PAPER *Special Issues on Photonic IP Network Technologies for Next Generation Broadband Access*

Optical Switching Middleware for the OptIPuter

T. DeFanti,[†] M. Brown,[†] J. Leigh,[†] O. Yu,[†] E. He,[†] J. Mambretti,[††] D. Lillethun,[††] and J. Weinberger[††]

Summary
The OptIPuter is a radical distributed visualization, teleimmersion, data mining and computing architecture. Observing that the exponential growth rates in bandwidth and storage are now much higher than Moore's Law, this major new project of several universities – currently six in the US and one in Amsterdam – exploits a new world of computing in which the central architectural element is optical networking. 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 a LambdaGrid. Providing applications-centric middleware to control the LambdaGrid on a regional and global scale is a key goal of the OptIPuter and StarLight Optical Switching projects.

Key words:
Optical switching, Grid, Optical Grid, LambdaGrid, Control planes, Optical networking, Lightpath provisioning, Dynamic wavelength provisioning, WDM, DWDM, Middleware, OptIPuter, StarLight, Optical control plane, Automated optical network, User-centric lightpath provisioning, Intelligent optical signaling.

1. Introduction

Consider the problem of deploying a 6000x3000-pixel display connected in real time to massive computing and storage resources over a 100 Gigabit-per-second (Gbps) network. Computing and storage vendors have long embraced parallelism as the way to gain speed and capability from commodity parts, so it is natural, given the current lack of single-screen 6000x3000 displays and 100Gbps network interfaces, to turn to parallelism for the visualization and networking as well.

The OptIPuter's integration of parallelized visualization, storage, computing and networking is a massive multi-year task involving scores of researchers, students and staff. OptIPuter experiments are underway in San Diego using routed campus and metro-area testbeds and in Chicago using switched metro and international networks.

The OptIPuter can be thought of as an array of PC processors connected to an array of PC graphics cards and disks via a system bus that happens to be a multi-channel high-speed optical network. It is a *virtual* parallel computer in which the individual *processors* are widely distributed clusters; the *backplane* is delivered over multiple dedicated 1-10 Gbps optical wavelengths or lightpaths (called *lambdas*); and, the *mass storage systems* are large distributed data repositories, fed by scientific instruments as peripheral devices, operated in near real-time. Collaboration tools are provided on super-high-definition, tiled, mono- or stereo-screens directly connected to the OptIPuter. All of this interconnectivity needs to be scheduled to maximize throughput, a feature common to supercomputers, scientific instruments and collaboration systems, but not bandwidth. The OptIPuter will provide sufficient bandwidth and middleware between its elements such that networking can be scheduled and relieved of its historical characterization as the chief non-deterministic element in distance computing. Latency still has to be managed, but in a metropolitan-scale OptIPuter, speed-of-light latency is less than disk seek time, and even long-distance networking should be quite predictable in a fully optically-switched LambdaGrid.

The opportunity to build and experiment with the OptIPuter arose because of major technology changes that occurred over the last five years. In the early '90s, Moore's Law growth curves for CPU processing dominated the growth of storage and bandwidth. Instruction rates were the important metric, while storage and bandwidth were the tail of the computing dog. Computing carefully conserved scarce bandwidth and storage, since they were slow *peripherals* to the computer. Now, in contrast, the growth rate of optical bandwidth and storage capacity is much higher than Moore's Law. The fact that bandwidth and storage exponentials are crossing Moore's Law turns the old computing paradigm on its head: that which was scarce is now abundant and vice versa. The OptIPuter strategy will reach its first milestone when the cost of adding wavelengths on fiber between processors is less than procuring the processors, storage and/or visualization devices. (The cost of the fiber itself is not included, as we consider it the same as providing other physical facilities, such as the buildings in which the computers are housed. However, the cost of equipment to light up the fiber is part of the equation.) In the metro scale, this milestone is near.

The OptIPuter capitalizes on the rapid advances in network bandwidth made available by Dense Wave Division Multiplexing (DWDM). In the OptIPuter model,

Manuscript received February 2003.

[†] University of Illinois at Chicago, Electronic Visualization Laboratory
[††] Northwestern University, International Center for Advanced Internet Research

endpoints and lambdas are dynamically configured in response to the needs of an application. Best-effort communications over the Internet today are typically multiplexed over telecommunications circuits that demand time-consuming manual provisioning and are therefore static in nature. Capturing the promise of dynamically configured lambdas, connected quickly at an application's request, requires advances in infrastructure, middleware, network control and signaling protocols. This paper covers initial work in Chicago in support of dynamically configured lambdas for the OptIPuter.

2. Intelligent Optical Networking

2.1 The Global LambdaGrid

High-performance data communications differs from traditional data communications in that e-scientists often want to do enormous (terabyte) data transfers during scheduled timeframes, rather than send small bursty traffic on a "best effort" basis. The International Center for Advanced Internet Research (iCAIR) at Northwestern University [1], the Electronic Visualization Laboratory (EVL) at the University of Illinois at Chicago [2], and their corporate research partners are developing intelligent optical networking technology to remove barriers to optimized high-performance data communication, using StarLight, a next-generation global optical networking exchange facility in Chicago [3]. The development of this Global LambdaGrid will provide new capabilities for many advanced applications [4,5], some of which were demonstrated at the iGrid 2002 conference in Amsterdam [6]. The OptIPuter [7] is an early test of the Global LambdaGrid, for which the boundaries between applications, computers and networks dissolve.

