Network Working Group
Internet Draft
Expiration Date: May 2001
Document: draft-bernstein-optical-bgp-01.txt

G. Bernstein, L. Ong
Ciena
B. Rajagopalan
Tellium
Angela Chiu
Celion
Frank Hujber
Alphion
John Strand
AT&T
V. Sharma
Metanoia
Sudheer Dharanikota
Nayna Networks
July 2001

Optical Inter Domain Routing Considerations

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026 [1].

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts. Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet- Drafts as reference material or to cite them other than as "work in progress."
The list of current Internet-Drafts can be accessed at
http://www.ietf.org/ietf/1id-abstracts.txt
The list of Internet-Draft Shadow Directories can be accessed at
http://www.ietf.org/shadow.html.

Abstract

This draft investigates the requirements for general inter-domain and inter-area routing in optical networks and reviews the applicability of existing route protocols in various optical routing applications.

Table of Contents:

1.1 Specification of Requirements.....3
1.2 Abbreviations.....3
2 Background.....3
2.1 Major Differences between Optical and IP datagram Routing.....4
2.2 Reachability.....5

2.3 Capability and Capacity Advertisement.....5
2.3.1 Subnetwork Capability Advertisement.....5

2.3.2 End System Capabilities.....6
2.4 Diversity in Optical Routing.....7
2.4.1 Generalizing Link Diversity.....8
2.4.2 Generalizing Node Diversity.....9
3 Applications of Optical Inter Domain Routing.....9
3.1 Inter-Area Routing.....9
3.1.1 Inter-Area Scalability.....9
3.1.2 Inter-vendor Inter-area.....10
3.1.3 Legacy Interoperability Inter-area.....10
3.1.4 Inter-Layer Partitioning.....12
3.2 Classical Inter-Domain (Inter-Carrier).....13
3.3 Multi-Domain Connection Control.....15
4 Multiple Layers of Routing.....16
4.1 Layers in Transport Networks.....16
4.2 Optical Physical Layer Routing.....17
4.2.1 Reconfigurable Network Elements.....17
4.2.2 Wavelength Routed All-Optical Networks.....18
4.2.3 More Complex Networks.....19
4.3 SDH/SONET layer Routing.....20
4.3.1 Switching Capabilities.....20
4.3.2 Switching Granularity.....20
4.3.3 Protection.....21
4.3.4 Available Capacity Advertisement.....22
4.4 Layer Integration.....23
4.5 Interaction with IP Layer Routing.....25
5 Existing Routing Protocol Applicability.....25
5.1 OSPF Applicability.....25
5.2 PNNI Routing.....26
5.2.1 PNNI overview.....26
5.2.2 PNNI Optical Applicability.....28
5.3 BGP Applicability.....29
5.3.1 Pick One! (route that is).....29
5.3.2 Reachability: Via Optical BGP like functionality.....29
5.3.3 Integrated with IP BGP?.....30
5.3.4 Policy Mechanisms.....30
6 Conclusion.....31
7 Security Considerations.....31
8 References.....31
9 Acknowledgments.....33
10 Author's Addresses.....33

1 Introduction

Multi Protocol Label Switching (MPLS) has received much attention recently for use as a control plane for non-packet switched technologies. In particular, optical technologies have a need to upgrade their control plane as reviewed in reference [2]. Many

different optical switching and multiplexing technologies exist and more are sure to come. For the purposes of this draft we only consider non-packet (i.e. circuit switching) forms of optical switching.
As the requirements for and extensions to interior gateway protocols such as OSPF and IS-IS have begun to be investigated in the single area case, e.g., reference [3], we consider the requirements that optical networking and switching impose in the inter-domain case.
By inter-domain in this draft we consider inter-area, inter-layer, and inter-vendor partitioning of routing and possibly other possibilities for partitioning routing in addition to administrative inter-domain (inter-carrier) partitioning. Comparisons of these requirements to existing functionality in BGP, multi-area OSPF and hierarchical PNNI will be made.
In particular, optical routing needs to provide for path diversity, switching capabilities, transport capabilities and impairments, and bandwidth/resource status reporting.
To add to the concreteness of these considerations we try to illustrate them with one or more specific examples from a particular optical networking layer or technology. This is not to reduce the generality of the requirement but to facilitate the understanding of the requirement or concept.

1.1 Specification of Requirements

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT","SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119.

1.2 Abbreviations

LSP	Label Switched Path (MPLS terminology)
LSR	Label Switched Router (MPLS terminology)
MPLS	Multiprotocol Label Switching
SDH	Synchronous Digital Hierarchy (ITU standard)
SONET	Synchronous Optical NETwork (ANSI standard)
STM(-N)	Synchronous Transport Module (-N)
STS(-N)	Synchronous Transport Signal-Level N (SONET)
TU-n	Tributary Unit-n (SDH)
TUG(-n)	Tributary Unit Group (-n) (SDH)
VC-n	Virtual Container-n (SDH)
VTn	Virtual Tributary-n (SONET)

2 Background

The motivation for inter domain routing in optical networks (circuit switched) is very similar to that in the case of IP datagram routing.

1. Distribute "reachability" information throughout an internetwork. An internetwork consists of an interconnected set of networks under different routing and/or administrative domains.

2. Maintain a clear separation between distinct administrative or routing domains.

3. Provide "information hiding" on the internal structure of the distinct administrative or routing domains.

4. Limit the scope of interior gateway routing protocols. This is for security, scalability reliability and policy reasons.

5. Provide for address/route aggregation.

2.1 Major Differences between Optical and IP datagram Routing

Let us first review the major difference between routing for optical (circuit switched networks) and IP datagram networks. In IP datagram networks packet forwarding is done on a hop-by-hop basis (no connection established ahead of time). While circuit switched optical networks end to end connections must be explicitly established based on network topology and resource status information. This topology and resource status information can be obtained via routing protocols. Note that the routing protocols in the circuit switch case are not involved with data (or bit) forwarding, i.e., they are not "service impacting", while in the IP datagram case the routing protocols are explicitly involved with data plane forwarding decisions and hence are very much service impacting.

This does not imply routing is unimportant in the optical case, only that its service impacting effect is secondary. For example, topology and resource status inaccuracies will affect whether a new connection can be established (or a restoration connection can be established) but will not (and should not) cause an existing connection to be torn down.

This tends to lead to a slightly different view towards incorporating new information fields (objects, LSA, etc.) into optical routing protocols versus IP routing protocols. In the optical circuit case, any information that can potentially aid in route computations or be used in service differentiation may be incorporated into the route protocol, as either a standard element or a vendor specific extension. Whether a route computation algorithm uses this information and whether two route computation algorithms use this information in the same way doesn t matter since the optical connections are explicitly routed (although perhaps loosely). The optical route computation problem is really a constraint-based routing problem. The basic route calculation is an atomic service that occurs, for a given connection, in a single network element. (In the case of loose explicit routing some details

may be filled in by other NE s.) This means that, even in a heterogeneous optical network, NEs from different vendors need not use the same algorithm.

Another difference - clear, hard blocking prevails in the optical world while some level of overloading is ok in the IP world, i.e., statistical multiplexing is not available with optical circuits. This also manifests itself in the commitment of the protection (or restoration) bandwidth. In a packet-based network although the protection path can be setup prior to any fault, the resources along the protection path are not used until the failure occurs. In circuit-based networks a protection path generally implies a committed resource. Such a basic difference restricts the direct applicability of some of the traffic engineering mechanisms used in a packet-based network to a circuit-based network.

2.2 Reachability

The main goal of path selection (route computation) is to find the best path(s) between a set of <source, destination> pairs satisfying a given set of constraints and possibly network optimality conditions. To aid in performing such path computation routing protocols carry information related to the topology of the network(characteristics of the links, nodes, subnetworks and domains).

Associated with a subnetwork we can ask what systems can be reached via this subnetwork. These systems can be nodes within the network, end systems (clients) to the network, or other subnetworks. Now this reachability information isn't too valuable unless there is at least one known path to reach that subnetwork.

2.3 Capability and Capacity Advertisement

2.3.1 Subnetwork Capability Advertisement

In addition to understanding what systems are directly reachable via a subnetwork it can be important to know about the capabilities or features offered by the subnetwork. Subnetwork information we will want to know includes:

1. Switching capabilities
2. Protection Capabilities
3. Available Capacity
4. Reliability Measures (if available)

Examples:

1. For example, in the SONET realm, one subnetwork may switch down to an STS-3c granularity while another switches down to an STS-1 granularity. Understanding what types of signals within a SDH/SONET multiplex structure can be switched by a subnetwork is important. Similar examples of granularity in switching apply to the waveband case.

