The OptIPuter: A National and Global-Scale Cyberinfrastructure for Enabling LambdaGrid Computing

Maxine Brown[1], Larry Smarr[2], Tom DeFanti[1,2], Jason Leigh[1],
Mark Ellisman[3], Phil Papadopoulos[4]

[1] Electronic Visualization Laboratory, University of Illinois at Chicago
[2] Calit2, University of California, San Diego (UCSD)
[3] National Center for Microscopy and Imaging Research, UCSD
[4] San Diego Supercomputer Center, UCSD

April 19, 2006

ABSTRACT

To facilitate the interactive visualization, analysis, and correlation of massive amounts of data from multiple sites, the NSF-funded OptIPuter project is designing a powerful distributed cyberinfrastructure to support data-intensive scientific research and collaboration. This research exploits a new world in which the central architectural element is optical networking, not computers. This transition is caused by the use of parallelism, as in supercomputing a decade ago. However, this time the parallelism is in multiple wavelengths of light, or lambdas, on single optical fibers, creating *supernetworks*. Dedicated 1- to 10-Gigabit deterministic network connections are being deployed internationally by the Global Lambda Integrated Facility (GLIF), nationally by the National LambdaRail (NLR), regionally by academic consortia, and locally on campuses, connecting scientists' laboratories to collaborators and/or data sources all over the world, providing researchers with guaranteed bandwidth for data movement, guaranteed latency for visualization/collaboration and data analysis, and guaranteed scheduling for remote instrument control. Bandwidth alone isn't the solution; the OptIPuter is working on new grid-computing paradigms − that is, new middleware, transport protocols and optical signaling, control and management software − to enable applications to dynamically manage lambda resources just as they do any grid resource, creating a *Lambda-Grid* of interconnected high-performance computers, data storage devices, and instrumentation. This paper summarizes some of the OptIPuter's developments over dedicated end-to-end lightpaths among partner sites in San Diego, Chicago and Amsterdam.

ENABLING E-SCIENCE

Doctors want to better study the flow dynamics of the human body's circulatory system. Ecologists want to better study entire ecosystems in estuaries, lakes and along coastlines. Biologists want to perform multi-scale, correlated microscopy experiments, zooming from an entire system, such as a rat cerebellum, to an individual spiny dendrite. And, crisis management strategists want an integrated joint decision support system across local, state, and federal agencies, combining massive amounts of high-resolution imagery, highly visual collaboration facilities, and real-time input from field sensors. [Leigh06]

In essence, computational scientists want to study and better understand complex systems − physical, geological, biological, environmental, and atmospheric − from the micro to the macro scale, in both time and space. They want new levels of persistent collaboration over continental and transoceanic distances, coupled with the ability to process, disseminate, and share information on unprecedented scales, immediately benefiting the scientific community and ultimately, everyone else, as well. These application drivers are motivating the development of large-scale collaboration and visualization environments, built on top of an emerging global LambdaGrid cyberinfrastructure that is based on optical networks. [Brown03]

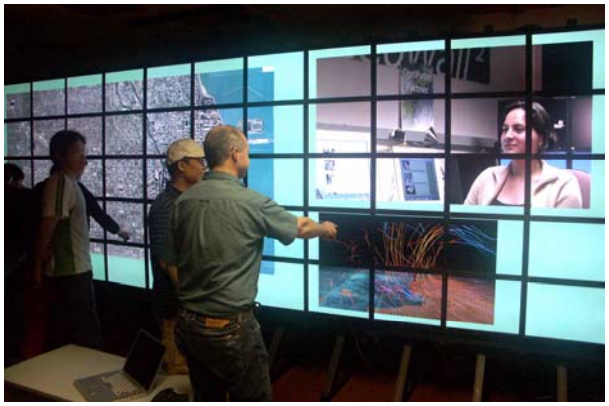

Figure 1: UIC's 100-Megapixel tiled display is managed by a software system called SAGE, which organizes the large-screen's "real estate" as if it were one continuous canvas [Jeong05], enabling researchers to view large-scale images while conducing high-definition video-teleconferences with remote colleagues.

Researchers want laboratories in which the walls are seamless ultra-high-resolution tiled displays fed by data streamed over ultra-high-speed networks, from distantly

located visualization and storage servers, enabling local and distributed groups of researchers to work with one another while viewing and analyzing visualizations of large distributed heterogeneous datasets (Figure 1).

