) and MPLS Working Groups
:ernet Draft
›ires: October 2001
:ument: draft-osu-ipo-mpls-issues-02.txt
:egory: Informational

S. Seetharaman
Ohio State University
A. Durresi
Ohio State University
R. Jagannathan
Ohio State University
R. Jain
Nayna Networks
N. Chandhok
Ohio State University
K. Vinodkrishnan
Ohio State University

April 2001

IP over Optical Networks: A Summary of Issues

ıtus of this Memo

Abstract

This draft presents a summary of issues related to transmission of IP packets over optical networks. This is a compilation of many drafts presented so far in IETF. The goal is to create a common document, which by including all the views and proposals will serve as a better reference point for further discussion. The novelty of this draft is that we try to cover all the main areas of integration and deployment of IP and optical networks including architecture, routing, signaling, management, and survivability.

Several existing and proposed network architectures are discussed. The two-layer model, which aims at a tighter integration between IP and optical layers, offers a series of important advantages over the current multi-layer architecture. The benefits include more flexibility in handling higher capacity networks, better network scalability, more efficient operations and better traffic engineering.

Multiprotocol Label Switching (MPLS) and its extension Generalized Multiprotocol Label Switching (GMPLS) have been proposed as the integrating structure between IP and optical layers. Routing in the non-optical and optical parts of the hybrid IP network needs to be coordinated. Several models have been proposed including overlay, augmented, and peer-to-peer models. These models and the required enhancements to IP routing protocols, such as, OSPF and IS-IS are provided.

Control in the IP over Optical networks is facilitated by MPLS control plane. Each node consists of an integrated IP router and optical layer crossconnect (OLXC). The interaction between the router and OLXC layers is defined. Signaling among various nodes is achieved using CR-LDP and RSVP-TE protocols.

The management functionality in optical networks is still being developed. The issues of link initialization and performance monitoring are summarized in this document.

With the introduction of IP in telecommunications networks, there is tremendous focus on reliability and availability of the new IP-optical hybrid infrastructures. Automated establishment and restoration of end to end paths in such networks require standardized signaling and routing mechanisms. Layering models that facilitate fault restoration are discussed. A better integration between IP and optical will provide opportunities to implement a better fault restoration.

The 02 revision fixes an error in the list of authors.

Conventions used in this document

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.

ntents:

Overview

. Introduction

Challenges presented by the exponential growth of the Internet have resulted in the intense demand for broadband services. In satisfying the increasing demand for bandwidth, optical network technologies represent a unique opportunity because of their almost unlimited potential bandwidth.

Recent developments in wavelength-division multiplexing (WDM) technology have dramatically increased the traffic capacities of optical networks. Research is ongoing to introduce more intelligence in the control plane of the optical transport systems, which will make them more survivable, flexible, controllable and open for traffic engineering. Some of the essential desirable attributes of optical transport networks include real-time provisioning of lightpaths, providing capabilities that enhance network survivability, providing interoperability functionality between vendor-specific optical sub-networks, and enabling protection and restoration capabilities in operational contexts. The research efforts now are focusing on the efficient internetworking of higher layers, primarily IP with WDM layer.

Along with this WDM network, IP networks, SONET networks, ATM backbones shall all coexist. Various standardization bodies have been involved in determining an architectural framework for the interoperability of all these systems.

One approach for sending IP traffic on WDM networks would use a multi-layered architecture comprising of IP/MPLS layer over ATM over SONET over WDM. If an appropriate interface is designed to provide access to the optical network, multiple higher layer protocols can request lightpaths to peers connected across the optical network. This architecture has 4 management layers. One can also use a packet over SONET approach, doing away with the ATM layer, by putting IP/PPP/HDLC into SONET framing. This architecture has 3 management layers. A few problems of such multi layered architectures have been studied.

```
     +---------------+
     |               |
     |   IP/MPLS     |
     |               |
     +---------------+   +---------------+
     |               |   |               |
     |     ATM       |   |    IP/MPLS    |
     |               |   |               |
     +---------------+   +---------------+    +--------------+
     |               |   |               |    |              |
     |    SONET      |   |    SONET      |    |  IP/MPLS     |
     |               |   |               |    |              |
     +---------------+   +---------------+    +--------------+
     |               |   |               |    |              |
     |     WDM       |   |      WDM      |    |     WDM      |
     |               |   |               |    |              |
     +---------------+   +---------------+    +--------------+

         (4 LAYERS)         (3 LAYERS)           (2 LAYERS)
```

Figure 1: Layering Architectures Possible

The fact that it supports multiple protocols, will increase complexity for IP-WDM integration because of various edge-interworkings required to route, map and protect client signals across WDM subnetworks. The existence of separate optical layer protocols may increase management costs for service providers.

One of the main goals of the integration architecture is to make optical channel provisioning driven by IP data paths and traffic engineering mechanisms. This will require a tight cooperation of routing and resource management protocols at the two layers. The multi-layered protocols architecture can complicate the timely flow of the possibly large amount of topological and resource information.

Another problem is with respect to survivability. There are various proposals stating that the optical layer itself should provide restoration/protection capabilities of some form. This will require careful coordination with the mechanisms of the higher layers such as the SONET Automatic Protection Switching (APS) and the IP re-routing strategies. Hold-off timers have been proposed to inhibit higher layers backup mechanisms.

Problems can also arise from the high level of multiplexing done. The optical fiber links contain a large number of higher layer flows such as SONET/SDH, IP flows or ATM VCs. Since these have their own mechanisms, a flooding of alarm messages can take place.

Hence, a much closer IP/WDM integration is required. The discussions, henceforth in this document, shall be of such an architecture. There exist, clouds of IP networks, clouds of WDM networks. Transfer of packets from a source IP router to a destination is required. How the combination does signaling to find an optimal path, route the packet, and ensure survivability are the topics of discussion.

Multi-Protocol Label Switching (MPLS) for IP packets is believed to be the best integrating structure between IP and WDM. MPLS brings two main advantages. First, it can be used as a powerful instrument for traffic engineering. Second, it fits naturally to WDM when wavelengths are used as labels. This extension of the MPLS is called the Multi-protocol lambda switching.

This document starts off with a description of the optical network model. Section 2 describes the correspondence between the optical network model and the MPLS architecture and how it can bring about the inter-working. Section 3 is on routing in this architecture. It also describes 3 models for looking at the IP cloud and the Optical cloud namely the Overlay model, the augmented model and the peer model. Sections 4 and 5 are on control, signaling and management, respectively. Section 6 is on restoration. Acronyms and glossary are defined in Sections 8 and 9.

2 Network Model

In this draft, all the discussions assume the network model shown in Figure 2 [Luciani00]. Here, we consider a network model consisting of IP routers attached to an optical core network and connected to their peers over dynamically switched lightpaths. The optical core network is assumed to consist of multi-vendor optical sub-networks and are incapable of processing IP packets. In this network model, a switched lightpath has to be established between a pair of IP routers for their communication. The lightpath might have to traverse multiple optical sub-networks and be subject to different provisioning and restoration procedures in each sub-network.

For this network model, two logical control interfaces are identified, viz., the client-optical network interface (UNI), and the optical sub-network interface (NNI). The UNI represents a technology boundary between the client and optical networks. And, the NNI represents a technology boundary across multi-vendor optical sub-networks [Luciani00].

The physical control structure used to realize these logical interfaces may vary depending on the context and service models, which are discussed later in this draft.

```
                           +---------------------------------------+
                           |          Optical Network              |
      +--------------+     |                                       |
      |              |     | +------------+        +------------+  |
      |   IP         |     | |            |        |            |  |
      |   Network     +--UNI--+  Optical  +---NNI--+   Optical  |  |
      |              |     | |Subnetwork  |        |Subnetwork  |  |
      +--------------+     | |            |  +-----+            |  |
                           | +------+-----+  |     +------+-----+  |
                           |        |        |            |        |
                           |       NNI      NNI          NNI       |
      +--------------+     |        |        |            |        |
      |              |     | +------+-----+  |     +------+-----+  |
      |   IP         +--UNI--|            +--+     |            |  |
      |   Network    |     | |  Optical   |        |  Optical   |  |
      |              |     | |Subnetwork  +---NNI--+Subnetwork  |  |
      +--------------+     | |            |        |            |  |
                           | +------+-----+        +------+-----+  |
                           |        |                     |        |
                           +-------UNI-------------------UNI-------+
                                    |                     |
                                    |                     |
                             +------+-------+     +------------+
                             |              |     |            |
                             |Other Client  |     |Other Client|
                             |  Network     |     |  Network   |
                             | (e.g., ATM)  |     |            |
                             +--------------+     +------------+
```

Figure 2: Network Model

Optical Switch Architecture

It has been realized that optical networks must be survivable, flexible, and controllable. And hence, there is ongoing trend to introduce intelligence in the control plane of optical networks to make them more versatile. There is general consensus in the industry that the optical network control plane should utilize IP-based protocols for dynamic provisioning and restoration of lightpaths within and across optical sub-networks. In the existing IP-centric data network domain, these functionalities are performed by the Multi Protocol Label Switching (MPLS) traffic engineering control plane and currently in the optical domain it is achieved by Multi Protocol Lambda Switching (MPLambdaS), where wavelength is used as a label for switching the data at each hop. In this section, we identify the similarities that exist between the all-optical crossconnects (OXCs) of the optical networks and the label switch routers (LSRs) of the MPLS networks and identify how the control plane model of MPLS traffic engineering (TE) can be applied to that of optical transport network.

. Multi Protocol Label Switching (MPLS)

Multi Protocol Label Switching is a switching method in which a label field in the incoming packets is used to determine the next hop. At each hop, the incoming label is replaced by another label that is used at the next hop. The path thus realized is called a Label Switched Path (LSP). Devices which base their forwarding decision based solely on the incoming labels (and ports) are called Label Switched Routers (LSRs). In an IP-centric optical internetworking environment, OXCs and LSRs are used to switch the LSPs in the optical domain and the IP domain respectively. The OXCs are programmable and may support wavelength conversion and translation. It is important here to enumerate the relations and distinctions between OXCs and LSRs to expose the reusable software artifacts from the MPLS traffic engineering control plane model. Both OXCs and LSRs emphasize problem decomposition by architecturally decoupling the control plane from the data plane.

2 Isomorphic Relations and Distinctions between OXCs And LSRs

While an LSR's data plane uses the label swapping paradigm to transfer a labeled packet from an input port to an output port, the data plane of an OXC uses a switch matrix to provision a lightpath from an input port to an output port. LSR's control plane is used to discover, distribute, and maintain relevant state information related to the MPLS network, and to instantiate and maintain label switched paths (LSPs) under various MPLS traffic engineering rules and policies. OXC's control plane is used to discover, distribute, and maintain relevant state information associated with the Optical Transport Network (OTN), and to establish and maintain lightpaths under various optical internetworking traffic engineering rules and policies [Awuduche].