2. Some networking technologies, particularly SONET/SDH, provide a wide range of standardized protection technologies. But not all subnetworks will offer all protection options. For example, a 2/4-F BLSR based subnetwork could offer extra data traffic, ring protected traffic and non-preemptible unprotected traffic,

(NUT)[4], while a mesh network might offer shared SONET line layer linear protection and some form of mesh protection.
3. Capacity information can be tricky to represent for an entire subnetwork. More than likely a subnetwork that provides a "transit" service would offer some type of summarized topological model from which capacity constrained routing decisions could be made.
4. Some subnetworks may be in locations that have lower incidences of link failure. Such information could be helpful in computing routes to statistically "share the pain".

The type of regeneration (if any) done at the NNI by each subnetwork will also need to be known. There are several reasons for this:
1. When entering or leaving an all-optical subnetwork, the impairment budget available for the next subnetwork will depend on this;
2. The routing process needs to be sensitive to the costs associated with "island-hopping".

This last point needs elaboration. It is extremely important to realize that, at least in the short to intermediate term, the resources committed by a single routing decision can be very significant: The equipment tied up by a single coast-to-coast OC-192 can easily have a first cost of $10**6, and the holding times on a circuit once established is likely to be measured in months. Carriers will expect the routing algorithms used to be sensitive to these costs. Simplistic measures of cost such as the number of "hops" are not likely to be acceptable.

Taking the case of an all-optical island consisting of an "ultra long-haul" system embedded in an OEO network of electrical fabric OLXC's as an example: It is likely that the ULH system will be relatively expensive for short hops but relatively economical for longer distances. It is therefore likely to be deployed as a sort of "express backbone". In this scenario a carrier is likely to expect the routing algorithm to balance OEO costs against the additional costs associated with ULH technology and route circuitously to make maximum use of the backbone where appropriate. Note that the metrics used to do this must be consistent throughout the routing domain if this expectation is to be met.

2.3.2 End System Capabilities

While properties of the subnetwork are very important when trying to decide which subnetwork to use to access a system (in the case of multi-homing), end systems also posses a wide variety of capabilities. Throwing end system capabilities such as a systems

ability to support SONET/SDH virtual concatenation for distribution into a routing protocol seems inappropriate since it counters the ability to summarize. If detailed end-system information is needed by another end system then a directory service or some type of

direct query between the end systems that does not impact the network seems more appropriate.

2.4 Diversity in Optical Routing

There are two basic demands that drive the need to discover diverse routes for establishing optical paths:

1. Reliability/Robustness
2. Bandwidth capacity.

Many times multiple optical connections are set up between the same end points. An important constraint on these connections is that they must be diversely routed in some way [5]. In particular they could be routed over paths that are link diverse, i.e., two connections do not share any common link. Or the more stringent constraint that the two paths should be node diverse, i.e., the two paths do not traverse any common node.

Additionally, insufficient bandwidth may exist to set up all the desired connection across the same path (set of links) and hence we need to know about alternative (diverse) ways of reaching the destination that may still have unused capacity.

"Diversity" is a relationship between lightpaths. Two lightpaths are said to be diverse if they have no single point of failure. In traditional telephony the dominant transport failure mode is a failure in the interoffice plant, such as a fiber cut inflicted by a backhoe.

Data network operators have relied on their private line providers to ensure diversity and so IP routing protocols have not had to deal directly with the problem. GMPLS makes the complexities handled by the private line provisioning process, including diversity, part of the common control plane and so visible to all.

Diversity is discussed in the IPO WG document [6]. A key associated concept, "Shared Risk Link Groups", is discussed in a number of other IETF (refs) and OIF (refs) documents. Some implications for routing that are drawn in [6] are:

- Dealing with diversity is an unavoidable requirement for routing in the optical layer. It requires dealing with constraints in the routing process but most importantly requires additional state information the SRLG relationships and also the routings of any existing circuits from which the new circuit is to be diverse to be available to the routing process.

- At present SRLG information cannot be self-discovered. Indeed, in a large network it is very difficult to maintain accurate

SRLG information. The problem becomes particularly daunting

whenever multiple administrative domains are involved, for instance after the acquisition of one network by another, because there normally is a likelihood that there are diversity violations between the domains. It is very unlikely that diversity relationships between carriers will be known any time in the near future.

- Considerable variation in what different customers will mean by acceptable diversity should be anticipated. Consequently we suggest that an SRLG should be defined as follows: (i) It is a relationship between two or more links, and (ii) it is characterized by two parameters, the type of compromise (shared conduit, shared ROW, shared optical ring, etc.) and the extent of the compromise (e.g., the number of miles over which the compromise persisted). This will allow the SRLG s appropriate to a particular routing request to be easily identified.

2.4.1 Generalizing Link Diversity

Optical networks may posses a number of hierarchical signaling layers. For example two routers interconnected across an optical network may communicate with IP packets encapsulated within an STS-48c SONET path layer signal. Within the optical network this STS-48c signal may be multiplexed at the SONET line layer into an OC-192 line layer signal. In addition this OC-192 may be wavelength division multiplexed onto a fiber with other OC-192 signals at different wavelengths (lambdas). These WDM signals can then be either lambda switched, wave band switched or fiber switched. Hence when we talk about diversity we need to specify the layer to which we are referring. In the previous example we can talk about diversity with respect to the SONET line layer, wave bands, and/or optical fibers. A similar situation arises when we consider the definition of node diversity. For example are we talking with respect to a SONET path layer switch or an optical switch or multiplexer?

The Shared Risk Link Group concept in reference [7] generalizes the notion of link diversity (general list of numbers). First it's useful with respect to major outages (cable cuts, natural disasters) to have a few more types of diversity defined:

1. Cable (conduit) diversity (allows us to know which fibers are in the same cable (conduit). This helps avoid sending signals over routes that are most vulnerable to "ordinary" cable cuts (technically known as backhoe fades).

2. Right of Way (ROW) diversity. This helps avoid sending signals over routes that are subject to larger scale

disasters such as ship anchor drags, train derailments, etc.

3. Geographic Route diversity. This type of diversity can help one avoid sending signals over routes that are subject to various larger scale disasters such as earthquakes, floods,

tornadoes, hurricanes, etc. A route could be approximately described by a piecewise set of latitude/longitude or UTM coordinate pairs.

We also have a form of link abstraction/summarization via the link bundling concept [8].

2.4.2 Generalizing Node Diversity

The concept of a node abstraction associated with GMPLS appears in reference [14] where it is used to generalize the concept of an explicitly routed path. In this case an abstract node can be a set of IP addresses or an AS number. From the point of view of node diverse routing specific concepts of interest include:

1. Nodes, i.e., individual switching elements.
2. Switching centers, i.e., a central office or exchange site.
3. Cities, or towns that contain more that one switching center.
4. Metro areas, or counties
5. States,
6. Countries, or
7. Geographic Regions

For example, although rumors of California's eventual slide into the Pacific Ocean have been greatly exaggerated, some telecommunications customers might prefer their Asia-bound traffic to egress at diverse US west coast locations such as Washington State, Oregon and/or California.

3 Applications of Optical Inter Domain Routing

3.1 Inter-Area Routing

Inter-area routing refers to a situation where the network that is to be partitioned into areas is under the control of one administrative entity. The main reasons for this partitioning in optical networks stem from scalability, inter-vendor interoperability, legacy equipment interoperability, and inter-layer partitioning.

3.1.1 Inter-Area Scalability

As networks grow it is useful to partition a routing domain into areas where limited or summarized information is shared between areas. This reduces the overhead of information exchange across the network as a whole, and reduces the convergence time of routing protocols within a particular area.

When the topology within the area is approximated then signaling and call processing at the area border must specify an approximated (loose) route and the border node must then translate this to a precise route through the area. Hence there is some linkage between

multi-domain connection control and inter-area/inter-domain routing.

Notes: This might also be valid in a multi domain case where there is trust between domains. This might arise, e.g., after one network

is acquired by another but not yet physically integrated; or between Metro and Core providers that are closely tied. One definition needing further refinement is that of "administrative entity" in the ON case.

