

THE ULTRALIGHT PROJECT: THE NETWORK AS AN INTEGRATED AND MANAGED RESOURCE FOR DATA-INTENSIVE SCIENCE

The UltraLight project treats the network interconnecting globally distributed data sets as a dynamic, configurable, and closely monitored resource to construct a next-generation system that can meet the high-energy physics community's data-processing, distribution, access, and analysis needs.

The high-energy physics (HEP) community engaged in the European Center for Nuclear Research's (CERN) Large Hadron Collider (LHC) is preparing to conduct a new round of experiments to probe the fundamental nature of matter and space-time, and to understand the universe's composition and early history. The accelerator's decade-long construction phase and associated experiments (see http://cern.ch/lhc/LHC_Experiments.htm) are approaching completion, and the design and development of the computing facilities and software are well under way. CERN expects the LHC to begin operations in 2007. The experiments face unprece-

dent engineering challenges because of the experimental data's volume and complexity and the need for collaboration among scientists located around the world.

The massive, globally distributed data sets that the LHC experiments will acquire, process, distribute, and analyze should grow beyond the 100-Pbyte level by 2010. Distributing these data sets to scientists and computing centers around the world will require network speeds of roughly 10 to 100 Gbits per second (Gbps) and above. The data volumes will likely rise to the exabyte range, and the corresponding network throughputs to the 100 Gbps to 1 Tbit-per-second range, by approximately 2015. In response to these challenges, collaborations in the US, Europe, and Asia—for example, Enabling Grids for E-science (EGEE, <http://public.eu-egee.org>), Open Science Grid (OSG, www.opensciencegrid.org), and Grid3 (www.ivdgl.org/grid2003/)—have developed grid-based infrastructures that provide massive computing and storage resources. However, the treatment of the interconnecting network as an external, passive, and largely unmanaged resource hampers the efficient use of these resources.

The UltraLight consortium of major HEP centers in the US (www.ultralight.org; see also the "UltraLight Consortium Member Organizations"

1521-9615/05/\$20.00 © 2005 IEEE
Copublished by the IEEE CS and the AIP

HARVEY NEWMAN, JULIAN BUNN, IOSIF LEGRAND, STEVEN LOW, DAN NAE, SYLVAIN RAVOT, CONRAD STEENBERG, XUN SU, MICHAEL THOMAS, FRANK VAN LINGEN, AND YANG XIA

California Institute of Technology

RICHARD CAVANAUGH

University of Florida

SHAWN MCKEE

University of Michigan

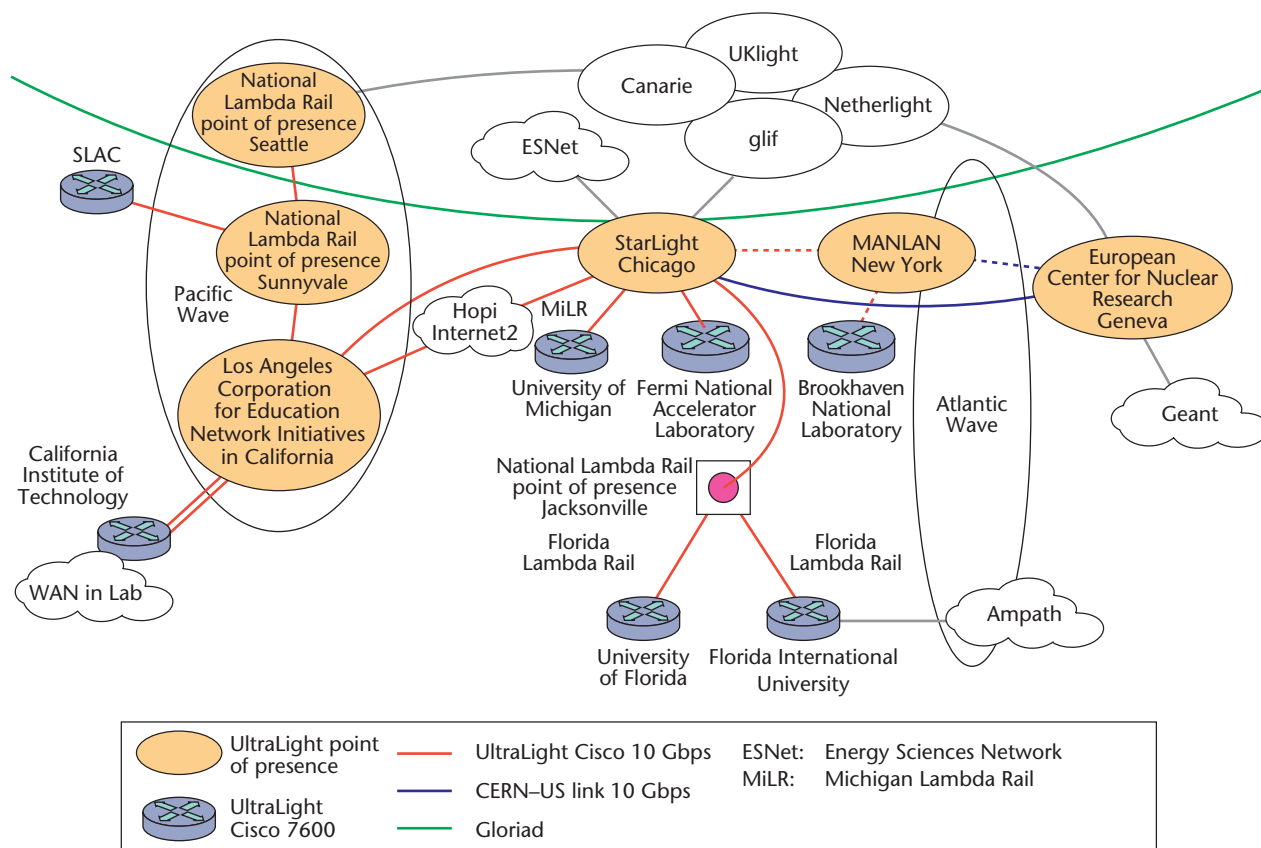


Figure 1. A schematic view of the initial UltraLight setup and connections to other major networks and Caltech’s WAN in Lab (<http://wil.cs.caltech.edu/>). The various sites are connected through points of presence (PoPs) and Cisco routers.