Current generation of OXCs and LSRs differ in certain characteristics. While LSRs are datagram devices that can perform certain packet level operations in the data plane, OXCs cannot. They cannot perform packet level processing in the data plane. Another difference is that the forwarding information is carried explicitly in LSRs as part of the labels appended to the data packets unlike OXCs, where the switching information is implied from the wavelength or the optical channel.

3 Isomorphic Relations and Distinctions between LSPs and Lightpaths

Both the explicit LSPs and lightpaths exhibit certain commonalties. For example, both of them are the abstractions of unidirectional, point-to-point virtual path connections. [Awuduche]. Another commonality is that the payload carried by both LSPs and lightpaths are transparent along their respective paths. They can be parameterized to stipulate their performance, behavioral, and survivability requirements from the network. There are certain similarities in the allocation of labels to LSPs and in the allocation of wavelengths to lightpaths.

There is one major distinction between LSPs and lightpaths in that LSPs support label stacking, but the concept similar to label stacking, i.e., wavelength stacking doesn't exist in the optical domain at this time.

4 General Requirements for the OXC Control Plane

This section describes some of the requirements for the OXC control plane with emphasis on the routing components. Some of the key aspects to these requirements are:
(a) to expedite the capability to establish lightpaths,
(b) to support traffic engineering functions, and
(c) to support various protection and restoration schemes.

Since the historical implementation of the "control plane" of optical transport networks via network management has detrimental effects like slow restoration, preclusion of distributed dynamic routing control, etc., motivation is to improve the responsiveness of the optical transport network and to increase the level of interoperability within and between service provider networks.

In the following sections, we give a brief overview of MPLS traffic engineering (MPLS-TE), summarize the enhancements that are required in the OXCs to support the MPLS TE as well as the changes required in the MPLS control plane to adapt to the OXCs.

4.1 Overview Of The MPLS Traffic Engineering Control

The components of the MPLS traffic engineering control plane model include the following modules [Awuduche]:

(a) Resource discovery.
(b) State information dissemination to distribute relevant information concerning the state of the network, like network topology, resource availability information.
(c) Path selection that is used to select an appropriate route through the MPLS network for explicit routing.
(d) Path management, which includes label distribution, path placement, path maintenance, and path revocation.

The above components of the MPLS traffic engineering control plane are separable, and independent of each other, and hence it allows an MPLS control plane to be implemented using a composition of best of breed modules.

4.2 OXC Enhancements to Support MPLS Control Plane

This section discusses some of the enhancements to OXCs to support wavelength switching. This extension, which has now been superceded with Generalized MPLS (GMPLS) was originally called Multiprotocol Lambda Switching or MPL(ambda)S. There are three key enhancements. First, there should be a mechanism to exchange control information between OXCs, and between OXCs and other LSRs. This can be accomplished in-band or quasi-in-band using the same links that are used to carry data-plane traffic, or out-of-band via a separate network. Second, an OXC should be able to provide the MPLS traffic engineering control plane with pertinent information regarding the state of individual fibers attached to that OXC, as well the state of individual lightpaths or lightpaths within each fiber. Third, even when an edge LSR does not have WDM capabilities, it should still have the capability to exchange control information with the OXCs in the domain.

4.3 MPLS Control Plane Enhancements

This section discusses the enhancements that are to be made in the MPLS control plane to support MPL(ambda)S.

An MPLS domain may consist of links with different properties depending upon the type of network elements at the endpoints of the links. Within the context of MPL(ambda)S, the properties of a link consisting of a fiber with WDM that interconnects two OXCs are different from that of a SONET link that interconnects two LSRs. As an example, a conventional LSP cannot be terminated on a link connected to a pure OXC. However, a conventional LSP can certainly be terminated on a link connected to a frame-based LSR. These differences should be taken into account when performing path computations to determine an explicit route for an LSP. It is also feasible to have the capability to restrict the path of some LSPs to links with certain characteristics. Path computation algorithms may then take this information into account when computing paths LSPs.

If there are multiple control channels and bearer channels between two OXCs, then there must be procedures to associate bearer channels to corresponding control channels. Procedures are required to de-multiplex the control traffic for different bearer channels if a control channel is associated with multiple bearer channels. Procedures are also needed to activate and deactivate bearer channels, to identify the bearer channels associated with any given physical link, to identify spare bearer channels for protection purposes, and to identify impaired bearer channels, particularly, in the situation where the physical links carrying the bearer channel are not impaired.

Signaling protocols (RSVP-TE and CR-LDP) need to be extended with objects that can provide sufficient details to establish reconfiguration parameters for OXC switch elements. Interior

Gateway Protocols (IGPs) should be extended to carry information about the physical diversity of the fibers. IGPs should be able to distribute information regarding the allocatable bandwidth granularity of any particular link.

5 MPLS Traffic Engineering Control Plane with OXCs

In IP-centric optical interworking systems, given that both OXCs and LSRs require control planes, one option would be to have two separate and independent control planes [Awuduche]. Another option is to develop a uniform control plane that can be used for both LSRs and OXCs. This option of having a uniform control plane will eliminate the administrative complexity of managing hybrid optical internetworking systems with separate, dissimilar control and operational semantics. Specialization may be introduced in the control plane, as necessary, to account for inherent peculiarities of the underlying technologies and networking contexts. A single control plane would be able to span both routers and OXCs. In such an environment, a LSP could traverse an intermix of routers and OXCs, or could span just routers, or just OXCs. This offers the potential for real bandwidth-on-demand networking, in which an IP router may dynamically request bandwidth services from the optical transport network.

To bootstrap the system, OXCs must be able to exchange control information. One way to support this is to pre-configure a dedicated control wavelength between each pair of adjacent OXCs, or between an OXC and a router, and to use this wavelength as a supervisory channel for exchange of control traffic. Another possibility would be to construct a dedicated out-of-band IP network for the distribution of control traffic.

Though an OXC equipped with MPLS traffic engineering control plane would resemble a Label Switching Router; there are some important distinctions and limitations. As discussed earlier, the distinction concerns the fact that there are no analogs of label merging in the optical domain, which implies that an OXC cannot merge several wavelengths into one wavelength. Another major distinction is that an OXC cannot perform the equivalent of label push and pop operation in the optical domain. This is due to lack of the concept of pushing and popping wavelengths is infeasible with contemporary commercial optical technologies. Finally, there is another important distinction, which is concerned with the granularity of resource allocation. An MPLS router operating in the electrical domain can potentially support an arbitrary number of LSPs with arbitrary bandwidth reservation granularities, whereas an OXC can only support a relatively small number of lightpaths, each of which will have coarse discrete bandwidth granularities.

6 Generalized MPLS (GMPLS)

The Multi Protocol Lambda Switching architecture has recently been extended to include routers whose forwarding plane recognizes neither packet, nor cell boundaries, and therefore, can't forward data based on the information carried in either packet or cell headers. Specifically, such routers include devices where the forwarding decision is based on time slots, wavelengths, or physical ports. GMPLS differs from traditional MPLS in that it supports multiple types of switching, i.e., the addition of support for TDM, lambda, and fiber (port) switching. The support for the additional types of switching has driven generalized MPLS to extend certain base functions of traditional MPLS [GMPLS].

While traditional MPLS links are unidirectional, generalized MPLS supports the establishment of bi-directional paths. The need for bi-directional LSPs comes from its extent of reach. Bi-directional paths also have the benefit of lower setup latency and lower number of messages required during setup. Other features supported by generalized MPLS are rapid failure notification and termination of a path on a specific egress port.

To deal with the widening scope of MPLS into the optical and time domain, several new forms of "label" are required. These new forms of label are collectively referred to as a "generalized label". A generalized label contains enough information to allow the receiving node to program its crossconnect. Since the nodes sending and receiving this new form of label know what kinds of link they are using, the generalized label does not contain a type field, instead the nodes are expected to know from context what type of label to expect. Currently, label formats supported by GMPLS are the Generalized Label, the Waveband Switching Label (which apparently uses the same Generalized Label format), the Suggested Label and the Label Set [GMPLS00].

(1) The Generalized Label: It extends the traditional Label Object in that it allows the representation of not only labels which travel in-band with associated data packets, but also labels which identify time-slots, wavelengths, or space division multiplexed positions. A Generalized Label only carries a single level of label, i.e., it is non-hierarchical. When multiple levels of label (LSPs within LSPs) are required, each LSP must be established separately.

(2) Waveband Switching Label: Waveband switching label format uses the same format as the generalized label. Waveband switching is a special case of lambda switching, where a set of contiguous wavelengths are represented as a waveband and can be switched together to a new waveband.

(3) Suggested Label: The Suggested Label is used to provide a downstream node with the upstream node's label preference, which permits the upstream node to start configuring its hardware with the proposed label before the label is communicated by the downstream node. This feature is valuable to systems where it takes non-trivial time to establish a label in hardware and thus reducing setup latency. One use of suggested labels is to indicate preferred wavelength.

(4) Label Set: The Label Set is used to limit the label choices of a downstream node to a set of acceptable labels. This limitation applies on a per hop basis. Label Set is used to restrict label ranges that may be used for a particular LSP between two peers. The receiver of a Label Set must restrict its choice of labels to one which is in Label Set. Conceptually, the absence of a Label Set implies a Label Set whose value is the set of all valid labels.

IP over Optical Networks

. Service models

The optical network model considered in this draft consists of multiple Optical Crossconnects (OXCs) interconnected by optical links in a general topology (referred to as an "optical mesh network"). Each OXC is assumed to be capable of switching a data stream from a given input port to a given output port. This switching function is controlled by appropriately configuring a crossconnect table. Conceptually, the crossconnect table consists of entries of the form <input port i, output port j>, indicating that data stream entering input port i will be switched to output port j. An "lightpath" from an ingress port in an OXC to an egress port in a remote OXC is established by setting up suitable crossconnects in the ingress, the egress and a set of intermediate OXCs such that a continuous physical path exists from the ingress to the egress port. Lightpaths are assumed to be bi-directional, i.e., the return path from the egress port to the ingress port follows the same path as the forward path.It is assumed that one or more control channels exist between neighboring OXCs for signaling purposes.

In this section two possible service models are discussed in brief. The first being at the optical UNI and the second being at the optical sub-network NNI[Luciani00].

..1 Client Server Model

Under this model the optical network primarily provides a set of high bandwidth pipes to the client requesting such. Standardized signaling can be used to invoke the following:

1. Lightpath creation.
2. Lightpath deletion.
3. Lightpath modification.
4. Lightpath status enquiry.

The continued operation of the system requires that the client systems continuously register with the optical network. Signalling extensions need to be added to allow clients to register, deregister and query other clients for an optical-networked administered address so that lightpaths can be established with other clients across the optical network. Along with these signaling extensions a service discovery mechanism needs to be added which will allow the client to discover the static parameters of the link along with the UNI signaling protocol being used on the link. In this service model the routing protocols inside the optical network are exclusive of what is followed inside the client network. Only a minimal set of messages need to be defined between the router and the optical network. RSVP-TE, LDP or a TCP based control channel can be used for the same. Within the optical cloud NNI interface is defined between the various optical subnetworks. Details of the UNI and NNI signaling requirements are provided further on in this document.