3.1.2 Inter-vendor Inter-area

Another example occurs when interoperability between two different optical vendors is desired. Vendors may use different protocols as the primary option between their own devices, adding specialized features or optimizing their performance based on their choice of protocol. Although one option is to force both vendors to adopt a new common protocol another is to only require a minimum subset of reachability/topology information to be shared between them.

Notes: A common model is that carriers tend to buy clusters of equipment from a common vendor. For example, it is unlikely that there will be a mixture of XXX, YYY, and ZZZ optical switches in the same subnetwork.

3.1.3 Legacy Interoperability Inter-area

A very important subcase of inter-vendor/inter-area is where some optical subnetworks (read: lots of existing installations) may not run a routing protocol at all, e.g., they rely strictly on EMS-based topology discovery/resource management. In this case it may be necessary to establish a "route proxy" to represent the sub-network and allow for interoperability with other subnetworks. Key in this case is the fact that we can't get the network elements in these subnetworks to run a distributed route protocol. However, we can have a separate software entity with access to the appropriate information, proxy routing information for this entire subnetwork.

The basic advantage here is that even though the vendor specific element management system (EMS) knows the topology of its subnetwork, it is better that information be exchanged automatically between adjoining areas (to avoid errors) via a neighbor discovery/link verification protocol such as those suggested in LMP [9], OIF-UNI [10], or G.disc [11]. Now these protocols will furnish basic node and port mapping information between the neighbor pairs but will need to supply additional information to let us know that these two elements belong to separate "vendor areas".

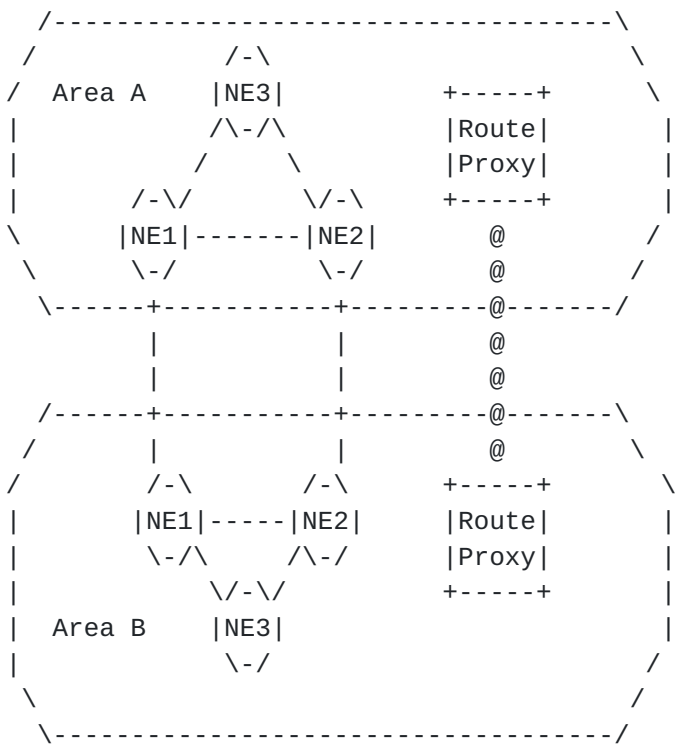

Figure 3-1: Route Proxy

Figure 3-1 shows an example of two areas with inter-connected NEs. Assume that neither of these areas runs a distributed routing protocol or desires to expose the details of its topology. Instead they may exchange routing proxy addresses through the neighbor discovery protocol, and then exchange routing information between route proxies. The functions of the route proxy would include: (a) direct reachability exchange -- what NEs can be reached directly from this area --, (b) verification of area connectedness -- how the two areas are inter-connected should be understood by both -- (also other areas), (c) area topology exchange and updates (possibly summarized topology), and (d) topology updates concerning other areas.

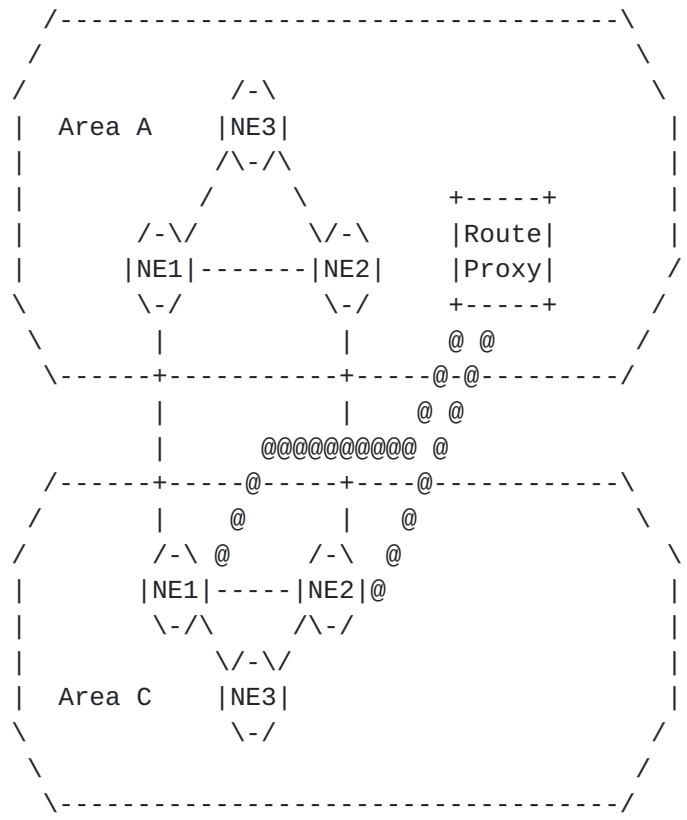

Figure 3-2: Route Proxy to Distributed Case

Figure 3-2 shows a case where one area runs a route proxy and the other area runs a distributed routing protocol. Once again a neighbor discovery procedure between area border NE's could be used to advertise route proxy address.

Flooding and summarization mechanisms could be applied by the route proxy as if it is a switching system. Since this is optical rather than IP routing, signaling would be carried by a control channel between the route proxy and the neighboring system, rather than being carried over the data link.

3.1.4 Inter-Layer Partitioning

In this situation the entities to be included in the route protocol all fall within the same administrative domain. However, the network is partitioned into sub-networks that operate at different switching layers. Not all the information from one layer is necessary or relevant to another layer. Hence, in this case, the flow of routing information between the layers may be asymmetric and also summarized. For example, between transparent optical switches and SDH/SONET path (VC) layer switches, not all the information at the

SONET layer is relevant to the optical layer. In addition optical networks may keep a lot more physical layer information (such as the properties of every optical amplifier on a WDM span) that is of no use to the SONET layer. One again this promotes scalability, but

also simplifies the implementation by reducing inter-layer information transfer to that which is actually useful.

Let us look at the kind of information that a lower network layer could make use of from its client (upper layer) subnetworks. In deciding where to place subnetwork connections in a given layer network it is very useful to have a view of the current higher layer traffic matrix [12] being satisfied and higher layer traffic trend measurements over time. Although we can somewhat see this in higher layer resource status changes over time, this represents a link level view when we really desire the trend (change in time) of the traffic matrices between sites. How this information gets distributed is an open issue. Currently individual nodes in a GMPLS network know only about connections that they source or sink.

Now looking the other way is initially simpler, i.e., it is easier to ask: what can a higher layer use for path selection from a lower layer. The first item that springs into mind is diversity information. Note from earlier discussions that there may be multiple layers of diversity information. For example in setting up a SONET STS-1 path we can talk about SONET line layer diversity but also about WDM fiber diversity. Other types of information maybe useful to share but may be very layer specific.

3.2 Classical Inter-Domain (Inter-Carrier)

In this case we are talking about dealing with outside entities, i.e., between service providers. There may be a range of levels of trust here; for example there might be some level of trust between two providers that have formed a marketing alliance or have some other form of business relationship. In general, however, trust can not be assumed. In this case, all the concerns of revealing too much information about one's network come into play. However, not revealing enough, say about diversity capabilities may also lead customers elsewhere. Also there are some other security issues not seen before. For example, in route distribution one carrier might not be inclined to pass on routing information that could point the way to competitive alternatives. This impacts the methods for route updates, etc.

With the interest in bandwidth trading [13] we can also look at this as an advertisement of network connectivity and capability with of course any "warts" covered up. This would include reliance on other carrier for fibers or lambdas. Also a fair amount of details such as "unused capacity" would not be advertised since this maybe financially sensitive information.

