Optical Exchanges

Freek Dijkstra and Cees de Laat
Universiteit van Amsterdam
Kruislaan 403, 1098 SJ, Amsterdam, The Netherlands
fdijkstr@science.uva.nl, delaat@science.uva.nl

Abstract– **Peering of network traffic is often concentrated at Internet Exchange Points. Current Internet Exchanges offer services at layer 2 and layer 3. There is a growing need for transport services where the traffic remains at lower layers. This article describes the rationale and model for an exchange, called an Optical Exchange, with services at these lower layers. It describes the basic interfaces and services that an Optical Exchange may provide.**

I. INTRODUCTION

A long term-goal of optical network research is to provide cheap optical routers. In attempts to achieve this goal, new technologies are being developed [1], starting with Optical Add-Drop Multiplexers (OADM), to photonic Multi-Protocol Label Switching (MPLS) routers and recently Optical Burst Switching (OBS).

It should be understood that technologies like MPLS and OBS aim at connection-less solutions. However, we belief there is a need for optical networks which aim at connect-oriented connections [2]. There are two strong arguments for this:

1) First, connection-oriented solutions are guaranteed congestion-free. There is no need to impose all kinds of Quality of Service (QoS) technologies. In addition, connection-oriented links allow users to use more efficient, but TCP unfriendly transport protocols.

2) Second, it is believed that connection-oriented technologies will remain an order of magnitude cheaper because there is no need to look into headers at the data stream itself. Even if the cost of these techniques will get a factor cheaper[1], other optical components will become cheaper as well, and the price ratio remains.

It is efficient to keep traffic that does not need any high-level functionality at the lowest layer possible. In particular, this will apply to relatively long (minutes or more) and high bandwidth (Gigabit/s or more) data streams between a limited number of destinations. Examples of these are data replication processes and instrumentation grids where huge data streams must be collected at a central place (like the eVLBI project [4]). The packets in these streams do not need routing. One knows the data has to go from this source to that destination. In general, this is called the service approach: only apply the required services, nothing more.

A. Current Internet Exchanges

There are currently three types of Internet Exchanges [5,6]:

1) *LAN-Based Internet Exchanges*: The most common exchange, typically with a layer 2 switch at the core, though some exchanges are distributed. This is the only stateless exchange. Blocking is possible if multiple peers want to send traffic to the same peer at the same time.

2) *ATM-Based Internet Exchanges*: A state full exchange, with permanent virtual circuits (PVCs) at the core. If variable bit rate (VBR) circuits are used, there is no guaranteed congestion-free transmission. Usage of constant bit rate (CBR) on the other hand, results in very inefficient usage of resources and poor scalability.

3) *MPLS-Based Internet Exchanges*: MPLS exchanges are both state full and packet based. Though the service may be congestion-free, it requires a lookup at the IP routing table at the Ingress Edge Label Switching Router (LSR). This is a relative expensive operation for very-high bandwidth data streams.

Given the properties of these types of exchanges, we can conclude that these architectures are not sufficient to support few high bandwidth flows that do not require routing. Either it is technically impossible (for LAN-based Internet Exchanges) or it yields unnecessary and costly overhead (for ATM- and MPLS-based Internet Exchanges).

B. Need for Optical Exchanges

Even with a limited number of destinations, it is very likely that the demanding data streams, as mentioned above, will traverse multiple network domains. Considering the scalability, it is likely that the peerings between the domains will cluster at peering locations.

A trivial co-location is one that provides no other functionality than rack space and the ability for providers to have bi-lateral peerings with other providers at the same co-locations. Peering of regular Internet traffic has let to switching-based Internet Exchanges. Similar, we belief that peering locations that do not require switching technology will emerge. We will call these *Optical Exchanges*[2].

This article first describes an Optical Exchange as a black box, with certain interfaces and services. Secondly, it will give an example of how it might be implemented.