sidebar) was formed to address this deficiency by treating the network as a dynamic, configurable, and closely monitored resource that’s managed end-to-end. Figure 1 shows UltraLight’s global capabilities and the extent of its networking resources (see the “UltraLight Networking Resources” sidebar for a list of networks and other resources involved). We believe UltraLight is viable for e-science because of the relatively limited scope of our virtual organizations (we focus only on distributed data analysis). The advances we’re proposing in the project’s current form will most likely not scale to the general Internet community. We’re aware of scalability as an issue and are targeting our developments toward a large but manageable set of resources and virtual organizations (approximately 20).

The Network as a Resource

Within HEP, there are typically two scenarios that use wide-area networks (WANs) extensively.

In the first scenario, (raw) data from the detector located at CERN (tier 0) produces so-called raw data at a rate of petabytes per year, which is

processed locally. CERN stores this reconstructed data, distributing it in part to tier-1 centers around the world. These centers, in turn, make the data available to tier-2 centers.

The second scenario is related to analysis of the data in scenario 1. Physics analysis represents a “needle in the haystack” problem in which physicists analyze large data sets and identify the data sets needed for physics discovery in an iterative manner. Hundreds of physicists perform various types of analysis at any time using data that can be distributed over several sites. During this process, certain data sets become very popular or “hot,” whereas others languish, or become “cold.” In time, data can change from hot to cold, or vice versa, depending on what data physicists are interested in. Data must be readily available for physicists wherever they’re working. This can require replicating the typically large data sets, which range from hundreds of gigabytes to several terabytes. This (potentially real-time) replication combined with scenario 1 can overwhelm the network infrastructure and create a network traffic jam, which

UltraLight Consortium Member Organizations

The US National Science Foundation funds the UltraLight project, which is based on a strong partnership between the following organizations:

- Brookhaven National Laboratory
- California Institute of Technology (Caltech)
- European Center for Nuclear Research (CERN)
- Fermi National Accelerator Laboratory
- Florida International University
- Internet2
- Stanford Linear Accelerator Center
- Universidade do Estado do Rio de Janeiro
- University of Florida
- University of Michigan
- University of Sao Paulo

can limit the use of computing resources because people can't get to their data.

The notion of treating the network as a managed resource is motivated by the general assumption that grid resources (CPU, storage, network, and so on) will always be insufficient to meet demand. This assumption is based on years of experience with HEP computing, and it has design implications for the system as a whole. Such a resource-constrained system requires policies that enforce fair sharing. "Fair" in this context means that the groups of users, or virtual organizations, involved agree on the terms under which they can use the available resources. Moreover, scheduling decisions must account for these policies and the desired turnaround-time profiles for each of several work classes. Properly scheduling and managing the system and adhering to these policies require detailed planning, monitoring, and enforcement procedures that take into account the relevant information.

To achieve the goal of treating the network as an actively managed resource, UltraLight focuses on four areas:

- End-to-end monitoring, which provides components with real-time status information on the entire system or on selected components. Autonomous components can use monitoring information to make decisions on the users' behalf or to optimize the system as a whole (for example, optimize data throughput and CPU utilization).
- Development and deployment of fundamental network and transfer protocols and tools such as

the Fast Active Queue Management Scalable Transmission Control Protocol (FAST TCP),^{1,2} and bandwidth and routing management technologies such as Multiprotocol Label Switching (MPLS) and quality of service (QoS).³

- Deployment, testing, and utilization of the UltraLight testbed and WAN in Lab (<http://wil.cs.caltech.edu/>).
- Application-level services for physics and other domains, providing interfaces and functionalities for physics applications to effectively interact with the networking, storage, and computation resources while performing (physics) analysis.

UltraLight combines results and applications developed in these four areas to provide an end-to-end service-based system that supports grid-enabled physics analysis. The system will utilize the network as an active managed resource to support thousands of users, and will exploit grid resources to allow the analysis of petabytes of data and contribute to active global collaboration on the discovery of new physics.

Results will be disseminated by the UltraLight consortium through OSG. Indeed, several applications and frameworks discussed in this article are already part of the OSG software distribution and are being deployed on the OSG testbed.

End-to-End Monitoring

The use of end-to-end monitoring is crucial to exposing the network as a managed resource. End-to-end monitoring lets applications and higher-level service layers account for a system's increasingly advanced and complex behavior, leading to a new class of proactive and reactive applications that dynamically adapt to new and unforeseen system behavior. For example, the distributed system could gracefully accommodate network congestion or hardware component failures and exploit the availability of new network routes or capabilities. These reactive applications would enhance the global system's resilience to malfunction and let it optimize resource use, thereby improving both overall task throughput and effective policy implementation and increasing the speed with which physics results are obtained. This is the UltraLight effort's ultimate goal.

