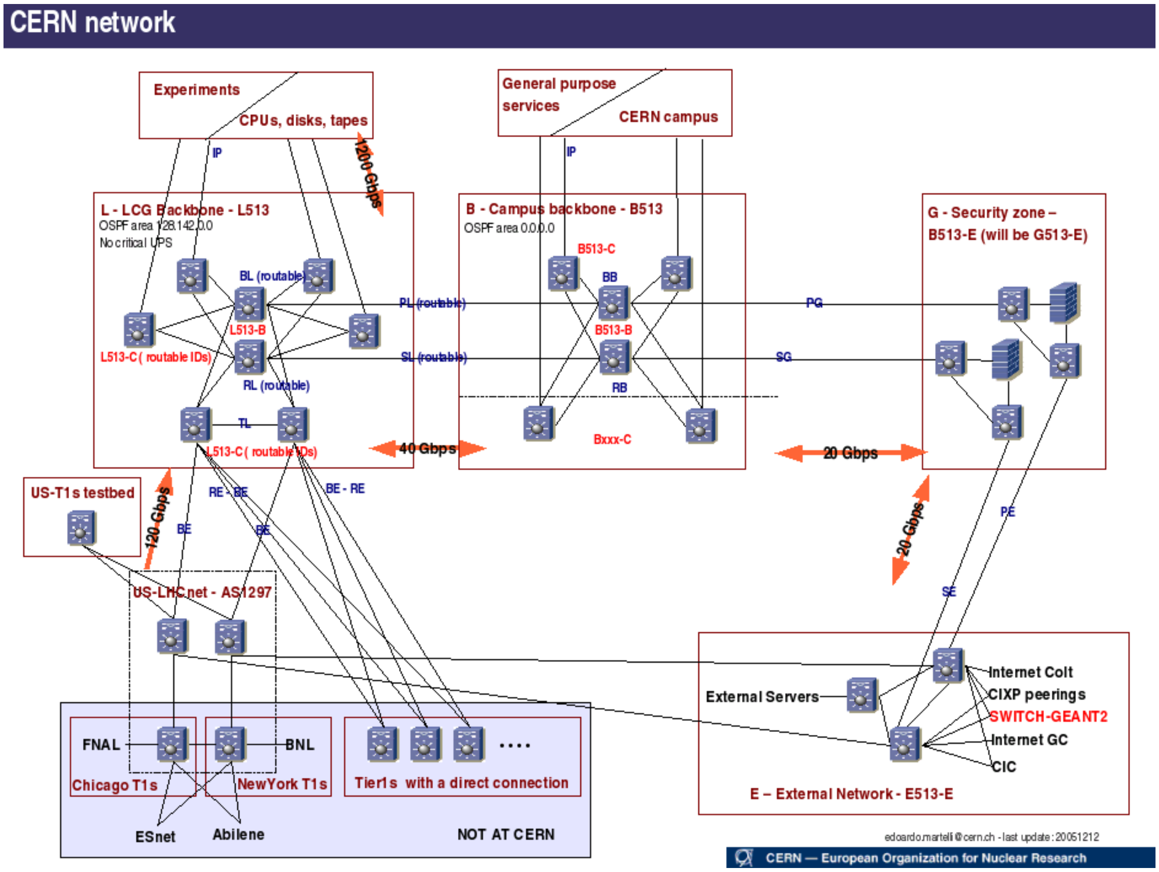


Figure 23 CERN Campus Network, showing the US LHCNet connections

The LHC OPN (**Figure 24**) is the sum of all links designed to carry Tier0 – Tier1 traffic. It is a star-shaped layer 1/2 network consisting of 10 Gbps links serving each Tier1 center plus additional links for redundancy. The US LHCNet network is the part of the LHC OPN serving the US Tier1s (BNL and FNAL).

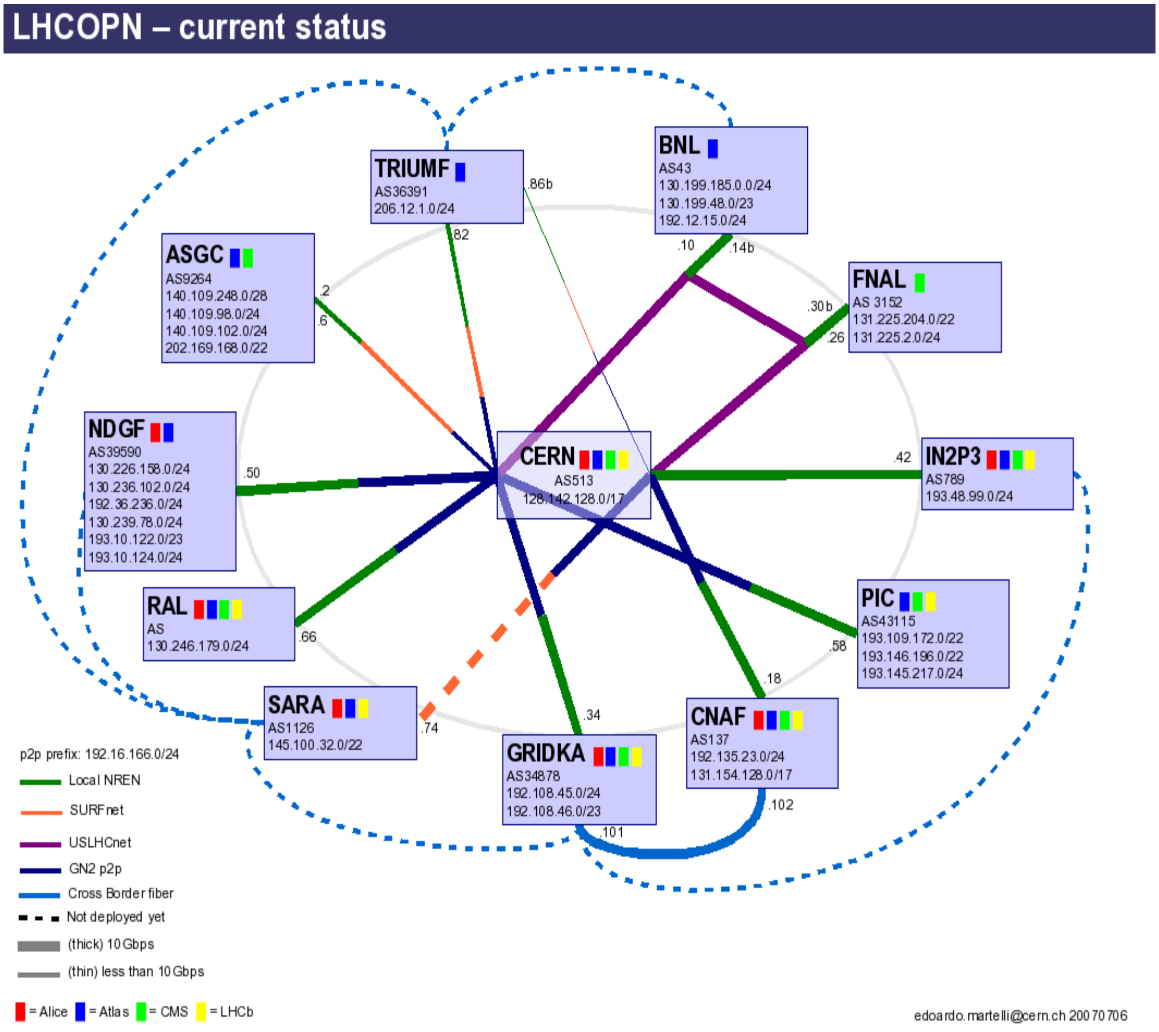


Figure 24: The LHC OPN

Annex E: ESNET and US LHCNet Bandwidth Planning for 2008-2010

As explained in Section 1.4 of the proposal, we will keep the LHCNet bandwidth and technology in line with the ESnet backbone and the DOE networking roadmap. We are working with ESnet and the DOE labs in the context of the U.S. LHC Network Working Group, to ensure that our network designs, operational modes and fallback strategies are mutually compatible and synergistic. In this Annex we summarize the new ESnet architecture and implementation strategy being developed.

The elements of the ESnet's architecture include a high reliability, national IP core and an independent, multi-lambda national core that provides circuit-like services – the Science Data Network (SDN). These two core networks will independently connect to Metropolitan Area Network (MAN) rings that connect to the ESnet sites (DOE Labs). The MAN rings are intended to provide redundant paths to the two core networks and to support the path management required for on-demand, high bandwidth point-to-point circuits. In particular, MAN rings will provide two independent physical paths from BNL and FNAL to respectively MANLAN and StarLight (i.e., to the LHCNet points of presence).

US LHCNet's connections to the planned ESnet4 infrastructure are shown in Figure 25 and Figure 26. In Chicago and New York, LHCNet will be connected to the MAN rings to directly access FNAL, BNL and the SDN core. A 10 Gbps connection to the national IP core will be maintained for general purpose Internet traffic between CERN and the DOE laboratories.

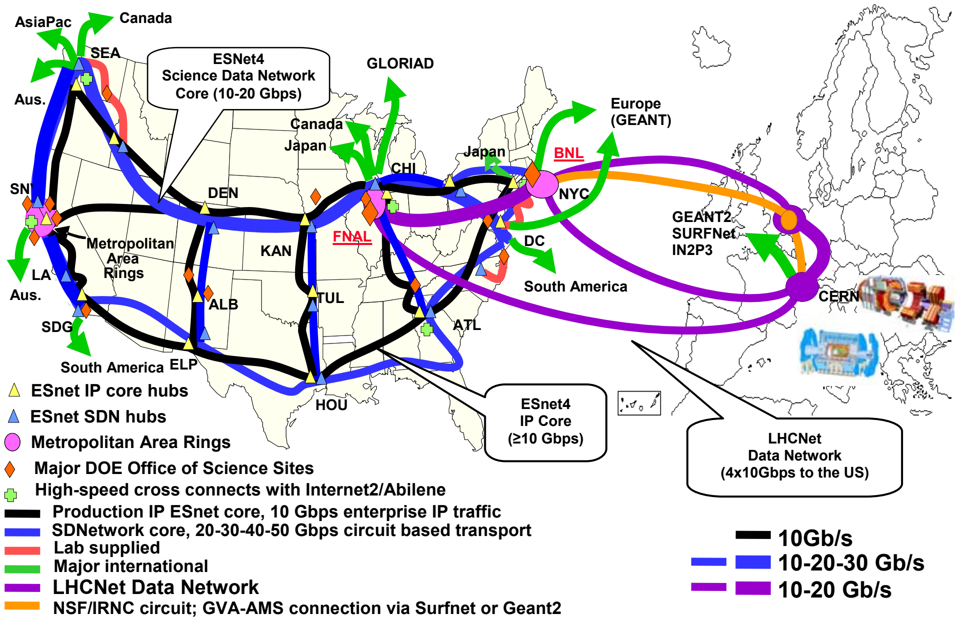


Figure 25: ESnet, LHCNet and the NSF/IRNC links – as planned for 2008

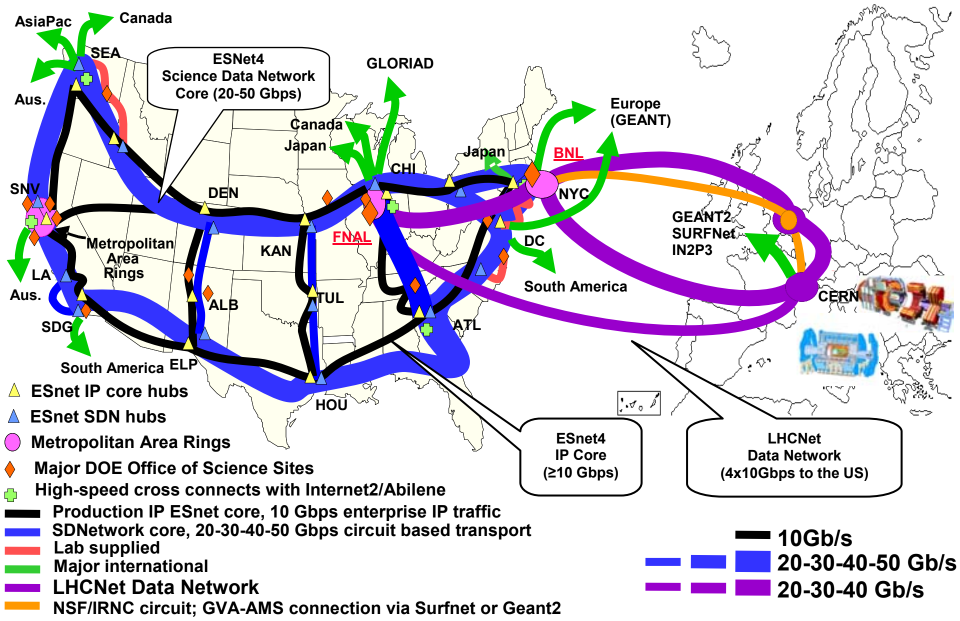


Figure 26: ESnet, LHCNet and the NSF/IRNC links – as planned for 2009/2010

The roadmap for deploying the new backbone will provide the required connectivity and redundancy for the US Tier1 just in time for the LHC startup next year, and will increase to accommodate future needs.

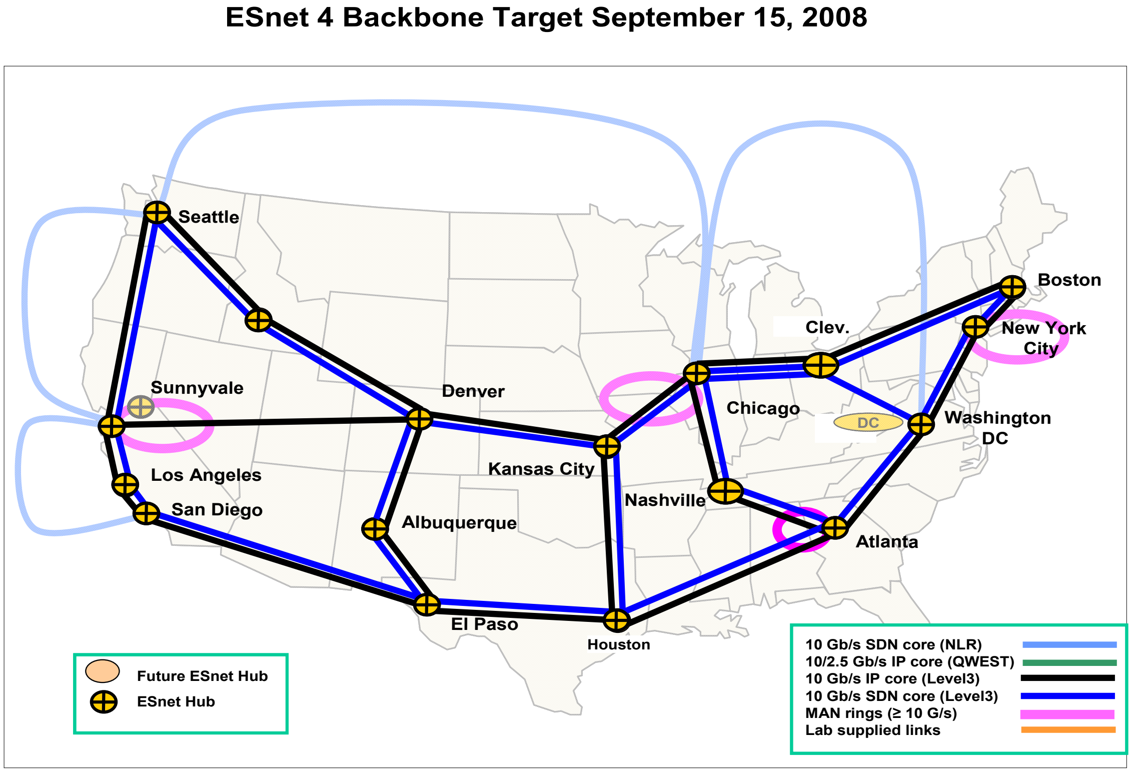
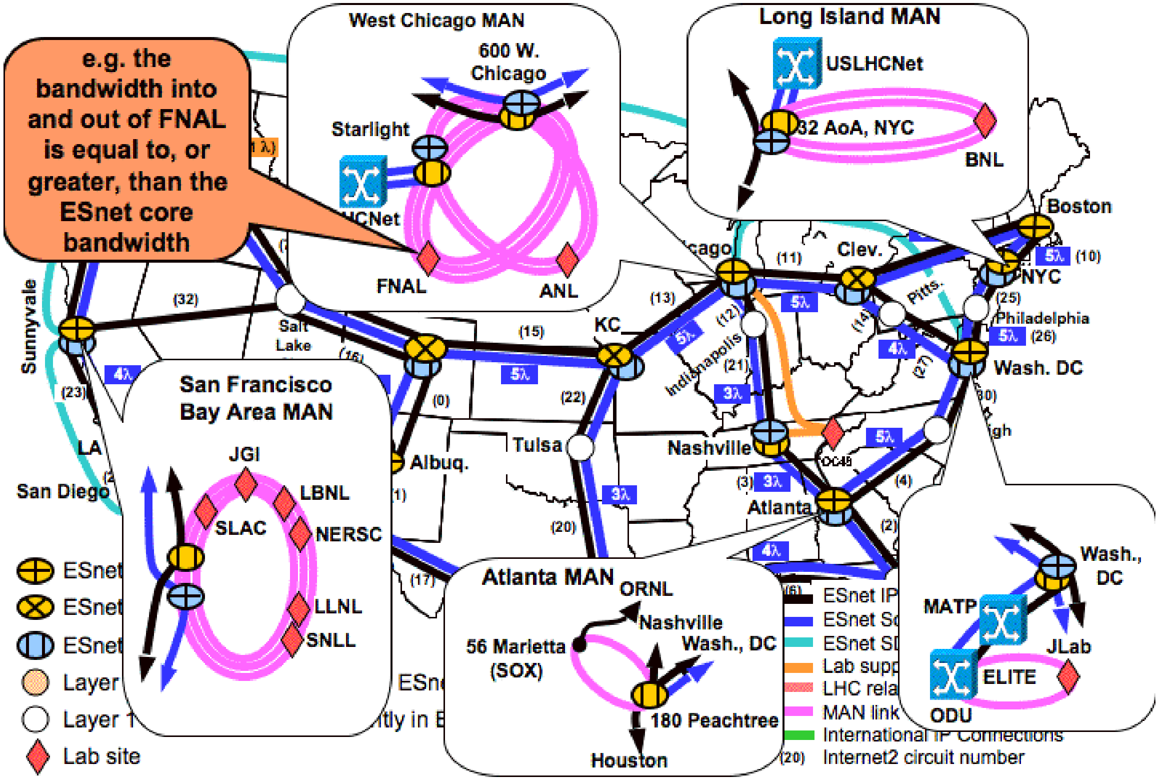


Figure 27 ESnet4 backbone as foreseen for September 15, 2008 (shortly after the LHC starts).

