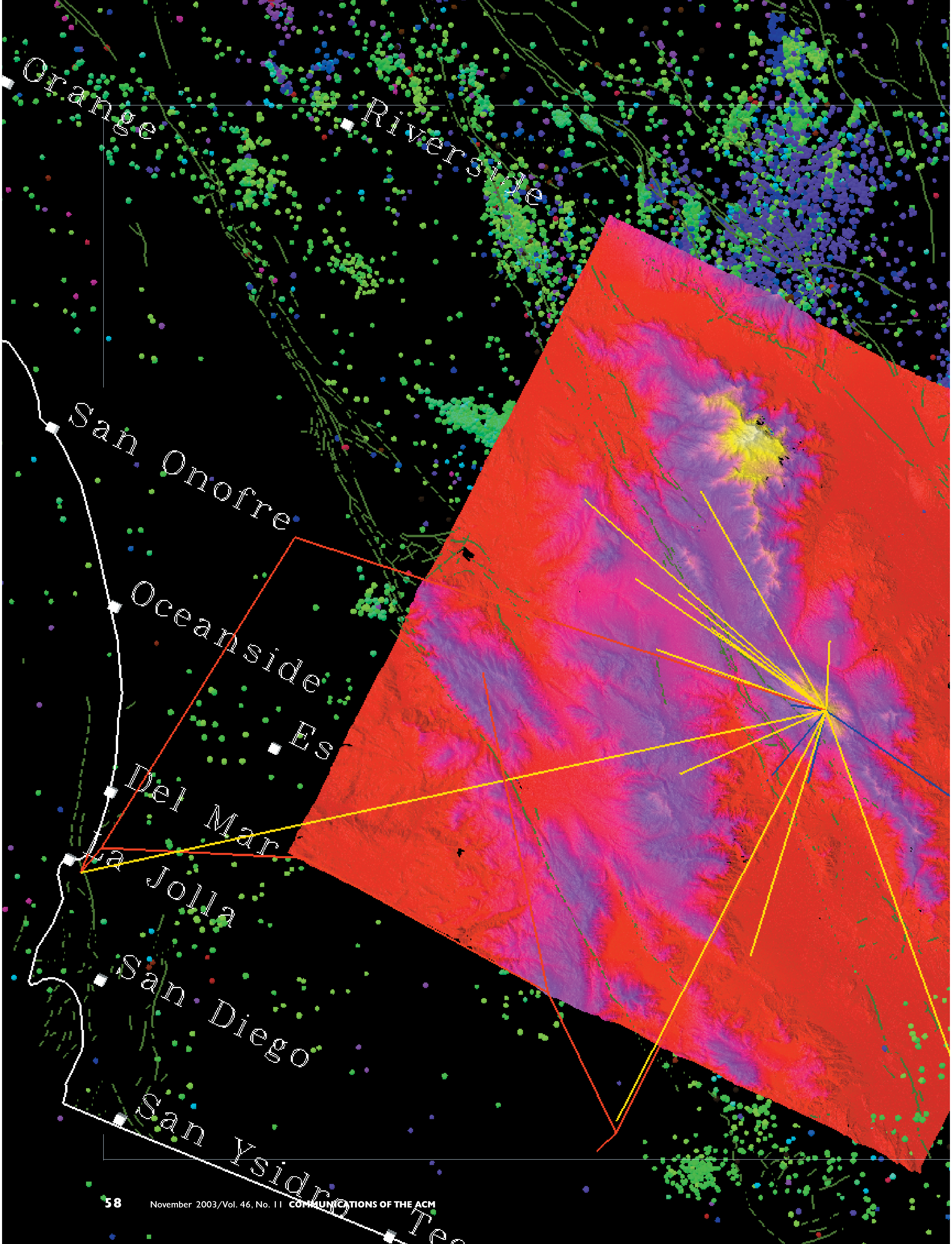
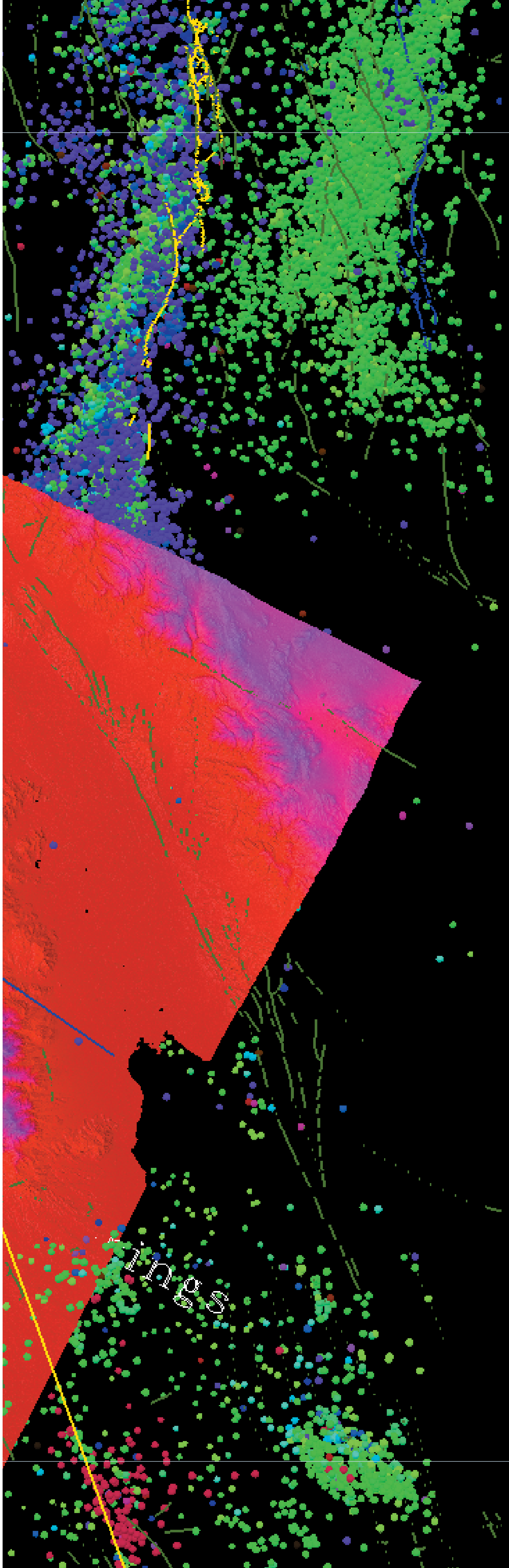

