Grid-Controlled Lightpaths for High Performance Grid Applications *

Raouf Boutaba, Wojciech Golab, Youssef Iraqi, Tianshu Li
({rboutaba, wgolab, iraqi, dtianshu}@bbcr.uwaterloo.ca)
University of Waterloo, Canada

Bill St. Arnaud
(bill.st.arnaud@canarie.ca)
CANARIE Inc., Canada

Abstract. Grid applications call for high performance networking support. One attractive solution is to deploy Grids over optical networks. However, resource management in optical domains is traditionally very rigid and cannot successfully meet the requirements of Grid applications, such as flexible provisioning and configuration. In this paper, we present a customizable resource management solution for optical networks where users can create lightpaths on demand and manage their own network resources. Thanks to a Grid-centric system architecture, lightpath resources can be shared among users and easily integrated with data and computation Grids.

Keywords: network support for Grids, high-performance networking, customer-controlled networks, optical networks, lightpath management

1. Introduction

Network support is a critical aspect of Grid environments. However, the best-effort delivery system of the Internet is severely restricting the deployment of Grids on wide-area scales. Concerned primarily with connectivity and fair sharing of bandwidth, the Internet is a vehicle too slow and unreliable for the masses of data being generated in emerging e-science applications. Take as an example the ATLAS particle collider under construction at CERN, which is expected to generate petabytes of data per year, or the LOFAR radio telescope in the Netherlands, which will produce data at an aggregate rate in the range of tens of terabits per second. Collection and storage of data from these experiments calls for a data Grid supported by a high-performance network.

Another concern in deploying Grids over the Internet is the cost of transmitting data [4]. Although projects like SETI@Home have already tapped into the power of idle desktop PCs for parallel number crunching, their success is due to a high ratio of computational demand to I/O demand (one day of computation per megabyte of input).

* This research is partially supported by CANARIE Inc. and Cisco Canada Inc.

In contrast, data scanning applications such as the Sloan Digital Sky Survey, which detects stars and galaxies, are more data intensive and do not share the same economic benefits. A network service that provides high capacity at low cost can open exciting opportunities in parallel computation, regardless of whether they involve idle desktops or geographically distributed supercomputers.

A promising solution to the communication needs of Grids is optical networking. In particular, optical circuit switching technologies can be used to provision bandwidth-guaranteed pipes that in addition feature minimal latency. The classic example of this is a dedicated wavelength, also known as a lightpath, carrying an optical signal with a capacity of 10 Gb/s. One can also think of sub-wavelength SONET circuits and ATM Constant Bit Rate circuits as lightpaths, since they provide the same QoS benefits and are typically provisioned over optical hardware.

In addition to superior performance compared to dedicated copper line or wireless technologies, lightpaths are economically attractive due to the convergence of two trends. One of these is Wavelength Division Multiplexing (WDM) technology, which involves the concurrent transmission of multiple optical signals over a common fibre. This makes possible a ten or even hundredfold increase in aggregate transmission capacity by converting legacy single-wavelength systems without having to install new fibre. The additional optical multiplexing hardware needed is based on passive optical components, and represents a disproportionately smaller increase in cost. The main economic concern becomes the cost of the switching elements, also referred to as lightpath cross-connect devices. However, with the emergence of micro-electro-mechanical systems (MEMS) technology, it is possible to perform all-optical switching at wavelength granularity and avoid the costs associated with high-speed electronic components.

The second important trend is the huge decrease in the cost of dark fibre that occurred around the year 2000, after a period of massive overprovisioning of optical infrastructure. Dark fibre is pre-installed optical fibre that is sold to customers, who light it up with their own terminating equipment [6]. It is an affordable alternative to the managed services offered by conventional network providers, that in addition allows customers to control their resources. For example, customers can interconnect directly to other customers in the region, using condominium-style shared equipment between their management domains. Similarly, customers can extend their infrastructure to a peering point where they can connect directly to a wide area network and avoid the ongoing fees charged by intermediate network providers.