Here we present the OptIPuter, a cyberinfrastructure research project that couples computational resources over parallel optical networks in support of data-intensive scientific research and collaboration. The OptIPuter has bioscience and Earth science application drivers, provided by the University of California, San Diego's (UCSD) National Center for Microscopy and Imaging Research (NCMIR) and NIH-funded Bio Informatics Research Network (BIRN) projects, and the UCSD Scripps Institution of Oceanography (SIO) and NSF-funded Earthscope projects. Lead institutions are UCSD and University of Illinois at Chicago (UIC); a list of partners and industry sponsors can be found at <www.optiputer.net>.

THE NETWORK AS BACKPLANE

The OptIPuter, so named for its use of optical networking, Internet Protocol (IP), computer storage, and processing and visualization technologies, is an infrastructure research effort that tightly couples computational resources over parallel optical networks using the IP communication mechanism. It is being designed as a *virtual* parallel computer in which the individual *processors* are distributed clusters; the *memory* is large distributed data repositories; *peripherals* are very-large scientific instruments, visualization displays and/or sensor arrays; and the *motherboard* uses standard IP delivered over multiple dedicated lambdas that serve as the *system bus* or *backplane.*

A *Grid* is a set of networked, middleware-enabled computing resources; a *LambdaGrid* is a Grid in which the lambda networks themselves are resources that can be scheduled like any other computing, storage and visualization resource. Recent major technological and cost breakthroughs in networking technology make it possible to send multiple lambdas down a single length of user-owned optical fiber. (A *lambda,* in networking parlance, is a fully dedicated wavelength of light in an optical network, capable of bandwidth speeds of 1-10 Gbps.) Metro and long-haul 10Gbps lambdas are 100 times faster than 100T-base Fast Ethernet local area networks used by PCs in research laboratories. The exponential growth rate in bandwidth capacity over the past 12 years has surpassed even Moore's Law due, in part, to the use of parallelism in network architectures. Now the parallelism is in multiple lambdas on single-strand optical fibers, creating *supernetworks,* or networks faster (and someday cheaper) than the computers attached to them [Brown03].

In the OptIPuter, all cluster nodes are on the network (unlike the log-in or head nodes of typical clusters), allowing experiments with multiple parallel communication paths. As OptIPuter networks scale up to multiple-10Gbps lambdas, the endpoints, too, must scale to *bandwidth-match* the network. The OptIPuter's dedicated network infrastructure has a number of significant advantages over shared Internet connections, including high bandwidth, controlled performance (no jitter), lower cost per unit bandwidth, and security.

OptIPuter endpoints at various partner sites are interconnected over the National LambdaRail (NLR) (Figure 2) and Global Lambda Integrated Facility (GLIF)-supported links (Figure 3). On a related note, UIC also has an NSF International Research Network Connections (IRNC) award for TransLight/StarLight, which provides two connections between the US and Europe for production science: a routed connection that connects the pan-European GÉANT2 to the USA Abilene and ESnet networks, and a switched connection between StarLight (in Chicago) and NetherLight (in Amsterdam) that is part of GLIF's LambdaGrid fabric.

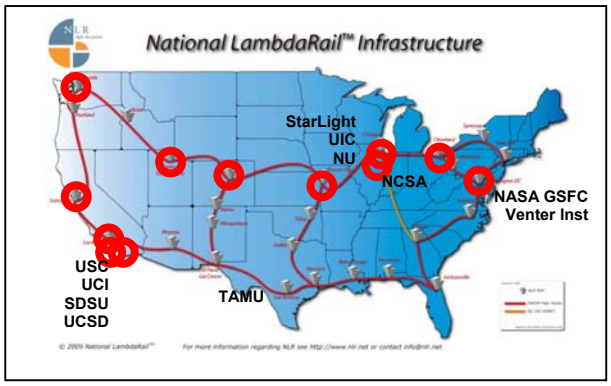

Figure 2: UIC has a persistent 10Gbps connection to University of Washington in Seattle and UCSD via a private wavelength on the NLR infrastructure called CAVEwave. This link is in the process of being extended to Washington DC to interconnect with NASA Goddard Space Flight Center (GSFC) and the Venter Institute.

Figure 3: GLIF is an international virtual organization that supports persistent data-intensive scientific research and middleware development on LambdaGrids. National Research & Education Networks, countries, consortia, institutions and individual research initiatives are providing the physical layer.

A supernetwork backbone is not the only requirement. One needs new software and middleware to provide data-delivery capabilities for the future lambda-rich world. The OptIPuter project is providing middleware and system software to harness the raw capabilities of the network hardware in a form that is readily available to and usable by applications [Smarr03].