Orange
Riverside
San Onofre
Oceanside
Es
Del Mar
La Jolla
San Diego
San Ysidro
Tec

By Larry L. Smarr, Andrew A. Chien, Tom DeFanti, Jason Leigh, and Philip M. Papadopoulos

The OptIPuter

This architecture/infrastructure of parallel optical networks couples data exploration, visualization, and collaboration technologies through IP at multi-gigabit speeds.

The OptIPuter exploits a new world of distributed Grid infrastructure in which the central architectural element is optical networking, not computers, creating "supernetworks," or networks faster than the computers attached to them. As in supercomputing a decade ago, parallelism makes this transition possible. But this time, parallelism takes the form of multiple wavelengths of light, or lambdas, capable of traversing individual strands of optical fiber.



Seismic data near Toro Peak in Southern California. Earthquake locations (colored spheres), seismic station telemetry paths (yellow/red/blue lines), and 30-meter Shuttle Radar Topography Mission topography. (Debi Kilb, Frank Vernon, and David T. Sandwell, Institute of Geophysics and Planetary Physics, Scripps Institution of Oceanography, University of California, San Diego)

So named for its use of optical networking, Internet Protocol, and computer storage, processing and visualization technologies, the OptIPuter is an envisioned infrastructure, available within five years, tightly coupling computational resources over parallel optical networks using the IP communication mechanism. Think of it as a "virtual" metacomputer in which the individual "processors" are widely distributed clusters; the backbone network is provided by IP delivered over multiple dedicated lambdas (each 1–10Gbps); and the "mass storage systems" are large distributed scientific data repositories, fed by scientific instruments as near-real-time peripheral devices.

The OptIPuter is a revolutionary architecture addressing many of the needs of e-scientists and distributed cyberinfrastructure (see the article by Newman et al. in this section). The implosion of the telecom industry over the past few years has made it possible for a consortium of universities to acquire long-term leases on previously unavailable dark fiber and wavelengths at metro, regional, national, and even international scales (see the article by DeFanti et al. in this section), thus enabling research projects like the OptIPuter. Individually controlled end-to-end lambdas, connecting clusters at a few key research laboratories to one another, as well as to remote instruments and data storage, can provide deterministic and guaranteed bandwidth to high-end users at multigigabit speeds.

