CTWatch QUARTERLY

http://www.ctwatch.org/quarterly/ ISSN 1555-9874 Volume 1 Number 2 May 2005

THE CYBERINFRASTRUCTURE BACKPLANE: THE JUMP TO LIGHT SPEED

Cyberinfrastructure Technology Watch
http://www.ctwatch.org/

Introduction

GUEST EDITORS
Philip Papadopoulos
Director, Advanced Cyberinfrastructure Lab, San Diego Supercomputer Center

Larry Smarr
Harry E. Gruber Professor, Department of Computer Science and Engineering, Jacobs School of Engineering and Director, California Institute for Telecommunications and Information Technology, University of California San Diego

Welcome to the second issue of the Cyberinfrastructure Technology Watch Quarterly. In this issue we focus on the state of one and 10 Gbps long-haul, optical circuits supporting the research community. It has been over a decade (1994) since the Very High-Speed Backbone Network Services (vBNS) connected major NSF centers and universities at OC-3 (155 Megabits/Sec Mbps). Soon thereafter, selected circuits were upgraded to OC-12 (622Mbps). In 1997, Internet2[1] was formed to connect a larger collection of universities at OC-12 and Gigabit speeds. Internet2 now operates with 10 Gigabit backbones. However, according to sites such as NASA's ENSIGHT,[2] end–to-end file transport from major scientific data repositories to end users laboratories across the shared internet is less than 1% of that: typically 50-100 mbps.

In the late 1990s, the future looked promising for long-haul research networks, and in three years vBNS/Internet2 dramatically increased long-haul research network capacity. But in 2000, the "Tech Bubble" turned into the "Tech Meltdown." During the Bubble phase, fiber-strands were being laid worldwide at the rate of over 8,000 km/hour (70 million kilometers in 1999)! Multiplying this build-out, Dense Wave-Division Multiplexing (DWDM) enabled carriers to run multiple 10-gigabit or even 40-gigabit channels (each termed a "lambda" for the wavelength bin in the infrared band it occupies) on a single fiber pair. This meant that the carriers had a functional way to multiply bandwidth without expensive trenching of new fiber. Current DWDM technology supports up to eighty 10-gigabit "waves" or "lambdas" on such networks (a mere 800 times the capacity of the Internet2 backbone of eight years ago).

As a number of long-haul network companies went bankrupt, a new opportunity became available because of the bandwidth overbuild—telecom carriers were willing to discuss long term leases of fiber or lambdas to individuals. Foreign countries took the lead, with Canada's CANARIE being in the vanguard. In the United States, NCSA, Argonne National Lab, EVL at UIC, and Northwestern convinced the state of Illinois in 1999 to extend the Illinois Century Network to construct a dark fiber state network, called I-WIRE,[3] to support Illinois researchers' needs for large amounts of bandwidth. In 2001, the Distributed TeraGrid Facility (DTF) proposed connecting large data nodes and computing clusters on a national scale using a 40 Gigabit dedicated backbone (four 10 Gb lambdas) among the four originating centers to form what we know now as the TeraGrid.[4] Also in 2001, NSF funded the international "point of entry" for research networks STARTAP to become StarLight[5] --a 1GigE and 10GigE switch/router facility for high-performance access to participating networks and a true optical switching facility for wavelengths.

We see these three state, national, and international initiatives as the catalyzing events. Researchers worldwide were convinced that large, dedicated optical circuit research networks were not only theoretically possible but were practically being put into service. A scant four years after the original TeraGrid award and three years after I-WIRE became operational, there are now eight sites connected to the extended TeraGrid, using the new formed National LambdaRail (NLR)[6] to create extensions of the original four lambdas, and by now over two dozen state and regional dark fiber networks exist and are interconnecting to NLR.

In this issue, we are fortunate to have three articles whose authors have all played critical roles in this new age of long-haul research networks. They delve deeper into the details and provide critical insights:

[1] http://www.internet2.org/

[2] http://ensight.eos.nasa.gov/active_net_measure.html

[3] http://www.iwire.org/

[4] http://www.teragrid.org/

[5] http://www.startap.net/starlight/

[6] http://www.nlr.net/

Introduction

Linda Winkler from Argonne National Laboratory has been in the trenches for both Teragrid and SciNet (the SCxy Conference's big monster network that exists for 5 days every November). In her article, "Does 10G Measure Up?", she describes the challenges of high-end research deployments and highlights heroic efforts in the Bandwidth Challenge where current state-of-the-art clocks in at over 100Gbps for an application using multiple 10G networks at SC2004. Linda also takes us through some of the Teragrid infrastructure.

In an article titled "The National LambdaRail" by Dave Farber and Tom West, the authors describe how regional research networks have taken advantage of the abundance of dark fiber to enable multiple, research-focused 10 Gigabit networks. Farber and West give a very nice condensed history of high-speed networking, how a variety of environmental factors made the NLR possible, how the NLR is being used today and what researchers might expect looking out 10 years.

Fast Research Networks are not a US-only concession. In fact, some would say that the US is a fast follower in the on-demand, lambda-based network. In "Translight, A Major US Component of GLIF", Tom DeFanti, Maxine Brown, Joe Mambretti, John Silvester, and Ron Johnson describe the optical interconections available between U.S. and international researchers. Truly, research networks have gone global and big/fast research networks are becoming prevalent. International partners are critical in the world of Team Science.

Here at UCSD, we are building a campus-level OptIPuter[7] interconnecting five laboratories and clusters of three functional types: compute, storage, and scalable tiled display walls. The total number of nodes in the OptIPuter fabric exceeds 500 and each lab has four fiber pairs that connect it to a central high-speed core switching complex that has both a Chiaro Enstara[8] (a large router based on a unique optical core) and standard Cisco[9] 6509 switch-router. The current instantiation supports both 10 gigabit and one gigabit signals running from each lab to the central core. In a follow-on NSF-funded proposal, Quartzite augments this structure by adding DWDM signaling on the established fiber plant, a transparent optical switch from Glimmerglass, and in 2006, a wavelength selective switch from Lucent. When complete, the Quartzite switching complex will be able to switch packets, wavelengths or entire fiber paths, allowing us to build different types of network layouts and capabilities to test OptIPuter research and other optical networking ideas. With reconfigurable networks and clusters, OptIPuter/Quartzite forms a campus-scale research instrument. At build-out, this instrument will support nearly half a Terabit of lambdas landing into a central, reconfigurable complex.

