THE DESIGN AND IMPLEMENTATION OF THE TRANSATLANTIC MISSION-ORIENTED PRODUCTION AND EXPERIMENTAL NETWORKS

Harvey Newman[1], Dimitri Bourilkov[2], Julian Bunn[1], Richard Cavanaugh[2], Iosif Legrand[1], Steven Low[1], Shawn McKee[3], Dan Nae[1], Sylvain Ravot[1], Conrad Steenberg[1], Xun Su[1], Michael Thomas[1], Frank van Lingen[1], Yang Xia[1]

[1]*California Institute of Technology*
{newman,conrad,xsu,thomas}@hep.caltech.edu
{ julian.bunn, slow, fvlingen,yxia}@caltech.edu
{iosif.legrand,dan.nae,sylvain.ravot}@cern.ch

[2]*University of Florida*
{bourilkov,cavanaug}@phys.ufl.edu

[3]*University of Michigan*
smckee@umich.edu

Abstract

In this paper we present the design and implementation of the mission-oriented USLHCNet for HEP research community and the UltraLight network testbed. The design philosophy for these networks is to help meet the data-intensive computing challenges of the next generation of particle physics experiments with a comprehensive, network-focused approach. Instead of treating the network as a static, unchanging and unmanaged set of inter-computer links, we are developing and using it as a dynamic, configurable, and closely monitored resource that is managed from end-to-end. In this paper we will present our work in the various areas of the project including infrastructure construction, protocol research and application development. Our goal is to construct a next-generation global system that is able to meet the data processing, distribution, access and analysis needs of the particle physics community.

I. Introduction

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 adoption of the grid "hierarchy" concept of Tier0, Tier1 and Tier2 centers (developed at Caltech in 1999) as the basis of the Computing Models of the major HEP experiments, along with the rapid development of affordable network technologies that support multiple high-bandwidth "wavelengths" on an optical fiber-pair, and advances by physicists working with network engineers and computer scientists in achieving multi-Gigabit per second throughput over long distances, are all accelerating HEP's large scale use and dependence upon long-range networks. This is particularly apparent in the ongoing series of "service challenges" involving increasingly large data transfers, marking the ramp-up of operations of the Tiered centers to the start of data-taking at the LHC.

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 10 Gbps backbone 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 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.

[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

Proceedings of the Second IEEE International
Conference on e-Science and Grid Computing (e-Science'06)
0-7695-2734-5/06 $20.00 © 2006 IEEE

In response to the significant challenges presented by data intensive e-science disciplines such as HEP, the Grid-based infrastructures developed by collaborations in the US, Europe and Asia such as OSG[3], Grid3[4] and EGEE[5] provide massive computing and storage resources. However, *efficient* use of these resources is hampered by the treatment of the interconnecting network as an external, passive, and largely unmanaged resource. The UltraLight project (www.ultralight.org) proposes to address this deficiency. We deployed the UltraLight hybrid packet/circuit-switched network testbed which is connected with various major research and education backbone networks, including LHCNet (www.datatag.org), National Lambda Rail (www.nlr.net), Internet2's Abilene network (abilene.internet2.edu), and StarLight (www.startap.net/startlight). Additional trans- and intercontinental wavelengths of our partner projects UltraScienceNet (http://www.csm.ornl.gov/ultranet/), Netherlight (http://www.surfnet.nl/innovatie/netherlight/), UKlight (http://www.uklight.ac.uk/), AMPATH (www.ampath.fiu.edu), and CA*Net4 (www.canarie.ca/canet4/) are used for network experiments on a part-time or scheduled basis.

In the rest of paper we will present the design and implementation of the network infrastructure, protocol test, and application development, including our experiences of the network setup, kernel building, application tuning and configuration used during the bandwidth challenge event at SC|05.

II. LHCNET: design and implementations

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.

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 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 and (c) develop new modes of monitoring and managing networks end-to-end, while incorporating their capabilities (on increasing scales) into integrated grid systems. This parallels the DOE Network Roadmap [6], where the "Production" network is accompanied by a "High Impact" network in which the next-round production capabilities are developed[7].

US LHCNet has been architected to ensure efficient and reliable use of the 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 more recently National Lambda Rail, as well as through regional and university-funded network initiatives. As described below and shown in Figure 1, 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 installing 10 Gbps connections to MANLAN.