End-to-end monitoring is thus vital within the UltraLight project to help optimize network resources. Part of the UltraLight planning is to implement a new set of global end-to-end managed monitoring services, building on our ongoing and rapidly advancing work with the MonAlisa agent-based system.⁴

MonAlisa will monitor and control network de-

UltraLight Networking Resources

UltraLight's resources include several major networks and network infrastructures:

- LHCNet (www.datatag.org), the transatlantic 10-Gbps backbone connecting CERN to the US in Chicago;
- transcontinental 10-Gbps wavelengths from National Lambda Rail (www.nlr.net), a major initiative of US research universities and private-sector technology companies (including Cisco Systems, www.cisco.com, and Internet2's Abilene network, <http://abilene.internet2.edu>) to provide a national-scale infrastructure for research and experimentation in networking technologies and applications; and
- partnerships with StarLight, a high-performance network exchange for many worldwide research and educational wide-area networks (WANs).

We use additional trans- and intercontinental wavelengths from several of our partner projects for network experiments on a part-time or scheduled basis. These include

- TransLight (www.startup.net/translight),
- Netherlight (www.netherlight.net),
- UKLight (www.uklight.ac.uk),
- Ampath (www.ampath.fiu.edu), and
- CA*Net4 (www.canarie.ca/canet4).

VICES, such as routers and photonic switches. Because it gathers system-wide information, MonAlisa can generate global views of the prevailing network connectivity to identify network or end-system problems and act on them strategically, or locally, as required. Additional services that take decisions based on these (global) system views produced by MonAlisa (such as schedulers) can be created and deployed. For example, a recent addition to MonAlisa is a mobile agent that provides optimized dynamic routing for distributed applications. At the time of this writing, the MonAlisa system includes 180 station servers (split between grid and virtual room videoconferencing system, or VRVS,⁵ sites) and monitors approximately 180,000 different operational parameters from 10,000 participating nodes and more than 60 WAN links. In MonAlisa, independent processes can publish and subscribe to the data of other processes in the globally deployed system. Using a low-level predicate mechanism within MonAlisa, it's possible to create filters in these processes and associate the filters with certain actions. We can view the combination of filters and associated actions as a rudimentary form of policy specification.

Consider a process subscribed to a network link's bandwidth utilization that identifies (by filtering the subscribed data) a high-priority data movement activity. The associated action can use MPLS and QoS to "throttle" the bandwidth usage for other processes through dynamically sized virtual pipes, improving the high-priority transfer's throughput. This would let the system assign a new higher-priority task the bandwidth it needs implicitly at the expense of less critical traffic. Each virtual organization would need to establish a set of policies to govern different actions' priority rankings. In another example, MonAlisa components could de-

tect a saturated network link and reroute additional traffic through other links. Using the traffic analogy, MonAlisa would supply the traffic lights to control network resource usage and direct high-priority tasks to dynamically created express lanes in the network.

As a first experiment in end-to-end resource monitoring, we've adapted MonAlisa and deployed it on VRVS reflectors to collect information about the system's topology, monitor and track traffic among reflectors and report communication errors with the peers, and track the number of clients and active virtual rooms. In addition, we monitored overall system information (such as CPU usage and total traffic in and out) and reported it in real time for each reflector. We developed agents within MonAlisa to provide and optimize dynamic routing of the VRVS data streams. These agents use information about the quality of alternative connections to solve a minimum spanning tree problem to optimize data flow at the global level. The latter is also important within the UltraLight project.

Figure 2 shows the MonAlisa system in action during Supercomputing 2004 (SC2004, www.sc-conference.org/sc2004). MonAlisa gathers arbitrarily complex monitoring information in the global system and processes it in its distributed agent framework. Its use of agents and a multi-threaded engine that hosts various loosely coupled and self-describing dynamic services, as well as each service's ability to register itself and then be discovered and used by other services or clients lets the framework scale well with system size.

Several grid projects (such as Grid3) have used MonAlisa. It's part of OSG and monitors several major networks, such as Abilene and the Global Ring Network for Advanced Applications Development (Gloriad, www.gloriad.org). One reason

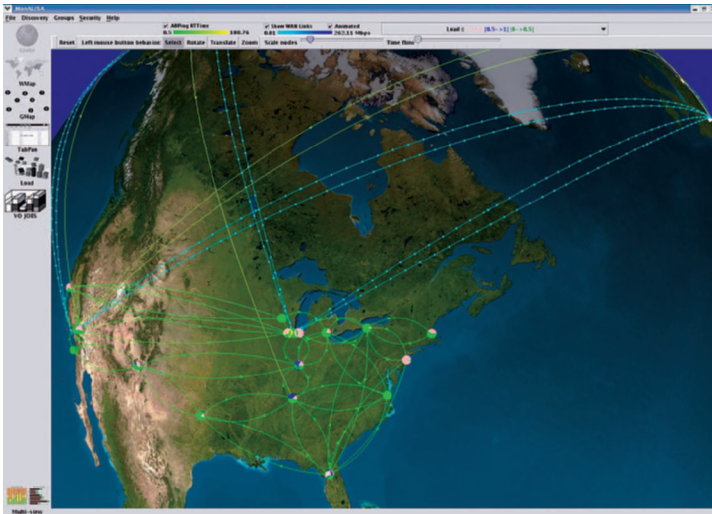


Figure 2. The MonAlisa framework. MonAlisa provides a distributed monitoring service system in which each station server hosts and can schedule many dynamic services. It thereby acts as a dynamic service system that can be discovered and used by any other services or clients requiring such information.

