Optical Network Infrastructure for Grid

Grid High Performance Networking Research Group	Dimitra Simeonidou & Reza Nejabati (Editors) *University of Essex*
GRID WORKING DRAFT	Bill St. Arnaud *Canarie*
draft-ggf-ghpn-opticalnets-1	Micah Beck *University of Tennessee*
Category: Informational Track	Peter Clarke *University College London*
http://forge.gridforum.org/projects/ghpn-rg/	Doan B. Hoang *University of Technology, Sydney*
	David Hutchison *Lancaster University*
	Gigi Karmous-Edwards *MCNC, Research & Development Institute*
	Tal Lavian *Nortel Networks Labs*
	Jason Leigh *University of Illinois at Chicago*
	Joe Mambretti *Northwestern University*
	Volker Sander *Research Centre Jülich, Germany*
	John Strand *AT&T*
	Franco Travostino *Nortel Networks Labs*

Status of this Memo
This memo provides information to the Grid community in the area of high performance networking. It does not define any standards or technical recommendations. Distribution is unlimited.

Comments: Comments should be sent to the GHPN mailing list (ghpn-wg@gridforum.org).

Copyright Notice

1. Introduction

During the past years it has become evident to the technical community that computational resources cannot keep up with the demands generated by some applications. As an example, particle physics experiments [1,2] produce more data than can be realistically processed and stored in one location (i.e. several Petabytes/year). In such situations where intensive computation analysis of shared large scale data is needed, one can try to use accessible computing resources distributed in different locations (combined data and computing Grid).

Distributed computing & the concept of a computational Grid is not a new paradigm but until a few years ago networks were too slow to allow efficient use of remote resources. As the bandwidth and the speed of networks have increased significantly, the interest in distributed computing has taken to a new level. Recent advances in optical networking have created a radical mismatch between the optical transmission world and the electrical forwarding/routing world. Currently, a single strand of optical fiber can transmit more bandwidth than the entire Internet core. What's more, only 10% of potential wavelengths on 10% of available fiber pairs are actually lit [3]. This represents 1-2% of potential bandwidth that is actually available in the fiber system. The result of this imbalance between supply and demand has led to severe price erosion of bandwidth product. Annual STM-1 (155 Mbit/sec) prices on major European routes have fallen by 85-90% from 1990-2002 [4]. Therefore it now becomes technically and economically viable to think of a set of computing, storage or combined computing storage nodes coupled through a high speed network as one large computational and storage device.

The use of the available fiber and DWDM infrastructure for the global Grid network is an attractive proposition ensuring global reach and huge amounts of cheap bandwidth. Fiber and DWDM networks have been great enablers of the World Wide Web fulfilling the capacity demand generated by Internet traffic and providing global connectivity. In a similar way optical technologies are expected to play an important role in creating an efficient infrastructure for supporting Grid applications [5].

The need for high throughput networks is evident in e-Science applications. The USA National Science Foundation (NSF) [6,7] and European Commission [8] have acknowledged this. These applications need very high bandwidth between a limited number of destinations. With the drop of prices for raw bandwidth, a substantial cost is going to be in the router infrastructure in which the circuits are terminated. "The current L3-based architectures can't effectively transmit Petabytes or even hundreds of Terabytes, and they impede service provided to high-end data-intensive applications. Current HEP projects at CERN and SLAC already generate Petabytes of data. This will reach Exabytes (10^{18}) by 2012, while the Internet-2 cannot effectively meet today's transfer needs."

The present document aims to discuss solutions towards an efficient and intelligent network infrastructure for the Grid taking advantage of recent developments in optical networking technologies.

2. Grid applications and their requirements for high speed, high bandwidth infrastructure

It is important to understand the potential applications and the community that would use lambda or optical Grids. In today's Internet we have a very rich set of application types. These applications can possibly be categorized as follows:

- Large file transfer between users or sites who are known to each other e.g. high energy physics, SANs
- Anonymous large file transfers e.g. music and film files
- Small bandwidth streams - e.g. audio and video
- Large bandwidth streams - e.g. Data flows from instrumentation like radio telescopes
- Low bandwidth real time interactive - e.g. web, gaming, VoIP, etc
- High bandwidth real time interactive e.g. large distributed computing applications
- Low bandwidth widely dispersed anonymous users - e.g. web pages

It is still unknown what will be the major applications for lambda or optical Grids. How many of these application types will require dedicated high speed optical links in the near future? It would seem unlikely that all the application types we see on the Internet today will require optical grids. One early obvious application is large data file transfers between known users or destinations. Some researchers have also hypothesized the need for bandwidth applications - such as interactive HDTV, e-health applications requiring remote screening, high performance computing and visualization. A brief outline of some applications is given below:

- **High Energy Particle Physics**

By the nature of its large international collaborations and data-intensive experiments, particle physics has long been a demanding user of leading edge networks. This tradition is set to continue into the future. The next generation of experiments at the Large Hadron Collider (LHC) in CERN will produce vast datasets measured in tens of Petabytes per year that can only be processed and analysed by globally distributed computing resources. High-bandwidth data transport between federated processing centres is therefore an essential component of the reconstruction and analysis chain.

The LHC experiments (ALICE, ATLAS, CMS, LHCb) [9] will collide intense proton bunches every 25 ns. These collisions (likened to colliding jars of strawberry jam at 1000 miles per hour in terms of the debris created) create many hundreds of tracks in the electronic detectors, leading to a raw data rate of ~1 PetaByte per second flowing from the interaction point. Most of these events are the uninteresting debris of “glancing interactions” between the protons, however buried in the data at the level of 1 in every 10^??? are the key high momentum events resulting from interactions between the constituent quarks. It is these rare events which will signal the presence of new physics, such as the elusive Higgs particle.

The online data reduction system reduces the raw data flow to the “manageable” headline figure of a few PetaBytes per year to be stored on tape and disk. These resultant stored data sets form the basis of the analysis which collaborating physicists will then perform over a period of many years. The volume of data is so great that it is impractical to process it all at any one site. It may be argued that the limitations are as much political as technical, but in any case the data will be processed at many major national centres spread throughout the globe. This is the reason for which the particle physics community has embraced Grid technology with a vengeance [ref EDG, PPDG, GriPhyn, LCG, EGEE]. In fact, the problems do not start at the LHC turn on date in 2007, for in order to prepare for this chain – in other words to be sure that this all works in practice and that the key events survive the processing chain - the LHC experiments are today engaged in a programme of high-volume data challenges to validate their computing infrastructures. This already leads to the demand for efficient and deterministic transport of 10-100 TeraByte datasets. A 100 Terabyte dataset requires a throughput of 10 Gbit/s for delivery within 24 hours.

We can therefore see why the advent of optical network services will be at least important to, and probably crucial to, this discipline. Resource schedulers will need to be able to schedule the convergence of data, storage and compute power resources which will require scheduled replication of Petabyte scale data sets. This will be best achieved using reservable delivery mechanisms where dedicated and guaranteed bandwidth is reserved for periods of days. This is beyond what is possible today, but is well within the capability of future wavelength-switched networks.

- **Very Long Baseline Interferometry**

Very Long Baseline Interferometry (VLBI) is used by radio astronomers to obtain detailed images of cosmic radio sources. The technique achieves the highest resolution of any astronomical instrument and provides astronomers with their clearest view of the most energetic phenomena in the universe. VLBI experiments invert the Particle Physics model by bringing data from a network of distributed but co-ordinated instruments to a central point in order to correlate the signals from individual telescopes, resulting in enhanced sensitivity and resolution. The combination of simultaneously acquired signals from two or more widely separated radio telescopes can effectively create a single coherent instrument with a resolving power proportional to their spatial separation. Such instruments can achieve a resolution of milliarcseconds, which exceeds the resolution of optical telescopes. Traditional VLBI experiments record data at separate sites with high-precisions timestamps, and then each site shipped tapes or disks holding this data to a central site where correlation was performed.

This laborious and costly transport is being supplanted by so-called eVLBI, where high-speed networks are used to transfer telescope data to a correlator, for example at the Joint Institute for VLBI in Europe (JIVE [JIV03]) located at Dwingeloo in the Netherlands. This will lead to faster turnaround of results, reduced from days or weeks to hours or minutes, which greatly increases the opportunities to study transient events, such as supernovae or gamma-ray bursts. A proof-of-concept project to connect 4 or 5 European telescopes in real-time to JIVE at 1 Gbit/s rates has been agreed to take place in 2004, with data transferred in a few runs of several (maximum 12) hours each. Tere are multiple important international Radio Telescope sites world wide [10].

The proof-of-concept trials will be very valuable, but todays network can only partially satisfy the true long term operational requirements due to limited sustainable rates (typically ~ 1 Gbit/s for short periods). eVLBI experiments could today use 10 Gbit/s and with some evolution of electronics easily move to 40 Gbit/s. The advent of optical networking services to enable schedulable data transports at multi-Gbit/s throughput will increase the capability enormously, leading to improved sensitivity, increasing as the square root of data rates. Many of the telescopes are already at practical and economic limits of physical size and theoretical noise levels, and so increased data rates are the simplest route to higher sensitivity.

- **High Performance Computing and Visualisation**

The advent of optical network services comes at an opportune time for High Performance Computing (HPC) and Visualization. High-end computational science has for some 15-20 years been focused on adapting and developing parallel codes for execution on massively parallel processors and, more recently, clusters of commodity systems. In general, these target systems have been assumed to possess high bandwidth and low latency interconnects, and it has been satisfactory to neglect the variance in these quantities.

However, it is increasingly urgent to revisit these assumptions. The advent of Grid computing is making it feasible to harness heterogeneous resources for distributed computations that are impractical on any single system. However, previous work has shown that even embarrassingly parallel problems do not transfer efficiently to a wide-area (trans-Atlantic) Grid environment without (a) predictable, low-variance network QoS, and (b) introspective and adaptive work scheduling algorithms. Naively transferring tightly coupled parallel codes to a metacomputer without addressing these issues can see order of magnitude (or worse) losses in parallel efficiency.

Visualisation is crucial to deriving insight from the terabytes of data generated by a broad class of modern HPC simulations. Visualisation systems that can keep pace with high-end simulations are not, nor are likely to become, commonplace. For this reason, the importance of a researcher being able to view, and interact with, visualisations remotely is increasing. The client, simulation and visualisation typically run on different systems. The flow of data from the simulation to the visualisation requires high bandwidth links ~ 1 Gbit/s. The flow of data from the visualisation to the remote observer in the form of compressed video requires some few hundred Mbit/s with low latency and jitter in order to maintain satisfactory interactivity. These requirements increase linearly with the number of remote observers if software multicasting is used, and are doubled again if remote stereoscopic rendering is employed.

For example, the two national UK HPC services, HPCx and CSAR [11] in collaboration with the Extensible Terascale Facility [12] in the USA have recently successfully performed a demonstration of computation in the UK, visualisation in the USA, and then check-pointing and transfer of computation to the USA. This was performed as part of the SC2003 meeting in Phoenix, Arizona, 2003, winning the SC2003 bandwidth challenges [13, 14]. This type of distribution will only be possible with the availability of high capacity schedulable links, and in this case was enabled through a collaboration of Uklight, Netherlight, Starlight, and Internet2 .

We can therefore see that point-to-point links between globally distributed sites is crucial to be able to connect these sites with "pseudo-backplane" capabilities (i.e tightly bound network characteristics), allowing facilities to function in a much more coherent way than is possible today.

- **eHealth applications: proof-of-concept of remote screening**

One in eight women in the western world will get breast cancer at some stage of their lives. The earlier the diagnosis, the better the prognosis: this is the fundamental principle that underpins the breast screening process. In the United Kingdom, the Breast Screening Programme currently invites women between the ages of 50 and 64 to attend a screening session every three years, with subsequent recalls to an assessment clinic if necessary. Approximately 1.5 million women are screened each year in the UK as part of the Breast Screening Programme. It is intended that the programme will be extended to include women up to and including the age of 70 by 2004. This is expected to lead to an increase in numbers to 2.5 million women per year by 2005. Given that by the end of 2005 every woman screened will have two views per breast taken, this will result in approximately 10 million mammograms per year being taken (and stored) by the Breast Screening Programme.

Mammography poses challenges for the deployment of supporting IT systems due to both the size and the quantity of images. Digitised film results in images of approximately 32MB when digitised at 50micron (mammography is predominantly performed using film within the Breast Screening Programme); this rises to approximately 75MB when full field digital machines are employed.

A mammography radiologist will typically read, analyse, and make a decision concerning the actions to be taken for a patient, in approximately thirty seconds. This means that, in general, a radiologist will perform in the region of 100 readings per one-hour session. This amounts to approximately 100GB of data per reading session (assuming full field digital mammography). In the future this will not all be co-located, and, even more demanding, there is a move to the concept of using remote radiographers – entailing the movement of such data sets across countries and perhaps national boundaries (assuming data protection issues could be resolved).

There are two key requirements for such transfer of information: speed and security. Speed is needed so that remote radiographers can access large images on demand with an acceptable latency, and security for the obvious reasons of patient confidentiality. A typical screening centre will screen approximately 100 women per day. The average woman will be having her fourth scan, so will have three previous screening sessions' worth of data (plus any other investigative data). In reality, if a clinic were to subcontract screening work, we can expect that it would do so on a batch basis, i.e., a days' work, which is 100 patients' worth of data. On average, this would mean three previous sets of four images plus one set of four images, i.e., 16 images, for each patient. If each image was fully digital, this would result in 16 * 75M = 1.2GB of data. So, for 100 patients to be screened remotely, the network would have to move 1.2GB of data every 30 seconds. Clearly the availability of optical network services offering real-time guarantees will be important in this field.

- **Logistical Networking**

Currently those who require lambda Grids for large data file transfers are well defined communities where the members or destination sites are known to each other. Such communities include the high energy physics facilities around the world (which are broken into smaller specific application communities - ATLAS (CERN), CMS (CERN), D0 (Fermilab), KEK (Japan). Other examples are the virtual observatories, SANs and very long base line interferometer projects. These communities are relatively small and maintain long lived persistent networked relationships. The need for "anonymous" large file transfer to unknown users outside of their respective communities is currently a limited requirement.

This is not to say there will be no need for optical networks for traffic engineering, aggregation and similar "network" requirements. Emerging network concepts such as Logistical Networking (described below) impose a new requirement for high bandwidth infrastructure and promise a wide range of applications.