[7] http://www.optiputer.net/

[8] http://www.chiaro.com/

[9] http://www.cisco.com/

OptIPuter and Quartzite preview what campuses need to evolve to: immense bandwidth, optical circuits on demand, and reconfigurable endpoint systems. Of critical importance is the evolution of large and network-capable storage clusters that can be accessed with clear paths from research labs scattered around campus. Using cluster management systems (we use the Rocks Clustering Toolkit[10]), most scalable systems (compute clusters, tiled display clusters, application servers) can be thought of as soft-state. However, as science moves to the inevitable data-intensive modes, information storage is critical to the campus research enterprise. This coming generation of campus networks allows storage silos (critical state) to be remote from labs, then managed and operated on behalf of researchers without losing performance or adaptability for the research scientists themselves. In essence, soft-state systems can be put anywhere on campus (notably in labs), and critical-state systems are not required to be in physical proximity. "Unlimited" campus network capacity allows universities to co-optimize the preservation of critical data and the ability to rapidly change soft-state systems to meet research challenges.

[10] http://www.rockclusters.org/

Introduction

We'd like to close with the following thought. Long haul, fast research networks are springing up everywhere and bandwidth is finally meeting the "it will be abundant" predictions that many of us have believed for nearly a decade. However, the missing link overall is the campus connectivity – some campuses are pioneering big networks, but most still operate on one gigabit backbones. It is a strange turn of events when the long-haul network is fatter and more capable than your campus network.

Arden Bement, the director of the National Science Foundation recently discussed this issue in the Chronicle of Higher Education.[11]

> *"Research is being stalled by 'information overload,' Mr. Bement said, because data from digital instruments are piling up far faster than researchers can study. In particular, he said, campus networks need to be improved. High-speed data lines crossing the nation are the equivalent of six-lane superhighways, he said. But networks at colleges and universities are not so capable. "Those massive conduits are reduced to two-lane roads at most college and university campuses," he said. Improving cyberinfrastructure, he said, "will transform the capabilities of campus-based scientists."*

[11] Kiernan, V. "NSF Has Plan to Improve 'Cyberinfrastructure,' but Agency's Director's Gives Few Details," The Chronicle of Higher Education, Volume 51 (36), May 2005. http://chronicle.com/prm/weekly/v51/i36/36a03001.htm

Does 10G Ethernet Measure Up?

Linda Winkler
Argonne National Laboratory

Introduction

Major challenges facing the scientific R&E community today involve insatiable needs for communications and collaborations at distance, as well as the ability to manage globally distributed computing power and data storage resources. Advances in telecommunications and networking technologies are up to the challenge of meeting these ever increasing demands for bandwidth.

Not long ago, it was common to find large disparities between the connection speeds of end systems to the local area network versus the speed of the shared wide area network. Fortunately, that is no longer common and instead we find end systems connected at speeds of 100 Mb/s to 1 Gb/s and wide area links operating at 1 Gb/s to multiples of 10 Gb/s. Many factors have contributed to the technology boom that resulted in wide area speeds becoming attainable. The dot com boom saw a huge build-out of infrastructure and technology only to have the bottom drop out of the marketplace. Many R&E organizations have taken advantage of the marketplace to acquire access to dark fiber and build customer owned and operated metropolitan, regional and wide-area infrastructures. This has put the fate of bandwidth availability squarely within the control of the end customers rather than the service providers and carriers. The result is a strategic opportunity for the R&E community.

Making multi-gigabit/second, end-to-end network performance achievable will lead to new models for how research and business are conducted. Scientists will be empowered to form virtual organizations on a global scale, sharing information and data in flexible ways, expanding their collective computing and data resources. These capabilities are vital for projects on the cutting edge of science and engineering, in data intensive fields such as particle physics, astronomy, bioinformatics, global climate modeling, geosciences, fusion, and neutron science.

This article addresses Ethernet advances leading to network convergence, end-to-end performance factors, and highlights examples of 10G early adoption, deployment, and wide-area infrastructure availability in the R&E community.

LAN/MAN/WAN speeds

We are witnessing the convergence of LAN/MAN/WAN data rates at 10 Gb/s, resulting in common equipment and interfaces to access the enterprise, metro and wide area network. As a result, we are experiencing reductions in the cost for implementation, ownership, support, maintenance and in some cases, recurring charges in the WAN.

Ethernet technology is currently the most deployed technology for high-performance LAN environments. Enterprises around the world have invested in cabling, equipment, processes, and training in Ethernet. In addition, the ubiquity of Ethernet keeps its costs low, and with each deployment of next-generation Ethernet technology, deployment costs have trended downward. Ethernet has gained worldwide acceptance as an enterprise infrastructure technology due to its considerable advantages in interoperability, scalability, simplicity, consistency, service ubiquity, provisioning speed, and price/performance. With the rapid price

decline in Gigabit Ethernet network interface cards, most servers come standard with Gigabit Ethernet network interface cards.

An Ethernet infrastructure supporting traditional networking applications, network storage, and clustering, enables greater compute density and physical consolidation of resources. Industry standardization of infrastructure components offers economies of scale that drive down deployment and management costs and leverage common training requirements. The simplified infrastructure reduces inventory and maintenance costs. Reconfigurable components offer the flexibility to make on-demand infrastructure changes.

A natural fit for 10G Ethernet technology is in the scalable uplink from the data center switches that connect server farms with 1G Ethernet interfaces. The price per gigabit for 10G Ethernet is projected to be 40% lower than for Gigabit Ethernet. As an open-standards-based, forward- and backward-compatible technology, Ethernet has been broadly adopted and understood by engineers worldwide. So instead of building networks of increasing complexity, and facing the increasingly more difficult task of finding engineers trained to manage them, enterprises and service providers can leverage the simple and familiar infrastructure that is second nature to a large majority of network engineers.