II. OPTICAL EXCHANGES

Different applications require different transport services [2], ranging from layer 0 to layer 3[3]. Instead of proposing

1 For example when optical shift registers [3] become a commodity.

[2] Not *Optical Internet Exchange*, because there is no need that the traffic going over such wires are IP packets, or is even packet based at all. The *Optical* adjective refers to the fact that on layer 0 and layer 1, where most peering will take place, are dominated by optical technologies like DWDM.

[3] We will use the layering terminology common in the telecom world: Thus, carriers and fiber switching is at layer 0, Wavelength

four or more different types of exchanges, we propose a generic architecture for an exchange, consisting of a black box with interfaces to connect to it, and services within. See figure 1.

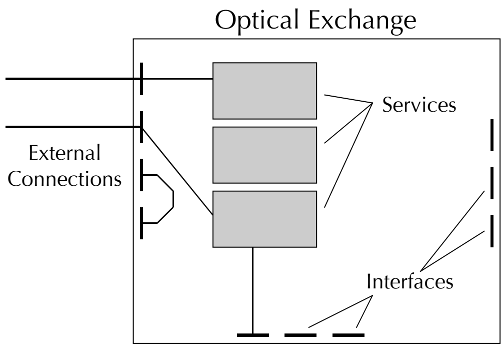

Figure 1. Components of an Optical Exchange

The interfaces may be of different type. For example, one interface may be used to carry undefined traffic over 32 wavelengths using Dense Wavelength Division Multiplexing (DWDM), while an other interface may only carry one signal at 1310 nanometer, which carries SDH-framed traffic. A third interface may be LAN-PHY based Ethernet, where traffic must reside in the 145.146.100.0/24 subnet.

The basic functionality of any exchange is to transport traffic from one provider to another provider. For an Optical Exchange, this implies that the main functionality is to be able to cross-connect each interface to another interface, forming a circuit. In addition, the different type of interfaces pose the requirement on an Optical Exchange to also be able to extract a certain signal from a carrier, and inject it in an other carrier using some kind of multiplexing. For example if one carrier uses DWDM with 50 GHz spacing, and another carrier with 100 GHz spacing, it will probably be necessary to first demultiplex (demux) and then multiplex (mux) the signal again. If a signal goes in at one layer and goes out at a different layer, the Optical Exchange effectively acts as an "elevator", lowering or higher traffic to different layers. For example if a signal comes in as a wavelength using DWDM, and is injected in a VLAN, the signal is elevated from layer 1 to layer 2.

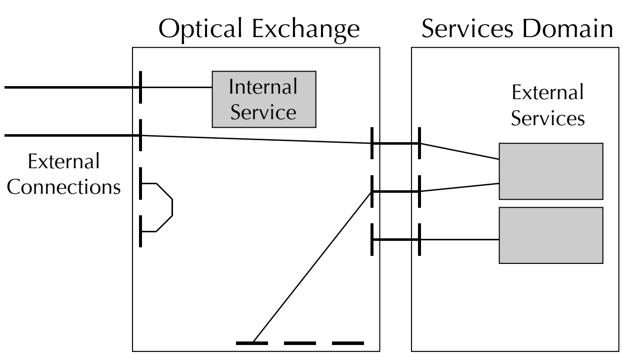

Figure 2. Internal and External Services

Division Multiplexing (WDM) and Time Division Multiplexing (TDM) is at layer 1, switching and aggregation is at layer 2, and IP routing is at layer 3.

The first type of services is making cross connects. The second type of services is mux and demuxing of bit streams. Finally, services that are even more complex may be required by users. Already mentioned is aggregation of traffic using a switch, but also optical multicast and store-and-forward services fall into this category. An Optical Exchange may only provide a subset of all possible services. Either the remaining services are not offered at all, or they may be offered by a service provider connected to the exchange, as figure 2 shows. It should be emphasized that there is no technical obligation to put one set of services in the Optical Exchange domain (the *Internal Services*) and another set in a Services Domain (the *External Services*). The decision as to which service should be an internal service and which one an external service is a business decision, and is out-of-scope of this article.