..2 Integrated Service Model

In the Integrated Service Model the IP and the optical networks are treated as a single network and there is no distinction between the optical switches and the IP routers as far as the control plane goes. MPLS would be the preferred method for control and routing and there is no distinction between the UNI, NNI or any other router-router interface. Under this model, optical network services are provisioned using MPLS signaling as specified in [GMPLS]. In this service model the edge router can do the creation and modification of the label switched paths across the optical network. In some sense this resembles the client server model just presented, but it seems to promise seamless integration when compared to the client server model. OSPF with TE extensions to support optical networks could be used to exchange topology information and do the routing .It might happen in an optical network that a LSP across the optical network may be a conduit for a lot of other LSPs. This can be advertised as a virtual link inside a forward adjacency in protocols like OSPF. Thus from the point of view of the data plane an overlay is created between two edge routers across the optical network.

2 IP Optical Interaction Models

The previous section presented possible service models for IP over optical networks. The models differ in the way routing is implemented. It is important to examine the architectural alternatives for routing information exchange between IP routers and optical switches. The aim of this exercise is to allow service discovery, automated establishment and seamless integration with minimal intervention. MPLS based signaling is assumed in the following discussion.

Some of the proposed models for interaction between IP and optical components in a hybrid network are [Luciani00]:

(1) Overlay model
(2) Integrated/Augmented model
(3) Peer model

The key consideration in deciding the type of model is whether there is a single or separate monolithic routing and signaling protocol spanning the IP and the Optical domains. If there are separate instances of routing protocols running for each domain then the following three considerations help determine the model:

1. What is the interface defined between the two protocol instances?
2. What kind of information is exchanged between the protocol instances?
3. What are policies regarding provisioning of the lightpaths across the optical domain between edge routers? This includes access control accounting and security.

2.1 Overlay Model

Under the overlay model, IP domain is more or less independent of the optical domain. That is IP domain acts as a client to the Optical domain. In this scenario, the optical network provides point to point connection to the IP domain. The IP/MPLS routing protocols are independent of the routing and signaling protocols of the optical layer. The overlay model may be statically provisioned using a Network Management System or may be dynamically provisioned. Static provisioning solution may not be scalable though.

2.2 Peer Model

In the peer model the optical routers and optical switches act as peers and there is only one instance of a routing protocol running in the optical domain and in the IP domain. A common IGP like OSPF or IS-IS may be used to exchange topology information. OSPF opaque Link State Advertisements (LSAs) and extended type-length-value encoded fields (TLVs) may be used in the case of IS-IS. The assumption in this model is that all the optical switches and the routers have a common addressing scheme.

2.3 Augmented Model

In the augmented model, there are actually separate routing instances in the IP and optical domains but information from one routing instance is leaked into the other routing instance. For example IP addresses could be assigned to optical network elements and carried by optical routing protocols to allow reachability information to be shared with the IP domain to support some degree of automated discovery.

3 Routing Approaches

3.1 Fully peered routing model

This routing model is used for the peer model described above. Under this approach there is only one instance of the routing protocol running in the IP and Optical domains. An IGP like OSPF or IS-IS with suitable optical extensions is used to exchange topology information. These optical extensions will capture the unique optical link parameters. The OXCs and the routers maintain the same link state database. The routers can then compute end-to-end paths to other routers across the OXCs. Such a Label Switched Path (LSP) can then be signaled using MPLS signaling protocols like RSVP-TE or CR-LDP. This lighpath is always a tunnel across the optical network between edge routers. Once created such lightpaths are treated as virtual links and are used in traffic engineering and route computation. As and when forwarding adjacencies (FAs) are introduced in the link state corresponding links over the IP Optical interface are removed from the link state advertisements. Finally the details of the optical network are completely replaced by the FAs advertised in the link state.

3.2 Domain Specific Routing

This routing model supports the augmented routing model. In this model the routing between the optical and the IP domains is separated with a specific routing protocol running between the domains. The focus is on the routing information to be exchanged at the IP optical interface. Interdomain routing protocols like BGP may be used to exchange information between the IP and optical domain. OSPF areas may also be used to exchange routing information across the two domains.

3.2.1 Routing using BGP

BGP will allow IP networks to advertise IP addresses within its network to external optical networks while receiving external IP prefixes from the optical network. Edge routers and OXCs can run External BGP (EBGP). Within the optical network EBGP can be used between optical subnetworks across the NNI and Internal BGP (IBGP) can be used within the optical network. Using this scheme it is essential to identify the optical network corresponding to the egress IP addresses. The reason is as follows. Whenever an edge router wants to setup a LSP across an optical network it is just going to specify the destination IP. Now if the edge router has to request another path to the destination it must know if there already exist lightpaths with residual capacity to the destination. To determine this it needs to know which ingress ports in an OXC correspond to which external destination. Thus a border OXC receiving external IP addresses by way of EBGP must include information about its IP address and pass it on to the edge router. The edge router must store this association between the OXCs and the external IP addresses and need not propagate the egress address further. Specific mechanisms to propagate the BGP egress addresses are yet to be determined.

3.2.2 Routing using OSPF

OSPF supports the concept of hierarchical routing using OSPF areas. Information across a UNI can be exchanged using this concept of a hierarchy. Routing within each area is flat. Routers attached to more than one areas are called Area Border Routers (ABR). An ABR propogates IP addressing information from one area to another using a summary LSA. Domain specific routing can be done within each area.

Optical networks can be implemented as an area with an enhanced version of OSPF with optical extensions running on it. IP client networks can be running OSPF with TE extensions. Summary LSAs exchanged between the two areas would provide enough information for the establishment of lightpaths across the optical network. Domain specific information in the optical network can be hidden from the client network.

OSPF or BGP help in route discovery and collecting reachability information. Determination of paths and setting up of the LSPs is a traffic engineering decision.

3.3 Overlay Routing

Overlay routing is much like the IP over ATM and supports the overlay connection model. IP overlays are setup across the optical network. Address resolution similar to that in IP over ATM is used. The optical network can maintain a registry of IP addresses and VPN identifiers it is connected to. On querying the database for an external IP address it would return the appropriate egress port address on the OXC. Once an initial set of lightpaths are created VPN wide routing adjacencies can be formed using OSPF. The IP VPN would then be "overlayed" on the underlying optical network which could have an independent way of routing.

4 Path Selection

A possible scenario for path selection is presented in Figure 3.

```
+--------+        +--------+        +--------+
|        |<------>|Path    |        |OSPF    |
| CR-LDP |    +-->|Selector|        |        |
|        |    |   |        |        |TE EXT  |
+--------+    |   +---+----+   +---|OPT EXT |
              |       |        |   +--------+
              |       v        |
              |   +--------+   |   +--------+
+--------+    |   |TE      |   |   |IS-IS   |
|        |    |   |Database|   |   |        |
| RSVP-TE|<---+   |        |   |   |TE EXT  |
|        |        |        |<--+---|OPT EXT |
+--------+        +--------+       +--------+
```

Figure 3: Lightpath Selection

These systems use CR-LDP or RSVP-TE to signal MPLS paths. These protocols can source route by consulting a traffic engineering database, which is maintained along with the IGP database. This information is carried opaquely by the IGP for constraint based routing. If RSVP-TE or CR-LDP is used solely for label provisioning, the IP router functionality must be present at every label switch hop along the way. Once the label has been provisioned by the protocol then at each hop the traffic is switched using the native capabilities of the device to the eventual egress LSR.

Path selection can be online or offline. An offline computation is normally centralized while as an online computation is normally distributed. An offline computation is facilitated by simulation or network planning tools and can be used to provide guidance to subsequent real time computations. An online computation may be done whenever a connection request comes in. A combination of offline and online computations may be used by a network operator. Offline computations are used when complicated traffic engineering, demand planning, cost planning and global optimization is a priority.

In case of online computations there can be two choices when it comes to routing.
1) Explicit routing using a global view of the network can be used to calculate the most optimal solution taking into consideration constraints other than link metrics.
2) Hop by hop routing using path calculation at every node. This may not be able to provide an optimal solution taking into consideration constraints other than non additive link metrics.

6 Constraints on Routing

The constraints highlighted here apply to any circuit switched networks but differences with an optical network are explained where applicable[GMPLS-CONTROL].

One of the main services provided by any transport network is restoration. Restoration introduces the constraint of physically diverse routing. Restoration can be provided by pre-computed paths or computing the backup path in real time. The backup path has to be diverse from the primary path at least in the failed link or completely physically diverse. A logical attribute like the Shared Risk Link Group (SRLG) is abstracted by the operators from various physical attributes like trench ID and destructive areas. Such an attribute may be needed to be considered when making a decision about which path to take in a network. Two links which share a SRLG cant be the backup for one another because they both may go down at the same time. In order to satisfy such constraints path selection algorithms are needed to find two disjoint paths in a graph. Suurballe's algorithm as discussed [Suurballe] is a good example of an algorithm to find two node disjoint paths in a network.

Another restoration mechanism is restoration in a shared mesh architecture wherein backup bandwidth may be shared among circuits. It may be the case that two link disjoint paths share a backup path in the network. This may be possible because a single failure scenario is assumed. A few heuristics to optimize the bandwidth allocated to a backup path in a mesh architecture have already been proposed [Bell-Labs]. Optimal routing requires considerable network level information and the most optimal solutions still require further study. Detailed protection and restoration mechanisms are discussed in later sections of this document.

Another constraint of interest is the concept of node, link, LSP inclusion or exclusion, propagation delay, wavelength convertibility and connection bandwidth among other things. A service provider may want to exclude a set of nodes due to the geographic location of the nodes. An example would be nodes lying in an area which is earthquake prone. Propagation delay may be another constraint for a large global network. Traffic from the US to Europe, shouldn't normally be routed over links across the Pacific ocean but instead should use links over the Atlantic ocean since propagation delay in this case would be much less.

Wavelength convertibility is a problem encountered in waveband networks. It refers to ability of OXC to crossconnect two different wavelengths. The wavelengths may be completely different or slightly different. Since wavelength convertibility currently involves an optical-electrical-optical (OEO) conversion, vendors may selectively deploy these converters inside the network. Therein lies the problem of routing a circuit over a network using the same wavelength. This requires that the path selection algorithms know the availability of each wavelength on each link along the route. With link bundling, this is difficult since information about all the wavelengths may be included in the same bundle. Link probing may have to be employed at the source router to find out the number of wavelengths available along the path.

Bandwidth availability is another consideration in routing. This is simplified in a wavelength optical network since requests are end to end. However, in a TDM transport network such as a SONET/SDH network, requests can be variable bandwidth. Routing needs to ensure that sufficient capacity is available end to end. There are further difficulties introduced due to the different concatenation schemes in the SONET/SDH schemes. An example would be a concatenated STS-3c channel, which would require three adjacent time slots to be allocated. This implies that a time slot map of the link has to be distributed to the entire network to facilitate a routing decision. Alternatively in a logical link representation one would need N different logical links to represent all possible STS-N signals. But then this would take up too much control bandwidth. A preferred approach would be to advertise just the largest block of time slots available on a logical link instead of the entire time slot map. This is sufficient to determine if a connection can be supported on a link. Detailed resource information on local resource availability is only used for routing decisions.

