

By Tom DeFanti, Cees de Laat, Joe Mambretti, ...... Kees Neggers, and Bill St. Arnaud .....

# TRANSLIGHT A GLOBAL-SCALE LAMBDAGRID FOR E-SCIENCE

This global experiment wants to see if high-end applications needing transport capacities of multiple Gbps for up to hours at a time can be handled through an optical bypass network.

ata-intensive e-science is far removed from transaction-based e-business and streaming-based e-entertainment, yet today's best-effort routed networks must serve all users. But these networks are swamped by huge flows of data mixed with normal-bandwidth, short-

lived traffic. Each type of traffic has a devastating effect on the other, Web pages take too long to open, and data is lost, requiring retransmission, especially when the networks do not offer scheduled services with guarantees of bandwidth or latency.

TRIANGULUM GALAXY. IMAGE COMBINES DATA FROM OPTICAL TELESCOPES ON KITT PEAK IN ARIZONA AND RADIO DATA FROM THE VERY LARGE ARRAY IN NEW MEXICO, THE ROBERT C. BYRD GREEN BANK TELESCOPE IN WEST VIRGINIA, AND THE WESTERBORK SYNTHISIS RADIO TELESCOPE IN THE NETHERLANDS. OPTICAL DATA (RED AND YELLOW HUES) SHOWS LIGHT FROM STARS IN THE GALAXY; RADIO DATA (BLUE TINTS) SHOWS DISTRIBUTION OF COOL GAS. (TRAVIS RECTOR, NATIONAL RADIO ASTRONOMY OBSERVATORY AND NATIONAL OPTICAL ASTRONOMY OBSERVATORY, DAVID THILKER, JOHN HOPKINS UNIVERSITY, AND ROBERT BRAUN, ASTRON, THE NETHERLANDS)

t National Science Foundation-sponsored workshops, e-scientists routinely request schedulable high-bandwidth, low-latency connectivity "with known and knowable characteristics" [2, 8]. Many of them, along with collaborating computer science researchers, have explored international multi-gigabit connectivity at special high-performance events like iGrid 2002 [3] in Amsterdam, The Netherlands. In response, government-sponsored research groups, including network engineers, in several nations are building TransLight, a global system of optical networks to move massive amounts of data directly under e-scientists' control (see Figure 1). TransLight aims to serve motivated scientific communities, notably high-energy physics, biology, astronomy, and Earth observation, with the services

they need well in advance of counterpart publicly available telecom carrier offerings. Trans-Light complements but does not replace global production research and education university- and government-supported best-effort networks.

Rather than emerging from traditional commercial telecommunications carriers, TransLight is being designed and implemented by interdisciplinary teams of scientists and engineers in government-funded research labs and universities in Europe, Japan, and North America (see the sidebar "TransLight Institutional Members and Bandwidth Contributions"). Traditional networks are complex at the core and governed by

hierarchical centralized management systems. In contrast, TransLight has a powerful, simplified core, with highly distributed peer-to-peer management systems. TransLight represents a paradigm shift that can be exploited for e-science years in advance of any market acceptance.

A lambda, in network parlance, is a fully dedicated wavelength of light in an optical network, capable of greater than 10Gbps bandwidth; lambda means large, desirable units of networking and is how the application scientists view them. A Grid is a set of distributed, networked, middleware-enabled computing, storage, and visualization resources. A LambdaGrid is a Grid in which lambdas form end-to-end connections that form connections (lightpaths) among com-

puting resources. The lambdas themselves are also treated by the control software (middleware) as allocatable Grid resources. Scores of lambdas are deployed in the form of a LambdaGrid. TransLight partners are providing at least 70Gbps of electronically and optically switched circuits (organized into Gigabit Ethernet, or GigE, channels, but are expected to grow to 10GigE channels, as the technology becomes more affordable and available) among major hubs in Chicago and Amsterdam, along with other cities in the figure. (The TransLight international experiment was first proposed by the authors and Michael McRobbie of Indiana University several years ago.)