The OptIPuter team has two e-science application drivers—the National Institutes of Health's Biomedical Informatics Research Network (BIRN, see www.nbirn.net) and the National Science Foundation's EarthScope (see www.earthscope.org)—each involving many multi-gigabyte-size individual data objects located in remote federated repositories. The related computer science effort is led by the University of California, San Diego, and the University of Illinois at Chicago. UCSD, through the California Institute for Telecommunications and Information Technology, heads up the Southern California partners: San Diego State University, the Information Sciences Institute at the University of Southern California, and the University of California, Irvine; and UIC heads up Chicago partner Northwestern University and affiliate Midwest and international partners U.S. Geological Survey Earth Resources Observation Systems Data Center and the University of the Amsterdam, The Netherlands. Industrial partners include IBM, Sun Microsystems, Telcordia Technologies, and Chiaro Networks.

The initial application now being built involves data objects, including gigazone seismic images of the East Pacific Rise Magma chamber [4] and 100-megapixel montages of rat cerebellum microscopy images. These very large volumetric data objects and visualizations are so big they exceed the capacity of the current shared Internet and laptop displays.

The OptIPuter opportunity is made possible by the more than 60× improvement in network bandwidth—from 155Mbps to 10Gbps—over the past eight years, a growth rate outpacing that of disk speed/capacity and processing power during the same period. This has produced a major technology mismatch, perhaps best illustrated by the fact that many PCs today are sold with Gigabit Ethernet (GigE) interfaces, even though typically available file transport speeds across the shared Internet are only 10–20Mb. Even with Dense Wave Division Multiplexing (DWDM) technology enabling immense amounts of bandwidth on a single pair of fiber strands, today's networks still clog up and slow to a crawl. For example, routing can cause some network paths to be overcrowded while others are empty, since routers usually optimize for shortest paths. Alternatively, lambdas dedicated to individual researchers create the equivalent of high-occupancy-vehicle expressway lanes, delivering more reliable and predictable network performance. The OptIPuter project posits a network in which optical circuits can be set up when and wherever they're needed, consuming huge amounts of bandwidth, as prophesied years ago by technology visionary George Gilder [2].

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.

OptIPuter Networking Strategies

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. Two rather different OptIPuter networking strategies are being implemented to schedule and configure custom applica-

THOUSANDS OF 1GIGE INTERFACES AND HUNDREDS OF 10GIGE INTERFACES CAN BE ROUTED AT FULL SPEED.

tion-based experiments that potentially (and routinely) overwhelm traditional networks and software stacks.

The Southern California team has built a router-centric OptIPuter (at UCSD) consisting of five endpoint clusters in various campus buildings (engineering, oceanography, medicine, computer science, and the San Diego Supercomputer Center), each linked by dedicated optical fibers to a central high-speed Chiaro Networks Enstara router (see www.chiaro.com). This architecture is an evolutionary approach; bandwidth-hungry biomedical (BIRN) and geoscience (EarthScope) applications use this multi-gigabit network, relying on the well-understood software stack of a normal routed network.

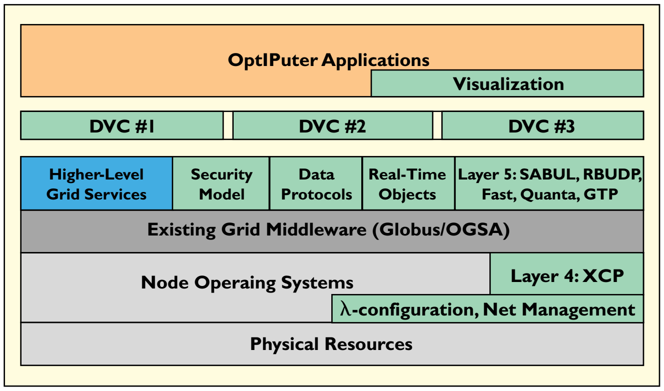
The Enstara router, which uses novel internal optical phased-array switching, provides essentially unlimited packet-forwarding capability with more than 6Tb of capacity; thousands of 1GigE interfaces and hundreds of 10GigE interfaces can be routed at full speed. Network capacity will grow over the next four years to 40Gbps per site and eventually to hundreds of Gbps per site, enabled by the availability of cheaper DWDM hardware and 10GigE interfaces. The description-based methods of the NPACI Rocks clustering toolkit [3, 8] are being extended so the hardware at the endpoints of the fibers is easily managed; included are switches, routers, storage, and other configurable elements. The BIRN project will measure the effectiveness of different transmission protocols, including the Explicit Control Protocol, the Reliable Blast User Data Protocol, and the Simple Available Bandwidth Utilization Library protocol, even when they require fundamentally different system configurations (see the article by Falk et al. in this section).