Difficult QoS requirements (for instance, latency lower than the speed of light allowing access to remote data) can in some cases be achieved by using large bandwidth, aggressively prefetching data across the network and storing it in proximity to the endpoint. If the data required by the application can be predicted "accurately enough" and "far enough in advance", and storage availability close to the endpoint and wide area bandwidth are high enough, then the latency seen by application may be reduced to the latency of local access, except for an initial delay in start-up. But, what if the data being prefetched is produced on demand by a cluster capable of filling the large pipe? Then the high bandwidth pipe is in fact tying together two halves of a distributed system, one the server and one the client, and the data being transferred may never exist in its entirety at the server, and it may never exist in its entirety at the client (if storage is limited, and prefetched data cannot be cached indefinitely). This is called a "terapipe," and it may have very broad applicability as an application paradigm for using high bandwidth networking and storage.This approach is an example of Logistical Networking that may have practical applications as shown in a data visualization application (Remote Visualization by Browsing Image Based Databases with Logistical Networking Jin Ding, Jian Huang, Micah Beck, Shaotao Liu, Terry Moore, and Stephen Soltesz Department of Computer Science, University of Tennessee, Knoxville, TN, to be presented at SC03). In this case the application had to be rewritten somewhat to produce data access predictions and supply them to a layer of Logistical Networking middleware that was responsible for the prefetching. In the experiments reported, the bandwidth of the pipe is not that high (20-40 Mbps) so the resolution of the images being browsed had to be limited (latency seen by the application was equivalent to local at 300x300, but not at 500x500). The size of the entire dataset was just 10GB. Increasing the resolution increases the storage and bandwidth requirements proportionately; full screen at 1400x1050 would require 100s of Mbps; serving a Power Wall at that resolution would easily require multiple Gbps of bandwidth and TBs of storage.

This "logistical" approach to using bandwidth can generate speculative transfers of data that are never used by the application. And if predictors are not good enough to mask circuit setup time, it may be necessary to keep a pipe open in order to respond to unexpected demands. On the other hand, it can allow an application to achieve latencies that are better than the lower bound imposed by the speed of light. It has the charm of not requiring a lot of detailed network programming - just a "good enough" predictor of data accesses and "high enough" bandwidth. If prestaging became a popular approach to achieving QoS, the demand for large pipes might increase greatly, particularly if good predictors were hard for application developers to supply.

2.1 Optical networking for high bandwidth applications

Grid applications can differ with respect to granularity of traffic flows and traffic characteristics such as required data transaction bandwidth, acceptable delay and packet loss. Here we specifically consider applications with high bandwidth requirements. Some of these applications (e.g. particle physics, CERN [15]) are sensitive to packet loss and require reliable data transmission. In contrast, there are high bandwidth Grid applications (e.g. radio astronomy [16]) that are sensitive to the packet loss pattern rather than the packet loss. There are also specific applications [17] that they may require bulk data transfers for database replication or load balancing and therefore packet loss minimisation is necessary to increase performance. Finally some emerging Grid applications (e.g. video-games for Grid [18]) require real time (short delay), long lived, relatively small bandwidth but potentially large number of users. Foster [19] proposes that Grid computing can support a heterogeneous set of "Virtual Organizations" (VO), each composed of a number of participants with varying degrees of prior relationship who want to share resources to perform some task.

Despite the above mentioned differences, there are two main common requirements generated by a large number of Grid applications:

- Large amounts cheap bandwidth provisioned and scheduled on-demand
- User or application management and control of the network resources (i.e. set-up self-organized distributed computing resources and facilitate bulk data transfers)

A number of other requirements concerning throughput, priority, latency, QoS and storage capacity will also influence the Grid network design but they are more specific to the type of application. Grid applications are also likely to differ in the number and type of participants, and also in the degree of trust between the participants [19].

A new network concept is now emerging to satisfy Grid application requirements. This is a network where resources such as ports, whole equipment, even bandwidth are controlled and maybe owned by the user. Furthermore, in contrast to traditional (telecommunications) networks where applications are allocated resources and routed over fixed network topologies, in Grid networks, resources under user/application control are organized in an automated way to provide connectivity without getting the permission from a carrier or a central authority. In other words, the user will drive its own virtual network topology.

Optical Technologies are best suited to fulfill some of these requirements, i.e. to offer huge capacity (theoretically up to 50 Tb/s/fiber) and relatively low latency. What's more, WDM & tunable technologies in combination with optical switching can provide dynamic control and allocation of bandwidth at the fiber, wavelength band, wavelength or sub-wavelength granularity in optical circuit, burst, or optical packet systems. Today's optical technologies support fast and dynamic response of bandwidth offering the capability to provide bandwidth services dynamically controlled by individual users/applications. This has been made possible by the development of a distributed control plane based on established IP/MPLS protocols. Based on this capability, future data-intensive applications will request the optical network to provide a point-to-point connection on a private network and not on the public Internet. The network infrastructure will have the intelligence to connect over IP network (packet) or to provide λ (circuit) to the applications. A λ service provided through OGSI will allow Virtual Organizations to access abundant optical bandwidth through the use of optical bandwidth on demand to data-intensive applications and compute-intensive applications. This will provide essential networking fundamentals that are presently missing from Grid Computing research and will overcome the bandwidth limitations, making VO a reality.

Despite these features, optical networks have been developed with telecommunications applications in mind and the implementation of a Grid optical network imposes a lot of new challenges. General requirements in this type of optical network can be summarized as follows:

- Scalable, flexible, and reconfigurable network infrastructure
  - It can be argued that initially optical grids are going to serve a small set of specialized applications and thus scaling becomes a minor and unimportant issue. However, we have already identified new applications requiring optical infrastructure and there seems to be a strong possibility that other applications will emerge. It is therefore significant addressing issues of scale. Scalability is an inherent attribute of the Grid vision, and enables the creation of ad hoc virtual organizations. Scalability considerations would be a big factor on the design and engineering decisions one would make in deploying an optical grid
- Ability to support very high capacity - Bulk data transfer
- Low cost bandwidth
- Bandwidth on demand capabilities for short or long periods of time between different discrete points across the network. Various schemes will be supported, for the management and exchange of information between Grid services (i.e. point and click provisioning, APIs and/or OGSI/OGSA services) that an application can use to exploit agile optical networks
- Variable bandwidth services in time
- Wavelength and sub-wavelength services (STS-n, optical packet/flow/burst)
- Broadcasting/multicasting capabilities
- Hardware flexibility to be able to support wide range of different distributed resources in the network
- High resilience across layers. In particular, a resilient physical layer will entail an number of features including resilient wavelengths, fast and dependable restoration mechanisms, as well as routing diversity stipulations being available to the user
- Enhanced network security and client-network relationship both at user-network level (UNI security) and network-network level (NNI and data path security)
- Ability to provide management and control of the distributed network resources to the user or application (i.e. set-up self-organized distributed computing resources and facilitate bulk data transfers)

2.2 Limitations of packet switching for data-intensive applications

In order to understand why optical networking for Grid, we need also to understand the current limitations of packet switching for Grid and data-intensive applications. The current Internet architecture is limited in its ability to support Grid computing applications and specifically to move very large data sets. Packet switching is a proven efficient technology for transporting burst transmission of short data packets, e.g., for remote login, consumer oriented email and web applications. It has not been sufficiently adaptable to meet the challenge of large-scale data as Grid applications require. Making forwarding decisions every 1500 bytes is sufficient for emails or 10k -100k web pages. This is not the optimal mechanism if we are to cope with data size of six to nine orders larger in magnitude. For example, copying 1.5 Terabytes of data using packet switching requires making the same forwarding decision about 1 billion times, over many routers along the path. Setting circuit or burst switching over optical links is a more effective multiplexing technique.

2.3 End-to-end Transport protocol Limitations

- **Responsiveness:**

  TCP works well in small Round Trip Time (RTT) and small pipes. It was designed and optimized for LAN or narrow WAN. TCP limitations in big pipes and large RTT are well documented. The responsiveness is the time to recover form single loss. It measures how quickly it goes back to using a network link at full capacity after experiencing a loss. For example, 15 years ago, in a LAN environment with RTT=2ms and 10Mbs the responsiveness was about 1.7ms. In today's 1Gbs LAN with RTT, if the maximum RTT is 2ms, the responsiveness is about 96ms. In a WAN environment where the RTT is very large the RTT from CERN to Chicago is 120ms, to Sunnyvale it is 180ms, and to Tokyo 300ms. In these cases the **responsiveness is over an hour** [15]. In other words, a single loss between CERN and Chicago on a 1Gbs link would take the network about an hour to recover. Between CERN and Tokyo on a 10GE link, it would take the network about **three hours to recover** [15].

- **Fairness:**

  In packet switching, the loss is an imperative mechanism for fairness. Dropping packets is in integral control mechanism to signal end-system to slow down. This mechanism was designed in multi streams sharing the same networking infrastructure. However, there is no sharing in dedicated optical link; thus, fairness is not an issue. There is no competition for network resources. Fairness need to be addressed in the level of reservation, scheduling and allocating the networking resources.

2.4 New transport protocols

In order to address some of the above packet switching limitations, new transport protocols have started to evolve. Examples are GridFTP FAST, XCP, Parallel TCP, and Tsunami. The enhancements in these protocols are done via three mechanisms: 1) tweaking the TCP and UDP settings; 2) transmitting over many streams; and 3) sending the data over UDP while the control is done in TCP.

Transmitting over TCP without the enhancements results in about 20Mbs over the Atlantic. Recent tests have seen GridFTP to achieve 512Mbs , Tsunami at 700Mbs , and in April 2003, FAST achieved 930Mbs from CERN to SLAC.

None of the above protocol can fully utilize OC-192 links. Statistical multiplexing of multiple streams of the above protocols can do current utilization of OC-192.

3. Photonic Network topology for Grid

The Grid enabled optical network will require the network topology to migrate from the traditional edge-core telecom model to a distributed model where the user is in the very heart of the network. In this type of network the user would have the ability to establish true peer-to-peer networking (i.e. control routing in an end-to-end way and the set up and teardown of light-paths between routing domains). To facilitate this level of user control, users or applications will be offered management/control or even ownership of the network resources of network resources from processing and storage capacity to bandwidth allocation (i.e. wavelength and sub-wavelength). These resources could be leased and exchanged between Grid users. The network infrastructure, including network elements and user interface, must enable and support OGSA. Through OGSA the Grid user can only have a unified network view of its owned resources on top of different autonomous systems. The resources can either be solely owned or shared with other users. Another topological alternative that could be used in conjunction with user-owned capacity is an OVPN. This means leasing wavelengths on commercial DWDM systems on a link-by-link basis. The status of these would be advertised to the Grid participants and they could dynamically connect capacity on a series of links together along a route they define by signaling messages.

These new topological solutions will have a direct impact on the design of optical network elements (optical cross-connects, add-drop multiplexers etc) and will impose new demands to the interface between the Grid user and network (GUNI[1]): i.e. The user through GUNI (see 3.3 for further for further details) will be able to access and manipulate the network elements. This requires propagation of significant network element information to the application interface, information that today resides almost exclusively in the provider's domain. It also implies new types of network processes for discovery, naming, and addressing. As an example:

- The optical network elements:
  - must be able to dynamically allocate and provision bandwidth on availability
  - have knowledge of adjacent network elements, overall network resources, and predefined user and network constrains
  - depending on application requirements, perform optical multicasting for high performance dynamic collaboration
  - The GUNI will be able to schedule huge bandwidth (i.e. OC768) over predefined time windows and establish optical connection by using control domain signaling (e.g. GMPLS)

[1] GUNI is the GRID User Network Interface (see section 7)

4. Optical switching technology and transport format considerations for Grid

An important consideration that would influence optical Grid network architecture is the choice of switching technology and transport format. Optical switching offers bandwidth manipulation at the wavelength (circuit switching) and sub-wavelength level through technologies such as optical packet and burst switching offering not only high switching granularity but also the capability to accommodate a wide variety of traffic characteristics and distributions.A number of optical switching technologies and transport formats can be considered:

- Wavelength switching: Wavelength switching (sometimes called photonic switching, or λ-switching) is the technology used to switch individual wavelengths of light onto separate paths for specific routing of information. In conjunction with technologies such as DWDM, λ-switching enables a light path to behave like a virtual circuit. λ-switching requires switching/reconfiguration times at the msec scale
- Hybrid router-wavelength switching:This architecture extends the wavelength switching architecture by adding a layer of IP routers with OC-48/192/768 interfaces between the Grid nodes and the optical network
- Optical burst switching: An optical transport technology with the capability of transmitting data in the form of bursts in an all-optical, buffer-less network, using either circuit switching (light paths), flow switching (persistent connection), or per-hop switching (single burst) services, depending on connection set-up message. The network is transparent to the content of a burst (analogue or any digital format) as well as to the data rate. Switching timescales will depend on the length/duration of bursts in a particular network scenario. Typical values vary from few μsec to several msec
- Optical flow switching: The switched entity is a set of consecutive packets in an active connection (ie packets form one source going to the same destination). Flow can be shorter than bursts (may be just 1 packet). A header is attached to the flow and it is routed and switched like a single packet. Buffering needed, which must be large enough to encompass the flow. Hop-by-hop path set-up. Advantages include integrity of transmitted sequence. The minimum flow duration will define the requirements for switching timescales. For optical networking at 10-40 Gb/sec, switching times at the nsec scale may be required
- Optical packet switching: The header is attached to the payload. At the switch the header is examined to determine whether payload is switched or buffered. Hop-by-hop path set up. Generally thought of as synchronous, but not necessarily so. Buffering may be a problem, due to lack of optical memory. Typical optical packet lengths vary from 50 bytes-15,000 or 30,000 bytes which clearly imposes a requirement for nsec switching technology

Most of the work to date assumes wavelength routing [20], because equipment such optical cross-connects (OXCs) is currently available. There is good evidence that optical burst or packet switching may eventually provide even better bandwidth and finer granularity [21]. In addition, application friendly switching such as optical flow switching can result in an improved end-to-end network performance [22].

The choice of format will be mainly driven by an understanding of the traffic characteristics generated by Grid applications. The expectation is that ongoing work on Grid will generate this information. It is likely that the right solution is going to vary between types of Grid applications. For example, wavelength switching may be the preferred solutions for moving terabytes of data from A to B, but appears to be inappropriate for video games applications, and the terabit router/OXC option may provide a competitive ready to deploy solution.

Decisions on switching and transport formats will also influence the design of optical network equipment as well as the management and the control of the network.

4.1 Wavelength Switching

Recent advances in Grid technology have promised the deployment of data-intensive applications. These may require moving terabytes or even Petabytes of data between data banks. However, the current technology used in the underlying network imposes a constraint on the transfer of massive amounts of data. Besides the lack of bandwidth, the inability to provide dedicated links makes the current network technology not well suited for Grid computing. A solution is needed to provide data-intensive applications with a more efficient network environment. This solution should provide higher bandwidth and dedicated links, which are dynamically allocated on-demand or by scheduled reservation. Wavelength switching (WS) is a promising solution, and the required infrastructure to realize this promise is now within reach.

Future data-intensive applications will ask the optical network for a point-to-point connection on a private network or an OVPN. Intelligent edge devices will decide to connect via a packet-based IP network or via circuit-based lambda allocations.