The current topology of US LHCNet network is shown in Figure 1. The OC-192 SONET [8] transatlantic circuits terminate on each side in Force 10 E600 switches. The technology used to cross the Atlantic is 10 GE WAN/PHY. The configuration 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" [9] point-to-point connections between CERN and the US-Tier1 centers, extensive Quality of

[3] Open Science Grid: (http://www.opensciencegrid.org/)
[4] Grid3: http://www.ivdgl.org/grid2003/
[5] EGEE: (http://egee-intranet.web.cern.ch/egee-intranet/gateway.html)

[6] The DOE Science Networking Challenge: Roadmap to 2008 report is at http://www.osti.gov/bridge/product.biblio.jsp?osti_id=815539
[7] Also see the report of the DOE High-Performance Network Planning Workshop at http://www.doecollaboratory.org/meetings/hpnpw/finalreport/
[8] SONET is a standard for synchronous data transmission over fiber optic networks http://www.sonet.com/
[9] 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

Service (QoS) configuration and policy-based routing (PBR). The virtual termination points of the R&D transatlantic link are on Cisco 7609 switches, so that the US LHCNet backbone is transparent to the R&D traffic. This configuration is particularly useful in that it clearly splits production from research traffic, and allows us to protect the production traffic.

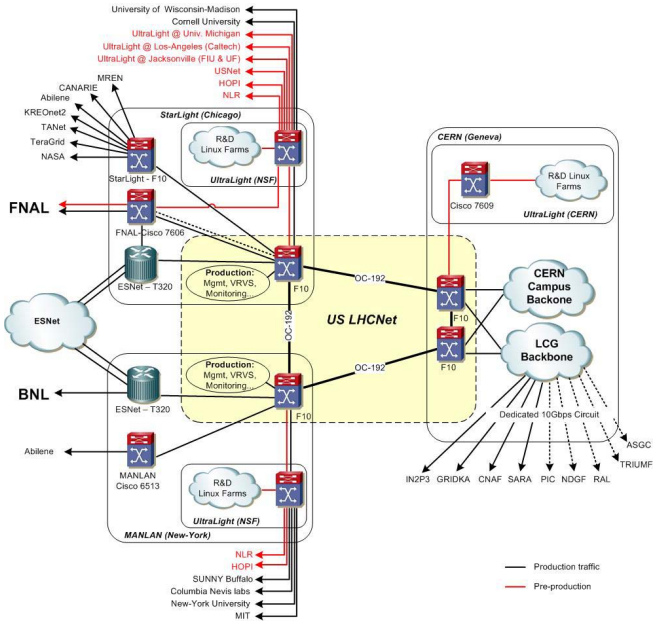

Figure 1: The current US LHCNet topology.

Since September 2004, US LHCNet has been directly connected to FNAL at 10 Gbps via the FNAL-StarLight dark fiber (note that this direct connection will be replaced in 2006 by a redundant 10 Gbps lambda provided by ESNet). The "service challenges" between CERN and FNAL have taken advantage of the 10 Gbps path to FNAL and have sustained multi-gigabit/s throughput for weeks between the two laboratories. The heavy usage of US LHCNet this year in support of these challenges is illustrated in Figure 2.

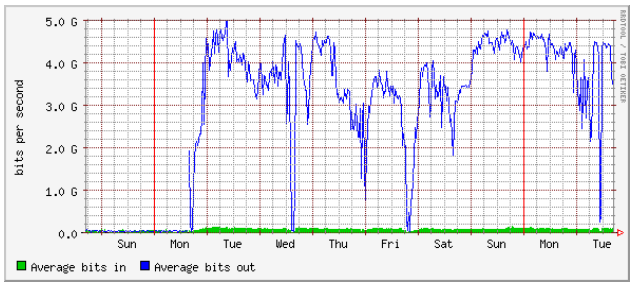

Figure 2: US LHCNet network traffic between CERN and CHICAGO during Spring 2006 service challenge.

An important element of the architecture is very high-speed connectivity to advanced optical network testbeds. In Chicago, US LHCNet is connected to the Internet2 Hybrid Optical and Packet Infrastructure (HOPI[10]), the DOE UltraScience experimental network testbed (USNet[11]) and the National Lambda Rail (NLR[12]) infrastructure. The goal is to examine, understand and design a coherent and scalable architecture for next-generation networks. US LHCNet testbed resources are made available to the GLIF[13] community via connections to NetherLight[14] (SURFnet), CANARIE and UKlight[15]. Our team is active in the GLIF organization, which brings together the world's premier research and education networking engineers, who are building an international LambdaGrid infrastructure by identifying equipment, connection requirements, and necessary engineering functions and services.

III. Ultralight: a hybrid optical testbed

To effectively manage the network resources on an end to end basis, it is essential to deploy an network monitoring systems that can both capture the current state of the network and provide a feedback mechanism to enable control actions [6][7][8]. Figure 3 shows the current topology for the experimental Ultralight network testbed. The testbed relies on a hybrid packet/circuit network infrastructure based on the NLR footprint and the USLHCNet Transatlantic links, and interconnect with other major networks such as Abilene.