In order to provide a reliable service to the US Tier1s, ESnet has been deploying metropolitan area networks in the Chicago (CHIMAN) and New York (LIMAN). These MANs (see Figure 28) provide ESnet with a resilient infrastructure which allow the delivery of uninterrupted service in case of a circuit, fiber or equipment failure, and ensure that the end-to-end path from the CERN Tier0 to the US Tier1 is fully redundant.

High Bandwidth all the Way to the End Sites – major ESnet sites are now effectively directly on the ESnet “core” network



e.g. the bandwidth into and out of FNAL is equal to, or greater, than the ESnet core bandwidth

Figure 28 ESnet map showing the metropolitan area networks

The scope of the planned ESnet4 SDN upgrades two to five years in the future (shown in Figure 29) is set by the requirements of the LHC experiments and the rest of the HEP community, along with the needs of the major DOE-funded science projects in many other fields. As shown in the and also with the development plans of US LHCNet. It should be noted though that for 2011-2012, the steep bandwidth increase implies a change of technology for continental DWDM systems, which may be delayed in the case of transatlantic circuits, due to technical limitations and the high costs of current submarine cables.

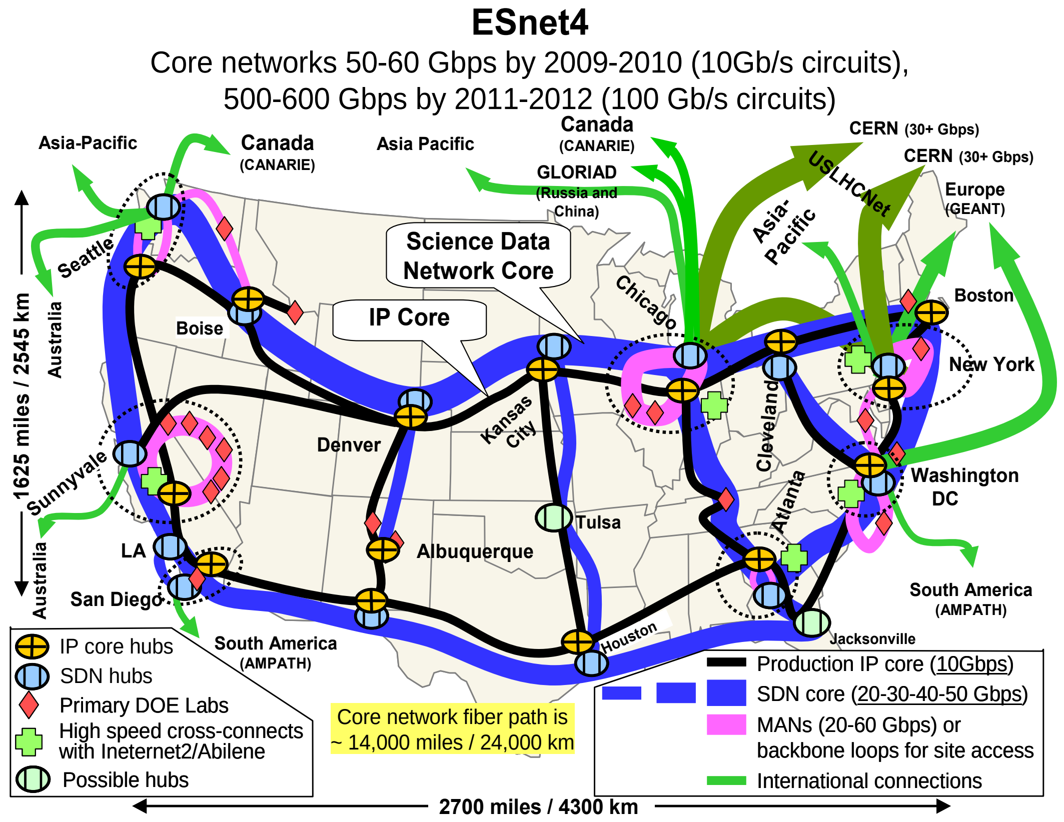
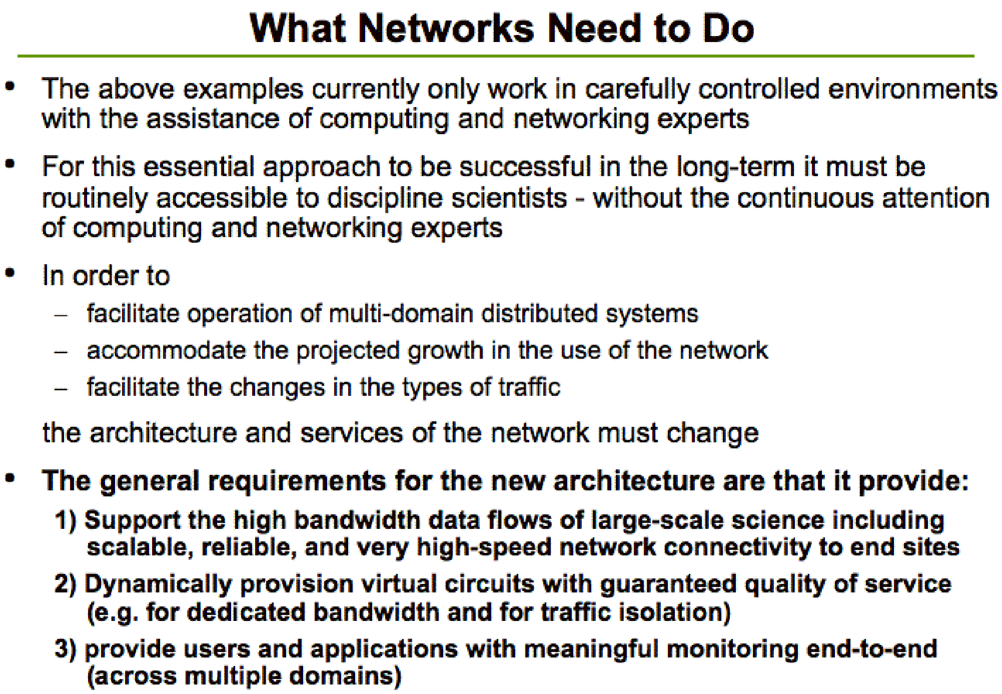


Figure 29 Planned ESnet4 network growth for 2009/2010 (50-60Gbps) and 2011-2012 (500-600Gbps)

Finally, it should be noted (as illustrated in Figure 30[7]) that the main points of the US LHCNet vision and the ESnet vision of what networks should do are essentially identical:

- support the high bandwidth data flows of large-scale science including scalable, reliable and very high-speed network connectivity to end sites
- dynamically provision virtual circuits with guaranteed quality of service (e.g. for dedicated bandwidth and for traffic isolation)
- provide users and applications with meaningful monitoring end-to-end (across multiple domains)

[7] This figure is taken from W. Johnston's presentation at the Joint Techs workshop at Fermilab, July 2007



What Networks Need to Do

- The above examples currently only work in carefully controlled environments with the assistance of computing and networking experts
- For this essential approach to be successful in the long-term it must be routinely accessible to discipline scientists - without the continuous attention of computing and networking experts
- In order to
  - facilitate operation of multi-domain distributed systems
  - accommodate the projected growth in the use of the network
  - facilitate the changes in the types of traffic

  the architecture and services of the network must change
- **The general requirements for the new architecture are that it provide:**

  1) Support the high bandwidth data flows of large-scale science including scalable, reliable, and very high-speed network connectivity to end sites

  2) Dynamically provision virtual circuits with guaranteed quality of service (e.g. for dedicated bandwidth and for traffic isolation)

  3) provide users and applications with meaningful monitoring end-to-end (across multiple domains)

Figure 30 ESnet’s vision for the future of networks for science (courtesy of ESnet)

Annex F: Network Services Infrastructure for managed high performance end-to-end data transfers

High performance networks form a critical resource in the computing model of each of the LHC experiments. The experiments have to distribute their data for analysis to various computing centers spread throughout the globe, and in doing so rely on the reliability of the network connections. Data transfer applications developed for this purpose are mostly network-unaware, focusing mainly on the storage system's performance, while assuming that the network is not a bottleneck. While this is true for today's high-performance networks, the situation is likely to change once the first data from the LHC becomes available.

In addition to the increased data volumes and data transfer rates foreseen by the LHC experiments once the LHC is in operation, the trend towards scarce network resources is driven by the high throughput data transfer methods currently employed at the US Tier1 and many US Tier2 sites, as well as recently-developed higher throughput transfer methods that are now being integrated with widely-used storage systems such as dCache. Once widely deployed and properly tuned (also outside the U.S., a process that is now underway in Europe, Brazil and Taiwan), the current and next-generation tools will provide a substantially greater data transfer capability than the foreseen transatlantic network bandwidth can support.

This leads to the clear result that wide area network bandwidth will need to be carefully allocated and managed. The methods presented in this Annex are designed to make this possible, and to ensure that the LHC experiments are able to make efficient use of the available network resources, up to high network occupancy levels.

In order to meet the need for high-performance data transfers for the LHC community, and the management of relatively scarce network resources on an increasing scale, starting in the near future, we are currently developing a system of network services which can be used by the data movement applications to set up a desired connection between end-hosts and monitor the progress of the data transfer. The system takes both the end-hosts (storage nodes) as well as network elements as resources into account, in order to provide means for truly robust data movement among the data centers. This global view will allow optimizing end-to-end transfer performance and facilitate problem tracking and resolution.

The basis for providing robust network services in our system is the ability to

(1) provide bandwidth guarantees to the application by means of provisioning circuits,

(2) schedule transfer requests based on their relative priority, and

(3) monitor all components of the system end-to-end.

These three features, together with the knowledge of the capabilities and state of the end-hosts, make it possible to determine a good (time-dependent) estimate of the time-to-completion for each transfer as it progresses, and thus to make truly managed data transfers possible. Later developments will include "strategic" elements that take policy and quotas assigned to transfers associated with different activities within the experiments' Computing Model into account.

Circuit Oriented Services

The new US LHCNet wide area infrastructure is built on SONET[8] technology. We use CIENA Core Director/CI multiservice switches, interconnected through OC-192 links. SONET is a circuit technology in that a logical connection (circuit) has to be established between two end-points before data can flow. The circuit can span multiple physical point-to-point connections. The bandwidth of a circuit is defined by the number of time slots it has been allocated. Virtual Concatenation (VCAT) permits a circuit to consist of any number of STS-1 frames without restrictions, i.e. a circuit can be allocated a bandwidth of N×51Mbps, with $1 \leq N \leq 192$. Using the Link Capacity Adjustment Scheme (LCAS), the bandwidth of a circuit can be subsequently modified, again in steps of 51Mbps.

On the side facing the local area networks at each site we use 10 gigabit Ethernet (10GbE) interfaces, which allows us to connect to relatively inexpensive Ethernet routers and switches. The CIENA CD/CI platform can either map an entire 10GbE interface onto a circuit in so-called tunnel mode, or map (a range of) VLAN(s) onto the circuit. The allocated bandwidth is defined by the circuit size.

With the installation of the CIENA CD/CI multiservice switches, US LHCNet has thus acquired the capability to construct circuits across the domain in a flexible way, and so to provide circuit-oriented services to the users. By reservation of a circuit to a data flow between two end-hosts, we can guarantee the bandwidth to be fully available for the transfer, something that routed IP networks can only approximate by means of QoS parameter tuning. Dynamic circuit setup, using control-plane software described below, gives us the possibility to engineer explicit paths through our domain, and this allows us in turn to match the requirements of the users in real-time.

By using the User/Application Interface, circuit provisioning does not require operator intervention, but can proceed in an automated way.

Managed Network Services

An overview picture of the system is shown in Figure 31, the top part of the picture shows the components, and main interactions between them, the bottom part indicates the underlying physical layer. The components are explained in the following paragraphs. All components of the system will be implemented using the MonALISA framework of distributed services and agents, which will avoid any single point of failure in the system. The communication between the components is implemented using MonALISA's proxy layer for secure interconnect of services and agents (See Annex G), as shown in Figure 32. The secure channels are composed of redundant set of 2-3 sockets for added reliability.

[8] See for example http://www.cisco.com/warp/public/127/sonet_tech_tips.html

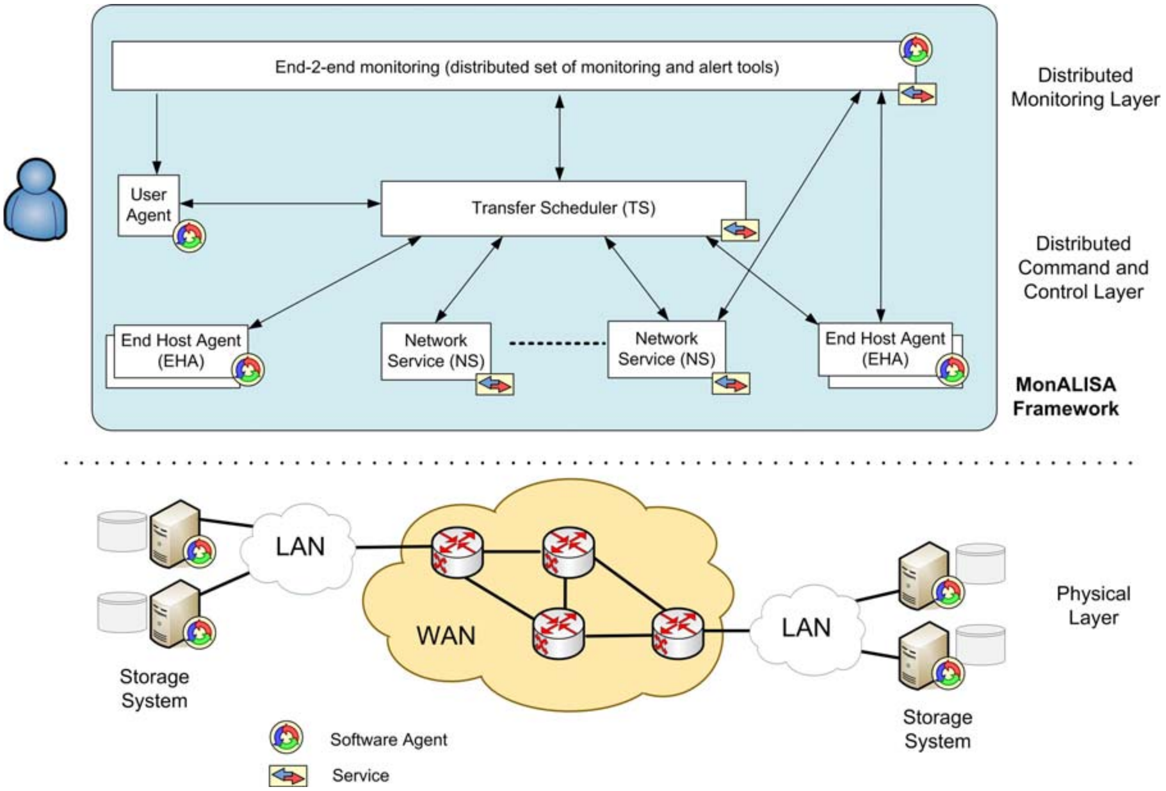


Figure 31: Overview of the proposed Network Services showing the main components (top) and the underlying physical layer (bottom).