Private line pricing today is based primarily on the service itself (bandwidth, end-points, etc.) and the holding time, and there is no reason to expect that this will change. When multiple service providers are involved the algorithm for dividing up the revenue

stream (which can be quite large even for a single connection) must be explicit by connect time. This could be done off-line or could be done at connect time. In either case, the entity or entities doing the routing will need to take provider pricing structures into account whenever there is a choice between providers that needs to be made. The routing logic could do this explicitly if the prices are captured in the advertised metrics or some other advertised data; alternatively it could be done by some sort of policy control, as it is today by BGP.

The essence of bandwidth trading is the existence of competing price structures that are known to the entity deciding which competitor to use. It is possible to create plausible bandwidth trading scenarios involving the UNI, the NNI, or both. If the NNI is involved, these price structures will need to be established across it. The situation is further complicated by the fact that bandwidth trading could be realized using any one of a number of business models, each with its own information requirements. To give two examples: If an auction model were used the buyer might repeatedly broadcast the lowest bid received to date and solicit lower bids from the competing providers. On the other hand, if there were a more formal market the providers might post their asking prices in some public fashion and a buyer would be matched by some third party with the lowest offer.

In the inter-carrier case notions of hierarchy seem rather sensitive, i.e., he who controls the summarization and advertisement may have an undue advantage over competitors. In addition, a "bandwidth aggregator" may want to advertise capabilities that he has put together via deals with multiple carriers...

Notes: We can attempt to extend the SRLG concept to links between ASs but we will need the two ASs to agree on the meaning and number of the list of 32 bit integers that comprise the SRLG, i.e., previously the SRLG concept was one of AS scope. And this is also where things get tricky since it may not be possible to distinguish diverse routes based upon differing path vectors (i.e., AS number traversal list). The reason for this is due the fact that many carriers "fill out" their networks by renting either dark fiber or "lambdas" from a WDM system and hence although the path vectors may be AS diverse they may not even be fiber diverse.

Hence there is a need for sharing of diversity information or constraints between ASs when setting up diverse connections across multiple ASs. This gets us somewhat into a quandary over which information needs to be public and how to coordinate its distribution. In this sense geographic link information may be the simplest and least contentious to get various players to disclose

and standardize.

Notes: (1) The real issue is consistency between the cloud/AS s since in many cases they are sharing conduit, ROW, etc. Getting this to happen could be very problematic. It would be preferable to

see a diversity option that doesn t require this. For example, ensure that there is diversity within each cloud and then do restoration separately within each cloud. (2) See the definition of SRLG in the Carrier Requirements an equivalence class of links, the extent of violation, and the level. (3) Flexibility in defining the level of violation seems very desirable these historically have drifted in time. There are many others eg, if the shared resources are SPRING protected that s less of a problem than otherwise.

Notes: Participation in the inter-domain network carries constraints on the carriers. First, in order to participate, each provider network MUST be willing to advertise the destinations that are reachable through his network at each entry point and advertise the formats available. Without providing such information, there is little motivation to participate since it is unlikely that others will be able to access services of which they are not aware. Second, every participating carriers MUST agree to fairly include the information made available by every other carrier so that each carrier has an equal opportunity to provide services. There may be specific exceptions, but the carrier claiming those exceptions MUST advertise the exceptions themselves. In this manner, other carriers that might otherwise be aware of distant services can be prompted to seek those services manually. Note a combination of minimal required information transferred with deferral to the originating subnetwork along with some basic security mechanisms such as integrity and non-repudiation may be useful in helping organizations to "play nice".

3.3 Multi-Domain Connection Control

MPLS loose routing capability allows one to specify a route for an optical connection in terms of a sequence of optical AS numbers. This, for example, is handled via RSVP-TE s abstract node concept [14]. Currently there is nothing in the GMPLS signaling specification that differentiates between intra AS boundaries, i.e., between two neighbor optical LSRs in the same AS, and inter AS boundaries, i.e. between two neighbor optical LSRs in different ASs. Note that these same notions can apply to separate routing domains within an AS. There may, however, be some useful reasons for differentiating these two cases:

1. Separation of signaling domains,
2. Separation of protection domains.

While routing protocols (used for their topology information) in the optical case are not "service impacting", signaling protocols most certainly are. It is desirable to build some type of "wall" between

optical ASs so that faults in one that lead to "signaling storms" do not get propagated to other ASs. Note that the same motivation applies for isolating other kinds of clouds, like vendors specific ones.

The natural situation where "signaling storms" would be most likely to arise is during network restoration signaling, i.e., signaling to recover connections during major network outages, e.g., natural disasters etc. In this case it may be very advantageous to break up general source reroute forms of restoration into per domain segments or to start reroute at domain boundaries rather than all the way back at the originating node. Note that this has the advantage of reducing the need for globally consistent SRLG s. (See earlier SRLG comment.) Such a capability requires some loose coordination between the local, intermediate and global protection mechanisms [15]. This is typically implemented via hold off timers, i.e., one layer of protection will not attempt restoration until a more fundamental (local) form has been given a chance to recover the connection [15].

In other words, prevention of restoration related signaling storms may require the breaking up of a large network into multiple signaling (and hence routing) domains. These domains could be within the same AS.

4 Multiple Layers of Routing

4.1 Layers in Transport Networks

In transport networks layering is a part of the multiplex and OA&M structure of the signals, playing a role in multiplexing, monitoring and general link management. Layering in the transport network is defined in fairly abstract terms in [G.805] and the concepts are applied to SDH in [G.803]. As explained in a recent ITU SG15 document (WD45 Q.14/15) not all the layers in the transport network are of interest to the control plane, or to routing in particular.

Some layers may not contain active switching elements, however this does not mean that information flow concerning a non-switching layer is not valuable in routing. For example in [GB-WDM-SRLG] static WDM layer information was used to set the SRLGs for SONET lines (i.e., information passed around by a link state protocol operating at the SONET line layer). It should be noted that much of the information available from non-switching layers relates to performance monitoring and fault management. Hence work in this area within CCAMP should take into account this layered approach.

Note that this is distinct from the layer idea used in the 7-layer OSI model or IP layer model. In the IP model, the term Layer means that, for example, the Application Layer entity requests services for delivering a message to an entity on another computer and it contacts the Transport Layer service, which in turn contacts the Internet Layer. Lower layers are successively contacted until an end-to-end service is provided. A key concept is that the

Application Layer cannot (or rather should not) contact the Internet Layer directly. In this model all the "layers" discussed in this document would lie in the "physical layer" (from an IP perspective).

For concreteness we first give a overview of routing at two of the various layers of interest in the optical network, transparent optical and SONET/SDH. We then discuss information sharing between layers in general and with the IP layer in particular.

4.2 Optical Physical Layer Routing

Routing in the optical layer is in general more than just finding a path that has the available wavelength. Besides possible distance and cost optimization and the diversity requirements as described in section 4.2, there are constraints arising from the design of new software controllable network elements as well as constraints in domains of transparency, i.e., all optical networks. Here, we summarize the main constraints in the two categories. See reference [16] for more detailed discussions.

4.2.1 Reconfigurable Network Elements

Besides OLXCs, there are other software reconfigurable elements on the horizon, specifically tunable lasers and receivers and reconfigurable optical add-drop multiplexers (OADM s). These elements are illustrated in the following simple example, which is modeled on announced Optical Transport System (OTS) products:

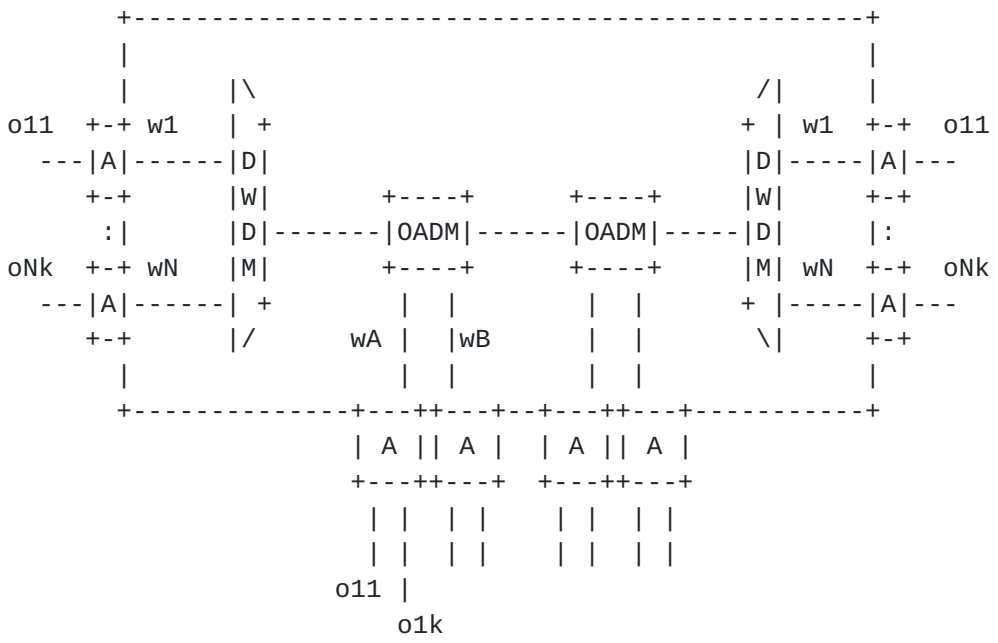

Figure 4-1: An OTS With OADM's - Functional Architecture

In Fig. 4-1, the part that is on the inner side of all boxes labeled "A" defines an all-optical subnetwork. Boxes labeled "A" provide

adaptation function that transform the incoming optical channel into the physical wavelength to be transported through the subnetwork as well as possible multiplexing function using either electrical or

optical TDM. These may result in the following constraints on routing:

- The adaptation function may force groups of input channels to be delivered together to the same distant adaptation function.
- Only adaptation functions whose lasers/receivers are tunable to compatible frequencies can be connected.
- The switching capability of the OADM s may also be constrained. For example:
  - o There may be some wavelengths that can not be dropped at all.
  - o There may be a fixed relationship between the frequency dropped and the physical port on the OADM to which it is dropped.
  - o OADM physical design may put an upper bound on the number of adaptation groupings dropped at any single OADM.

4.2.2 Wavelength Routed All-Optical Networks

The optical networks presently being deployed may be called "opaque" [17]: each link is optically isolated by transponders doing O/E/O conversions. They provide regeneration with retiming and reshaping, also called 3R, which eliminates transparency to bit rates and frame format. These transponders are quite expensive and their lack of transparency also constrains the rapid introduction of new services. Thus there are strong motivators to introduce "domains of transparency" - all-optical subnetworks - larger than an OTS, where signal passes through the domain optically. There are two unique types of constraints on routing: one that is due to limited (or no) wavelength conversion, and the other that is due to physical impairments.

Within an all-optical domain, "wavelength conversion" (changing the wavelength of a connection) is still expensive and not yet practical without an OEO conversion. Therefore it is important to understand the routing implications of limited (or no) wavelength conversion. This requires us to look at what is called the "Routing and Wavelength Assignment (RWA) Problem" [18]: Given one or more connections that need to be established in an all-optical domain, determine the routes over which each connection should be routed and also assign each connection a color. If the routes are already known, the problem is called the "Wavelength Assignment (WA) Problem".

As domains of transparency get larger and bit rates of 10 Gb/sec and higher become common physical impairments including amplifier spontaneous emission (ASE), polarization mode dispersion (PMD), and others may become a routing issue. We consider a single domain of transparency. Additionally due to the proprietary nature of DWDM

transmission technology, we assume that the domain is either single vendor or architected using a single coherent design philosophy, particularly with regard to the management of impairments. Specifically:

- ASE noise which accumulates imposes a limit on the maximum number of spans for the transparent segment in a lightpath, which is bit rate dependent: the higher the bit rate the fewer the spans. A span refers to a segment between two optical amplifiers.
- PMD imposes a limit on the maximum transmission distance for the transparent segment that is inversely proportional to the square of the bit rate of the signal. For typical installed fibers the limits are 400km and 25km for bit rates of 10Gb/s and 40Gb/s, respectively. With newer fibers assuming PMD parameter of 0.1 ps/.km, the limits are 10000km and 625km, respectively.
- Crosstalk and effective passband narrowing due to filtering effects can be treated approximately as a constraint on the maximum allowable number of OADMs/OXCs in the transparent segment of the lightpath.
- Other impairments including chromatic dispersion, nonlinear impairments are assumed to be treated at the transmission system level and/or as additional system margin on OSNR (optical signal to noise ratio).

4.2.3 More Complex Networks

An optical network composed of multiple domains of transparency optically isolated from each other by OEO devices (transponders) is more plausible. A network composed of both "opaque" (optically isolated) OLXC's and one or more all-optical "islands" isolated by transponders is of particular interest because this is most likely how all-optical technologies are going to be introduced. We now consider the complexities raised by these alternatives.

The first requirement for routing in a multi-island network is that the routing process needs to know the extent of each island. There are several reasons for this:

- When entering or leaving an all-optical island, the regeneration process cleans up the optical impairments discussed.
- Each all-optical island may have its own bounds on each impairment.
- The routing process needs to be sensitive to the costs associated with "island-hopping".

The first-order implications for GMPLS seem to be:

- Information about island boundaries needs to be advertised.
- The routing algorithm needs to be sensitive to island transitions and to the connectivity limitations and impairment constraints particular to each island.
- The cost function used in routing must allow the balancing of transponder costs, OXC and OADM costs, and line haul costs

across the entire routing domain.

Several distributed approaches to multi-island routing seem worth investigating:

. Advertise the internal topology and constraints of each island globally; let the ingress node compute an end-to-end strict explicit route sensitive to all constraints and wavelength availabilities. In this approach the routing algorithm used by the ingress node must be able to deal with the details of routing within each island.
. Have the EMS or control plane of each island determine and advertise the connectivity between its boundary nodes together with additional information such as costs and the bit rates and formats supported. As the spare capacity situation changes, updates would be advertised. In this approach impairment constraints are handled within each island and impairment-related parameters need not be advertised outside of the island. The ingress node would then do a loose explicit route and leave the routing and wavelength selection within each island to the island.
. Have the ingress node send out probes or queries to nearby gateway nodes or to an NMS to get routing guidance.

4.3 SDH/SONET layer Routing

An overview of link state intra domain routing applied to SONET/SDH networks can be found in reference [3]. We will give a very short review here with an emphasis on the multiple-layer aspects.

4.3.1 Switching Capabilities

The main switching capabilities that characterize a SONET/SDH end system and thus get advertised into the link state route protocol are: the switching granularity, supported forms of concatenation, and the level of transparency.

4.3.2 Switching Granularity

The signals switched in SONET/SDH can be divided in to two main categories: lower order signals and higher order signals as shown in Table 2.

Table 2. SDH/SONET switched signal groupings.

Signal Type	SDH	SONET
Lower Order	VC-11, VC-12, VC-2	VT-1.5 SPE, VT-2 SPE, VT-3 SPE, VT-6 SPE
Higher Order	VC-3, VC-4	STS-1 SPE
	VC-4-Xc (concatenated)	STS-Nc SPE (concat.)

For transport across a SONET network the lower order signals must be

multiplexed into a non-concatenated higher order signal. Hence a higher order "connection" is required between "lower order" switches before the lower order traffic can be switched.

A network element capable of switching one type of lower order signal is not required to support switching of all the other types of lower order signals and a similar notion holds true for the higher order signals. Hence there is a need to distribute the switching capabilities (granularity) of a node on one end of a link.

4.3.3 Protection

SONET and SDH networks offer a variety of protection options at both the SONET line (SDH multiplex section) and SONET/SDH path level. This means that we can protection mechanisms directly for the lower order or higher order switching layers defined in the previous section, i.e., the SONET line (SDH MS) techniques protect the higher order signals on a per line layer link basis. While the path layer protection mechanisms protect either the lower order signals on a per higher order link basis or the higher order signals on a subnetwork connection basis.

Standardized SONET line level protection techniques include Linear 1+1 and Linear 1:N automatic protection switching (APS) and both two-fiber and four-fiber bi-directional line switched rings (BLSRs). At the path layer, SONET offers uni-directional path switched ring protection. Both ring and 1:N line protection also allow for "extra traffic" to be carried over the protection line when that line is not being used, i.e., when it is not carrying traffic for a failed working line. These protection methods are summarized in Table 5.

Table 5. Common SONET/SDH protection mechanisms.