10G Ethernet also meets several criteria for efficient and effective high-speed network performance, which makes it a natural choice for expanding, extending, and upgrading existing Ethernet networks: A customer's existing Ethernet infrastructure is easily interoperable with 10G. Existing Ethernet standards, such as 802.1q for virtual LANs, 802.1p for traffic prioritization and 802.3ad for link aggregation, also apply to 10G Ethernet, making the deployment of 10G Ethernet simply plug-and-play for most enterprises and service providers. The new technology provides lower cost of ownership including both acquisition and support costs versus current alternative technologies. Using processes, protocols, and management tools already deployed in the management infrastructure, 10G draws on familiar management tools and a common skills base. Multi-vendor sources of standards-based products provide proven interoperability.

In the metropolitan area network (MAN) the predominant leader continues to be 1Gb/s Ethernet; however, 10 Gb/s metro systems are beginning to emerge as prices of optical components continue to drop. 10 Gigabit Ethernet is on the roadmap to enable cost-effective, Gigabit-level connections between customer access gear and service provider POPs in native Ethernet format, simple, high-speed, low-cost access to the metropolitan optical infrastructure, metropolitan-based campus interconnection over dark fiber, targeting distances of 10 to 40 km, and end-to-end optical networks with common management systems. While there are pockets of new MANs, many locations are still waiting for better market conditions. MANs implemented using Ethernet technology are inexpensive and ideal for seamlessly interconnecting distributed Ethernet LANs because they require no protocol conversion.

The IEEE 802.3ae 10G standard, ratified in mid 2002, features an interface speed at 10 Gb/s at the media access layer, along with two families of Physical Layer Specifications (PHY): LAN PHY operating at 10 Gb/s and WAN PHY operating at 9.29 Gb/s compatible with the payload of OC-192c/SDH. The 10G standard not only increases the speed of Ethernet from 1 Gb/s to 10 Gb/s, but also extends its interconnectivity and its operating distance up to 40 km. Using 10G WAN PHY allows service providers to use the installed-base of SONET Layer 1 transport gear to provision 10G Ethernet traffic. Because the 10G Ethernet WAN PHY avoids the costly aspects of the traditional SONET, such as stringent grid laser specifications, jitter requirements

and stratum clocking, it offers a compelling alternative to traditional SONET interfaces with better price/performance. The ability to send Ethernet directly from an Ethernet switch over a WAN PHY link eliminates the need for expensive Packet over SONET router interfaces.

It's important to note that 10G LAN systems offer fewer alarms and indicators than 10G WAN systems. Unlike on the WAN side, there are currently no explicit standards that prescribe techniques for carrying forward error correction (FEC) to extend the reach of 10G LAN PHY. As a result, equipment manufacturers are developing proprietary interfaces to carry 10G LAN PHY with FEC for metro applications. This lack of explicit standard raises the question of interoperability and third party testing. Efforts are under way to develop standards that will make Ethernet services "carrier class" by incorporating operations, administration, and maintenance capabilities.

End-to-end Issues

Before 10G Ethernet will gain broad adoption, some technology barriers in end systems need to be addressed. For servers, a major issue is protocol processing. Using conventional network interface card (NIC) architecture simply scaled to 10G would result in the CPU's processing power being the bottleneck. Ongoing efforts seek to offload some TCP processing (TOE) from the system CPU onto the NIC hardware.

Another source of processing overhead is data copying. In a conventional networking stack, incoming packets are stored in operating system memory and later copied to application memory. The copy function consumes CPU cycles and introduces delay. For parallel processing applications that use small buffers, data copying is a major performance hit. Commonly known as iWARP, the protocols for RDMA-over-IP will enable data to be written directly into application memory, eliminating costly copy operations. For applications which use small packets, 10G NICs that implement iWARP will provide lower latency by eliminating memory copies.

10G Ethernet performance has been constrained by the limits of end system interfaces and I/O interconnects. First-generation 10G NICs with partial TCP offloads and PCI-X system interface delivered peak performance of 6-8 Gb/s. Using large packet sizes, these NICs consume less than 100% of a typical server CPU. Second generation 10G NICs with TOE are available and achieve throughput similar to first generation NICs while lowering CPU utilization. Third-generation 10G NICs should achieve full line rate with large packets when combined with end systems with a 3GIO I/O interconnect such as PCI Express.

The smooth inter-working of 10G interfaces from multiple vendors, the ability to successfully fill 10 Gb/s paths both on local area networks, cross-continent and internationally, the ability to transmit greater than 10 Gb/s from a single host, and the ability of TCP offload engines to reduce CPU utilization all illustrate the maturity of the 10 Gb/s Ethernet market. The current performance limitations are not in the network but rather in the end systems.

SC Conference

The annual International Conference for High Performance Computing and Communications (SC)[1] is co-sponsored by ACM SIGARCH and the IEEE Computer Society in November each year. Networks are an integral piece of modern high performance computing. SCinet is the very high-performance network built to support the SC conference.

[1] http://www.nlr.net/

SCinet features both a high-performance production-quality network and an extremely high performance experimental network connecting to all the major national scientific networks and supercomputer centers. 2001 was the first year SCinet deployed two pre-standard 10G LAN interfaces in the showfloor production LAN. In 2002, 10 10G LAN interfaces were deployed. In 2004, 48 10G LAN interfaces were used to satisfy bandwidth requirements.

The Bandwidth Challenge event held during SC invites participants to stress the SCinet network infrastructure while demonstrating innovative applications across the multiple research networks that connect to SCinet. The ability to maximize network throughput is an essential element to the success of high performance computation. The primary measure of performance is the verifiable network throughput. In the five year history of the Bandwidth Challenge during SC, the peak throughput achieved by the winning individual application are shown below-

2000: **1.7 Gb/s** 2001: **3.3 Gb/s** 2002: **16.8 Gb/s** 2003: **23.2 Gb/s** 2004: **101 Gb/s**

Achievable bandwidth rates are directly related to the number and capacity of WAN circuits brought into the SC venue. You may ask "why is the bandwidth challenge significant?" The Bandwidth Challenge 1) offers an opportunity to test the next generation network capacity as early as two years before production; 2) provides the opportunity to test software ideas that will be required to make use of the next generation network two years in advance; 3) creates an opportunity to test future network engineers giving them a two year lead on the problems with future networks.