The Chicago team is building a fully optically switched OptIPuter under the control of novel lambda-management software allowing for significantly more flexibility and granularity in provisioning optical network resources. Related switch research covers new types of signaling, control-plane techniques based on these signaling methods, and mechanisms for dynamically provisioning lightpaths within individual domains and across multiple domains [1]. This optically switched OptIPuter is being built with Calient DiamondWave and GlimmerGlass Reflexion 3D micro-electro-mechanical system (MEMS) switches (see www.calient.net and www.glimmerglassnet.com), as well as conventional routers, for out-of-band control. (Each cluster element has two GigE interfaces, one to a standard router, the other to the optical switch.) The cost of MEMS switches is modest, less than $1,000 per port. A big part of the OptIPuter project is testing the situations in which optical switches might be appropriate for supporting data-intensive science and engineering research. For example, UIC is connected to the University of Amsterdam via 15Gbps lambda-dedicated circuits. Another Calient switch and computer cluster is housed at the NetherLight optical network facility in Amsterdam. Similar experiments are certainly possible with research networks based in Canada, the European Laboratory for Particle Physics (CERN), and the U.K., as well as with TeraGrid sites in the U.S.

As the OptIPuter networks are scaled up to multiple-10Gbps lambdas, the endpoints, too, must scale to "bandwidth-match" the network. A variety of OptIPuter cluster nodes will emerge, including visualization nodes and parallel data-capable systems to form a distributed set of caches, as well as traditional high-performance computing clusters. OptIPuter endpoints will be built at SDSU, UIC, USC/ISI, and NU, and interconnected to UCSD and UIC, applying the results of the various switched and routed strategies.

Critical to many of the OptIPuter applications is the availability of a high-speed, parallel-access storage abstraction, or model. The current commercial practice is to export only low-level disk block interfaces over storage area network/wide area network (SAN/WAN) proprietary protocol bridges. But this approach does not scale to e-science needs; OptIPuter researchers are thus investigating high-speed cluster storage services where each node contributes to an aggregated globally accessible storage pool to exploit the fundamental parallelism in the network, client, and server over both campus and wide-area networks.

Figure 1. OptIPuter software architecture. Green boxes indicate OptIPuter research areas; gray boxes represent existing technologies; and the blue box reflects the ongoing evolution in Grid middleware from lower-level to higher-level Grid services.



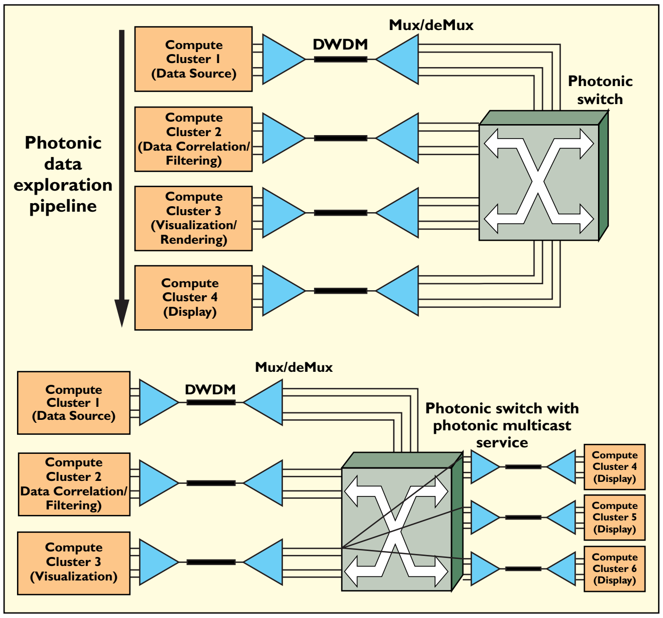

Compute Cluster 1 (Data Source)
Compute Cluster 2 (Data Correlation/ Filtering)
Compute Cluster 3 (Visualization/ Rendering)
Compute Cluster 4 (Display)
DWDM
Mux/deMux
Photonic switch
Photonic data exploration pipeline
Compute Cluster 1 (Data Source)
Compute Cluster 2 Data Correlation/ Filtering)
Compute Cluster 3 (Visualization)
DWDM
Mux/deMux
Photonic switch with photonic multicast service
Compute Cluster 4 (Display)
Compute Cluster 5 (Display)
Compute Cluster 6 (Display)

Figure 2. Photonic switch architecture in the OptIPuter. (top) A number of distributed compute clusters are interconnected via a photonic switch rather than via traditional routers, creating a photonic data exploration pipeline; the optical transport platform multiplexes and demultiplexes multiple lightpaths over a DWDM network. (bottom) Photonic multicasting is integrated into the photonic data exploration pipeline.

