

The Ultralight Network Testbed: Motivation, Architecture and Experience

Harvey Newman¹, Julian Bunn¹, Richard Cavanaugh², Iosif Legrand¹, Steven Low¹,
Shawn McKee³, Dan Nae¹, Sylvain Ravot¹, Conrad Steenberg¹, Xun Su¹, Michael
Thomas¹, Frank van Lingen¹, Yang Xia¹

¹*California Institute of Technology, United States*
{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*
cavanaugh@phys.ufl.edu

³*University of Michigan*
smckee@umich.edu

Keywords: Network Protocols, Distributed Systems, Network Monitoring, Physics

Abstract

In this paper we describe progress in the NSF-funded UltraLight project, and a recent demonstration of UltraLight technologies at SuperComputing 2005 (SC|05). The goal of the Ultralight project is to help meet the data-intensive computing challenges of the next generation of particle physics experiments with a comprehensive, network-focused approach. UltraLight adopts a new approach to networking: instead of treating it traditionally, 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. Thus we are constructing a next-generation global system that is able to meet the data processing, distribution, access and analysis needs of the particle physics community. In this paper we present the motivation for, and an overview of, the UltraLight project. We then cover early results in the various working areas of the project. The remainder of the paper describes our experiences of the UltraLight network architecture, kernel setup, application tuning and configuration used during the bandwidth challenge event at SC|05. During this Challenge, we achieved a record-breaking aggregate data rate in excess of 150 Gbps while moving physics datasets between many sites in the WAN. The exercise highlighted the benefits of UltraLight's research and development efforts that are enabling new and advanced methods of distributed scientific data analysis.

Introduction

Physicists are conducting a new round of experiments to probe the fundamental nature of matter and space-time, and to understand the composition and early history of the universe. The decade-long construction phase of the accelerator and associated LHC¹ experiments is now approaching completion. These experiments face unprecedented engineering challenges due to the volumes

¹ Large Hadron Collider (<http://lhc.web.cern.ch/lhc>)

and complexity of the data, and the need of collaboration among scientists working in the very diverse regions in the world. The massive, globally distributed datasets to be acquired by these experiments, expected to grow to the 100 Petabyte level by 2010 and rise to the Exabyte range, will require data throughputs on the order of 10-100 gigabits per second (Gbps) between sites located around the globe. In response to these challenges, the Grid-based infrastructures developed by collaborations in the US, Europe and Asia such as OSG², Grid3³ and EGEE⁴ 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.

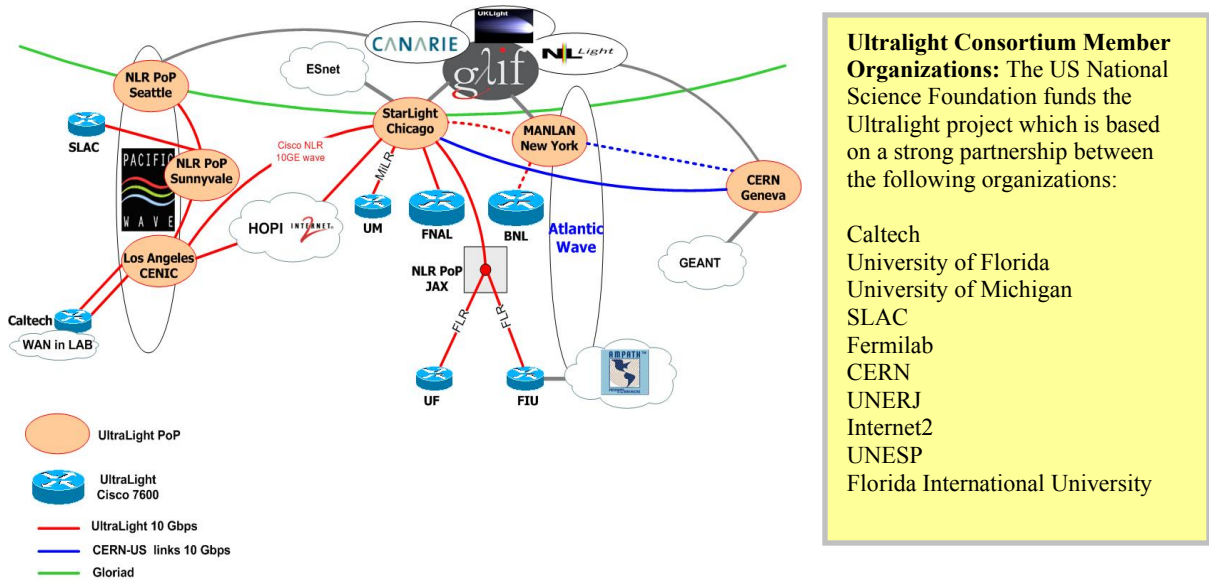


Figure 1: A schematic view of the initial UltraLight setup and connections to other major networks and WAN in Lab.

The NSF funded UltraLight project (www.ultralight.org) was proposed to address this deficiency. The consortium (see sidebar) deployed the UltraLight hybrid packet/circuit-switched network infrastructure (Figure 1) 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.startup.net/startlight). Additional trans- and intercontinental wavelengths of our partner projects UltraScience Net (<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.

