The OptIPuter, Quartzite, and Starlight Projects: A Campus to Global-Scale Testbed for Optical Technologies Enabling LambdaGrid Computing

Larry Smarr, Joe Ford, Phil Papadopoulos, Shaya Fainman
University of California at San Diego, La Jolla, CA, USA

Thomas DeFanti, Maxine Brown, Jason Leigh
University of Illinois at Chicago, Chicago, IL, USA

Abstract: Dedicated optical connections have significant advantages over shared internet connections. The OptIPuter project (www.optiputer.net) uses medical and earth sciences imaging as application drivers. Quartzite (UCSD) and Starlight (Chicago) create unique combinations of OEO routers and OOO and wavelength-selective optical switches.

OCIS codes: (060.2330) Fiber optic communications, (170.0110) Imaging systems

Over campus and state fiber networks, the National Lambda Rail, and internationally, scientists are beginning to use private, 1 or 10 Gbit/s (Gbps) light pipes (termed "lambdas") to create deterministic network connections coming right into their laboratories. These dedicated connections have a number of significant advantages over shared internet connections, including high bandwidth, controlled performance (no jitter), lower cost per unit bandwidth, and security. By connecting scalable Linux clusters with these lambdas, one creates “metacomputers” on the scale of a nation or even the planet Earth. One of the largest research projects on LambdaGrids is the NSF- funded OptIPuter (www.optiputer.net), which uses large medical and earth sciences imaging as application drivers. The OptIPuter has two regional cores, one in Southern California and one in Chicago, which has now been extended to Amsterdam. One aim of the OptIPuter project is to make interactive visualization of remote gigabyte data objects as easy as the Web makes manipulating megabyte-size data objects today. This requires parallel scaling up PCs to 100-1000x in display, storage, and compute power, while maintaining personal interactivity.

The NSF has recently funded an extension of the OptIPuter project, called Quartzite, so that we can efficiently investigate and compare campus-scale terabit-class lambda network architectures that span from optical-circuits-only to packet-switched-only networks and a range of hybrid combinations in between. Quartzite connects over 300 individual cluster nodes on the UCSD campus with a novel switching core. The Quartzite core is comprised of a leading edge commercial packet switch tightly coupled to a commercial MEMS passive optical switch and then to an experimental wavelength-switchable device. Laboratory clusters supporting specialized instruments, computation, visualization, and storage serve as ideal parallel endpoints. Because Quartzite enables soft reconfiguration (from optical circuit to optical packet) of an endpoint, we will be able to better understand the where, how, and why of the packet vs. circuit architectural tradeoff, what protocols (both optical signaling and higher-level messaging) are effective, and how dynamic virtual collections at the campus scale can be knitted together with high-speed parallel networks to form an effective analysis platform for the next-generation of scientific research.

We see a future in which a scientist plugs their lab instrument's or cluster's fiber uplink into the campus core, just as they do today, but with an important difference: the backbone fiber carries multiple "stand-by" allocatable wavelengths (lambdas) in addition to the common shared and routed internet traffic. Some of these "spare" lambdas are initially disconnected from the internet fabric, but are available for on-demand capacity provisioning. This LambdaGrid architecture allows us to explore significant new dimensions in system architecture in support of data intensive science, such as earth science and biomedical research.

A central component of Quartzite will be a hybrid circuit/packet switch that in a few years will have enough network bisection to handle 100s of 10-gigabit networked endpoints. Physically, a star topology, the central Quartzite switch is a "stack" that is made up of a transparent optical MEMS-based switch (OOO Switch), a wavelength selective switch (WS-Switch) and a leading edge packet switch. The construction of Quartzite is unique in that the same physical fibers can be software reconfigured to be all packet-based, hybrid packet and circuit, or all circuit based. Purposefully, we are not building the highest-capacity network, but one in which a variety of different communication techniques can be compared and contrasted. We will be able to model a number of virtual topologies on top of our physical core.