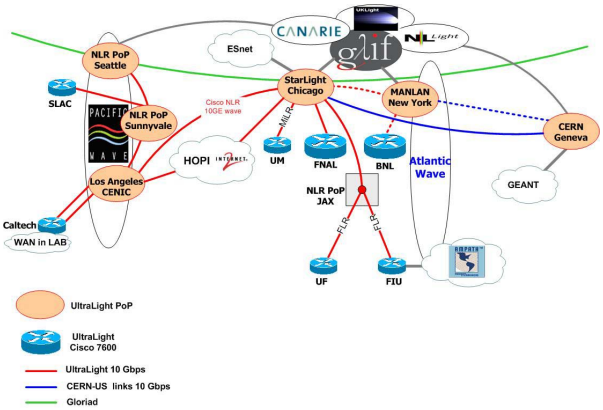

Figure 3: UltraLight topology and connections to other major networks.

[10] http://networks.internet2.edu/hopi/
[11] http://www.csm.ornl.gov/ultranet/
[12] http://www.nlr.net/
[13] http://www.glif.is/
[14] NetherLight is an advanced optical infrastructure with high speed international connectivity. http://www.surfnet.nl/innovatie/netherlight/
[15] UKlight is a national facility to support projects working on developments towards optical networks http://www.uklight.ac.uk/

In the Ultralight testbed, we have deployed and continue to develop Caltech's MonALISA system (Monitoring Agents in A Large Integrated Services Architecture)[4], which provides a distributed real-time services architecture that is suitable for this task. While its initial target field of application is networks and Grid systems supporting data processing and analysis for global high energy and nuclear physics collaborations, MonALISA is broadly applicable to many fields of data intensive science, and to the monitoring and management of major research and education networks.

Figure 4 is a snapshot of the MonALISA monitoring network for Abilene. It shows all the active nodes running MonALISA services for this particular network, discovered automatically by a graphical MonALISA client. The client can display the real time global views and connectivity, as well as the usage and load of the network. MonALISA operates in an analogous fashion for grid facilities, monitoring the load and other state parameters for each of the compute nodes as well as their interconnections.

The core of the MonALISA monitoring service is based on a set of multi-threaded engines that perform the data collection tasks in parallel, independently. The modules used for collecting different sets of information, or interfacing with other monitoring tools, are dynamically loaded and executed in independent threads. In order to reduce the load on systems running MonALISA, a dynamic pool of threads is created once, and the threads are then reused when a task assigned to a thread is completed. This allows one to run a large number of monitoring modules concurrently and independently, and to dynamically adapt to the load and the response time of the components in the system. If a monitoring task fails or hangs due to I/O errors, the other tasks are not delayed or disrupted, since they are executing in other, independent threads. A dedicated control thread is used to properly stop the threads in case of I/O errors, and to reschedule those tasks that have not been successfully completed. A priority queue is used for the tasks that need to be performed periodically.

Using a low level predicate mechanism within MonALISA, it is possible to create filters in any given processes and associate these filters with certain actions. An example of end-to-end monitoring of resources has been the integration of MonALISA and Caltech's Virtual Room Videoconference System [5]. MonALISA was adapted and deployed on the 83 VRVS reflectors situated at sites around the world, to collect information about the topology of the VRVS reflector-network, to monitor and track traffic among the reflectors, to report communication errors among the peers, and to track the number of clients and active virtual rooms. Agents within MonALISA have been developed to provide and optimize dynamic routing of the VRVS data streams. These agents acquire information about the quality of alternative connections and solve a minimum spanning tree problem to optimize data flow at the global level.

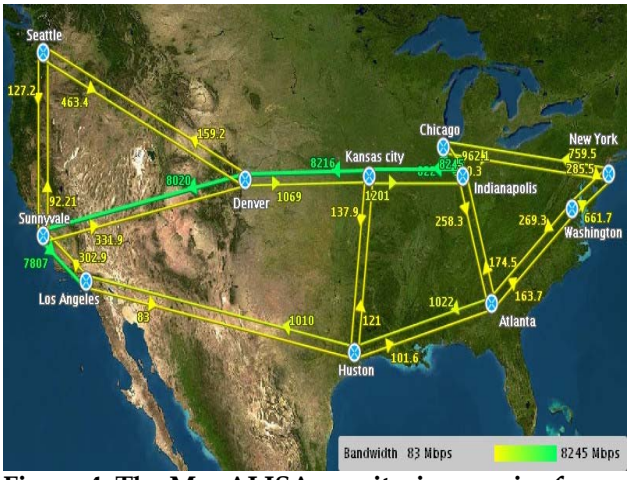

Figure 4: The MonALISA monitoring service for Abilene (with 8Gbps injected traffic)