The motivation for treating the network as a managed resource is based on years of prior experience with computing in HEP, where resources within a Grid (CPU, storage, network) will always be insufficient to meet the demand. This has significant implications on the overall system design. Specifically it requires the implementation of the fair-sharing policies in a resource-constrained system and agreement on the part of the users on the terms under which they may use the available resources. As an example consider the following scenario in which the HEP network

² OSG: Open Science Grid (<http://www.opensciencegrid.org/>)

³ Grid3: <http://www.ivdgl.org/grid2003/>

⁴ EGEE: Enabling Grids for E-science (<http://egee-intranet.web.cern.ch/egee-intranet/gateway.html>)

resource is used to support physics discoveries. The detector located at CERN (Tier 0) produces raw data at a rate of Petabytes per year which will be “reconstructed” locally. The reconstructed data will be stored at CERN and distributed in part to the Tier 1 centers located around the world. Tier 1 centers in turn make it available to Tier 2 centers. With these datasets located at various Tier1/2 centers hundreds of physicists will perform various types of analysis at any time, using data that may potentially be distributed over several sites. During this process certain datasets become very popular or “hot” while other datasets languish or become “cold”, over time data can change from “hot” to “cold”, or vice versa depending on what datasets physicists are interested in. The management of the most relevant datasets, making them readily available to all physicists of diverse research interests represents a major challenge on the network infrastructure and if not properly addressed can severely limit the utility of the computing resources.

Our approach of an end-to-end monitored/managed network might not be scalable to the Internet in its current form [3], especially considering the complications involving inter-domain resource coordination, economic policy issues, and the lack of scalable control-plane support etc. However, we believe the end-to-end managed network model developed by UltraLight is viable for e-science because of the relatively limited scope of the Virtual Organizations (VOs). Hence we are targeting our developments towards a large but manageable set of resources in practice, that represents a wide range of e-science projects involving dozens of VOs. Moreover, our design can be viewed as an exploration on what is potentially useful for a “clean-slate” design of the future Internet, and what is needed for a successful transition towards that design.

In the following sections we describe our work in the four focus-areas: (1) end-to-end monitoring, that provides components with real time status information of the system as a whole; (2) fundamental network protocols and tools such as FAST TCP [1] [2], and development of the WAN in Lab testbed; (3) application level services, that allow for the physics applications to effectively interact with the networking, storage and computation resources; (4) the bandwidth challenge demonstration at SC05, which not only showcased the capability of the current Ultralight infrastructure, but also highlighted the significant challenges we still must overcome to achieve the goals of the Ultralight project: deploying an advanced integrated system of network and grid services for production use.

End-to-End Monitoring

To effectively manage the network resources on an end to end basis, it is essential to deploy a network monitoring system that can both capture the current state of the network and provide a feedback mechanism to enable control actions [6][7][8]. In Ultralight, we have deployed and continue to develop Caltech’s MonALISA (Monitoring Agents in A Large Integrated Services Architecture) system[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.

MonALISA is based on a scalable Dynamic Distributed Services Architecture, and is implemented in Java using JINI and WSDL technologies. The scalability of the system derives from the use of a multi-threaded engine to host a variety of loosely coupled self-describing dynamic services, and the ability of each service to register itself and then to be discovered and used by other services or clients that require such information. The framework integrates many existing monitoring tools and procedures to collect parameters describing computational nodes,

applications and network performance. Specialized mobile agents are used in the MonALISA framework to perform global optimization tasks or help improve the operation of large distributed systems by performing supervising tasks for different applications. MonALISA is currently running around the clock monitoring several Grids and distributed applications on approximately 200 sites around 14,000 participating nodes using over 60 WAN links, and monitoring approximately 250,000 different operational parameters.



Figure 2: The MonALISA monitoring service for Abilene (with 8Gbps injected traffic)

Figure 2 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.

In this particular instance we captured a highly intensive data transfer event on June 19th, 2004 where a group of 12 disk servers in CERN concurrently sent TCP traffic via LHCNet and Abilene to their destinations in Caltech. Note that in this case MonALISA reported a throughput reaching 8.4 Gbps on the Abilene links from Chicago → Kansas City → Denver → Sunnyvale → Los Angeles.

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 (VRVS; see <http://www.vrvs.org>) [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.

High-speed transport protocol development and WAN-in-Lab

To facilitate high-throughput data movement for collaborative e-science applications, such as HEP distributed data analysis, we are deploying and testing "ultra-scale" TCP stacks, e.g., Caltech's FAST TCP, while following a systematic development path addressing issues of protocol performance and fairness. This is especially suited for the long-distance transfers of Terabyte-scale and larger datasets that are required in large collaborative e-science projects.

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 and collaborators at Caltech 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 3 shows a comparison between the achievable throughput of FAST TCP and RENO TCP [11], in the presence of packet loss.