The Global LambdaGrid will require novel methods for application-level dynamic control of resource discovery, allocation and adjustment to allow more flexibility in service provisioning, infrastructure deployment and service resource management, oriented toward dynamic multi-wavelength lightpath provisioning and supported by more flexible DWDM-based networking technology than implemented in today's static point-to-point optical networks [8,9]. These methods will allow applications to be more optical-network *aware*; that is, they will have a capability for directly discovering and signaling use of the networking resources they require, including signaling for the provisioning of lightpaths.

2.2 Survey of Emerging Architectures

This research is being conducted within the context of multiple emerging architectures being developed within standards bodies, including the ITU [10,11,12]. The communications industry has been moving toward architectural models that have fewer hierarchical layers and increased network transparency, such as providing for IP over DWDM, including IP control of optical transport networks [13]. There are four primary models (with many hybrids and variations): *overlay, signaled overlay, peering,* and *integrated.* For the overlay model, IP is supported ATM-style, with separate control and management planes for each layer.

The signaled overlay model is the focus of much current industry effort; for example, IETF's Link Management Protocol [14], and the control plane efforts of the Generalized Multi-Protocol Label Switching (GMPLS) [15,16,17,18,19] emerging standard. The GMPLS architecture identifies mechanisms for resource discovery, link provisioning, label switched path creation, deletion, and property definition, traffic engineering, circuit routing, channel signaling, and path protection and recovery.

GMPLS provides extensions of the MPLS concept, which uses an IP-based control plane. MPLS adds label headers to IP packets in order to facilitate forwarding via signaled label paths rather than simply via the destination IP address. GMPLS-specific extensions were introduced in order to extend the MPLS concept to forwarding planes that are not capable of recognizing packet boundaries, such as traditional devices based on time-division multiplexing (e.g., SONET ADMs) and newer devices, based on wavelengths and spatial switches [20]. GMPLS allows dynamic path creation across these types of circuit-oriented technologies based on information gathered about resources such as timeslots, wavelengths or ports. Path determination and optimization are based on Labeled Switched Path (LSP) creation, which gathers information to establish a lightpath and to determine its characteristics, including descriptive information (address identifiers, reachability, etc.) [21]. This type of IP control plane provides extremely high-performance capabilities for a variety of functions, such as optical node identification, service level descriptions (e.g., request characterizations), managing link state data, allocating and re-allocating resources, establishing and revising optimal lightpath routes, and determining responses to fault conditions [22, 23,24,25]. Traffic engineering extensions allow for specific CR-LDP formats and mechanisms and for RSVP-TE signaling [26,27]. Path protection is a key requirement, and requires continual monitoring of state information [28].

2.3 LambdaGrid Signaling Requirements

Although the research described herein is proceeding within the context of these standards initiatives, it departs from traditional approaches in several ways. First, it is oriented toward emerging infrastructures that envision global services based on a data communications infrastructure that is primarily dependent on layer 1 and 2 transport as opposed to routed paths. Second, it envisions a much closer integration of all infrastructure components. Also, it anticipates capabilities for applications *directly signaling their own resources*. OptIPuter projects are focused on specialized signaling methods to allow very-large-scale distributed applications to directly manipulate a wide range of optical networking functions. Such signaling will enable applications to provision and control their own dynamically provisioned lightpaths, which could be implemented as Global Dynamic VPNs (GDVPNs). These new signaling methods could be used to create Optical VPNs (OVPNs), and to extend lightpaths to edge resources through other types of dynamically provisioned layer 2 Gigabit Ethernet (GigE) links, including complete Ethernet vLANs. Some of these techniques were demonstrated at iGrid 2002, like the Photonic Data Services demonstration that set a new high-performance record for transatlantic data transit [29].

2.4 TeraAPI, ODIN, THOR and DEITI

OptIPuter applications must be much more "network aware" than most current applications, especially about changing network and edge resource dynamics. The Simple Lightpath Control Protocol (SLCP) specification [30] is a preliminary application protocol for making requests of low-level network service layers, which led to TeraAPI, a User Network Interface (UNI) that is a complete API interface between the application and low-level optical networking resources, through the following service intermediaries, currently in development:

- Optical Dynamic Intelligent Network (ODIN) [31], which provides a single point of control for a defined set of network service requests within a single administrative domain.
- TeraScale High Performance Optical Resource-Regulator (THOR), which manages the optical network control plane and resource provisioning, including dynamic provisioning, deletion, and attribute setting of lightpaths
- Dynamic Ethernet Intelligent Transit Interface (DEITI), which can extend lightpaths to other layer 2 links, currently GigE links (e.g., to allow applications to access edge resources, such as compute clusters and data storage repositories).