OptIPuter Middleware

While network hardware, including DWDM, optical fiber, and high-speed routers and switches provide raw capabilities, the middleware and system software must harness it in a form readily available to and usable by applications. A LambdaGrid is a set of networked, middleware-enabled computing resources, in which the lambda networks themselves are resources that can be scheduled like computing resources. The OptIPuter LambdaGrid software architecture allows applications to dynamically provision dedicated lambdas in seconds, where these optical paths span the domains of metro, national, and international network service providers. However, such dynamic connections raise questions about routing, security, cross-vendor signaling, management, and presentation. But they potentially offer dramatically new levels of performance, quality of service, and distributed resource abstractions that simplify construction of distributed applications.

A key element of our approach is to develop a resource abstraction—a distributed virtual computer, or DVC—usable by an application as a simple model for complex distributed environments. DVCs are collections of resources with fixed capabilities, trust relationships, network provisioning, and other forms of performance quality of service. These virtual metacomputers are likely to span multiple administrative domains and be coupled by high-speed dedicated optical connections. A DVC typically includes: a virtual computer (processor, storage device, display device); a virtual cluster (homogeneous collection of processors, storage, and display devices); a heterogeneous collection of components (processors, storage devices, displays, and other peripherals), or any collection of distributed network devices.

DVCs encapsulate the difficult problem of configuration management, easing the secure configuration of resources from multiple administrative domains to perform useful scientific or commercial applications. A DVC can be configured, named, and instantiated on a variety of time scales. Consider three cases:

Dynamic. Configured on-demand within the DVC, applications flexibly share resources directly. A

DISTRIBUTED VIRTUAL COMPUTERS ENABLE APPLICATIONS TO VIEW THEIR WORLD AS A SAFE LOCAL CLUSTER ENVIRONMENT RATHER THAN AS A HOSTILE, BEST-EFFORT, OPEN INTERNET.

dynamic DVC serves as a single administrative domain with centralized resource control.

Pseudo-static. Configured through negotiation with multiple site resource administrators, then instantiated on demand. Once configured and instantiated, its use is similar to that of a dynamic DVC.

Federated Grid resources. Configured using a virtual organization and full Grid Services Infrastructure security infrastructure. Resource access is limited to three types of interface: Open Grid Services Infrastructure (see www.gridforum.org/ogsi-wg/) level interfaces and Open Grid Services Architecture (see www.globus.org/ogsa/) interfaces.

Viewing collections of distributed resources in DVCs enables construction of simplified distributed resource abstractions for applications. In the network, they guarantee performance and ensure physical security and high-speed, in-order, low-loss communication attributes. Displays and storage devices without significant Grid middleware capability can be tightly coupled into a DVC and used like a physically secure, localized LAN/SAN environment. Finally, within a DVC, peer device performance, security, and resource management can be simplified. DVCs enable applications to view their environment as a safe local cluster rather than as the relatively hostile, best-effort, open Internet.

DVCs thus enable the use of a simpler network model—a layered single administrative domain with a collection of resources at the e-science user's disposal—but with predictable performance. Research efforts must still, however, address the following areas:

- DVC models and abstractions that simplify distributed application programming;
- High-speed protocols and communication layers that help deliver network performance;
- Dynamic network optical-path configuration, management, routing, and de-configuration;
- Multipoint (parallel and multicast) communication abstractions and APIs that enable scalable compute elements to be matched to 10–100Gbps wide-area connections;
- Storage and file systems to support high-speed access to remote data;
- Security protocols and models for high-bandwidth, long-latency environments with millions of resources;
- Adaptive and peer-to-peer data access, streaming, and integration protocols; and
- Real-time object models enabling predictable performance across the entire system.