The basic elements of the network services system are

1. **Transfer Classes and Priorities**,
2. **Network Services (NS)**, for monitoring and configuration of the network elements
3. **End-Host Agent (EHA)**, for monitoring and configuration of the end-host systems
4. **Transfer Scheduler (TS)**, dealing with requests for circuit setup
5. **User Agent (UA)**, for operator and/or application access to the services.
6. **Global Monitoring** of the system

These components are described in the following sections.

Transfer Classes, Priorities and Preemption

There are four transfer classes (reserved, normal, scavenger, and general traffic), and three priorities (high, normal and low) defined in the system.

- Reserved transfers have the starting time allocated (reserved) by the scheduler at request time. This class is reserved for mission-critical transfers, such as production traffic from Tier 0, or for data set synchronization between Tier 1's.

Once allocated, the starting time cannot be modified, apart from transfer cancellation.

- Normal transfers have only an estimated starting time. The requests are stored in priority queues, the queues being processed at invocation time of the scheduler, when the highest priority, run-able (i.e. end-to-end circuit can be established) transfer is selected for execution.
- Scavenger class is meant primarily for transfers which are not finished when their allocated time window has expired. They are guaranteed only the minimal bandwidth (one STS-1 channel, i.e. 51 Mbps), but are allowed to use the unallocated bandwidth. In this way we avoid that a long transfer which for any technical reason could not finish on time is aborted and would need retransmission of the complete data. While not encouraged, a user can also request a transfer to fall into this class.
- General traffic class is intended to provide non-circuit oriented IP service. It will be used for transfers between end-systems which are not using the Network Services described here, for transfers deemed too short to profit from the queuing system (to be defined based on experience), and for unmanaged traffic, like web access etc. This class will be implemented as a set of permanent circuits, their bandwidth will be adjusted according to availability, with a guaranteed minimum, to be defined based on experience.

All transfers can be preempted by a higher priority or class transfer, with the exception of high priority fixed time transfers (highest possible class/priority). Preemption in our context means that the allocated bandwidth is reduced to make space for another transfer. The circuit is not cut; the transfer still continues at a lower rate. TCP connections are not closed. When a transfer is preempted, the scheduler notifies the End-Host Agent of the bandwidth reduction, which is in particular important for transfers using UDP as transport protocol, in order to avoid wasting CPU resources.

At any time, a user can query for the estimated start time, which is obtained by running the scheduler algorithm in simulation mode, or directly from the allocated start time in case of the reserved transfer class.

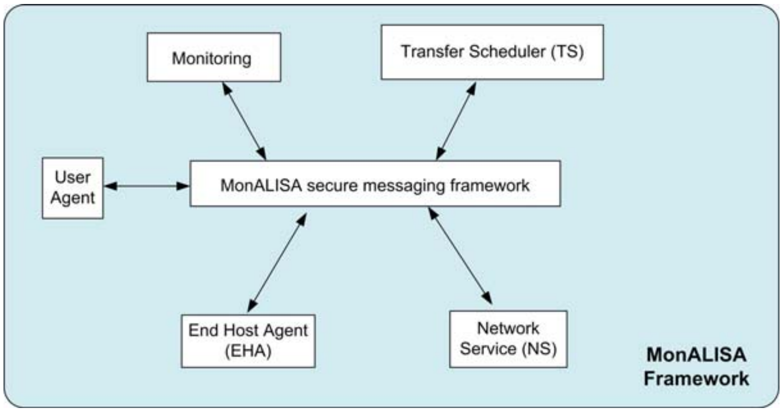


Figure 32: Message exchange between all components pass through a uniform framework of MonALISA services.

Network Services

The Network Services (NS) are responsible for setting up the required path when authorized to do so by the scheduler.

The configuration of the network devices (CIENA CD/CI) is done using the TL1 engine developed for the VINCI project, which has been proven to work reliably. The functionality will include setting up and tearing down of circuits, adjusting allocated bandwidth and configuring protection mechanisms for the circuits.

An important part of the Network Services will be the path discovery service. Once a request for bandwidth allocation has been received and authorized, a physical path has to be found which matches the request criteria. VINCI's topology services[9] will be used to obtain a list of possible paths. A path is described as a tuple {N, L}, with N being a vector of nodes, and L a vector of links interconnecting the nodes. The available bandwidth on each link has to be determined. Depending on the transfer class and request priority, the available bandwidth can be the fraction of link capacity not allocated to any circuit, the bandwidth available after squeezing lower priority circuits, or, in case of general traffic, simply a part of a circuit shared with other transfers.

For construction of end-to-end virtual circuits, the path discovery service will interface to domain controllers of the other circuit oriented networks, as described in a following section of this Annex. Depending on the policies of the domains, the path information can range from fully detailed (peer model of the network) to fully abstracted (overlay model). In the latter case, the only information obtained is the entry and exit points of the domain. In any case, the information obtained will be sufficient to find a network path from the source to the destination end-system.

The path discovery service will provide the scheduler with the best suited path(s) between source and destination, together with information on bandwidth, earliest reservation time and maximum duration for which the bandwidth can be guaranteed.

End-Host Agent

The End-Host Agent (EHA) performs several functions. It is able to control applications where needed (such as FDT for fast data transfer, for example), to monitor the end-host state and take mitigating actions in case of problems, and to report all the relevant information to the Network Services, which can take appropriate action when certain trigger conditions (such as too high a packet loss rate) are detected.

The EHA also can adjust the configuration of the end-hosts, where possible, using LISA[10] agents. The configuration parameters are typically the IP address of the data interface, routing table entries, and VLAN membership. It should be highlighted that while the monitoring of all the devices including end-hosts is crucial for providing quick and, where possible, transparent problem resolution, the end-host configuration is optional. We are fully aware that this means privileged access to the end-host (root access), which might be against the computing rules of some sites. In such cases we will

[9] Described in Annex H.

[10] Described in Annex G.

work together with the local administrators to provide an alternative and satisfactory solution.

Transfer Scheduler (TS)

The following events form invocation points of the Transfer Scheduler: *transfer request, timed start of transfer, end-of-transfer*, *transfer status change* (e.g. disk problem), *transfer cancellation*, and *change of topology* (e.g. link down).

The scheduler interacts with several other component services in the course of its operation:

- the AAA services for authentication, authorization and accounting
- the path discovery services (part of the MonALISA services in the figure) to find the best suitable path/bandwidth/start time and duration of the transfer
- the End-Host Agent to start a transfer, to notify the end-hosts of bandwidth allocation change and get notification of completed transfer, and
- the monitoring services to recognize and act upon problems in the network or end-host systems.

A User Agent acting on behalf of a user that needs a data transfer will request a circuit between the source and destination end-hosts. The request goes to the scheduler front-end, specifying the IP addresses, protocols, port numbers, total amount of data to be transferred, desired bandwidth and transfer priority. Upon verifying the requester's credentials, the scheduler checks the reachability and capabilities of both end-hosts involved in the transfer. The scheduler estimates the bandwidth that can be allocated to the transfer, using the information it has received from the End Host Agents on the end-host capabilities and configuration, and from the Network Services on the network configuration and status, and it then queues the request. By taking into account the topology, the scheduler can either take the decision as to the best route to take, or present options to the user, specifying the possible routes and estimated starting time and bandwidth available on each route[11]. In the latter case, the user can decide which option he/she prefers, or take the default (i.e., the scheduler's decision).

System Messages

Table 1 shows the main messages exchanged between the components of the system, and the information passed.

[11] These decisions may at some stage also include consideration of the relative cost for each choice. This will apply if quotas for the use of network resources have to be implemented to handle the demand.

From	To	Message	Parameters	Return values
UA	TS	Transfer Request (TREQ)	• Source IP address • source port • destination IP address • destination port • protocol • Dataset Identifier (DID) • transfer class • transfer priority • desired start time • desired bandwidth	• Transfer ID • Start Time (estimated or fixed) • Allocated Bandwidth • Transfer class • Transfer priority
		Transfer Cancellation (TC)	• Transfer Identifier (TID)	• Bytes transferred
		Request Modification (RMOD)	• TID • desired class • desired priority • Desired start time • Desired bandwidth	• Start Time (estimated or fixed) • Allocated bandwidth • Transfer class • Transfer priority
EHA	TS	End of Transfer (EOT)	• TID • Bytes transferred	
		Transfer Status Change (TSC)	• TID • Possible throughput • *Other parameters will be implemented based on future experience*	
TS	NS	Configure	• List of nodes and link identifiers • Bandwidth	• Acknowledge, when circuit is provisioned and ready to transmit data
NS	TS	Change of Topology (COT)	• List of link identifiers	
TS	EHA	Verify Transfer	• TID • DID • Desired start time • Desired bandwidth	• TID • DID • Possible start time • Possible bandwidth
		Register Transfer	• TID • DID • Start time (estimated or fixed) • Allocated bandwidth	• Acknowledge
		Start Transfer	• TID	
		Cancel Transfer	• TID	
		Change of Bandwidth	• TID • Current Bandwidth	

Table 3: System messages used for the setup and running of a transfer.

Other messages not listed in Table 1 are monitoring messages, including path discovery, and messages related to AAA functions.

Global Monitoring

We are using existing MonALISA services for end-to-end monitoring[12]. Routers and switches are monitored by means of the SNMP protocol, while the CIENA Core Directors are again accessed using the TL1 interface described above. LISA agents are used for end-host monitoring and configuration. We gather statistics on the CPU and memory utilization, I/O load, disk usage and network interface utilization.

Problem scenarios can be as various as simple misconfiguration of a network device or end-host, software crash, loss of optical signal quality (increase in Bit Error Rate), and the many different hardware failure modes. In many cases it is very hard to track down the real root of the problem, all too often due to the missing global view of the system. As is done now in other applications using MonALISA, our monitoring system will monitor all relevant and available parameters, and by means of detecting deviations from expected values and correlations between such events, will allow us to pinpoint the source of an observed problem.

Through the MonALISA services working with the End Host Agents, the system will be able to take independent actions in simple and well-known situations. For example a failed link can lead to automatic re-routing of the traffic without operator intervention. While this is an operation any IP router does perform nowadays, here the re-routing can be accompanied by a reduction in the bandwidth allocated to the transfer, followed by a notification (sent from the monitoring system) to the End Host Agents of the reduced bandwidth, which in turn can take appropriate action.

In the extreme case where the circuit cannot be re-routed, for example due to no available bandwidth at the requested (typically lower) priority on a secondary link, the End Host Agent can abort the impaired transfer, resuming it when the circuit becomes available again. In the EHA might proceed with another transfer request for a destination not affected by the outage.

As part of the managed transfer capability, if a higher priority transfer requests a segment on the same path (preemption) as a transfer already in progress, then the EHAs can temporarily reduce bandwidth allocated to the transfer in progress. This could happen in a scheduled way if a higher priority request arrives during the transfer, or because of a problem on a different link that causes some higher-priority traffic to be re-routed.

User Agent (Operator/Application Interface)

Users and applications that want/need to interact with the Network Services do so through the User Agent (UA). The UA takes the necessary parameters from the user or application, and sends the transfer request to the TS. It receives an acknowledge message carrying information described above, and continues monitoring the transfer progress until finished.

[12] See Annexes G and H.

There are two types of User Agents provided. The main UA is a web applet, providing a uniform GUI to the whole system. The second UA is a command line tool, which can be used more easily in scripts and for easy integration of existing data management tools.

The user can use MonALISA at all times for monitoring of transfer status, including the Estimated Time to Completion (ETC), network status (link utilization, alarm conditions, topology and route taken), end-host status (CPU/memory/network utilization), etc.

Operational Scenarios

In this section we present some operational scenarios, used during the design phase of the project. These scenarios do not show the complete architecture of the system, and as such there is no claim of completeness. In particular we refer to the MonALISA framework for distributed agents and services, without making it explicit in the scenarios, in the interests of keeping the discussion relatively simple. However, following our established practice, all services will be implemented either in a distributed or redundant way, so as to avoid any single point of failure.

Scenario 1: Transfer request and circuit setup

The following procedure, depicted in Figure 33, is applied when an application or user wants to transfer a data set:

1. The application (or user) sends a transfer request to the transfer scheduler (TS). The TS is the “entry point” in using the network services. The application requests network resources from the TS, specifying the destination IP address, protocol to be used and the port numbers as well as the total amount of data to be transferred.
   Upon verifying the requester credentials through authentication and authorization (AA) services, the TS checks the reachability and capabilities of both end-hosts, through the end host agents (EHAs).
2. The scheduler estimates from the request and the information on the network configuration and its status obtained from the Network Services, the possible bandwidth. Part of this (global) system view is retrieved through the distributed monitoring system. The TS queues the request and responds to the requestor with an acknowledge message that contains an estimated waiting time (EWT) and bandwidth forecast. The queued request contains the following information: entry point, exit point, transfer size, transfer priority, desired bandwidth.
3. Once the network (and storage) resources are available for the request, the Network Services (NS) are used to configure the circuit for the transfer. NS are responsible for setting up the requested path when authorized to do so by the TS. The NS also provide information on topology, utilization and error status to the monitoring agent. A change of topology is propagated to the scheduler through the MonALISA monitoring system.
4. When the path setup is complete, the TS notifies the EHA by issuing a *start-transfer* message, that the transfer can start. The EHA monitors the progress,

checking that the requested bandwidth can indeed be maintained. If this is not the case, it takes corrective measures if possible, e.g. by adjusting the end-host configuration or (as in the case of FDT) by adjusting parameters in the data transfer application. If the problem cannot be resolved, it notifies the scheduler through the monitoring system.

5. During the transfer the user can utilize MonALISA for monitoring the transfer status, including the Estimated Time to Completion (ETC), network status (link utilization, alarm conditions, topology and route taken), end-host status (CPU/memory/network utilization).

6. Once the transfer is finished the TS receives an *end-of-transfer* message which is propagated to the application or user. Upon reception of an *end-of-transfer* message, the scheduler calculates the new network configuration for the next pending transfer, and sends the new configuration to the NSs.

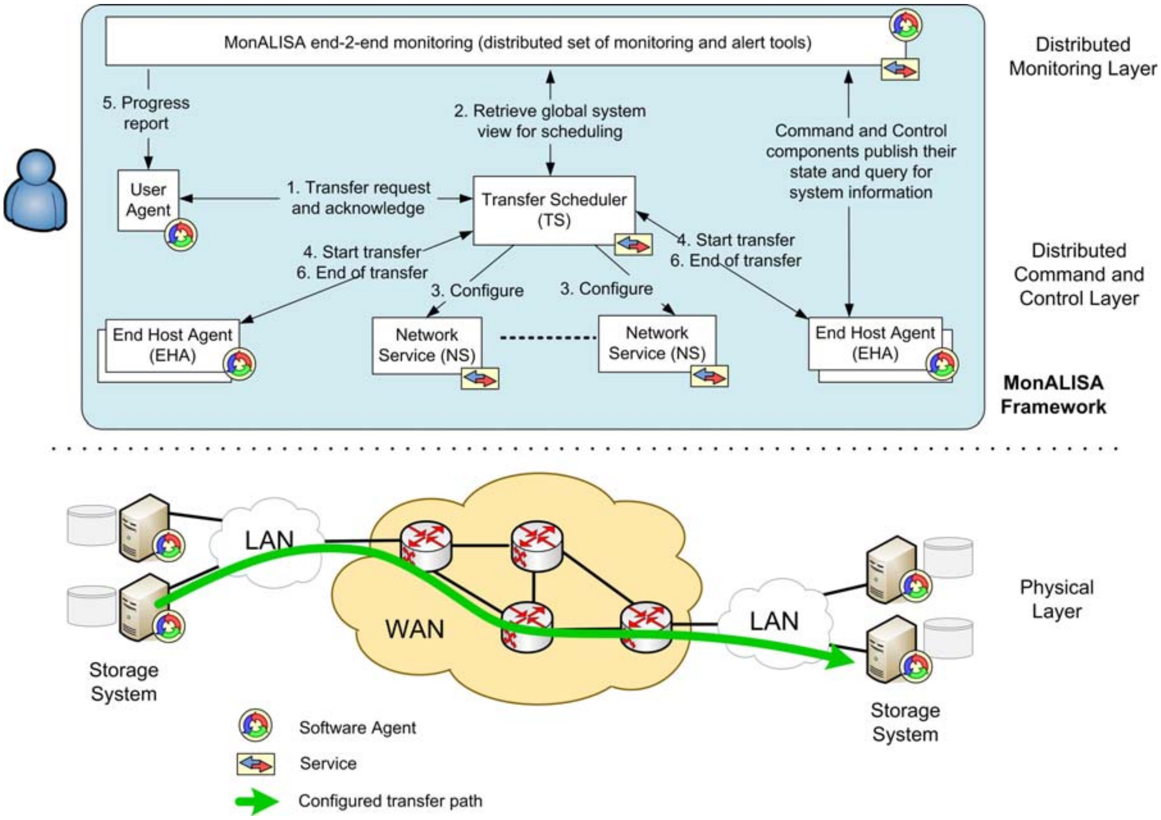


Figure 33: Circuit setup for a data transfer.