TeraAPI accesses the ODIN service layer, middleware between high-performance distributed applications and lower-level network service layers. Collectively, these service layers allow for dynamic, integrated coordination among the applications that may reside on a client network with various processes and resources at the optical network layer. This approach requires a policy engine; a candidate implementation is being developed at the University of Amsterdam based on the basic principles of the IETF Generic Authentication, Authorization and Accounting (AAA) architecture [32,33]

ODIN's single point of control is incorporated within a process that resides on a control server. The process has a complete understanding of the topology and current resource allocations within the administrative domain. ODIN accepts requests for resource allocations from applications over the network, listening on a TCP socket for requests from applications, and responding over a connected session linked to the applications. When resources are allocated to fulfill those requests, the process tells the requisite network switches to configure themselves to meet the application's requirements. These switches can be optical-domain switches, Ethernet switches and/or IP routers. In summary, ODIN can:

- Accept requests from clients for resources (the client requests a resource, implying a request for a path to the resource, although the specific path need not be known to the client)
- Determine an available path, possibly an optimal path if there are multiple available paths
- Create the mechanisms required to route the data traffic over the defined optimal path (virtual network)
- Notify the client and target resources to configure themselves for the configured virtual network (ODIN returns a new IP and subnet mask in response to a resource request)

Provisioning is accomplished directly by applications through THOR, an optical route allocation and management system. THOR is an interface between the ODIN service layer and a signaled overlay control plane that does the primary work of the dynamic lambda provisioning. THOR is a process that establishes and deletes lightpaths based on an understanding of application requirements, physical optical network topology, potential capabilities for resource allocations within that topology, and performance optimization. THOR components include mechanisms for receiving and

fulfilling requests, such as allocating and managing network resources (e.g., routes), and for monitoring state information such as route configuration data. After receiving application request(s) through ODIN, THOR determines the state of the lambda-based lightpaths in the optical network, determines the most optimal lightpath(s) for a particular request, creates a lightpath (e.g., an OVPN), by configuring the photonic node switches, notifies the client and the target resource to configure themselves (e.g., for use of the OVPN), and finally reallocates optical resources when they are no longer being used. THOR has an understanding of the network configuration to such a degree that it can allocate lambda-switched lightpaths representing resources at the level of multiple Gbps, but also provide for lambda resource sharing (i.e., multiple paths on a single lambda). THOR currently controls the DWDM layer by direct calls to a UNI API that Nortel Research Labs developed for OMNInet, the testbed being used for this research.

2.5 OMNInet

Much of the research described herein is being conducted on OMNInet [34], the Optical Metropolitan Network Initiative, currently deployed in Chicago and Evanston, Illinois. The OMNInet testbed was established, in part, to create a reference model for next-generation optical metro networks. It is a joint effort of SBC, Nortel Networks, iCAIR, EVL, the Math and Computer Science Division at Argonne National Laboratory, and the Canadian Network for the Advancement of Research, Industry, and Education (CANARIE). SONET-based networks are optimized for traditional communication services, not data communications; but, because OMNInet was designed to optimize for metro-area data services, it contains no SONET components. OMNInet is optimized for highly asymmetric, high-performance data communication services. In part, this initiative addresses issues such as optimization for network resources and high-performance scalability protocols as traffic flows transition from multiple Gbps, to 10s of Gbps, to 40 Gbps and above.

The research described in this section focuses on intra-domain networks. However, ODIN can be extended to inter-domain connectivity. In anticipation of such provisioning, OMNInet has been linked to StarLight and, via a transatlantic high-performance link provided by SURFnet [35], to NetherLight in Amsterdam [36]. Photonic-enabled applications are possible not only in metro-area networks such as OMNInet, but can also be extended to global networks – and the Global LambdaGrid [37,38].

3. Intra-domain and Inter-domain Dynamic Lightpath Provisioning

3.1 Optimizing Lambda Utilization

User domains typically aggregate traffic at the edge of an optical core network. The edge devices support lightpaths that are either statically provisioned or dynamically signaled across the core optical network. Statically-provisioned lightpaths are appropriate and economical for large-scale traffic aggregation, as such lightpaths need to be persistently maintained to service any active traffic. Dynamically-signaled lightpaths incur much control plane complexity and signaling overhead.

The scale of aggregation decreases as application traffic bandwidth demand increases. In the extreme case, user domains and edge devices are reduced to individual multi-gigabit applications (e.g., bandwidth-intensive science