Development of OptIPuter system software is properly viewed against a background of emerging rich and complex Grid middleware (see Foster's and Grossman's article in this section), providing high-performance distributed applications with basic elements of security, resource access, and coordination, all built on an IP network fabric. The OptIPuter software builds on this Grid middleware infrastructure, providing unique capabilities derived from dedicated optical paths (see Figure 1).

OptIPuter Data Exploration, Visualization, Collaboration

Data exploration, visualization, and collaboration are primary application enablers on LambdaGrids for the OptIPuter project. Dedicated optical networking affects classical assumptions about the ways applications operate. Motivation for using MEMS-

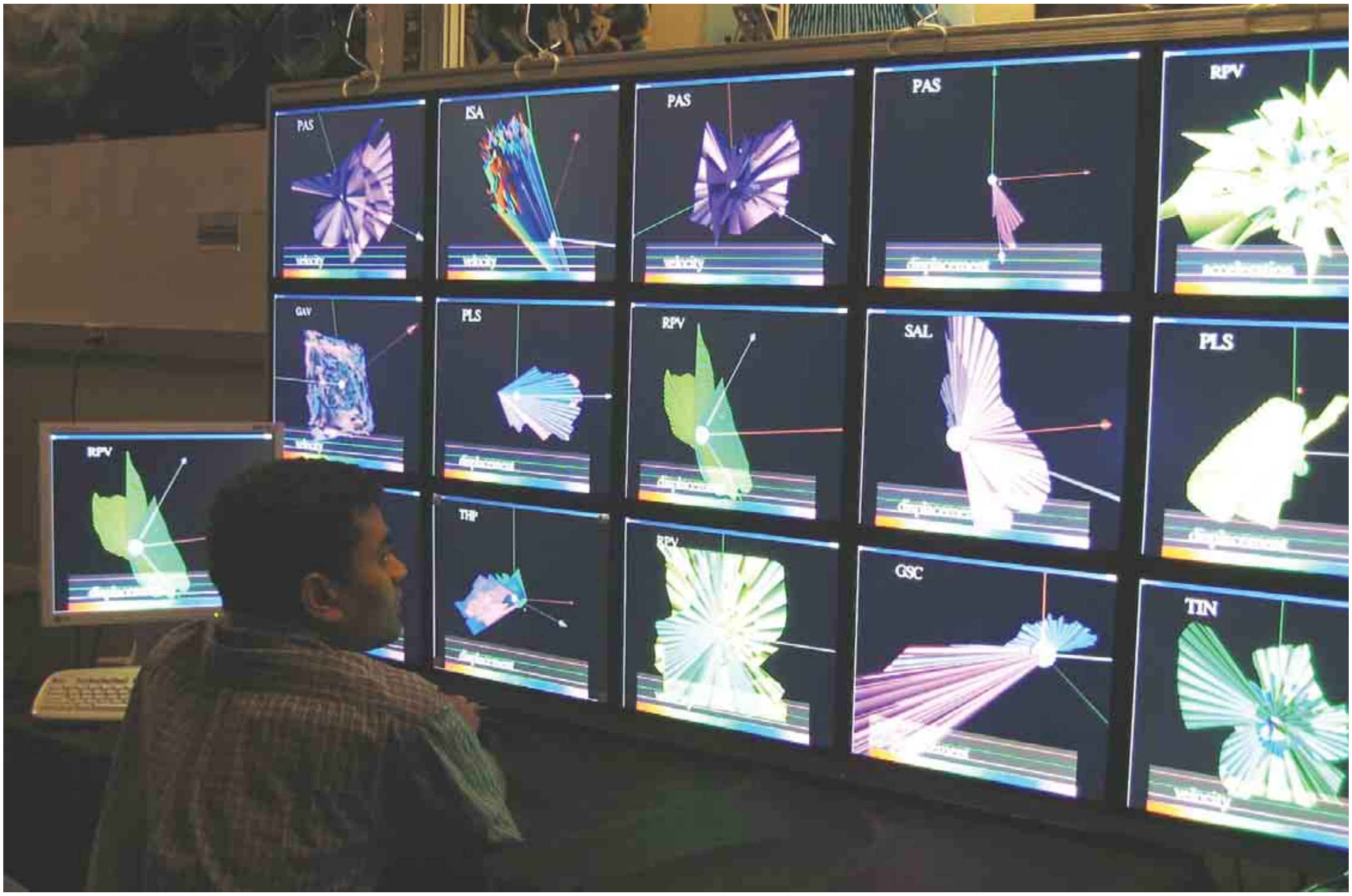


Figure 3. High-resolution digital montages on tiled displays. (left) A series of seismoglyphs summarizes real-time seismic data from remote seismometers distributed around the world; data is retrieved in real time from databases managed by the Incorporated Research Institutions for Seismology consortium (see www.iris.edu) [7]. (right) JuxtaVision shows planetary bathymetry data.