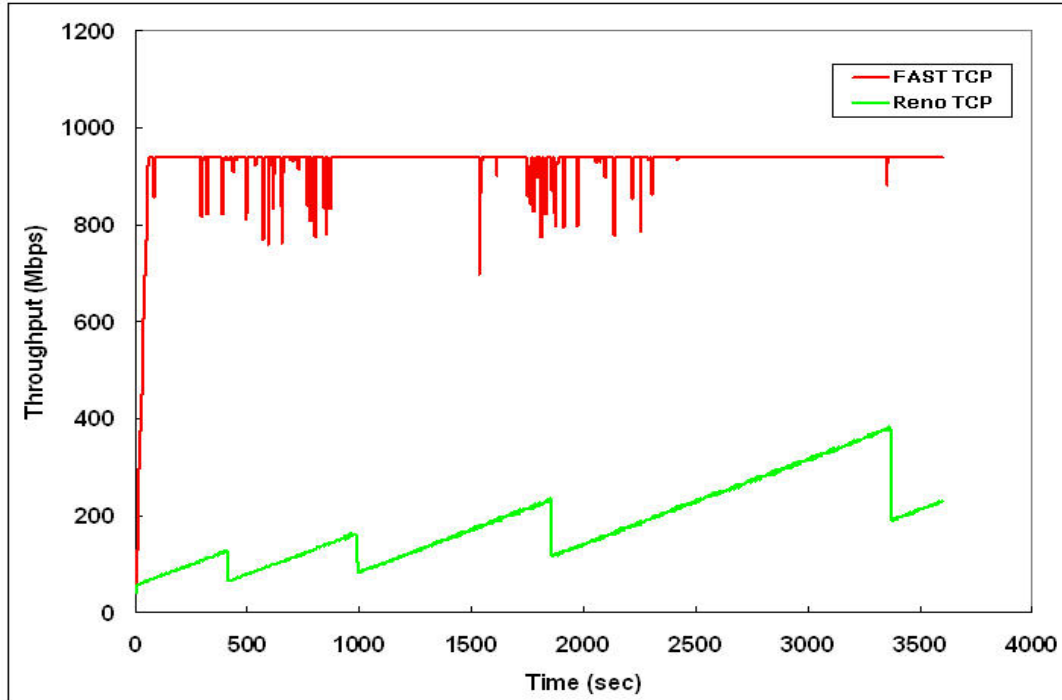


Figure 3. The throughput of FAST flows compared with RENO, in the presence of packet loss.

WAN in Lab⁵ 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 4) and thus becoming an integral part of 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 CalREN2, Abilene, and HEP's major networks. This both greatly expands the scope and usefulness of WAN-in-Lab and makes it a complementary component of this global research infrastructure. 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.

⁵ <http://wil.cs.caltech.edu/>

WAN-In-Lab Extended Layout

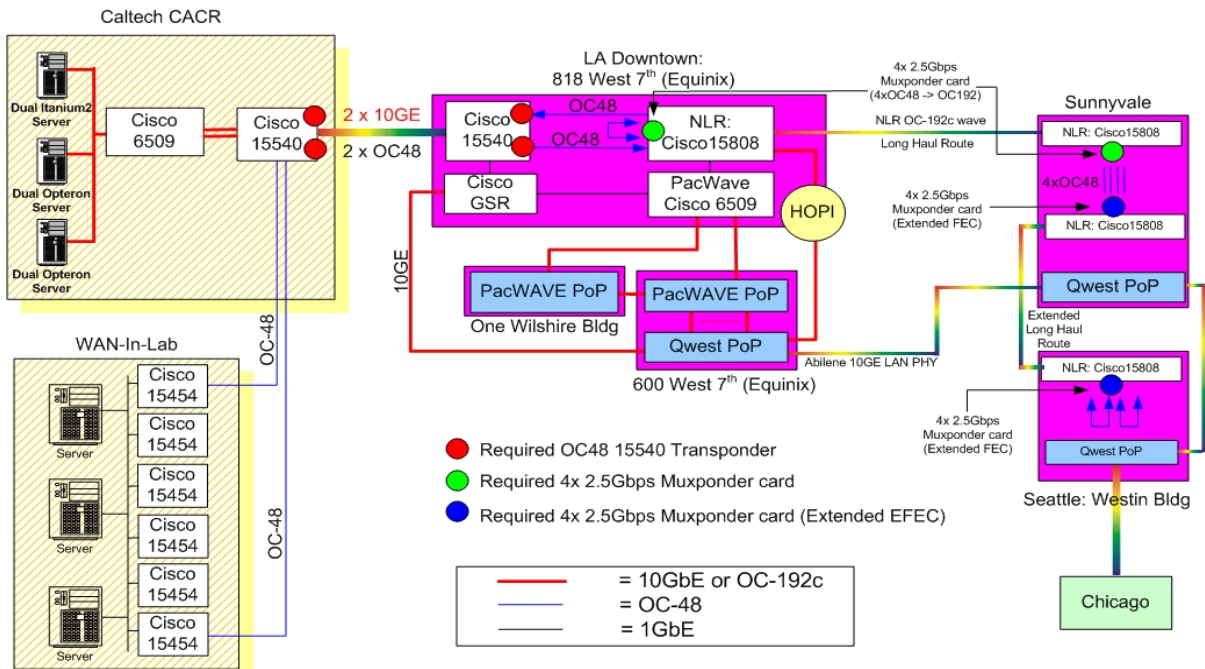


Figure 4: WAN in Lab extension: to LA-Sunnyvale-Seattle-Chicago.

Application level services development

Within the scope of 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).