National 10G Resources

National Lamba Rail (NLR) Inc is a consortium of leading U.S. research universities and private sector technology companies "lighting" a national networking infrastructure to foster the concurrent advancement of networking research. Simultaneously, NLR will enable the next generation of network-based applications in science, engineering and medicine.[2]

[2] http://www.nlr.net/

NLR is the first national scale network to deploy transcontinental circuits based upon ubiquitous Ethernet technology end-to-end. The use of 10G LAN PHY standards-based facilities in NLR represents a generational shift in the nature, usability and cost of technologies in long-haul circuits. This is a powerful capability that enables the allocation of affordable, independent, dedicated, deterministic ultra-high performance network services for research projects.

Marking a new era in control over and accessibility to national-scale optical networking capabilities for the U.S. research community, the Electronic Visualization Laboratory (EVL) at the University of Illinois at Chicago (UIC) has acquired a dedicated 10G LAN PHY circuit on the NLR infrastructure from Chicago to San Diego via Seattle. The 3,200-mile wavelength, known as the CAVEwaveTM, will initially support the National Science Foundation-funded OptIPuter project shared between UIC and the University of California, San Diego.

"CAVEwave provides researchers with a deterministic network, with guaranteed bandwidth, schedulable times and known latency characteristics, in order to understand requirements for the real-time visualization, analysis and correlation of terabytes and petabytes of data from multiple storage sites," explained EVL director Tom DeFanti. "All this bandwidth,

supplements our existing network infrastructure, for less than the cost of a 32-node cluster at each end!"

The OptIPuter[3], so named for its use of Optical networking, Internet Protocol, computer storage, processing and visualization technologies, is an envisioned infrastructure that will tightly couple computational resources over parallel optical networks using the IP communication mechanism. The OptIPuter exploits a new world in which the central architectural element is optical networking, not computers - creating "supernetworks". This paradigm shift requires large-scale applications-driven, system experiments and a broad multidisciplinary team to understand and develop innovative solutions for a "LambdaGrid" world. The goal of this new architecture is to enable scientists who are generating terabytes and petabytes of data to interactively visualize, analyze, and correlate their data from multiple storage sites connected to optical networks.

[3] http://www.optiputer.net/

Conclusion

Ethernet has withstood the test of time to become the most widely adopted networking technology in the world. Due to its proven low implementation cost, reliability, and relative simplicity of installation and maintenance, Ethernet's popularity has grown to the point that nearly all traffic on the Internet originates or terminates with an Ethernet connection. As the demand for ever-faster network speeds has increased, the Ethernet standard has been adapted to handle these higher speeds. 10G Ethernet is the natural evolution of the well-established IEEE 802.3 standard in speed and distance. In addition to increasing the line speed, it extends Ethernet's proven value set and economics to metropolitan and wide area networks by providing: lowest total cost-of-ownership; straight-forward migration to higher performance levels; proven multi-vendor interoperability; and a familiar network management interface.

The10G WAN PHY standard allows service providers to use the installed-base of SONET Layer 1 transport gear to provision 10G Ethernet traffic. For the customer, this eliminates the need for expensive Packet over SONET router interfaces, lowering the barrier to entry into the 10G WAN market.

Third generation 10G NICs combined with 3GIO I/O subsystems in end systems promise to deliver full line rate performance for servers. The maturing of the 10G Ethernet market is demonstrated by the smooth interoperability of 10G interfaces from multiple vendors, the ability to successfully fill 10 Gb/s paths both on local area networks, cross-continent and internationally, and the ability to transmit greater than 10 Gb/s from a single host.

Showcase events such as the SC conference have successfully demonstrated the interoperability of 10G technology as well as its capability of meeting the ever-increasing demand for bandwidth. Participants as well as attendees have an opportunity to witness the next generation network two years in advance.

NLR is probably the most ambitious research and education networking initiative since the ARPANET and the NSFnet, both of which led to the commercialization of the Internet. In the spirit of these great success stories, NLR strives to stimulate and support innovative network research to go above and beyond the current incremental evolution of the Internet. The results of such endeavors are expected to facilitate further commercial development and creation of new technologies and markets, thereby stimulating economic development and contributing to U.S. national competitiveness.

TransLight, a Major US Component of the GLIF

An Optical Web Connecting Research Networks in North America, Europe and the Pacific Rim

Tom DeFanti
University of Illinois, Chicago

Maxine Brown
University of Illinois, Chicago

Joe Mambretti
Northwestern University

John Silvester
University of Southern California

Ron Johnson
University of Washington

The US National Science Foundation (NSF) funds two complementary efforts through its International Research Connection Networks (IRNC) program – TransLight/StarLight and TransLight/Pacific Wave – that provide multi-gigabit links and supporting infrastructure to interconnect North American, European and Pacific Rim research & education networks, as well as to supplement available bandwidth that is provided by other countries.