Additional lambdas are expected from Europe and Japan, including donations through the Internet Educational Equal Access Foundation (see www. ieeaf.org). These circuits are available for scheduled

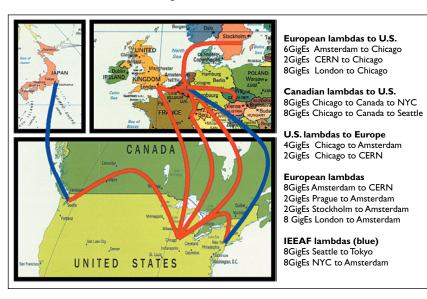


Figure 1. TransLight lambdas.

use by the e-science community to move unprecedented amounts of data in reasonable timeframes (see the article by Newman et al. in this section).

The long-term goal is to discover novel uses of lambdas (such as providing congestion-less backplane interconnects for distributed computational clusters) serving as transient memory buffers for distributed storage and computational devices, perhaps to the point of conceiving post-von Neumann computer architectures. TransLight's switched optical networks promise vastly increased transport capacity with predictable latency, determined largely by the speed of light, and development of new methods of provisioning that offer control of lightpaths, their characteristics, and traffic behavior to the application level.

It is unclear whether routers will ever provide deterministic guaranteed bandwidth; huge marketing and technical barriers stand in the way of offering

## TRANSLIGHT'S SWITCHED OPTICAL NETWORKS PROMISE VASTLY INCREASED TRANSPORT CAPACITY WITH PREDICTABLE LATENCY, DETERMINED LARGELY BY THE SPEED OF LIGHT.

widespread guarantees. However, disrupting production-routed research and education production networks to find out is not conscionable; moreover, convincing proofs-of-concept for new technology

would never emerge from modest laboratory-scale testbed networks or simulations alone. Perceiving the need, then, for a flexible global-scale laboratory, and bringing together international resources, TransLight is being created to enable applications Grid interact with network elements treated by e-scientists' software as Grid resources. These resources can be reserved, concatenated, consumed, released, repeated if need be, and reused. Applications can

control and characterize the performance of bulk data transfer over end-to-end lightpaths.

Radio astronomers use very long baseline interferometry (VLBI) projects to study radio signals from deep space. VLBI, as the name implies, involves connecting telescopes far apart to central signal-correlation processors; the angular resolution improves with the distance among the telescopes used in the analysis. The value to the astronomer emerges in the correlation, so the signals cannot usefully be preprocessed at

the telescopes. Today's mode of operation is to digitize the radio signals and put them on many high-capacity magnetic tapes, dismount them, then transport them to and remount them at the processing site.

> Estimations required bandwidth per radio receiver are approximately 8Gbps per telescope. Sending amount of data through a routed research network backbone is a waste of routing capacity and expense; a more efficient switched connection is justified. Distributing high-energy physics data also involves extraordinary data processing needs, as does the OptI-

Puter with its biomedical and geoscience applications (see the article by Smarr et al. in this section); additional high-bandwidth applications include data mining, emergency response, military imaging, and digital cinema (Hollywood).

The links in the experimental TransLight network can be scheduled by e-science researchers on a regular basis. The countries/consortia/networking groups paying for these links "donate" a portion of their bandwidth to the global experiment. The initial TransLight collection of international links is built from 70Gbps of capacity, as outlined in Figure 1. Additional lambdas accessible to the research community are expected from Europe and Japan, including

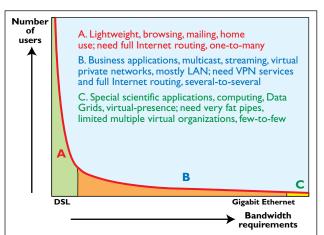


Figure 2. Numbers of Class A, B, and C users and their bandwidth appetites.