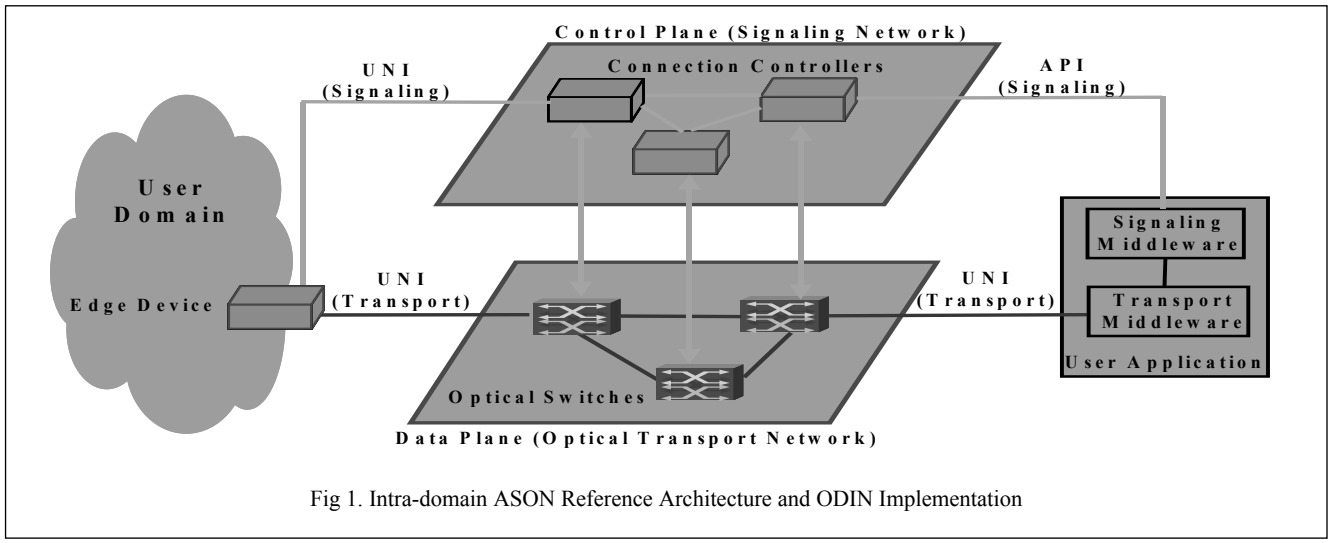

Fig 1. Intra-domain ASON Reference Architecture and ODIN Implementation

applications), which require the carrying capacity of individual lightpaths. To optimize lambda utilization within the optical core network, it seems more opportune to empower multi-gigabit user applications with dynamic lightpath signaling.

Defined by the ITU-T, the Automatically Switched Optical Networks (ASON) [39] (Figure 1) is a control plane reference architecture for enabling user domains to execute dynamic provisioning of lightpaths or optical-switched connections over the optical core network. As a reference architecture, ASON does not specify signaling and routing protocols, it only defines the components in an optical control plane and the interactions among them.

The ODIN control plane extends the ASON reference architecture by collapsing user domains and edge devices into individual user applications, which are enabled to execute dynamic provisioning of lightpaths. ODIN employs OIF Optical UNI (O-UNI) [40] between edge devices and optical switches, and IETF GMPLS [41] protocols for the control plane supported by an out-of-band signaling network.

3.2 Photonic Inter-domain Negotiator (PIN)

There is increasing interest to investigate dynamic provisioning of inter-domain lightpaths for optical networks. The Optical Border Gateway Protocol (OBGP) [42] has been proposed to enable edge devices of user domains to execute dynamic provisioning of inter-domain lightpaths over multi-domains with homogeneous local control planes. OBGP extends BGP routing with lightpath connection signaling to support lightpath route selection, setup and management. In a multi-domain environment, security management is critical and optical networks may employ different control plane and signaling protocols. For this reason, the PIN architecture is proposed to enable individual applications to execute secure dynamic provisioning of lightpaths over multi-domains with heterogeneous local control planes. PIN will also enable the dynamic scheduling of lightpaths for future deployment through a grid scheduling scheme based on an adaptation of the Globus Architecture for Reservation and Allocation (GARA) [43] for optical networks.

PIN specifies distributed domain agents to realize inter-domain routing and signaling schemes over heterogeneous optical network domains. The PIN inter-domain routing scheme consists of two routing protocols:

- Domain-level source routing to complement application-centric signaling
- BGP routing to adapt to lightpath topology changes

The PIN inter-domain signaling scheme is realized by signaling dispatchers and translators of the distributed PIN agents located in heterogeneous domains. The signaling dispatcher decides whether incoming signaling messages are terminated locally or forwarded to remote domains. For local termination, the signaling translator maps PIN inter-domain signaling messages into corresponding messages understood by local domains and passes them to the local control planes. For example, the OMNInet PIN agent will pass translated GMPLS signaling messages to ODIN.

PIN supports policy based secure inter-domain routing and signaling controls via the IETF AAA architecture [44].

3.3 Resource Sharing Optical Networks

Optical network traffic engineering functions have several basic components. First, topology information distribution is needed to advertise up-to-date information about optical links to all optical switches. Second, lightpath selection uses the topology information to compute reachability information between optical switches, and selects a route to establish an end-to-end lightpath that optimizes the use of resources. Third, lightpath connection signaling is needed to setup and release lightpaths.

Resource-sharing optical networks are differentiated according to the extent of participation of user domains in traffic engineering functions of the control plane. In general, user domains act as associated partners in traffic engineering functions. Examples of resource sharing optical networks include the TeraGrid [45] and OptIPuter, each with cluster-based user domains.

The TeraGrid is an IP-over-optical multiple network layer model with computer clusters linked by a DWDM-based optical network core. The OptIPuter is a single optical network layer model with computer clusters interconnected primarily by direct lambda links, and dedicated lightpaths through an optical-switching core or fiber-switching core with a patch panel. The OptIPuter is distinguished by a simplified, unlayered optical core network, and is based on the resource model with computing resources being the limiting factors as optical bandwidth resources become abundant and readily available.

For an aggressive, resource-trading optical network, the user domains must act as full partner in traffic engineering functions. A prominent example is CANARIE's CA*net4 network [46], which supports very fine granularity of shared optical resources; for example, user domains can share individual ports of optical switches.

4. Quanta: Adaptive Data Transport Middleware for the OptIPuter

The OptIPuter assumes it has dedicated lightpaths with multiple gigabits of bandwidth available to its applications. Full use of this bandwidth is the goal of Quanta, in concert with enabling coordinated sharing of computing resources. Reservation of optical bandwidth resources is accomplished through the PIN architecture (described in Section 3) with secure access enabled by the AAA architecture. Reservation of computing resources is accomplished through the Globus Toolkit with secure access enabled by the Grid Security Infrastructure (GSI).