IV. High-speed transport protocol development and wan-in-lab

A central issue in networking is how to allocate bandwidth to flows *efficiently* and *fairly*, in a decentralized manner. A recent body of work by S. Low et. al. has shown that as long as traffic sources adapt their rates to the aggregate congestion measure in their paths, they are implicitly maximizing the utility of the overall network. Maintaining high throughput in the presence of packet loss has been a significant problem for existing TCP protocols. Traditionally TCP uses packet loss as a signal to slow down, assuming the loss is due to overflowing router buffers caused by congestion. However, packets can also be lost due to channel error, such as from interference in wireless networks. In these environments TCP performs poorly due to lost packets being misinterpreted as network congestion. FAST on the other hand uses delay as the congestion signal, rather than packet loss as is case for TCP RENO. This allows FAST TCP to stabilize at a steady throughput, and to reach equilibrium quickly. As a result, FAST avoids having long queues of waiting packets accumulate which lead to buffer overflows and additional packet loss, as inevitably occurs with loss-based schemes [9][10]. The decoupling of loss and congestion in FAST facilitates the development of far more efficient loss recovery algorithms. Figure 5 shows a comparison between the achievable throughput of FAST TCP and RENO TCP [11], in the presence of packet loss.

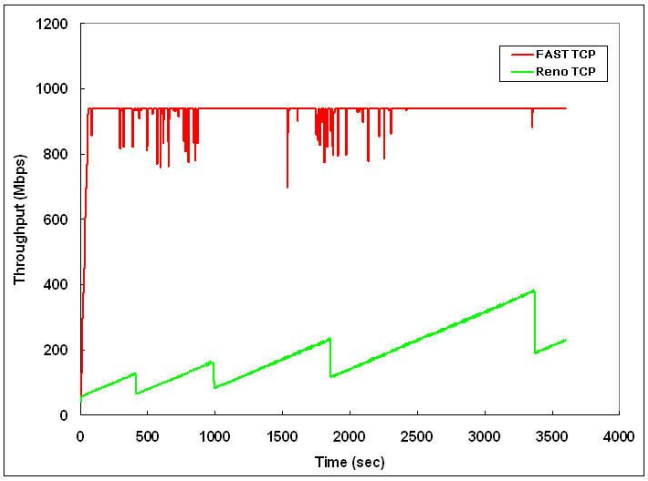

Figure 5: The throughput of FAST flows compared with RENO, in the presence of packet loss.

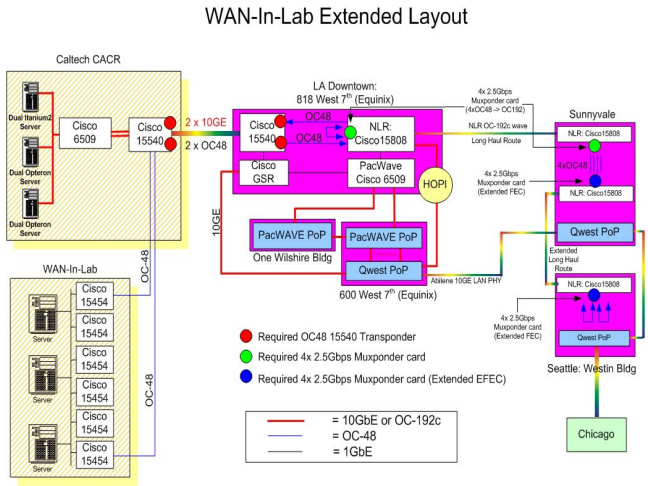

Figure 6: WAN-in-Lab extension: to LA-Sunnyvale-Seattle-Chicago.

WAN in Lab[16] provides the controlled in-lab experimental facility that is critically needed to complement our theoretical understanding, simulation studies, and long range field tests of ultra-scale transport protocols such as FAST TCP. It is literally a wide-area-network – it includes 24,000 kilometers of fibers, optical amplifiers, dispersion compensation modules, WDM (Wavelength Division Multiplexing) gear, optical switches, routers, and servers - but it is housed in a single laboratory at Caltech. By connecting it to the Sunnyvale and Seattle GigaPoPs (see Figure 6) and thus becoming an integral part of the Ultralight, we can extend the round-trip time of an end-to-end connection between a server in WAN-in-Lab and one in a global production network to more than 300ms. This is larger but of the same scale as the largest round-trip times we expect in the "real" networks.

WAN-in-Lab also will be directly connected to the international research and production networks, such as Abilene, and USLHCNet. The integrated infrastructure will provide a uniform environment for the development, testing, demonstration and deployment of new protocols, that facilitates the transition among these stages, and from laboratory to the market place. It will also allow us to study the interaction of new protocols with existing protocols, in a realistic production environment, and without the need to modify any equipment not in the Lab. This not only minimizes the disruption to other groups on the shared network, but also offers a unique environment to explore issues in incremental deployment.