### TRANSLIGHT IS AN EXPERIMENT TO DETERMINE IF THE ENORMOUS LOADS ANTICIPATED FROM CLASS C USERS CAN BE TAKEN OUT OF BAND AND HANDLED THROUGH AN OPTICAL BYPASS NETWORK.

donations through the Internet Educational Equal Access Foundation. A TransLight Governance Board meeting three times a year at international networking conferences makes collective policy regarding usage. To access these international lambdas, national, regional, and metro lambdas are connected to the two initial international hubs—Amsterdam's NetherLight and Chicago's StarLight.

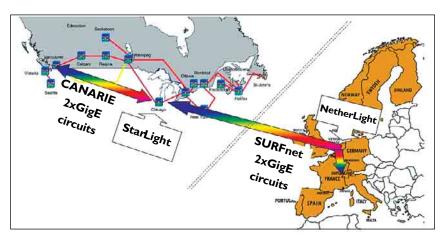


Figure 3. Lambdas among CERN, NetherLight, StarLight, and Vancouver during iGrid2002, Sept. 2002. A terabyte of research data was transferred disk-to-disk in less than three hours over a newly established lightpath extending 12,000 kilometers from TRIUMF (Tri-University Meson Facility) at Canada's National Laboratory for Particle and Nuclear Physics in Vancouver to CERN in Geneva.

#### **Three User Communities**

Network users are classified into three communities (see Figure 2) [4]. Class A includes the typical home user with Mbps DSL or cable-modem speeds who expects commodity consumer services, including good Web access, email with MB-size attachments, downloads of streaming media, messaging, and peer-to-peer (music, gaming) applications. Class A users typically need full Internet routing; their flows are generally short-lived and go from anywhere to anywhere (and back). Class B consists of corporations, enterprises, universities, and Grid-based vir-

tual organizations and laboratories operating at Gbps LAN speeds. Class B connectivity uses mostly switched services, virtual private networks, and full Internet routing uplinks, often through firewalls. This community typically needs protected environments, many-to-many connectivity, and collaboration support. Most of the traffic typically stays within the virtual organization. However, class B

users are also connected to perhaps several thousand other sites via routed high-performance networks, some dedicated to specific communities. Class C represents a few hundred truly high-end applications coming this decade that need transport capacities of multiple Gbps for from minutes to hours, originating from only a few places, destined to arrive at only a few other places. Class C traffic does not require routing, since it always takes the same route

from source to destination. However, it does require dynamic path provisioning, because most of the applications require the gathering and utilization of resources at several sites.

If we estimate for the moment that the backbone load of the total sum of class A is of the same order of magnitude as the class B traffic in a region—approximately 1Gbps, indeed what a typical business or university provider's infrastructure is built support—then the needs of a 5Gbps class C user is a distinctly disruptive event. Providers start to plan for upgrading when their lines and interfaces are loaded with traffic exceeding 33%-50% of their built capacity. Yet it may not make sense to invest in another full round of bigger, faster routers if the major new load on a backbone comes from class C users who do not need full Internet routing. TransLight is an alternative, an experiment to determine if the enormous loads anticipated from class C users can be taken out of band and handled through an optical bypass network—one that uses a scheduled LambdaGrid to do the heavy transport.

TransLight hubs, including StarLight and NetherLight, make rack space, power, and connections available for computers and servers, as well as for networking equipment, so first experiments now taking place have not had to wait for completion of fiber

to e-scientists' labs. Some experiments, including large-scale data caching and optical multicasting (splitting data to go on several paths simultaneously), are best done from a hub. Success might motivate wide deployment of fiber, and thus lambdas "to the lab" this decade.

#### TransLight Institutional Members and Bandwidth Contributions

ere are the initial TranLight member organizations and their contributions to overall project bandwidth:

CA\*net4. This fourth-generation Canadian Network for the Advancement of Research, Industry, and Education (CANARIE) includes trans-Canadian wavelengths and dual OC-192 connectivity into StarLight to support TransLight. Eventually, it will provide 10Gbps in Layer 2 circuits to connect Chicago to New York, Chicago to Seattle, and Chicago to Canadian locations of its choosing (see www.canarie.ca/canet4/).