Scenario 2: Automatic transfer re-direction

In case of an outage on one of the links, the system is capable of taking corrective measures. Assuming a backup path exists, the system proceeds in the following way (illustrated in Figure 34):

1. When the Network Service detects a link failure, a *change-of-topology* event is sent to the MonALISA monitoring service, which propagates the information to the scheduler.
2. The scheduler re-examines the new topology, and decides on a secondary path through the network, taking into account the priorities of the transfers on the failed circuit, as well as of those on all alternate paths.
3. The network elements are reconfigured to establish a new connection.

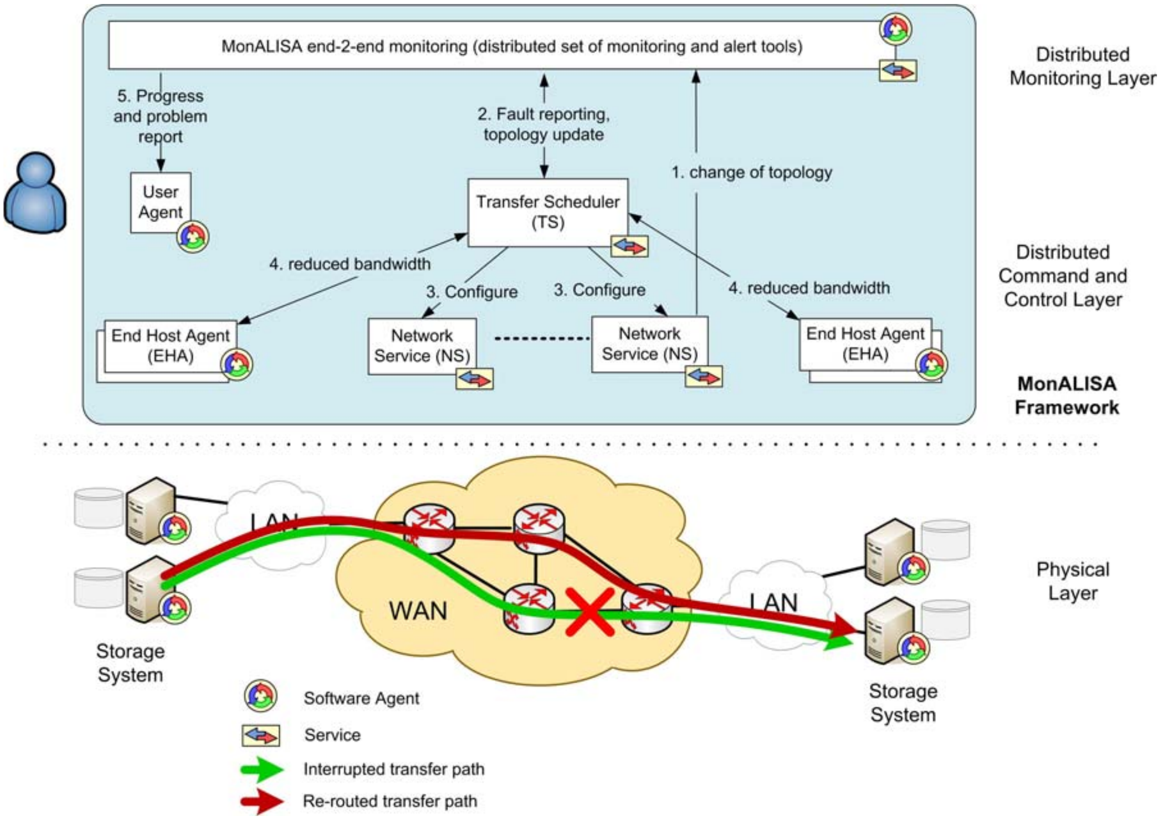


Figure 34: Automatic recovery from link failure.

4. If the new circuit provides less bandwidth, the End-Host Agents are notified. This is in particular useful when UDP is used as transport protocol. Without knowing the allocated bandwidth, the application will push data as fast as it can, wasting CPU resources while packets are dropped in the network.
5. At all times, the operator has a global view of the events, he observes the reduced bandwidth, and other real-time information provided by the monitoring system, and as a result knows that this is due to a change in the network.

While automatic route changes are present in routed IP networks, and different protection features exist in SONET networks to deal with link failure scenarios, it should be stressed that the system presented here takes the transfer priority as requested by the end-system into account. Circuits that have already been provisioned on the backup path might be squeezed (using LCAS for example) in order to accommodate the higher priority transfers following the re-routing, and in the case where more than one alternate path exists, the priorities as well as the bandwidth allocated to all transfers in progress along these paths will be taken into account in determining the "best" solution.

Scenario 3: Preemption

All transfers are marked with a class and a priority. As described above, a higher class and/or priority transfer can preempt a lower priority one. This is done in the US LHCNet domain by using the LCAS scheme, i.e. by adjusting the bandwidth of a circuit. In this way, the preempted transfer is not aborted, but can continue at a controlled lower rate, thus ensuring more resilient operation. The chain of events in case of preemption is

1. A high priority transfer request is placed by the User Agent to the Transfer Scheduler. Authorization and Authentication is checked by the TS.

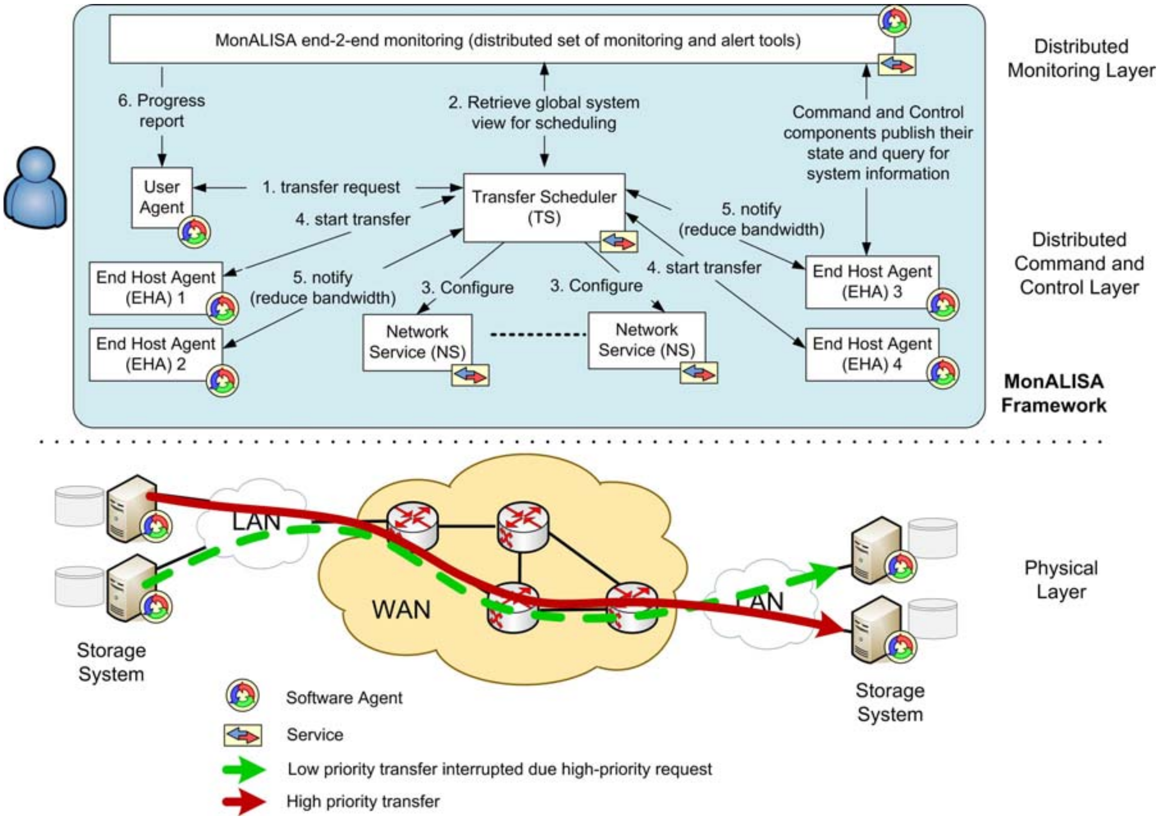

MonALISA end-2-end monitoring (distributed set of monitoring and alert tools)
Distributed Monitoring Layer
6. Progress report
2. Retrieve global system view for scheduling
Command and Control components publish their state and query for system information
User Agent
1. transfer request
4. start transfer
Transfer Scheduler (TS)
5. notify (reduce bandwidth)
Distributed Command and Control Layer
End Host Agent (EHA) 1
5. notify (reduce bandwidth)
3. Configure
3. Configure
4. start transfer
End Host Agent (EHA) 3
End Host Agent (EHA) 2
Network Service (NS)
Network Service (NS)
End Host Agent (EHA) 4
MonALISA Framework
LAN
WAN
LAN
Physical Layer
Storage System
Storage System
Software Agent
Service
Low priority transfer interrupted due high-priority request
High priority transfer

Figure 35: Example of transfer preemption.

2. The scheduler consults the topology information, and finds that there is no path available with the requested bandwidth, since all the bandwidth on these paths has already been allocated. But by squeezing the bandwidth allocated to a set of lower priority transfers on a given segment, enough bandwidth can be made available for the new request.
3. When the start time of the new transfer arrives, the TS sends a (set of) *configure* message(s) to the NS, to change the bandwidth of the running transfers, and provision the new circuit.
4. TS sends a start of transfer message to the EHA.
5. The TS notifies all concerned EHAs of the change in available bandwidth.
6. At all times the user or application is kept up-to-date on the progress though the User Agent.
7. Once the transfer is finished the TS receives an *end-of-transfer* message which is propagated to the application. Upon reception of an *end-of-transfer* message, the scheduler calculates the new network configuration for the next pending transfer, as well as to restore the originally requested bandwidth to the preempted transfers, and sends the new configuration to the NSs. *(Not in picture[13])*
8. TS notifies the concerned EHAs of the restored bandwidth. *(Not in picture)*

Scenario 4: End-host performance problem

1. One possible scenario demonstrating the advantages of end-to-end monitoring, including the end-hosts, is shown in Figure 36.
   An End-Host Agent monitoring the receiving system detects an error condition, e.g. disk performance problem. The receiver is not able to continue the transfer at the full rate. It notifies the Transfer Scheduler and the MonALISA monitoring service.
2. The scheduler notifies the sender of the problem condition, the transfer is put on hold.
3. The scheduler performs a look-up of the next transfer sharing the same path.
4. The problem path is reduced in bandwidth to a sustainable rate, and circuit for the new transfer is provisioned.
5. The TS sends a start transfer command to the EHAs.

[13] We omit pictorial representation of some of the steps, as well as discussion of some of the “strategic” elements of the scheduler in optimizing use of its resources, to maintain a degree of simplicity in this Annex.

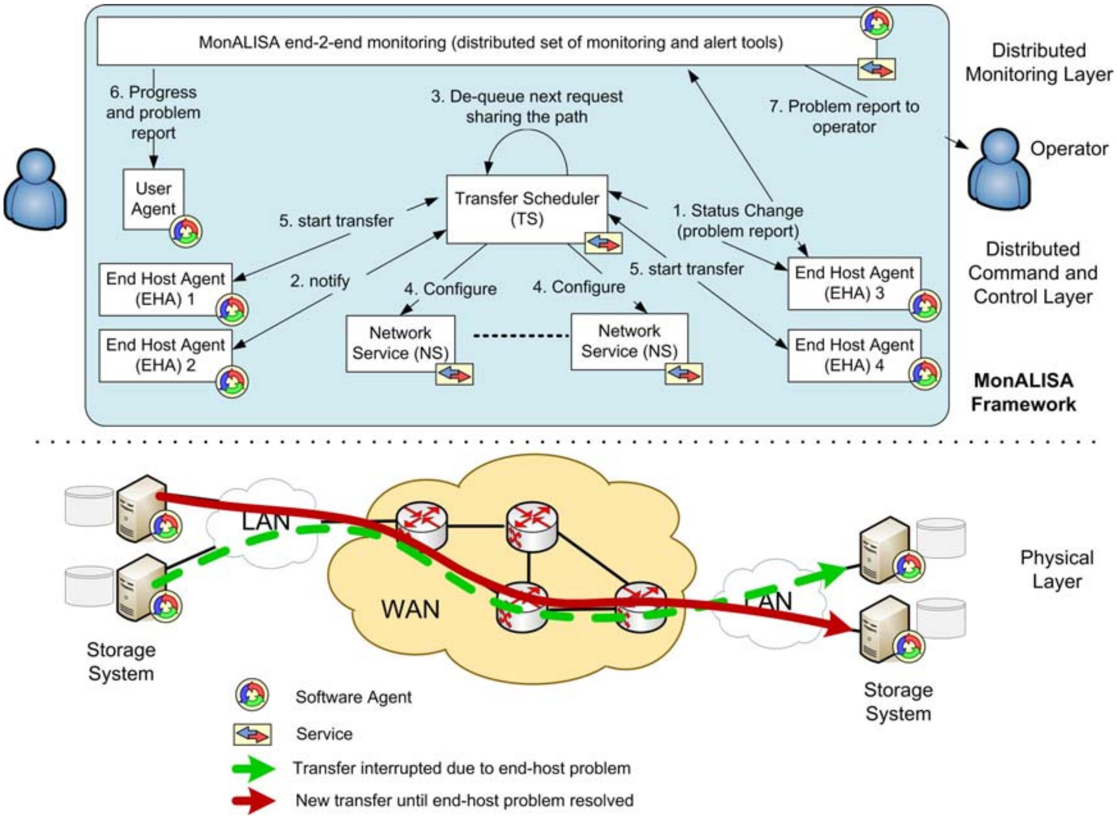

MonALISA end-2-end monitoring (distributed set of monitoring and alert tools)
Distributed Monitoring Layer
6. Progress and problem report
3. De-queue next request sharing the path
7. Problem report to operator
Operator
User Agent
Transfer Scheduler (TS)
5. start transfer
1. Status Change (problem report)
Distributed Command and Control Layer
End Host Agent (EHA) 1
2. notify
4. Configure
4. Configure
5. start transfer
End Host Agent (EHA) 3
End Host Agent (EHA) 2
Network Service (NS)
Network Service (NS)
End Host Agent (EHA) 4
MonALISA Framework
LAN
WAN
LAN
Physical Layer
Storage System
Storage System
Software Agent
Service
Transfer interrupted due to end-host problem
New transfer until end-host problem resolved

Figure 36: Example of system response to end-host related problems.

6. At all time the user is kept up-to-date with the transfer progress and the problem status.
7. If the problem cannot be resolved automatically, the operator responsible for the failed system is notified (e.g. by an e-mail to the Operations Centre).

Authentication, Authorization, and Accounting (AAA)

The system will use external AA (Authentication and Authorization) services used by the respective VOs. Accounting will be implemented using the MonALISA services (as already done in some other grid applications), and will be used to provide usage statistics. At a later stage it is foreseen to implement a quota-based system to help guarantee fair-sharing of the resources among the users.

Inter-domain provisioning and local connectivity at data centers

Data transfers to and from Tier-2 centers have to cross other networks, notably Internet2 and ESNet. For true end-to-end service, including Tier-2 hosts, the system has to interact with the circuit services provided by these domains. Internet2 is currently building its circuit services using the DRAGON[2] project, while ESNet is constructing the OSCARS [3] system for this purpose.

The main challenge stems from the fact that the different domains will implement different data and control plane technologies. For example, OSCARS is based on MPLS, while DRAGON has implemented a GMPLS control plane and uses Ethernet or SONET in the data plane. As proposed in [4], a unified Inter-Domain Communication mechanism based on Web Services technology can provide the necessary functionality for inter-domain circuit provisioning. We are planning to interface to OSCARS and DRAGON by using the WS E-NNI as described in [4].

A recent proposal from [7] for a Web-Services based control plane for inter-domain provisioning is shown in Figure 37.

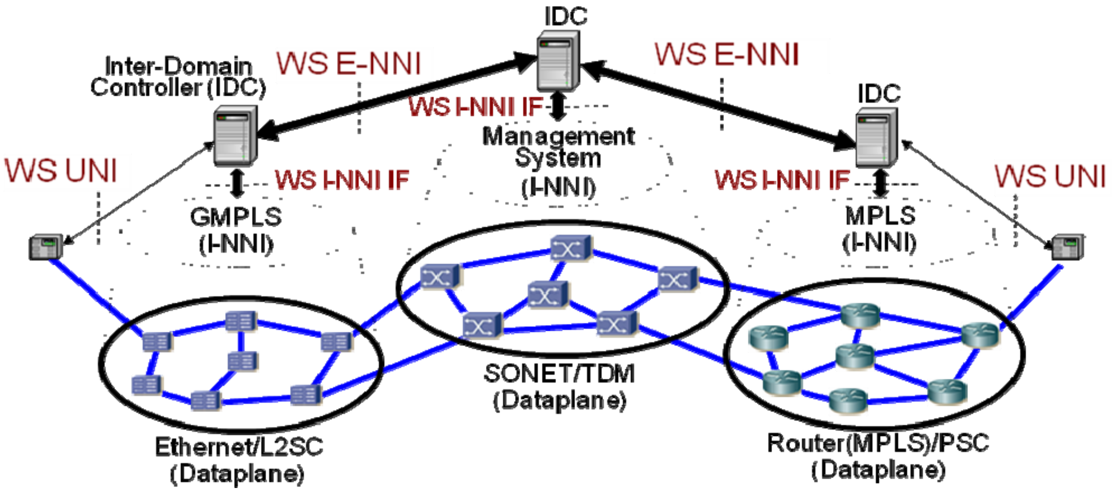


Figure 37: Proposal for Web-Services based control plane for inter-domain provisioning in hybrid networks. (T. Lehman et al.)