4.2 Wavelength Switching – Hardware Infrastructure

In this architecture the long haul networking backbone would be provided by agile all-optical networking equipment such as ultra long-haul DWDM with integrated optical cross-connects (IOXC's) providing OADM-like functionality with extensions to support degree n (n>2) nodes. Fiber could be user-owned, obtained via an IRU (Irrevocable Right to Use) agreement, or carrier owned; in the latter case the Grid network would contract for the number of wavelengths on each link which they need. Bandwidth would be available in increments of OC-48, OC-192, and eventually OC-768. Optical maintenance and optical fault isolation/recovery would primarily by the responsibility of the EMS and control plane software provided by the optical vendors. The backbone network would be controlled by a distributed control plane using GMPLS or similar technology, with sub-second connection set-up time. To allow control by the Grid infrastructure, internal network state information needed for routing and capacity management would be advertised by the network to the infrastructure. Connection changes would be controlled by signaling messages (RSVP or CR-LDP in the case of GMPLS) initiated by the Grid infrastructure. When capacity is shared between applications where there is not trust the OVPN mechanism could be used to provide firewalls and prevent unwanted contention for resources. In the event that all nodes involved in a single Grid application could not be connected to the same optical network, inter-domain connectivity would be provided using an ONNI. The ONNI would also be used to provide interworking between dissimilar technologies or different vendors where necessary. The strengths of this architecture include:

- The hardware and control technologies exist or are low-risk extensions of current work. Many vendors are at work in this space, as are the standards bodies.
- Little doubt about scalability.
- Compatible commercial networks providing the necessary functionality already have a large footprint in the U.S. and elsewhere.
- Likely to be the lowest cost, fastest, most secure, and most reliable way of transporting vary large (multi terabyte) data sets between two points (or from 1 to N points) on demand.
- Transmission times should have less variance than any of the options using packet or flow switching. This might allow improved scheduling.
- Compatible with both users owned and carrier provided networks, and also hybrids.
- Short-lived Grid relationships can establish and then tear down their optical infrastructure by use of carrier OVPN's.

The issues for this architecture include:

- Not competitive for small (< ?? GB) data transfers.
- Not appropriate for highly interactive applications involving a large number of nodes or for N-to-N multipoint applications (large N).
- Vendors need to be persuaded to make the necessary control plane extensions, and (for use of carrier facilities) carriers need to be persuaded to offer OVPN's at a reasonable price.

4.3 Wavelength Switching–Software Infrastructure for Network Scheduling

In many circumstances, Grid applications will need to make similar requests for bandwidth at specific times in the future (“future scheduling”). For these applications, there should be a facility for scheduling future allocations of wavelengths without knowledge of the underlying network topology or management protocols. In addition, other applications will need traditional “on-demand” allocations, and both models must be supported. Grid applications typically need to schedule allocation of computing and data resources from multiple sources. With the advent of wavelength switching, network bandwidth is another such resource that requires scheduling.

Services such as the Globus Resource Allocation Manager (GRAM) job scheduler have been developed to coordinate and schedule the computing and data resources needed by Grid applications. Some Grid network allocation proposals are based on DiffServ configuration and do not take into account the optical layers. These services will need to be extended to handle network resources as well. To do so, they will require facilities for scheduled allocation of wavelengths. Simple coordinating and scheduling services may need only high-level facilities. However, services that attempt to optimize network resources will need a richer interface. For example, optimization of schedules with multiple possible paths and replicas will require the ability to schedule individual segments of wavelength paths. A facility for scheduled allocation of wavelengths on switched optical networks should present a standardized, high-level, network-accessible interface. A natural choice for Grid applications is an Open Grid Service Interface (OGSI). Such interfaces are compliant with the GGF's OGSA specification and conform to widely used Web Services standards (WSDL, SOAP, XML). In addition to presenting an OGSI-compliant interface, the wavelength service should have a standard way of representing wavelength resources for communicating with clients. Unfortunately no such standard currently exists. For the Grid community, a promising approach would be to extend the XML form of the Resource Specification Language (RSL). This RSL schema is currently used by GRAM to schedule other resources. Adding network extensions to RSL would make it possible to enhance GRAM to handle network resources as well.

The General-purpose Architecture for Reservation and Allocation (GARA) provides advance reservations and end-to-end management for quality of service on different types of resources, including networks, CPUs, and disks [23, 24]. It defines APIs that allows users and applications to manipulate reservations of different resources in uniform ways. For networking resources, GARA implements a specific network resource manager which can be viewed as a bandwidth broker. The current model of GARA supports the co-allocation of multiple resources. However, since GARA is an advance reservation framework, it does not implement the services that actually perform co-allocation. For example, GridFTP is a mechanism to copy the data from remote storage to the local storage near the computation. This process is called “data pre-staging”. The GARA design supports to schedule the start of computation once the data is available locally. However, it does not actually submit the computation. A particular problem that arises in such a scenario is associated with the underlying resources. While most storage and computation exist within a single administrative domain or “points”; a network connection may cross administration boundaries and can be thought of as a “line”. A network path has a start point and an end point. This makes network resources different from CPU, and storage resources. CPU and storage resources are isolated and local, while network resources are combined and global. For example, a network path between a CPU and storage may involve a number of small networks. GARA discuses two approaches to this problem: Treating the network reservation as special case of the general co-allocation problem, or relying on appropriate signaling mechanisms in the network (i.e. bandwidth broker to bandwidth broker signaling). The first approach follows a virtualization of network services, i.e. it composes end-to-end network services out of single domain service. . Hence, this network service layer must interact with the optical network discovery facility, find the availability of network resources, and optimize the schedule and availability of the optical network resources. This service layer interfaces with the optical control plane and make the decision to use traditional IP networks or optical networks.

4.4 Wavelength Switching – Economics

Recent cost structure changes have generated new economic considerations that drive fundamentally different architecture principles for high bandwidth networking.

- Inexpensive optical bandwidth: DWDM provides multiple Lambdas, and each one of them accommodates high bandwidth over long distances. Thus, now the transmission cost per data unit is extremely low. This is a departure from the assumptions prevalent for the past 20 years. When the bandwidth is almost free, old assumptions must be reconsidered.

- Optical HW costs: Depending on the specific Grid application, simplifications and cost reductions may be possible. These include use of dumb CWDM optics rather than agile IOXC or OBS optical networks. For example, a star network with a small number of simple MEMS OXC in the center (and OBGP as protocol), might be adequate in many situations. When all the GRID nodes are close together, there are no trust issues, and the relationships are expected to be long-lasting.
- Optical costs: L3 routers can look into packets and make routing decisions, while optical transmissions do not require this functionality. Therefore, the L3 architecture in traditional routing requires substantially more silicon budget. The routing architecture in OC-192 costs about 10x more than the optical transmission equivalent. Specifically, an OC-192 router port costs about 5x as much as the Optical Cross Connect (OXC) equivalent. Furthermore, at intermediate nodes the router ports are in addition to the optical costs.
- Connectivity costs: Until recently, an OC-192 connection coast-to-coast has cost about one million dollars. The design of the new optical ultra-long-haul connection reduces the economic fundamentals of big-pipe, long-haul connections.
- Last mile costs: Previously, the last-mile connections were expensive and very narrow. Due to recent technology advances and economic restructuring, Optical Metro service has changed the principles of the access. Therefore, we believe that eventually last mile big optical pipes will be affordable for many Grid Computing and data-intensive applications.
- Inexpensive LAN bandwidth: 1GE NICs become extremely inexpensive with a new price point of $50 for copper and $100 for optical. 1 GE becomes a commodity for servers and the desktop, while the cost per port of 1Gbs switching port has fallen substantially. With the aggregation of 1 Gbs ports, we believe that this will drive a domino effect into 10GE. With this price point per bit, bandwidth is almost free in the LAN.
- Storage costs: Presently, disk prices are very inexpensive. One terabyte currently costs less than $1,000. This affordability has encouraged Grid applications to use larger amounts of data. In particular, 1 Petabyte storage systems cost approximately $2-3 million, which is within the budget of large organizations. With this new economic cost structure and affordability, it is reasonable that many Grid projects will build large data storage.
- Computation costs: Many Grid applications require massive amounts of computational power, which is nonetheless inexpensive. The computational power that we have on our desks is larger than a super computer of 10 years ago, and at a price point which is orders of magnitude lower. This phenomenon drives massive amounts of computation at low prices and in many cases require massive amounts of data transfer.

Based on these fundamental cost structure changes in many dimensions, we can expect substantial growth. It looks like Grid applications will be the first to use these new inexpensive infrastructures. The design of optical networking infrastructure for Grid applications must address these challenges in order to allow for predicted growth.

4.5 Hybrid Router/Wavelength Switching

This architecture extends the wavelength switching architecture just discussed by adding a layer of IP routers with OC-48/192/768 interfaces between the Grid nodes and the optical network. The GRID node would connect optically to these interfaces, as would the optical network. In addition there might also be connectivity directly from the Grid nodes to the optical network so that the previous architecture could be used where appropriate. The routers would be capable of providing full line-rate packet switching. Connectivity between the routers would be dynamically established by use of the UNI or extensions. This could be done under control from the Grid connectivity API, presumably. Packet routing/forwarding from the Grid node, through the router and the optical network, and to the remote Grid node could be controlled by the Grid node by use of GMPLS. The strengths of this architecture are:

- Full IP packet networking at optical speeds.
- Delay, packet loss, and costs associated with intermediate routers can be minimized by dynamically establishing direct router-router pipes for periods when they are needed.
- Can be used in conjunction with the wavelength switching architecture.
- The necessary networking capabilities are mostly commercially available.

The weaknesses include:

- Uses more resources than wavelength switching if the routers are used for giant file transfers.
- The Grid/router control interface needs definition.
- The addition of another layer will complicate OAM.

4.6 Optical Burst Switching

Many in the networking research community believe that optical burst switching (OBS) can meet the needs of the scientific community in the near term (2-3 years). For clarification, the 2-3 years timescale is relevant to early adopters such as Universities and government institutions (usually the same organizations pushing the technology envelope to meet their un-met applications' requirements), pre-standardization. The Grid community seems to fit this definition. Large carrier deployment for the public arena will come later, in practice, since network management and standards need to be in place prior to widespread deployment.

OBS brings together the complementary strengths of optics and electronics [25,26, 27, 28, 29, 30, 31 ,32]. The fundamental premise of OBS is the separation of the control and data planes, and the segregation of functionality within the appropriate domain (electronic or optical). This is accomplished by an end-user, an application, or an OBS edge node initiating a set-up message (control message) to an OBS ingress switch. The ingress switch is typically a commercial off-the-shelf (COTS) optical cross-connect (OXC). The control processor forwards the message along the data transmission path toward the destination. Control messages are processed at each node (requiring OEO conversions); they inform each node of the impending data burst, and initiate switch configurations to accommodate the data burst. The data burst is launched after a small offset delay. Bursts remain in the optical plane end-to-end, and are typically not buffered as they transit the network core. A burst can be defined as a contiguous set of data bytes or packets. This allows for fine-grain multiplexing of data over a single lambda. Bursts incur negligible additional latency. The bursts' content, protocol, bit rate, modulation format, encoding (digital or analog) are completely transparent to the intermediate switches. OBS has the potential of meeting several important objectives: *(i)* high bandwidth, low latency, deterministic transport required for high demand Grid applications; *(ii)* all-optical data transmission with ultra-fast user/application-initiated light path setup; *(iii)* implementable with cost effective COTS optical devices.

There are several major OBS variants. They differ in a number of ways: **(*i*)** how they reserve resources (*e.g.,* 'tell-and-wait', 'tell-and-go'), **(*ii*)** how they schedule and release resources (*e.g.*, 'just-in-time' 'just-enough-time'), **(*iii*)** hardware requirements (*e.g.*, novel switch architectures optimized for OBS, commercial optical switches augmented with OBS network controllers), **(*iv*)** whether bursts are buffered (using optical delay lines or other technologies), **(*v*)** signaling architecture (in-band, out-of-band), **(*vi*)** performance, **(*vii*)** complexity, and **(*viii*)** cost (capital, operational, $/Gbit, *etc*.).

Most OBS research has focused on edge-core, overlay architectures [33, 34, 35]. However, some research is focusing on OBS network interface cards (NICs) for peer-to-peer, distributed networking.

TCP and UDP variants will almost certainly be the predominant transport protocols for data communications. However, some high demand applications might require novel transport protocols which can better take advantage of OBS. OBS allows for bursts of unlimited length, ranging from a few bytes to tens or hundreds of gigabytes. This has led some in the OBS research community to rethink some of the IP protocols to better take advantage of OBS technology – no buffering, ultra-high throughput, ultra-low error rates, etc. Others are investigating simplified constraint-based routing and forwarding algorithms for OBS (e.g., that consider dynamic physical impairments in optical plane when making forwarding decisions [36, 37, 38, 39]) and on methods based on GMPLS. OBS is deployed in several laboratory test-beds and in at least one metropolitan area dark fiber network test-bed (with a circumference of about 150 Km). Proof-of-concept experiments are underway, and will continue to provide further insights into OBS technology. Also, there is an effort underway to extend GridFTP to utilize Just In Time (JIT) TAG protocol for possible improvements in performance.

Many in the scientific research community are of the opinion that today's production, experimental and research networks do not have the capabilities to meet the needs of some of the existing e-science and Grid applications. Many of these applications have requirements of one or more of these constraints: determinism (guaranteed QoS), shared data spaces, real-time multicasting, large transfer of data, and latency requirements that are only achievable through dedicated lambdas, as well as the need to have user/application control of these lambdas. Key for OBS technology is to determine early on, how the technology, protocols, and architecture must be designed to provide solutions to these requirements. This is an opportunistic time within the development stage (pre-standardization) of OBS to incorporate these solutions. Key concepts of interest to the OBS community are as follows:

- Network feedback mechanisms to user
    - Status
    - Alarms
    - Availability and reach

  - Creation of hooks to provide policy based control of network behavior
- Policy based routing algorithms: user or carriers decide on how forwarding tables are created.
- Integrating security concerns at both the protocol level as well as control and management plane.
- Incorporating necessary inter-domain information exchange in protocol definitions.
- Providing necessary flexibility in architectures to meet both carrier-owned and user-owned networks.
- Understanding the requirements for both physical layer QoS and application layer QoS and incorporating them into protocol definitions.
- Determine how users will get billed for the Grid network service
- Determine what is meant by Grid SLAs and how the network can provide them.

5. Optical switching nodes for photonic Grid

The network nodes combine edge and core switch functionalities. The edge nodes provide the interface between the electrical domain and optical domain in different layers (i.e. from control layer to physical layer). The core switches, based on the control information configure the switch matrix to route the incoming data to the appropriate output port, and resolve any contention issues that may arise.

A generic structure of an optical switch consists of an input interface, a switching matrix and an output interface. The input interface performs delineation and retrieves control information, encoded in the control packets. The switching block is responsible for the internal routing the wavebands/wavelengths or bursts/packets - depending on technology used - to the appropriate output ports and resolving any collision/contention issues, while the output interface is responsible for control update and any signal conditioning that may be required such as power equalization, wavelength conversion or regeneration. The optical switch architecture will offer features such as:

- dynamic reconfiguration with high switching speed (<ms, although a more relaxed requirement will be acceptable for very large data transfers and long duration of optical connectivity)
- strictly non-blocking connectivity between input and output ports
- broadcasting and multicasting capabilities in dedicated devices (i.e. near the source or destination)
- capability to address contention issues
- scalability
- protection and restoration capabilities
- minimum performance degradation for all paths and good concatenation performance

In terms of optical switch architectures there are a number of options already proposed in the literature, but the different proposals need to be adjusted to the set of requirements imposed by this new application framework. Especially, waveband and transparent switching are challenging issues. Features such as broadcasting/multicasting are central and need to be addressed by the proposed solution. The broadcast and select architecture may be the obvious choice, but architectures utilizing tunable wavelength converters and wavelength routing devices offer an alternative solution as optical wavelength converters may offer capabilities such as creation of multiple replicas of a single optical signal.