TransLight/StarLight's mission is to best serve established US/European production science, including support for scientists, engineers and educators who have persistent large-flow, real-time, and/or other advanced application requirements. Two OC-192 circuits are being implemented between the US and Europe. One circuit is a 10 Gb/s link that connects Internet2/Abilene and the pan-European GÉANT2 via a routed network connection. The second circuit is a 10 Gb/s link that connects US hybrid networks, which can provide high performance, dedicated optical channels, such as the National LambdaRail (NLR), to similar European networks at NetherLight (configured as either one 10 Gb or eight 1 Gb switched circuits, or lambdas). Considerations related to security and measurement/monitoring will carefully be addressed under this award for both circuits.[1]

TransLight/Pacific Wave's mission is the development of a distributed Open Exchange along the US west coast, from Seattle to San Diego, to interconnect North American, Asian, Australian and Mexican/South American links.[2]

Across North America, NLR, Canada's CA*net4, and Internet2's Abilene and Hybrid Optical and Packet Infrastructure (HOPI) projects connect the combined TransLights, from New York (Manhattan Landing, or MAN LAN), to Chicago (StarLight), to Seattle (Pacific Northwest GigaPoP). Pacific Wave carries the connection from Seattle down the US west coast to Los Angeles and on to San Diego and Tijuana (via CalREN - the California Research and Education Network, which is operated by CENIC - the Corporation for Educational Network Initiatives in California). These locations are the sites that support the vast majority of international connections to the US and form the fabric by which most international networks peer and exchange traffic with Abilene and the US Federal Research Networks. The TransLight team is the global community of people and groups who have most advanced the art, architecture, practice, and science of Open Exchange interconnectivity among high-performance networks. TransLight's approach is based not just on backbone connectivity, but end-to-end connectivity and activism in advanced networking and applications, with a proven track record in attracting new technologies and stimulating collaborations, especially among leading domain scientists at end sites.

TransLight enables grid researchers and application developers to experiment with deterministic provisioning of dedicated circuits and then compare results with standard, aggregated "best-effort" Internet traffic. Multi-gigabit networks are referred to as "deterministic" networks, as they guarantee specific service attributes, such as bandwidth (for researchers who need to move large amounts of data), latency (to support real-time collaboration and visualization), and the time of usage (for those who need to schedule use of remote instrumentation or computers). Only through deployment of an integrated research and production infrastructure at network layers 1 through 3 will the various technical communities be able to address the major challenges of large-scale and complex systems research in peer-to-peer

[1] http://www.startap.net/translight/

[2] http://www.pacificwave.net/

systems, Grids, collaboratories, peering, routing, network management, network monitoring, end-to-end QoS, adaptive and ad hoc networks, fault tolerance, high availability, and critical infrastructure to support advanced applications and Grids.[3]

[3] Aiken, R., Boroumand, J., Wolff, S. "Back to the Future," CACM, Volume 46, No. 1, January 2004.

The current monoculture of the Internet is already being replaced by a diversity of options for interconnecting researchers and educators, enabled by scalable wavelength technologies. By the year 2010, this trend will have substantially transformed networking, enabling multiple additional capabilities. Large-scale sensor nets and huge scientific instruments will generate extraordinary amounts of data. Cheap 1000-processor clusters will serve globally distributed science projects, interconnecting at tens of gigabits per second, working on computational problems, data-intensive applications, and visualization of massive datasets – if and only if there are sufficient, affordable and predictable networks by then. New research activities like the OptIPuter and new deployments like NLR in the US, as well as similar activities in other countries, and economically affordable trans-oceanic submarine capacity (up to 10Gb) are rapidly becoming essential components of the research and education landscape.

The ability to schedule and reserve lambda networks using advanced grid services is creating an advanced cyberinfrastructure, termed the LambdaGrid. A production-class, application-centric LambdaGrid, comprised of electronically and optically switched circuits and advanced grid services, is being built by teams of programmers, networking engineers, electrical/computer engineers, computer scientists and discipline scientists who are attacking the challenging research issues and helping develop innovative solutions.

The Global Lambda Integrated Facility (GLIF)[4] is an international virtual organization that supports this decade's most advanced data-intensive scientific research and middleware development for the LambdaGrid. GLIF participants include National Research Networks (NRNs), countries, consortia and institutions that have adequate bandwidth for research and education production traffic, and that also have additional capacity they are willing to make available for use by global teams of discipline scientists, computer scientists and engineers. The GLIF community shares a common vision of building a new grid-computing paradigm, in which the central architectural element is optical networks, not computers, to support this decade's most demanding applications. To ensure the worldwide interoperability and interconnectivity of optical networks, GLIF has taken the lead in advanced facilities innovation and is developing architectural standards, or models, for open optical exchanges, which are being adopted by NRNs worldwide. The GLIF community is pioneering the concept of creating international, national, and regional distributed facilities, based on optical technologies, which departs from the traditional concept of a dedicated network that provides limited, non-deterministic services. For example, this new approach allows Grid applications to ride on dynamically configured networks based on optical wavelengths concurrent with normal Internet paths for the remaining traffic mix.

[4] http://www.glif.is/

TransLight, and all the IRNC awardees, participate in GLIF and provide connectivity between multi-gigabit international networks and US/GLIF participants, including NLR, ESnet/UltraScience Net, and Internet2/HOPI.

Reaching Out to Broader Communities of Interest

TransLight's goal is not only to make enough international bandwidth available to try a myriad of application-serving solutions, leveraging all nations' and science, engineering and education needs, but also to empower access to a diversity of networking strategies. TransLight

has proven that simply providing additional bandwidth with traditional networks, which provide only for a narrowly defined monolithic "best effort" service to all communities, is not a solution for long-term requirements. TransLight was the first advanced international infrastructure project that demonstrated the potential of agile optical networking in meeting the needs of these communities. TransLight has been developing powerful, sophisticated new techniques for matching specific community requirements with required infrastructure resources. TransLight continues to develop new techniques for precisely matching capability to each science and engineering research and education community served and, learning from the successes and failures of new models, continues to transfer best practices between these communities. Hybrid network services are desired by the international science community and are, in fact, required to advance science over the next decade. TransLight is using lambdas to advantage, using packets when expedient, and dedicated circuits when necessary.

International carriers have extraordinary overcapacity and they are actively seeking market development. According to an article in the Business section of the May 10, 2004 edition of the New York Times, "11 percent of available undersea bandwidth globally is being used."[5] TransLight can lead the broadest science and engineering research and education communities to exploit this bandwidth to communicate and collaborate while demonstrating to the carrier community that market opportunities will emerge.

[5] Belson, K. "Technology; New Undersea Cable Projects Face Some Old Problems", *New York Times,* Section C, Page 4, May 10, 2004.