In order to provide a system capable of end-to-end provisioning and monitoring, it is also mandatory to provide adequate connectivity at the sites where the end-hosts are located, either through configuration of dedicated VLANs or by providing QoS guarantees. To achieve this, we plan to interface to systems such as LambdaStation [5] and TeraPaths [6] for configuration of the local connectivity at the data centre sites. In fact, collaboration with LambdaStation, TeraPaths, ESNet (OSCARS) and Internet2 (DRAGON) has been already established, with the goal to converge on a common web-services based interface for cross-domain interoperability, as described above.

On sites which will not deploy LambdaStation or TeraPaths, the only requirement is the provisioning of an adequate data network connection. Depending on the local area network (LAN) topology, our path services will treat the local network as one or several fixed-bandwidth segments of the circuit.

Data transfer and management tools

The final aim of our network services is to provide to the experiments specific data transfer and management tools with a uniform interface to the network, and to enable the network to be used as a managed part of the data movement infrastructure. In this we are collaborating with teams from CMS and ATLAS in order to integrate our services with their data management software.

Milestones

The milestones summarized below aim at deploying the complete system before the start-up of the LHC, in June 2008. We plan an early deployment of a set of basic services, for testing and development purposes. For this purpose we have selected a set of sites which have expressed interest in participating in the test during development. This shall also serve to demonstrate the capabilities of the system to the experiments. The detailed schedule is listed below:

October 2007:

- Deploy EHAs on participating test sites, and a first set of NS and TS components.
- EHA functionality: monitoring, basic end-host configuration (routing table)
- NS functionality: monitoring, "static" circuit setup
- Scheduler functionality: UI/API, simple scheduling algorithm, only on request and end-of-transfer, keeps circuit active in case of problems
- MonALISA monitoring framework

Demonstrator: Data transfer between pairs of participating sites. Users can monitor the transfers through MonALISA. EHAs monitor basic parameters.

December 2007:

- Enable transfers from mass storage system to mass storage system.
- Monitor end-host parameters important for the data transfer through the EHAs.
- Upgrade EHA functionality: add configuration features (VLAN configuration)
- Upgrade NS functionality: dynamic change of circuit parameters (in- or decrease bandwidth, route change)
- Upgrade Scheduler functionality: authentication, react to errors (e.g. link down)

Demonstrator: Transfer an experiment's dataset using a simple (User Agent) client that initiates a third party transfer from storage-system source to storage-system target. The User Agents negotiates with the EHAs on necessary network resources.

February 2007

- Interface with the experiment's data transfer application (that is enable the transfer application to perform third party transfers using the TS interface)
- Further refinement of monitoring parameters for end-hosts and actions that need to be taken in case of transfer problems. Learn from what the current transfer

operators experience as important parameters and what failure modes they encounter (user feedback).

Demonstrator: Experiment's data transfer applications initiate a third party transfer through the TS interface.

March 2008:

Measure the quality of the data transfers that use the EHAs through the expriment's data transfer application, versus conventional data transfers. Interface to LambdaStation, TeraPaths, DRAGON (Internet2) and OSCARS (ESNet). This will allow us to extend the functionality of the system to provide end-to-end transfer capability across other network domains.

May 2008:

Based on experience throughout the previous months, incrementally improve automated exception handling of transfers (mainly end-host problems) where possible. It is difficult to say what exactly needs to improved as it depends on the experience with data transfers during the previous months.

Summary

The network services system presented in this Annex will enable movement of large datasets between data centers in a truly managed way. It will provide network resource allocation for end-to-end data transfers, and a global monitoring covering the network as well as the involved end-systems. Problem resolution will be automated where possible, and facilitated otherwise by providing the operator with a global view of the ongoing and scheduled transfers.

In order to construct an end-to-end system, we collaborate closely with other network domains and systems (ESNet, Internet2, LambdaStation, TeraPaths), groups involved in development of storage systems (dcache, fts) as well as the LHC experiment's data management systems (for example PheDEx in CMS, FTD in ALICE, etc.).

First deployment of a prototype system to selected sites is foreseen for October 2007, leading to a full-featured release including inter-domain provisioning for May 2008, in time for the startup of the LHC in June 2008.

The longer term plan is to implement additional functionality in the Scheduler, taking advantage of higher-level MonALISA services that use the monitoring information to make "best use" of the available network resources, shared among the many VOs, while simultaneously attempting to minimize the typical transfer-times in response to individual requests. Experience with analogous problems in the use of MonALISA to underpin the VRVS/EVO system, and the dynamic management of Layer 0 and 1 end-to-end paths as already implemented in VINCI[14] using MEMS[15]-based purely optical switches, gives us confidence that a consistent solution can be developed.

References

[1] Office of Science, U.S. Department of Energy, "High-Performace Networks for High-Impact Science", Report of the High Performance Network Planning Workshop 2002, available at http://www.es.net/pub/esnet-doc/2-3high-performance_networks.pdf

[2] DRAGON (Dynamic Resource Allocation via GMPLS Optical Networks), http://dragon.east.isi.edu

[3] DOE ESNet, "OSCARS: On-demand Secure Circuits and Advance Reservation System", http://www.es.net/oscars

[4] T. Lehman , X. Yang, C. Guok, N. Rao, A. Lake, J. Vollbrecht, N. Ghani, "Control Plane Architecture and Design Considerations for Multi-Service Multi-Layer, Multi-Domain Hybrid Networks", INFOCOM 2007, IEEE (TCHSN/ONTC), also available at http://www.es.net/OSCARS/documents/papers/2007hsn-infocom-paper-lehman-etal.pdf

[5] Bobyshev, X. Su, H. Newman, M. Crawford, P. DeMar, V. Grigaliunas, M. Grigoriev, A. Moibenko, D. Petravick, R. Rechenmacher, M. Thomas, Y. Xia, J. Bunn, C. Steenberg, F. van Lingen, S. Ravot, "Lambda Station: On-demand Flow Based Routing for Data Intensive Grid Applications over Multitopology Networks", In proceedings of GRIDNETS 2006, San Jose, California, October 1-2, 2006 (see also: http://www.lambdastation.org/)

[6] Gibbard, D. Katramatos, D. Yu, S. McKee, "TeraPaths: End-to-End Network Path QoS Configuration Using Cross-Domain Reservation Negotiation", In proceedings of GRIDNETS 2006, San Jose, California, October 1-2, 2006 (see also: http://www.atlasgrid.bnl.gov/terapaths/)

[7] T. Lehman, C. Guok, A. Lake, J. Vollbrecht, "Hybrid Network Control Plane Interoperation Between Internet2 and ESnet", Joint Techs Workshop, Fermilab, Batavia IL, July 16 2007 (see also: http://events.internet2.edu/2007/jt-batavia/sessionDetails.cfm?session=3348&event=272)

[14] See Annex H.

[15] Micro-Elctro-Mechanical Systems, see for example www.**mems**net.org/**mems**/what-is.html. We used such switches manufactured by Glimmerglass and Calient starting in 2005, in UltraLight.

Annex G: 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 CMS, ALICE, ATLAS, LHCb the Open Science Grid (OSG), and the Russian LCG sites. It actively monitors US LHCNet and the UltraLight testbed, as well as the Enlightened project. It collects traffic measurement information on all segments of the Internet2 backbone, part of the GLORIAD network, and several other major links used by the HEP community (CERN-GEANT, Taiwan – Starlight, CERN –IN2P3, etc.). MonALISA also is used to monitor, control and administer all of the VRVS/EVO reflectors, and to help manage and optimize their interconnections.

As of this writing, more than 350 MonALISA services are running throughout the world. These services monitor more than 20,000 compute servers, and thousands of concurrent jobs. More than 1,000,000 parameters are currently monitored in near-real time with an aggregate update rate of approximately 10,000 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.

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 38*,*Figure 38* 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.

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 three 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.

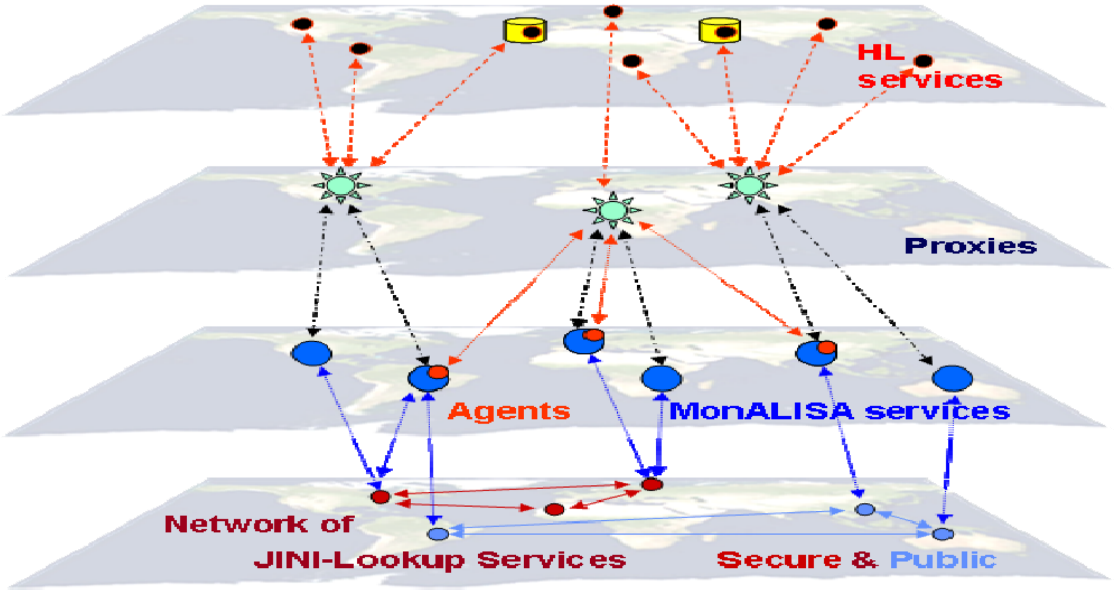


Figure 38. The four layers, main services and components of the MonALISA framework

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.

ATLAS and LHCb jointly developed GANGA, a tool which is designed to give physicists a simple and consistent way to organize and execute analysis programs. GANGA is instrumented using a MonALISA API. The monitoring information is used as an automatic feedback from different user communities, and it can be used by users or system administrators to understand how the system is functioning, and to detect problems. DIANE is a job execution framework used by ATLAS and LHCb, based on a master-worker processing model. DIANE's application plugins make use of MonALISA monitoring sensors to report application-specific data during execution. This information can be used by the submitter to follow the application's progress and the computer resources acquired, or it can be used by the framework itself to optimize the (re)scheduling decisions.

MonALISA is also used to monitor the network traffic on more than 100 WAN links from several major networks (Internet2, US LHCnet, Ultralight and Roedunet). 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 (as discussed in) 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 like 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.

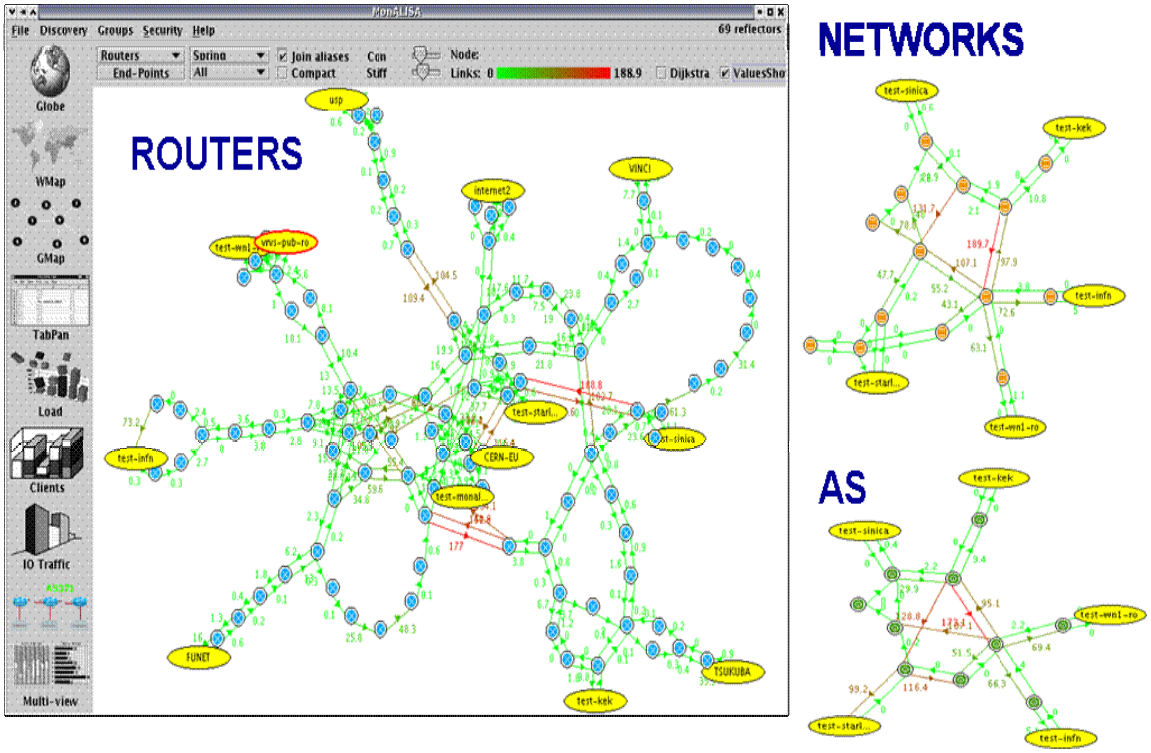
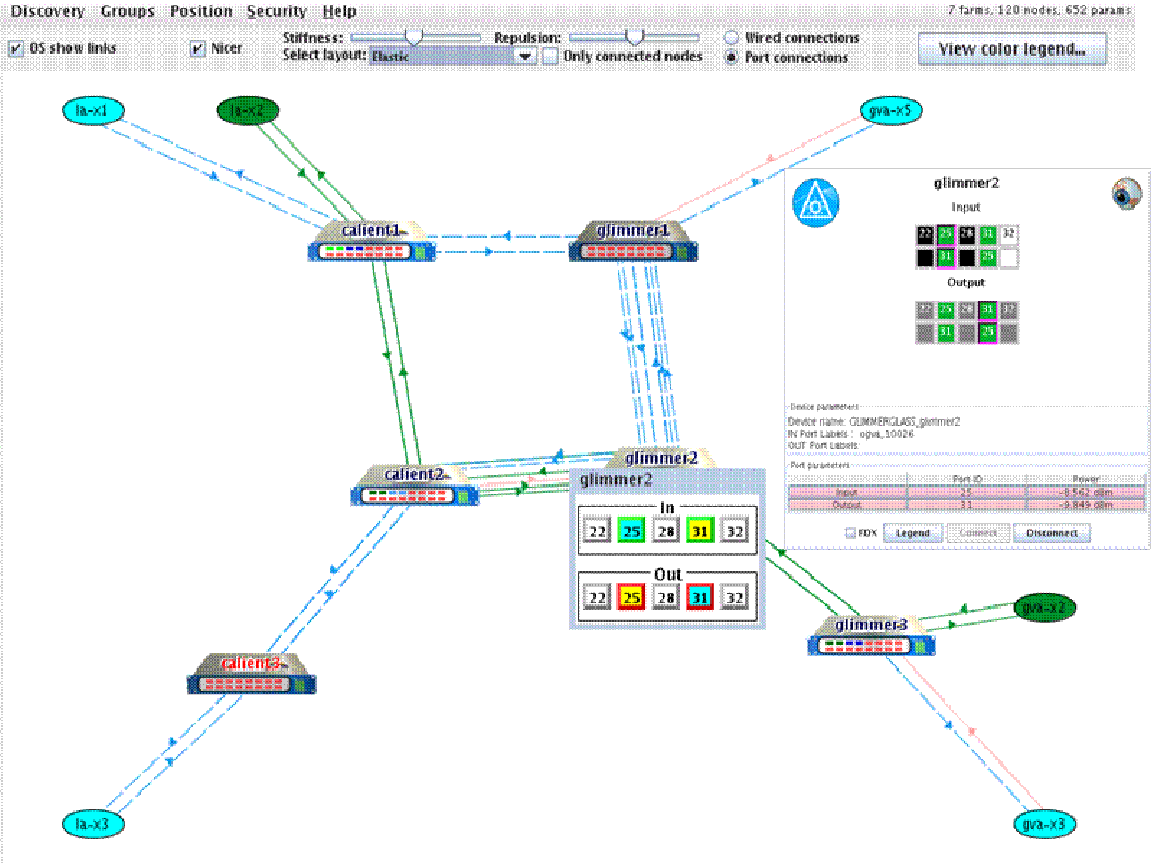


Figure 39. MonALISA real time view of the topology of WANs used by HEP. A view of all the routers, or just the network or "autonomous system" identifiers can be shown.