V. Application services development

Within the scope of the Ultralight project we explore how to best make available the end-to-end managed network resource to the globally distributed e-science applications. As an example UltraLight is extending the Grid Analysis Environment (GAE) [12], an application level Service Oriented Architecture (SOA) supporting end-to-end (physics) analysis, to the UltraLight Analysis Environment (UAE). UAE integrates the components identified in the GAE and exposes the network as a managed resource. UAE will interact with monitor applications, replicate data, schedule jobs, and find optimal network connections in an autonomous manner that would result in a self organizing Grid that minimizes single point of failures, in which thousands of users are able to get fair access to a limited set of distributed resources of the Grid in a responsive manner. Many of the Web Service implemented within the UAE will be made available through and developed in CLARENS[13] and MonALISA, that offers several additional features: X.509 Certificate based authentication when establishing a connection, access control on Web Services, remote file access (and access control on files), discovery of services and software, virtual organization management, high performance (measured 1400 calls/second), role management, and support for multiple protocols (XML-RPC, SOAP, Java RMI, JSON-RPC).

VI. SC|05 bandwidth challenge

Using the Ultralight testbed, the team from C**altech-CERN-Florida-FNAL-Michigan-Manchester-SLAC** participated and won the SC05 bandwidth challenge (BWC) with an overall bandwidth usage exceeding **131 Gbps**. This number is an average measured by the jury over a period of 15 minutes on 17 of the 22 10 Gbps waves being used by the team entry. The team is a collaboration of institutes including Caltech, University of Michigan,

[16] http://wil.cs.caltech.edu/

SLAC and FNAL, CERN, and Manchester. Note that the bandwidth challenge involves not only networks, but also servers on the receiving and sending side that are connected via the wide area network. In the Caltech booth at SC05 4 server racks were placed especially for this purpose. A detailed server and router configuration is as shown in Figure 7.

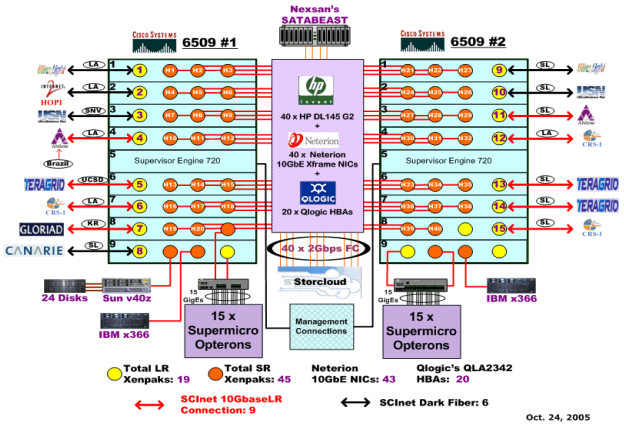

Figure 7: Server/switches at SC05 Caltech booth.

Our entry used real-world applications where real physics data was transferred based on *ROOT*[17] file, a format frequently used by physicists. As such the bandwidth result, and lessons learned from it, will have some lasting benefits for transfer, and management of large amounts of scientific data. Several different protocols were used for transferring data, including *bbcp*[18], *xrootd*[19], and *gridftp*[20]. Part of the data was transferred between remote *SRM*[21] *dcache*[22] deployments, and ones deployed at the show floor using gridftp. The extraordinary achieved bandwidth usage was made possible in part through the use of the FAST TCP protocol.

Figure 8 illustrates the traffic flows to/from Caltech booth that were involved in the bandwidth challenge, as well as the array of research and education backbone networks that are enlisted to support this effort (Ultralight, USN, Pacific Wave, Internet2, TeraGrid, NLR, GLORIAD). Figure 9 shows the traffic flows and network paths used by the SLAC/Fermi Lab booth. This includes four waves to FNAL via StarLight, two to SLAC via ESnet, and one to UKLight. Our Brazilian partners involved in the exercise, namely UNESP and UERJ, set a Brazilian research and education network speed record of 2Gbps from Brazil to US (and 1Gbps from US to Brazil) over the WHREN-LILA link connecting AMPATH[23] at Miami and ANSP[24] at Sao Paulo. Our international partners also include KEK Japan and KNU Korea, which by utilizing JGN2 and GLORIAD networks was able to transmit 6Gbps to the SC05 (1.5Gbps on the reversed direction).