CERN. The European Laboratory for Particle Physics (CERN) provides experimental facilities and networks for particle physics experiments, mainly in the domain of high-energy physics. It serves as the prime contractor and manager of the European Union's TransAtlantic Grid, or DataTAG, initiative to enable next-generation scientific exploration requiring intensive computation and analysis of shared large-scale databases across distributed scientific communities. The DataTAG link makes it possible to send and receive 2.5Gbps between Geneva and Chicago and is being upgraded to 10Gbps this year (see cern.ch and www.datatag. org).

NetherLight. Located on the campus of the Amsterdam Science and Technology Centre, NetherLight international connectivity includes dedicated lambdas to the StarLight facility in Chicago and to CERN in Switzerland. It is being built and operated by SURFnet, the Dutch Research Network organization, and funded via the government-sponsored GigaPort project (see www.surfnet.nl).

StarLight. This U.S. National Science Foundation-supported carrier-neutral co-location facility at Northwestern University in Chicago is run by the Electronic Visualization Laboratory at the University of Illinois at Chicago, the International Center for Advanced Internet Research at Northwestern University, and the Mathematics and Computer Science Division at Argonne National Laboratory (see www.startap.net/starlight).

NorthernLight. Located at the Royal Institute of Technology in Stockholm, Sweden, its international

connectivity consists of a dedicated lambda to
NetherLight in Amsterdam; dedicated lambdas to
Copenhagen, Helsinki, and Oslo are planned. NorthernLight is funded by NORDUnet, the Nordic research
network collaboration (see www.nordu.net).

CzechLight. Managing a dedicated 2.5Gbps circuit between Prague and NetherLight in Amsterdam, it involves plans to connect to other Central and Eastern European countries and to other Czech Republic cities. It is supported by the Czech government's Research and Development Council and run by CESNET, the country's National Research and Education Network organization (see www.cesnet.cz).

UKLight. This national facility is located at the SuperJANET operations centre at the University of London Computing Centre; fiber links to a nearby research access point are to be provided in 2004 by University College London, as well as possible links to remote sites via a national development network. It will also have links to both StarLight and NetherLight. UKLight is funded by the Higher Education Funding Council and managed via the Joint Information Systems Committee and UKERNA the U.K.'s government-funded educational and research network (see www.ja.net).

Internet Educational Equal Access Foundation. The IEEAF is a U.S.-based nonprofit organization whose mission is to obtain donations of telecommunications capacity and equipment, then make them available to the global research and education community. Through partnerships and alliances among government, private-sector entities, educational institutions, and other nonprofit organizations, IEEAF fosters global educational collaboration and equitable access to network resources via a plan called the Global Quilt (see www.ieeaf.org). The Seattle-Tokyo link is under the stewardship of the Pacific Wave in Seattle (see pacificwave.net) and the Widely Integrated Distributed Environment, or WIDE, Project in Tokyo (see www.wide.ad.jp). The New York City-Amsterdam link is under the stewardship of the University Corporation for Advanced Internet Development (see www.internet2.edu) in the U.S. and SURFnet in The Netherlands.

ONE GOAL MOTIVATING TRANSLIGHT IS TO MINIMIZE END-TO-END LATENCY AND JITTER BY NEEDING LITTLE TIME BEYOND THAT REQUIRED FOR PROCESSING AND TRANSMITTING THE LIGHT ITSELF.

#### **Scaling Networks for E-Science**

Within the past two years, 1Gb and 10Gb Ethernet (GigE) technology has come to dominate research and education networks. As 10GigE technology, available and affordable in metro areas today, becomes more widely deployed, exponential penetration in the market will likely occur, and prices will drop. An intense debate in the networking research community rages as to whether switched lambdas can possibly scale to usable numbers or if router capacity can economically scale by orders of magnitude. Today, a 10Gbps lambda between major cities in North America and across the Atlantic Ocean costs about \$120/hour (\$1 million per year). A nationwide, 20-year, multi-10Gbps capability can be built for the cost of a typical university building or two. Metro-scale fiber can be long-term leased and populated with customer-lit wavelengths; a GigE across Chicago for 20 years (as a one-time investment) costs what a typical carrier-managed service charges per month. Government-funded science programs and university research projects can afford to experiment, but these economies are realizable today if, and only if, the communities of e-scientists themselves provide all the services and manage them, as was done at iGrid 2002 (see Figure 3).