Real-world demonstration: SC05 Bandwidth Challenge

Using Ultralight testbed, the team from Caltech-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, 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 5. Figure 6 shows the overall WAN topology we set up for to support the bandwidth challenge, including the various research and education backbone networks we used.

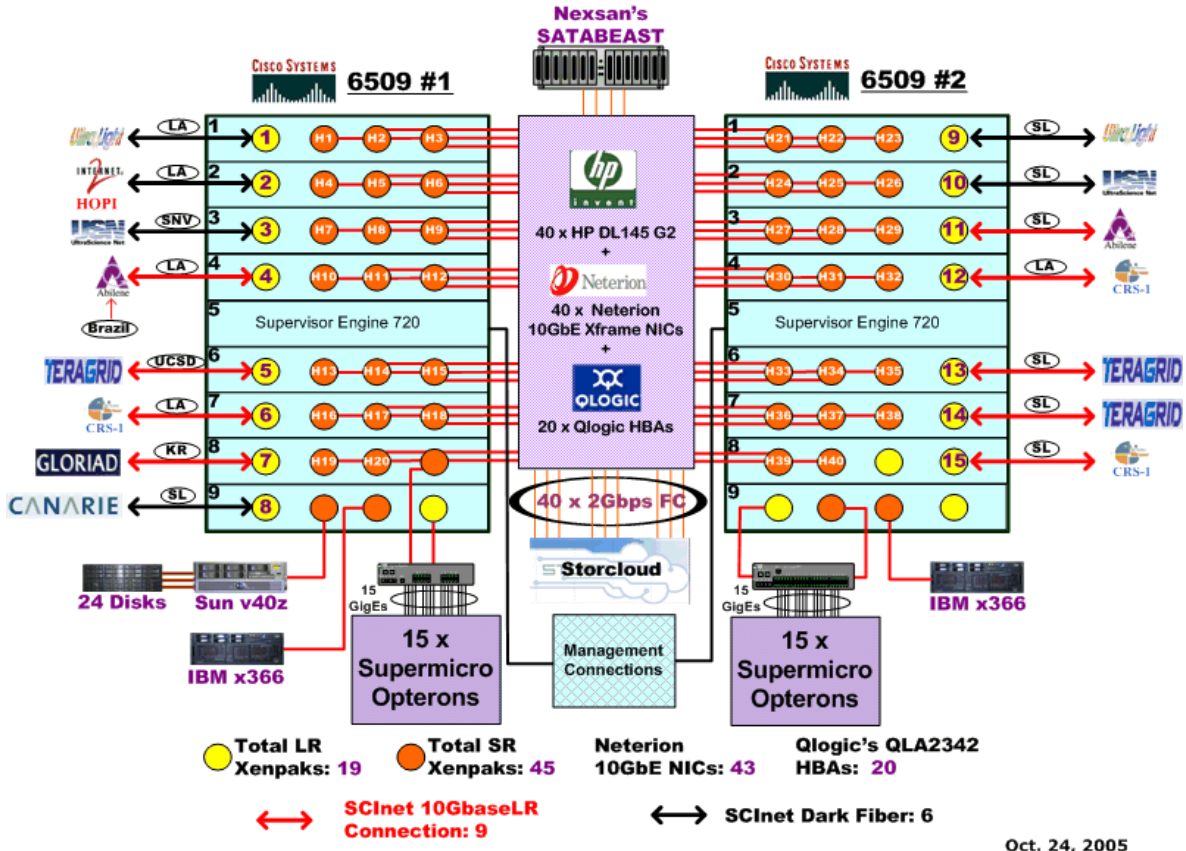


Figure 5: Switch and server interconnections at SC05 Caltech booth.