The combination of WDM technology and dark fibre makes it possible for network customers such as research institutions to build and operate their own high-performance networks. Thus, the hardware technology needed to support data-intensive Grids is in place. One final problem that remains is the lack of suitable management and control software to enable dynamic provisioning and sharing of lightpaths in support of geographically distributed virtual organizations. Current optical network management systems are geared toward conventional network providers who offer managed services and do not allow customers to control their own resources. For example, while customers are able to lease lightpaths, they have no flexibility in configuring, i.e. partitioning and composing, these resources because they have no control over the appropriate lightpath cross-connect devices. Since collaboration and resource sharing are the foundation of the Grid concept, a new management technology is needed that allows users to acquire resources on demand, control interconnections among lightpaths, and share unused bandwidth in a flexible and collaborative fashion.

In this paper, we present a solution to the problem of managing next-generation customer-owned optical networks. Through virtualization of hardware devices distributed across multiple management domains, our system empowers customers to both own and control lightpaths. Our implementation is compliant with open Grid Services standards, which build on Web Services to support the management functionality needed in computation and data Grids. This feature facilitates integration of our system with emerging Grid infrastructures, and takes advantage of the universality of XML for interoperability in a global community of network users.

This paper is organized as follows. In Section 2 we present the architecture of our User-Controlled Lightpath Management System. Section 3 then explores how the system can be used by Grid applications. We conclude the paper with a discussion of possible extensions to the system and research directions.

2. A Grid-Centric Architecture for Lightpath Management

In this section we describe the architecture of our User-Controlled Lightpath Management System. We begin with a high-level overview in Section 2.1 and go on to examine the Service Provisioning Layer in detail in Section 2.2.

2.1. High-Level Overview

The architecture of our User-Controlled Lightpath Management System consists of three layers, as shown in Figure 1.

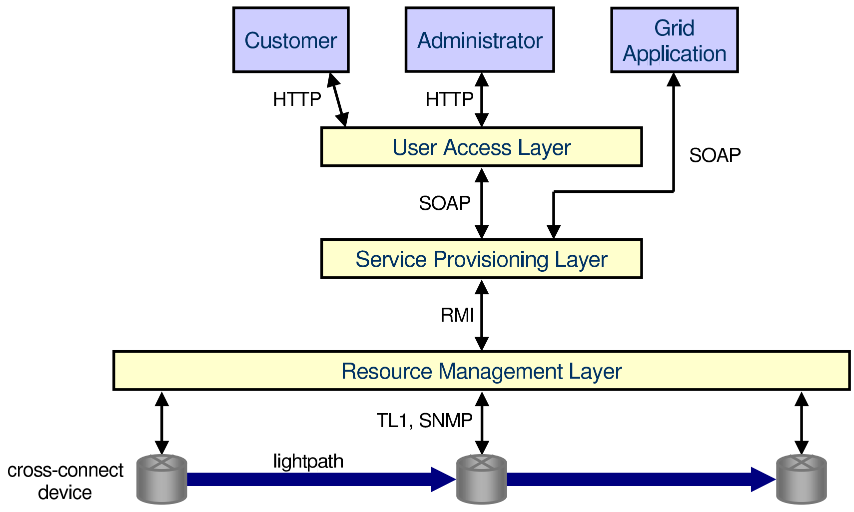

Figure 1. High-level architecture of the lightpath management system.

The User Access Layer (UAL) exposes an HTML interface to the human user. It is concerned with translating user requests into operations on the services provided by the Service Provisioning Layer (SPL). The implementation is based on the Java Web Services Developer Pack (Java WSDP) from Sun Microsystems [7]. This suite of tools includes Apache Tomcat, which is the official reference implementation for Java Servlet and JavaServer Pages technologies, as well as an implementation of the Java API for XML Messaging (JAXM), which is used to construct and parse SOAP messages exchanged with the SPL.

The Service Provisioning Layer (SPL) is concerned with hosting a set of Grid services that carry out the high-level operations defined in the system. The Grid hosting environment consists of Globus Toolkit 3 (GT3) in conjunction with the JBoss application server, where the service implementations are deployed as Enterprise JavaBeans (EJBs). GT3 provides a Grid platform that inherits the benefits of the XML-based Web Services framework [10]. The most important of these is interoperability, which arises from the application-independent and platform-independent nature of XML. This is a critical feature of the system, which is intended to be used by a wide variety of users, dispersed across multiple institutions and possibly across multiple nations.