In terms of switching technology, different options are available. Among the main selection criteria would be the switching speed. Depending on the transport format, options may include certain switching technologies such as opto-mechanical or micro-electromechanical system (MEMS) supporting slower switching speeds (typically μsec-msec). For faster switching speeds, more appropriate switch choices are based on electro-optic or SOA technologies supporting ns switching times. These technologies commonly suffer by reduced switch matrix dimensions that can be overcome using multistage architectures. The alternative solution based on the broadcast and select architecture utilizes passive splitters/couplers and tunable filters instead of a switch fabric and in this case the challenging technology choice is associated with the tunable filtering function. A third option in terms of switching functionality is provided through the use of tunable wavelength converters and wavelength routing devices.

5.1 Multicasting in optical switching nodes a requirement for photonic Grid

Multicasting has traditionally found greatest use in multi-site video conferencing, such as on the AccessGrid where each site participating in the conference multicasts or broadcasts several 320x200 video streams to each other. However in the context of Grid computing new uses for extremely high speed multicast are emerging. These are usually data-intensive applications for which there is a real time data producer that needs to be accessed simultaneously by multiple data consumers. For example, in collaborative and interactive Grid visualization applications, extremely high resolution computer graphics (on the order of 6000x3000 pixels and beyond,) that are generated by large visualization clusters (such as the TeraGrid visualization server at Argonne,) need to be simultaneously streamed to multiple collaborating sites (we call this egress multicasting). In another example, data from a remote data source may need to be “cloned” as it arrives at a receiving site and fed into distinct compute clusters to process the data in different ways. Again using large scale data visualization as an example, a single data stream could be used to generate two or more different visual representations of the data using distinct compute clusters running different visualization algorithms (we call this ingress multicasting).

5.2 Photonic Multicasting

Strictly speaking photonic multicasting is 1:N broadcasting rather than N:N as in the classical router-based multicast. Hence this 1:N broadcast is often called a Light Tree. A Multicast-capable photonic switch (also called a multicast-capable optical cross connect switch) is a photonic switch that uses optical splitters, also referred to as power splitters, to split a lightpath into N>1 copies of itself. For an N-way split, the signal strength in each split is reduced by at least 1/N. In practice there is always a few dB loss as the light beam passes through the splitter. Hence depending on the size of N and the distance to the termination point, optical amplifiers may need to be incorporated to boost the signal. However optical amplifiers may also amplify any noise in the signal. Rouskas, Ali and others [40, 41, 42] have proposed several possible designs for power-efficient multicast-capable photonic switches and Leigh [43] in collaboration with Glimmerglass

Networks, is building a low-cost multicast-capable photonic switch to support collaborative Grid visualization applications.

To support multiple wavelengths, wavelength demultiplexers can be used to split the light into W individual wavelengths which can then be fed into W multicast-capable photonic switch units. The outputs would then reconverge onto a set of W wavelength multiplexers. This solution would support any permutation of photonic multicast and unicast in a non-blocking manner, however its use of W photonic switches with W inputs makes this solution prohibitively expensive to build [40]. Hence simpler and more modularly approaches, such as the one proposed in [43], are needed in the interim until we gain a clearer understanding of practical use-patterns for data-intensive Grid multicast applications.

5.3 Controlling Light Trees

It is well known that the problem of Routing and Wavelength Assignment (RWA) in photonic networks is far more difficult than electronic routing. When establishing a lightpath between two endpoints one needs to select a suitable path AND allocate an available wavelength. Dutta [44] shows that optimal solutions for point-to-point RWA cannot be practically found. The Multicast RWA (MC-RWA) problem is even more challenging because, if wavelength conversion is not employed, wavelength assignment must also ensure that same wavelength is used along the entire photonic multicast tree [45]. This will require the development of new control plane algorithms and software in three areas: Firstly the topology and resource discovery algorithms must be extended to include consideration for the availability and location of the multicast switches and their relevant attributes such as maximum splitter fan-out. Secondly multicast extensions to classical RWA algorithms must be made to support both lightpath and lighttree route and wavelength determination. Some excellent initial simulation-based research has already been done by [46, 47, 48, 49, 50, 51]. Thirdly, control plane software needs to be extended to handle setup and teardown of lighttrees. Consequently GMPLS protocols such as CR-LDP and RSVP-TE must be augmented to handle lighttrees.

5.4 Application of Photonic Switches as Cluster-interconnects and Ingress Multicasting for Data Replication

The use of photonic switches as interconnects for compute clusters [43] is sparked by the growing trend to move optics closer to the CPU. Savage [52] believes that in 2-5 years optical connections will move between circuit boards inside computers, and in 5-10 years chip-to-chip optical connections will emerge. Today, using multiple optical gigabit network interface cards in each node of a Grid compute cluster, it is possible and potentially advantageous to create dedicated connections between compute nodes using a photonic switching [43]. Since the paths do not go through any electronics, higher speed optical gigabit NICs (at 10G and perhaps 40G) can be used as they become affordable. Furthermore the application-level programmability of the photonic switch allows for the creation of a variety of computing configurations- for example one could connect a collection of compute nodes in several parallel chains or as a tree. This allows applications to reconfigure computing resources to form architectures that are best suited for the particular computing task at hand.

In the photonic cluster-interconnect paradigm, photonic multicasting can be an effective way to take incoming data from a remote source, duplicate it and pass it on to a number of parallel computing units that may be performing different tasks on the same data (for example, generating different types of visualizations at the same time). What this suggests is that the photonic control plane software that is currently focused on assigning wavelengths between remote domains will in the future also need to provide control for a hierarchy of subdomains at a finer granularity level than previously anticipated. That is, RWA for lightpaths and lighttrees will need to be extended to support lambda allocation in the photonic cluster-interconnect paradigm.

6. Optical network control and signalling

It is well known that a separation into a control plane and a data transport plane is necessary for an agile network. The control plane typically refers to the infrastructure and distributed intelligence that controls the establishment and maintenance of connections in the network, including the protocols and mechanisms for discovering, updating available (optical) resources in the data plane; the mechanisms to disseminate this information; and algorithms for engineering an optimal path between end points. In particular, it requires protocols for routing, protocols for establishing paths between end points, and protocols for configuring and controlling the OXCs (optical cross-connects).

Another given is the rapid replacement of centralized network control with a much more distributed model. In this paradigm, functions like provisioning new circuits and recovering from failures are performed in a distributed fashion by intelligent network elements (NEs). The network state information needed is "discovered" by the NE's communicating with each other.

An enormous amount of work on transport architectures and protocols based on these two fundamentals has been underway in both the major standards bodies (IETF, ITU-T, OIF (Optical Interworking Forum)) and in research groups. In addition many vendors have their own proprietary control plane implementations, which tend to be partially standards- based.

The Grid community will need to decide the extent to which their transport control plane will be standards-based, and the extent to which customized or less standardized protocols will be used. The next section describes the relevant work underway in the major standards bodies. There follows a section on OBGP, a relevant protocol being developed outside of these bodies. Finally a section discusses the applicability of these alternatives to the Grid world.

6.1 Standardization Activities

The IETF has long championed distributed control. More recently it has been developing IP switching methods such as Multi-Protocol Label Switching (MPLS), which provides a signaling protocol that separates forwarding information from IP header information [53, 54, 55, 56, 57]. Forwarding, therefore, can be based on label swapping and various routing options. The concept of a "label" has been generalized to include TDM time slots and optical wavelength frequencies. The IETF is now developing mechanisms, derived from these concepts, for IP-based control planes for optical networks as well as for other IP-optical networking processes [58]. This has culminated in the development of the Generalized Multi-Protocol Label Switching protocol (GMPLS), which, being conceptually a generalized extension of MPLS, expanding its basic concepts to switching domains. [59, 60, 61, 62, 63, 64, 65].

GMPLS is an important emerging standard. GMPLS provides for a distinct separation between control and data planes. It also provides for simplified management of both these functions, for enhanced signaling capabilities, and for integration with protection and survivability mechanisms. GMPLS can be used for resource discovery, link provisioning, label switched path creation, deletion, and property definition, traffic engineering, routing, channel signaling, and path protection and recovery.

GMPLS has extensions that allow it to interface with traditional devices, including L2 switch devices (e.g., ATM, FR, Ethernet), devices based on time-division multiplexing (e.g., SONET/SDH) and newer devices, based on wavelength switching and fiber (spatial) switches [66]. Therefore, GMPLS allows forwarding decisions to be based on time slots, wavelengths, or ports. Path determination and optimization are based on Labeled Switched Path (LSP) creation. This process gathers the information required to establish a lightpath and determines its characteristics, including descriptive information [67]. This type of IP control plane provides for 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, especially for rapid revisions, allocating and re-allocating resources, establishing and revising optimal lightpath routes, and determining responses to fault conditions.

GMPLS is actually an architecture which is realized in a suite of protocols, some new (e.g., Link Management Protocol (LMP [LMP ID [68]), others extensions of existing protocols (e.g., RSVP-TE - [RFC 3473 [69]). It should be noted that what is called "GMPLS routing" actually is limited to things like automatic topology and resource discovery; path computation is not presently in-scope.

ITU-T, the internationally-sanctioned telecommunications standards body, has been working on the architecture, functional models, and protocols of what it calls the Automatic Switched Optical Network (ASON), which presently is limited to connection-oriented optical networking. The ASON architecture (G.8080) is being fleshed out in a series of recommendations:

- Neighbor Discovery (G.7714)
- Signaling and Connection Management (G.7713): Defines signaling interfaces and functional requirements, including specific signaling messages, objects and procedures to realize connection provisioning. Protocol-specific recommendations are in the series G.7713.x, including some based on GMPLS (e.g., G.7713.2 which is based on GMPLS RSVP-TE).
- Routing and Topology Discovery (G.7715). Protocol specifics based on IETF protocols are expected.

In general, it appears that the ITU and IETF are moving in a consistent fashion. The ITU is increasingly relying on the IETF to provide the needed protocol expertise. The IETF in turn seems to be listening to the functional requirements coming from the ITU and the OIF, which have more input from the telecom carriers.

The GMPLS vision is that of a wide variety of technologies (including packet-switched, lambda-switched, TDM, fiber-switched) smoothly interworking. The reality is much more complex. Even though GMPLS protocols are being widely implemented, end-to-end provisioning through a single GMPLS-based domain is not a realistic solution because of vendor and technological incompatibilities, administrative and ownership constraints, and scalability [70].

The expected outcome is a control plane divided into discrete "clouds" (domains) on the basis of vendor, ownership and administration, scalability, and technology. Within clouds there will be trust and complete information sharing as needed; between clouds there may be limits on information flow based on trust, scalability issues, and/or technical differences. These control planes will interwork through "User-Network Interfaces" (UNIs) at the edges of the optical transport cloud between a client and the optical network, and "Network-Network Interfaces" (NNIs) between domains within the optical transport cloud.

Before turning to the UNI and NNI standards, two important general points need to be made:

- It is essential not to overemphasize the UNI/NNI distinction. Indeed, in the GMPLS signaling architecture these interfaces are treated as one, with a recognition that the specific information flows will differ between interface types. As we will see, there is almost a continuum of interfaces possible.
- No assumption is made about the control planes running on either side of the interface.

Turning first to the UNI: A UNI can be categorized as public or private depending upon context and service models. Routing information (i.e., topology state information) can be exchanged across a private UNI. On the other hand, such information is not exchanged across a public UNI interface, or such information may be exchanged with very explicit restrictions. The most restrictive UNI can be compared to voice telephone "dial tone": After handshakes (the dial tone) the client sends the called party's address over the UNI. The network may then respond with progress signals, culminating in a call established message.

The OIF UNI 1.0 Implementation Agreement [71]: This is based on GMPLS signaling specification (RSVP-TE and LMP, with a CR-LDP option). UNI 1.0 specifies methods for neighbor and service discovery, and allows a client to request a connection, including bandwidth, signal type, and routing constraints (diversity). UNI signaling can be between the network elements. It is also possible for one or both of them to be represented by a proxy.

Work continues on OIF UNI 2.0. Notable features under consideration include dynamic in-service modification of the connection bandwidth, multi-point connections, and Gigabit Ethernet support. The UNI could legitimately be extended in a number of other dimensions. Within the OIF, for example, proposals have been made to allow the client to specify the routing to use and to pass some topology information from the network to the client. The applications in mind were for a private UNI (in the sense discussed above). A limited NNI capability, suitable for a "private NNI" such as might be needed for a L1 VPN (see below), was identified as an application of this sort of UNI.

When these were discussed (2001-2) there was not sufficient demand and so they were not prioritized for an early OIF release.

The OIF is considering forming an "end user" working group to allow non-carrier clients to be represented in the identification and prioritization of needs.There are many OIF UNI 1.0 implementations, by both system vendors and protocol stack vendors. The OIF has held two successful Interop events, at Supercomm 2001 (with 25 vendors), and most recently at OFC 2003, where interoperability between these implementations was demonstrated. In both cases more detailed testing was hosted by the University of New Hampshire (UNH) prior to the public event.

Carriers have identified a clear need for an NNI to define administrative boundaries, to allow for scalability of routing and signaling, to isolate partitions of the network for security or reliability, and to accommodate technology differences between systems, for example, by partitioning a single carrier's network into separate single vendor sub-networks. NNI drivers and issues are discussed in [70] In addition, the Grid community and others have identified needs for the rapid establishment of connections spanning multiple carriers.

An NNI raises issues not seen in a public UNI:

- Reachability information and sufficient topology and capacity availability information to allow adequate routing must be exchanged over the NNI, but to avoid "signaling storms", especially when there is a significant failure, it is important to limit the volume of this information.
- It may be difficult or impossible to determine whether a link in one domain is diverse (i.e., has no common failure points) from a link in another domain; this greatly complicates diverse routing.
- When there is only limited trust or there are business issues involved, there may be further information sharing constraints. As can be seen in the BGP protocol used for inter-AS routing in the Internet, this can lead to considerable complexity and manual policy management [72]

Multi-domain optical routing also differs from the corresponding IP problem, most notably because the cost impact of each routing decision can be far greater. As more complex all-optical domains come into existence, additional considerations arise [73,74].

NNI architecture is based on some assumptions: independence from the protocols used within a domain; internal domain operation invisible outside the domain; and independence of intra-domain protection and restoration methods.