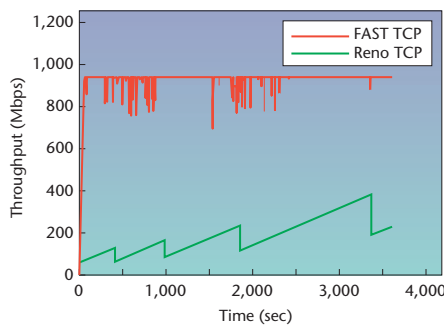


Figure 3. Throughput of FAST TCP flows compared with Reno in the presence of packet loss. FAST TCP stabilizes at a steady high throughput whereas Reno ramps up much slower at a lower throughput.

for using MonAlisa within the UltraLight project is its features, such as global scalability, low-level policy specifications, autonomous publish-subscribe functionality, and the ability to steer other applications based on monitor information.

These features aren't (or are only partly) available in other grid-based monitoring systems such as Ganga,⁶ the Relational Grid-Monitoring Architecture,⁷ and GridICE.⁸

Protocols and Tools

FAST TCP is a TCP implementation with a new congestion-control algorithm optimized for high-speed long-distance transfers. Whereas the current

TCP implementation's congestion-control algorithm uses packet loss as a measure of congestion, FAST TCP uses round-trip delay (the time from sending a packet to receiving its acknowledgment). This lets FAST TCP stabilize at a steady throughput without having to perpetually push the queue to overflow, as loss-based schemes inevitably do. Moreover, scaling delay with link capacity enhances stability as networks grow in capacity and geographical size.⁹

In addition to many experimental evaluations of FAST TCP in real networks and emulated testbeds, we've modeled it mathematically and analyzed its equilibrium and stability properties. In equilibrium, FAST TCP allocates bandwidth among competing flows in general networks according to proportional fairness, which favors small flows but less extremely than maxmin fairness.¹⁰ Moreover, the equilibrium point always exists and is unique for an arbitrary network. Previous work has shown that the equilibrium point is stable under various assumptions.⁹ Figure 3 compares FAST TCP with Reno TCP, which is based on the Reno fast retransmit algorithm.¹¹

UltraLight intends to exploit recent advances in robust control theory and convex optimization to guide the operation of the large-scale network we will build. New techniques at the California Institute of Technology (Caltech) use sum-of-squares optimization methods to provide convex polynomial time relaxations for many NP-hard problems involving positive polynomials. The observation that it's possible to use semidefinite programming to efficiently compute sum-of-squares decompositions of multivariate polynomials has initiated the development of software tools (such as SOSTools¹²) that let you formulate semidefinite programs from their sum-of-squares equivalents. UltraLight will develop methods of incorporating SOSTools into the MonAlisa monitoring and control software system, which will in turn let us calculate stability regions for a given network operation regime and derive control actions that will steer the network so that it remains within a desired performance regime.

UltraLight High-Speed Networking

The UltraLight hybrid packet- and circuit-switched network infrastructure uses both ultrascale protocols, such as FAST TCP, and the dynamic creation of optical paths for efficient fair sharing on networks in the 10-Gbps range. In November 2004, UltraLight project members, led by researchers at Caltech, broke the Internet2 land-speed record (<http://lsr.internet2.edu>) by sending 2.9 Tbytes of data across 26,950 kilometers of network in one hour using Caltech's FAST TCP,¹ at an average rate of 6.86 Gbps. The same week, the team captured the Supercomputing Band-