Signaling & Control

Signaling refers to messages used to communicate characteristics of services requested or provided. This section discusses a few of the signaling procedures. It is assumed that there exists some default communication mechanism between routers prior to using any of the routing and signaling mechanisms.

. MPLS Control Plane

A candidate system architecture for an OXC equipped with an MPLS control plane model is shown in Figure 4. The salient feature of the network architecture is that every node in the network consists of an IP router and a reconfigurable OLXC. The IP router is responsible for all non-local management functions, including the management of optical resources, configuration and capacity management, addressing, routing, traffic engineering, topology discovery, exception handling and restoration. In general, the router may be traffic bearing, or it may function purely as a controller for the optical network and carry no IP data traffic. Although the IP protocols are used to perform all management and control functions, lightpaths may carry arbitrary types of traffic.

The IP router implements the necessary IP protocols and uses IP for signaling to establish lightpaths. Specifically, optical resource management requires resource availability per link to be propagated, implying link state protocols such as OSPF. Between each pair of neighbors in the network, one communication channel exists that allows router to router connectivity over the channel. These signaling channels reflect the physical topology. All traffic on the signaling channel is IP traffic and is processed or forwarded by the router. Multiple signaling channels may exist between two neighbors and some may be reserved for restoration. Therefore, we can assume that as long as the link between two neighbors is functional, there is a signaling channel between those neighbors.

```
        --------------------------------
       |   OXC WITH MPLS CONTROL PLANE  |
       |                                |
       |        -------------------     |
       |       |                   |    |
       |       | MPLS Control Plane|    |
       |       |                   |    |
       |        -------------------     |
       |                |               |
       |        -------------------     |
       |       |                   |    |
       |       |Control Adaptation |    |
       |        -------------------     |
       |       |    OXC Switch     |    |
       |       |    Controller     |    |
       |        -------------------     |
       |                |               |
       |        -------------------     |
       |       |                   |    |
       |       | OXC Switch Fabric |    |
       |       | OXC Data Plane    |    |
       |        -------------------     |
       |                                |
```

```
      |________________________________|
       --------------------------------
```

Figure 4: OXC Architecture

The IP router module communicates with the OLXC device through a logical interface. The interface defines a set of basic primitives to configure the OLXC, and to enable the OLXC to convey information to the router. The mediation device translates the logical primitives to and from the proprietary controls of the OLXC. Ideally, this interface is both explicit and open. We recognize that a particular realization may integrate the router and the OLXC into a single box and use a proprietary interface implementation. Figure 5 illustrates this implementation.

The following interface primitives are examples of a proposal for communication between the router and the OLXC within a node:

a) Connect(input link, input channel, output link, output channel): Commands sent from the router to the OLXC requesting that the OLXC crossconnect input channel on the input link to the output channel on the output link.

b) Disconnect(input link, input channel, output link, output channel): Command sent from the router to the OLXC requesting that it disconnect the output channel on the output link from the connected input channel on the input link.

```
                   +-------------------------------+
                   |                               |
                   |        Router   module        |
                   |                               |
                   |                               |
                   |  |                         |  |
                   |  |        Primitives       |  |
                   |  |        /|\   |          |  |
                   |  |         |    |          |  |
                   +--+---------+----+----------+--+  Control
Channel            |  |         |    |          |  |
         ----------+--+         |    |          +--+---------
         ----------+----------  |   \|/  ----------+---------
         ----------+----------           ----------+---------
         ----------+----------           ----------+---------
         ----------+----------           ----------+---------
                   |             OLXC              |
                   |                               |
                   |                               |
                   +-------------------------------+
```

Figure 5: Control Plane Architecture

c) Bridge(input link, input channel, output link, output channel): Command sent from the router controller to the OLXC requesting the bridging of a connected input channel on input link to another output channel on output link.

d) Switch(old input link, old input channel, new input link, new input channel, output link, output channel): Switch output port from the currently connected input channel on the input link to the new input channel on the new input link. The switch primitive is equivalent to atomically implementing a disconnect(old input channel, old input link, output channel, output link) followed by a connect(new input link, new input channel, output link, output channel).

e) Alarm(exception, object):
Command sent from the OLXC to the router informing it of a failure detected by the OLXC. The object represents the element for which the failure has been detected.

For all of the above interfaces, the end of the connection can also be a drop port.

2 Addressing

Every network addressable element must have an IP address. Typically these elements include each node and every optical link and IP router port. When it is desirable to have the ability to address individual optical channels those are assigned IP addresses as well. The IP addresses must be globally unique if the element is globally addressable. Otherwise domain unique addresses suffice. A client must also have an IP address by which it is identified. However, optical lightpaths could potentially be established between devices that do not support IP (i.e., are not IP aware), and consequently do not have IP addresses. This could be handled either by assigning an IP address to the device, or by assigning an address to the OLXC port to which the device is attached. Whether or not a client is IP aware can be discovered by the network using traditional IP mechanisms.

3 Path Setup

This section describes a protocol proposed for setting up an end-to-end lightpath for a channel. A complete path might contain the two endpoints and an array of intermediate OXCs for transport across the optical network. This section describes the handshake used for ad-hoc establishment of lightpaths in the network. Provisioning an end-to-end optical path across multiple sub-networks involves the establishment of path segments in each sub-network sequentially. Inside the optical domain, a path segment is established from the source OXC to a border OXC in the source sub-network. From this border OXC, signaling across the NNI is performed to establish a path segment to a border OXC in the next sub-network. Provisioning continues this way until the destination OXC is reached.

The link state information is used to compute the routes for lightpaths being established. It is assumed that a request to establish a lightpath may originate from an IP router (over the UNI), a border node (over the NNI) , or a management system. This request carries all required parameters. After computing the route, the actual path establishment commences. However, once path setup is complete the data transfer happens passively and is straightforward without much intervention from the control plane. The connection needs to be maintained as per the service level agreements.

To automate this process, there are certain initiation procedures so as to determine the route for each segment (viz. IP host - IP border router, IP border router - border OXC, between border OXCs).

* Resource Discovery

Routing within the optical network relies on knowledge of network topology and resource availability. The first step towards network-wide link state determination is the discovery of the status of local links to all neighbors by each OXC. The end result is that each OXC creates a port state database.

Topology information is distributed and maintained using standard routing algorithms, e.g., OSPF and IS-IS. On boot, each network node goes through neighbor discovery. By combining neighbor discovery with local configuration, each node creates an inventory of local resources and resource hierarchies, namely: channels, channel capacity, wavelengths, and links.

* Route calculation

Different mechanisms for routing exist [Luciani00]. The route computation, after receiving all network parameters in the form of link state packets, reduces to a mathematical problem. It involves solving a problem of Routing and Wavelength Assignment (RWA) for the new connection. The problem is simplified if there exists a wavelength converter at every hop in the optical network.

3.1 UNI Path Provisioning

The real handshake between the client network and the optical backbone happens after performing the initial service & neighbor discovery. The continued operation of the system requires that client systems constantly register with the optical network. The registration procedure aids in verifying local port connectivity between the optical and client devices, and allows each device to learn the IP address of the other to establish a UNI control channel. The following procedures may be made available over the UNI:

* Client Registration: This service allows a client to register its address(es) and user group identifier(s) with the optical network.
* Client De-Registration: This service allows a client to withdraw its address(es) and user group identifier(s) from the optical network.

The optical network primarily offers discrete capacity, high bandwidth connectivity in the form of lightpaths. The properties of the lightpaths are defined by the attributes specified during lightpath establishment or via acceptable modification requests. To ensure operation of the domain services model, the following actions need to be supported at the UNI so as to offer all essential lightpath services. The UNI signaling messages are structured as *requests* and *responses* [UNI00].

1. Lightpath creation: This action allows a lightpath with the specified attributes to be created between a pair of termination points. Each lightpath is assigned a unique identifier by the optical network, called the lightpath ID. Lightpath creation may be subject to network-defined policies and security procedures.

2. Lightpath deletion: This action, originating from either end, allows an existing lightpath (referenced by its ID) to be deleted.

3. Lightpath modification: This action allows certain parameters of the lightpath (referenced by its ID) to be modified. Lightpath modification must not result in the loss of the original lightpath.

4. Lightpath status enquiry: This service allows the status of certain parameters of the lightpath (referenced by its ID) to be queried.

5. Notification: This action sends an autonomous message from the optical network to the client to indicate a change in the status of the lightpath (e.g., non-restorable lightpath failure).

Thus, the above actions provision both edges of the overall connection, while NNI provisioning builds the central portion of the setup

8.2 Basic Path Setup Procedure for NNI

The model for provisioning an optical path across optical sub-networks is as follows. A provisioning request may be received by a source OXC from the client border IP router (or from a management system), specifying the source and destination end-points. The source end-point is implicit and the destination endpoint is identified by the IP address. In both cases, the routing of an optical path inside the optical backbone is done as follows [Pendarakis00]:

* The source OXC looks up its routing information corresponding to the specified destination IP address. If the destination is an OXC in the source sub-network, a path maybe directly computed to it. If the destination is an external address, the routing information will indicate a border OXC that would terminate the path in the source sub-network. A path is computed to the border OXC.

* The computed path is signaled from the source to the destination OXC within the source sub-network. The destination OXC in the source sub-network determines if it is the ultimate destination of the path. if it is, then it completes the path set-up process. Otherwise, it determines the address of a border OXC in an adjacent sub-network that leads to the final destination. The path set-up is signaled to this OXC using NNI signaling. The next OXC then acts as the source for the path and the same steps are repeated.

Thus, NNI provisioning involves looking up in the routing table computed by various schemes mentioned previously and performing path setup within an optical sub-network. Techniques for link provisioning within the optical sub-network depends upon whether the OXCs do or do not have wavelength conversion. Both these cases are discussed below.

8.2.1 Network with Wavelength Converters

In an optical network with wavelength conversion, channel allocation can be performed independently on different links along a route. A lightpath request from a source is received by the first-hop router. (The term router here denotes the routing entity in the optical nodes or OXCs). The first-hop router creates a lightpath setup message and sends it towards the destination of the lightpath where it is received by the last-hop router. The lightpath setup is sent from the first-hop router on the default routed lightpath as the payload of a normal IP packet with router alert. A router alert ensures that the packet is processed by every router in the path. A channel is allocated for the lightpath on the downstream link at every node traversed by the setup. The identifier of the allocated channel is written to the setup message.

Note that the lightpath is established over the links traversed by the lightpath setup packet. After a channel has been allocated at a node, the router communicates with the OLXC to reconfigure the OLXC to provide the desired connectivity. After processing the setup, the destination (or the last-hop router) returns an acknowledgement to the source. The acknowledgment indicates that a channel has been allocated on each hop of the lightpath. It does not, however, confirm that the lightpath has been successfully implemented (or configured).