The SPL maintains information concerning users and lightpaths in a relational database. Three classes of users are defined: ordinary users, who are able to perform basic operations on lightpaths; domain administrators, who in addition can create accounts for ordinary users and can register bandwidth resources in the system that are available in their respective management domains; and system administrators, who have access to the full functionality of the system. Lightpath Objects (LPOs), which represent lightpaths, can be categorized according to how they arise in the system. Root LPOs are manually registered by domain administrators, as mentioned above. They represent unused bandwidth between a pair of physically adjacent cross-connect devices. All other LPOs are created through partitioning and concatenation operations. These operations will be explained in detail in the next section.

Finally, the Resource Management Layer (RML) serves as an interface to the switching elements in the network, i.e. the lightpath cross-connect devices. It comprises a set of Resource Agents, each associated with exactly one lightpath cross-connect device. In the case of our testbed network, CA*net4 [1], each such device is a Cisco ONS 15454 SONET Multiservice Provisioning Platform. The high-level role of the Resource Agent is to enable shared customer control over the hardware device, while hiding technology-specific details such as how the endpoints are addressed (e.g. slot / port / channel in SONET), or what protocols are used for configuration and performance monitoring.

The Resource Agent includes a programmable component that allows users to customize the system by installing binary Java code. In our prototype implementation, this feature allows users to realize custom performance monitoring policies by installing modules that collect and filter performance data. The benefit of this is a reduction in management traffic, which follows from the ability to make intelligent filtering decisions at a location close to the network hardware. The programmable component is supported by an LPO space, which stores all LPOs corresponding to lightpaths that originate at the associated cross-connect device. This way, user code can retrieve the set of LPOs owned by the corresponding user without having to contact the SPL.

2.2. Grid-based Service Provisioning Layer

2.2.1. *Grid Services*

A Grid-centric Service Provisioning Layer is a natural choice for a system that is meant to provide network support for Grid computing. In essence, the Service Provisioning Layer in our system can be viewed as a service Grid that can be easily integrated with other data and computation Grids. For example, a user can acquire some shared storage space from another Grid and set up an end-to-end lightpath to transfer data to the storage space in a fast and secure fashion.

In our Service Provisioning Layer, lightpath management logic is deployed in the form of Grid Services, which encapsulate the resource discovery, allocation, and access functionalities, providing a simple and efficient service interface to the end users. All the services adhere to open Grid standards and service discovery can be achieved by querying the Service Data Description.

Specifically, the following services hosted by the Service Provisioning Layer empower customers to own and control lightpaths.

- CreateRootLPO: One of the most fundamental services is root LPO creation, which was mentioned in the last section. This service allows resources from various management domains to be enlisted in the system by the appropriate administrative users.

- AdvertiseLPO and LeaseLPO: Lightpath resources are advertised for lease to other users using the AdvertiseLPO service, and the ownership of lightpaths can be transferred among users via the LeaseLPO service.

- PartitionLPO and ConcatenateLPO: These two services provide lightpath partitioning and composition functionality. The former allows the bandwidth of a lightpath to be divided among multiple child lightpaths. The latter enables formation of longer lightpaths by cross-connecting a series of shorter lightpaths of uniform bandwidth.

- AccessLPO: The access service is used to prepare a lightpath for routing data traffic, for example by cross-connecting it to Ethernet LANs at each end.

- ReconfigureLPO: The reconfiguration service allows users to specify policies concerning how lightpaths can be accessed. For example, to prevent access to private LANs, access options for a lightpath can be restricted (e.g. the set of Ethernet ports that can be used to access the lightpath can be reconfigured) before it is advertised to others.

- EstablishEndToEndLPO: Finally, the end-to-end lightpath establishment service builds on many of the others by first identifying the resources necessary to form a lightpath of a given bandwidth between a particular pair of endpoints, then reserving these resources, as well as partitioning and concatenating them as necessary. This service is the main service exposed to Grid applications and end users.

2.2.2. *Service Implementation: Data and Computation*

The Service Provisioning Layer tracks the states of all lightpaths managed by the system. As mentioned in the last section, lightpath information is represented as a set of Lightpath Objects (LPOs). The fields of an LPO include the following: a unique ID; the IDs of the current and previous owner; advertisement and lease expiry dates; a status indicating whether the LPO is reserved for use by the owner or has been advertised, partitioned, or concatenated; bandwidth in kbps; as well as hardware parameters specifying each endpoint, e.g. slot / port / SONET channel.