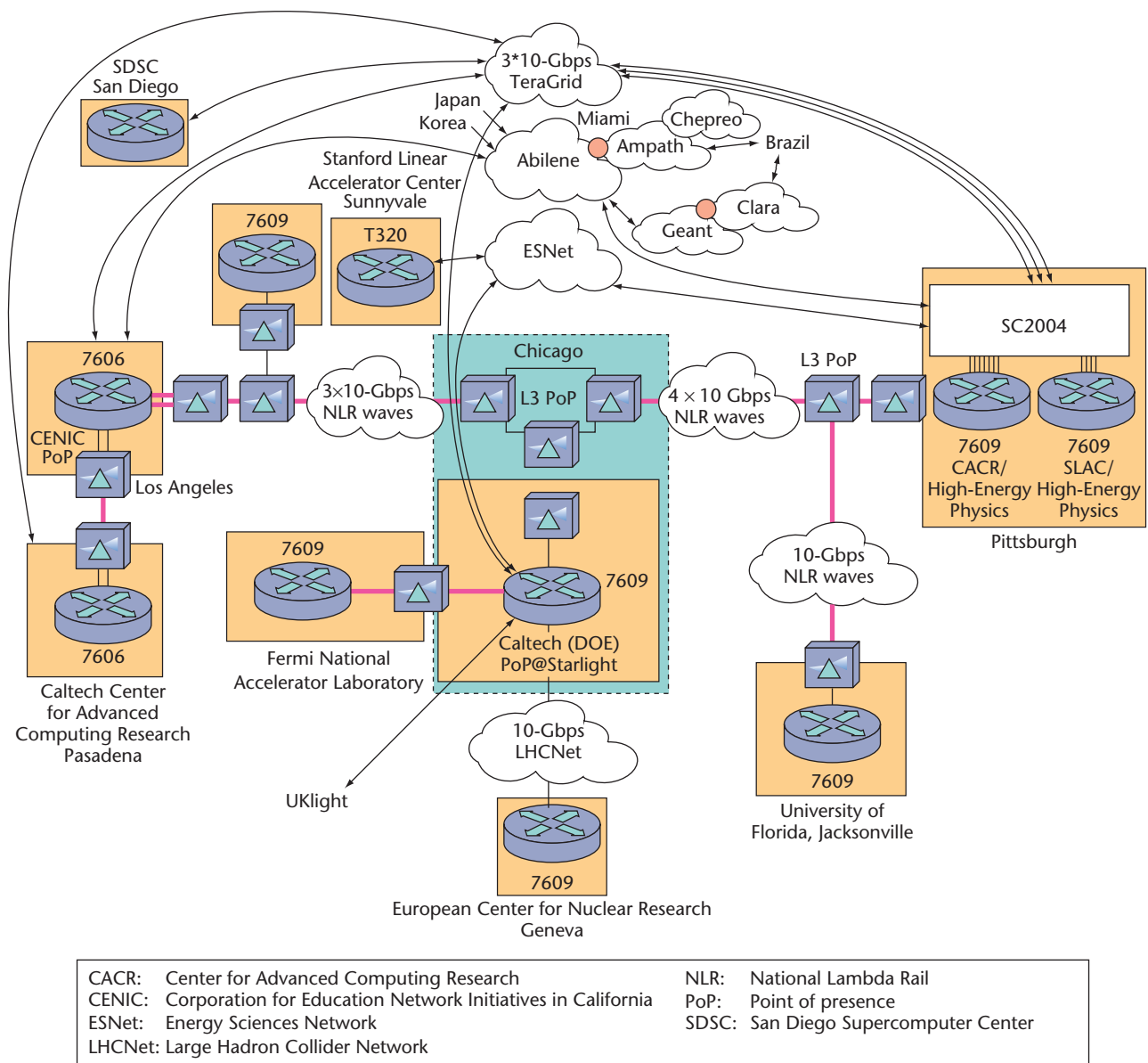


Figure 4. High-speed terabyte transfers for physics demonstration at Supercomputing 2004. Using FAST TCP and links provided by various organizations, we achieved an aggregate rate of 101 Gbps to and from the conference floor.

width Challenge award (http://pr.caltech.edu/media/Press_Releases/PR12620.html) for sustained bandwidth by generating an aggregate rate of 101 Gbps to and from the conference floor—the same level of data throughput expected to occur in the early operation phases of the LHC experiments.

As Figure 4 illustrates, we were able to achieve this bandwidth by using FAST TCP together with

- seven 10-Gbps links to the Caltech Center for Advanced Computing (CACR) booth;
- three 10-Gbps links to the Stanford Linear Ac-

- celerator Center (SLAC)/Fermilab booth;
- wide-area links provided by National Lambda Rail, Abilene, LHCNet, the Energy Sciences Network (ESNet; www.es.net), and TeraGrid (www.teragrid.org);
- links to Brazil over the Ampath and Chepreo links (www.chepreo.org) and via an EU-funded Clara link (www.redclara.net) and the Geant (www.geant.net) pan-European network; and
- a link to Korea.

Other key elements in the diagram include two

10-Gbps links between Caltech and Los Angeles provided by Cisco Systems, connections to the Jacksonville location on Florida Lambda Rail, and a 10-Gbps dark-fiber link between StarLight and Fermilab. Other than the connections in Pittsburgh, most of the network connections shown in the figure persist as part of the UltraLight network testbed.

These bandwidth challenges aim to show that given the combination of hardware and software, we can create a data-transfer superhighway that can meet the challenges of the physics community. A next step in the bandwidth challenges will be disk-to-disk transfers and the analysis of potentially remote data as if it were on a local disk. Although this type of analysis won't be available to everyone, we'll make it available to high-priority users and processes. Thus, we'll be able to offer a more agile system with the flexibility to assign resources (CPU, storage, and network) to promising physics analysis processes or groups, which will speed new scientific discoveries.

WAN in Lab

The SC2004 Bandwidth Challenge provided an ideal environment for dedicated access to network resources for a limited amount of time. Although we deployed monitoring applications on many of the network resources, we had no control over others, which made diagnosing low-level component failures difficult. To analyze every aspect of the network and applications interacting with network resources, the UltraLight consortium will use an unusual facility: WAN in Lab.

WAN in Lab is a unique testbed being built at Caltech and funded by the US National Science Foundation, the US Army Research Office, Cisco, and Caltech. WAN in Lab is literally a wide-area network—it includes 2,400 kilometers of fibers, optical amplifiers, dispersion-compensation modules, wavelength-division multiplexing gear, optical switches, routers, and servers—but it's housed in a single laboratory at Caltech. The initial hardware, anticipated to be operational in Fall 2005, will have six Cisco ONS 15454 switches, four Cisco 7609 routers, and a few dozen high-speed servers. We intend to connect it to UltraLight by a 10-Gbps link, making it an integral part of the system (see Figure 1). This will extend the round-trip time of an end-to-end connection between a WAN in Lab server and a server in a global production network to more than 300 ms. This number is important because it's larger than, but of the same scale as, the largest round-trip times we expect in real networks, ensuring that our work is relevant for global networks. We also intend to connect WAN in Lab to the Sunnyside and Seattle Gigapops, as Figure 5 illustrates.

WAN in Lab—the “wind tunnel of networking research”—offers a unique environment for developing and testing protocols for sharing resources across high-speed WANs and complements the UltraLight infrastructure synergistically. The UltraLight community can develop, debug, and test distributed systems tools on WAN in Lab before deploying them progressively on UltraLight. This will greatly shorten the design, development, testing, and deployment cycle for the community.

Application-Level Services

Ultimately, we must integrate the network as a managed resource into existing globally distributed systems. The Grid Analysis Environment (GAE)¹³ describes an application-level service-oriented architecture (SOA) to support end-to-end (physics) analysis. It describes the ensemble of services (and their interactions)—discovery,¹⁴ scheduling,¹⁵ submission,¹⁶ and job tracking, for example—that the SOA will expose to users and domain applications. UltraLight is extending the GAE to the UltraLight Analysis Environment. The UAE focuses on integrating components identified in the GAE and components that expose the network as a managed resource.