III. Interfaces and Protocols

An Optical Exchange may accept one or more types of interfaces. The list bellow defines what we belief will be the most common interfaces.

At layer 0, we only consider single-mode fibers. We specifically ignore multi-mode fiber and electrical (UTP) carriers, because in our experiences with early pseudo Optical Exchanges like NetherLight [7] and StarLight [8], there is a trend towards single mode fiber, away from multimode fiber and UTP[4].

At layer 1, each fiber may either carry

1) a single bit stream, within the 1260-1675 nm range, or

2) multiple bit streams of data using separate DWDM wavelengths (lambdas).

Each bit-stream (lambda) uses one the following sampling and framings:

1) A bit-stream, at a certain wavelength, with up to 10.1 GHz or up to 40 GHz sampling, without other known properties.

2) A bit-stream, at a certain wavelength, with SONET/SDH framing, either OC-48 (2.5 Gbit/s), OC-192 (10 Gbit/s) or OC-768 (40 Gbit/s)

3) A bit-stream, at a certain wavelength, with 1 Gbit/s Ethernet,

4) A bit-stream, at a certain wavelength, with LAN PHY Ethernet,

5) A bit-stream, at a certain wavelength, with WAN PHY Ethernet,

On the local up to the national scale, the dominant technology is Ethernet. Each fiber typically carries one wavelengths or it carries 32 or 64 wavelengths using DWDM. The speed is typically either 1 Gb/s or 10 Gb/s LAN-PHY variant.

[4] There are three likely causes for this trend. First, Optical Cross Connects with MEMS switches absorb light at 800 nm, and thus cannot deal with multi-mode fiber. Secondly, DWDM usage is increasing, and DWDM is only possible with single mode fiber. Third, single-mode has a wider range of operation (few hundred km) then multi-mode (about 2 km).

For regional up to worldwide scale (there is some overlap), typically one SONET/SDH or WAN PHY signal is used per carrier. The SONET speed may be OC-48 (2.5 Gb/s), OC-192 (10 Gb/s) or OC-768 (40 Gb/s). The reason that TDM (with SONET or SDH) is used on a worldwide scale is because on trans-oceanic links the providers of these links require customers to comply with SONET/SDH framing. Note that WAN PHY is compatible with SONET framing and in practice also with SONET optical equipment [9]. Thus, WAN PHY can be used on trans-oceanic links

There are many ways to multiplex signals in one carrier using DWDM, due to the variety of wavelength bands and channel spacings. Similarly not all vendors encapsulate Ethernet in the same way over a SONET/SDH connection. An Optical Exchange must need this kind of information to determine the correct service to use. Of course, if no demultiplexing is required, then the Optical Exchange does not need any of this information.

A major issue when connecting two fibers together is to correctly tune the power levels. Especially with single mode fibers, a single spec of dust may ruin signal strength. When making and breaking connections between two carriers on the fly, using an Optical Cross Connect[5], we recommend that the signal strength has a known, pre-defined value when leaving the Optical Exchange domain.

The protocols used on each circuit is either unknown to the Optical Exchange, or it may be specified by:

1) IP over Ethernet. Note that if all interfaces are of this type, the Optical Exchange would be reduced to a LAN-based Internet Exchange.
2) Ethernet with 802.1q (VLAN-tagging) on top of it [10].

Most end-to-end connections are full-duplex, even if the application does not require dedicated bandwidth in both directions. Most applications do not support a one-way connection, but a relatively easy way to optimize resource usage is to support asynchronous connections. At minimum, an Optical Exchange must be aware if an interface is always full-duplex, or supports one-way connections.

IV. SERVICES

A. Services List

We have identified the services below for an Optical Exchange. However, some of these services do not have to be offered by the Optical Exchange itself, but may be outsourced to a services provider connected to the Optical Exchange, or may not be offered at all. The abbreviations are used in the service matrix below.