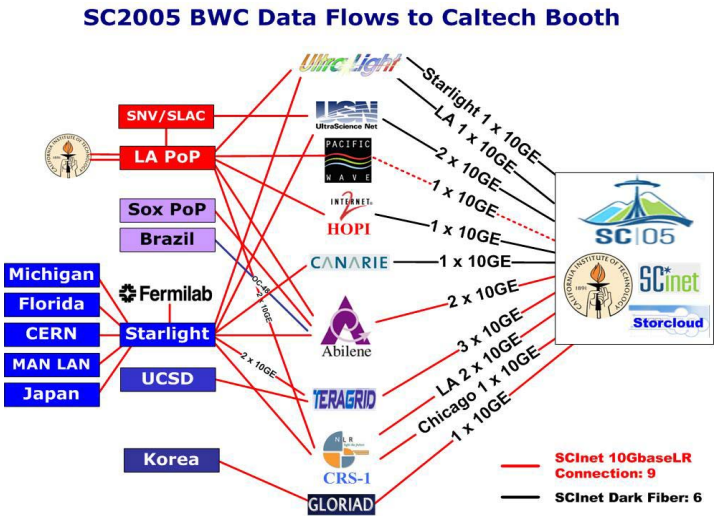

Figure 8: The traffic flows to/from Caltech booth in the SC|05 Bandwidth Challenge.

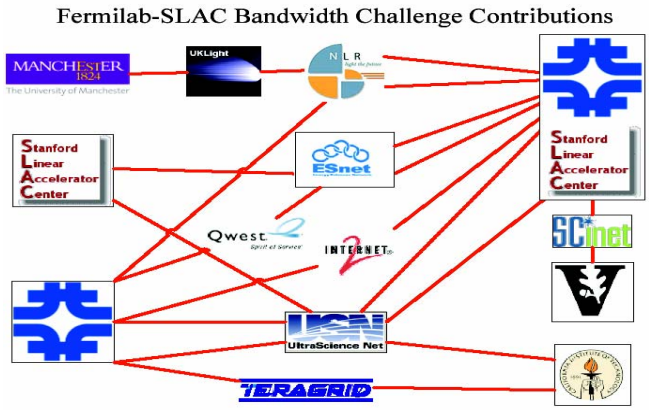

Figure 9: The traffic flows to/from SLAC/Fermi Lab booth in the SC05 Bandwidth Challenge.

The Bandwidth Challenge is an interesting benchmark of what is possible with high performance networking. It is especially important for the LHC experiments, which will generate Petabytes to Exabytes of data per year to be analyzed by physicists around the world. In the near future most of ATLAS and CMS Tier-2's and even some Tier-3's will have 10 Gigabit connections and will want to be able

[17] http://root.cern.ch
[18] http://www.slac.stanford.edu/~abh/bbcp/
[19] http://xrootd.slac.stanford.edu/
[20] http://www.globus.org/grid_software/data/gridftp.php
[21] http://sbm.lbl.gov/srm-wg
[22] http://www.dcache.org
[23] http://www.ampath.fiu.edu
[24] http://nara.org.br

to utilize them effectively. Activities like calibration and alignment of detectors for these experiments will rely upon being able to quickly move large amounts of data from CERN (the place where the LHC resides and Tier 0) to the sites responsible for that data's reduction. Part of how these huge data transfers take place is depicted in the LHC *data hierarchy scheme*[25], which will be augmented with many transfers between Tier-2's. The Bandwidth Challenge demonstrates what is possible with current networks when a focused effort is undertaken and will prepare us for enormous amounts of data that will generate increasingly more *network traffic*[26]. The result of this challenge is part of the larger picture for LHC physics. We need to continue to make progress, especially "end-to-end". Efforts like this are just a step on the way to providing a robust high performance infrastructure for LHC science and other global data intensive science collaborations.

Figure 10 shows measurements of individual and aggregate waves as measured by MonALISA during BWC. In about 3 hours an aggregate of **142.8 TB (Terabyte)** was transferred, with sustained transfer rates ranging from **90 Gbps** to **150 Gbps** and a measured peak of **151 Gbps**. Figure 11 shows the aggregate data volume transferred during the bandwidth challenge. For the whole day (24 hours) on which the bandwidth challenge took place approximately **475 TB** where transferred. This number (475 TB) is lower than what the team was capable of, based on the estimation by extrapolating our BWC throughput, as we did not always have exclusive access to waves, outside the bandwidth challenge time slot. Multiplying the 142.8 TB observed by 8 corresponds to approximately **1.1 PB (Petabyte) per day**. This is equivalent to approximately 4 (DVD) movies per second, assuming an average size of 3.5 GB per movie. On a related note, during the bandwidth challenge we also used StorCloud, a high performance storage facility set up for use by the SC05. Using bbcp we transferred physics data from 20 nodes in Caltech to StorCloud at a rate around 320~350MByte/s for each node and in some cases it reached as high as 380MByte/s for some nodes. The aggregate rate for 20 nodes was over 6GByte/s.