Most users will access grid resources through grid portals that hide much of the grid's (Web service) infrastructure and resource complexity. GAE components will monitor applications, replicate data, schedule jobs, and autonomously find optimal network connections, resulting in a self-organizing grid that minimizes single points of failure. This would let thousands of users get fair access to a limited set of distributed grid resources in a responsive manner. Many UAE Web service implementations will be made available through and developed in Clarens, a Web-service-based framework (such as Globus¹⁷ and Glite¹⁸) available in Python and Java implementations. Clarens offers several additional features:

- X.509-certificate-based authentication when establishing a connection,
- access control on Web services,
- remote file access and access control,
- services and software discovery,
- virtual organization management,
- high performance (measured at 1,400 calls per second),
- role management (which we're extending to interoperate with GUMS servers¹⁹), and
- support for multiple protocols (such as XML Remote Procedure Call, SOAP, JavaScript Object Notation, and Java remote method invocation).

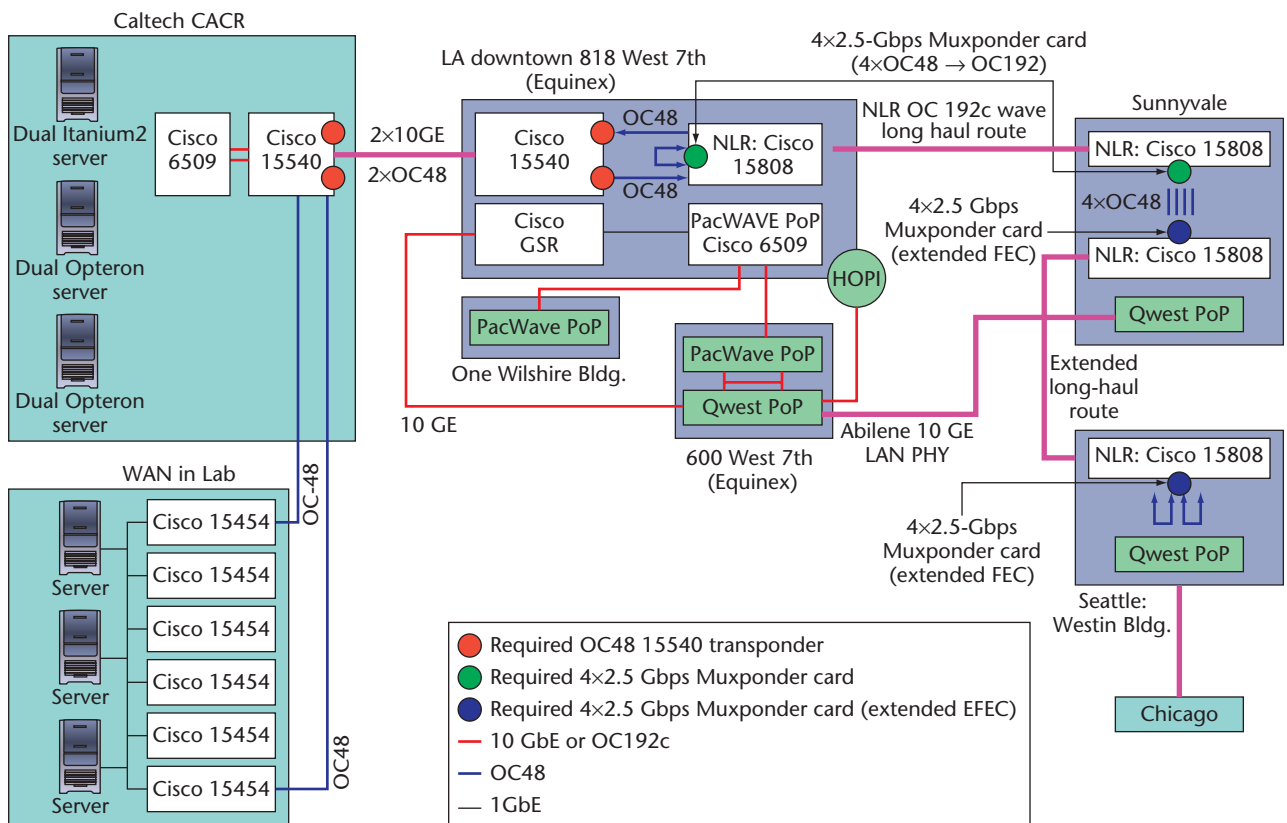


Figure 5. WAN in Lab extended layout. The lower left part shows the WAN in Lab connected to Caltech Center for Advanced Computing Research (CACR), which in turn connects to the point of presence (PoP) in Los Angeles. This figure is a magnification of part of the left side of Figure 1.

We'll integrate the network resources and Web services using MonAlisa.¹⁴

At SC2003 and SC2004, we demonstrated an early GAE prototype that let users submit jobs to a site through the Sphinx Grid scheduler.¹⁵ Sphinx differs from other schedulers such as Pegasus²⁰ in that it supports a decentralized policy-based environment for scheduling. Sphinx bases scheduling decisions on MonAlisa monitoring data, which includes the site's queue length and the speed at which a job finishes once it starts running. Sphinx then actively monitors the job's progress in the queue and, if necessary, reschedules it to another site. An extension to the prototype provided a wrapper around the job that sent job state information to MonAlisa and the Batch Object Submission System (BOSS), which provides job wrappers for monitoring detailed job status.¹⁶ BOSS stores detailed information on the job's state, including error logs, whereas MonAlisa stores real-time information regarding the job's state (for example, submitted, started, running, and finished). We visualized this information by plotting state against time in a lifeline plot.