1) *Cross (Connect):* Given two interfaces of equal type, be able to make a cross-connect between these interfaces. Typically, this should be done in a user- or traffic-initiated way by a software control plane. However, we do not impose this limitation.
2) *Regenerate:* Amplify or attenuate the power levels, to match a certain output power level; amplify and reshape; or reamplify, reshape and retime (3R).
3) *λ convert:* Wavelength conversion, by regenerating the signal or by physically altering the wavelength. Regenerating, using tunable transponders, may allow network engineers to monitor the Bit-Error-Rate (BER) of a signal, but requires the regenerator to understand the modulation and possible framing of the signal.
4) *WDM mux/demux*: Multiplex wavelengths of different color into a single carrier, and demultiplex different signals in a single carrier into many separate fibers. This process does not need to convert the wavelengths itself. An advanced demultiplexer may first demux the signal into sets of wavelengths, called wavebands, before the wavebands are demuxed into individual wavelengths [11]. Also, not all multiplexers are compatible, since different optical bands and channel spacings may be used.
5) *(Optical) multicast*: The ability to duplicate an optical signal as-is. Of course, this can only be done one-way. Possible usages include visualization.
6) *TDM mux/demux:* The ability to extract an Ethernet signal from a SONET or SDH carrier or to insert one or more Ethernet connections in a SONET or SDH carrier. It should be noted that not all SONET/SDH multiplexers are compatible.
7) *SONET (switching):* The ability to combine and switch SONET or SDH circuits, without knowing the contents of the circuits.
8) *Aggregate*: There may be different needs for an Ethernet switch in an Optical Exchange. First, it allows aggregation of traffic. This may cause congestion.
9) *(Ethernet) conversion*: A second use for an Ethernet switch is the conversion between different types of framing. The most useful conversion is perhaps the ability to convert between LAN PHY and WAN PHY.
10) *VLAN encap/decap*: Ethernet encapsulation: A third use for an Ethernet switch is the encapsulation of traffic in a VLAN trunk [10]. This allows combining different, separable data streams on a single link.

The combination of DWDM support, wavelength conversion and cross-connects will effectively make an Optical Add-Drop Multiplexing (OADM) facility.

Multiple services may be combined in a single device.

B. External Services

It is expected that an Optical Exchange itself does not offer any services on layer 3. However, service domains connected to an Optical Exchange may provide layer 3 services. For example:

11) *L3 exit/entry*: The ability to exit to the layer 3 Internet (L3 exit/entry), coming from a dedicated connection (lambda) or visa versa. This service may not only include the establishment of a physical link, but also negotiation of IP addresses to use. Care should be taken when offering this service, because it may allow easy creation of bypasses on the regular Internet, causing BGP-ripples if such bypasses are continuously created and destroyed.
12) *Store-and-forward:* One way to reduce blocking chances is to transport large chunks of data on a hop-by-hop

[5] With an optical device, we refer to an OOO device, not to an OEO device.

basis. A location near an Optical Exchange would be the ideal place for a terabyte storage facility.

This list of external services is not complete. For example, MPLS and ATM services were not considered in this list.

C. *Service Matrix*

Services are transition from one type of interface to another or the same type of interface, or sometime to multiple interfaces. In the Service Matrix below, the conversion from an interface listed on the left to an interface on top is listed.