The OIF is working on an NNI Implementation Agreement. An interoperability event, with 12 systems and software vendors participating, was held at OFC 2003 with preliminary testing at UNH. Preliminary NNI signaling and routing specs were used as the basis for this:

- Signaling was based on the GMPLS extensions of RSVP-TE.
- Routing was based on the ITU-T G.8080 routing architecture with some details as defined in the G.7715. These require support of hierarchy using a link-state based protocol at each routing level. The protocol used was OSPF with extensions for Traffic Engineering and GMPLS, and some new (sub-) TLVs.

The OIF architecture allows several types of domain abstraction. One, comparable to that used by BGP in the Internet, replaces each domain with a single "abstract node". This can cause seriously non-optimal routing in some cases, however, so the capability of representing a domain by its "border nodes" (where inter-domain connections occur) and abstract intra-domain links connecting them is also provided.

The initial OIF NNI targets the multi-domain/single carrier space. However if there are not serious trust issues conceptually it should be more generally applicable.

Another area receiving considerable attention in all the standards bodies are Layer 1 Virtual Private Networks (L1 VPNs). There are many different types of L1 VPNs possible (see ITU Y.1312). A rough realization of a L1 VPN at the wavelength level might be as follows:

- The Grid application contracts for specific numbers of wavelengths on designated links in a provider network, including OXC capacity.
- The detailed state of this capacity, including service-affecting outages, is advertised back to the application in real time by the network.
- The application (and only the application) can add and remove connections routed over the contracted capacity. Routing can be controlled if desired by the application.

In effect this would behave like a customer-owned network.

Standards for L1 VPNs are starting to emerge. ITU Y.1312 provides requirements; Y.l1vpnarch an architecture; and the Telemanagement Forum (TMF) TMF 814 covers some important control interfaces.

Initial work in the ITU and OIF targets VPNs within a single provider network. This could be extended by use of an NNI or by putting application-owned OXCs at the network connect points.

6.2 OBGP

If a path is wholly contained within an administrative domain, it is possible to engineer an optimal path with GMPLS. However, if the path traverses multiple administrative domains, more complicated negotiation is necessary. OBGP [75] is needed to bridge the path between end points that are in different domains, and each domain may deploy a different strategy to allocate its resources.

Optical Border Gateway Protocol (OBGP) builds on the Border Gateway Protocol (BGP), the well established inter-autonomous routing system protocol [76]. OBGP is very much oriented toward the Grid concept of enabling applications to discover and utilize all required resources, including light-paths. OBGP was designed in part to motivate the migration from today's centralized networking environments with their complex hierarchies of protocols and control methods to an environment where optical network resources are shared and managed by individual organizations and communities [77]. OBGP is an interdomain lightpath management tool with capabilities for discovery, provisioning, messaging, and adjustment.

OBGP is an OGSA service that automates the establishment of new forwarding paths in the edge routers or servers on a networks as a result of the creation of optical path across one or more optical clouds. If the forwarding tables are not updated then the edge IP routers or servers will not see the new path. To date most routing "first hop" interface topology configuration is hand coded into routers and servers. OBGP automates that process. OBGP may in some cases may be part of work flow process for the establishment of an optical path where a Grid application signals individual network elements or network service abstractions (such as UNI) in, for example, CANARIE's User Controlled LightPath Software (UCLP) Grid Service instantiation.

In many cases, some higher authority may be involved to solve or arbitrate various problems concerning policies within a domain.

OBGP can be used in conjunction with GMPLS to interconnect networks and maintaining the lightpath between end-to-end connections. OBGP can also perform some optimisation in term of dynamically selecting autonomous domains and therefore improving the performance of Grid.

The combination of GMPLS, OBGP and/or other multi-domain protocols under evaluation will enable control of optical nodes, peer-to-peer connections, secure data exchange and QoS required by the Grid.

6.3 Control Plane Summary

In a dedicated optical Grid network where high volume data transfers between well known users and/or sites are the major application there are two approaches as to how an optical network could be deployed:

(a) A shared optical "cloud" with rapid switching of lambdas between users (OBS, GMPLS, ASON)

(b) A fixed optical point to point (partial) mesh between users with slow "automatic fiber patch panel" switching (OBGP)

An important infrastructure choice that will confront the grid community is deciding when/where to use standardized optical networking control plane architectures/protocols and when/where to create their own customized protocols or use some existing alternative such as OBGP.

Using standardized architectures and protocols has a number of advantages:

- These protocols will need to be supported by applications, like some of the "Virtual Organization" examples given in [78], which need temporary reconfigurable connectivity to locations best reached by use of carrier facilities.
- Likewise, if rapid reconfigurability is desired or if an application might scale to a significant size (tens of nodes) these protocols seem to be the only plausible path.
- Software implementing these protocols is frequently available on the web or from software vendors. When this can be done, the need to build up optical control plane expertise at the expense of investing in the application may be mitigated.
- Post-bubble, the vendors who do the bulk of the work in the standards forums are likely to be eager to extend standards to meet the needs of a large user community with specific needs.

Using these architectures and protocols also has disadvantages:

- Applications whose connectivity needs can be met with dark fiber owned or under IRU/long lease, and whose network size is modest and stable, and who do not need rapid reconfigurability, will likely gain little from advanced, feature-rich solutions developed for much larger and more volatile applications.

- Applications requiring unique or extremely demanding optical networking capabilities may not be able to get their needs met through the standardization process.
- To date, standardized protocols and architectures are only available for connection-oriented networking. No help for optical packet switching or optical burst switching, for example, is yet available.

7. Grid User Network Interface (GUNI)

To facilitate on demand access to Grid services, interoperable procedures between Grid users and optical network for agreement negotiation and Grid service activation have to be developed. These procedures constitute the Grid User Network Interface (GUNI). The GUNI functionalities and implementation will be influenced by:

- Service invocation scenarios
    - Direct service invocation : user/client directly requests from the optical network for a Grid service
    - Indirect service invocation : user/client requests for a Grid service through an agent on the optical network
- Control plane architecture
    - Overlay model: In this model the user sees the optical network topology as a black box and user protocols are separated from network protocols. Under this model, the optical network provides a set of well-defined services to clients
    - Peer model: network acts like a single collection of devices including user and single protocol runs by both user and optical nodes for the optical path placement and setup
- Optical transport format : it determines how to send signalling and control messages as well as data from user/client to the optical network
    - Circuit/Wavelength/frame dependent switching: signalling is sent in conjunction with the data or over dedicated signalling connection (e.g. dedicated wavelength or SDH/SONET connection)
    - Flow/burst/packet switching : signalling is send using signalling packet or control burst
    - Hybrid switching (combination of two former approaches)

There are several standard organisations working on evaluation of the optical network toward optical Internet. Among all of them the ITU-T (through ASTN frame work), Optical Domain Service Interconnect (ODSI), Optical Internetworking Forum (OIF) and IETF are involved with development of the User-Network Interface (UNI) mechanism. The UNI standards and definitions from these standard bodies can be used as a basic platform (model) for the GUNI [71,79].

7.1 Background

- **Network control model:**

Within the ODSI, OIF (UNI 1.0) and ITU-T (G.ASTN) standard bodies, the UNI is addressed considering overlay control architecture. The OIF-UNI (an extension to UNI

1.0 and later UNI 2.0) in conjunction with IETF through GMPLS (MPλS) also addresses UNI mechanism in a peer control model.

  - **G.ASTN:** It addresses the control and management architecture for an automatically switched optical transport network including the control plane of UNI and its requirements for signaling
  - **OIF UNI 1.0:** Within the Optical Internetworking Forum (OIF), the User Network Interface (UNI) 1.0 specification addresses the demand for defining a set network services, the signaling protocols used to invoke the services, the mechanisms used to transport signaling messages, and procedures that aid signaling. UNI 1.0 particularly focuses on the ability to create and delete point-to-point transport network connections on-demand. The service exposure is accomplished according to the UNI service reference configurations that rely on a client-side (UNI-C) and a network-side (UNI-N) signaling agent. Similarly to the client-side, proxy mechanisms on the network side are supported by the OIF document. The UNI-N implement is either provided by the network element itself, or by some management system. The UNI in OIF has been extended in conjunction with IETF-GMPLS to support MPλS thus it can be used in optical networks with unified control plane (peer model)
  - **ODSI:** It defines the service interface for management of optical trails as well as transaction control points and a message set used to interface with the optical network. The UNI within the ODSI standard provides a service discovery mechanism to invoke connection creation, deletion, modification and query.

- **Service invocation scenario:**

Under the direct invocation model, the client is directly attached to the transport network and is itself a member of the service signaling process. It therefore implements the signaling and, optionally, the neighbor discovery functions. Under the indirect invocation model, the client invokes transport network services using proxy signaling. Here, the proxy implements the signaling functionality and exposes the services according to its service exposure mechanisms. As a consequence, the indirect invocation model allows for an integration of UNI-based services without claiming UNI-based functionality in each client.

- **Optical transport format consideration:**

All of the UNI standards support SDH/SONET and wavelength switching transport format in optical domain. Thus there is lake of support for the Grid services that use optical flow/packet/burst in optical transport network.

While all of fore mentioned UNI standards offer a way to request a particular point-to-point connection with rather limited flexibility, it does not support a more complex agreement negotiation process such as the one proposed by the Web Service Agreement draft document of the Grid Resource Allocation Agreement Protocol (GRAAP) Working Group (www.ggf.org). Here, an agreement provider negotiates the details of a potentially future service with an agreement initiator. This process covers the admission control decisions, including policy and security information to be processed by AAA functions. Once an agreement is observed, the related service can be activated under the constraints listed in the agreement. Hence, an agreement can be used to model an advance reservation.

In the Grid enabled optical network with heterogeneous types of services and user demands, it is essential to support various types of UNI signaling and control (peer and overlay model), service invocation scenarios (direct and indirect) as well as different data transport formats (SDH/SONET and optical packet/burst/follow). This wide variety of requirements suggests that GUNI must be implemented using a combination of the various UNI standards explained before plus extra functionalities that is required to support Grid networking services.

7.2 Goals

While a high-level agreement negotiation process such as WS Agreement addresses the demand of a Grid resource management infrastructure, the signaling and data transport also needs to be developed between Service provider and the underlying optical transport network. These procedures constitute the GUNI, i.e. the service interface between a Grid service provider (indirect service invocation) or Grid user (direct service invocation) and optical transport network. The GUNI functionalities are grouped in the following categories:

- Signalling
  - Flexible bandwidth allocation
  - Support for claiming existing agreements including
    - Scheduled services, i.e. advance reservations
    - Incorporation of AAA-information
  - Automatic and timely provisioning
    - light-path setup
      - Automatic neighbour hood discovery
      - Automatic service discovery
  - Fault detection, protection and restoration
  - Propagation of service and agreement related events
- Transport
  - Traffic classification, grooming, shaping and transmission entity construction
  - Data plane security

The signalling mechanism will be responsible for requesting, establishing and maintaining connectivity between Grid users and Grid resources while the data transport mechanism will provide a traffic/bandwidth mapping between the Grid service and the optical transport network.

7.3 Functionalities

- **Flexible bandwidth allocation:**

GUNI will provide a mechanism for allocation of the required bandwidth (i.e. Wavelength or sub-wavelength) for the Grid user/service. The attribute “flexible” is used to indicate that GUNI will in principle support various bandwidth services requiring multiple wavelength, single wavelength or sub-wavelength (Burst, packet) such as a Private Line, a Relay service, a Wire service, as well as multipoint and VPN services. Finally, the term “flexible” also gives an indication the ability to control the actual service at multi-homed end-systems. The UNI 2.0 interim assessment of candidates (OIF2002.024.3) already lists various specific functions in this area, particularly:

a. Multi- and Dual-Homing
b. Optical VPNs
c. Ethernet Support (including optical)
d. G.709 Support
e. Point-to-Multipoint Connection Setup

- **Support for claiming existing agreements:**

GUNI is not aiming to support complex agreement negotiations such as proposed by WS Agreement. Instead, GUNI is supposed to be the interface to claim the service of an existing agreement. Hence, GUNI must allow for the incorporation of information that relates to an existing agreement. This covers the support of a lambda time-sharing mechanism to facilitate scheduling of bandwidth over predefined time windows for the Grid users/service (i.e. lambda time-sharing for efficient/low cost bandwidth utilization). The GUNI signaling also would be required to support ownership policy of bandwidth and the transport of authentication and authorization related credentials.

- **Automatic and timely light-path setup:**

Users can automatically schedule, provision, and set up light-paths across the network. To setup a light-path for a particular Grid service, user must be able to discover and invoke the Grid service (automatic service discovery). Note that this setup might be related to an agreement that covers a future time interval.

- **Fault detection, protection and restoration:**

As Grid services have wide variety of requirements and different level of sensitivity to transport network faults (see section 2) the GUNI must be able to support/invoke different protection and restoration signaling schemes.

- **Propagation of service and agreement related events:**

GUNI will have to address the particular demand of Grid Users/Services. The support of propagating asynchronous events allows for the development of adaptive applications and services. Also, the support of scheduled services requires the ability to notify the requester about events that result in service provisioning problems.

- **Traffic classification, grooming, shaping and transmission entity construction:**

The GUNI performs traffic classification and aggregation under supervision of service control and management plane. At transport layer (physical layer) the GUNI must be able to map the data traffic to a transmission entity (e.g. optical burst). In case of in band signaling the GUNI will provide a mapping mechanism for transmission of control messages (e.g. control wavelength allocation).

- **Security:**

The GUNI would be necessary to support a security mechanism for both control plane (signaling) supporting security credentials and policy information sourced by an agreement provider and data plane (transport). (See section 10)

7.4 Implementation (technology consideration)

The GUNI implementation will be influenced mainly by the transport network switching paradigm described in section 2.2. For example OBS technology will require a fast tuneable and reconfigurable GUNI to facilitate dynamic bandwidth allocation and lambda sharing between users.

In terms of GUNI technology, fast tuneable laser and high-speed reconfigurable hardware (e.g. fast field programmable gate arrays) are promising technology for realizing required functionality at the user interface of the optical enabled Grid network. They can meet hardware requirements for a hybrid GUNI that supports different type of signaling and transmission formats.

8. Optical Networks as Grid service environment

Optical networks can be viewed as essential building blocks for a connectivity infrastructure for service architectures including the Open Grid Service Architecture (OGSA) [80], or as "network resources" to be offered as services to the Grid like any other resources such as processing and storage devices.

This section offers some definitions of a Grid service, explores how optical network resources can be created and encapsulated as a Grid service.

8.1 Grid Services

Grid services are self-contained, self-describing applications that can be published, located, and invoked over an internet. Grid services can perform a range of functions, from simple resource requests to complicated business or scientific procedures. Once a Grid service component is deployed, other Grid services can discover and invoke the published service via its interface. A Grid service must also possess three additional properties. First, it must be an instance of a service implementation of some service type. Second, it must have a Grid Services Handle (GSH), which might be the Web Service Description Language (WSDL) document (or some other representations) for the service instance. Third, each Grid Service instance must implement a port called "GridService" which has three operations:

- FindServiceData. This operation allows a client to discover more information about the service's state, execution environment and other details that are not available in the GSR.
- Destroy. This operation allows an authorized client to terminate the service instance.
- SetTerminationTime. This operation allows the lifetime of a service to be set

OGSA defines the semantics of a Grid service instance including service instance creation, naming, lifetime management and communication protocols. The creation of a new Grid service instance involves the creation of a new process in the hosting environment, which has the primary responsibility for ensuring that the services it supports adhere to defined Grid service semantics.