If no channel is available on some link, the setup fails, and a message is returned to the first-hop router informing it that the lightpath cannot be established. If the setup fails, the first-hop router issues a release message to release resources allocated for the partially constructed lightpath. Upon failure, the first-hop router may attempt to establish the lightpath over an alternate route, before giving up on satisfying the original user request. The first-hop router is obligated to establish the complete path. Only if it fails on all possible routes does it give a failure notification to the true source.

3.2.2 Network without wavelength converters

However, if wavelength converters are not available, then a common wavelength must be located on each link along the entire route, which requires some degree of coordination between different nodes in choosing an appropriate wavelength.

Sections of a network that do not have wavelength converters are thus referred to as being wavelength continuous. A common wavelength must be chosen on each link along a wavelength continuous section of a lightpath. Whatever wavelength is chosen on the first link defines the wavelength allocation along the rest of the section. A wavelength assignment algorithm must thus be used to choose this wavelength. Wavelength selection within the network must be performed within a subset of client wavelengths.

Optical non-linearity, chromatic dispersion, amplifier spontaneous emission and other factors together may limit the scalability of an all-optical network. Routing in such networks may then have to take into account noise accumulation and dispersion to ensure that lightpaths are established with adequate signal qualities. Hence, all routes become geographically constrained so that they will have adequate signal quality, and physical layer attributes can be ignored during routing and wavelength assignment.

One approach to provisioning in a network without wavelength converters would be to propagate information throughout the network about the state of every wavelength on every link in the network. However, the state required and the overhead involved in maintaining this information would be excessive. By not propagating individual wavelength availability information around the network, we must select a route and wavelength upon which to establish a new lightpath, without detailed knowledge of wavelength availability.

A probe message can be used to determine available wavelengths along wavelength continuous routes. A vector of the same size as the number of wavelengths on the first link is sent out to each node in turn along the desired route. This vector represents wavelength availability, and is set at the first node to the wavelength availability on the first link along the wavelength continuous section. If a wavelength on a link is not available or does not exist, then this is noted in the wavelength availability vector (i.e. the wavelength is set to being unavailable). Once the entire route has been traversed, the wavelength availability vector will denote the wavelengths that are available on every link along the route. The vector is returned to the source OXC, and a wavelength is chosen from amongst the available wavelengths using an arbitrary wavelength assignment scheme, such as first-fit.

The construction of a bi-directional lightpath differs from the construction of a unidirectional lightpath above only in that upon receiving the setup request, the last-hop router returns the setup message using the reverse of the explicit route of the forward path. Both directions of a bi-directional lightpath share the same characteristics, i.e., set of nodes, bandwidth and restoration requirements. For more general bi-directional connectivity, a user simply requests multiple individual lightpaths.

A lightpath must be removed when it is no longer required. To achieve this, an explicit release request is sent by the first-hop router along the lightpath route. Each router in the path processes the release message by releasing the resources allocated to the lightpath, and removing the associated state. It is worth noting that the release message is an optimization and need not be sent reliably, as if it is lost or never issued (e.g., due to customer premise equipment failure) the softness of the lightpath state ensures that it will eventually expire and be released.

4 Signaling protocols

The OXCs in the optical network are responsible for switching streams based on the labels present. The MPLS architecture for IP networks defines protocols for associating labels to individual paths. The signaling protocols are used to provision such paths in the optical networks. There are two options for MPLS-based signaling protocols-*Resource reSerVation Protocol Traffic Engineering Extensions (RSVP-TE)* or *Constraint Based Routing Label Distribution Protocol (CR-LDP)*.

There are some basic differences between the two protocols, but both essentially allow hop-by-hop signaling from a source to a destination node and in the reverse direction. Each of these protocols are capable of providing quality of service (QoS) and traffic engineering. Not all features present in these protocols are necessary to support lightpath provisioning. On the other hand, certain new features must be introduced in these protocols for lightpath provisioning, including support for bi-directional paths, support for switches without wavelength conversion, support for establishing shared backup paths, and fault tolerance.

The connection request may include bandwidth parameters and channel type, reliability parameters, restoration options, setup and holding priorities for the path etc. On receipt of the request, the ingress node computes a suitable route for the requested path, following applicable policies and constraints. Once the route has been computed, the ingress node invokes RSVP-TE / CR-LDP to set up the path.

4.1 CR-LDP Extensions for Path Setup

Label Distribution Protocol (LDP) is defined for distribution of labels inside one MPLS domain. CR-LDP is the constraint-based extension of LDP. One of the most important services that may be offered using MPLS in general and CR-LDP in particular is support for constraint-based routing of lightpaths across the routed network. Constraint-based routing offers the opportunity to extend the information used to setup paths beyond what is available for the routing protocol. For instance, an LSP can be setup based on explicit route constraints, QoS constraints, and other constraints. Constraint-based routing (CR) is a mechanism used to meet traffic-engineering requirements that have been proposed.

A Label Request message is used by an upstream LSR to request a label binding from the downstream LSR for a specified forwarding equivalency class (FEC) and CR-LSP. In optical networks, a Label Request message may be used by the upstream OXC to request a port (and wavelength) assignment from the downstream OXC for the lightpath being established. Using downstream-on-demand and ordered control mode, a Label Request message is initially generated at the ingress OXC and is propagated to the egress OXC. Also, a protocol is required to determine the port mappings.

To incorporate the above mentioned constraints, the following extensions to current version of CR-LDP have been proposed:

* Inclusion of Signaling Port ID
* Signaling Optical Switched Path Identifier
* Signaling the two end points of the path being set up
* Signaling requirements for both span and path protection
* Recording the precise route of the path being established

4.2 RSVP-TE Extensions for Path Setup

Resource reSerVation Protocol with Traffic Engineering extensions (RSVP-TE) is a unicast and multicast signaling protocol designed to install and maintain reservation state information at each routing engine along a path [Luciani00]. The key characteristics of RSVP are that it is simplex, receiver-oriented and soft. It makes reservations for unidirectional data flows. The receiver of a data flow generally initiates and maintains the resource reservation used for that flow. It maintains "soft" state in routing engines. The "path" messages are propagated from the source towards potential recipients. The receivers interested in communicating with the source send the "Resv" messages.

The following extensions to RSVP-TE have been proposed to support path setup :

- Reduction of lightpath establishment latency
- Establishment of bi-directional lightpaths
- Fast failure notification
- Bundling of notifications

5 Stream Control Transmission Protocol (SCTP)

There is further discussion on which transport layer protocol to use for the signaling messages encapsulated in CR-LDP / RSVP-TE. The requirements of the transport layer is to provide a reliable channel for transmitting information (both data / control). The IETF Sigtran workgroup came up with designs for a new protocol called SCTP, which could be used in lieu of TCP, and is designed especially for signaling purposes[SCTP]. Like TCP it runs directly over IP but offers some signaling tailored features:

* Datagram-oriented (TCP is byte-stream-oriented)
* Fragmentation and re-assembly for large datagrams
* Multiplexing of several small datagrams into one IP packet
* Support of multi-homing (an SCTP endpoint may have several IP addresses)
* Path monitoring by periodic heartbeat messages
* Retransmission over a different path, if available
* Selective acknowledgements
* Fast retransmit
* 32 bit checksum over the whole payload
* Avoids IP fragmentation due to MTU discovery
* Protection against SYN attacks and blind masquerade attacks

SCTP is far from complete and is quite immature compared to its nemesis TCP. Current implementation of the signaling protocol shall thereby use TCP for its reliable transmissions.

5 Configuration Control using GSMP

In a general mesh network where the OXCs do not participate in topology distribution protocols, General Switch Management Protocol (GSMP) can be used to communicate crossconnect information. This ensures that the OXCs on the lightpath maintain appropriate databases. The first hop router having complete knowledge of LP, L2 and L3 topology acts as the "controller" to the OXCs in the lightpath.

GSMP is a master-slave protocol [GSMP]. The controller issues request messages to the switch. Each request message indicates whether a response is required from the switch (and contains a transaction identifier to enable the response to be associated with the request). The switch replies with a response message indicating either a successful result or a failure. The switch may also generate asynchronous Event messages to inform the controller of asynchronous events.

' Resource Discovery Using NHRP

The Next Hop Resolution Protocol (NHRP) allows a source station (a host or router), wishing to communicate over a Non-Broadcast, Multi-Access (NBMA) subnetwork, to determine the internetworking layer addresses and NBMA addresses of suitable "NBMA next hops" toward a destination station [NHRP]. A subnetwork can be non-broadcast either because it technically doesn't support broadcasting (e.g., an X.25 subnetwork) or because broadcasting is not feasible for one reason or another (e.g., a Switched Multi-megabit Data Service multicast group or an extended Ethernet would be too large).

If the destination is connected to the NBMA subnetwork, then the NBMA next hop is the destination station itself. Otherwise, the NBMA next hop is the egress router from the NBMA subnetwork that is "nearest" to the destination station. NHRP is intended for use in a multiprotocol internetworking layer environment over NBMA subnetworks.

In short, NHRP may be applied as a resource discovery to find the egress OXC in an optical network. To request this information, the existing packet format for the NHRP Resolution Request would be used with a new extension in the form of a modified Forward Transit NHS Extension. The extension would include enough information at each hop (including the source and destination)

* to uniquely identify which wavelength.
* to use when bypassing each routed/forwarded hop and which port that the request was received on.

Essentially a shortcut is setup from ingress to egress using this protocol.

Optical Network Management

The management functionality in all-optical networks is still in the rudimentary phase. Management in a system refers to set of functionalities like performance monitoring, link initialization and other network diagnostics to verify safe and continued operation of the network. The wavelengths in the optical domain will require routing, add/drop, and protection functions, which can only be achieved through the implementation of network-wide management and monitoring capabilities. Current proposals for link initialization and performance monitoring are summarized below.

. Link Initialization

The links between OXCs will carry a number of user bearer channels and possibly one or more associated control channels. This section describes a link management protocol (LMP) that can be run between neighboring OXCs and can be used for both link provisioning and fault isolation. A unique feature of LMP is that it is able to isolate faults independent of the encoding scheme used for the bearer channels. LMP will be used to maintain control channel connectivity, verify bearer channel connectivity, and isolate link, fiber, or channel failures within the optical network.

..1 Control Channel Management

For LMP, it is essential that a control channel is always available for a link, and in the event of a control channel failure, an alternate (or backup) control channel should be made available to reestablish communication with the neighboring OXC. If the control channel cannot be established on the primary (fiber, wavelength) pair, then a backup control channel should be tried. The control channel of a link can be either explicitly configured or automatically selected. The control channel can be used to exchange:

a) MPLS control-plane information such as link provisioning and fault isolation information (implemented using a messaging protocol such as LMP, proposed in this section),

b) path management and label distribution information (implemented using a signaling protocol such as RSVP-TE or CR-LDP), and

c) topology and state distribution information (implemented using traffic engineering extended protocols such as OSPF and IS-IS).