The proposed architecture for Quanta (Figure 2) consists of four main components. First, a set of *Application-Centered Data Sharing* services provides the communications abstractions (such as remote procedure call, remote file I/O, etc.) and communications protocols by which data is transmitted over the networks. Each of these communications mechanisms is instrumented to record throughput, latency, burstiness and jitter. The gathered performance data is maintained by the next major component of Quanta, the *Resource Monitor*, responsible for gathering information about resources that affect communications performance, such as: available system bus bandwidth, available bandwidth on the network and CPU utilization. The Physical Network Resource Monitor keeps track of the status of the underlying networks to which an application may have access. The knowledge it develops depends on the infrastructure available for providing feedback about the network. For example, in our OptIPuter project, the PINs may regularly publicize how many wavelengths are available for use over a given fiber of the photonic network. Over a traditional electronically routed network, the available bandwidth might be determined by regular sampling of active measurement devices attached to each router.

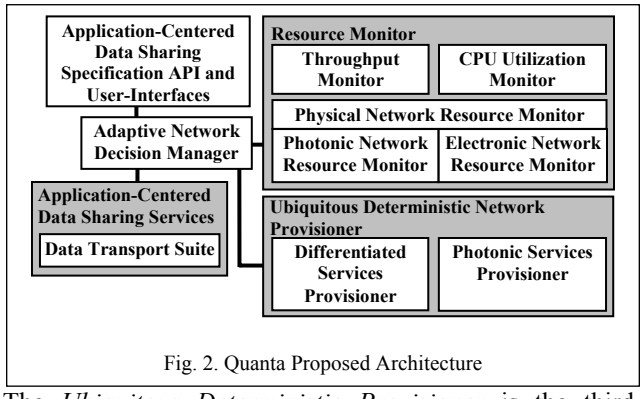

Fig. 2. Quanta Proposed Architecture

The *Ubiquitous Deterministic Provisioner* is the third component. Its role is to provide a uniform interface to allow applications to take advantage of any quality-of-service capabilities that might be available to it, such as Differentiated Services, or provisionable lightpaths. The fourth component, the *Adaptive Network Decision Manager,* examines the previous three components to make the best resource decisions on the application's behalf. The application specifies its communications requirements at a high level using the *Application-Centered Data Sharing Specification API and User-Interfaces*. Then the Adaptive Network Decision Manager translates these requirements into the set of physical network services and data sharing services needed to meet or exceed the application's requirements. During the execution of the application, the adaptive manager constantly monitors the performance of the application to make any necessary adjustments to attempt to optimize performance. This information also forms the basis of a service that the application user and developer can query if the application fails to perform well.

The data sharing services consist of a suite of C++ classes that simplify socket-level programming of TCP, UDP and multicast communications. Details are found in the Quanta API manual [47]. All the data transport classes have performance monitoring built into them so that an application can determine how much bandwidth it is using and how much latency it is experiencing. Quanta is a cross-platform toolkit; it provides a data packing API that allows applications to ensure that their transmissions are correctly translated into the format of the target computer system. It provides a set of threading and mutual exclusion classes and a number of data sharing abstractions, described next.

Data reflection is a unicast method for emulating multicast, and is one of the most heavily-used capabilities to support data sharing in collaborative applications. Clients send information to a central server rather than a single multicast address and the reflector repeats/reflects that same information to all other subscribing clients. The UDP reflector provides both unicast reflection and multicast bridging. This enables groups of clients to operate multicast within separated domains and to share information across them using a bridge rather than having to set up a multicast tunnel, which often requires system administrator privileges. The TCP reflector is similar to the UDP reflector in that it places boundaries on TCP messages (making them discrete) instead of broadcasting them as a continuous stream.

Quanta provides persistent distributed shared memory emulation via a client/server database with automatic data reflection. Hence, any updates to the database are propagated to all subscribers of the database. Clients are

notified either via a traditional callback function or via a subject/observer mechanism [49], essentially an object-oriented replacement for callbacks. The subject maintains a list of its observers for specific events and each observer is triggered whenever the specific event occurs.

Remote File I/O classes have the capability of uploading and downloading files from a remote server. The provision of both 32-bit and 64-bit versions as well as parallel socket versions allows for the efficient delivery of all file sizes, including those larger than 2 Gigabytes. The 64-bit version effectively allows delivery of terabyte-size files.

For long-distance, international networks, latencies are high (on the order of hundreds of milliseconds). In advanced collaborative applications, state updates in a shared environment should occur with a minimum amount of latency and a high degree of reliability. Data must therefore be transmitted reliably over long distances without the acknowledgement typically used in protocols such as TCP. We have applied Forward Error Correction (FEC) to achieve this [50]. FEC collects between 1 and N (typically 2 or 3) data packets and performs a bit-wise operation on the packets (such as XOR), to produce a "redundant" packet. This packet is delivered along with the regular UDP traffic as a separate UDP stream. If any data packets are lost, FEC packets can be used to reconstruct the missing packet. By using such a scheme, latency and jitter can be reduced for reliable transmission over long-distance networks.