8.2 Optical Network Resources

If optical networks are considered as network resources to be shared among virtual organizations one needs to specify exactly what are meant by optical network resources, how to encapsulate these resources into services, how to manage these services.

So what would be a meaningful optical network resource that could be offered at a level most useful to an application? In optical networks, possible resources may include optical an cross connect (OXC) or a photonic switching device (i.e. OBS, OPS), a fiber, a wavelength, a waveband, a generalized label, an optical timeslot, an interface, etc. [68]. These and other choices are normally coupled tightly with the intended application. For the purpose of this document, let's assume some typical network resources: 1) an optical path with a specific bandwidth requirement across two end points and 2) an optical tree with adequate bandwidth across multiple end points in a multicast situation. To be more specific, one may specify QoS constraints on these paths in terms reliability, delay, jitter, protection, alternative path, or even the exact time and duration for which the resource is needed.

Whatever the choices, it can be seen that an optical resource (as defined) will involve two or more network entities, not wholly contained within a network element. This makes the situation a bit more complicated since any reservation and allocation will involve cooperation of more than one network elements. Other Grid services such as processors, storage devices can be simply controlled and allocated (booked, reserved) by one network element without external constraints.

The situation is further complicated when a desired path traverses multiple heterogeneous administrative domain. Local management of the resource at the originating end of the path may not able to negotiate a path without involvement of some higher authority. Issues involved security and cooperation among different administrative domains have to be considered.

8.3 Optical network as a Grid service

OGSA framework demands that a service be represented as a self contained, modular entity that can be discovered, registered, monitored, instantiated, created and destroyed with some form of life cycle management.

To be OGSA-compliant, an optical network resource has to be wrapped up into an object that has name, characteristics, and facilities for invocation, monitoring. It is thus necessary for a local Grid Resource Allocation and Management [81], situated above the Optical Control Plane, to manage its resources. The local Grid Resource Allocation and Management is responsible to create as well as manage the required optical resources using GMPLS or other form of signaling.

To assist the messaging, discovery, instance creation and lifetime management functions required by a Grid service, the OGSA standard Grid Service ports include

- *NotificationSource and NotificationSink ports*. These services constitute a simple publish-subscribe system.
- *HandleMap*. This service provides the mapping between the Grid Service Handle and the current Grid Service Reference.
- *Registry*. This service allows a service instance to be bound to a registry. The Registry port also allows services to be unregistered.

- *Factory*. A Factory service is a service that can be used to create instances of other services. In Grid applications the factory service can create instances of transient application services.

A Grid service hence always requires a hosting environment to provide supplementary functions including Global Information Services, Grid Security Infrastructure, and to ensuring that the services it supports adhere to defined Grid service semantics.

8.4 Grid Resource Management issues

Few people in the Grid community thought of network as a resource in the same way as processing or storage. They are inclined either to view the network as a bottleneck or, if bandwidth resources are plentiful, to take the network for granted without the need for reserving options for their applications. This view was reflected in the early architecture of the Globus Resource Allocation Management (GRAM) architecture. Advances have been made, however, the "network resources" managing problem is far from being solved. This section takes a look at various issues concerning the encapsulation and allocation of optical network resources.

In the network community, network resources are often statically allocated, or allocated on-demand. In the Grid community, resources are often reserved, allocated, and even scheduled. In cases where only on-demand allocation is required, existing reservation techniques may be adequate. In other cases, co-reservation and co-allocation may be necessary to cope with staging in a heterogeneous environment [24]. In cases where flexible scheduling is necessary to resolve conflicting requests for resources, additional protocols involving cooperation are required to make sure a scheduled plan is acceptable among all participants.

- ***Globus Resource Management Architecture.***

A Resource Management Architecture for Metacomputing Systems [82] was proposed to deal with the co-allocation problem where applications have resource requirements that can be satisfied only by using resources simultaneously at several sites. In this architecture, an extensible resource specification language (RSL) is used to communicate requests for resources between components: from applications to resource brokers, resource co-allocators and resource managers. A Monitoring and Discovery Service (MDS) is a service that houses information pertaining to the potential computing resources, their specifications, and their current availability. Resource brokers are responsible for taking high-level RSL specifications and transforming them into more concrete specifications (ground requests) that can be passed to a co-allocator which is responsible for coordinating the allocation and management of resources at multiple sites. Resource co-allocators break a multirequest that involves resources at multiple sites, into its constituent elements and pass each component to the appropriate resource manager. Each resource manager (GRAM, Globus Resource Allocation Manager) in the system is responsible for taking a RSL request and translating it into operations in the local, site-specific resource management system.

- ***Advance reservation.***

The realization of end-to-end quality of service (QoS) guaranteed in emerging network-based applications requires mechanisms that support first dynamic discovery and then advance or immediate reservation of resources that will often be heterogeneous in type and implementation and independently controlled and administered.

The GRAM architecture does not address the issue of advance reservations and heterogeneous resource types. The absence of advance reservations means that we cannot ensure that a resource can provide a requested QoS when required. The lack of support for network, disk, ands other resource types makes it impossible to provide end-to-end QoS guarantees when an application involves more than just computation.

To address this problem, the General-purpose Architecture for Reservation and Allocation (GARA) was proposed [24]. By splitting reservation from allocation, GARA enables advance reservation of resources, which can be critical to application success if a required resource is in high demand. Also, if reservation is cheaper than allocation, lighter-weight resource reservation strategies can be employed rather than expensive and immediate allocation of actual resources.

- ***Service scheduling and Agreement-based Service Management.***

The most challenging issue in the management of resources in Grid environments is the scheduling of dynamic Grid services where negotiation may be required to adapt application requirements to resource availability, particularly when requirements and resource characteristics change during execution. The deployment of such environments requires the ability to create Grid services and adjust their policies and behavior based on organizational goals and application requirement.

WS-Agreement negotiation model was proposed [83] allowing management in these environments where centralized control is impossible. The WS-Agreement model uses agreement negotiation to capture the notion of dynamically adjusting policies that affect the service environment without necessarily exposing the details necessary to enact or enforce the policies.

WS-Agreement is based on Agreement services that represent an ongoing relationship between an agreement initiator (a user, a client or an application manager) and an agreement provider. It also defines the behavior of a delivered service to the client.

WS-Agreement model uses agreement negotiation to arrive at a mutual understanding of service provider behavior. Negotiation is a stateful dialogue. It may be as simple as a single request message being allowed (or not) by policy, or it may involve a complicated scenarios where the policies and intermediate commitments of the two parties are revealed piece by piece over a long sequence of message exchanges, resulting in an agreement capturing an intersection in their policies.

The WS-Agreement model defines two essential portTypes: the Agreement service and the AgreementFactory service. The AgreementFactory supports the creation of the Agreement servicve. A client negotiates agreements by invoking the createService operation of the AgreementFactory service of an agreement provider with appropriate argument content (requested terms). Some input CreationParameters are fixed while others may be negotiated.

As a result of the negotiation process, an Agreement service may be created if all the agreement terms are acceptable (observed) by the provider, otherwise a fault response is returned. An Agreement service should always relate to a "delivered service "behavior which may involve a Grid service. It may relate to an "existing service" known by the agreement provider. In this case the Agreement represents an aspect of policy affecting the behavior of that service. Alternatively, the Agreement service may relate to a "new service" which will be created due to the agreement. In this case, the Agreement service may represents, on the part of the agreement provider, both a commitment to create the new service and policy affecting the behavior of the new service.

Realizing that relationships can be formed between services and the relationships between Agreement services may be dynamically changed during the agreement lifecycle, a WS-Agreement service is endowed with the capability to expose rich, dynamic relationships to other services. This flexibility is achieved by inheriting the ServiceGroup portType and defining ServiceGroup entry content to characterize the relationship of member services to the service presenting the member in a ServiceGroup entry. An interesting relationship is the agreement composition relationship which provides a coordinated interface to multiple Agreement services.

As mentioned earlier, flexible scheduling is necessary to resolve conflicting user requests for resources, additional protocols involving cooperation are required to make sure a scheduled plan is acceptable among all participants. WS-Agreement is a Grid interface that is being specified by the GGF as a protocol and interface for managing Grid services.

It is believed that WS-Agreement model presents a very useful framework for effective scheduling of Grid resources. Adopting this model of cooperating agreement is essential in providing interoperability in the Grid heterogeneous environments. However, it is equally important to ensure that an OGSI-Agreement model remains simple and realistic. It has the potential of evolving into an over-complicated model which cannot be deployed effectively.

8.5 Network Specific Issues

- ***Level of abstraction for encapsulating network resource Grid service.***

As defined earlier, a Grid service is a self-contained, self described application that can be published, located and invoked over a network. By this definition, capacities offered by a network end point do not constitute a network Grid service. Multiple end points must cooperate to establish a network Grid service. For example, reservation along an end-to-end path is required to establish a network pipe between end points with a certain bandwidth capacity. By comparison, other resources such as storage capacity or processing capacity can be offered by a node without cooperation with other nodes. For this reason we believe that a different of abstraction is required to model network resource as a Grid service.

Attempts have been made to treat network resources as first class citizens like other resources. However, very little attempt has been made to understand the differences in the nature of a network Grid service relative to other Grid services. Simple solution is often offered whereby the source domain is given authority to take care of everything in establishing a network resource Grid service. This is not always a feasible solution across different administrative domains.

Another approach (the network community approach or IETF-like approach) is to use signaling mechanisms of the control plane (such as GUNI) of a network to establish a quantifiable network resource. The problem here is how to agree on AAA functions, integrate signaling mechanisms, and negotiate policies across multiple domains.

We believe that with WS-Agreement services, elegant solutions may be found. We suggest that Agreement services be established at several levels depending on the nature of the resources. For example, in the case of network resources within a virtual organization, an Agreement service at the VO level is necessary to establish the overall policy over its multiple domains and other Agreement services at domain level are required to negotiate and encapsulate an end-to-end network resource satisfactorily. By doing so, network resource can easily be encapsulated as a Grid service.

- ***Agreement negotiation and service initiation.***

Another issue is the need to distinguish between the agreement negotiation and the negotiated service initiation. It is believe that the separation is necessary for a number of considerations:

- The two activities belong to different phases of establishing a network Grid service. Agreement negotiation can take care of AAA functions and other policy matters and service initiation allows GUNI or other interfaces to invoke the negotiated network service.
- WS-Agreement-based services are generic and can be used with services other than network services.
- The separation makes the design of components cleaner, reusable and efficient

8.6 High level GUNI services

It is assumed in this section that the WS-Agreement-based set of services has been satisfied. These services as mentioned above will handle all policy and AAA related negotiations and agreements. This could also contain the policy of inter-domain type of interactions from the originating end. Since most connections will involve inter-domain interactions as well as different signaling protocols, the Grid service interface shall be generic enough to not exclude the different protocols (e.g. UCLP, GMPLS, JIT, etc.) or the different technologies. (SONET, GigE, PXC (photonic cross-connects)). The Grid service will basically request optical connectivity, the service implementation will translate that request to the local context and user WS-Agreement for the signaling protocol of choice. A single optical connection can include a combination of several signaling protocols and technologies, that the user may or may not be aware of.

This section assumes that the first phase of establishing a network Grid service, the WS-Argreement has occurred with an end user which takes care of all policy matters as mentioned above. What follows, are some basic generic services imitated by the application for connection establishment, connection monitoring, connection notifications, and advanced scheduled connections.

Initially we will break these tasks into three Grid services: 1) CreateOpticalConnection, 2) CreateScheduledOpticalConnection, 3) QueryOpticalNetwork.

Instances of these services will interact with each other as well as other Grid services before resolving the request. An example of operations and Service Data elements for an optical network service follows:

CreateOpticalConnection Service Operations

Operation Name	InputMessage	OutputMessage	Coments
getConnection	DestinationAdress SourceAddress wavelength QoSData BW Duration DataSize ApplicationQoS Protocols	ErrorCode ConnectionId	Other than the destination address, many of these input parameters should have default values.
getSourceRoutedConnection	DestinationAdress RoutingInfo wavelength QoSData BW Duration DataSize Protocols	ErrorCode ConnectionId	This request may be required for lightpaths that require a certain route due to policy, negotiation, or quality reasons.
getAllPhotonicConnection	DestinationAdress wavelength QoSData BW Duration DataSize Protocols	ErrorCode ConnectionId	Connection that does not undergo OEO conversion on the data – all-photonic connection
get AddressTranslation	protocolAddress	commonAddress	An example: translation from an ATM address to an IP address may be useful.
isReachable	destinationAddress	boolean	Within the WS-Agreement, is a particular destination address reachable or not.
isAvailable	destinationAddress wavelength BW QoS	boolean	Is the network connection available at this time.

CreateOpticalConnection Service Data Elements

Service Data	Type	Sequence
connectionStatus	:boolean	
connectionBW	int	
DestinationAdress		
wavelength		
QoSData	Sequence	TransprotBER ReastorationTime Priority PreEmption
ApplicationQoS	Sequence	Jitter Delay BW
BW	int	
Duration		
DataSize		
Protocols	Enum	

9. QoS for the Grid optical transport

QoS of an optical transport network will play an important role in the future of high-demand Grid computing. Optical connections in a Grid environment will be initiated on an as needed basis by the Grid applications, and that each connection request will have an associated set of optical transport QoS requirements. The following are potential QoS parameters for which a Grid application may request: i) optical layer restoration times, ii) priority and preemption of a connection, iii) physical layer signal degradation (application BER). The Grid application's connection request will contain the appropriate QoS parameters to meet the application's needs. Physical layer impairments are a key concern in high-datarate optical networks and will play a significant role in future Grid networks and SLAs. This section discusses some of the issues related to physical layer QoS.

Advances in optical technologies, faster transceivers, higher quality fibers, faster photonic switches, will generate significant changes in future optical networks. Most of today's currently deployed optical transport networks have the following characteristics: i) small all-optical islands, ii) relatively Low bit rates (less than 10Gig), iii) static wavelength configuration, iv) over engineered to reflect a more homogeneous (from a physical layer QoS perspective) network (all routes have low BER), v) more OADMs than photonic switches. In contrast, it is predicted that future optical networks will migrate towards the following characteristics: i) large All-optical islands (no OEO regeneration) – end-to-end optical connections, ii) heterogeneous signals (modulation format, datarates, protocols), iii) higher bitrates > 10 Gig, iv) dynamically reconfigurable at wavelength, and sub-wavelength levels, v) multiple physical layer QoS levels.