:e a control channel is configured between two OXCs, a Hello protocol ı be used to establish and maintain connectivity between the OXCs and detect link failures. The Hello protocol of LMP is intended to be a ghtweight keep-alive mechanism that will react to control channel .lures rapidly. A protocol similar to the HDLC frame exchange is ed to continue the handshake. [Lang00]

..2 Verifying Link Connectivity

In this section, we describe the mechanism used to verify the physical connectivity of the bearer channels. This will be done initially when a link is established, and subsequently, on a periodic basis for all free bearer channels on the link. To ensure proper verification of bearer channel connectivity, it is required that until the bearer channels are allocated, they should be opaque.

As part of the link verification protocol, the control channel is first verified, and connectivity maintained, using the Hello protocol discussed in Section 5.1.1. Once the control channel has been established between the two OXCs, bearer channel connectivity is verified by exchanging Ping-type Test messages over all of the bearer channels specified in the link. It should be noted that all messages except for the Test message are exchanged over the control channel and that Hello messages continue to be exchanged over the control channel during the bearer channel verification process. The Test message is sent over the bearer channel that is being verified. Bearer channels are tested in the transmit direction as they are unidirectional, and as such, it may be possible for both OXCs to exchange the Test messages simultaneously [Lang00].

..3 Fault Localization

Fault detection is delegated to the physical layer (i.e., loss of light or optical monitoring of the data) instead of the layer 2 or layer 3. Hence, detection should be handled at the layer closest to the failure; for optical networks, this is the physical (optical) layer. One measure of fault detection at the physical layer is simply detecting loss of light (LOL). Other techniques for monitoring optical signals are still being developed.

A link connecting two OXCs consists of a control channel and a number of bearer channels. If bearer channels fail between two OXCs, a mechanism should be used to rapidly locate the failure so that appropriate protection/restoration mechanisms can be initiated. This is discussed further in Section 6.10.

: Optical Performance Monitoring (OPM)

Current-generation WDM networks are monitored, managed, and protected within the digital domain, using SONET and its associated support systems. However, to leverage the full potential of wavelength-based networking, the provisioning, switching, management and monitoring functions have to move from the digital to the optical domain.

The information generated by the performance monitoring operation can be used to ensure safe operation of the optical network. In addition to verifying the service level provided by the network to the user, performance monitoring is also necessary to ensure that the users of the network comply with the requirements that were negotiated between them and the network operator. For example, one function may be to monitor the wavelength and power levels of signals being input to the network to ensure that they meet the requirements imposed by the network. Current performance monitoring in optical networks requires termination of a channel at an optical-electrical-optical conversion point to detect bits related to BER of the payload or frame (e.g., SONET LTE monitoring). However, while these bits indicate if errors have occurred, they do not supply channel-performance data. This makes it very difficult to assess the actual cause of the degraded performance.

Fast and accurate determination of the various performance measures of a wavelength channel implies that measurements have to be done while leaving it in optical format. One possible way of achieving this is by tapping a portion of the optical power from the main channel using a low loss tap of about 1%. In this scenario, the most basic form of monitoring will utilize a power-averaging receiver to detect loss of signal at the optical power tap point. Existing WDM systems use optical time-domain reflectometers to measure the parameters of the optical links.

Another problem lies in determining the threshold values for the various parameters at which alarms should be declared. Very often these values depend on the bit rate on the channel and should ideally be set depending on the bit rate. In addition, since a signal is not terminated at an intermediate node, if a wavelength fails, all nodes along the path downstream of the failed wavelength could trigger an alarm. This can lead to a large number of alarms for a single failure, and makes it somewhat more complicated to determine the cause of the alarm (alarm correlation). A list of such optical parameters to be monitored periodically have been proposed . Optical cross talk, dispersion, and insertion loss are key parameters to name a few.

Care needs to be taken in exchanging these performance parameters. The vast majority of existing telecommunication networks use framing and data formatting overhead as the means to communicate between network elements and management systems. It is worth mentioning that while the signaling is used to communicate all monitoring results, the monitoring itself is done on the actual data channel, or some range of bandwidth around the channel. Therefore, all network elements must be guaranteed to pass this bandwidth in order for monitoring to happen at any point in the network.

One of the options being considered for transmitting the information is the framing and formatting bits of the SONET interface. But, it hampers transparency. It is clear that truly transparent and open photonic networks can only be built with transparent signaling support. The MPLS control plane architecture suggested can be extended beyond simple bandwidth provisioning to include optical performance monitoring.

Fault restoration in Optical networks

Telecom networks have traditionally been designed with rapid fault detection, rapid fault isolation and recovery. With the introduction of IP and WDM in these networks, these features need to be provided in the IP and WDM layers also. Automated establishment and restoration of end-to-end paths in such networks requires standardized signaling, routing, and restoration mechanisms.

Survivability techniques are being made available at multiple layers in the network. Each layer has certain recovery features and one needs to understand the impact of interaction between these layers. The central idea is that the lower layers can provide fast protection while the higher layers can provide intelligent restoration. It is desirable to avoid too many layers with functional overlaps. The IP over MPLS scheme can provide a smooth mapping of IP into WDM layer, thus bringing about a integrated protection/restoration capability, which is coordinated at both the layers.

. Layering

Clearly the layering and architecture for service restoration is a major component for IP to optical internetworking. This section summarizes some schemes, which aid in optical protection at the lower layers, SONET and Optical.

..1 SONET Layer Protection

The SONET standards specify an end-to-end two-way availability objective of 99.98% for inter office applications (0.02% unavailability or 105 minutes/year maximum down time) and 99.99 % for loop transport between the central office and the customer's premises. To conform to these standards, failure/restoration times have to be short. For both, point-to-point and ring systems, automatic protection switching (APS) is used, the network performs failure restoration in tens of milliseconds (approximately 50 milliseconds).

Architectures composed of SONET add-drop multiplexers (ADMs) interconnected in a ring provide a method of APS that allows facilities to be shared while protecting traffic within an acceptable restoration time. There are 2 possible ring architectures:

* UPSR: Unidirectional path switched ring architecture is a 1+1 single-ended, unidirectional, SONET path layer dedicated protection architecture. The nodes are connected in a ring configuration with one fiber pair connecting adjacent nodes. One fiber on a link is used as the working and other is protection. They operate in opposite directions. So there is a working ring in one direction and a protection ring in the opposite direction. The optical signal is sent on both outgoing fibers. The receiver compares the 2 signals and selects the better of the two based on signal quality. This transmission on both fibers is called 1+1 protection.

* BLSR: In bi-directional line switched ring architecture, a bi-directional connection between 2 nodes traverses the same intermediate nodes and links in opposite directions. In contrast to the UPSR, where the protection capacity is dedicated, the BLSR shares protection capacity among all spans on the ring. They are also called Shared Protection ring (SPRing) architectures. In BLSR architecture, switching is coordinated by the nodes on either side of a failure in the ring, so that a signaling protocol is required to perform a line switch and to restore the network. These architectures are more difficult to operate than UPSRs where no signaling is required.

The disadvantage of the SONET layer is that it is usually restricted to ring type architectures. These are extremely bandwidth inefficient. The bandwidth along each segment of the ring ahs to be equal to the bandwidth of the busiest segment. It does not incorporate traffic priorities. It cannot detect higher layer errors.

..2 Optical Layer Protection

The concept of SONET ring architectures can be extended to WDM self-healing optical rings (SHRs). As in SONET, WDM SHRs can be either path switched or line switched. In recent testbed experiments, lithium niobate protection switches have been used to achieve 10-microsecond restoration times in WDM Shared protection Rings. Multi-wavelength systems add extra complexity to the restoration problem. Under these circumstances, simple ring architecture may not suffice. Hence, arbitrary mesh architectures become important. Usually, for such architectures, restoration is usually performed after evaluation at the higher layer. But this takes a lot of time.

..2.1 Point-to-Point Mechanisms

In case of point to point, one can provide 1+1, 1:1 or 1:N protection. In 1+1, the same information is sent through 2 paths and the better one is selected at the receiver. The receiver makes a blind switch when the selected (working) path's signal is poor. Unlike SONET, a continuous comparison of 2 signals is not done in the optical layer. In 1:1 protection, signal is sent only on the working path while a protection path is also set but it can be used for lower priority signals that are preempted if the working path fails. A signaling channel is required to inform the transmitter to switch path if the receiever detects a failure in the working path. A generalization of 1:1 protection is 1:N protection in which one protection fiber is shared among N working fibers. It is usually applied for equipment protection [JOHNSON99].

..2.2 Ring systems [MANCHESTER99]

Ring mechanisms are broadly classified into: Dedicated linear protection and Shared protection rings.

Dedicated linear protection is an extension of 1+1 protection applied to a ring. It is effectively a path protection mechanism. Entire path from source to the destination node is protected. Since each channel constitutes a separate path, it is also called Optical Channel Subnetwork Connection Protection (OCh-SNCP). From each node, the working and protection signals are transmitted in opposite directions along the 2 fibers. At the receiving end, if the working path signal is weak, the receiver switches to the protection path signal. Bidirectional traffic between two nodes, travels along the same direction of the ring. The ring through put is restricted to that of a single fiber. This is usually applied to hubbed transport scenarios, near access rings. For other types of connections, it is very expensive [GERSTEL00].

Shared protection rings (SPRings) protect a link rather than a path. This is conceptually similar to the SONET BLSR architecture. Bidirectional traffic between 2 nodes travels along opposite directions and between the same intermediate nodes. The wavelength used in one part of the network can be re-used in another non-overlapping part of the ring also. Thus this permits reusing of wavelengths. Moreover, unless there is a fault, only half the capacity is used at any time. So the protection bandwidth can be used by a some other traffic. These are also easier to setup and are the more common ring protection mechanisms.

In a 2-fiber SPRings case with two counter-rotating rings, half the wavelengths in each fiber are reserved for protection. If a link failure occurs, the OADM adjacent to the link failure bridges its outgoing channels in a direction opposite to that of the failure and selects its incoming working channels from the incoming protection channels in the direction away from the failure. This is called ring switching.

In a 4-fiber SPRing, two fibers each are allocated for working and protection. The operation is similar to that of the 2-fiber SPRing. However, this system can allow span switching in addition to ring switching. Span switching means that if only the working fiber in a link fails, the traffic can use the protection fiber in the same span. In case of 2-fiber systems, it will have to take the longer path around the ring.

The 2 fiber and 4 fiber SPRing architectures have signaling complexities associated with them, because these rings perform switching at intermediate nodes.

Sometimes the need arises to protect against isolated optoelectronic failures that will affect only a single optical channel at a time. Thus, we need a protection architecture that performs channel level switching based on channel level indications. The Optical Multiplexed Section (OMS) SPRings, discussed so far, switch a group of channels within the fiber. The Optical Channel (OCh) SPRings are capable of protecting OChs independent of one another based on OCh level failure indication. An N-Channel OADM based 4-fiber ring can support upto N independent OCh SPRings.

SPRing architectures are referred to as Bidirectional line switched ring (BLSR) architectures. OCh SPRings are referred to as Bidirectional Wavelength Line Switched Ring technology, (BWLSR). ITU-T draft recommendation G.872 describes a transoceanic switching protocol for 4-fiber OMS SPRings. This protocols requires that after a span switching a path should not traverse any span more than once. When ring switching occurs, this may not be true. This protocol is essential in long-distance undersea transmissions to avoid unnecessary delay.