TABLE I. SERVICE MATRIX

From \ To	WDM (multiple λ)	Single λ, any bitstream	SONET/ SDH	1 Gb/s Ethernet	LAN PHY Ethernet	WAN PHY Ethernet	VLAN tagged Ethernet	IP over Ethernet
WDM (multiple λ)	cross-connect, multicast, regenerate, multicast	WDM demux	WDM demux*	WDM demux *	WDM demux *	WDM demux *	WDM demux *	WDM demux *
Single λ, any bitstream	WDM mux	cross-connect, λ conversion, regenerate multicast	N/A *	N/A *	N/A *	N/A *	N/A *	N/A *
SONET/SDH	WDM mux	N/A *	SONET switch, +	TDM demux *	TDM demux[6]	SONET switch	TDM demux *	TDM demux *
1 Gb/s Ethernet	WDM mux	N/A *	TDM mux	aggregate, Ethernet conversion, +	aggregate, eth. convert	aggregate, Ethernet conversion	aggregate, VLAN encap	L3 entry *
LAN PHY Ethernet	WDM mux	N/A*	TDM mux[6]	aggregate, Ethernet conversion	aggregate, Ethernet conversion, +	Ethernet conversion	aggregate, VLAN encap	L3 entry *
WAN PHY Ethernet	WDM mux	N/A *	SONET switch	aggregate, Ethernet conversion	Ethernet conversion	aggregate, Ethernet conversion, +	aggregate, VLAN encap	L3 entry *
VLAN tagged Ethernet	WDM mux	N/A *	TDM mux	aggregate, VLAN decap	aggregate, VLAN decap	aggregate, VLAN decap	Aggregate, VLAN decap & encap, +	N/A
IP over Ethernet	WDM mux	N/A *	TDM mux	L3 exit *	L3 exit *	L3 exit *	N/A	Store & forward, L3 entry/exit, +

Additional legend for table I:

N/A *Not Available*. However, it may be possible to go from one to the other interface by applying two services in series, using an intermediate interface or protocol.

\+ These functions may also be possible between one interface and another interface of the same type: cross connect, regeneration, and λ conversion.

* Only possible if the interface type is correct. For example, it may be possible to demux DWDM carrier to LAN PHY Ethernet, but of course only if one of the wavelengths already contains LAN PHY Ethernet.

V. IMPLEMENTATION

A. *Cost Optimization: How Low Can You Go.*

In the end, the reason for building Optical Exchanges next to regular Internet Exchanges is a matter of cost-optimization. You should offer only the services required for your application, in the most cost efficient way possible.

If a non-blocking Ethernet switch or router with the same amount of 10 Gbit/s interfaces would be as cheap as an Optical Cross Connect (OXC), it would technically be sound to use switches and routers instead of OXCs, because that is a lot more flexible. However, if the price difference between an OXC and a switch or router is high enough, it is cost-wise more optimal to use OXCs as much as possible.

Determining the exact trade-off is out-of-scope for this article. As a rule of thumb, on short distances (2 ms scale), fibers are inexpensive, while ports on switches and routers are expensive. For these scales, it often pays off to add an Optical Cross Connect in front of more expensive

[6] Unlike WAN PHY, It is not possible to put LAN-PHY in an OC-192 channel, since LAN PHY has a higher bitrate. However, it is technically possible, albeit uncommon, to put LAN PHY in an OC-256 or OC-768 channel, so it is listed here.

equipment so that you may be able to save on expensive ports. For example, if you have eight 10 GbE LAN PHY connections coming in, and 60% of the time they just need to forward the traffic, an OXC could take care of that, and having six 10 GbE LAN PHY connections on the switch instead of eight may offer sufficient flexibility.

For longer distances, like worldwide scale (200 ms), fibers are relatively expensive, compared to the cost of network devices, in particular when it applies to trans-Pacific or trans-Atlantic connections. For the regional/national scale (20 ms), we see a move from expensive fibers to relative inexpensive fibers and expensive ports.

B. DWDM Optimization

A service that may be offered by an Optical Exchange is to demux multiple DWDM wavelengths on a single fiber into streams on multiple fibers. If the signal sometimes needs to be demuxed, but at other times all wavelengths are just forwarded to an egress interface, it may be cost-effective to pass the full (not demuxed) signal through an OXC, so that at these times, the demultiplex can be used for other purposes. The added cost is two ports on an OXC. Rough calculations have shown that doing this may be cost-effective when multiplexing is needed in roughly less then 90% of the time, for DWDM with 32-channels.