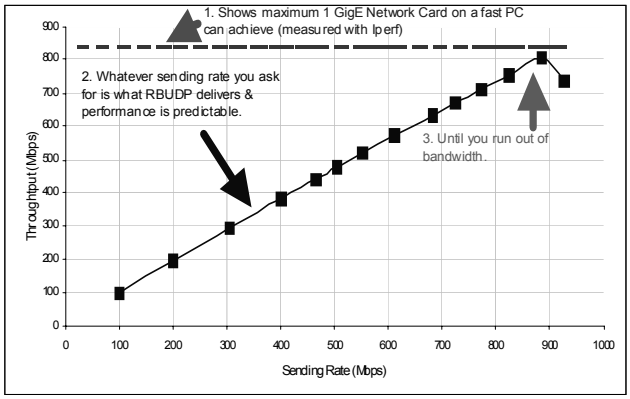

Fig. 3: Performance of Quanta's RBUDP Bulk Data Transfer Protocol

When operating over dedicated networks, the probability of packet loss is low. To take advantage of this opportunity, one can use UDP augmented with acknowledgements. The Reliable Blast UDP (RBUDP) scheme works by "blasting" the contents of a data file at just below the available bandwidth without asking the remote site to acknowledge any of the packets [51]. Hence, all the available bandwidth is used for pure data transmission. At the remote site, a tally is kept of all the packets that have arrived and, after some timeout period, a list of missing packets is sent back to the sending client. The sender reacts by resending all the missing packets and again waiting for another negative acknowledgement, and so on until done.

Figure 3 shows how well RBUDP performs on a single RBUDP stream transmitted between a pair of computers with GigE network interface cards, over a dedicated 2.5 Gbps link from Chicago to Amsterdam. The dashed horizontal line shows the maximum UDP throughput the network card is able to achieve. The diagonal line shows that RBUDP is able to utilize almost all this bandwidth for useful data transmission.

We have developed a prediction function that allows an application to predict, given the sending rate and round-trip time, the expected throughput of RBUDP [52]. The importance of this prediction function is illustrated in Figure 4. In this example, a graphics application wants to reliably stream a sequence of animations from Chicago to Amsterdam (say, with 140ms round-trip delay) [53].

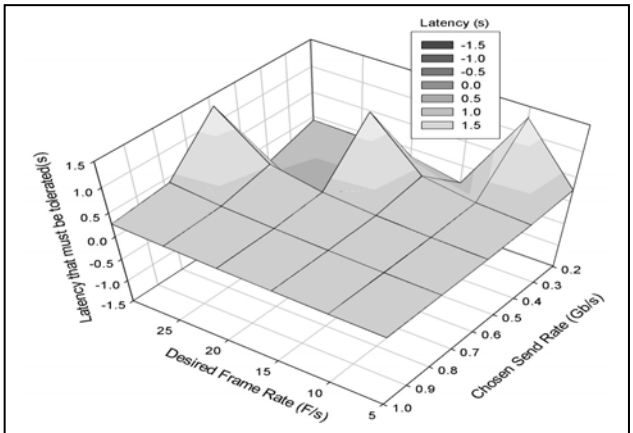

Fig 4: Quanta's prediction function predicts how much latency must be tolerated by a graphics streaming application to pack a large enough payload to achieve a desired frame rate using a chosen sending rate. Negative latency values indicate the frame rate is impossible. (Assumptions: 1024x768 24-bit image, Round Trip Time = 140ms.)

One of the caveats of RBUDP is that throughput is high only for large payloads (because a single acknowledge takes at least half the round trip time). Several 1024x768 24-bit color animation frames must therefore be packed together to form a large payload. This means that the viewer at the endpoint will experience a certain amount of latency depending on the number of frames that need to be packed. High degrees of latency may be tolerable for passive viewing, but if the goal is to stream *interactive*

graphics, achieving low latency overhead is important. Using our RBUDP prediction model, we can create a graph like the one in Figure 4 that allows the application to ask questions such as: given the desired frame rate, and the desired send rate for RBUDP, how much latency will be incurred to achieve the desired frame rate? Such prediction models, coupled with information about the physical links, are what will allow Quanta's Adaptive Network Decision Manager to help an application select the best transmission services to meet its performance requirements.

5. Conclusions

The OptIPuter is a complex project that seeks to dramatically advance the capabilities of visualization and teleimmersion through coordinated parallelized networking, storage, computing and visualization technologies. The OptIPuter relies on relatively inexpensive component visualization parts, modest computer clusters, remote data mining and computing capabilities, and extreme bandwidth to allow the user to favor resources other than networking. This is not to say, however, that networking is not the hardest part or that procuring the network bandwidth alone is sufficient or trivial; making the bandwidth work in the context of the OptIPuter involves the scheduling of lambdas on the fly, doable if and only if the middleware commands a suitable amount of networking. Bandwidth is fortunately getting cheaper much faster than disk space, which is getting cheaper much faster than computers, which are, of course, much cheaper than people.

The middleware described here will be next deployed to control Calient [54] and GlimmerGlass Networks [55] 3D MEMS switches over I-WIRE, a State of Illinois-sponsored DWDM network [56]. Further anticipated deployment in the Trans-Light project will connect Canada, Europe and Asia via StarLight [3] to form a Global LambdaGrid for the OptIPuter and many other extreme bandwidth experiments.

Acknowledgments

The advanced networking research at EVL and iCAIR are made possible by major funding from the US National Science Foundation (NSF), awards EIA-9802090, EIA-0115809, ANI-9980480, ANI-0229642, ANI-9730202, ANI-0123399, ANI-0129527, EAR-0218918, and ACI-9619019. In addition, funding is received from the State of Illinois, Microsoft Research, General Motors Research and Pacific Interface on behalf of NTT Optical Network Systems Laboratory in Japan. *StarLight* is a service mark of the Board of Trustees of the University of Illinois and the Board of Trustees of Northwestern University.