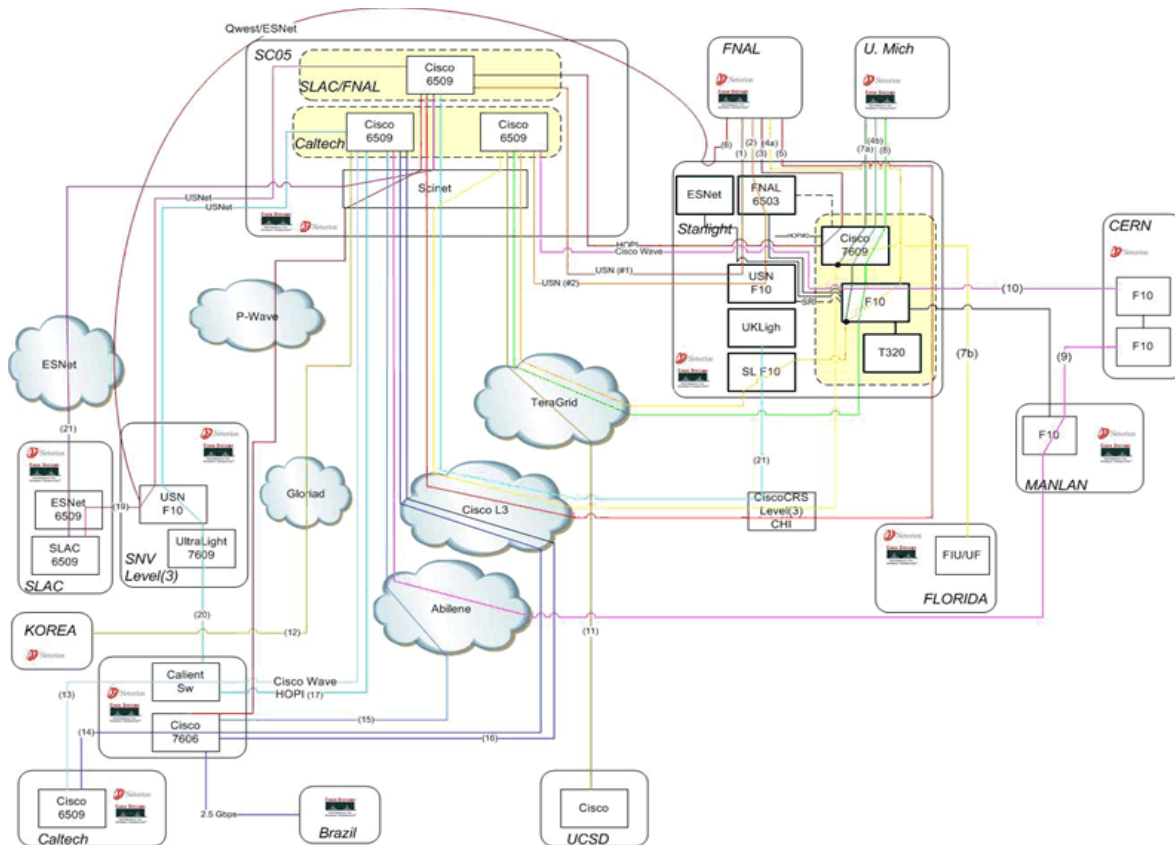


Figure 6: The wide area network circuits used by our SC05 Bandwidth Challenge exercise.

Our entry used real-world applications where real physics data was transferred based on *ROOT*⁶ 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*⁷, *xrootd*⁸, and *gridftp*⁹. Part of the data was transferred between remote *SRM*¹⁰ *dcache*¹¹ 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 7 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 8 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. Figure 9 is MonALISA graph showcasing the Brazilian sites involved in the exercise, namely UNESP and UERJ, which has been a participant of the BWC since SC04, when they set a Brazilian research network speed record of 2Gbps from Brazil to US (and 1Gbps from US to Brazil) over the

⁶ <http://root.cern.ch>

⁷ <http://www.slac.stanford.edu/~abh/bbcp/>

⁸ <http://xrootd.slac.stanford.edu/>

⁹ http://www.globus.org/grid_software/data/gridftp.php

¹⁰ <http://sbm.lbl.gov/srm-wg>

¹¹ <http://www.dcache.org>

WHREN-LILA link connecting AMPATH¹² at Miami and ANSP¹³ 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).

SC2005 BWC Data Flows to Caltech Booth

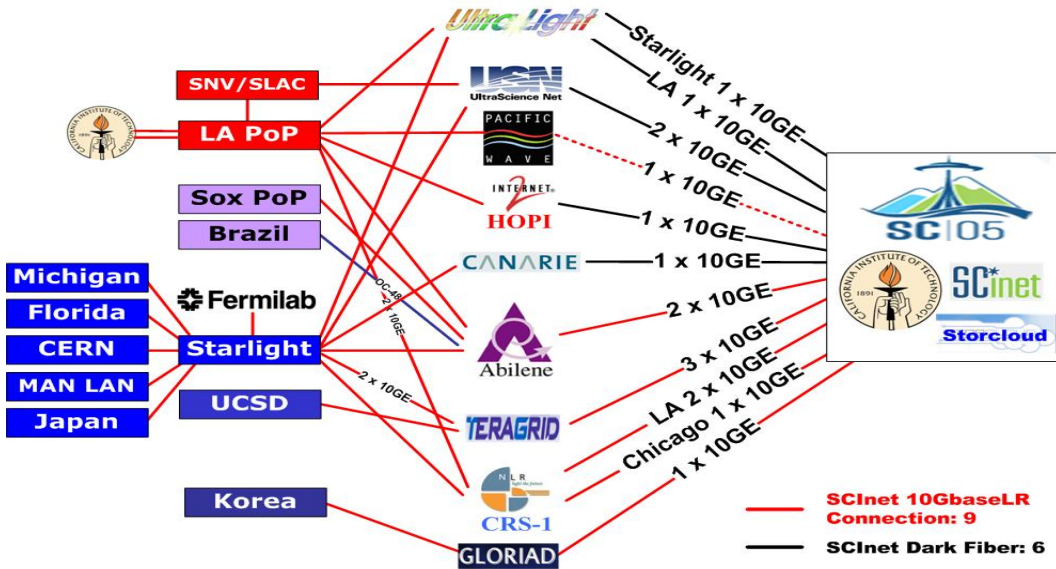


Figure 7: The traffic flows to/from Caltech booth involved in the SC|05 Bandwidth Challenge.