..2.3 Mesh Architectures

Along a single fiber, any two connections cannot use the same wavelength. The whole problem of routing in a WDM network with proper allocation of a minimum number of wavelengths is called the routing and wavelength assignment (RWA) problem. It is found that in arbitrary mesh architectures, where the connectivity of each node is high, the number of wavelengths required greatly decreases. This is the advantage of having a mesh architecture. Moreover addition of new nodes and removing existing nodes becomes very easy. However, with mesh architectures, finding an alternate path every time a failure occurs would be a time consuming process. Hence, an automatic protection switching mechanism, like that for the rings, is required. Three alternatives are briefly discussed here:

Ring Covers

The whole mesh configuration is divided into smaller cycles in such a way that each edge comes under atleast one cycle. Along each cycle, a protection fiber is laid. It may so happen that certain edges come under more than one cycle. In these edges, more than one protection fiber will have to be laid. Hence, the idea is to divide the graph into cycles in such a way that this redundancy is minimized [WU]. However, in most cases the redundancy required is more than 100%.

Protection Cycles [ELLINAS]

This method reduces the redundancy to exactly 100%. The networks considered have a pair of bi-directional working and protection fibers. Fault protection against link failures is possible in all networks that are modeled by 2-edge connected digraphs. The idea is to find a family of directed cycles so that all protection fibers are used exactly once and in any directed cycle a pair of protection fibers is not used in both directions unless they belong to a bridge.

For planar graphs, such directed cycles are along the faces of the graph. For non-planar graphs, the directed cycles are taken along the orientable cycle double covers, which are conjectured to exist for every digraph. Heuristic algorithms exist for obtaining cyclic double covers for every non-planar graph.

Thus, the main advantage of optical layer mechanisms is the fast restoration. It also has the capacity of large switching granularity in the sense that it can restore a large number of higher layer flows by a single switching.

The disadvantage is that it cannot carry traffic engineering capabilities. It can only operate at the lightpath level and cannot differentiate between different data types. Also the switching speed comes into play only if all the nodes which can detect a fault have switching capabilities. Building such an architecture is extremely expensive.

..3 IP Layer Protection

The IP layer plays a major role in the IP network infrastructure. There are some advantages of having survivability mechanism in this layer. It can find optimal routes in the system. It provides a finer granularity at which protection can be done, enabling the system to have priorities. It also possesses load balancing capabilities.

However the recovery operations are very slow. It also cannot detect physical layer faults.

..4 MPLS Layer Protection

The rerouting capability of the optical layer can be expanded and newer bandwidth efficient protection can be facilitated if there is some controlled coordination between the optical layer and a higher layer that has a signaling mechanism.

Similarly, the optical layer which cannot detect faults in the router or switching node, could learn of the faults if the higher layer communicated this to it. Then, the optical layer can initiate protection at the lower layer.

Fast signaling is the main advantage of the MPLS layer in protection. Since MPLS binds packets to a route (or path) via the labels, it is imperative that MPLS be able to provide protection and restoration of traffic. In fact, a protection priority could be used as a differentiating mechanism for premium services that require high reliability. The MPLs layer has visibility into the lower layer. The lower layer can inform this layer about faults by a liveness message, basically signaling.

When we talk of the IP/MPLS over WDM architecture, we may seal off SONET APS protection from the discussion and the WDM optical layer can provide the same kind of restoration capabilities at the lower layer. Thus there has to be interaction only between the MPLS and optical layer and not with the SONET layer.

The following sections present a summary of techniques being proposed for implementing survivability in the MPLS layer. These include signaling requirements, architectural considerations and timing considerations.

: Failure detection [OWENS00]

Loss of Signal (LOS) is a lower layer impairment that arises when a signal is not detected at an interface, for example, a SONET LOS. In this case, enough time should be provided for the lower layer to detect LOS and take corrective action.

A Link Failure (LF) is declared when the link probing mechanism fails. An example of a probing mechanism is the Liveness message that is exchanged periodically along the working path between peer LSRs. A LF is detected when a certain number k of consecutive Liveness messages are either not received from a peer LSR or are received in error.

A Loss of Packets (LOP) occurs when there is excessive discarding of packets at an LSR interface, either due to label mismatches or due to time-to-live (TTL) errors. LOP due to label mismatch may be detected simply by counting the number of packets dropped at an interface because an incoming label did not match any label in the forwarding table. Likewise, LOP due to invalid TTL may be detected by counting the number of packets that were dropped at an interface because the TTL decrements to zero.

3 Failure Notification [OWENS00]

Protection switching relies on rapid notification of failures. Once a failure is detected, the node that detected the failure must send out a notification of the failure by transmitting a failure indication signal (FIS) to those of its upstream LSRs that were sending traffic on the working path that is affected by the failure. This notification is relayed hop-by-hop by each subsequent LSR to its upstream neighbor, until it eventually reaches a PSL.

The PSL is the LSR that originates both the working and protection paths, and the LSR that is the termination point of both the FIS and the failure recovery signal (FRS). Note that the PSL need not be the origin of the working LSP.

The PML is the LSR that terminates both the working path and its corresponding protection path. Depending on whether or not the PML is a destination, it may either pass the traffic on to the higher layers or may merge the incoming traffic on to a single outgoing LSR. Thus, the PML need not be the destination of the working LSP.

An LSR that is neither a PSL nor a PML is called an intermediate LSR. The intermediate LSR could be either on the working or the protection path, and could be a merging LSR (without being a PML).

3.1 Reverse Notification Tree (RNT)

Since the LSPs are unidirectional entities and protection requires the notification of failures, the failure indication and the failure recovery notification both need to travel along a reverse path of the working path from the point of failure back to the PSL(s). When label merging occurs, the working paths converge to form a multipoint-to-point tree, with the PSLs as the leaves and the PML as the root. The reverse notification tree is a point-multipoint tree rooted at the PML along which the FIS and the FRS travel, and which is an exact mirror image of the converged working paths.

The establishment of the protection path requires identification of the working path, and hence the protection domain. In most cases, the working path and its corresponding protection path would be specified via administrative configuration, and would be established between the two nodes at the boundaries of the protection domain (the PSL and PML) via explicit (or source) routing using LDP, RSVP, signaling (alternatively, using manual configuration).

The RNT is used for propagating the FIS and the FRS, and can be created very easily by a simple extension to the LSP setup process. During the establishment of the working path, the signaling message carries with it the identity (address) of the upstream node that sent it. Each LSR along the path simply remembers the identity of its immediately prior upstream neighbor on each incoming link. The node then creates an inverse crossconnect table that for each protected outgoing LSP maintains a list of the incoming LSPs that merge into that outgoing LSP, together with the identity of the upstream node that each incoming LSP comes from. Upon receiving an FIS, an LSR extracts the labels contained in it (which are the labels of the protected LSPs that use the outgoing link that the FIS was received on) consults its inverse crossconnect table to determine the identity of the upstream nodes that the protected LSPs come from, and creates and transmits an FIS to each of them.

4 Protection options [SHARMA00]

When using the MPLS layer for providing survivability, we can have different options, just like in any other layer. Each has its own advantages depending on requirements.

4.1 Dynamic Protection

These protection mechanisms dynamically create protection paths for restoring traffic, based upon failure information, bandwidth allocation and optimized reroute assignment. Thus, upon detecting failure, the LSPs crossing a failed link or LSR are broken at the point of failure and reestablished using signaling. These methods may increase resource utilization because capacity or bandwidth is not reserved beforehand and because it is available for use by other (possibly lower priority) traffic, when the protection path does not require this capacity. They may, however, require longer restoration times, since it is difficult to instantaneously switch over to a protection entity, following the detection of a failure.

4.2 Pre-negotiated Protection

These are dedicated protection mechanisms, where for each working path there exists a pre-established protection path, which is node and link disjoint with the primary/working path, but may merge with other working paths that are disjoint with the primary. The resources (bandwidth, buffers, processing) on the backup entity may be either pre-determined and reserved beforehand (and unused), or may be allocated dynamically by displacing lower priority traffic that was allowed to use them in the absence of a failure on the working path.

4.3 End-to-end Repair

In end-to-end repair, upon detection of a failure on the primary path, an alternate or backup path is re-established starting at the source. Thus, protection is always activated on an end-to-end basis, irrespective of where along a working path a failure occurs. This method might be slower than the local repair method discussed below, since the failure information has to propagate all the way back to the source before a protection switch is accomplished.

4.4 Local Repair

In local repair, upon detecting a failure on the primary path, an alternate path is re-established starting from the point of failure. Thus protection is activated by each LSR along the path in a distributed fashion on an as-needed basis. While this method has an advantage in terms of the time taken to react to a fault, it introduces the complication that every LSR along a working path may now have to function as a protection switch LSR (PSL).

4.5 Link Protection

The intent is to protect against a single link failure. For example, the protection path may be configured to route around certain links deemed to be potentially risky. If static configuration is used, several protection paths may be pre-configured, depending on the specific link failure that each protects against. Alternatively, if dynamic configuration is used, upon the occurrence of a failure on the working path, the protection path is rebuilt such that it detours around the failed link.

4.6 Path Protection

The intention is to protect against any link or node failure on the entire working path. This has the advantage of protecting against multiple simultaneous failures on the working path, and possibly being more bandwidth efficient than link protection.

4.7 Revertive Mode

In the revertive mode of operation, the traffic is automatically restored to the working path once repairs have been affected, and the PSL(s) are informed that the working path is up. This is useful, since once traffic is switched to the protection path it is, in general, unprotected. Thus, revertive switching ensures that the traffic remains unprotected only for the shortest amount of time. This could have the disadvantage, however, of producing oscillation of traffic in the network, by altering link loads.

4.8 Non-revertive Mode

In the non-revertive mode of operation, traffic once switched to the protection path is not automatically restored to the working path, even if the working path is repaired. Thus, some form of administrative intervention is needed to invoke the restoration action. The advantage is that only one protection switch is needed per working path. A disadvantage is that the protection path remains unprotected until administrative action (or manual reconfiguration) is taken to either restore the traffic back to the working path or to configure a backup path for the protection path.

4.9 1+1 Protection

In 1+1 protection, the resources (bandwidth, buffers, processing capacity) on the backup path are fully reserved to carry only working traffic. This bandwidth is used to transmit an exact copy of the working traffic, with a selection between the traffic on the working and protection paths being made at the protection merge LSR (PML).

4.10 1:1, 1:n, and n:m Protection

In 1:1 protection, the resources (bandwidth, buffers, and processing capacity) allocated on the protection path are fully available to preemptable low priority traffic when the protection path is not in use by the working traffic. In other words, in 1:1 protection, the working traffic normally travels only on the working path, and is switched to the protection path only when the working entity is unavailable. Once the protection switch is initiated, all the low priority traffic being carried on the protection path is discarded to free resources for the working traffic. This method affords a way to make efficient use of the backup path, since resources on the protection path are not locked and can be used by other traffic when the backup path is not being used to carry working traffic.