More recently, we implemented a software and service discovery Web service as Clarens Web Services. Both services use MonAlisa to disseminate discovery information. The services provide a dynamic real-time view of the Web services and software applications available within the distributed system.

The Sphinx scheduler and the discovery services are two early examples of integrating application-level services with an end-to-end monitoring system to provide a global view of the system. The discovery services are part of the OSG and have been deployed on the OSG testbed.

Recent work has begun on creating a reservation service in collaboration with the Lambda Station project (www.lambdastation.org). The service allocates bandwidth (where possible) in multiple network domains from a source to a target when a user requests it. Such a service is important for grid operators who move large amounts of data through multiple network domains. In most cases, this data movement can be scheduled in advance. In the near future, we can replace manual operation of these data transfers with autonomous applications (such

as agents) that monitor data access and decide when to reserve bandwidth.

The UltraLight project marks the entry into a new era of global real-time responsive systems in which we can monitor and track all three sets of resources—computational, storage, and network—to provide efficient, policy-based resource use and optimize distributed system performance on a global scale. 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.

UltraLight paves the way for more flexible, efficient sharing of data by scientists in many countries and could be a key factor in the next round of discoveries at the HEP frontier. Advancements in UltraLight, through Caltech's VRVS system, could also have profound implications for integrating information sharing and on-demand audiovisual collaboration in our daily lives. As a testament that UltraLight addresses real needs within the science community, EGEE lists scenarios that will benefit from the work being undertaken in the UltraLight project at <https://edms.cern.ch/document/476742>.

Acknowledgments

This work is partly supported by US Department of Energy grants DE-FC02-01ER25459, DE-FG03-92-ER40701, and DE-AC02-76CH03000 as part of the Particle Physics DataGrid project and DE-FG02-04ER-25613 as part of Lambda Station; US National Science Foundation grants ANI-0230967, PHY-0218937, PHY-0122557, PHY-0427110, ANI-0113425, ANI-0230967, and EIA-0303620; US Army Research Office grants DAAD19-02-1-0283 and F49620-03-1-0119; and US Air Force Office of Scientific Research grant F49620-03-1-0119. We also acknowledge Cisco's generous support. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and don't necessarily reflect the views of the Department of Energy or the NSF.

References

1. C. Jin, D.X. Wei, and S.H. Low, "FAST TCP: Motivation, Architecture, Algorithms, Performance," *Proc. IEEE Infocom*, IEEE Press, 2004, <http://netlab.caltech.edu/FAST>.
2. C. Jin et al., "FAST TCP: From Theory to Experiments," *IEEE Network*, vol. 19, no. 1, 2005, pp. 4–11.
3. X. Xiao and L.M. Ni, "Internet QoS: A Big Picture," *IEEE Network*, vol. 13, no. 2, 1999, pp. 8–18.
4. H.B. Newman et al., "MonAlisa: A Distributed Monitoring Ser-

vice Architecture," *Proc. Computing for High Energy Physics (CHEP)*, paper MOET001, 2003; www.slac.stanford.edu/econf/C0303241/proc/papers/MOET001.pdf.

5. D. Adamczyk et al., "A Globally Distributed Real Time Infrastructure for World Wide Collaborations," *Proc. Computing for High Energy Physics (CHEP)*, paper 88, 2004; <http://indico.cern.ch/contributionDisplay.py?contribId=88&sessionId=11&confId=0>.
6. M.L. Massie, B.N. Chun, and D.E. Culler, "The Ganglia Distributed Monitoring System: Design, Implementation, and Experience," *Parallel Computing*, vol. 30, no. 7, 2004, pp. 817–840.
7. A. Cooke et al., "R-GMA: An Information Integration System for Grid Monitoring," *Proc. 11th Int'l Conf. Cooperative Information Systems (CoopIS 2003)*, Springer-Verlag, 2003, pp. 462–481.
8. S. Andreatti et al., "GridICE: A Monitoring Service for Grid Systems," *Future Generation Computer Systems J.*, vol. 21, no. 4, 2005, pp. 559–571.
9. J. Wang, D.X. Wei, and S.H. Low, "Modeling and Stability of FAST TCP," *Proc. IEEE Infocom*, IEEE Press, 2005, pp. 938–948.
10. F.P. Kelly, A.K. Maulloo, and D.K.H. Tan, "Rate Control in Communication Networks: Shadow Prices, Proportional Fairness and Stability," *J. Operational Research Soc.*, vol. 49, no. 3, 1998, pp. 237–252.
11. W. Stevens, M. Allman, and V. Paxson, *TCP Congestion Control*, Internet Eng. Task Force, RFC 2581, Apr. 1999; www.ietf.org/rfc/rfc2581.txt.
12. S. Prajna et al., "SOSTools and Its Control Applications," *Positive Polynomials in Control*, D. Henrion and A. Garulli, eds., Springer-Verlag, 2005.
13. F. van Lingen et al., "Grid Enabled Analysis: Architecture, Prototype and Status," *Proc. Computing for High Energy Physics (CHEP)*, paper 182, 2004; <http://indico.cern.ch/contributionDisplay.py?contribId=182&sessionId=9&confId=0>.
14. F. van Lingen et al., "The Clarens Web Service Framework for Distributed Scientific Analysis in Grid Projects," *Proc. Int'l Conf. Parallel Processing*, IEEE CS Press, 2005, pp. 45–52; <http://clarens.sourceforge.net>.
15. J. In et al., "Policy-Based Scheduling for Simple Quality of Service in Grid Computing," *Proc. 18th Int'l Parallel and Distributed Processing Symp.*, IEEE CS Press, 2004.
16. C. Grandi and A. Renzi, "BOSS: A Tool for Batch Monitoring and Book-Keeping," *Proc. Computing for High Energy Physics (CHEP)*, paper THET001, 2003; www.slac.stanford.edu/econf/C0303241/proc/papers/THET001.pdf.
17. I. Foster and C. Kesselman, "Globus: A Metacomputing Infrastructure Toolkit," *Int'l J. Supercomputer Applications*, vol. 11, no. 2, 1997, pp. 115–128.
18. M. Lamanna et al., "Experiences with the gLite Grid Middleware," *Proc. Computing for High Energy Physics (CHEP)*, paper 305, 2004; <http://indico.cern.ch/contributionDisplay.py?contribId=305&sessionId=7&confId=0>.
19. G. Carcassi et al., "A Scalable Grid User Management System for Large Virtual Organizations," *Proc. Computing for High Energy Physics (CHEP)*, paper 122, 2004; <http://indico.cern.ch/contributionDisplay.py?contribId=122&sessionId=12&confId=0>.
20. E. Deelman et al., "Pegasus: Mapping Scientific Workflows onto the Grid," *Proc. European Across Grids Conf.*, LNCS 3165, Springer-Verlag, 2004, pp. 11–20.