Protection Type	Extra Traffic Optionally Supported	Comments
1+1 Unidirectional	No	Requires no coordination between the two ends of the circuit. Dedicated protection line.
1+1 Bi-directional	No	Coordination via K byte protocol. Lines must be consistently configured. Dedicated protection line.
1:1	Yes	Dedicated protection.
1:N	Yes	One Protection line shared by N working lines.

4F-BLSR (4 fiber bi-	Yes	Dedicated protection, with alternative ring path.

directional line switched ring)		
2F-BLSR (2 fiber bi-directional line switched ring)	Yes	Dedicated protection, with alternative ring path
UPSR (uni-directional path switched ring)	No	Dedicated protection via alternative ring path. Typically used in access networks.

It may be desirable to route some connections over lines that support protection of a given type, while others may be routed over unprotected lines, or as "extra traffic" over protection lines. Also to assist in the configuration of these various protection methods it can be extremely valuable to advertise the link protection attributes in the route protocol. For example suppose that a 1:N protection group is being configured via two nodes. One must make sure that the lines are "numbered the same" with respect to both end of the connection or else the APS (K1/K2 byte) protocol will not operate correctly.

4.3.4 Available Capacity Advertisement

Internal to each SDH/SONET LSR interface, a table is maintained indicating each signal allocated in the multiplex structure. This internal table is the most complete and accurate view of the link usage and available capacity.

This information needs to be advertised in some way to all the other SONET/SDH switches/multiplexers in the same domain for use in path computation. There is a trade off to be reached concerning: the amount of detail in the available capacity information to be reported via a link state routing protocol, the frequency or conditions under which this information is updated, the percentage of connection establishments that are unsuccessful on their first attempt, the extent to which network resources can be optimized. There are different levels of summarization that are being considered today for the available capacity information. At one extreme all signals that are allocated on an interface could be advertised, or on the other extreme, a single aggregated value of the available bandwidth could be advertised. It makes the most sense to keep at least the bandwidth reporting for the lower order and

higher order signals separate since these are working at different layers in the multiplex hierarchy.

Consider first the relatively simple structure of SONET and its most common current and planned usage. DS1s and DS3s are the signals most

often carried within a SONET STS-1. Either a single DS3 occupies the STS-1 or up to 28 DS1s (4 each within the 7 VT groups) are carried within the STS-1. With a reasonable VT1.5 placement algorithm within each node it may be possible to just report on aggregate bandwidth usage in terms of number of whole STS-1s (dedicated to DS3s) used and the number of STS-1s dedicated to carrying DS1s allocated for this purpose. This way a network optimization program could try to determine the optimal placement of DS3s and DS1s to minimize wasted bandwidth due to half-empty STS-1s at various places within the transport network.

Similarly consider the set of super rate SONET signals (STS-Nc). If the links between the two switches support flexible concatenation then the reporting is particularly straightforward since any of the STS-1s within an STS-M can be used to comprise the transported STS-Nc. However, if only standard concatenation is supported then reporting gets trickier since there are constraints on where the STS-1s can be placed.

4.4 Layer Integration

As previously discussed, there are multiple layers of signals included in what in the IP model one would call the Physical Layer. One could separate the layers by creating sublayers in the Physical Layer. For example, sublayers in the Physical Layer might be, top to bottom: LOVCs, HOVCs, and Lambdas. If a system supports only one of the three, then isolation of the sublayers is a given; it's geographical. But there are systems which will support more than one physical sublayer, therefore, it is necessary to establish whether or not there is a need to isolate the sublayers in the same manner. Or put another way is there a reason to "integrate" the sublayers for the purposes of routing (topology dissemination).

If they are isolated, then there will be separate topological models for each sublayer: one mesh for the LOVC, one for the HOVC, one for the Lambda, and possibly others. The appropriate way to access a sublayer is via the use of sublayer SAPs (service access points). For example, in this way, one may find that use of Lambdas is more efficient because each sublayer can assess the availability of services at its own layer before searching for coarser-granularity services. On the other hand, the control plane must accommodate three separate routing protocols, or at least three separate instances of the same routing protocol, all operating at both intra and inter-domain level.

Section 4.4.2, herein, states "For transport across a SONET network, the lower order signals must be multiplexed into a non-concatenated

higher order signal." Given that this is true, LOVCs are not routed independently, but only as tributaries of HOVCs. In addition in the SDH hierarchy there is a signal, VC3, that can be treated (multiplexed) as either a LOVC or a HOVC. With this tight and

somewhat confused coupling of these layers it may beneficial to sometimes combine them into the same route protocol instance.

Use of the terms LOVC and HOVC infers that all of the services to be supported by inter-domain routing are those formally associated with the terms in SONET and SDH standards. However, among the optical systems emerging in today s market are rate and format independent systems, which claim to offer services that do not rely on SONET/SDH framing. Their intent is to support Ethernet, ATM, and OTN framing without the need for electronics specifically targeted at the signal of interest. The question arises whether or not to include these "clear channel" services as a separate sublayer of the Physical Layer.

The alternative to separate routing protocols per sublayer is the original notion behind GMPLS routing and the forwarding adjaciency concept [19]. Rather than separating the route protocols into separate layers (or sublayers) with distinct topologies, each ONE would advertise the services it can provide, along with its topology information. For example, a ONE (optical network element) might advertise that it carries a route to node A with STS-N service and clear-channel lambda service and carries multiple routes to node B with STS-N service. It might, alternatively, advertise its entire network with summarized link capacity information for every included link. Neighboring carriers would, implicitly, be allowed to summarize that information for internal advertisement via its IGP. Further consideration could be given to a query service, where a carrier advertises the geographical area it serves without detailed reachability or capacity information. A second carrier desiring service could query the first carrier as to reachability for a specific destination, and the first carrier would respond with availability and capacity information.

Integrating multiple layers into the same routing protocol instance leaves us fewer routing protocols to manage. The downside of this is that more information must be exchanged via this routing protocol and more network elements participate in this single instance of the routing protocol which can lead to scalability concerns. If the equipment working on the different sublayers comes from different vendors there would be little incentive to integrate multiple layers into the routing protocol for a single layer product.
Regardless of whether multiple layers are integrated into the same routing protocol instance it can be very useful to share information between layers as illustrated by the following examples:

- Drop side links between layers: Capabilities of the links that are between the (client and server) layers need to be

propagated into the routing protocol.

- Summarize link capabilities: Summarizing the server layer capabilities in the client layer will reduce the amount of information required for multi-layer constraint based path computation.

- Send only that are required: Sending only the capabilities that are useful in the constraint path computation in the client layer.

4.5 Interaction with IP Layer Routing

The applicability of IP-based routing protocols has, over the years, been constantly expanded to increasingly more circuit-oriented layers. The community began with pure datagram routing, gradually expanded to cover virtual-circuit switched packet routing (for e.g., MPLS), and is finally looking at the application of routing protocols to real circuit switching, e.g. the optical layer.

However, as pointed out earlier in this document, it is not clear that the different layers should necessarily share the same instance of the IP routing protocols. Indeed, there may be significant reasons for not doing so. For example, IP-layer reachability information is not particularly useful for the optical layer, so it seems an overkill to burden the optical equipment with storing and distributing that information. (It is an extra expense on memory and processing for information that the optical layer does not really care about, so there is little incentive for a vendor to want to do so.) Likewise, information on physical plant (fibers, conduits, ducts) diversity, which is crucial at the optical transport layer, is very unlikely to be used directly by the IP layer. So, it would be quite wasteful of resources to burden the IP layer routing with distributing and manipulating this information.

Thus, the extent of interaction or integration with IP layer routing (if any) requires careful consideration.

5 Existing Routing Protocol Applicability

Here we look at the applicability of OSPF, PNNI and BGP to various aspects of the general optical inter domain routing problem. All protocols provide reachability information. The questions to be investigated are how they deal with partitioning the network, diverse routing, summarized/abstracted topology information sharing, and suitability for the inter-carrier environment.

5.1 OSPF Applicability

[THIS SECTION IS UNDER CONSTRUCTION]
Notes: Interested here in OSPF areas their capabilities and limitations.

For example in OSPF [RFC2328, section 3] no topology information is shared between areas only summarized address information. A key

property of OSPF areas is that all areas must be attached to the "backbone" area in some manner, either via physical links or virtual links.