9.1 Physical Layer Impairments

A number of publications [84,85,86,87,88,89,90]state that physical optical impairments play a more significant role in signal degradation at bit rates greater than 10Gb/s. Most carriers have started experimentation with 40Gb/s and research is well underway for 160Gb/s. The following are a list of some of the physical layer impairments, which cause signal degradation:

Linear impairments:

- ASE - Amplifier Spontaneous Emission
- PMD - Polarized Mode Dispersion
- CD - Chromatic Dispersion

Nonlinear impairments:

- SPM - Self-phase modulation
- XPM - Cross-phase modulation
- FWM - Four-wave mixing
- SRS - Stimulated Ramman scattering effects
- SBS - Stimulated Brillion

9.2 All-photonic networks

A goal of most optical switching technologies (lambda, packet, burst, etc.) is to increase the all-photonic island (no OEO). Having an all-photonic network connection provides the following advantages: i) a unique capability where only the two end-point transceivers need to understand the format, protocol, data rate, etc. of the data transmitted, ii) low latency across the network (assuming application level latency and jitter requirements are handled at the edges) iii) no OEO (reduced NE costs). Examples include: raw data sent from instrumentation to remote processing systems, non-IP applications (HDTV), analog data, etc. This is not the case when OEO is involved. This could be very useful in high-end Grid applications, where the sharing of data requires mainly compatible transceivers. The network is completely unaware of the contents of the transmitted signal. Alongside these benefits, exists an increase in physical layer impairments resulting in higher BER for applications.

As stated earlier, the goal for future optical networking technologies is to increase the size of the all-photonic island as well as increase datarates, both of which increases signal degradation. Increased signal degradation forces a reduction in the all-photonic island, which puts the above concepts at odds with each other. One strategy that may allow the two to co-exist is to 1) integrate physical layer quality monitoring information into dynamic routing algorithms, 2) provide network connection services for different levels of physical layer QoS based on loss-sensitivity of the application. The rationale being, that different application streams have different BER requirements, e.g. voice connections can tolerate BERs as high as 10^{-4}, while real time quality video require 10^{-9}. Several mechanisms exist which compensates for error (signal loss) on optical connections, among them is Forward Error Correction (FEC) is a mechanism, re-transmission, etc. High-end Grid workstations may have FEC mechanisms available for their optical connections. Each application will request a connection with the appropriate physical layer QoS parameter to meet its BER tolerance. This will allow some applications to transmit in the all-photonic plane (no OEO) at higher BER for longer distances.

9.3 Physical layer monitoring and the control plane

Today's optical networks use optical layer monitoring for determining the max # of hops, max length of spans, and max # of Amplifiers before regeneration in order to maintain low BER (10^{-9} to 10^{-12}). It may be necessary to also utilize optical monitors throughout the network and integrate quality monitoring information into the control plane for routing and forwarding. The routing algorithm can incorporate link-based as well as channel-based quality monitoring information and provide the following benefits: i) provide dynamic compensation per channel/link – as needed basis (research stage), ii) pre-determine end-to-end physical layer QoS (BER) of a route based on quality monitoring information iii) allow data to be maintained in the optical plane for longer distance than current practice. There are key challenges regarding the assurance of a requested physical layer QoS was met on a per connection bases. Grid SLA agreements will require end-to-end assurances of connection QoS. Monitoring information capture and flow for the Grid environment is a current topic of research and analysis.

Grid users shall be provided query mechanisms (Grid network monitoring services) for determining a route's BER (source to destination); based on returned information, the Grid application may choose whether a particular route is suitable for the applications loss requirements. Grid applications should be made aware of their loss tolerance for end-to-end connections from the network. It is realized that most of today's application have not been tested for their BER requirements. However, due to the potential high BER of a wireless networks, some applications are now being analyzed for their loss tolerance. It would be useful for Grid applications to follow the wireless model. For Grid environments, many in the Grid community tend to think that routing decisions should be left up to the network, but providing a mechanism for the Grid user/application to request their required level of physical layer QoS (end-to-end BER).

In a Grid environment as mentioned earlier, one end-to-end connection can traverse multiple domains, multiple technologies, including the signaling protocols (UCLP, JIT, OIF-UNI, etc). Each domain will have monitoring capability; it is not clear if connection-based QoS information will need to be collected only at the endpoints or throughout the network. WS-agreements are Grid based service SLAs, monitoring information is necessary to reveal whether or not an agreement has been violated. SLA violation should utilize the OGSI grid services notifications (both pull and push models) to alert end users of where and how the violation occurred and the resulting action.

An end-to-end connection traversing multiple technologies (GMPLS, UCPL, etc.) will require adequate translation of the connection requested QoS parameters. Inter-operation will be an important challenge for the grid community to resolve.

9.4 Potential Optical Grid Networking Architectures

The handling of optical layer QoS for the Grid environments will be dependent on existing and future Grid networking architectures. The Grid community must first determine key characteristics of a Grid (VO). Will the Grid community target specific research communities (e.g. high energy physics) which have the following characteristics: i) relatively small numbers of participating locations, ii) long lived relationships (years), iii) participants have a high degree of trust. Or, is the target more of a ad hoc "virtual organizations" (as defined in Foster et al, "Anatomy Of The Grid"), which has the following characteristics: i) participating locations determined by VO needs – unpredictable, ii) number of simultaneous VO's could be large, iii) trust levels, longevitiy of the relationships, etc. will vary by VO. Many in the Grid community are leaning more towards the latter which will require the following optical networking strategy: i) networking protocols must be scalable, robust, not assume trust, ii) VO optical infrastructures likely to vary (customer owned, IRU, leased), iii) multiple optical control domains may need to cooperate to support A given VO.

Many in the Grid community might be converging on the peer model versus the overlay model. Optical control and network state information in the peer model is shared between users and the network. This will require user software for network security and robustness, including participating multi-user optical network providers to set up firewalls to keep any user from compromising other users. Since, commercial protocol development to date has been overwhelmingly focused on overlay models; it is highly recommended that the GGF work with the IETF and the OIF to define mutually acceptable form of peer model (OVPN).

10. Security Considerations

10.1 Threats

Active/passive attacks are grouped in the following three categories (A, B, and C) according to their target.

A. Attacks on out-of-band user-network and network-network signaling:

A.1) Acquire confidential data and identities by snooping traffic

A.2) Modify packets (e.g., a downgrade attack to lessen security agreements)

A.3) Inject new packets

A.4) Man-in-the-middle attack at setup time, with user or network impersonation, and hijacking of traffic

A.5) Mount DoS attack against legitimate signaling traffic

A.6) Disrupt the security negotiation process

A.7) Traffic analysis

A.8) Covert channels

A.7 and A.8 are the most speculative ones (no evidence of grid communities with sensitivity to these types of attack).

B. Attacks on in-band user-network signaling (as seen in flavors of OBS):

B.1) A malicious user can wreak havoc by abusing semantics (e.g., get authorization to proceed with "tell and wait" and use "tell and go" instead). A stratum of strong up-front authentication/authorization is required, and out-of-band solutions make the most sense (e.g. due to heavy-duty crypto processing and database handling). This is vulnerable to the threats identified in out-of-band user-network signaling (see [A]).

B.2) Past this barrier, a user must be trusted to use the lightweight in-band signaling in a sensible way. Therefore, "door-rattling" attacks on the control processor (e.g., by announcing silly burst sizes) are ruled out.

C. Attacks on the data plane (assuming that L3 and above data are already end-to-end authenticated, with integrity, confidentiality, and replay prevention):

C.1) Forging of logical capabilities granting access to lightpaths (hence circumventing signaling)

C.2) Violation of non-TDM sharing rules (e.g., OBS) within a lightpath

10.2 Strengths

When compared to packet switching, the circuit-oriented technologies described in this paper show noteworthy points of strength in security. Chiefly, a circuit is a practical way to limit trust relationships to a small, tractable set of users (e.g., the two peers in a dedicated lightpath, or a small set of peers in an OBS setup).

Conversely, in a packet-switched network a user must trust any and all of its users to "play nice" and execute their end-to-end protocols in the IETF sanctioned terms only. For instance, experimental, faulty, or outright malicious TCP implementations [91] can dramatically alter fairness, often reaching the extreme case of (D)DoS attack. Access capacity, QoS, and policy boundaries are known to lessen this exposure, though in practice these boundaries are soft-boundaries when compared to a circuit's “hard” boundaries. As a case in point, research testbeds can be easy exploitation targets due to the mix of experimentation, high access capacity, and non-commercial-grade QoS/policy stipulations.

The circuit-oriented optical technologies are seen having the following strengths:

I. isolation and non-interference among users
II. compartmentalization in the face of failure or compromise
III. friendly end-to-end protocol experimentation with a limited trust base
IV. traceable and accountable access (no need for firewalls)
V. hitless circuit setup/teardown

Like any other networked solution, the combination of grid applications and optical networks is not risk free. Attackers can resort to three broadly-defined exploitation areas, which apply to the case of applications handling lightpaths directly as well as the case of applications delegating lightpaths handling to an intervening router:

A. out-of-fiber signaling, if any (e.g., the attacker impersonates either a user or a network, with ensuing hijacking of traffic, or downgrading of security defenses leading to further exploitations);
B. in-fiber signaling, if any (e.g., within an OBS setup, the attacker obtains authorization to proceed with "tell and wait", but switches to "tell and go" instead);
C. the data plane and its correlation with in-fiber/out-of-fiber signaling (e.g., the attacker forges capabilities to the data plane, and circumvents the signaling plane altogether).

For these, an attacker compromises one or more elements among application, network services, network elements, or the link through which the in-fiber/out-of-fiber handshakes occur.

Well-known authentication, authorization, and accounting (AAA) techniques, and their correct implementation/operation, provide an effective first line of defense. More defenses are in order (e.g., against attackers tampering with network elements and/or the link).

In scenarios with out-of-fiber signaling, the separation of concerns in signaling vs. data has merits as well as inherent risks. The key strength is that security measures can now be designed to custom fit signaling channel and data channels. That is, the a-priori knowledge of their two different traffic patterns can lead to a security schema with tighter protection. A key risk is that the signaling plane represents a manifest and highly rewarding target to attackers. It is easy to imagine that an intrusion into signaling and control planes can generate catastrophic failures. While optical networks typically use physically isolated networks for the signaling/control functions, it is also the case that researchers are advocating greater and more direct control of the network (with potential vulnerabilities at the testbed level at least). .

10.3 Design options

Out-of-fiber signaling can effectively occur through a legacy IP network. In that case, network-level security (e.g., IPsec [92]) can thwart the attacks falling in the a) realm. [93] describes a possible implementation.

With regard to in-fiber signaling and type c) attacks, rate-control fixtures can force traffic to fit into agreed-upon envelopes. This aptly complements the trust granted to the (small) set of users sharing a lightpath via, say, OBS techniques (e.g., a user can still be faulty).

Some other type c) attacks require that the capability to a lightpath (i.e., the outcome of successful signaling to the network) be closely guarded. In optically-attached systems, the point of ingress to a lightpath is integral part of the TCB, and standard OS security considerations apply. In setups where traffic is groomed on lightpaths one or more hops away (e.g., in a cloud by-pass situation), an attacker can infer that, for instance, VLAN IDs correspond to lightpaths, and sweep the VLAN ID space with spurious traffic until a lightpath is found. These setups can be secured by protecting the access ramps to lightpaths from traffic injection, or using on-the-wire IDs stronger than VLAN IDs.

OVPNs [94] are an emerging solution to increase the granularity of a circuit's capacity (e.g., to scale a circuit in STS-1 increments). Additionally, they can restrict connectivity and isolate domains of addressing/routing. As such, they are a powerful step towards securing these circuit-oriented optical technologies.

When optical resources are exposed as an OGSI-based service, the above-mentioned security techniques can be thought of as operating in the back-end of the service. The front-end of the service should conform to the GGF’s Grid Security Infrastructure, enabling a seamless integration of the optical resource with other resources.

The scoreboard of strengths vs. risks hints that Grid experimentation can proceed on optical networks with a remarkably good security potential, starting with the early research testbeds.

11. Authors Information

1. Dimitra Simeonidou (Editor), Photonic Networks Laboratory, University of Essex,UK, dsimeo@essex.ac.uk
2. Bill St. Arnaud, Canarie, Canada, bill.st.arnaud@canarie.ca
3. Micah Beck, Logistical Computing and Internetworking, Computer Science University of Tennessee, USA, mbeck@cs.utk.edu
4. Peter Clarke, University College London, clarke@hep.ucl.ac.uk
5. Doan B. Hoang, Department of Computer Systems, University of Technology, Sydney, dhoang@it.uts.edu.au
6. David Hutchison, Lancaster University, d.hutchison@lancaster.ac.uk
7. Gigi Karmous-Edwards, Advanced Network Research, MCNC Research and Development Institute, USA, gigi@anr.mcnc.org
8. Tal Lavian, Nortel Networks Labs, tlavian@nortelnetworks.com
9. Jason Leigh, Electronic Visualization Lab, University of Illinois at Chicago, spiff@evl.uic.edu
10. Joe Mambretti, International Center for Advanced Internet Research, Northwestern University, Illinois,USA, j-mambretti@northwestern.edu
11. Reza Nejabati, Photonic Networks Laboratory, University of Essex, ,UK, rnejab@essex.ac.uk
12. Volker Sander, Research Centre Jülich, Germany, v.sander@fz-juelich.de
13. John Strand, AT&T Transport Network Evolution Dept, jls@research.att.com
14. Franco Travostino, Nortel Networks Labs, travos@nortelnetworks.com

12. Intellectual Property Statement

The GGF takes no position regarding the validity or scope of any intellectual property or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; neither does it represent that it has made any effort to identify any such rights. Copies of claims of rights made available for publication and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the GGF Secretariat. The GGF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights which may cover technology that may be required to practice this recommendation. Please address the information to the GGF Executive Director (see contacts information at GGF website).

13. Full Copyright Notice

12. References

[1] Information about the Large Hadron Collider at CERN: lhc-new-homepage.web.cern.ch, Information about the BarBar experiment: www.slac.stanford.edu/BFROOT/

[2]Harvey B. Newman, Mark H. Ellisman, and John A. Orcutt, “Data Intensive E-Science Frontier Research,” Special Issue, Communications of the ACM, “Blueprint for the Future of High Performance Networking,” Nov. 2003, Vol. 46, No. 11, pp. 68-77.

[3] World Economic Forum, New York 2001, Digital Divide Report

[4] Telegeography Inc, Terrestrial bandwidth 2002

[5]“The GRID2, Blueprint for a New Computing Infrastructure”, 2nd Edition, Ian Foster and Carl Kesselman, Eds, Morgan Kaufmann Publishers, Elsevier Press, 2004

[6] Tom DeFanti, Cees de Laat, Joe Mambretti, Kees Neggars, Bill St. Arnaud, “TransLight, A Global-Scale LambdaGrid for E-Science, Special Issue, Communications of the ACM, “Blueprint for the Future of High Performance Networking,” Nov. 2003, Vol. 46, No. 11, pp. 43-41.