The packet switch is a Chiaro Enstara router that internally uses an optical phase array (OPA) architecture to switch standard packets at line rate. The Enstara also supports virtual routing instances with access to the programming interfaces, the network processors that sit in each line card. Both capabilities give an experiment full control over its virtual router. The second major component in Quartzite switching core is a 64-port optically transparent switch using MEMS-based mirrors (an OOO provisioning switch—eg from Calient or Glimmerglass), which are completely protocol, frequency, and signal speed independent. Our initial intention is to use this switch as an optical "patch panel." Finally, the Wavelength-Selective switch (WS-Switch) is also MEMS-based but uses an internal optical grating demultiplexer and MEMS tilt-mirror arrays to further split the signal into wavelengths and then switch wavelengths (either individually, or more commonly in groups) to the desired destination. The 4x4 WS-Switch being developed by Lucent for this project has 4 physical fiber pair inputs, each carrying multiple wavelength channels. In this system data will be carried on 32 lambdas.

The Quartzite core is sized to handle up to 32 ten-Gbps packet circuits reaching 0.6 Terabits of bidirectional bandwidth, a sufficient amount of bandwidth to approximate a Terabit campus.

Complementing San Diego's Quartzite infrastructure is Chicago's StarLight facility. StarLight's key systems for supporting the OptIPuter project include a Force10 switch which services tens of gigabits of routed network traffic. We are also experimenting with two MEMS-based OOO switches, which provide the dynamic allocation of lightpaths between computing clusters at each of the three locations—a Calient, which connects to an equivalent switch at the University of Amsterdam (an OptIPuter partner), and a Glimmerglass switch at the Electronic Visualization Laboratory (EVL) at the University of Illinois. At EVL, a 30-node Opteron PC cluster driving a 17-foot wide 55-tile LCD display system (called LambdaVision) serves as one of the end systems. Applications create lightpaths by communicating through a scheduling agent with the Photonic Interdomain Negotiators (PINs) attached to each of the MEMS switches. At the recent

SC2003 conference we were able to demonstrate application-controlled photonic multicasting using the Glimmerglass photonic switch.

A key problem under investigation is how to architect the edges of the LambdaGrid so that the large amounts of aggregated bandwidth that is being created through the allocation of multiple lightpaths, can be *cost-effectively* distributed to the computing nodes to provide full bisectional bandwidth. One solution is to place OEO switches at the edges and to upgrade them to match the available or needed bandwidth. For this approach to be cost-effective computing elements must use older generation networking interfaces (for example 1Gbps NICs rather than 10Gbps NICs).

An alternative approach is to bring several 10Gbps lightpaths into an OOO switch and point them directly at the computing nodes. The computing nodes can then forward the data as needed over the cluster's existing backplane (at EVL the plan is to use Infiniband.) This scheme has the advantage that the MEMS switches are only a fraction of the cost of OEO switches that are capable of switching multiple 10Gbps flows. The use of Infiniband, or a similar cluster backplane technology, provides higher bisection bandwidth within the cluster than existing Ethernet technologies.

A significant drawback of this scheme, however, is that switching times on the MEMS devices are on the order of tens of milliseconds. A way to address this latency is to use caching at the computing nodes. Work is underway to develop LambdaRAM, a network-memory middleware system that uses high bandwidth networking to aggressively pre-fetch and cache data by predicting an application's data retrieval patterns. The first generation of the system has been tested in a high-resolution 2D image-display application called JuxtaView. The demonstration showed that LambdaRAM was able to retrieve data from a remote data server between Chicago and Amsterdam as fast as it was able to retrieve data from a local disk system [Krishnaprasad2004].

Work is underway to provide a detailed comparison of the two competing schemes under a variety of application-centered network traffic patterns. It is expected that in the short-term the MEMS-based scheme may *not* perform as well as an OEO scheme. However in the end, the small performance difference may be offset by the large cost savings. Furthermore the performance difference will likely decrease as the number of available 10Gbps lightpaths increases.

Krishnaprasad, N., Vishwanath, V., Venkataraman, S., Rao, A., Renambot, L., Leigh, J., Johnson, A., Davis, B., "JuxtaView – a Tool for Interactive Visualization of Large Imagery on Scalable Tiled Displays," Cluster 2004.