Fermilab-SLAC Bandwidth Challenge Contributions

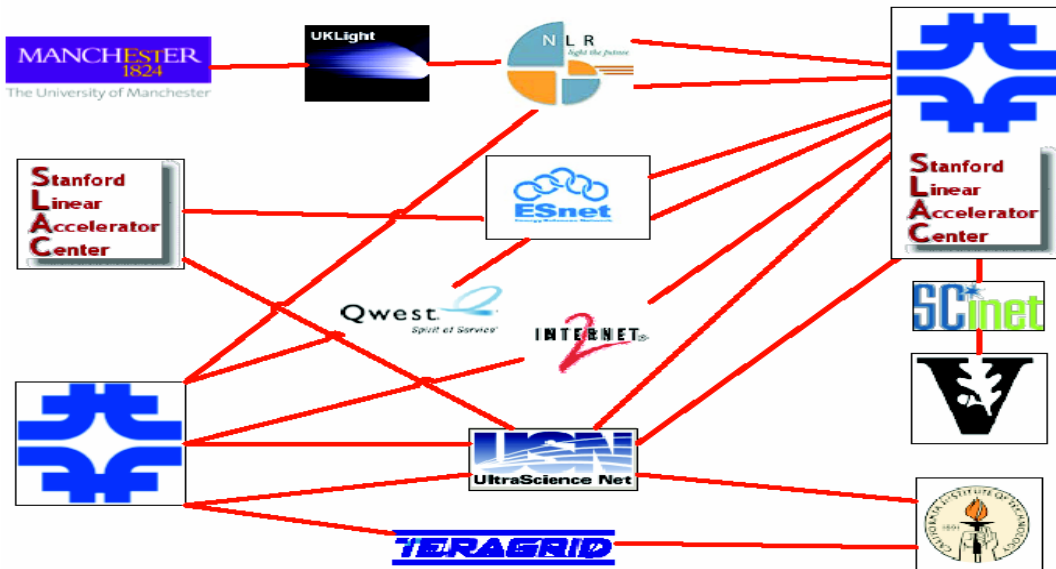


Figure 8: The traffic flows to/from SLAC/Fermi Lab booth in the SC05 Bandwidth Challenge.

¹² <http://www.ampath.fiu.edu>

¹³ <http://nara.org.br>

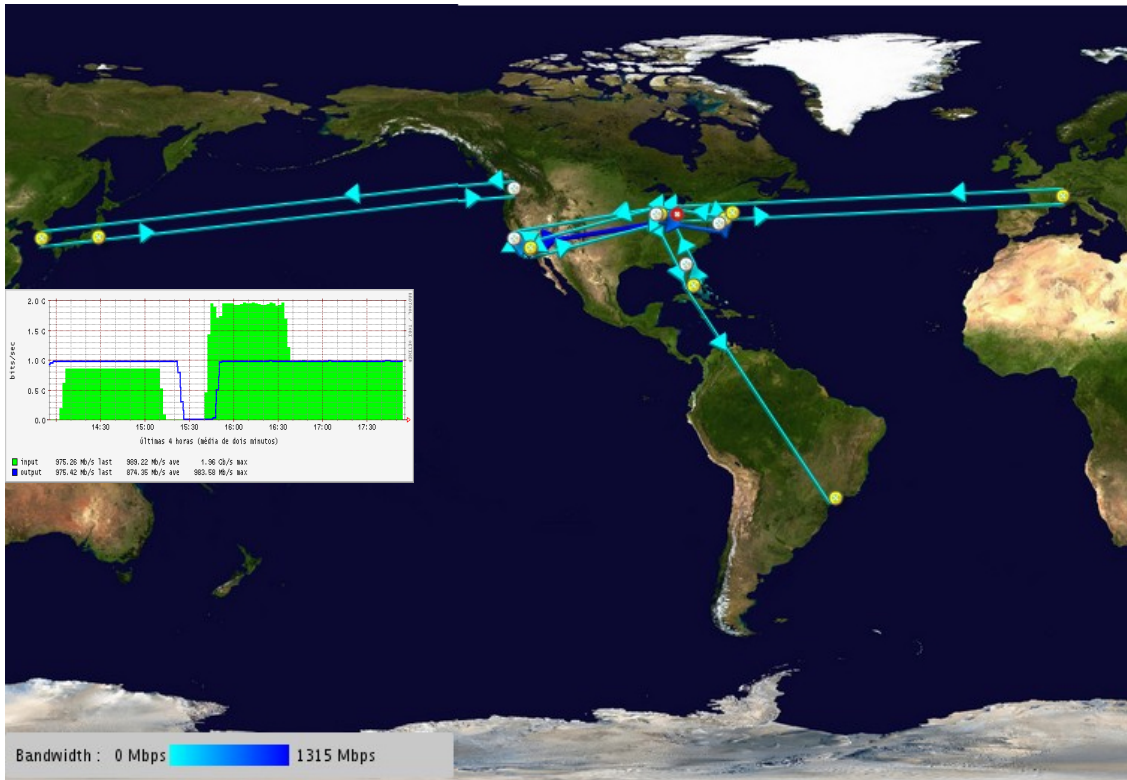


Figure 9: Brazilian participants of bwc: UNESP and UERJ, and the Brazilian R&E network record.

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 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*¹⁴, 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*¹⁵. 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**.