MIDDLEWARE

The OptIPuter middleware uses the concept of a Distributed Virtual Computer (DVC) to integrate a wide range of unique OptIPuter component technologies (high-speed transport protocols, dynamic optical-network configurations, real-time, and visualization packages) with externally developed technologies (Globus grid resource management services and security infrastructure) that are increasingly being adopted in the grid community. A key benefit to applications includes control of a distributed resource abstraction, which includes network configuration, grid resource selection, and a simple uniform set of Application Programming Interfaces (APIs) for communication. DVC enables large-scale, flexible experimentation with a wide range of application configurations, enabling better evaluation and more rapid research progress.

VISUALIZATION AND COLLABORATION

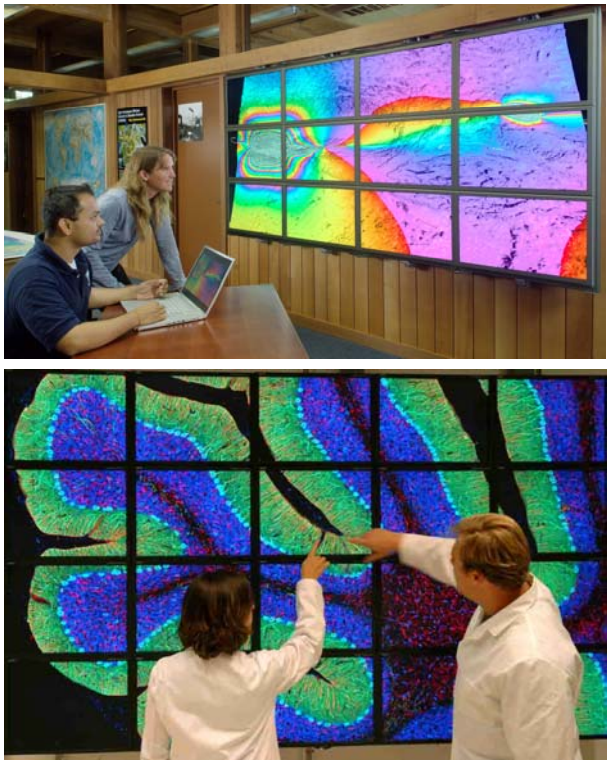

Figure 4: (a) SIO researchers study a 3D visualization of displacement after the 1906 San Francisco earthquake. (b) NCMIR researchers study a 2D cross-section image of a rat cerebellum.

The Scalable Adaptive Graphics Environment (SAGE) combines hardware-based, real-time, multi-resolution, view-dependent rendering techniques with software-based high-resolution rendering in a seamless manner that scales with the abilities of the computer hardware, the resolution of the display screens, and the size of the data. Particular attention was given to supporting real-time visualization of time-varying datasets that are potentially distributed at remote sites.

The OptIPuter project's two main visualization applications, JuxtaView and Vol-a-Tile, run on clusters of computers and stream the results to tiled displays (see Figure 4) [Krishnaprasad04, Schwarz04, Jeong05]. JuxtaView is a tool to view and interact with time-varying large-scale 2D montages, such as images from confocal or electron microscopes or satellite and aerial photographs. Vol-a-Tile is a tool to view large-scale time-varying 3D data, such as seismic volumes. These applications are unique in that they attempt to anticipate how the user will interact with the data, and use the available network capacity to aggressively *pre-fetch* the needed data, thus reducing the overall latency when data is retrieved from distantly located storage systems.

OPTIPUTER AND TERAGRID

UIC's LambdaStream is an application-level data transport protocol being developed as part of the OptIPuter project to support high-bandwidth multi-Gbps streaming for network-intensive applications, especially those requiring reliability at high bandwidth and with low jitter. An *application-level protocol* enables users to select and modify available protocols. Systems are capable of sending data with TCP or UDP protocols, but typically special sys-admin privileges or machine modifications are necessary to change and/or optimize them. LambdaStream enables the user, or application, to achieve high throughput over high-speed networks by building on top of existing TCP and UDP protocols without any machine modification or special sys-admin privileges.

In January 2006, LambdaStream was configured as a reliable UDP-based protocol and used to stream image data over both the TeraGrid and CAVEwave networks, where the former is a routed infrastructure and the latter is a switched infrastructure. Specifically, the tests compared throughput over a TeraGrid 10Gbps SONET-routed network using MultiProtocol Label Switching (MPLS) between Chicago and San Diego, and CAVEwave, a 10Gbps switched LAN PHY network between the same locations. [Vishwanath06]

An application file-transfer program invoked LambdaStream to send multiple 1Gbps data streams over each of these networks to saturate the 10Gbps links for 30-60 minute intervals. The dataset

used was a 1-foot-resolution map of 5,000 square miles of the city of Chicago provided by the US Geological Survey's National Center for Earth Resources Observation and Science (EROS). The map consists of 3,000 files of tiled images that are 75MBytes each, for a total of 220GBytes of information.