How much topology information gets lost in the virtual link case? An ABR connected to the backbone via a virtual link will know the topology of the backbone (via virtual link) and the (at least) two non-backbone areas that it is connected to. By being connected to the back bone we find out reachability. Each ABR advertises its directly attached areas into the backbone. Note that it doesn't seem possible to discern the area structure from the summary LSAs. Notes: we could obtain via a backbone router on the network info to identify all the ABRs [Check this] then go to each of these ABRs get their link state route tables and put together a complete picture of the network. It would, however be nice to get updates from the areas as to major changes (i.e., bandwidth info) without polling.

Notes: draft-kompella-mpls-multiarea-te-01.txt has some analysis. They look at the issue of who knows what, the fact that ABRs know about topology of all areas that they connect, they use "crankback" like methods to pick another ABR in case of failure. They don't hit the diversity case.

5.2 PNNI Routing

The routing portion of ATM's Private Network-to-Network Interface (PNNI) [20] is a link state routing protocol like OSPF and IS-IS with a general inter-area hierarchy capability. We explore the characteristics of PNNI here because, as a link-state protocol, it was designed at the outset with several features that are attractive in a connection-oriented network:

1. Distribution of topology information,
2. Distribution of resource (bandwidth) status information,
3. Establishment of hierarchical groups,

In addition, PNNI routing was designed to work with PNNI signaling which is very similar in functionality to the traffic engineering capable forms of GMPLS (MPLS) signaling (label distribution protocols). In particular PNNI signaling is based upon:

1. Source routing (exact and loose forms),
2. Crankback and Alternate Routing.

A brief summary of PNNI routing capabilities and a brief assessment of its applicability to the Optical Inter-Domain Routing problem follows.

5.2.1 PNNI overview

PNNI routing supports the provisioning of the network into a hierarchy that can include up to 104 levels. The PNNI Hierarchy starts at the lowest level where the lowest-level nodes are organized into "peer groups", a generalization of OSPF's area concept. A peer group is a collection of nodes, each of which

exchanges information with other members of the group so that all members of the group maintain an identical view of the group. Each

peer group is identified with a "peer group identifier," which is provisioned when the network is configured. Each peer group elects a "peer group leader" through a system of provisioned priorities and arbitrated using the NSAP addresses of the nodes. This system is applied at each successively higher layer in the hierarchy. The peer group leader is also known as the "logical group node" and it represents the peer group in the next higher layer in the hierarchy.

Information is fed in both the upward and downward directions in the hierarchy. The logical group node feeds reachability and topology aggregation information upward. Reachability information includes a summary of the addresses that are reachable through this peer group. Topology aggregation includes the information needed to route into and through this peer group. In the downward direction, the logical group node feeds information that gives the lower-level nodes knowledge of how to route to all destinations reachable within the routing domain. Completion of the hierarchy is achieved by creating ever higher levels until the entire network is encompassed in a single peer group.

As discussed above, topology and reachability information is distributed throughout the network so that the every switch in the domain maintains a consistent picture of the network. Information is exchanged among the peer groups and between each peer group and the group immediately above it in the hierarchy. At the lowest level in the hierarchy, the nodes exchange Hello packets with immediate neighbors to determine local state information. Each node bundles the information it collects from Hellos into a "PNNI Topology State Element" (PTSE) and the PTSEs are reliably flooded throughout the peer group. The flooding continues until every node in the domain has a consistent picture of the network. As previously discussed, PTSEs are fed downward to the next lower level in the hierarchy.

Nodes actively taking part in PNNI routing are addressed using ATM End System Addresses, which are modeled after NSAP addresses. PNNI router addresses include a prefix that specifies the Peer Group Identifier. Peer group identifiers are encoded using 14 octets: a 1 octet level indicator followed by 13 octets of identifier information. The value of the level indicator must be between 0 and 104. The value set in the identifier information field must be encoded with the 104-n right-most bits set to zero, where n is the level. The identifier information is formatted left-to-right in a hierarchical manner. The structure of the hierarchy is defined by the peer group identifiers. Address assignment has a hierarchy that, for proper scaling, should generally correspond to the topological hierarchy. This will allow address summarization where an address

prefix represents reachability to all addresses that begin with the stated prefix. When summarizing reachable addresses for advertisement, addresses that are exceptions are described by longer prefixes.

PNNI signaling works with the topology and resource status information provided by PNNI routing via source route control. The precise name for this in PNNI signaling is a Designated Transit Lists (DTL), where the ingress switch decides the entire path across the PNNI routing domain. The DTL consists of a list of Node IDs and/or Port IDs traversing the peer group. To get across specific peer groups, the first node (a border node) in each peer group selects the specific path, in detail, across the local peer group. Since the ingress node uses currently available information, there are occasions when a path being processed according to a DTL may be blocked along the route. When a route cannot be processed according to the DTL, it is "cranked back" to the creator of that DTL, with an indication of a problem. This node may choose an alternate path for the route or it may crank the route back further.

5.2.2 PNNI Optical Applicability

As we saw PNNI routing has a general hierarchy and was designed to work with an explicit source routed signaling protocol. This resulted in a couple of key properties. First, no specific path computation is tied with PNNI routing or even included within the specification. Second, due to the fact that the routes must be computed outside of the routing protocol, even as we move up the peer group (generalized area) hierarchy, the link state nature of the protocol is preserved. In particular we still get topology and resource status information, albeit at an increasingly coarser level. Hence, information needed for diverse routing is still available even at the inter-area (higher order peer group) level.

Other items to note are the automated set up of control channels between peer group leaders, i.e., logical nodes at the next layer up in the hierarchy and a process where peer group leaders can be changed (due to failures or maintenance). Peer groups at a common hierarchical level can be connected arbitrarily; the notion of Area 0 does not apply in PNNI as it does in OSPF.

PNNI was not originally intended for use in an inter-carrier environment. It was originally intended for use in a private ATM network in a network-to-network or a node-to-network capacity. While PNNI provides summarization, it does not provide the means to "hide" the topology of a peer group. In addition, its topology summarization capability is limited to the "complex node

representation" which includes "exceptions". While this hub-spoke plus extensions is a bit better than an unstructured "blob" model, in the inter-carrier case we may wish to give more information than a single complex node but not have any of the "lower peer groups" actually share information. A typical example maybe the representation of the network in terms of city pairs visited and services offered, but not detailed link information (number of links between city pairs or available capacity).

5.3 BGP Applicability

The most basic functionality that we need to know about is which end systems are attached to each area or domain. In addition, we need at least one method for reaching each domain or area if we are to set up an optical circuit.

5.3.1 Pick One! (route that is)

With datagram routing we need to pick one route to a destination and make sure this choice is consistent throughout the AS. In particular BGP specifically reduces the number of choices according to the following rule [21]:

> Fundamental to BGP is the rule that an AS advertises to its neighboring AS's only those routes that it uses. This rule reflects the "hop-by-hop" routing paradigm generally used by the current Internet.

In the optical circuits case we are not using a "hop-by-hop" routing paradigm. Hence it seems that BGP constrains our knowledge of diverse routes in the optical case. This hits a major difference in use between the optical and IP datagram forwarding cases. In the optical case we are really interested in topology information that allows an optical connection path to be computed based on whatever criteria is desired for that connection. In the IP datagram case we are interested in a consistent set of routes for use in hop-by-hop forwarding. Hence the optical case has tended to favor link state protocols since they furnish raw topology information that can be used in computing routes as opposed to distance vector protocols whose output is a set of routes (without necessarily providing complete topology information).

5.3.2 Reachability: Via Optical BGP like functionality

BGP is "the" reachability protocol. The Update message contains a path (AS_PATH) that furnished at least one possible route to reach the destinations summarized (via prefixes) in the Network Layer Reachability Information (NRLI) field. Note that the NEXT_HOP attribute can be used in terms of the next optical hop (rather than

IP hop). Hence as it stands BGP can be used for arbitrary optical reachability. The BGP sessions are set up via the IPCC addresses (IP routable) but the information exchanged pertains to the optical network not the IP control channel network.

One of the first issues that arises in such an approach is that not all optical end systems are the same, i.e., they support different types of signals (TDM, lambdas, etc...). It has been suggested that the BGP communities attribute [RFC1997] can be used to differentiate optical equipment (end systems) with different types of termination capabilities. For example to differentiate a 10Gbps WAN interface (where the data is carried in SONET OC-192 framing) from an OC-192 interface terminating an STS-192c SONET signal (carrying POS or other types of payload such as GFP). This can also be used to differentiate the packet switch capable systems from the non-packet switch (hasn't Kireeti or Yakov written something on this with MPLS and BGP?)