The set of LPOs stored in the SPL is used in path computations during the end-to-end lightpath establishment operation. Specifically, a logical topology is formed using the set of lightpaths that are accessible to the calling user, and are eligible given the bandwidth and duration requested by the user. A path computation is then performed on this topology to identify a chain of constituent lightpaths that can be concatenated to yield the desired end-to-end path. The system supports multiple routing engines for flexibility in this respect.

Our prototype implementation includes a shortest path routing engine based on Dijkstra's algorithm, that supports multiple routing metrics. The default metric defines the weight of a lightpath in terms of the hop count, namely the number of cross-connect devices traversed less one. The amount of bandwidth in excess of the user's requirement is also incorporated into the metric, with a lesser weight, in order to break ties among lightpaths having equal hop counts. Additional metrics can be constructed that favour end-to-end paths that can be sustained for a longer period of time, or that prioritize the user's own lightpaths over those advertised by others.

2.2.3. *Security*

Security in our lightpath management system takes advantage of the Grid Security Infrastructure (GSI) provided by Globus Toolkit 3.0 [9, 3]. In the prototype lightpath management system, we implemented several security mechanisms as shown in Figure 2 to ensure secure communication, authentication, and authorization between the UAL / applications and Service Provisioning Layer.

GSI is based on public key encryption, X.509 certificates, and the Secure Sockets Layer (SSL) communication protocol [3]. Mutual Au-

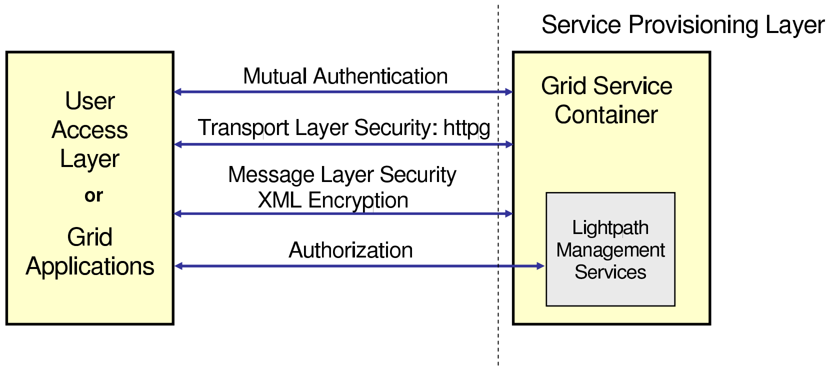

Figure 2. Security Mechanisms in the lightpath management system

thentication is provided by Globus Toolkit based on public key encryption and Secure Socket Layer (SSL). GSI also provides a transport level security mechanism, namely the httpg protocol, to secure the communication channel, as well as a message level security mechanism, namely XML encryption and XML signatures, to secure the messages themselves.

However, these security mechanisms cannot deal with user privileges in our lightpath management system. In other words, a user might be a member of the Grid but still should have no access to lightpath resources if the user is not a member of our system. In this case, the user may still have access to the service registry but will not be able to invoke an unauthorized operation on a lightpath management service instance. In our prototype system, we implemented a separate authorization mechanism for fine-grained access control.

Note that only the authorization and membership management functionality is customized, while other mechanisms are already provided by GSI. This rich set of built-in security mechanisms is also one of the main reasons why a Grid-centric approach was chosen for our Service Provisioning Layer.

2.3. Fault Tolerance

Ensuring consistency between the state maintained by the Service Provisioning Layer and the configuration of the network hardware is a key concern in the design of the system. However, achieving this is a challenge due to a variety of fault conditions that can occur under normal operation of the software. Naturally, the distributed architecture of the system gives rise to the possibility of faults due to loss of network connectivity among the three layers of the system, or even between a Resource Agent and a lightpath cross-connect device. In addition, resource contention conditions can arise. For example, two users may issue simultaneous requests to cross-connect their lightpaths to the same Ethernet port during the Access LPO operation. Similarly, concurrent attempts to lease the same lightpath can occur.

The design of the system must take into account the variety of possible fault scenarios in order to ensure transactional behaviour of the supported high-level operations. In particular, the interaction of the Service Provisioning Layer with the distributed Resource Management Layer is of critical importance due to the signalling complexity. To this end, we have implemented a two-phase commit (2PC) protocol between the SPL and RML in support of atomic execution of operations that involve multiple Resource Agents.