The week-long exercise at the SC05 allowed us to access the IT challenges of the next generation e-science at the HEP Frontier, this includes (1) Petabyte-scale datasets; (2) Tens of national and transoceanic links at 10 Gbps (and up) (3) 100+ Gbps aggregate data transport sustained for hours. The team set the scale and learned to gauge the difficulty of the global networks and transport systems required for the LHC mission through an intensive process of setting up, shaking down and successfully running the system in < 1 week.

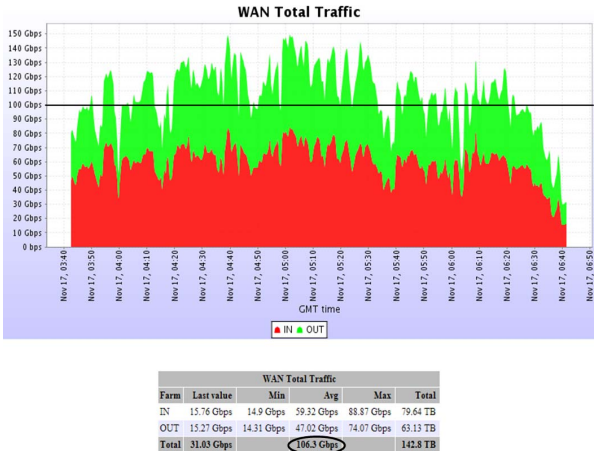

Figure 10: Three hour snapshot of total bandwidth usage, with an average throughput of more than 100 Gbps.

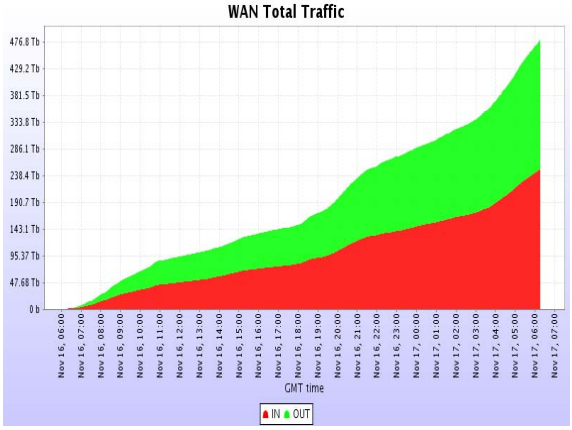

Figure 11: Total WAN traffic volume during SC05 bandwidth challenge for a 24 hour period.

VII. Integration with internet2 NewNet concept

A recent significant development in the US R&E network is Internet2's NewNet project, an innovative and cost-effective hybrid optical and packet network. As an extension of the Internet2's Abilene backbone network, NewNet is designed to provide production services as well as serve as a platform for the development of new networking ideas and techniques. NewNet will be deployed nationally over 13,000 miles of dedicated fiber, providing complete control of the optical layer and highly granular lightpath services that can be provisioned dynamically. It will provide short-term and long-term waves, as well as on demand or advanced reservation "lightpath" scheduling. The NewNet IP network, corresponding to the current Abilene network, will be built on the optical network using advanced optical ROADM

[25] http://ultralight.caltech.edu/web-site/sc05/pictures/misc/data_grid_hierarchy.jpg

[26] http://ultralight.caltech.edu/website/sc05/pictures/misc/traffic_trends.jpg

and long-haul DWDM devices. The basic connectivity is expected to include two 10 Gbps waves, one for IP and one for point-to-point optical services.

From LHCNet/Ultralight's point of view, the NewNet present ample opportunities for synergistic development on many aspects of resource management, infrastructure development and service delivery. In particular as a potential user and peer network we believe the system approach taken by LHCNet/Ultralight for end-to-end management of network resource as part of the overall Grid system is indispensable for the effective use of the NewNet resources. The monitoring and dynamic provisioning schemes developed in the MonALISA/VINCI project can be integrated with the MPLS/GMPLS tools developed in NewNet's HOPI project, enabling truly end-to-end integration of the applications, host systems, and network devices. Moreover, in LHCNet/Ultralight we are experimenting with the use of VCAT/LCAT/GFP capable optical provisioning platforms and photonic switches. This allows the development of a bandwidth-efficient and cost-effective Ethernet-based provisioning strategy. As indicated in the NewNet's mission statement an important feature of its services is the granular pre-scheduled or on-demand lightpath provisioning capability. We believe this fits well with the goals of the LHCNet/Ultralight, and as a consequence our work in the lightpath provisioning that can be beneficial to other users of the NewNet.

VIII. Conclusion

The LHCNet and UltraLight projects mark the entry into a new era of global real time responsive systems where all three sets of resources - computational, storage and network - are monitored and tracked to provide efficient, policy-based resource usage, and optimized distributed system performance on a global scale. In addition to building highly advanced network infrastructure, we also develop sophisticated applications built on top of advanced network protocols such as FAST, and autonomous service-oriented frameworks such as MonALISA. By consolidating with other emerging data-intensive Grid systems, LHCNet/UltraLight will drive the next generation of Grid developments, and support new modes of collaborative work. This paves the way for more flexible, efficient sharing of data by scientists in many countries which operate in a resource constraint environment, and could be a key factor enabling the next round of discoveries soon to be explored at the LHC.