Also can we use the communities attribute to indicate the path is also compatible with the end systems capabilities. What about an STS-3c granularity end system, would we want to indicate this via a communities attribute of some value and then have the AS_PATH attribute be a valid AS_PATH along switches of STS-3c granularity. Or consider the STS-1 granularity case.

Do we have a forwarding adjacency concept defined yet in the inter-domain case? For example a DWDM lambda connecting two SONET LTE boxes. Or an OEO-PXC-OEO type of switch connecting two SONET LTEs.

To promote optical interworking a common set of attributes and their meanings could be defined. [This seems like a good project. But how much are communities attributes used? And how is their meaning shared between ASs?]

5.3.3 Integrated with IP BGP?

Given the fairly modest initial demands of the emerging routing controlled optical network on a BGP implementation and its management, it is reasonable to ask if there is any benefit to integrate this optical layer routing information with the IP layer routing information? An optical subnetwork using the communities attributes to differentiate optical equipment from IP equipment and various types of optical equipment would just filter out all communities not of interest (right?). And hence the IP/Optical BGP integration would stop there? There isn't too much written on using communities attribute besides [RFC1998]?

5.3.4 Policy Mechanisms

BGP-4 [22] provides a number of policy mechanisms that relate to how routing information is used and disseminated. In particular the E-BGP border router model keeps distinct the routing information received from each of a border routers autonomous systems external

peers (Adj-RIBs-In -- Adjacent Routing Information Base In), the routing information that the Autonomous System (AS) itself is using (Loc-RIB -- Local Routing Information Base), and the routing information that the AS forwards onto its external peers (Adj-RIBs-Out -- Adjacent Routing Information Base Out). Via this model one

can develop policies with regards to which routes get chosen for use in the AS, i.e., which routes from the Adj-RIBs-In are chosen to populate the Loc-RIB. One also develops policies concerning what routing information gets advertised to external peers, i.e., which routes from Loc-RIB gets exported to each of the Adj-RIBs-Out.

The choice of which routes get imported for local routes generally is concerned with the "quality" of those advertising the routes since not too much else is known (besides the AS path vector). In deciding which routes to advertise to external peers "transit policies", i.e., whose traffic is allowed to transit this AS is the prime consideration.

In the MPLS and in particular the explicitly routed optical case we have a very strong additional policy mechanism, that of connection admission control (CAC). Although an optical AS probably shouldn t advertise transit capabilities that it doesn t wish to support, CAC during connection establishment will be the final arbiter of any transit policy. In addition, some areas that are being addressed by policies in the IP datagram case such as load balancing are much easier to implement via CAC and/or explicit routing.

Notes: Seems like a key choice is when policies are applied. One choice is CAC do it at connection establishment. This seems to force crankback however: If a request goes thru multiple AS s A->B->C-> and C doesn t do business with A, for example. Or with 2 domains A, B B might not want to let A use up last slot to some destination. This suggests that at least in this case policy could be applied by updated advertisements; if B decides it doesn t want any external AS using some particular link it advertises a changed connectivity or metric.

6 Conclusion

This draft highlighted some of the considerations for an inter-domain route protocol for use in optical internetworking. The main differences between optical routing and datagram routing were highlighted. Additional requirements to be addressed in an optical inter-domain route protocol were discussed and several applications of inter-domain routing were highlighted. A summary of optical sublayer specific routing information was furnished for both the transparent optical sublayer and the SONET/SDH sublayer. Finally a review of the applicability of several existing route protocols to the optical inter-domain route problem was given.

7 Security Considerations

Security considerations are not discussed in this version of the

document.

8 References

[1] Bradner, S., "The Internet Standards Process -- Revision 3", BCP 9, RFC 2026, October 1996.

[2] G. Bernstein, J. Yates, D. Saha, "IP-Centric Control and Management of Optical Transport Networks", IEEE Communications Magazine, October 2000.

[3] G. Bernstein, E. Mannie, V. Sharma, "Framework for MPLS-based Control of Optical SDH/SONET Networks", <draft-bms-optical-sdhsonet-mpls-control-frmwrk-01.txt>, July 2001.

[4] ANSI T1.105.01-1995, Synchronous Optical Network (SONET) Automatic Protection Switching, American National Standards institute.

[5] Ramesh Bhandari, Survivable Networks: Algorithms for Diverse Routing, Kluwer Academic Publishers,1999.

[6] Strand, J. (ed.) "Impairments And Other Constraints On Optical Layer Routing", work in progress, draft-ietf-ipo-impairments-00.txt, May 2001.

[7] Kompella, K., et. al. "IS-IS Extensions in Support of Generalized MPLS", Work in Progress, draft-ietf-isis-gmpls-extensions-01.txt, November 2000.

[8] K. Kompella, et. al. "Link Bundling in MPLS Traffic Engineering", Work in Progress, draft-kompella-mpls-bundle-05.txt, February 2001.

[9] J. Lang, et. al. "Link Management Protocol (LMP)", Work in Progress, draft-ietf-mpls-lmp-02.txt, March 2001.

[10] B. Rajagopalan (ed.), "User Network Interface (UNI) 1.0 Signaling Specification", OIF2000.125.5, The Optical Internetworking Forum, June 4th, 2001.

[11] T1X1.5-160 G.disc draft version 0.1.

[12] Robert S. Cahn, Wide Area Network Design: Concepts and Tools for Optimization, Morgan Kaufmann Publishers, Inc., 1998.

[13] Meghan Fuller, "Bandwidth trading no longer a case of 'if' but 'when' says report", Lightwave, June 2001. (www.light-wave.com)

[14] Awduche, D., et. Al., "RSVP-TE: Extensions to RSVP for LSP Tunnels", Work in Progress, draft-ietf-mpls-rsvp-lsp-tunnel-08.txt, February 2001.

[15] K. Owens, V. Sharma, M. Oommen, "Network Survivability Considerations for Traffic Engineered IP Networks", Work in Progress, draft-owens-te-network-survivability-01.txt, July 2001.

[16] A. Chiu, et. al., "Features and Requirements for The Optical Layer Control Plane" work in progress, draft-chiu-strand-unique-olcp-02.txt, February 2001.

[17] Tkach, R., Goldstein, E., Nagel, J., and Strand, J., "Fundamental Limits of Optical Transparency", Optical Fiber Communication Conf., Feb. 1998, pp. 161-162.

[18] Ramaswami, R. and Sivarajan, K. N., Optical Networks: A Practical Perspective, Morgan Kaufmann Publishers, 1998.

[19] K. Kompella and Y. Rekhter, "LSP Hierarchy with MPLS TE", draft-ietf-mpls-lsp-hierarchy-02.txt, Internet Draft, Work in Progress, February 2001.

[20] ATM Forum, "Private Network-Network Interface Specification (PNNI 1.0), Version 1.0," af-pnni-0055.000, March 1996.

[21] Rekhter, Y., and P. Gross, "Application of the Border Gateway Protocol in the Internet", RFC 1772, T.J. Watson Research Center, IBM Corp., MCI, March 1995.

[22] Rekhter Y., and T. Li, "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, T.J. Watson Research Center, IBM Corp., cisco Systems, March 1995.

9 Acknowledgments

10 Author's Addresses

Greg Bernstein, Lyndon Ong
Ciena Corporation
10480 Ridgeview Court
Cupertino, CA 94014
Phone: (510) 573-2237
Email: greg@ciena.com, lyong@ciena.com

Bala Rajagopalan
Tellium, Inc

2 Crescent Place
Ocean Port, NJ 07757
Email: braja@tellium.com

John Strand
AT&T Labs
200 Laurel Ave., Rm A5-1D06
Middletown, NJ 07748
Phone:(732) 420-9036
Email: jls@research.att.com

Angela Chiu
Celion Networks
1 Shiela Dr., Suite 2
Tinton Falls, NJ 07724
Phone:(732) 747-9987
Email: angela.chiu@celion.com

Frank Hujber
Alphion Corporation
4 Industrial Way West
Eatontown, NJ 07724
fhujber@alphion.com

Vishal Sharma
Metanoia, Inc.
335 Elan Village Lane, Unit 203
San Jose, CA 95134
Phone: +1 408 943 1794
Email: v.sharma@ieee.org

Sudheer Dharanikota
Nayna Networks Inc.
475 Sycamore drive,
Milpitas, CA 95035
Email : sudheer@nayna.com