¹⁴ http://ultraviolet.caltech.edu/web-site/sc05/pictures/misc/data_grid_hierarchy.jpg

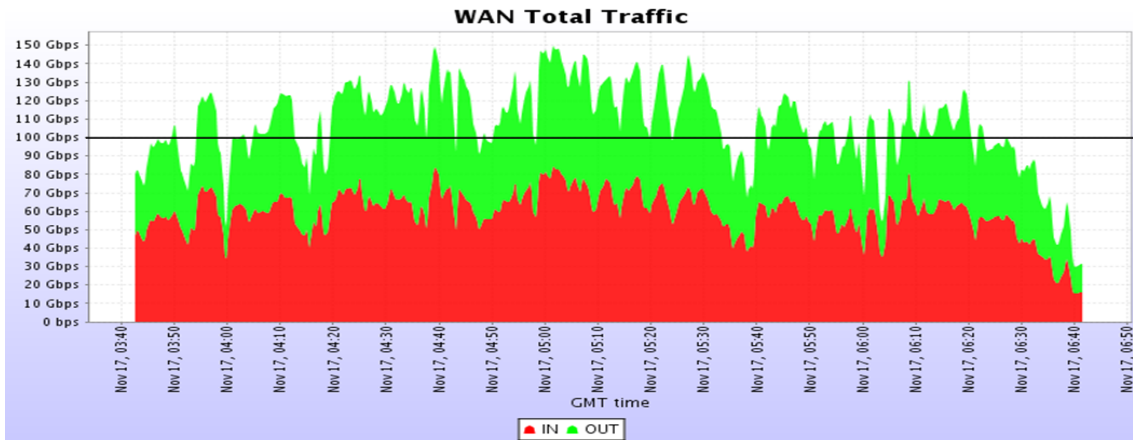
¹⁵ http://ultraviolet.caltech.edu/web-site/sc05/pictures/misc/traffic_trends.jpg

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** were 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. Some interesting take-aways from this exercise include,

- (1) An optimized Linux (2.6.12 + FAST + NFSv4) kernel for data transport after 7 full kernel-build cycles in 4 days. This package also included scripts for publishing information to MonALISA. The intention is to further develop this kernel package and make it available through collaborations such as the Open Science Grid. Such a package would make it easier to install a software stack needed to support high performance data transfers and utilities to publish monitoring data for diagnosing potential problems and actively managing end systems in a network environment.
- (2) A newly optimized application-level copy program BSCP was tested that matches the performance of iperf under some conditions. Figure 12 shows the throughput achieved using bbcp during bandwidth challenge using LHCnet from Chicago to CERN, with an average rate of approximately 420MByte/s. Other transfer applications included Xrootd, an optimized low-latency file access application for clusters, across the wide area. SLAC recorded about 3.2 TByte of data to StorCloud in 1649 files at the same time as it transferred over 18 TBytes in 257913 files via Xrootd on the network between SLAC and SC05.

Systems such as SRM DCache offer the use of GridFTP as mechanism for data transfer and the SRM interface specification allows for multiple protocols to be available in SRM systems. It is most likely that in future network aware Grid systems there will be multiple transfer protocols based on GridFTP, BSCP, XRootD,..., supporting data intensive sciences, some of the higher level transfer protocols will be supported by a lower level FAST protocol.



WAN Total Traffic						
Farm	Last value	Min	Avg	Max	Total	
IN	15.76 Gbps	14.9 Gbps	59.32 Gbps	88.87 Gbps	79.64 TB	
OUT	15.27 Gbps	14.31 Gbps	47.02 Gbps	74.07 Gbps	63.13 TB	
Total	31.03 Gbps		106.3 Gbps		142.8 TB	