EVL performed three different LambdaStream tests over the TeraGrid and CAVEwave networks: moving data from memory to memory, from disk to memory, and from disk to disk; the three ways scientists typically use networks to access data. Using *iperf,* the maximum throughput achieved over the switched 10-Gigabit Ethernet (GE) (10Gbps) CAVEwave network was 9.75Gbps, and the maximum throughput over the routed OC-192 (9.6Gbps) TeraGrid was 9.45Gbps. Using LambdaStream, CAVEwave achieved speeds of 9.23Gbps reliable memory-to-memory, 9.21Gbps reliable disk-to-memory, and 9.30Gbps reliable disk-to-disk transfers; and, TeraGrid achieved speeds of 9.16Gbps reliable memory-to-memory, 9.15Gbps reliable disk-to-memory, and 9.22Gbps reliable disk-to-disk transfers. Bidirectionally, CAVEwave achieved 18.19Gbps and TeraGrid achieved 18.06Gbps doing reliable memory-to- memory transfers.

The measured performance across CAVEwave and TeraGrid was comparable, proving that a routed and non-routed infrastructure could both sustain near 10Gbps and behave the same when a specialized protocol is used.

CONCLUSIONS

Lambdas are a simplistic (some network architects would say overkill) means of achieving guaranteed quality of service, that is, end-to-end deterministic, scheduled connectivity. However, dedicated lambdas allow OptIPuter researchers to experimentally allocate entire end-to-end lightpaths and devote OptIPuter middleware research to enabling applications, rather than perfecting congestion control. In the same way, 20 years ago, software shifted from optimizing mainframe timesharing to human factors on workstations and PCs. Thus, the OptIPuter project is not optimizing toward scaling to millions of sites, a requirement for commercial profit, but empowering networking at a much higher level of data volume, accuracy, and timeliness for a few high-priority research and education sites.

ACKNOWLEDGMENTS

Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the funding agencies and companies. Major funding for the OptIPuter is provided by a National Science Foundation (NSF) cooperative agreement (OCI-0225642) to UCSD. Funding for TransLight/ StarLight is provided by NSF award (OCI-0441094) to the UIC's Electronic Visualization Laboratory (EVL).

The CAVEwave is so named because funds derived from the licensing of the EVL's CAVE® virtual reality room were used to procure the 10GE wavelength from NLR. CAVE and CAVEwave are trademarks of the Board of Trustees of the University of Illinois.

REFERENCES

[Brown03] M.D. Brown (guest editor), "Blueprint for the Future of High-Performance Networking (Intro)," Communications of the ACM, Volume 46, Number 11, November 2003, pp. 30-33.

[Jeong05] B. Jeong, R. Jagodic, L. Renambot, R. Singh, A. Johnson, J. Leigh, "Scalable Graphics Architecture for High-Resolution Displays", Proceedings of the Using Large, High-Resolution Displays for Information Visualization Workshop, IEEE Visualization 2005, Minneapolis, MN, October 2005.

[Krishnaprasad04] N. Krishnaprasad, V. Vishwanath, S. Venkataraman, A. Rao, L. Renambot, J. Leigh, A. Johnson, B. Davis, "JuxtaView – a Tool for Interactive Visualization of Large Imagery on Scalable Tiled Displays," Proceedings of IEEE Cluster 2004, San Diego, September 20-23, 2004.

[Leigh06] Jason Leigh, Luc Renambot and Maxine Brown, "Grid Network Requirements for Large-Scale Visualization and Collaboration," Grid Network Requirements Defined By Driver Applications (chapter 2), Grid Networks: Enabling Grids with Advanced Communication Technology, John Wiley & Sons, Ltd., 2006, to appear

[Schwarz04] N. Schwarz, S. Venkataraman, L. Renambot, N. Krishnaprasad, V. Vishwanath, J. Leigh, A. Johnson, G. Kent, A. Nayak, "Vol-a-Tile − A Tool for Interactive Exploration of Large Vol. Data on Scalable Tiled Displays" (poster), IEEE Visualization 2004, Austin, TX, October 2004.

[Smarr03] L. Smarr, A. Chien, T. DeFanti, J. Leigh, P.M. Papadopoulos, "The OptIPuter," Communications of the ACM, Vol. 46, Issue 11 (November 2003), pp. 58 – 67.

[Vishwanath06] Venkat Vishwanath, "Wide Area Network Experiments with LambdaStream over Dedicate High Bandwidth Networks," IEEE Infocom 2006 Conference, Barcelona, April 24, 2006, to be presented