TransLight History

TransLight was an outgrowth of the NSF Euro-Link award, which funded high-performance connections between the US and Europe from 1999 through June 2005. TransLight, conceived in 2001 and started in 2003, was and continues to be a rational global network architecture that achieves great economy of scale and provides links to the largest communities of interest with the broadest services. In 2002, the goal was to give researchers who participated in iGrid 2002[6] as much international bandwidth as they could use. We engaged network managers, carriers, artists, network engineers, computer scientists, and domain scientists. We persisted at Supercomputing (SC) conferences in November 2002 and November 2003. Those who participated in these events published many journal and conference papers based on their results.[7] By mid-2003, TransLight became the first persistent LambdaGrid.

[6] http://www.startap.net/igrid2002/ and http://www.igrid2002.org/

[7] DeFanti, T., Brown, M., De Laat, C., eds. "iGrid 2002: The International Virtual Laboratory," Journal of Future Generation Computer Systems (FGCS), Elsevier, Volume 19, No. 6, August 2003.

In 2003, in partnership with SURFnet, NSF Euro-Link funds were used to purchase an OC-192 transatlantic circuit between Chicago and Amsterdam that provided both Layer 2 and Layer 3 connectivity. This hybrid network architecture provided the research and education community with both packet-switched (Layer 3) routed paths for many-to-many usage, as well as circuit-switched (Layer 2) lightpaths (or lambdas) for high-speed few-to-few usage.

TransLight quickly became a global partnership among institutions, organizations, consortia or country NRNs who wished to make additional bandwidth on their links available for scheduled, experimental use.[8] TransLight evolved into a two-year experiment to develop a governance model of how the US and international networking collaborators would work together to provide a common infrastructure in support of scientific research. In September 2004, as more and more countries began sharing bandwidth among one another, TransLight members dissolved the TransLight Governance body in favor of having the GLIF, an international body, assume this role.

[8] DeFanti, T., De Laat, C., Mambretti, J., Neggers, K., St. Arnaud, B. "TransLight: A Global Scale Lambda Grid for E-Science," Special Issue on "Blueprint for the Future of High Performance Networking," Communications of the ACM, Volume. 46, No. 11, November 2003, pp. 34-41.

GLIF partners are now organizing institutions of the iGrid 2005 event, to be held at the new Calit2 building in San Diego, September 26-30, 2005.[9] Emphasis is on demonstrating

[9] http://www.startap.net/igrid2005/ and http://www.igrid2005.org/ (coming soon)

applications research and middleware development that utilize new architectural approaches to next-generation Internet design and development using optical networking. Communities of interest will create their own private networks or share networks, creating on-demand LambdaGrids of interconnected, distributed computing, sensor and instrument resources that enable new infrastructures for advanced science.

NSF International Research Network Connections Program

The NSF IRNC Program funds international network links to connect US and foreign science and engineering communities, encourage the investigation and incorporation of advanced architectures needed to support the advanced and developing needs of science and engineering, encourage rational development and leveraging of deployed infrastructure to meet current and anticipated needs, and enable network engineers to engage in system and technology demonstrations and rigorous experimentation. The IRNC Program supports efforts to network North America to Europe, South America, the Pacific Rim, and to a global ring (via The Netherlands, Russia, Korea and China). It also supports network measurement and monitoring.

Relevant Websites

- TransLight/StarLight - http://www.startap.net/translight/
- TransLight/Pacific Wave - http://www.pacificwave.net/
- Western Hemisphere Research and Education Networks (WHREN) - http://www.ampath.fiu.edu/
- Global Ring Network for Advanced Applications Development (GLORIAD) - http://www.gloriad.org/gloriad/index.html
- Consortium of International Research and Education Networks (CIREN) - http://www.transpac.org/
- SGER: Exploratory Research on Network Measurement for International Connections - http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0513437
- SGER: Network Measurement for International Connections - http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0457404

Acknowledgments

TransLight/StarLight receives major funding from National Science Foundation (NSF) International Research Network Connections (IRNC) program, award SCI-0441094 to the University of Illinois at Chicago (UIC), for the period February 2005 – January 2010; Tom DeFanti is principal investigator and Maxine Brown is co-principal investigator. Previous funding for Euro-Link, from the NSF High Performance International Internet Services (HPIIS) program, award SCI-9730202 to UIC, was for the period April 1999 – September 2005. SURFnet bv is a Key Institutional Partner of the TransLight/StarLight award, continuing an already long and successful partnership between the UIC, StarLight in Chicago, and NetherLight/SURFnet in Amsterdam.

TransLight/Pacific Wave receives major funding from NSF IRNC award SCI-0441119 to the University of Southern California (USC) for the period March 2005 – February 2010; John Silvester of USC is principal investigator and Ron Johnson, University of Washington, is co-principal investigator.

The National LambdaRail

Cyberinfrastructure for Tomorrow's Research and Education

David Farber
National LambdaRail
Carnegie Mellon University

Tom West
National LambdaRail

Introduction

The demands of leading-edge science increasingly require networking capabilities beyond those currently available from even the most advanced research and education networks. As network-enabled collaboration and access to remote resources become central to science and education, researchers often spend significant time and resources securing the specialized networking resources they need to conduct their research. As a result, there is less time and fewer resources available to conduct the research itself.

New technology holds the promise of providing more easily the networking capabilities researchers require. Increasingly, the best option for ensuring the technology and these capabilities are available seems to be for the research and education community to own and manage the underlying network infrastructure. This movement towards a facilities-owned approach is relatively unprecedented in the history of research and education networking, yet holds the promise for unique benefits for research and education. Ownership provides the control and flexibility, as well as the efficiency and effectiveness needed to meet research and education's uniquely demanding networking requirements.

A new global network infrastructure owned and operated by the research and education community is being developed, deployed, and used. In the United States, a nationwide infrastructure is being built by the National LambdaRail (NLR) organization, in collaboration with scientists and network researchers, with leadership from the academic community, and in partnership with industry and the federal government. Furthermore, NLR both leverages, and provides leverage for, existing and new regional and local efforts to deploy academic-owned network infrastructure.