The roles of the SPL and RML in transaction management are as follows. The SPL acts as a transaction coordinator. Here we build on the existing container-managed transaction mechanism of the JBoss application server, which protects the consistency of the main database, by communicating transaction states to the appropriate Resource Agents. The agents, in turn, assist in transaction isolation by reserving hardware resources. In addition, an LPO locking scheme is used to preserve the consistency of the LPO space. Lock requests in the Resource Agent are subject to a short timeout in order to prevent deadlock.

Hardware resource reservation in the RML is based on a simple hard reservation scheme. For example, in the Access LPO operation, the lightpath is cross-connected to Ethernet ports in the prepare phase, and the cross-connections are torn down in the abort phase. In the inverse operation, the Ethernet ports are not released until the commit phase. This strategy provides a simple method of resolving resource contention conditions, while not requiring transactional operation of the protocol used to control the lightpath cross-connect device.

3. Interaction with Grid Applications

3.1. Lightpath-aware GridFTP

As mentioned before, Grid applications normally require high performance networking support (e.g. high throughput), which makes our lightpath management system a superior enabling technology for Grid computing compared to the Internet. Because the services deployed in our system are compliant with Grid service standards, integrating our lightpath management system with Grid applications becomes natural.

In this section, we study a case scenario where a file is transferred using standard GridFTP service but the data channel is established over an end-to-end lightpath constructed using our lightpath management system. In the following sections, we start by describing the case scenario briefly and explaining the challenges faced during the design phase. We then examine the possible solutions and propose our lightpath-aware GridFTP solution.

3.1.1. *GridFTP and RFT*

GridFTP and RFT are two of the main data management services provided by Globus [5]. GridFTP is built on top of the standard FTP protocol with the following main additional features to facilitate transferring files in a Grid environment:

- Grid Security Infrastructure (GSI)
- Multiple Data Channels (for parallel file transfer)
- Third-Party Transfer (server-to-server)

Reliable File Transfer (RFT) is essentially a reliable data transfer service that is based on the GridFTP Third-Party Transfer, but with additional monitoring and control features in order to support sophisticated failure recovery during file transfer.

Since client-server GridFTP can be considered as a special case of Third Party Transfer, we will restrict our discussion to the Third-Party Transfer scenario only. To perform a file transfer between two servers A and B, two control channels are first established (i.e. client–server A and client–server B). Control messages are exchanged through these two control channels. Prior to a file transfer operation, an additional data channel between the two servers is established through GridFTP commands. In a normal GridFTP Third-Party Transfer, the data channel is established over the Internet, and hence the performance of the file transfer cannot be guaranteed due to the best-effort nature of the Internet. Here, our objective is to establish the data channel over an end-to-end lightpath created through our lightpath management system, and immensely improve the performance of the GridFTP and RFT services. Figure 3 illustrates this scenario where the data channel shown utilizes the end-to-end lightpath in the optical domain instead of the Internet.

This lightpath-aware GridFTP scenario is an example of how Grid applications can be integrated with our lightpath management system. As mentioned earlier, since our lightpath management services are Grid