Major funding is also provided by an NSF Information Technology Research (ITR) cooperative agreement (ANI-0225642) to the University of California San Diego (UCSD) for *The OptIPuter*. OptIPuter lead institutions are UCSD and UIC, in partnership with University of Southern California, the University of California, Irvine, Northwestern University, San Diego State University and University of Amsterdam.

The authors wish to thank Caren Litvanyi of the Math and Computer Sciences Division of Argonne National Laboratory for her helpful comments on the manuscript.

References

[1] www.icair.org
[2] www.evl.uic.edu
[3] www.startap.net/starlight
[4] I. Foster and C. Kesselman (editors), The Grid: Blueprint for a Future Computing Infrastructure, Morgan Kaufmann Publishers, 1999
[5] www.globalgridforum.org
[6] www.igrid2002.org
[7] www.calit2.org, www.evl.uic.edu
[8] T. Stern and K. Bala, Multiwavelength Optical Networks: A Layered Approach. Addison-Wesley, 1999.
[9] R. Ramaswami and K. Sivarajan, Optical Networks: A Practical Perspective, Morgan Kaufmann Publishers, 1998.
[10] "G.872: Architecture of optical transport networks," Telecommunication Standardization Sector of the International Telecommunication Union (ITU), Nov. 2001.
[11] "G.874: Management aspects of the optical transport network element," Telecommunication Standardization Sector of the ITU, Nov. 2001.
[12] "G.871: Framework of Optical Transport Network Recommendations," Telecommunication Standardization Sector (ITU-T) of the ITU, Oct. 2000.
[13] G. Bernstein, J. Yates, D. Saha "IP-Centric Control and Management of Optical Transport Networks," IEEE Communications Magazine, October 2000.
[14] www.ietf.org/proceedings/01aug/slides/ccamp-1/
[15] E. Mannie, GMPLS Signaling Extension to Control the Conversion between Contiguous and Virtual Concatenation for SONET and SDH, IETF, draft-mannie-ccamp-gmpls-concatenation-conversion-00.txt
[16] E. Mannie, et. al., Generalized Multi-Protocol Label Switching (GMPLS) Architecture, IETF, draft-ietf-ccamp-gmpls-architecture-00.txt
[17] A. Bellato, G.709 Optical Transport Networks GMPLS Control Framework, IETF, draft-bellato-ccamp-g709-framework-00.txt
[18] A. Bellato, GMPLS Signaling Extensions for G.709 Optical Transport Networks Control, IETF, draft-fontana-ccamp-gmpls-g709-00.txt