History of Research Networking

To understand how the most recent movement in research and education (R&E) networking differs from those of the past, and how unique the capabilities it provides are, let us take a look back at how R&E networking has developed in the United States over the past 35 years. In 1987, the initial NSFNET backbone provided just 56 kilobits per second of bandwidth. Even in 1991 only 1.5 megabits per second were available on the backbone, less than many current home broadband connections. Today, nationwide R&E networks have links of 10 Gigabits per second (Gbps), nearly 7000 times their capacity just 15 years ago. Yet it is increasingly apparent that even this is not enough capacity to meet emerging demands.

It is also important to realize that, tracing the development of today's Internet back to the ARPANET of 1969, pioneers from the university community, with the support of government and industry, have provided leadership for network development to meet the needs of research and education. While many share in the development and evolution of the Internet as we know it today, university-based researchers played a key role both in developing fundamental Internet technologies and in providing large-scale testbeds that put those technologies to work and drove their further development.

In the early 1990s, the research and education community realized that it had lost control of this critical resource to the telecommunications industry. In conjunction with funding for universities' high-performance networking connections from the National Science Foundation, and in partnership with industry and other government agencies, Internet2 was formed in 1996 and launched the nationwide Abilene network as a 2.4 Gbps backbone in 1998. Regional networks extended Abilene's capabilities to university campuses, including an upgrade of Abilene to 10 Gbps in 2003.

Today, we have multiple networks for research and education running over multiple global, national, regional, and institutional infrastructures. Some of these networks, such as the Internet2 Abilene backbone, are high performance networks that are shared by researchers from numerous disciplines. Other, more specialized networks, such as ESnet, are designed to serve the needs of a subset of the entire research community.

An important common characteristic of all the R&E networks deployed over the past three decades is that most of the underlying infrastructure has not been owned by the research and education community. Rather, the networks have been built from leased circuits from traditional telecommunications companies. It has also been true that the capabilities required by leading-edge science and education have often not been available as off-the-shelf services by commercial providers. Therefore, to meet its requirements the R&E community has needed to cobble together circuits and services from multiple providers. As a consequence, research groups have historically spent significant amounts of time and energy developing and securing the networking capabilities they need before conducting the scientific research that is their end goal.

Sea of Change

There are two significant forces that are fundamentally changing the nature of research and education networking and providing an opportunity to reduce the amount of effort needed to provide scientists with the networking capabilities they need.

First, there is a growing urgency to develop new network technologies that scale to the growing needs of the worldwide R&E community and, later, to commodity Internet users. Undertaking this development requires an experimental testbed where network researchers can experiment with new approaches to all levels of networking technology. The results of this research will enable networks capable of supporting scientific projects in fields such as high-energy nuclear physics and radio astronomy, which require real-time collaboration among scientists and manipulation of enormous data sets. Already, individual projects in these fields can usefully consume a majority of the largest network links available. Together, even a few of them could potentially overwhelm existing advanced research and education networks. And, these kind of bandwidth-hungry applications are spreading. Applications in almost every discipline are now emerging with the same need for big, broadband networks.

The second big development is the fortuitous availability of dark fiber in the United States and elsewhere as a result of the downturn in the telecommunications industry. The last four years have provided the research and education community a historic opportunity to migrate from leasing circuits from traditional telecommunications carriers to owning fiber outright. This fiber, combined with optical dense wave division multiplexing (DWDM) technology enables multiple R&E networks to be built and run over the same fiber pair. Taken together, fiber ownership and DWDM change the dynamics of deploying and managing dedicated

research networks to support demanding scientific applications and large-scale network research. In short, we now have the components for owning and controlling a robust optical network infrastructure that will support multiple, disparate networks.

U.S. Optical Network Infrastructure

Nationwide networks in several other countries and continents already have leveraged the combination of optical fiber and DWDM to deploy operational network infrastructures. Notable among these are CA*Net 4 in Canada through the CANARIE organization, SURFnet6 in the Netherlands through Stichting SURF, and AARNet in Australia. And others are emerging. For example, DANTE will soon be deploying the pan-European GÉANT2 network.

In the United States, campus and regional networks have led developments in this area since around 1995. A large number of institutions, especially research institutions such as the University of California Berkeley, have established on-campus, fiber-based network infrastructures that serve multiple networks. In many instances the institution goes far beyond the physical dimensions of the main campus and reaches out to university facilities in the surrounding community. Increasingly, these facilities are developed as part of the institutional infrastructure, such as that deployed by the University of California San Diego.

At the regional level in the United States, consortia of institutions within states like Texas and Florida have formed a not-for-profit corporation to undertake regional infrastructure development. California pioneered this model beginning in 1997 with the formation of the Corporation for Education Network Initiatives in California (CENIC), which brought together public and private universities, and community colleges. CENIC's CalREN optical fiber-based infrastructure provides multiple networks to serve this wide range of constituencies. Across the United States, roughly 15,000 miles of fiber optic cable is controlled by regional network organizations. FiberCo, an organization created by Internet2, has been very instrumental in facilitating the acquisition of much of this fiber.

Although regional optical network infrastructure development emerged a few years ago, the formation of NLR has spurred a virtual explosion in the number of regional efforts. Less than three years ago the NLR was just a glimmer in the eyes of very few people in the United States. It started as a grassroots effort on the west coast to link Seattle with San Diego. It then evolved to have a redundant path via Seattle to Denver and Denver to Los Angeles. NLR evolved from a regional effort but recently NLR has stimulated new regional developments.

National LambdaRail

The mission of the NLR is to build an advanced, nationwide network infrastructure that will support many types and levels of networks for research, clinical, and educational fields. This infrastructure consists of 11,500 miles of fiber and optical networking equipment, all of it owned by NLR. The infrastructure supports both experimental and production networks, fosters networking research, promotes next-generation applications, and facilitates interconnectivity among regional and international high-performance research and education networks.

The NLR infrastructure is composed of 30 segments. Each segment can support at least 32 individual channels of light. On the northern routes, from Sunnyvale, California to Jacksonville, Florida an additional eight waves can be added. Each wave in each segment can

support 10 Gbps, so there is the potential for 1072 channel-segments, each with 10 Gbps of capacity. Equally significantly is that each channel-segment operates independently and, therefore, can support networks with different operational characteristics.