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 39. The combined information from all the sites allows one to detect asymmetric routing 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.

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 40 we show the MonALISA GUI that is used to monitor the topology on Layer 0/1 connections and the state of the links.



7 farms, 120 nodes, 652 params

☑ OS show links
☑ Nicer
Stiffness:
Repulsion:
Select layout: Elastic
□ Only connected nodes
○ Wired connections
● Port connections

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

□ FDX

Figure 40. Monitoring and autonomous control for optical switches and optical links

MonALISA is used to monitor all of the traffic in US LHCnet, including:

- The traffic on all interfaces (and peering) using SNMP on the Force 10 switches.
- State of the links and error counters
- sFlow analysis, and aggregation per sites and applications
- Status of links from the CIENA CD/CI links and circuits, via TL1 and trigger alarms

Figure 41 shows the real-time topology graph, indicating the traffic and link state on each segment, along with two example plots (of many available using data in the MonALISA repositories): of the historical traffic-data on the links and the aggregated total traffic.

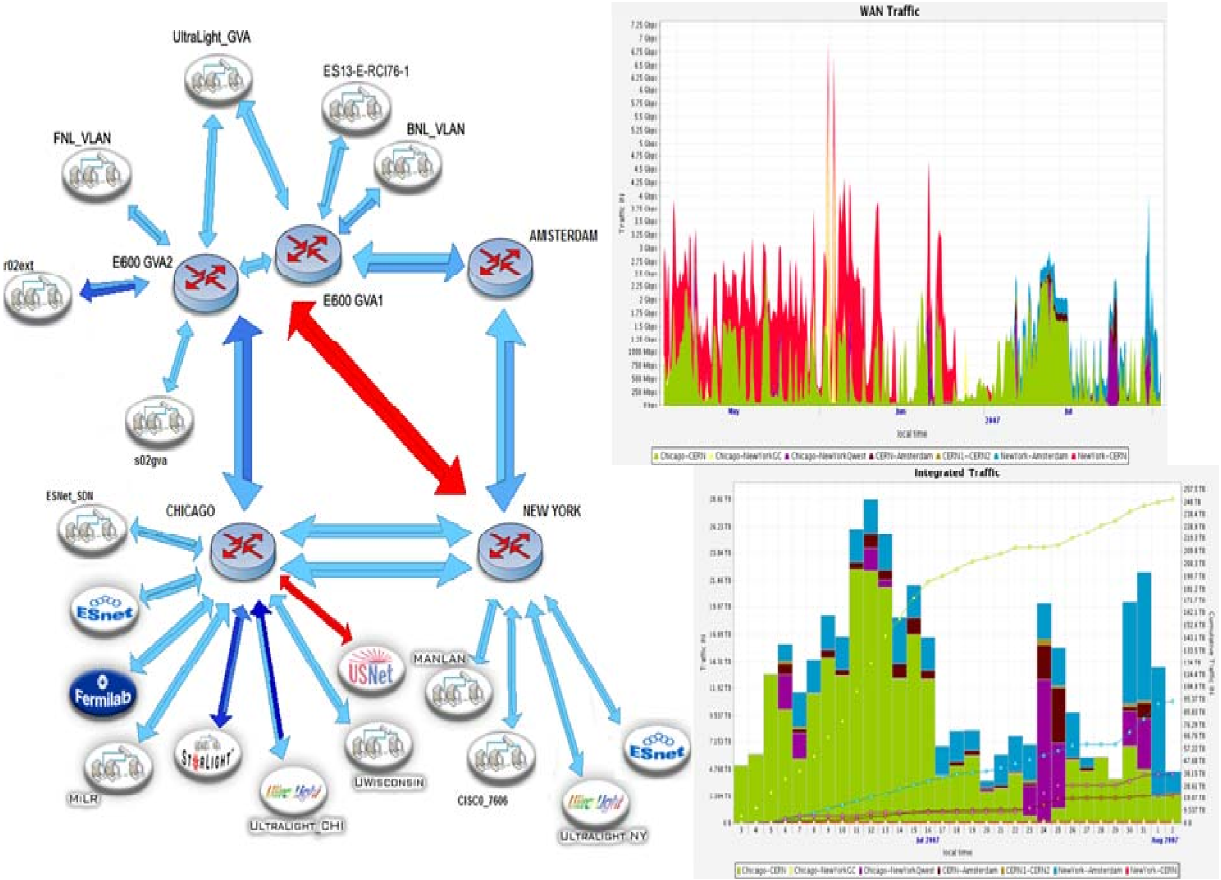


Figure 41 . Example MonALISA monitoring panels for US LHCnet

Annex H: VINCI: Virtual Intelligent Networks for Computing Infrastructures

To satisfy the demands of data intensive grid applications it is necessary to move to more synergetic relationships between users' applications and networks. The main objective of the VINCI project16 is to enable users' applications, at the LHC and in other fields of data-intensive science, to use networks more effectively by coordinating the use of packet-switched and circuit-oriented network resources to form end-to-end network paths with reliable performance. As described in this Annex, several of VINCI's services which have already been developed and tested are part of the Network Services system described in Annex F. Other VINCI services and functions provide the foundation for more-strategic scheduling and overall system performance optimization using "global" monitoring information, that will be used in future releases of the Network Services system over the next three years.

Field trials of VINCI to support large data flows by dynamically building optical paths, to automatically re-route traffic over alternative paths in case of failure, and to develop an optical control plane that can integrate across optical (Layer 1) and VLAN (Layer 2) segments, were successfully carried out over the last two years, as described in this Annex.

End To End Optical Control Plane

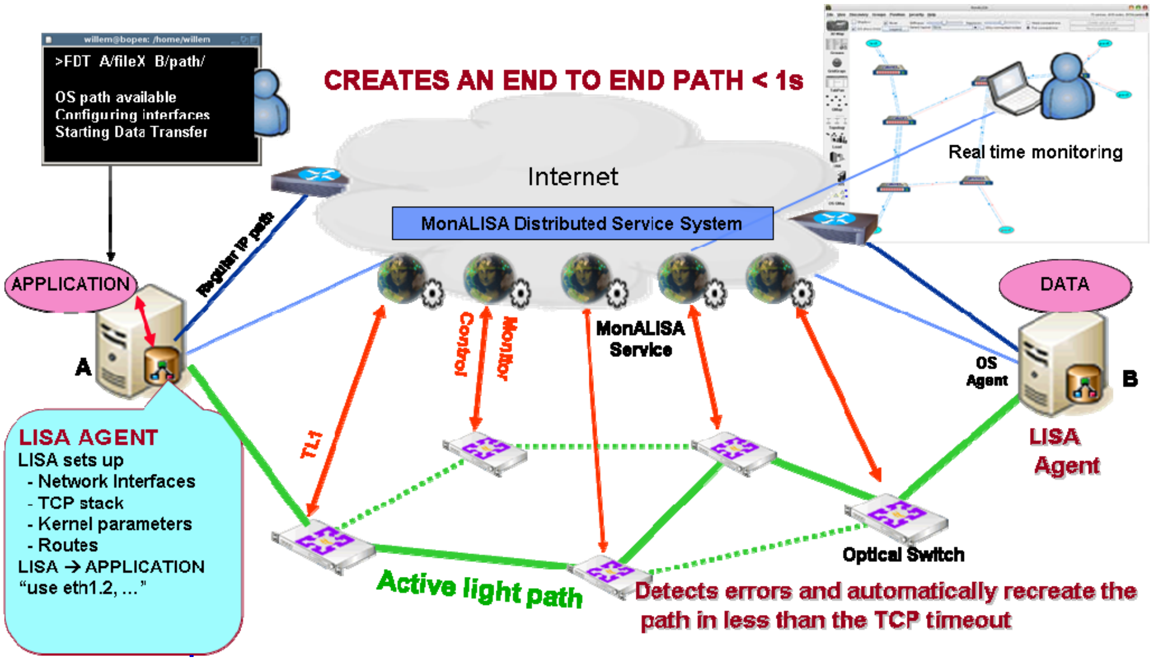
The MonALISA / VINCI framework has been applied to develop an integrated Optical Control Plane system (OCPS) that controls and creates end-to-end optical paths on demand, using optical switches.

As part of the development of end-to-end circuit-oriented network management services, we developed dedicated modules and agents to monitor, administer and control Optical Switches; specifically the purely photonic switches from Calient and Glimmerglass. The modules use TL1 commands to monitor the connectivity matrix of each switch, as well as the optical power on each port. Any change in the state of any link is reported to dedicated agents. If a switch is connected to the network, or if it ceases to operate, or if a port's light level changes, these state changes are detected immediately and are reflected in the topology presented by the MonALISA Graphical User Interface (GUI). By using the GUI, an authorized administrator also can manually construct any light path, and monitor the optical power on each new link as it is created.

The distributed set of MonALISA agents was used to control the optical switches, and to create an optical path on demand. The agents use MonALISA's discovery layer to "discover" each other, and then communicate among themselves autonomously, using the Proxy services. Each proxy service can handle more than 1,000 messages per second, and several such services are typically used in parallel. This ensures that the communications among the agents is highly reliable, even at very high message-passing rates.

[16] http://monalisa.cern.ch/monalisa__Service_Applications__Vinci.html

The set of agents also is used to create a global path or tree, as it knows the state and performance of each local area and wide area network link, and the state of the cross connections in each switch. The routing algorithm provides global optimization by considering the "cost" of each link or cross-connect. This makes the optimization algorithm capable of being adapted to handle various policies on priorities, and pre-reservation schemes. The time to determine and construct an optical path (or a multicast tree) end-to-end is typically less than one second, independent of the number of links along the path and overall the length of the path. A schematic view of how the MonALISA agents are used to create an optical path for an (authorized, authenticated) end-user application is presented in *Figure 42*.



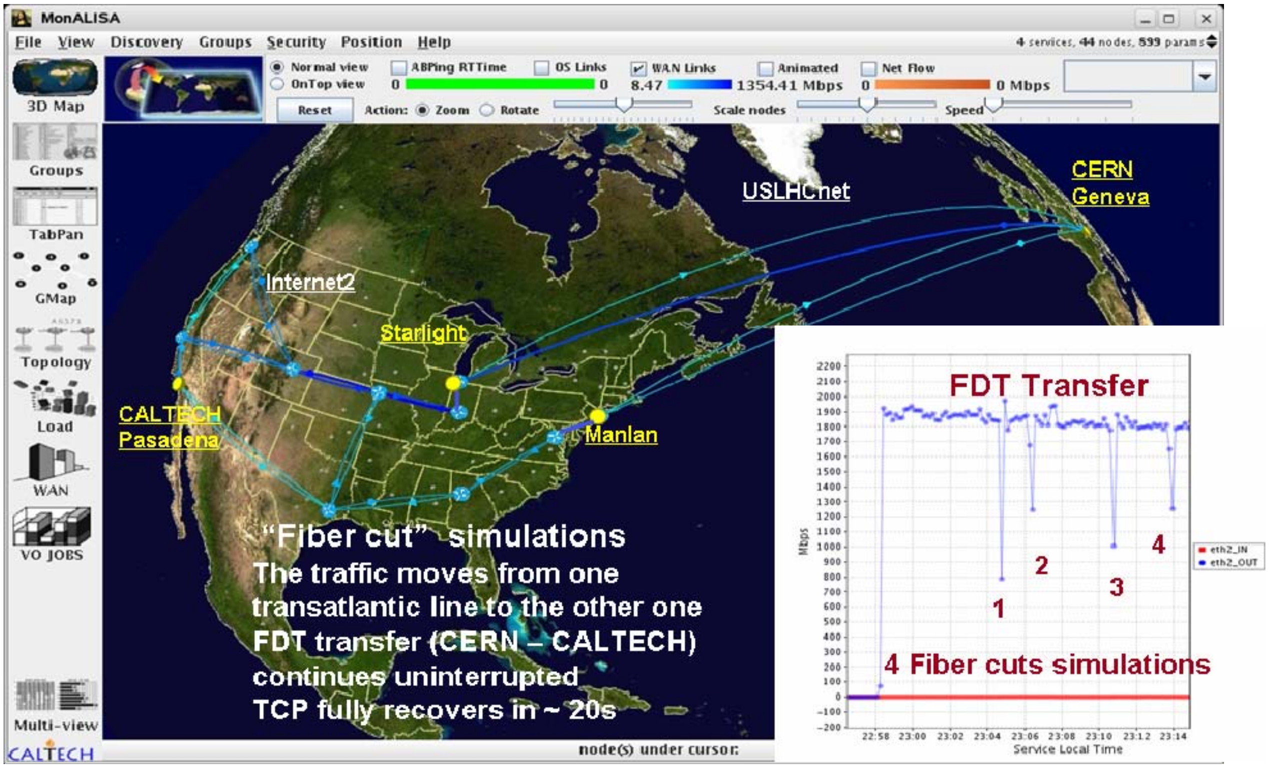
```
willem@bopex: /home/willem
>FDT  A/fileX  B/path/

OS path available
Configuring interfaces
Starting Data Transfer
```

Figure 42 MonALISA agents are used to monitor and control optical switches. The agents interact with end-user applications to provision an optical path on demand.

If network errors are detected, an alternative path is set up rapidly enough to avoid a TCP timeout, so that data transfers will continue uninterrupted. This functionality will be important in the construction of the virtual circuit-oriented network services, mentioned in previous section

Figure 43 shows an example of how MonALISA is used to create dynamically, on demand, a path between two end- systems CERN and Caltech. The topology, the cross-connections, the ports and the segments where light is detected and the end-to-end path created by the system all are displayed in real-time. The end to end path is created in approximately 0.5 seconds, and then disk-to-disk data transfer using FDT is started. We simulated 4 consecutive "fiber cuts" in the circuits over the Atlantic. The agents controlling the optical switches detect the optical power lost and they created another complete path in less than 1 second.   The alternative path was set up rapidly enough to avoid a TCP timeout, so that data transfers continue uninterrupted (Figure 43, right). As soon as the transfer initiated by the end-user application was completed, the path was released.



“Fiber cut” simulations

The traffic moves from one transatlantic line to the other one

FDT transfer (CERN – CALTECH) continues uninterrupted

TCP fully recovers in ~ 20s

Figure 43 MonALISA / VINCI agents used to create an end to end path. Four “fiber cut” simulations were done for the transatlantic circuits. The alternative path was created rapidly enough to avoid TCP timeout and the FDT traffic continued uninterrupted.

Annex I: US LHCNet Request for Proposals

REQUEST FOR PROPOSAL NO. KS100

FOR

Supply of multiple 10Gbit/s Transparent SONET Circuits interconnecting Geneva, Chicago, New York, Amsterdam, London, and Paris

DATE OF ISSUANCE: July 17, 2007

PROPOSALS TO BE RECEIVED AT CALTECH NO LATER THAN

DATE: August 29, 2007

TIME: 17:00 PDT

COMMUNICATIONS IN REFERENCE TO THIS RFP

Send two (2) completed copies of this RFP to the attention of:

California Institute of Technology

Attn: Harvey Newman
256-48 HEP
1200 E. California Blvd.
Pasadena, CA 91125
USA

Phone: +1 626 7982323
Fax: +1 626 795 3951
E-mail: Harvey.Newman@cern.ch

CERN

Attn: Artur Barczyk
31 R-016
CH-1211 Genève 23
Switzerland

Phone: +41 22 7675801
Fax: +41 22 7670160
E-mail: Artur.Barczyk@cern.ch

All other communications in reference to this RFP shall be directed to the following contacts, and must be identified with the RFP number (KS100):

COMMERCIAL:	Harvey NEWMAN	Tel: +1 626 798 23 23 E-mail: newman@hep.caltech.edu
TECHNICAL:	Artur BARCZYK	Tel: +41 22 76 75801 E-mail: Artur.Barczyk@cern.ch
	Dan NAE	Tel: +41 22 76 74415 E-mail: Dan.Nae@cern.ch

Table of Contents

I. Project Scope 1

II. Introduction 1

III. Definitions 3

IV. Service Requirements 3

V. Selection Process 4

VI. Subcontractors 5

VII. Combination of Firms 5

VIII. Late Proposals 6

IX. Alternative Proposals 7

X. Sealing and Marking of Bids 7

XI. Deadline for Submission of Bids 8

XII. Proposal Timetable 8

XIII. Confidentiality and Proprietary Rights 9

XIV. Term of Proposal 9

XV. Supplier Questionnaire 10

Attachment A – Bid Statement of Work 14

Attachment B – Bid Sheets 21

Attachment C – Terms of Agreement 23

Attachment D - SECTION 1 - COMMERCIAL ITEMS OR SERVICES AGREEMENT GENERAL PROVISIONS 26

Attachment E - SECTION 2 - GOVERNMENT FUNDED GRANT PROVISIONS 27

Attachment F – Acceptance of Terms and Conditions 28

I. Project Scope

The Large Hadron Collider (LHC) now being built at CERN (Geneva – Switzerland) is a particle accelerator which will probe deeper into matter than ever before. This will allow scientists to penetrate still further into the structure of matter and recreate the conditions prevailing in the early universe, just after the "Big Bang". Once it starts operation in 2007, it will generate tens of petabytes of experimental data each year.