[19] O. Aboul-Magd, A Framework for Generalized Multi-Protocol Label Switching (GMPLS), IETF, draft-many-ccamp-gmpls-framework-00.txt
[20] B. Davie, P. Doolan and Y. Rekhter, Switching in IP Networks: IP Switching, Tag Switching, and Related Technologies, The Morgan Kaufmann Series in Networking. New York: Academic Press, 1998.
[21] J. Lang, et. al., Generalized MPLS Recovery Mechanisms, IETF, draft-lang-ccamp-recovery-01.txt, draft-mannie-ccamp-gmpls-concatenation-conversion-00.txt
[22] Robert Doverspike and Jennifer Yates "Challenges for MPLS in Optical Network Restoration," IEEE Communications Magazine, Feb 2001, pp. 89-96.
[23] K. Kompella, et. al., OSPF Extensions in Support of Generalized MPLS, IETF, draft-ietf-ccamp-ospf-gmpls-extensions-00.txt
[24] K. Kompella, Routing Extensions in Support of Generalized MPLS, IETF, draft-ietf-ccamp-gmpls-routing-00.txt
[25] A. Banerjeet et. al., "Generalized Multiprotocol Label Switching: An Overview of Routing and Management Enhancements," IEEE Communications Magazine, March 2001, pp. 144-150.
[26] P. Ashwood-Smith, et al, "Generalized MPLS Signaling: CR-LDP Extensions," IETF Network Working Group Draft Report, May, 2001.
[27] P. Ashwood-Smith, et. al., "Generalized MPLS Signaling: RSVP-TE Extensions," IETF Network Working Group Draft Report, May 2001.
[28] J. Lang, et. al., Link Management Protocol (LMP), IETF, draft-ietf-ccamp-lmp-00.txt
[29] R. Grossman, Y. Gu, D. Hanley, X. Hong, J. Levera, M. Mazzucco, D. Lillethun, J. Mambretti, J. Weinberger, "Photonic Data Services: Integrating Path, Network and Data Services to Support Next Generation Data Mining Applications," Proceedings of the NGDM '02, Nov. 2002.
[30] D. Lillethun, J. Weinberger, "Simple Lightpath Control Protocol Specification," in preparation.
[31] D. Lillethun, J. Weinberger, J. Mambretti, "ODIN: Path Services for Optical Networks," in preparation.
[32] J. Vollbrecht, et. al., AAA Authorization Framework, RFC–2904, August 2000.
[33] C. deLaat, E. Radius, S. Wallace, "The Rationale of the Current Optical Networking Initiatives," Journal of Future Computer Systems, Elsevier Press, to appear 2003.
[34] www.icair.org/omninet
[35] www.surfnet.nl
[36] www.uva.nl
[37] J. Mambretti, "Next Generation Optical Metro Networks and the Global LambdaGrid," Annual Review of Communications, Vol. 55, International Engineering Consortium, 2002, pp. 531-534.
[38] J. Mambretti, J. Weinberger, J. Chen, E. Bacon, F. Yeh, D. Lillethun, B. Grossman, Y. Gu, M. Mazzuco, "The Photonic TeraStream: Enabling Next Generation Applications Through Intelligent Optical Networking at iGrid 2002," Journal of Future Computer Systems, Elsevier Press, to appear 2003.
[39] "Architecture for the Automatically Switched Optical Network (ASON)," ITU-T Rec. G.8080/Y.1304, Nov. 2001.
[40] "User Networking Interface (UNI) 1.0 Signaling Specification," Optical Internetworking Forum, OIF-UNI-01.0, <http://www.oiforum.com>
[41] B. Berger et. al., "Generalized MPLS – Signaling Functional Description, IETF, draft-ietf-mpls-generalized-signaling-0.9.txt, February 2003.
[42] M. Blanchet, F. Parent, B. St-Arnaud, "Optical BGP (OBGP): InterAS lightpath provisioning," Internet Draft, Jan. 2001.
[43] Ian Foster, Volker Sander and Alain Roy, "A Quality of Service Architecture that Combines Resource Reservation and Application Adaptation," Proceedings of the Eighth International Workshop on Quality of Service, June 2000.
[44] J. Vollbreecht et. al., "AAA Authorization Framework," RFC 2904, August 2000.
[45] www.teragrid.org
[46] www.canarie.ca/canet4/index.html
[47] www.evl.uic.edu/cavern/quanta
[48] E. Gamma, E. Helm, R. Johnson, and J. Vlissides, Design Patterns – Elements of Resuable Object-Oriented Software, Reading, MA: Addison-Wesley, 1995, pp. 293-303.
[49] R. Fang, D. Schonfeld, R. Ansari, J. Leigh, Forward Error Correction for Multimedia and Tele-immersion Streams, EVL Technical Report, 2000 <www.startap.net/images/PDF/RayFangFEC1999.pdf>
[50] J. Leigh, O. Yu, D. Schonfeld, R. Ansari, et al., "Adaptive Networking for Tele-Immersion," Proc. Immersive Projection Technology/Eurographics Virtual Environments Workshop (IPT/EGVE), May 16-18, Stuttgart, Germany, 2001.
[51] E. He, J. Leigh, O. Yu, T. DeFanti, "Reliable Blast UDP: Predictable High Performance Bulk Data Transfer," Proc. IEEE Cluster Computing 2002, Chicago, Illinois, September 2002.
[52] E. He, J. Alimohideen, J. Eliason, N. K. Krishnaprasad, J. Leigh, O. Yu, T.A. DeFanti, "QUANTA: A Toolkit for High Performance Data Delivery over Photonic Networks," Journal of Future Computer Systems, Elsevier Press, to appear 2003.
[53] www.calient.net
[54] www.glimmerglassnet.com
[55] www.iwire.org

Thomas A. DeFanti, PhD, is director of the Electronic Visualization Laboratory (EVL) and a distinguished professor in the department of Computer Science at the University of Illinois at Chicago. He is PI of the NSF StarLight that provides long-term interconnection and interoperability of advanced international networking, and he is co-PI of the NSF OptIPuter project.

Maxine D. Brown is an associate director of EVL responsible for the funding, documentation, and promotion of its research activities. She is co-principal investigator of the NSF STAR TAP/StarLight and Euro-Link initiatives and project manager of the NSF OptIPuter project.

Jason Leigh, PhD, received his PhD from UIC in 1998, specializing in Tele-Immersion. Leigh is the primary architect of Quanta. He is an Associate Professor of Computer Science at UIC, and co-PI of the NSF OptIPuter project.

Oliver Yu, PhD, received his PhD in electrical and computer engineering from the University of British Columbia. He held technical positions at Microtel Pacific Research, Nortel and Hughes. He last worked in Motorola as the section manager for the GPRS wireless mobile network development. He is an assistant professor in the ECE department at UIC.

Eric He received his Masters degree in Optical Engineering from Beijing Institute of Technology in 1997. He is currently working toward a Ph.D. degree in Computer Science at EVL in High Performance Network Transport Protocols, Optical Networking, Quality of Service, Cluster Computing and Virtual Reality.

Joe Mambretti, PhD, is director of the International Center for Advanced Internet Research (iCAIR) at Northwestern University, the director of the Metropolitan Research and Education Network (MREN), a partner in the StarLight/STAR TAP initiative, a member of I-WIRE, and a principal researcher on OMNInet.

David Lillethun received his BS degree in computer science from Northwestern University and is currently a research associate at iCAIR. He is a principal architect, software engineer and protocol developer for ODIN, THOR and DEITI, and has performed a number of innovative intelligent signaling techniques on the OMNInet testbed.

Jeremy Weinberger received his BS degree in computer science from Northwestern University and is currently a research associate at iCAIR. He is manager of Operations for OMNInet and a principal architect and developer for ODIN, THOR and DEITI and related protocols, and investigator of new intelligent signaling methods.