C. Blueprint for an Optical Exchange

In the previous section, we explained how the adding a low-cost Optical Cross Connect (OXC) in front of a service might reduce the number of expensive ports required at service equipment. An OXC, essentially an automated patch panel, is the simplest type of device available, making it one of the cheapest devices available. In addition, since it is a true optical device, ignorant of transmission speed, a 10 Gb/s port costs as much as 1 Gb/s port.

Therefore, one vision for an Optical Exchange is to place an OXC at the core, with each interface connected to it. This applies both for external interfaces, as well as interfaces to the different service devices. Figure 3 shows an example of such an exchange. Of course, it is possible that an exchange can contain multiple OXCs.

D. Global Lambda Integrated Facility

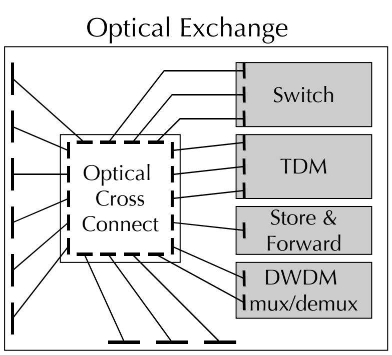

Figure 3. Possible implementation of an Optical Exchange with an Optical Cross Connect at the core and providing miscellaneous services.

It is expected that Optical Exchange, along with network connections (lambdas) in between, and grid services on the edges will together form a global layer 1 data network. The Global Lambda Integrated Facility [12] is an attempt to make this integration a reality.

VI. Web Services Approach

A. Policy Enforcement

An Optical Exchange is connection oriented. Thus, a user or application must set up a circuit prior to its usage. A provider connects with a limited number of bit-stream interfaces to an Optical Exchange, so a policy must be enforced to prevent denial-of-service attacks. Since the Optical Exchange is state full, the policy must be or must also be enforced at the exchange itself.

A message (path setup, status query, path tear down, etc.) will be sent from a end-user user or application or from the provider. This can be in-band, or over a separate control plane. With regular Internet Exchange, routing information is sent in-band using BGP. An extension to BGP for lightpaths, called OBGP, has been proposed by [13], but has not been implemented so far.

It is the job of the Optical Exchange to verify if the user is allowed to make the specified request, and if so, fulfill the request. The verification may involve sending an inquiry to other domains if the user wants to connect to an other domain.

This control model maps seamlessly onto the Authentication, Authorization and Accounting framework as described in [14]. If we apply the framework to this situation, the Optical Exchange takes the role of an administrative domain, just like each provider also has the role of an administrative domain. Thus, as far as the AAA control plane is concerned, the concept of an Optical Exchange would be reduced to the concept of many private peerings.

If a peer does not wish to add an authorization phase to the path set-up, it can choose to enforce a trivial policy: thus, allowing each connection, regardless of who made the request.

B. Grid Integration

To facilitate Grid integration of an Optical Exchange, the network elements and the services offered by the Optical Exchange should be exposed as Web Services [15] as specified by [16]. Typically, on the lowest level, each network element is virtualized by a Web Services. On higher levels, the functions of these Web Services are combined to form high level services. For example, if an Optical Exchange and a Service Domain connected to the Optical Exchange both expose their services as web services, then it is possible for a broker domain to combine both services and expose them as a single service to the users.

It is hard to design a control plane with a resource manager, which can combine all the different resources on the fly in an intelligent way.

If two providers want to be able to make reservations of circuits in time, the Optical Exchange must support this as well. Thus, the peering policy must not only support current state (check if a circuit is available), it must know future states as well. Clearly, at least some intelligence must go in the plane that controls the network. A typical example of the message passing to support policy enforcement is given in the previous section.