U.S. physicists involved in the LHC depend on reliable high speed networks if they are to contribute effectively to the LHC physics program and take part in the physics discoveries. A robust and performant transatlantic network interconnecting U.S. institutions and CERN is an essential resource for U.S. participation in the LHC experiments. The California Institute of Technology is mandated by the Department of Energy science (DOE) to design, deploy and operate a performant and reliable transatlantic network.

The purpose of this RFP is to select Telecom Operators having the capability and the experience to provide reliable and cost effective solutions for up to eight 10Gbit/s circuits between CERN (Geneva) and the USA.

II. Introduction

The European Laboratory for Particle Physics CERN on the Franco-Swiss border near Geneva provides experimental facilities for particle physics experiments, mainly in the domain of high energy physics (HEP). The next particle accelerator, due to start operation in year 2007, is the 14 TeV (1 Tera electron volt = 1 trillion (10^{12}) electron volts) Large Hadron Collider (LHC). The LHC is being built using high powered 14 meter-long superconducting magnets and is in the process of being installed. The four LHC experiments are expected to produce tens of petabytes per experiment each year. To enable U.S. participation into the next generation experiments taking place at CERN, the California Institute of Technology has been mandated by the DOE to design, deploy and operate a transatlantic network.

Many petabytes of physics data generated by experiment will be distributed in near real time to approximately 10 LHC regional computer centers (Tier1) around the world, including two Tier1 centers in the USA located at Fermi National Laboratory (FNAL[17]) near Chicago and Brookhaven National Laboratory (BNL[18]) near New-York city respectively. There are also approximately 100 smaller "Tier2" centers planned, most of which are already in operation at universities or laboratories, that will access and in some cases exchange large quantities of data

[17] http://www.fnal.gov

[18] http://www.bnl.gov

with the Tier1s: within a region and in some cases crossing between regions, including between the US and Europe. The sheer volume of the data combined with the complexity of the analysis to be performed, and the requirement that the processing of the data shall be done in a fully distributed manner, places heavy demands on the High Energy & Physics (HEP) computing and networking infrastructure.

The traffic produced by the exchange of data between CERN and US Tier1 regional computing centers presents enormous IT challenges. Communities of thousands of scientists, distributed globally and served by networks of varying bandwidths, will need to exchange information, with data volumes in the range of 100 Terabytes to 100 Petabytes and possibly more, over the next decade.

According to the current estimates, the minimum bandwidth requirements between CERN and the USA are 4 x 10Gbit/s in 2008, rising to 8 x 10 Gbps in 2010.

In order to be ready to tackle these challenges on time, Caltech is reviewing the interconnection topology of its current transatlantic network infrastructure which consists of four points of presence in Chicago, New-York, Amsterdam and Geneva interconnected with 10 Gb/s SONET circuits.

Caltech is also considering making use of the European research and education network GEANT2[19] infrastructure, to provide OC-192 connectivity from CERN to their points of presence in London and Paris. Caltech may use the GEANT2 links, together with transatlantic links terminating in London or Paris that are part of this request for proposal, as parts of the New York – Geneva and Chicago – Geneva circuits, in order to reduce operational costs as well as to guarantee path diversity on the continental links in Europe and the US. GEANT2 is managed and operated by DANTE[20].

The purpose of this RFP is to:

1. Renew the contracts of the existing infrastructure; this does not exclude new Bidders.
2. Select two to three contractors for a period of three (3) years.
3. Increase the transatlantic bandwidth to four (4) 10 Gbps in 2008, to six (6) in 2009 and eight (8) in 2010.

The current topology is shown in Attachment A – Statement of Work, Figure 1.

[19] http://www.geant2.net

[20] http://www.dante.net

Definitions

10 Gb/s Circuit(s) shall refer to un-protected, fully transparent 10 Gbit/s (i.e. OC-192/STM-64 with SONET/SDH framing) transatlantic wavelengths.

1. NY shall refer to New York, New York.

2. CHI shall refer to Chicago, Illinois.

3. AMS shall refer to Amsterdam, Netherlands.

4. GVA shall refer to Geneva, Switzerland.

5. LON shall refer to London, England.

6. PAR shall refer to Paris, France.

7. Caltech shall refer to the California Institute of Technology

Service Requirements

The service requested in accordance with Attachment A – Statement of Work, consists of the following items:

1. GVA-CHI 10Gb/s Circuit (up to two (2) circuits)

2. GVA-NY 10Gb/s Circuit (up to three (3) circuits)

3. GVA-AMS 10Gb/s Circuit (up to three (3) circuits)

4. AMS-NY 10Gb/s Circuit (up to two (2) circuits)

5. NY-CHI 10Gb/s Circuit (up to three (3) circuits)

6. CHI-LON 10Gb/s Circuit (up to two (2) circuits)

7. CHI-PAR 10Gb/s Circuit (up to two (2) circuits)

8. NY-PAR 10Gb/s Circuit (up to three (3) circuits)

9. NY-LON 10Gb/s Circuit (up to three (3) circuits)

The evaluation process will consider each item independently. Bidders will not be disqualified for submitting proposals which do not include all items listed above. Caltech will select a subset of these services, such as to obtain the configuration shown in Attachment A – Statement of Work, Figure 1, including the existing and 2008 circuits.

In addition to bids on the individual items listed above, Bidders are also encouraged to propose "Packages" including several items if circuit costs can be reduced. Proposals for Packages taking into account the upgrade plans as shown in Attachment A – Statement of Work, Figures 2 and 3 are welcome. In the case a Package is proposed, Bidders shall ensure that circuit physical paths are complaint with Attachment A – Statement of Work.

III. Selection Process

Evaluations will be ranked based on clarity and thoroughness of response, cost, accessibility of facilities, favorable references, and acceptance of Caltech's Terms and Conditions.

The technical expressions, terminology, and abbreviations used in this RFP are considered known by the Bidders.

This RFP does not commit Caltech to pay any costs incurred in submitting your proposal, making studies or designs for preparing the proposal or in procuring or subcontracting for services or supplies related to the proposal.

If the proposal contains data that either you or your subcontractors do not wish to be disclosed for any purpose other than proposal evaluation, you must mark your cover sheet with the legend below:

"Data contained in pages _________ of this proposal furnished in connection with RFP No. KS100 shall not be used or disclosed, except for evaluation purposes, provided that if an Agreement is awarded to this offer or as a result of or in connection with the submission of this proposal, Caltech shall have the right to use or disclose this data to the extent provided in the contract. This restriction does not limit Caltech's right to use or disclose any data obtained from another source without restriction."

During the proposal preparation period, all requests for clarification and/or additional information, must be submitted in writing to the individual referenced by "Attention:" on the cover page of this RFP. When appropriate, responses to requests, as well as any Caltech initiated changes, will be provided to all prospective proposers in writing as addenda to the RFP. (NOTE: You must include reference to all addenda on your proposal cover letter.)

Unnecessarily elaborate brochures or presentation layouts, other than those sufficient to present a complete and effective proposal, are not desired.

Contractors must develop and present a response that represents a complete and binding offer. While further negotiations and due diligence will be performed, Contractor finalists will be selected based on these responses.

Caltech reserves the right to retain all proposal information submitted in response to this RFP.

Caltech reserves the right to reject all proposals, to award contract(s) based on initial proposals (without proposal clarifications) or conduct oral discussions (for the purpose of proposal clarification) prior to making source selection.

IV. Subcontractors

The bidder shall specify in Section XIV, Question 6, any part of the obligations which he/she would, if awarded the contract, want to subcontract. The bidder shall specify the name of the proposed subcontractor(s) and sub-subcontractor(s), the nature of the subcontracting, the address(es) of the premises where the subcontracted or sub-subcontracted obligations would be performed as well as their respective value.

Caltech shall be entitled to reject in whole or in part the bidder's proposal(s) concerning the subcontracting of his obligations, it being understood that in any event:

- the proposed subcontract(s) shall not represent more than 51% of the value of the contractual obligations in total and 30 % per subcontract,
- the management and follow-up of the contract shall not be sub-contracted,
- any change in subcontracting/subcontractors under the contract shall be subject to prior permission in writing by Caltech.

V. Combination of Firms

The firms comprising the combination of firms shall jointly appoint one firm amongst them as their sole representative in all matters concerning the proposal and the contract save for the signing thereof, and that firm shall provide Caltech with evidence in writing of its appointment. If it was appointed for the market survey, that appointment shall also be valid for the tender and the contract.

The firms comprising the combination of firms shall be jointly and severally liable for the performance of the bidder's obligations under the invitation to tender.

The bidder shall specify in the tender the percentage shares of the contract allocated to each firm of the combination of firms.

Any change under the contract in the composition of the combination of firms or the percentage shares of the contract allocated to each firm shall be subject to the prior written permission in writing by Caltech.

VI. Late Proposals

Any proposal, portion of a proposal, or unrequested proposal revision received at Caltech after the time and date specified on the cover page of this RFP will be considered late. Late proposals will not be considered for award, except under the following circumstances:

(1) Caltech determines that the late receipt was due solely to a delay by the U.S. postal service for which the Bidder was not responsible. Timely postmark or receipt of registered, certified, or express mail "next-day service," establishing the time of deposit, must be evidenced.

(2) Caltech determines that the proposal was late due solely to mishandling by the Institute after receipt at Caltech, provided that the timely receipt at Caltech is evidenced.

(3) The complete proposal has been submitted in electronic form by the date specified in Section XI, followed by paper copy sent by overnight express to the addresses specified in Section X.

(4) No acceptable proposals are received in a timely manner.

VII. Alternative Proposals

The Bidder is required to submit a bid on Attachment B - Bid Sheets. The bid shall be based on the documents included in this RFP (subject to any Addenda issued by Caltech) without any variation or alternative to the RFP documents (hereinafter referred to as a "Conforming Bid").

Any Bidder wishing to offer an alternative proposal, whether technical or otherwise must also submit a Conforming Bid. If technical or other alternatives are offered in addition to a Conforming Bid then such alternatives must be accompanied by all information (including, without limitation, technical, contractual and financial) to enable a complete evaluation of the alternative by Caltech. The Bidder shall indicate in his proposal any increases or deductions in the total price of the Conforming Bid that would result from the acceptance of any alternative proposal. Alternative Proposals must be submitted on separate sheets of paper, marked “RFP KS100 - Alternative Proposal,” and included with Bidder’s Conforming Bid RFP response.

Only the alternatives submitted by the successful Bidder, if any, as a result of the application of the selection and evaluation criteria described herein, shall be considered by Caltech, and it shall be within the sole discretion of Caltech whether to accept the Conforming Bid or the alternative submitted by that Bidder, regardless of the result of the comparative evaluation. The Bidder must, therefore, be prepared to enter into an Agreement with Caltech on the basis of its Conforming Bid notwithstanding any alternative offered.

VIII. Sealing and Marking of Bids

The RFP response and all accompanying documents must be sent **in duplicate** to the following addresses:

California Institute of Technology
Attn: Harvey Newman
256-48 HEP
1200 E. California Blvd.
Pasadena, CA 91125
USA
e-mail: Harvey.Newman@cern.ch

CERN
Attn.: Artur Barczyk
31 R-016
CH-1211 Genève 23
Switzerland
e-mail: Artur.Barczyk@cern.ch

The RFP shall be sent by registered mail or by courier service in one package comprising all documents, as stipulated in the cover letter and the name and address of the bidder for easy identification. Electronic submission to the e-mail addresses listed in Section VIII is accepted, if followed by a paper copy sent by overnight express to the addresses listed above.

IX. Deadline for Submission of Bids

The RFP response shall be sent to the addresses given in Section VIII and received no later than: **17:00 PDT, August 29, 2007**. The postmark, or the validated dated receipt of the courier service will be accepted as proof of date of posting.

A Bidder requiring any clarification of the RFP documents may notify Caltech in writing, e-mail, or by fax only. *All communications in reference to this RFP shall be directed* <u>*via e-mail or fax only*</u> *to the above contact, and must be identified with the RFP number (KS100). Communications received regarding this RFP via telephone will not be acknowledged. Drop-In visits to this office during the RFP process are not allowed.*

Caltech will respond to any request for clarification that it receives earlier than fourteen (14) calendar days prior to the deadline for submission of bids. Copies of Caltech's response will be forwarded to all bidders, including a description of the inquiry but without identifying its source. All bidders are required to seek clarification from Caltech for any apparent discrepancy discovered.

X. Proposal Timetable

The timetable is as follows:

Request For Proposal	July 17, 2007
Last date for questions	August 15, 2007
Last date for answers and corrections	August 22, 2007
Last date for submissions	August 29, 2007
End of selection and awarding process	September 26, 2007
Start of test[21]	November 26, 2007
Start of service	December 3, 2007

(These dates are subject to change by Caltech at any time. Caltech will use its best efforts to notify Contractor of any changes when made.)

XI. Confidentiality and Proprietary Rights

[21] A delivery and acceptance test of at least 72 hours is required to allow Caltech to verify the compliance of the service.

Bidders shall keep confidential and shall not without prior permission in writing by the Caltech disclose confidential information to any third party, or use it for purpose other than the performance of their obligations under the Invitation to Tender. Bidders shall limit the circle of recipients of confidential information on a need-to know basis.
Bidders shall continue to comply with its obligations as defined above for a period of five years from the date upon which the information is received.

Notwithstanding above clauses, a bidder is entitled to disclose confidential information for which it is required by law to disclose or which, in a lawful manner, it has obtained independently of confidential information, or which has become public knowledge other than as a result of a breach by that Bidder of above clauses.

Information disclosed by Caltech shall not create any proprietary right in respect of that information, bidders shall only use such information in so far as is necessary for the performance of its obligations under, and for the purposes of the Invitation to Tender.

XII. Term of Proposal

The proposal must be valid for a period of at least six (6) months. Each proposal will be treated confidentially. Each item of the proposal sent to Caltech is considered to be the property of Caltech unless otherwise specified and agreed by both parties. Caltech reserves the right to contract partially for the proposed service as well as not to contract all without justification.

XIII. Supplier Questionnaire

Your answers will be used to assist in determining your participation in the final round of selection.

General Company Information

1. Complete address, phone and fax numbers, e-mail addresses, and key contacts.

Company Name ______________________

Address ______________________

____

City, State, Zip ______________________

Phone/Fax ___________/___________

Key Contacts (Include name, title, phone number, and e-mail address(es)

____

____

2. What are the main products and/or services your company provides to customers?

____

____

____

3. How many years has your company been in business?________________________

4. Has your business been previously known by a different name? Yes No

If Yes, please explain ____

____

____

5. Please provide three references (other than Caltech) that we may contact. Include Company Name, contacts, and phone numbers

____

____

____

____

____

____

____

____

____

____

____

6. Caltech requires the successful start of service by December 3 2007, at the latest provided that a formal notification is given by Caltech 10 weeks earlier.

   Does the bidder agree with the delivery schedule? Yes No

7. Do the proposed services comply with the Request For Proposal in all respects? Yes No

   If not, please indicate the exceptions:

____

____

____

____

____

8. Please list names of all proposed subcontractors, the nature of subcontracting, the addresses of the premises where subcontracted obligations would be performed as well as their respective value.

____

____

____

____

____

____

____

____

____

9. What additional information can you provide that will help us make a decision in your favor? Include any additional services you provide at no added cost.

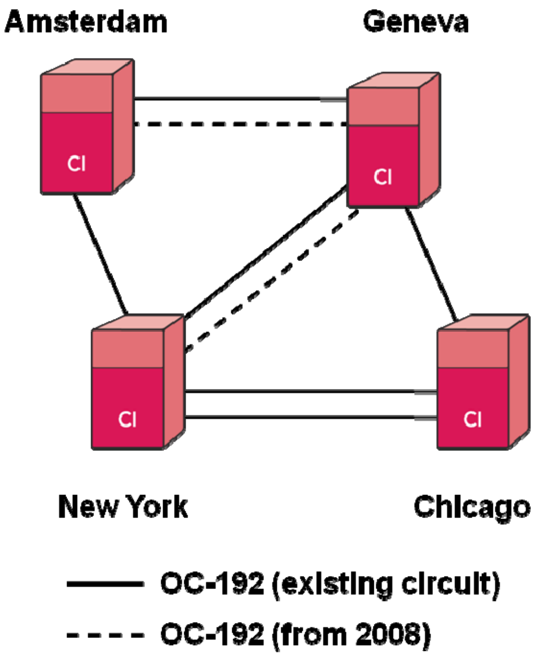
ATTACHMENT A
STATEMENT OF WORK
RFP KS100

Service Requirements

The service requested consists of the following items:

GVA-CHI 10Gb/s Circuit (up to two (2) circuits)
GVA-NY 10Gb/s Circuit (up to three (3) circuits)
GVA-AMS 10Gb/s Circuit (up to three (3) circuits)
AMS-NY 10Gb/s Circuit (up to two (2) circuits)
NY-CHI 10Gb/s Circuit (up to three (3) circuits)
CHI-LON 10Gb/s Circuit (up to two (2) circuits)
CHI-PAR 10Gb/s Circuit (up to two (2) circuits)
NY-PAR 10Gb/s Circuit (up to three (3) circuits)
NY-LON 10Gb/s Circuit (up to three (3) circuits)

Figure 1 (below) shows the circuits comprising the subject of this RFP, indicating the existing circuits and circuits planned to be taken in operation by 2008. The planned upgrade paths for 2009 and 2010 are shown in Figure 2 and Figure 3, respectively, on the following page.