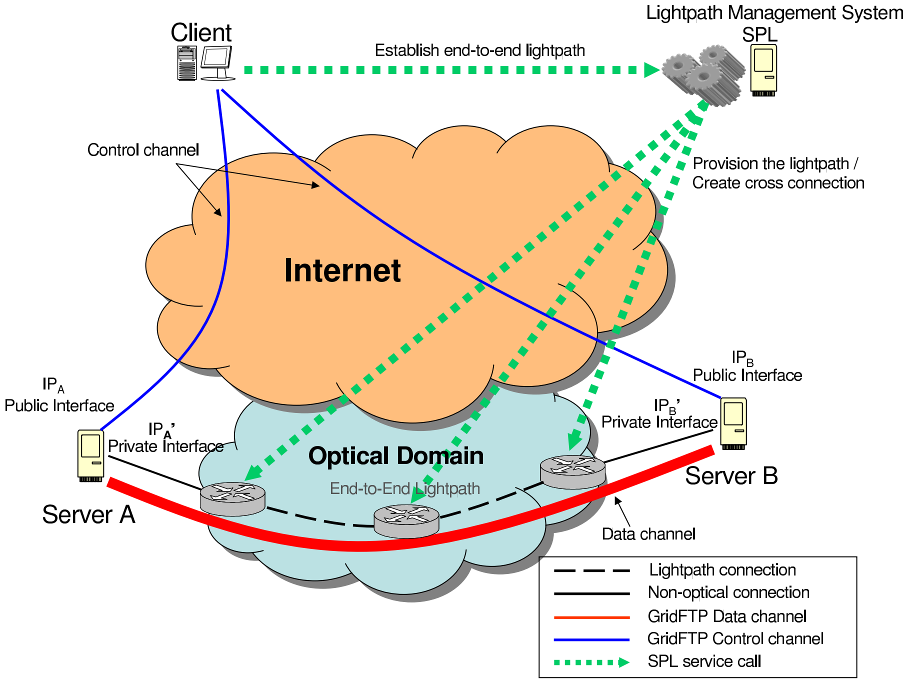

Figure 3. Lightpath-aware GridFTP third-party transfer scenario

services themselves, they can be easily integrated with other Grid services such as data and computation services to provide a rich combination to the users and meet various application requirements. This case scenario was demonstrated during the 2003 CANARIE Advanced Networks Workshop [2]. The lightpath-aware GridFTP client was launched from a laptop at the workshop in Montreal, Canada, and used to establish a lightpath between two Cisco ONS 15454 switches located at a CANARIE lab in Ottawa and perform a third-party GridFTP file transfer between two servers. The SPL of the Lightpath Management System was deployed at the University of Waterloo.

During the demonstration, we compared the performance of an STS-3 (155 Mb/s raw rate) SONET lightpath with a lightly loaded 10 Mb/s Ethernet LAN for transferring a 260 MB file. Using the lightpath, we were able to complete the file transfer in approximately fifteen to twenty seconds, while more than five minutes were typically needed using the 10 Mb/s LAN. Real Internet transfers are expected to be slower than the 10 Mb/s LAN transfer used during the demonstration. For example, using the netperf utility we determined that the throughput between our lab at the University of Waterloo and the CANARIE lab in Ottawa is approximately 5 Mb/s.

3.2. Automatic Virtual Network Creation

Whereas the last section discussed the dynamic creation of end-to-end lightpaths in support of individual file transfers, the basic functionality offered by the system can also be used to create entire virtual networks. In particular, lightpaths serve as an excellent basis for virtual private networks by providing dedicated bandwidth that is isolated from the rest of the network. This property implies greater physical security, as well as protection against congestion hot spots and distributed denial of service attacks that menace the public Internet.

Creating a virtual network using our lightpath management system involves constructing and accessing a set of end-to-end lightpaths. For example, this might be done in preparation for transferring data from multiple geographically distributed sources in a radio astronomy experiment to a remote central processing station. The relevant operations can be conveniently scripted inside a custom Grid application that builds the virtual network incrementally by executing a sequence of end-to-end lightpath establishment and lightpath access operations. In order to achieve better resource utilization, one can also consider a joint computation of the relevant paths, as done in [8] using a multi-commodity flow formulation. To this end, users are free to implement the optimization scheme of their choice, building on the lease, partitioning, and concatenation services offered by the system.

The use of a custom Grid application to configure lightpaths was demonstrated along with our lightpath-aware GridFTP during the 2003 CANARIE Advanced Networks Workshop [2]. The application was used to set up the demonstration environment by creating and partitioning a series of root lightpaths, concatenating a subset of the children, and performing a series of advertisements. A physical topology consisting of five nodes was simulated using a lab equipped with two Cisco ONS 15454 SONET platforms connected by an OC-192 facility. This was accomplished by executing multiple resource agents on each switch, and assigning a subset of the 192 available channels to each logical link in the simulated topology.

4. Conclusion and Future Work

In this paper, we have presented our User-Controlled Lightpath Management System as a solution that fulfills the networking needs of Grids.

Our prototype implementation provides the essential services needed to establish lightpaths on demand, and can be integrated into other Grid applications, such as GridFTP, to provide high-performance network support. However, several important problems remain to be explored.