Harvey Newman, the UltraLight project's principal investigator, is a professor of physics at the California Institute of Technology (Caltech) and the board chair of the US CMS collaboration. His research interests include elementary particle physics, global distributed

grid systems, and intercontinental networks for data-intensive sciences. He co-led the physics collaboration that discovered the gluon in 1979 and was part of the NSFNet Technical Advisory Group in 1986. He's led the Caltech team that has won 11 Internet2 Land Speed records since 2003. Newman has an ScD in physics from MIT. Contact him at newman@hep.caltech.edu.

Richard Cavanaugh, UltraLight's project manager, is a research physicist at the University of Florida, where he is a member of the CMS collaboration and several grid computing projects. His research interests include establishing experimental signatures for theories predicting cold dark matter candidates that might be produced at the Large Hadron Collider, and effective co-utilization of resources (computational, storage, networking) for HEP data analysis in a grid environment accounting for differential policies within and across multiple large virtual organizations. Cavanaugh has a PhD in physics from Florida State University. Contact him at cavanaugh@phys.ufl.edu.

Julian James Bunn is a faculty associate and senior scientist at the Center for Advanced Computational Research at Caltech. His research interests include Web-based information systems, petabyte-scale distributed databases, grid middleware, high-performance computing and networking, object-oriented technology, and biological systems modeling. Bunn has a PhD in physics from the Universities of Sheffield and Manchester in England. Contact him at julian.bunn@caltech.edu.

Iosif Legrand is a senior software engineer at Caltech, Division of Physics, Mathematics, and Astronomy. His research interests include distributed network services, monitoring systems, and grid-related activities. Legrand has a PhD in physics from the National Institute for Physics, Bucharest. Contact him at iosif.legrand@cern.ch.

Steven H. Low is an associate professor at Caltech, where he leads the FAST Project, and a senior fellow of the University of Melbourne. His research interests are in the control and optimization of networks and protocols. Low has a PhD in electrical engineering from the University of California, Berkeley. Contact him at slow@caltech.edu.

Dan Nae is a network engineer at Caltech. His research interests include network protocols, control, and monitoring. Nae has an MSc in computer science from Politehnica University of Bucharest. Contact him at dan.nae@cern.ch.

Sylvain Ravot is a network engineer at Caltech, Division of Physics, Mathematics, and Astronomy, and is currently based at CERN in Geneva. His research interests include protocols for fast long-distance networks, high-performance networking, and network monitoring. Ravot has a degree in communication systems from the Swiss Federal Institute of Technology (Lausanne). Contact him at sylvain.ravot@cern.ch.

Conrad D. Steenberg is a scientific software engineer at the Space Radiation Lab. His research interests include improving access to, and analysis of, HEP data through the use of high-speed networks and grids. Steenberg has a PhD in cosmic ray physics from the University of Potchefstroom in South Africa. Contact him at conrad@hep.caltech.edu.

Xun Su is a network engineer at Caltech, where he's working on several networking-related projects for the Large Hadron Collider experiments. His research interests include network modeling and measurement, protocol design and analysis, wireless networking, and peer-to-peer systems. Su has a PhD in electrical engineering from the University of Texas, Austin. Contact him at xsu@hep.caltech.edu.

Michael Thomas is a scientific software engineer in the HEP group at Caltech. His research interests include grid-related software for the CMS and other high-energy physics experiments at CERN, as well as classical encryption techniques and Tcl-related open-source software. Thomas has a BS in physics from Caltech. Contact him at thomas@hep.caltech.edu.

Frank van Lingen is a scientific software engineer with the HEP group at Caltech. His research interests include distributed systems, high-performance computing, graphs in distributed systems, and software design. van Lingen has a PhD in computer science from the Eindhoven University of Technology, the Netherlands. Contact him at fvlingen@caltech.edu.

Yang Xia is a network engineer at Caltech, Division of Physics, Mathematics, and Astronomy. His research interests include network and data transfer behavior over high-bandwidth/latency networks. Xia has MS degrees in electrical engineering and physics from the University of Missouri. Contact him at yxia@caltech.edu.

Shawn McKee is a high-energy astrophysicist in the University of Michigan physics department. His research interests include particle astrophysics, neutrino physics, HEP, and observational cosmology. McKee has a PhD in physics from the University of Michigan. Contact him at smckee@umich.edu.