C. *Monitoring*

A disadvantage of handling traffic at layer 0 is that is not possible to monitor error counters, even if signal strength can be measured. Typically, traffic is kept at layer 0 within a domain, but OEO conversion takes place at the edges. If errors occur, this enables network engineers to at least pinpoint the administrative domain where the error originates. If the optical boundaries stretch between domains, this may not be possible. There are multiple solutions to this problem:

1) Keep the OEO conversion at domain boundaries, at the expense of higher costs;

2) Make sure that power levels are correct at domain boundaries, and be able to measure error rates, for example using optical multiplexing and a specific device;

3) Expose error counters as web services to allow other domains to still monitor the end-to-end quality of a circuit. This assumes that the domain publishing the error counters does not have an incentive to tamper with this data. This is generally the case if it is only used for QoS monitoring and not for billing purposes.

4) A provider or Optical Exchange may offer two types of services for a circuit: a cheap one at layer 0, where the client should expose their error counters, or an expensive circuit at layer 1.

VII. Conclusion

Optical Exchanges allow the exchange of data traffic below layer 2 of the OSI model. An Optical Exchange may accept multiple type of interfaces and protocols, and may offer services to connect the different interfaces with each other. Services may also be offered by a service provider connect to the exchange. For Grid-integration, exchanges may expose their networks elements as web services.

Acknowledgement

This work was funded by the GigaPort NG project [17]. In the first place, a word of thank you goes to Erik-Jan Bos for his excellent remarks and for proofreading the article. In addtion, thanks to Dennis Paus and Erik Radius for their discussions, to Bert Andree for and to Bo Tian for providing interesting references. In addition, the authors wish to express their gratitude towards Tom DeFanti, Bill St. Arnaud, Kees Neggers and Joe Mambretti for their long-term discussions and support.

References

[1] Masafumi Koga and Ken-ichi Sato, "Recent Advances in Photonic Networking Technologies", Proceedings of International Conference on Optical Internet, July 2002

[2] Cees de Laat, Erik Radius and Steven Wallace, "Rational of Current Optical Networking Initiatives", Future Generation Computer System 19, August 2003

[3] Bo Tian, Wim van Etten and Wim Beuwer, "Ultra Fast All-Optical Shift Register and It's Perspective Application for Optical Fast Packet Switching", IEEE Journal of Quantum Electronics, Vol. 8, no. 3, 2002

[4] Consortium for Very Long Baseline Interferometry in Europe, http://www.evlbi.org/

[5] Geoff Huston, "Interconnection, Peering, and Settlements", Proceedings of Inet'99, June 1999

[6] Ikuo Nakagawa, Hiroshi Esaki, Yutaka Kikuchi and Kenichi Nagami, "Design of Next Generation IX Using MPLS Technology", IPSJ Journal, November 2002

[7] NetherLight, http://www.netherlight.net/

[8] StarLight, http://www.startap.net/starlight/

[9] Catalin Meirosu, et al., "Native 10 Gigabit Ethernet Experiments between Amsterdam and Geneva", Future Generation Computer Systems. In press.

[10] IEEE 802.1q, "VirtualBridged Local Area Networks" IEEE standard, 2001

[11] Xiaojun Cao, Vishal Anand and Chunming Qiao, "Waveband Switching in Optical Networks", IEEE Communications Magazine, April 2003

[12] Global Lambda Integrated Facility, http://www.glif.is/

[13] Bill St. Arnaud, Rene Hatem, Wade Hong, Marc Blanchet and Florent Parent, "Optical BGP Networks", March 2001, http://www.canet4.net/library/canet4design.html

[14] John Vollbrecht, et al., "AAA Authorization Framework", RFC 2904, August 2000

[15] Web Services Resource Framework, http://www.globus.org/wsrf/

[16] Ian Foster, Hiro kishimoto (editors), "The Open Grid Services Architecture", Work in Progress, http://forge.gridforum.org/projects/ogsa-wg/

[17] GigaPort NG, http://www.gigaport.nl/