[7]"Revolutionizing Science And Engineering Through Cyberinfrastructure", Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, January 2003, http://www.cise.nsf.gov/evnt/reports/atkins_annc_020303.htm**“NSF CISE Grand Challenges in e-Science Workshop Report,” Report to the National Science Foundation, Directorate for Computer and Information Science and Engineering (CISE), Advanced Networking Infrastructure & Research Division, T. DeFanti, M. Brown et al, Jan. 2002.**

[8] "Research Networking in Europe - Striving for global leadership", European Commission, 15 sep 2002, http://www.cordis.lu/ist/rn/rn-brochure.htm

[9] http://lhc-new-homepage.web.cern.ch

[10] http://www.evlbi.org/network/network.http , http : // www.jb.man.ac.uk / vlbi / gallery / radtel.html

[11] http://www.hpcx.ac.uk/, http://www.csar.cfs.ac.uk/

[12] www.teragrid.org

[13] http://www.sc-conference.org/sc2003/
[14] http://scinet.supercomp.org/2003/bwc/results/index.html

[15] Information about CERN, The CERN Grid Deployment group: http://it-div-gd.web.cern.ch/it-div-gd/

[16] Ralph Spencer, Steve Parsley and Richard Hughes-Jones "The resilience of e-VLBI data to packet loss", 2nd eVLBI workshop, 15-16 May 2003, Netherlands

[17] Allcock, W., A. Chervenak, I. Foster, C. Kesselman, C. Salisbury, and S. Tuecke, "The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Datasets", Network and Computer Applications, 2002.

[18] David Levine, "Grid Computing for the Online Video Game Industry ", GlobusWorld January 14, 2003

[19] Foster, I., Kesselman, C. and Tuecke, S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. International Journal of High Performance Computing Applications, 15 (3). 200-222. 2001. www.globus.org/research/papers/anatomy.pdf.

[20] www.canarie.ca

[21] M. J. O'Mahony, D. Simeonidou, D. K. Hunter, A. Tzanakaki, " The application of optical packet switching in future communication networks", IEEE Communications Magazine, pp. 128-135, March'01

[22] J. He, D. Simeonidou, "Flow routing and its performance analysis in optical IP networks", Photonic Network Communications, Vol 3, pp 49-62 (Special issue for IP over WDM), 2001

[23] Foster, I., Fidler, M., Roy, A., V, Sander, Winkler, L., "End-to-End Quality of Service for High-end Applications." Elsevier Computer Communications Journal, , 2004

[24] Roy, A. and Sander, V., "GARA: A Uniform Quality of Service Architecture", Resource Management: State of the Art and Future Trends, Kluwer Academic Publishers, pp. 135-144, 2003, Edited by Jarek Nabrzyski, Jennifer M. Schopf and Jan Weglarz

[25] J. S. Turner. Terabit burst switching. Journal of High Speed Networks, 8(1): 3{16, January 1999

[26] C. Qiao and M. Yoo. Choices, features, and issues in optical burst switching. Optical Networks, 1(2): 36{44, April 2000.

[27] J. Y. Wei and R. I. McFarland. Just-in-time signaling for WDM optical burst switching networks. Journal of Lightwave Technology, 18(12):2019{2037, December 2000.

[28] Y. Xiong, M. Vandenhoute, and H.C. Cankaya. Control architecture in optical burst-switched WDM networks. IEEE Journal on Selected Areas in Communications, 18(10):1838{1851, October 2000.

[29] C. Qiao and M. Yoo. Optical burst switching (OBS)-A new paradigm for an optical Internet. Journal of High Speed Networks, 8(1):69{84, January 1999.

[30] Baldine, G. N. Rouskas, H. G. Perros, and D. Stevenson. JumpStart: A just-in-time signaling architecture for WDM burst-switched networks. IEEE Communications, 40(2):82{89, February 2002.

[31] Kozlovski E., M. Duser, I. De Miguel, P. Bayvel, “Analysis of burst scheduling for dynamic wavelength assignment in optical burst-switched networks”, IEEE, Proc. LEOS’01, vol. 1, 2001.

[32] Dolzer K. “Assured horizon - an efficient framework for service differentiation in optical burst switched networks.” Proc. SPIE OptiComm. 2002. 1 page. (Proc. SPIE Vol 4874.)

[33] http://www.ind.uni-stuttgart.de/~gauger/BurstSwitching.html#Publications

[34] http://www.utdallas.edu/~vinod/obs.html

[35] http://www.cs.buffalo.edu/~yangchen/OBS_Pub_year.html

[36] Dimitri Papadimitriou, Denis Penninckx, “ Physical Routing Impairments in Wavelength-switched Optical networks”, Business Briefing: Global Optical Communications, 2002.

[37] John Strand, Angela Chiu and Robert Tkach, “Issues for Routing in the Optical Layer,” IEEE Communications Magazine, February 2001.

[38] Daniel Blumenthal, “Performance Monitoring in Photonic Transport Networks”, Bussiness Breifing: Global Photonics Applications and Technology 2000.

[39] Byrav Ramamurthy, Debasish Datta, Helena Feng, Jonathan P. Heritage, Biswanath Mukherjee, “Impact of Transmission Impairments on the Teletraffic Performance of Wavelength-Routed Optical networks”, IEEE/OSA Journal of Lightwave Technology Oct ’99.

[40] G. N. Rouskas, “Optical Layer Multicast: Rationale, Building Blocks, and Challenges,” IEEE Network, Jan/Feb 2003, pp. 60-65.

[41] M. Ali, J. Deogun, “Power-efficient Design of Multicast Wavelength-Routed Networks”, IEEE JSAC, vol. 18, no. 10, 2000, pp. 1852-1862.

[42] M. Ali, J. Deogun, “Allocation of Splitting Nodes in Wavelength-routed Networks,” Photonic Net. Comm., vol. 2, no. 3, Aug. 2000, pp. 245-263.

[43] Leigh et al. An Experimental OptIPuter Architecture for Data-Intensive Collaborative Visualization, the 3rd Workshop on Advanced Collaborative Environments (in conjunction with the High Performance Distributed Computing Conference), Seattle, Washington, June 22, 2003 http://www- unix.mcs.anl.gov/fl/events/wace2003/index.html

[44] R. Dutta, G. N. Rouskas, “A Survey of Virtual Topology Design Algorithms for Wavelength Routed Optical Networks,” Opt. Net., vol. 1, no. 1, Jan 2000, pp.73-89.

[45] G. N. Rouskas, “Light-Tree Routing Under Optical Layer Power Budget Constraints,” Proc. 17th IEEE Comp. Comm. Wksp., Oct. 14-16, 2002.

[46] X.H. Jia et al., “Optimization of Wavelength Assignment for QoS Milticast in WDM Networks,” IEEE Trans. Comm., vol. 49, no. 2, Feb. 2001, pp. 341-350.

[47] G. Sahin, M. Azizoglu, “Milticast Routing and Wavelength Assignment in Wide-Area Networks.” Proc. SPIE, vol. 3531, Nov. 1998, pp. 196-208.

[48] A. E. Kamal, A. K. Al-Yatama, “Blocking Probabilities in Circuit-switched Wavelength Division Multiplexing Networks Under Multicast Service,” Perf. Eval., vol. 47, no. 1, 2000, pp.43-71.

[49] S. Ramesh, G. N. Rouskas, H. G. Perros, “Computing Call Blocking Probabilities in Multi-class Wavelength Routing Networks with Multicast Traffic,” IEEE JSAC, vol. 20, no.1, Jan. 2002, pp. 89-96.

[50] K.D., Wu, J. C., Wu, C.S. Yang, “Multicast Routing with Power Consideration in Sparce Splitting WDM Networks,” Proc. IEEE ICC, 2001, pp. 513-517.

[51] X. Zhang, J. Y. Wei, C. Qiao, “Constrained Multicast Routing in WDM Networks with Sparce Light Splitting,” J. Lightwave Tech., vol. 18, no. 12, Dec. 2000, pp. 1917-1927.

[52] Savage, N. “Linking with Light”, IEEE Spectrum, pp. 32-36, Aug, 2002.

[53] R. Callon, et al. 1999. A Framework for Multiprotocol Label Switching. ID: draft-ietf-mpls-framework-03.txt

[54] E. Rosen, et al. 1999. Multiprotocol Label Switching Architecture. ID: draft-ietf-mpls-arch-05.txt

[55] E. Rosen, A. Viswanaathan, R. Callon, “Multiprotocol Label Switching Architecture,” IETF RFC - 3031, January 2001.

[56] T. Nadeau, C. Srinivasan, A. Farrel "Multiprotocol Label Switching (MPLS) Management Overview", July 23, 2003

[57] D. Awduche and Y. Rekhter, “Multi-Protocol Lambda Switching: Combining MPLS Traffic Engineering Control with Optical Crossconnects,” IEEE Communications Magazine, March 2001, pp. 111-116.

[58] B. Rajagopalan, J. Luciani, D. Awduche, “IP over Optical Networks: A Framework,” draft-ietf-ipo-framework-03.txt

[59] E. Mannie, et al, GMPLS Extensions for SONET and SDH Control draft-ietf-ccamp-gmpls-sonet-sdh-01.txt

[60] E. Mannie GMPLS Signaling Extension to Control the Conversion between Contiguous and Virtual Concatenation for SONET and SDH draft-mannie-ccamp-gmpls-concatenation-conversion-00.txt

[61] E. Mannie, et al, Generalized Multi-Protocol Label Switching (GMPLS) Architecture draft-ietf-ccamp-gmpls-architecture-00.txt

[62] A. Bellato G.709 Optical Transport Networks GMPLS Control Framework draft-bellato-ccamp-g709-framework-00.txt

[63] A. Bellato GMPLS Signaling Extensions for G.709 Optical Transport Networks Control draft-fontana-ccamp-gmpls-g709-00.txt

[64] O. Aboul-Magd A Framework for Generalized Multi-Protocol Label Switching (GMPLS) draft-many-ccamp-gmpls-framework-00.txt

[65] G.8080/Y.1304, Architecture for the Automatically Switched Optical Network (ASON), ITU-T

[66] B. Davie, P. Doolan, and Y. Rekhter. 1998. Switching in IP Networks: IP Switching, Tag Switching, and Related Technologies. The Morgan Kaufmann Series in Networking. New York: Academic Press.

[67] J. Lang, et al, Generalized MPLS Recovery Mechanisms draft-lang-ccamp-recovery-01.txt, draft-mannie-ccamp-gmpls-concatenation-conversion-00.txt

[68] J.P. Lang, et al., “Link Management Protocol,” IETF Internet Draft, draft-ietf-ccamp-lmp-wdm-03.txt

[69] RFC 3473, L. Berger (ed.), Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions , draft-ietf-ccamp-gmpls-**rsvp-te**-ason-01.txt

[70] J. L. Strand, A. L. Chiu, “Control Plane Considerations for All-Optical and Multi-Domain Optical Networks and Their Status in OIF and IETF”, Optical Networks Magazine, Vol. 4 No. 1 (January/February 2003), pp. 26-35.

[71] Optical Interworking Forum, “User Network Interface (UNI) 1.0 Signaling Specification”, http://www.oiforum.com/public/documents/OIF-UNI-01.0.pdf

[72] John W. Stewart III, “BGP4: Inter-Domain Routing in the Internet”, Addison-Wesley, 1999.

[73] J. L. Strand, A. L. Chiu, “Control Plane Considerations for All-Optical and Multi-Domain Optical Networks and Their Status in OIF and IETF”, Optical Networks Magazine, Vol. 4 No. 1 (January/February 2003), pp. 26-35.

[74] J. L. Strand, J.; A. L. Chiu, , R. Tkach, . “Issues For Routing In The Optical Layer”, IEEE Communications Magazine, 2/2001, vol. 39, no. 2, pp. 81 –87

[75]“Optical BGP networks”, Canarie OBGP, Internet draft: http://obgp.canet4.net/

[76] Y. Rekhter "A Border Gateway Protocol 4 (BGP-4)", IETF March, 2003 The Border Gateway Protocol, IETF [RFC1518, RFC1519].

[77] M. Blanchet, F. Parent, B. St Arnaud “Optical BGP (OBGP): InterAS lightpath provisioning draft-parent-obgp-01.txt March 2001

[78] I. Foster, C. Kesselman, S. Tuecke. , “The Anatomy of the Grid: Enabling Scalable Virtual Organizations”. *International J. Supercomputer Applications*, 15(3), 2001.

[79] ITU-T SG13 Draft Recommendation "G.astn; Architecture for the Automatic Switched Transport Network", November 2000.

[80] Foster, I., Kesselman, C., Nick, J., Tuecke, S., “The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration,” Open Grid Service Infrastructure WG, Global Grid Forum, June 22, 2002

[81] Foster, I., Fidler, M., Roy, A., V, Sander, Winkler, L., “End-to-End Quality of Service for High-end Applications.” *Computer Communications, Special Issue on Network Support for Grid Computing,* 2002

[82] Czajkowski, K., Foster, I., Karonis, N., Kesselman**,** C., Martin, S., Smith, W, and Tuecke, S., A “A Resource Management Architecture for metacomputing systems**,**” *In the 4th Workshop on Job Scheduling Strategies for Parallel Processing, pp. 62-82. Springer-Verlag LNCS 1459, 1988.*

[83]Czajkowski, K., Dan, A., Rofrano, J., Tuecke, S., and Xu, M., “Agreement-based Grid Service Management (WS-Agreement),” GWD-R draft-ggf-graap-agrement-1. December 2003.

[84] John Strand, Angela Chiu and Robert Tkach, "Issues for Routing in the Optical Layer," IEEE Communications Magazine, February 2001.

[85] Dimitri Papadimitriou, Denis Penninckx, “ Physical Routing Impairments in Wavelength-switched Optical networks”, Business Briefing: Global Optical Communications, 2002.

[86] Angela Chiu and John Strand, "Control Plane Considerations for All-Optical and Multi-Domain Optical Networks and Their Status in OIF and IETF," to appear in Optical Networks Magazine.

[87] John strand, Angela Chui, draft-ietf-ipo-impairments-03.txt

[88] John Strand, Angela Chiu and Robert Tkach, "Issues for Routing in the Optical Layer," IEEE Communications Magazine, February 2001.

[89] Daniel Blumenthal, “Performance Monitoring in Photonic Transport Networks”, Bussiness Breifing: Global Photonics Applications and Technology 2000.

[90] Byrav Ramamurthy, Debasish Datta, Helena Feng, Jonathan P. Heritage, Biswanath Mukherjee, “Impact of Transmission Impairments on the Teletraffic Performance of Wavelength-Routed Optical networks”, IEEE/OSA Journal of Lightwave Technology Oct ’99.

[91] Stefan Savage, Neal Cardwell, David Wetherall and Tom Anderson, TCP Congestion Control with a Misbehaving Receiver, ACM Computer Communications Review, 29(5):71-78, October 1999.

[92] RFC 2401, The Internet Engineering Task Force

[93] The Optical Internetworking Forum, Security Extensions for UNI and NNI, http://www.oiforum.com/public/documents/Security-IA.pdf

[94] Service Requirements for Optical Virtual Private Networks, Internet Draft, http://www.ietf.org/internet-drafts/draft-ouldbrahim-ppvpn-ovpn-requirements-01.txt