Begun in 2004, phase one of deployment is complete, with the entire infrastructure on schedule to be finished by October 2005. Already, nearly 25 percent of the capacity is in use and it is anticipated that nearly 60 percent of the total capacity will be in use by late 2007. Planning for increased capacity and enhanced capabilities is already underway to ensure that NLR is always ready to meet the most demanding requirements of the research and education community.

NLR members and associates span a wide geographic and organizational range including individual universities; boards of regents; consortiums of institutions; not for profit corporations; a supercomputing center; a limited liability corporation; and Internet2, a not-for-profit organization that represents more than 300 organizations, including all the NLR members. In stark contrast to the government support provided to most R&E networking in the United States, NLR has been funded by the direct contributions of more than two dozen members and key corporate participants. The major strategic, corporate participant has been Cisco Systems. NLR would not have happened without the commitment of Cisco Systems to provide major resources, including optical equipment, routers, and switches. Cisco also provided early and ongoing support for, and focus on, advancing the network research. Level 3 Communications and WilTel Communications, as NLR's predominate providers of fiber and related services, provided consideration in the acquisition of fiber and in providing the related services.

There are two main audiences for the NLR: network researchers and researchers involved in big science applications, including supercomputing. The focus on network researchers is a distinguishing characteristic of NLR. Fifty percent of NLR capacity is being devoted to support network research projects under the auspices of a network research advisory council led by NLR Chief Scientist David Farber of Carnegie Mellon University.

The NLR Network Research Council gathers thought leaders to guide NLR's support of network research and provides a direct and enduring link to the community at the forefront of conceiving, developing, and testing revolutionary, not just evolutionary, networking technologies and capabilities. Directly engaging the network research community ensures that as the fundamental shift in networking to increasingly leverage optical technology continues, the NLR infrastructure will continue to be in the best possible position to support the investigations of cutting-edge network research—work that is not possible in the laboratory or any other national-scale network.

NLR already provides a unique, world-class nationwide testbed for network research. Dramatic experiments in new technologies such as dynamic wavelength provisioning and quantum encryption can be conducted without concerns about interrupting production services. Furthermore, the usage and performance of existing production services that use the NLR infrastructure can be examined in detail, providing the possibility for improving the capabilities of other networks that use those technologies.

The NLR infrastructure is already being used to support national-scale projects that require capabilities that today only NLR can provide:

- The Extensible Terascale Facility (ETF) supported by the National Science Foundation, is a multi-million dollar, multi-year effort that has built and deployed the TeraGrid, a world-class networking, computing and storage infrastructure designed to engage the science and engineering community to catalyze new discoveries. The Pittsburgh Supercomputing Center, one of the original TeraGrid participants, was the first organization to use NLR to connect its facilities to the nationwide TeraGrid facility. Recently, the Texas Advanced Computing Center acquired a 10 Gbps wave from NLR to connect Austin to Chicago. Oak Ridge National Laboratory also is using NLR for back-up waves between Atlanta and Chicago as part of ETF.

- The HOPI project of Internet2 is using NLR to explore the evolution of the Internet's core. This project is engaging industry, regional, and international partners to examine a hybrid of packet switching and dynamically provisioned lambdas. It is using a wavelength on the entire NLR infrastructure footprint.

- The CENIC organization and the Pacific Northwest GigaPOP are undertaking a joint project that uses NLR infrastructure to create, deploy, and operate Pacific Wave, an advanced, extensible peering facility along the entire Pacific Coast of the United States. Pacific Wave will create a new peering paradigm by removing the geographical barriers of traditional peering facilities. Pacific Wave will enable any U.S. or international network to connect at any location along the U.S. Pacific Coast facility, as well as the option to peer with any other Pacific Wave participant regardless of their physical connection.

- The U.S. Department of Energy (DOE) UltraScience project is using NLR infrastructure to link Sunnyvale, Seattle, with Chicago. The UltraScience Net is an experimental research test bed funded by DOE's Office of Science to develop networks with unprecedented capabilities to support distributed, large-scale science applications.

- The OptIPuter is a powerful, distributed cyberinfrastructure supporting two major data-intensive scientific research and collaboration efforts in the Earth sciences and bioscience. OptIPuter is a five-year research program led by the University of California, San Diego and the University of Illinois at Chicago with several partners. NLR waves support the OptIPuter from University of California, San Diego and San Diego State University in the Southwest, to the University of Washington in the Northwest, to the University of Illinois at Chicago in the Midwest.

NLR and the Future of Research and Education

Cooperation and collaboration on common goals are the hallmark of the NLR. NLR provides a unique nationwide infrastructure that is able to provide the networking capabilities that are an increasingly critical part of the cyberinfrastructure required by the U.S. R&E community. This includes stable and reliable production networks at the regional, national, and international levels, as well as "breakable" experimental networks in support of network research. NLR also provides a locus for the symbiotic relationship between researchers using networking capabilities, and networking researchers looking to develop and test new network capabilities.

Large scale scientific applications, many driven by supercomputing, are becoming more routine. However, there is a looming collision between application requirements and network capacity. Ownership and control of the basic infrastructure can provide the most cost-effective way to meet the full range of networking needs. It provides a platform for researchers to spend the least amount of time possible working on connecting participants in large-scale research efforts.

An historic opportunity exists for the R&E community to leverage technology and achieve control over advanced network resources. This is an opportunity not only to meet today's needs but also to lay the foundation for a new round of innovation. The R&E community has historically led the way in advanced networking and it can continue to do so.

CTWatch QUARTERLY

May 2005

THE CYBERINFRASTRUCTURE BACKPLANE: THE JUMP TO LIGHT SPEED

GUEST EDITORS

Philip Papadopoulos, SDSC

Larry Smarr, UCSD

http://icl.cs.utk.edu/

http://www.ncsa.uiuc.edu/

http://www.sdsc.edu/

http://www.ctwatch.org/quarterly/ quarterly@ctwatch.org