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

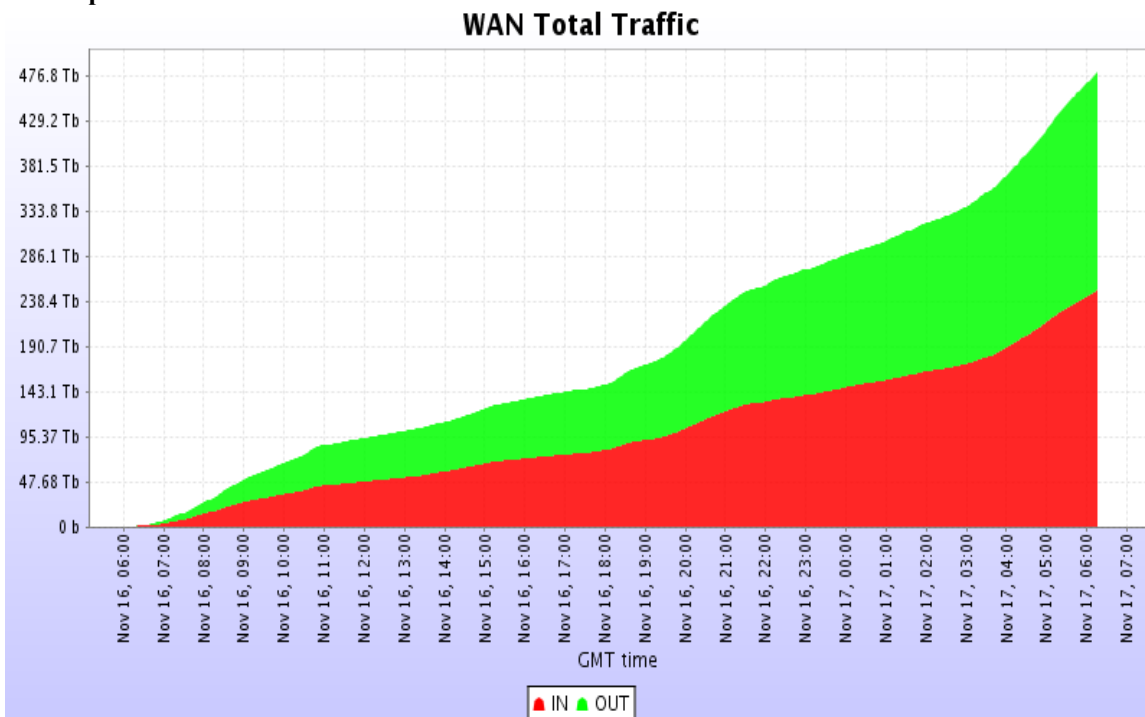


Figure 11: Total WAN traffic volume during SC05 bandwidth challenge for a 24 hour period.

- (3) The BWC also enabled us to understand the limits of 10 Gbps-capable systems under stress, especially how to effectively utilize 10GE and 1GE connected systems to drive 10 gigabit wavelengths in both directions. For example, University of Michigan was able to reach about 30 Gbps over 3 waves.

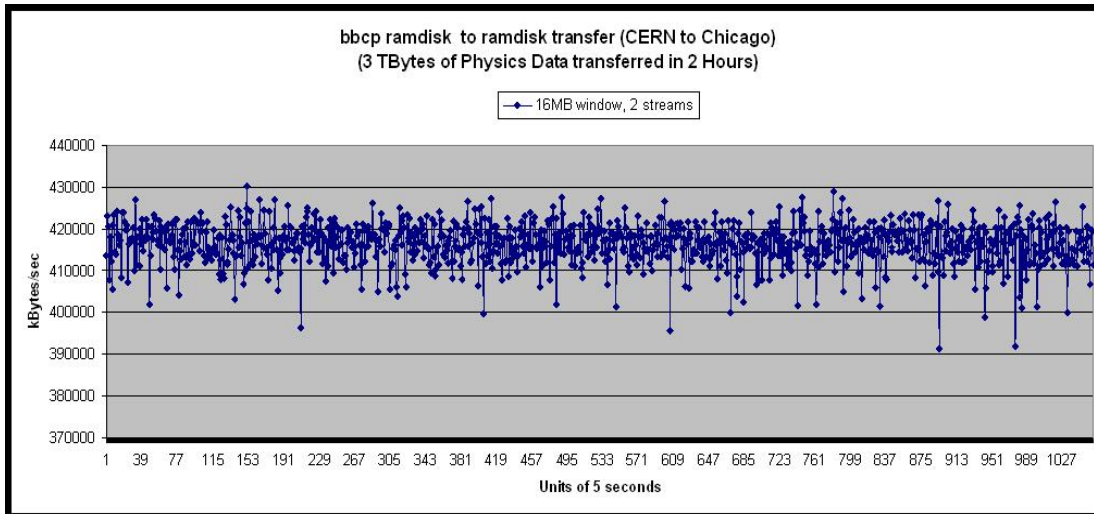


Figure 12: bbcp performance example: 16MB windows, 2 streams over LHCnet.

- (4) Practical experience was gained during BWC towards the use of production and test clusters at FNAL reaching more than 20 Gbps of network throughput. Clusters based on SRM/DCache, like the ones used in the BWC are also used in a production environment for LHC experiments. The BWC event also led to useful hands-on information on the stability limits of server and network interfaces. Several of these interfaces crashed under the heavy loads, which was sometimes attributed to overheating. Crucial during the BWC was therefore a collection of high powered fans that were cooling the server set up at the Caltech booth.
- (5) Significant challenges remain from the perspective of supporting e-science discovery. This entails substantial progress to be made in management, integration and optimization of network resources, as well as the development of capabilities to utilize these network resources end to end to effectively integrating scientific applications and IO devices (disk and storage systems).

Conclusion

The UltraLight project marks 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 a network testbed of unprecedented scope, both in the field and in the laboratory, UltraLight relies on 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, UltraLight will drive the next generation of Grid developments, and support new modes of collaborative work. Such globally distributed systems will serve future advanced applications in many disciplines, bringing great benefit to science and society. UltraLight 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 at the HEP frontier, soon to be explored at the LHC.

While the SC05 demonstration required a major effort by the teams involved and their sponsors, in partnership with major research and education network organizations in the U.S., Europe, Latin America and Asia Pacific, it is expected that networking on this scale in support of the largest

science projects (such as the LHC), will be commonplace within the next three to five years. By demonstrating that many 10 Gbps wavelengths can be used efficiently over continental and transoceanic distances (often in both directions simultaneously), the HEP team showed that this vision of a worldwide dynamic Grid supporting many Terabyte or larger data transactions is feasible and that issues such as End-2-End monitoring and management need to be addressed in an integrated orchestrated manner between CPU, storage and network resources.

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 would also like to acknowledge the generous support of our many 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 Department of Energy or the National Science Foundation.

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. Buhmaster, 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. Latka, 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/>)