E-SCIENTISTS AND THEIR PROGRAMMERS NEED NEW or greatly improved protocols. For example, the Transmission Control Protocol (TCP), needed mainly for congestion control on shared routed networks, suffers from throughput problems. Early tests using lambdas between NetherLight and StarLight have shown that unexpected TCP traffic behaviors surface when real applications are put on the infrastructure, which was originally developed for congestion control on shared routed networks [1]. TCP is

highly susceptible to packet loss, possibly resulting from congestion or transmission errors. In order to achieve Gbps performance with standard TCP on a 4,000-kilometer path with a round trip time of 40ms and 1,500B packets, the loss rate must not exceed 8.5×108 packets. End-to-end bit error rate (BER) for optical networks is typically 1012 to 1015 hits, giving a packet loss rate of approximately 108 to 1011. One suggested way to limit the packet loss rate is to increase the size of the packets, but the larger the packet, the greater the packet loss rate for a given BER, since the probability of having a bad bit in a packet increases linearly with packet size. Consequently, with today's optical BERs, the theoretical limit of traditional TCP for achieving any significant throughput on IP-routed networks is close to being reached.

In uncongested networks, including those provided by TransLight, TCP is not as obligatory as it is in aggregated networks. A survey of transport protocols for high-performance networks, designed to circumvent some of TCP's limitations, are discussed by Falk et al. in this section.

Other factors also affect throughput performance on routed networks, the combination of which may result in poor throughput performance, even with well-tuned TCP implementations. Recent theoretical evidence from advanced queuing theory points to the possibility of an underlying pathology where routed networks become unstable if large single server-to-server flows are mixed with normal aggregated IP traffic [5]. Even the most promising and innovative of the proposed TCP enhancements do not preclude the necessity of exploiting the capacity of LambdaGrids, because dedicated lightpaths can set application performance goals and expectations for more widely available networks. These technologies will surely be used in complementary ways.

### Known and Knowable Latency and Bandwidth

Latency is an unavoidable problem on long links due to the speed of light in fiber and regeneration. Typical metro-scale latency, with no regeneration of signals, is less than 2ms round-trip time. Halfway around the globe takes about 250ms. More than 100ms of latency makes it difficult for e-scientists to do interactive visualization, simulation steering, and collaboration. Random variations in latency, common in congested networks, are particularly debilitating to collaboration system users. One goal motivating TransLight is to minimize end-to-end latency and jitter by needing little time beyond that required for processing and transmitting the light itself.

Another research focus is to identify and develop ways to enable multi-domain-path provisioning and integration of lambdas with Grid environments. This is further complicated by the complexity of coordinating lambda switching with computational Grid scheduling, disk storage availability, and collaboration, especially as the entire process becomes dynamic and the time slices become finer. Authorization, authentication, and accounting (AAA) services need to interoperate across organizational boundaries and across a variety of Internet services, enabling an AAA transaction spanning many stakeholders. This activity grows out of the work of the authorization team of the Internet Research Task Force AAA Research Group and is carried forward in the Global Grid Forum, a worldwide group of 500 researchers (see www.Gridforum.org) [6]. Canada's advanced Internet development organization, known as the Canadian Network for the Advancement of Research, Industry, and Education, or CANARIE, funds development and implementation of optical network management tools and protocols based on the new Open Grid Services Architecture and other space-based distributed protocols, including Jini and JavaSpaces, that allow end users to independently manage their own portions of a condominium wide-area optical network. The Chicago-Amsterdam team is designing Photonic Interdomain Negotiation, or PIN, middleware for applications of the OptIPuter project.