Presently, we are investigating the problem of distributing the Service Provisioning Layer. The motivation behind this is to allow each management domain to administer its own, potentially customized, instance of the system. For example, this allows different domains to use different Grid hosting environments and implementation languages. In addition, a distributed SPL offers greater fault tolerance and scalability. The XML-based service description, discovery, and invocation mechanisms of Web Services technology simplifies the problem of interoperability between instances.

In support of a distributed SPL, one requires a method of disseminating the availability of inter-domain lightpaths. BGP cannot be used for this as its route advertisements carry connectivity information only, and do not provide a mechanism for QoS routing. However, the hierarchical routing paradigm of BGP is a natural basis for more sophisticated inter-domain peering mechanisms.

As mentioned in Section 3, we assume that users have direct access to optical networks. However, a more general solution where users do not have direct access is also of interest, because this will further improve the applicability of Grids and our lightpath management system. The integration of our lightpath-aware GridFTP or other Grid applications/services with explicit routing mechanisms is currently under investigation. For example, if IPv6 is adopted, then a way of specifying IP sequences inside Grid applications can be introduced.

A fundamental problem in using lightpaths for communication is controlling quality of service between the end host and the cross-connect device used to access the optical network. Unless this last mile consists of a dedicated LAN, one faces the problem of supporting the guaranteed bandwidth of the lightpath using whatever QoS mechanisms are present in the routing and switching equipment at hand. In the case of IEEE 802 LANs, one can consider the possibility of applying basic MAC level QoS as defined in the 802.1P and 802.1Q standards. The former can be used to provide expedited forwarding for lightpath traffic in the event that ambient traffic is present. The latter allows the formation of Virtual LANs, whereby the segments of a LAN can be isolated in such a way that broadcast and multicast frames in the lightpath traffic flow do not flood the entire LAN. At the same time, this provides increased security for lightpath traffic. Together, these technologies provide a way to extend lightpaths beyond the core optical network and into the LAN.

Another possible extension is to add support for provisioning of survivable lightpaths. This functionality could be incorporated into the end-to-end lightpath establishment operation. For example, users could be given the choice of using path protection, if it is supported by the network hardware, a slower but more resource-efficient restoration mechanism, or no protection at all, depending on the requirements of the traffic.

5. Acknowledgement

The following persons are hereby acknowledged in alphabetical order for their contribution to the development of the lightpath management system: Shivani Aggarwal, Bassam Aoun, Mohamed Dadi, Acila Derbal, Adel Ghlamallah, Boris Jabes, Darcy Quesnel, Basem Shihada, Wayne Szeto, Thomas Tam, and Iban Touchet.

References

1. CANARIE Inc., 'CA*net4 connected institutions map'. http://www.canarie.ca/canet4/connected/canet4_map.html.
2. CANARIE Inc.: 2003, 'CANARIE Advanced Networks Workshop 2003'. http://www.risq.qc.ca/risq2003-canw2003/en/index.html.
3. Foster, I., C. Kesselman, G. Tsudik, and S. Tuecke: 1998, 'A Security Architecture for Computational Grids'. In: *Proc. 5th ACM Conference on Computer and Communications Security Conference*. pp. 83–92.
4. Gray, J., 'Distributed Computing Economics'. http://www.clustercomputing.org/content/tfcc-5-1-gray.html.
5. Gray, J.: 2003, 'Data Management Services'. http://www-unix.globus.org/developer/data-management.html.
6. St Arnaud, B.: 2002, 'Frequently Asked Questions about Customer Owned Dark Fiber, Condominium Fiber, Community and Municipal Fiber Networks'. http://www.canarie.ca/canet4/library/customer /frequentlyaskedquestionsaboutdarkfiber.pdf.
7. Sun Microsystems: 2003, 'Java Web Services Developer Pack'. http://java.sun.com/Webservices/downloads/Webservicespack.html.
8. Szeto, W., Y. Iraqi, and R. Boutaba: 2003, 'A Multi-Commodity Flow Based Approach to Virtual Network Resource Allocation'. In: *Proc. IEEE Global Telecommunications Conference (GLOBECOM'03)*. San Francisco, CA (USA).
9. The Globus Alliance, 'Grid Security Infrastructure'. http://www-unix.globus.org/security/.
10. Xue, G., M. J. Fairman, and S. J. Cox: 2002, 'Exploiting Web Technologies for Grid Architecture'. In: *Proc. 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGRID'02).*