switched networking to support visualization and data exploration is the observation that the predominant model for large-scale data-intensive applications employs the following data-exploration pipeline:

Data sources —> Data correlation/filtering system —> Visualization system —> Display system

Large collections of real or simulation data are fed into a data-correlation or filtering system that generates subsampled or summarized results from which a visual representation can be created and displayed on individual screens, tiled displays, and virtual-reality systems. In the context of collaborative data exploration, the results of the pipeline may need to be multicast to several end points. Unlike Web browsing (which tends to involve users jumping from Web site to Web site), the connectivity between the computing components in a large-scale data-exploration pipeline is typically static, once connections are established. Hence, costly packet-by-packet, multi-gigabit routing is unnecessary.

Multiple photonic switches can be combined to create very high-performance mesh networks supporting fully dedicated high-speed data transmissions between network elements (see Figure 2). Combined with proper scheduling of cluster computing resources and lightpaths, this architecture allows applications to create multiple, simultaneous, distributed computing pipelines.

For collaborative data exploration, the computational pipeline must support multicast-to-multiple endpoints. Prior research suggests how network engineers might build such a system [6, 10]. A multicast-capable photonic switch is being developed at UIC in partnership with Glimmerglass Networks [5].

Motivated by the emergence of photonic networking techniques, the Chicago OptIPuter team is also developing the Continuum, a collaborative project room [9] uniting a range of display technologies, including passive stereo displays (such as GeoWalls, see www.geowall.org), high-resolution tiled displays, AccessGrids (see www.accessgrid.org), and large-for-

mat shared digital whiteboards (see www.evl.uic.edu/cavern/optiputer/). Designed to take advantage of the OptIPuter, the Continuum employs the following prototype visualization and collaboration tools:

TeraScope. Monitoring the performance of remote disk systems, networks, data mining clusters, and graphics-rendering clusters, this adaptive framework intelligently selects a remote visualization strategy appropriate for the size of the data being visualized and the end system used to visualize the data [12].

JuxtaVision. Though it employs the TeraScope concept, this tool is designed to display extremely high-resolution digital montages on tiled displays (see Figure 3). Incorporating a networked, read-only memory system called LambdaRAM [12] to provide predictive paging of image data for visualizations, it uses the OptIPuter to create large memory pools serving as data caches to help overcome access latency in long-distance networks.

TeraVision. Designed to format high-resolution graphics on a visualization cluster, this application toolkit for collaborative high-resolution graphics streaming uses photonic multicasting to distribute visualizations to multiple collaborating sites with tiled displays [11].

The overall Continuum goal is to develop new display technologies allowing collaborators to seamlessly manipulate content on the displays and conduct human-centered research to understand how display-rich environments influence the effectiveness of distance collaboration. The OptIPuter uses two very different tens-of-megapixels display technologies—the Geowall2, a tiled (3×5) PC LCD display and the dual-screen IBM 9-megapixel T221 Bertha display (see www.research.ibm.com/deepview)—to high-resolution data. In a year or so, the project will attempt to develop displays with resolutions in excess of 100 megapixels using new display media, including flexible organic LEDs and large arrays of borderless LCDs, whichever is developed first at reasonable cost.

Conclusion

The OptIPuter is a radically new distributed visualization, data mining, and computing architecture that goes to the end of the technological rainbow to exploit a new world of network infrastructure in which the central architectural element is networking, not computers. The OptIPuter project aims to learn how to, as George Gilder [2] suggests, waste bandwidth and storage in order to conserve increas-

The project will attempt to develop displays with resolutions in excess of 100 megapixels using new display media, including flexible organic LEDs and arrays of borderless LCDs.

ingly scarce high-end computing and people time in an emerging world of inverted values.