#### **Conclusion**

Connection-based optical networks pose fundamental challenges in throughput performance, capacity provisioning, and resource control, promising huge amounts of cheap tailored bandwidth to specific escience research applications. TransLight enables Grid researchers to experiment with deterministic provisioning of dedicated circuits, then compare

results with standard and research Internet traffic. Methods to be tested and compared over the next several years include moving large amounts of data, supporting real-time collaboration and visualization, and enabling globally distributed computing at rates equaling the fast proliferation of e-science needs for at least the next 10 years.

#### REFERENCES

- 1. Antony, A., Lee, J., de Laat, C., Blom, J., and Sjouw, W. Microscopic examination TCP flows over trans-Atlantic links. In iGrid 2002 special issue. *Future Gen. Comput. Syst. 19*, 6 (Aug. 2003); 1017–1030 see www.science.uva.nl/~delaat/techrep-2003-2-tcp.pdf.
- DeFanti, T. and Brown, M., Eds. NSF CISE Grand Challenges in e-Science Workshop Report. National Science Foundation Directorate for Computer and Information Science and Engineering (CISE) Advanced Networking Infrastructure & Research Division (supported by NSF grant ANI-9980480) (University of Illinois at Chicago, Dec. 5–6, 2001); see www.evl.uic.edu/activity/NSF/index.html.
- 3. de Laat, C., Brown, M., and DeFanti, T., Guest Eds. Special issue on iGrid 2002. Future Gen. Comput. Syst. 19, 6 (Aug. 2003), 803–1062.
- de Laat, C., Wallace, S., and Radius, E. The rationale of the current optical networking initiatives. In iGrid 2002 special issue. Future Gen. Comput. Syst. 19, 6 (Aug. 2003); see www.science.uva.nl/-delaat/ techrep-2003-1-optical.pdf.
- He, E., Leigh, J., Yu, O., and DeFanti, T. Reliable blast UDP: Predictable high-performance bulk data transfer. In *Proceedings of IEEE Cluster Computing 2002* (Chicago, Sept. 24–26). IEEE Computer Society Press, 2002; see www.evl.uic.edu/cavern/papers/cluster2002.pdf.
- Li, G.-L. and Li, V. Networks of Queues: Myth and Reality. Tech. rep., University of Hong Kong, Apr. 2003; see www.eee.hku.hk/staff/down-load/webpaper.pdf.
- Smarr, L., Clapp, G., DeFanti, T., and Brown, M., Eds. NSF ANIR Workshop on Experimental Infostructure Networks. Report to the National Science Foundation Directorate for Computer and Information Science and Engineering, Advanced Networking Infrastructure & Research Division (supported by NSF grant ANI-0227640) (University of California, Irvine, Nov. 5, 2002); see www.calit2.net/events/2002/ nsf/index.html.
- 8. van Oudenaarde, B., Taal, A., de Laat, C., and Gommans, L. Authorization of a QoS path based on generic AAA. In iGrid 2002 special issue. *Future Gen. Comput. Syst. 19*, 6 (Aug. 2003), 1009–1016; see www.science.uva.nl/-delaat/techrep-2003-3-aaa.pdf.

**TOM DEFANTI** (tom@uic.edu) is director of the Electronic Visualization Laboratory and a professor of computer science at the University of Illinois at Chicago.

CEES DE LAAT (delaat@science.uva.nl) is an associate professor in the Informatics Institute at the University of Amsterdam, The Netherlands. JOE MAMBRETTI (j-mambretti@northwestern.edu) is director of the International Center for Advanced Internet Research at Northwestern University, Evanston, IL.

KEES NEGGERS (kees.neggers@surfnet.nl) is managing director of SURFnet, the national computer network for higher education and research, Utrecht, The Netherlands, and director of the GigaPort national project.

BILL ST. ARNAUD (bill.st.arnaud@canarie.ca) is senior director of advanced networks at CANARIE, Inc., Canada's advanced Internet development organization, Ottawa, Canada.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

© 2003 ACM 0002-0782/03/1100 \$5.00