Acknowledgements

This work is partly supported by the Department of Energy grants: DE-FC02-01ER25459, DE-FG02-92-ER40701, DE-AC02-76CH03000 (Particle Physics DataGrid project), DE-FG02-04ER-25613 (Lambda Station project), DE-AC02-76SF00515 (Terapaths project) and DE-FG02-05ER41359 (LHCnet project), and by the National Science Foundation grants: PHY-0122557, PHY-0427110 (Ultralight project) , ANI-0113425, EIA-0303620(WAN in Lab project). We acknowledge the generous support of our sponsors and contributors (http://ultralight.org/web-site/sc05/html/contributors.html). Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the DOE or the NSF.

References

[1] Cheng Jin, David X. Wei and Steven H. Low. "*FAST TCP: motivation, architecture, algorithms, performance",* Proceedings of the IEEE Infocom, Hong Kong, March 2004. (see also: http://netlab.caltech.edu/FAST)

[2] C. Jin, D. X. Wei, S. H. Low, G. Buhrmaster, J. Bunn, D. H. Choe, R. L. A. Cottrell, J. C. Doyle, W. Feng, O. Martin, H. Newman, F. Paganini, S. Ravot, S. Singh. "*FAST TCP: From Theory to Experiments*", IEEE Network, 19(1):4-11, January/February 2005.

[3] X. Xiao, L. M. Ni, "Internet QoS: A Big Picture", in IEEE Network, 13(2):8-18, March, 1999

[4] H.B. Newman, I.C. Legrand, P. Galvez, R. Voicu, C. Cirstoiu "*Monalisa: A Distributed Monitoring Service Architecture.*" In proceedings of Computing for High Energy Physics (CHEP), Paper ID: MOET001, La Jolla, California, June 2003. (see also: http://monalisa.caltech.edu/)

[5] D. Adamczyk, G. Denis, J. Fernandes, P. Farkas, P. Galvez, D. Lattka, I. Legrand, H. Newman, J. Sucik, K. Wei,"*A Globally Distributed Real Time Infrastructure for World Wide Collaborations*", In proceedings of Computing for High Energy Physics (CHEP), Paper ID:88, Interlaken, Switzerland, September 2004.

[6] M. L. Massie, B. N. Chun, D.E. Culler, "*The Ganglia Distributed Monitoring System: Design, Implementation, and Experience*", Parallel Computing 30(7):817-840, July 2004.

[7] A. Cooke, A. Gray, L. Ma, et al. "*R-GMA: an Information Integration System for Grid Monitoring*", In proceedings of the 11th International Conference on Cooperative Information Systems (CoopIS 2003) pp 462-481, Catania, Italy, November 2003.

[8] S. Andreozzi, N. De Bortoli, S. Fantinel, A. Ghiselli, G. Rubini, G. Tortone, M. Vistoli, "*GridICE: a Monitoring Service for Grid Systems*", Preprint. to appear in Future Generation Computer Systems journal, Elsevier.

[9] J. Wang, D. X. Wei and S. H. Low. "*Modeling and stability of FAST TCP."* In proceedings of the IEEE Infocom, Miami, Florida,, March 2005.

[10] F. P. Kelly, A.K. Maulloo and D. K. H. Tan. "*Rate Control in Communication Networks: Shadow Prices, Proportional Fairness and stability*", Journal of the Operational Research Society 49 (1998), 237-252.

[11] W. Stevens, M. Allman, V. Paxson, *"TCP congestion Control"* RFC2581 , April 1999.

[12] F. van Lingen, J. Bunn, I. Legrand, H. Newman, C. Steenberg, M. Thomas, P. Avery, D. Bourilkov, R. Cavanaugh, L. Chitnis, M. Kulkarni, J. Uk In, A. Anjum, T. Azim "*Grid Enabled Analysis: Architecture, Prototype and Status*" in proceedings of Computing for High Energy Physics (CHEP) Interlaken, Switzerland September 2004. (see also: http://ultralight.caltech.edu/gaeweb/portal)

[13] F. van Lingen, J. Bunn, I. Legrand, H. Newman, C. Steenberg, M. Thomas, A. Anjum, T. Azim, "*The Clarens Web Service Framework for Distributed Scientific Analysis in Grid Projects"*, In proceedings of the International Conference on Parallel Processing pp 45-52, Oslo, Norway, June 14-17, 2005. (see also: http://clarens.sourceforge.net/)