References

1. DeFanti, T., Brown, M., Leigh, J., Yu, O., He, E., Mambretti, J., Lillethun, D., and Weinberger, J. Optical switching middleware for the OptIPuter. In a special issue on photonic IP network technologies for next-generation broadband access. *IEICE Transact. Commun.* E86-B, 8 (Aug. 2003), 2263–2272.
2. Gilder, G. *Telecosm: How Infinite Bandwidth Will Revolutionize Our World.* Free Press, New York, 2000.
3. Katz, M., Papadopoulos, P., and Bruno, G. Leveraging standard core technologies to programmatically build Linux cluster appliances. In *Proceedings of CLUSTER 2002: IEEE International Conference on Cluster Computing* (Chicago, Sept. 23–26). IEEE Computer Society Press, Los Alamitos, CA, 2002, 47–53.
4. Kent, G., Singh, S., Harding, A., Sinha, M., Orcutt, J., Barton, P., White, R., Bazin, S., Hobbs, R., Tong, C., and Pye, J. Evidence from three-dimensional seismic reflectivity images for enhanced melt supply beneath mid-ocean-ridge discontinuities. *Nature 406* (Aug. 10, 2000), 614–618.
5. Leigh, J., Renambot, L., DeFanti, T., Brown, M., He, E., Krishnaprasad, N., Meerasa, J., Nayak, A., Park, K., Singh, R., Venkataraman, S., Zhang, C., Livingston, D., and McLaughlin, M. An experimental OptIPuter architecture for data-intensive collaborative visualization. In *Proceedings of the Workshop on Advanced Collaboration Environments* (Seattle, WA, June 22–24, 2003).
6. Leuthold, J. and Joyner, C. Multimode interference couplers with tunable power splitting ratios. *IEEE/OSA J. Lightwave Tech. 19,* 5 (May 2001), 700–706.
7. Nayak, A., Leigh, J., Johnson, A., Russo, R., Morin, P., Laughbon, C., and Ahern, T. WiggleView: Visualizing large seismic data sets. *EOS Transact. American Geophys. Union 83,* 47 (fall 2002).
8. Papadopoulos, P., Papadopoulos, C., Katz, M., Link, W., and Bruno, G. Configuring large high-performance clusters at lightspeed: A case study. In *Proceedings of Clusters and Computational Grids for Scientific Computing 2002.*
9. Park, K., Renambot, L., Leigh, J., and Johnson, A. The impact of display-rich environments on enhancing task parallelism and group awareness in advanced collaborative environments. In *Proceedings of the Workshop on Advanced Collaboration Environments* (Seattle, WA, June 22–24, 2003).
10. Rouskas, G. Optical layer multicast: Rationale, building blocks, and challenges. *IEEE Network 17,* 1 (Jan./Feb. 2003), 60–65.
11. Singh, R., Leigh, J., DeFanti, T., and Karayannis, F. TeraVision: A high-resolution graphics streaming device for amplified collaboration environments. *J. Future Gen. Comput. Syst. 19,* 6 (Aug. 2003), 957–972.
12. Zhang, C., Leigh, J., and DeFanti, T. TeraScope: Distributed visual data mining of terascale data sets over photonic networks. *J. Future Gen. Comput. Syst. 19,* 6 (Aug. 2003), 935–944.

Larry L. Smarr (lsmarr@ucsd.edu) is director of the California Institute for Telecommunications and Information Technology at the University of California, San Diego, and the University of California, Irvine.

Andrew A. Chien (achien@ucsd.edu) is the Science Application International Corporation Chair and a professor in the Department of Computer Science and Engineering at the University of California, San Diego.

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.

Jason Leigh (spiff@evl.uic.edu) is an associate professor in the Electronic Visualization Laboratory at the University of Illinois at Chicago.

Philip M. Papadopoulos (phil@sdsc.edu) is program director of Grid and cluster computing in the San Diego Supercomputer Center at the University of California, San Diego.

Major funding for the OptIPuter is provided by a National Science Foundation Information Technology Research cooperative agreement (ANI-0225642) to the University of California, San Diego. The advanced networking facilities at the Electronic Visualization Laboratory and Northwestern University are made possible by major funding from the NSF, specifically awards EIA-9802090, EIA-0115809, ANI-9980480, ANI-0229642, ANI-9730202, ANI-0123399, ANI-0129527, EAR-0218918, and ACI-9619019.

Andromeda Galaxy. Most detailed image yet of the cool gas in any galaxy. (Made with the Westerbork Synthesis Radio Telescope; Robert Braun, ASTRON, The Netherlands; E. Corbelli, Arcetri Observatory, Italy; Rene Walterbos, New Mexico State University; and David Thilker, John Hopkins University)