The Bidder must provide precise geographical information about the routing of the circuits end-to-end specifying landing points, including their GPS coordinates, and transatlantic cables used, as well as fibre routing maps showing street level details for all terrestrial segments.

When possible, the Bidder should provide for each item described in the Service Requirements two diverse end-to-end physical circuits specifying the cost for each of the circuits.

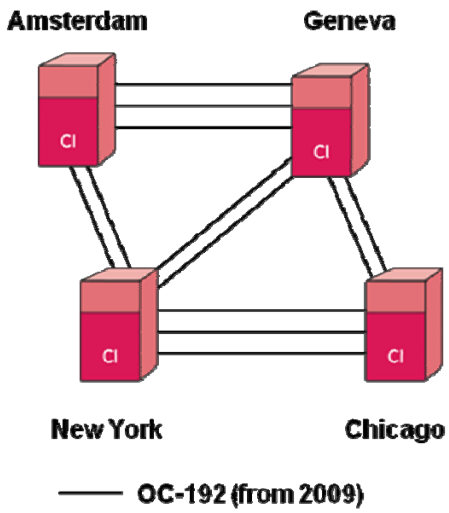


Figure 2: Planned topology for 2009.

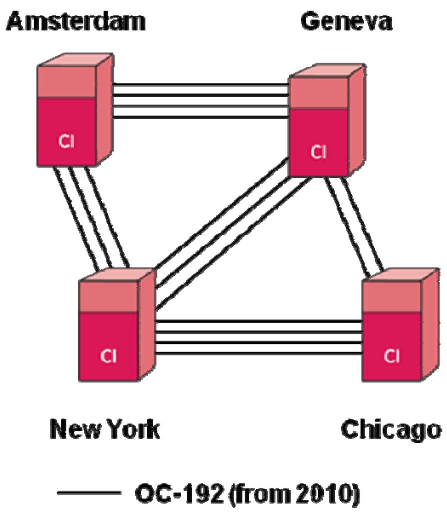


Figure 3: Planned topology for 2010.

1. *Service Reliability and Protection*

Given that the 10 Gb/s circuits will be delivered as un-protected wavelengths, the Caltech engineering team will preferentially select circuits provisioned over diverse end-to-end physical paths, in order to maximize the chance that at least one circuit to each point-of-presence is available at any time, and hence to ensure non-stop operation of the network connecting the US Tier1 centers and CERN. We emphasize that circuits that share the same conduit at any point cannot be considered to be diverse. The automatic reconfiguration of circuit termination equipment owned by Caltech should guarantee that at least two 10 Gb/s paths between CERN and the USA are available nearly 100 % of the time.

2. Service Delivery

A. European End-Points

(1) CERN (Geneva) demarcation point

CERN is located across the Franco-Swiss border, therefore the service in Europe can be terminated on either side of the border (building 513, Telecom Room, 513-R-012):

Geneva: (CERN, CH-1211 GENEVA 23 - SWITZERLAND)

Telephone number: +41 227675801

Prevessin: (CERN, F-01631 CERN CEDEX - FRANCE)

(2) SARA (Amsterdam) connection information

SARA Computing and Networking Services
Kruislaan 415
1098 Sj Amsterdam
Netherlands

Telephone number: +31 206683167

(3) DANTE (London) connection information

Telecity 2
8-9 Harbour Exchange Square
London E14 9GB
United Kingdom

Telephone number: +44 207 510 0400

(4) DANTE (Paris) connection information

InterXion-1
Batiment 260
rez-de-chaussée
45 avenue Victor Hugo
93534 Aubervilliers Cedex
France

Telephone number: +33 1 53 56 36 19

B. USA End-Points

(1) MAN-LAN(New-York) demarcation point

The MAN-LAN itself is located in the collocation space operated by the New York State Education and Research Network (NYSERNet).

NYSERNet, Inc.
24th Floor
32 Sixth Ave.
New York, NY 10013

Telephone number: +1 2122263213

Note that NYSERNet is able to provide dark fibers to some other locations in NYC including the carrier hotels at 60 Hudson St. & 111 8th Ave.

(2) STAR LIGHT (Chicago) demarcation point

2nd Floor Communications Center
Abbott Hall
710 North Lake Shore Drive
Chicago, Illinois 60611, USA

Telephone number: +1 3125034254

Access to STARLIGHT is available through a number of Telecom Operators and dark fiber providers as documented at the following address:

http://www.startap.net/starlight/CONNECT/connectCarrierInfo.html

Technical Requirements

A. SONET/SDH circuit specifications

Circuits are "un-protected, fully transparent, 10 Gbit/s (i.e. OC-192/STM-64 with SONET/SDH framing) wavelengths".

At the 4 points of presence operated by Caltech (GVA, NY, CHI, AMS) the circuits will be connected to Ciena CoreDirector/CI OC-192 interfaces. The interface type is ITU/Bellcore I.64-2/SR2 (1550nm). The circuits are required to support non-standard concatenation, and advanced SONET/SDH features such as VCAT/LCAS, which are being used by

Caltech. At the DANTE PoPs (LON/PAR), the circuits will be patched through to GEANT2 circuits, terminating in GVA (on the Ciena CD/CI).

Given the “research and development” nature of some of the circuits, we may for periods of time connect the circuits to Force 10 switches equipped with 10 GE WAN/PHY[22] interfaces[23].

B. *Performance*

RTT[24], measured from our service entry point on the system provided by the Bidder, must not deviate significantly, i.e. less than 10%, from what is customarily acceptable in the industry. This can and will be measured using ICMP[25] echo requests.

For information, on the existing 10 Gbit/s service that Caltech has in operation, the measured RTT is approximately 140ms between Geneva and Chicago and 100ms between Geneva and New-York. Caltech considers this as acceptable.

The end to end bit error rates must be no worse than 10^{-15}, however given the technical specifications of most DWDM equipment we can tolerate temporary degradation up to 10^{-12} for short periods of time to be mutually agreed during the acceptance tests.

C. Operational Requirements

(1) Availability

The service shall be available 24 hours a day and 7 days a week, throughout the Agreement period. Caltech is entitled to claim compensation if the service is unavailable for more than 7.2 hours during one calendar month as detailed in Attachment C, Section 3.

(2) Preventive maintenance

Preventive maintenance can be arranged at 5 working days notice to take place during our preferred maintenance window (the Caltech preferred maintenance window is during early morning hours, typically between 3am and 7am local Geneva time, on weekdays). Different maintenance windows must be used for each of the circuits provided, so that no two of the provided circuits are taken out of service simultaneously.

Due to the nature of the application (high-speed, scheduled data transfers with deadline requirements), Caltech cannot schedule

[22] A white paper about the WAN/PHY technology is available at http://www.force10networks.com/products/w_oc192wan1.asp

[23] Interface specification are available at http://www.force10networks.com/products/lce12e6_4p10ge.asp

[24]Round Trip Time

[25] Internet Control Message Protocol

transfers on circuits which are inside a maintenance window. For this reason, the announcement of a maintenance window shall be precise, its duration kept to a minimum, and not to exceed 6 hours. For longer maintenance window duration, the time above 6 hours will be counted as 20% unavailability), in addition to the real experienced downtime. E.g. an eight hours window in which the link was down for 1 hour, will be counted as 60+24=84 minutes unavailability.

In any case, the Bidder should be aware that downtime, even during a maintenance window, counts as unavailability.

(3) Network management

A report providing statistics on the performance and availability of each circuit shall be provided every month. This report shall arrive at Caltech not later than the 5th working day of the month following the month that is reported about. This reporting must be done electronically, preferably in HTML format. Reports of every fault with information of cause, duration and treatment shall be provided, within 24 hours after the closing of the trouble ticket. This reporting must be done electronically, preferably in HTML format.

5. Implementation Requirements

A. Providers responsibility

The provider must carry out all installation and preparation for setting the equipment into operation, including providing all required cabling up to the demarcation point at each point of presence.

B. Acceptance test

After the provider declares the service ready for Caltech to test, the acceptance test will start. The acceptance test duration is 72 hours.

In order to pass the acceptance test the service must have performance, availability and reliability as defined in paragraphs 4.1, 5.1 and 8.2. 10 Gb/s circuits will be tested individually but the service will only be accepted when all items composing the service are made available.

In case the acceptance test fails anytime during the period of 72 hours, the circuit will be handed back to the provider to fix any problems accompanied by a report stating the problems. The provider has to declare the service ready for testing again after fixing problems and a new period of 72 hours starts. This process will be repeated until the acceptance test passes for a maximum period of 30 working days, following which the Agreement will be automatically cancelled without compensation to the provider.

6. Support Requirements

A. The Network Operations Centre (NOC) of the provider must have 24 hours/day and 7 days/week operations staff responding to emergency calls, faxes and trouble tickets arriving via Internet e-mail from our operations staff. The NOC of the provider must confirm the receipt of such a notice stating the NOC's trouble ticket number and with reference to Caltech's trouble ticket number.

B. The provider must provide procedures for dealing with unscheduled interruptions of the service, including maximum response times. Target response time to the opening of a trouble ticket by Caltech is 20 minutes maximum.

C. The NOC of the provider shall report back with an explanation of the cause of the failure and the expected duration and shall with no delay report back when the error situation has been fixed.

D. During fault situations the NOC of the provider shall report the status of the failure by e-mail (noc@uslhcnet.org) at least once every hour.

E. Key personnel in engineering and operations departments shall be named, and a project team must be formed as quickly as possible after adjudication.

ATTACHMENT B
BID SHEETS
RFP KS100

The Bidder shall indicate in the table below total and firm prices (including VAT) in US dollars (USD), not subject to revision, drawn up in accordance with the requirements stated in the RFP. The Bidder shall enter "No Bid" when the corresponding circuit(s) can not be offered. Non-recurring charges (called NRC in the table) shall include installation costs.

Caltech reserves the right to split the services among different suppliers.

Unprotected, fully transparent, 10 Gbit/s circuit (i.e. OC-192c/STM-64 with SONET/SDH framing)	Charge USD	Price for one (1) circuit	Price for two (2) circuits	Price for two (2) circuits with independent physical end-to-end paths	Price for three (3) circuits	Price for three (3) circuits with two (2) independent physical end-to-end paths
GVA-CHI	NRC					
	Yearly					
	TOTAL					
GVA – NY	NRC					
	Yearly					
	TOTAL					
GVA – AMS	NRC					
	Yearly					
	TOTAL					
AMS – NY	NRC					
	Yearly					
	TOTAL					
NY – CHI	NRC					
	Yearly					
	TOTAL					
CHI-LON	NRC					
	Yearly					
	TOTAL					
CHI-PAR	NRC					
	Yearly					
	TOTAL					

NY-PAR	NRC					
	Yearly					
	TOTAL					
NY-LON	NRC					
	Yearly					
	TOTAL					

ATTACHMENT C
TERMS OF AGREEMENT
RFP KS100

The terms detailed in Attachments C, D, and E (to follow) will become a material part of any Agreement or purchase order (Agreement) resulting from this RFP.

Coverage

The Agreement shall for the whole Agreement period cover the bandwidth agreed by the parties and all bandwidth upgrades offered by the Bidder in its proposal.

1. Definitions

A. Faults and disturbances

A fault exists when data cannot flow between the two Caltech's termination points of a circuit for a continuous period of more than 180 seconds.

When the data flow is degraded by errors, it shall be considered as a fault when more than 1 out of 100 packets are damaged over a period of 180 seconds. Every such period shall be counted as a fault with the duration of 180 seconds.

Disturbance exists when data cannot flow between the termination points as defined above for fault, for a continuous period of more than 20 seconds. More than one (1) disturbance within 180 seconds shall be considered as a fault and shall be counted as a fault with the duration of 180 seconds.

A link will be considered to be "flapping" (changing link status from up to down) when logs of Caltech's termination equipment report a circuit went down and up more than two (2) times within 180 seconds. A link flapping should be counted as a fault with the duration of 180 seconds, for each transition in status from up to down.

In addition, the performance of the service, as defined in Attachment A, Section 4B must be available during the whole duration of the contract.

B. A circuit is considered to be unavailable when it is faulty as defined in section 2A (above).

The total unavailability time is from the opening of a trouble ticket by either party or from the time Caltech reports a fault until the trouble ticket is cleared.

Availability calculation

The availability of the service is defined as follows:

$$\%Availability = \frac{Total\_Hours - Unavailability\_Hours}{Total\_Hours}$$

Availability is measured by the provider and reported monthly. In case of discrepancy between the availability measured by the provider and Caltech the parties will make serious efforts to explain such differences and come to an agreement.

C. Duration of Agreement

The duration of the Agreement shall be three (3) years with an initial duration of 12 months. At Caltech's request only, the 12 month Agreement may be renewed twice, under the same or more favorable pricing conditions. However, given the possibility of future downward pricing-trends in the Telecommunications market for unprotected 10Gbit/s wavelengths, Caltech plans to renegotiate the Agreement every year in order to obtain the best price conditions, and to progressively upgrade our network by adding additional 10Gbit/s circuits (see Attachment A, Section 2.).

D. Applicable Law

The application and interpretation of the terms of the Agreement shall in all respects be governed by the laws of the State of California.

Difficulties arising from the interpretation and application of the Agreement shall be solved by amicable settlement the Parties; failing such amicable settlement litigation will be settled in Los Angeles, California, USA.

2. Compensation

A. Compensation for Late Delivery

If the provider fails to deliver any item of the service on the agreed schedule Caltech is entitled to claim compensation. The compensation will start after one week of late delivery for the item of the service. The amount of compensation per item per delayed week will be set to 5% of annual amount of the Agreement of the specific item. The compensation shall be limited to a maximum of 20% of the total amount of the Agreement for the specific item. Without prejudice to other remedies, which it may have under the Agreement or otherwise, Caltech shall have the right to cancel the Agreement without prior notice when the maximum amount is reached.

B. Compensation for Service Unavailability/Service Level Agreement (SLA)

Caltech is entitled to claim compensation for any circuit of the service that is unavailable for more than 7.2 hours during one calendar month according to the table below:

Total Monthly Availability A in %	Circuit Unavailability (hours)	% Credit of Monthly Service Charge
99% ≤A	0 – 7.2 hours	N/A
98.5% ≤A < 99%	7.3 – 10.8 hours	5%
98% ≤A < 98.5%	10.9 - 14.4 hours	10%
95 ≤A < 98%	14.4 - 36 hours	25 %
90 ≤A < 95%	36 – 72 hours	40 %
80 ≤A < 90%	72 – 144 hours	60 %
A < 80 %	144 hours	100%

C. Non-Fulfillment

If the provider fails to deliver all items of the specified service, Caltech is entitled to terminate the Agreement with immediate effect, without penalty or other claims from the provider.

In case the monthly availability of the 10Gbit/s circuits has been below 98 % for more than two (2) consecutive months after the date of the acceptance, Caltech is entitled to terminate the Agreement with immediate effect, without penalty or other claims from the provider.

ATTACHMENT D
SECTION 1 - COMMERCIAL ITEMS OR SERVICES CONTRACT GENERAL PROVISIONS RFP KS100

Terms and Conditions referenced above can be found on:

http://procurement.caltech.edu/purchasing/Ts&Cs/CommericalItemsorServices.pdf

ATTACHMENT E

SECTION 2 - GOVERNMENT FUNDED GRANT PROVISIONS

RFP KS100

Terms and Conditions referenced above can be found on:

http://procurement.caltech.edu/purchasing/Ts&Cs/GovntFundedGrantProvisions.pdf

ATTACHMENT F
ACCEPTANCE OF TERMS AND CONDITIONS
RFP KS100

The following terms, included in detail in the Attachments referenced below will become a material part of any Agreement resulting from this RFP.

Bidder agrees to the following terms in their entirety

ATTACHMENT C - TERMS OF AGREEMENT

ATTACHMENT D - SECTION 1 - COMMERCIAL ITEMS OR SERVICES AGREEMENT GENERAL PROVISIONS Rev 0906

ATTACHMENT E - SECTION 2 - GOVERNMENT FUNDED GRANT PROVISIONS RMar03

Authorized Signature & Title

Bidder takes the following exceptions (***please list on separate sheet***)

Authorized Signature & Title