Similarly, in 1:n protection, up to n working paths are protected using only one backup path, while in m:n protection, up to n working paths are protected using up to m backup paths.

4.11 Recovery Granularity

Another dimension of recovery considers the amount of traffic requiring protection. This may range from a fraction of a path to a bundle of paths.

4.11.1 Selective Traffic Recovery

This option allows for the protection of a fraction of traffic within the same path. The portion of the traffic on an individual path that requires protection is called a protected traffic portion (PTP). A single path may carry different classes of traffic, with different protection requirements. The protected portion of this traffic may be identified by its class, as for example, via the EXP bits in the MPLS shim header or via the cell loss priority (CLP) bit in the ATM header.

.11.2 Bundling

Bundling is a technique used to group multiple working paths together in order to recover them simultaneously. The logical bundling of multiple working paths requiring protection, each of which is routed identically between a PSL and a PML, is called a protected path group (PPG). When a fault occurs on the working path carrying the PPG, the PPG as a whole can be protected either by being switched to a bypass tunnel or by being switched to a recovery path.

Signaling Requirements related to restoration [SAHA00]

Signaling mechanisms for optical networks should be tailored to the needs of optical networking.

Some signaling requirements directed towards restoration in optical networks are:

1. Signaling mechanisms should minimize the need for manual configuration of relevant information, such as local topology.

2. Lightpaths are fixed bandwidth pipes. There is no need to convey complex traffic characterization or other QoS parameters in signaling messages. On the other hand, new service related parameters such as restoration priority, protection scheme desired, etc., may have to be conveyed.

3. Signaling for path establishment should be quick and reliable. It is especially important to minimize signaling delays during restoration.

4. Lightpaths are typically bi-directional. Both directions of the path should generally be established along the same physical route.

5. OXCs are subject to high reliability requirements. A transient failure that does not affect the data plane of the established paths should not result in these paths being torn down.

6. Restoration schemes in mesh networks rely on sharing backup path among many primary paths. Signaling protocols should support this feature.

7. The interaction between path establishment signaling and automatic protection schemes should be well defined to avoid false restoration attempts during path set-up or tear down.

Pre-computed, Priority-Based Restoration

The previous sections have discussed so far, the different requirements for restoration in optical networks and has seen a number of methods possible. This section tries to summarize that and brings together the best of the options.

A simple restoration strategy is possible for rings. But the mesh architectures promise flexible use of bandwidth. Hence the goal is to find a solution to provide fast alternate paths in a mesh based optical network.

The optical network will be surrounded by edge switches, which are the entry and exit points for wavelength paths. Hence, these edge switches will compute the path through the optical network from source to destination. They shall also have the task of having an alternate path ready, incase of faults in the network. Each of the switches inside the network are called core switches. In case of a fault, these switches should propagate the information back to the entry switch.

If the edge switch tries to obtain an alternate path on the spur of the moment, it will be time consuming. Hence a pre-computation strategy would work better.

Link based restoration methods re-route disrupted traffic around the failed link. This mechanism saves some signaling time, but it requires alternate paths from each node. Computation is tougher. Also restoration would be tougher in case of a node failure. So a path based re-routing is sought, which replaces the whole path between two end points.

The selection of the protection path should be such that, the links along working and protection path should be mutually exclusive. Also, in case of any single failure, the total bandwidth on any of the diverted links should not exceed its capacity. Algorithms for alternate path finding are discussed in [Bell-Labs].

The next thing is to incorporate priority inside the procedure. When an edge swich gets a request to route a traffic through the optical network, it will include priority information. Let these be categorized into 3 levels, 1, 2 and 3 from highest to lowest.

For traffic no. 1, the switch will compute 2 paths. Also, if the protection path is not found, it will preempt a lower priority traffic and establish 2 paths during the creation phase itself. In other words, this would be a 1+1 style protection.

For traffic no.2, a working path will be established. The protection path will be pre-computed, but need not necessarily be available in terms of bandwidth or wavelength at the time of creation. All the switches along the path would be pre-configured with the information. In case of failure, the lower priority path along those switches would be preempted and the traffic no. 2 would be restored.

If this preemption capacity is built into the switches itself, that would be the fastest and at the optical layer itself. Otherwise, some time is lost in signaling. But still, this can meet the 50mS requirement set by SONET.

Traffic no.3 has the lowest priority. It has no requested back up paths. It is set up along the backup paths of the existing traffic 2 working paths, unless extra bandwidth is available.

Thus bandwidth is used efficiently, in providing restoration and classes are considered in the above mechanism.

Other protection priorities like longer protection paths and shorter paths can also be taken into account while setting up the paths. Also, given a steady traffic flow, with no new paths being created, algorithms to optimize the paths selected would enhance the performance of the network.

The next thing required is the signaling messages to set up these paths and release them. [Hahm00]

' RSVP-TE/CR-LDP Support for Restoration [BALA00]

Special requirements for protecting and restoring lightpaths and the extensions to RSVP-TE and CR-LDP have been identified. Some of the proposed extensions are as follows:

a. A new SESSION_ATTRIBUTE object has been proposed, which indicates whether the path is unidirectional/bi-directional, primary/backup. Local protection 1+1 or 1:N can also be specified.
b. Setup Priority: The priority of the session with respect to taking resources. The Setup Priority is used in deciding whether this session can preempt another session.
c. Holding Priority: The priority of the session with respect to holding resources. Holding Priority is used in deciding whether this session can be preempted by another session.

Note that for the shared backup paths the crossconnects can not be setup during the signaling for the backup path since multiple backup paths may share the same resource and can over-subscribe it. The idea behind shared backups is to make soft reservations and to claim the resource only when it is required.

Security Considerations

This document raises no new security issues for MPL(ambda) Switching implementation over optical networks. Security considerations are for future study.

Acronyms

```
3R   - Regeneration with Retiming and Reshaping
AIS  - Alarm Indication Signal
APS  - Automatic Protection Switching
BER  - Bit Error Rate
BGP  - Border Gateway Protocol
BLSR  - Bi-directional Line-Switched Ring
CR-LPD  - Constraint-Based Routing LDP
CSPF  - Constraint Shortest Path First
FA   - Forwarding Adjacency
FA-LSP  - Forwarding Adjacency Label Switched Path
FA-TDM  - Time Division Multiplexing capable Forwarding Adjacency
FA-LSC  - Lambda Switch Capable Forwarding Adjacency
FA-PSC  - Packet Switch Capable Forwarding Adjacency
FA-FSC  - Fiber Switch Capable Forwarding Adjacency
FEC  - Forwarding Equivalence Class
FIS  - Failure Indication Signal
FRS  - Failure Recovery Signal
GSMP  - General Switch Management Protocol
IGP  - Interior Gateway Protocol
IS-IS  - Intermediate System to Intermediate System Protocol
ITU-T  - International Telecommunications Union - Telecommunications
        Sector
LDP  - Label Distribution Protocol
LF  - Link Failure
LMP  - Link Management Protocol
LMT  - Link Media Type
LOL  - Loss of Light
LOP  - Loss of Packets
LOS  - Loss Of Signal
LP  - Lightpath
LSA  - Link State Advertisement
LSC  - Lambda Switch Capable
LSP  - Label Switched Path
LSR  - Label Switched Router
MPLS - Multi-Protocol Lambda Switching
MTG  - MPLS Traffic Group
NBMA - Non-Broadcast Multi-Access
NHRP - Next Hop Resolution Protocol
OCT  - Optical Channel Trail
OLXC  - Optical layer crossconnect
OMS  - Optical Multiplex Section
OPM  - Optical Performance Monitoring
OSPF  - Open Shortest Path First
OTN  - Optical Transport Network
OTS  - Optical Transmission Section
OXC  - Optical Crossconnect
PML  - Protection Merge LSR
PMTG  - Protected MPLS Traffic Group
PMTP  - Protected MPLS Traffic Portion
PPG  - Protected Path Group
```

PSC - Packet Switch Capable
PSL - Protection Switch LSR
PTP - Protected Traffic Portion
PVC - Permanent Virtual Circuit
PXC - Photonic Crossconnect
QoS - Quality of Service
RNT - Reverse Notification Tree
RSVP - Resource reSerVation Protocol
RSVP-TE - Resource reSerVation Protocol with Traffic Engineering
SHR - Self-healing Ring
SPRing - Shared Protection ring
SRLG - Shared Risk Link Group
TDM - Time Division Multiplexing
TE   - Traffic Engineering
TLV - Type Length Value
TTL - Time to Live
UNI - User to Network Interface
UPSR - Unidirectional Path-Switched Ring
VC   - Virtual Circuit
WDM - Wavelength Division Multiplexing

Terminology

Channel:
A channel is a unidirectional optical tributary connecting two OLXCs. Multiple channels are multiplexed optically at the WDM system. One direction of an OC-48/192 connecting two immediately neighboring OLXCs is an example of a channel. A channel can generally be associated with a specific wavelength in the WDM system. A single wavelength may transport multiple channels multiplexed in the time domain.

Downstream node:
In a unidirectional lightpath, this is the next node closer to destination.

Failure Indication Signal:
A signal that indicates that a failure has been detected at a peer LSR. It consists of a sequence of failure indication packets transmitted by a downstream LSR to an upstream LSR repeatedly. It is relayed by each intermediate LSR to its upstream neighbor, until it reaches an LSR that is setup to perform a protection switch.

Failure Recovery Signal:
A signal that indicates that a failure along the path of an LSP has been repaired. It consists of a sequence of recovery indication packets that are transmitted by a downstream LSR to its upstream LSR, repeatedly. Again, like the failure indication signal, it is relayed by each intermediate LSR to its upstream neighbor, until is reaches the LSR that performed the original protection switch.

First-hop router:
The first router within the domain of concern along the lightpath route. If the source is a router in the network, it is also its own first-hop router.

Intermediate LSR:
LSR on the working or protection path that is neither a PSL nor a PML.

Last-hop router:
The last router within the domain of concern along the lightpath route. If the destination is a router in the network, it is also its own last-hop router.

Lightpath:
This denotes an Optical Channel Trail in the context of this document. See "Optical Channel Trail" later in this section.

Link Failure:
A link failure is defined as the failure of the link probing mechanism, and is indicative of the failure of either the underlying physical link between adjacent LSRs or a neighbor LSR itself. (In case of a bi-directional link implemented as two unidirectional links, it could mean that either one or both unidirectional links are damaged.)

Liveness Message:
A message exchanged periodically between two adjacent LSRs that serves as a link probing mechanism. It provides an integrity check of the forward and the backward directions of the link between the two LSRs as well as a check of neighbor aliveness.

Loss of Signal:
A lower layer impairment that occurs when a signal is not detected at an interface. This may be communicated to the MPLS layer by the lower layer.

Loss of Packet:
An MPLS layer impairment that is local to the LSR and consists of excessive discarding of packets at an interface, either due to label mismatch or due to TTL errors. Working or Active LSP established to carry traffic from a source LSR to a destination LSR under normal conditions, that is, in the absence of failures. In other words, a working LSP is an LSP that contains streams that require protection.

MPLS Traffic Group:
A logical bundling of multiple, working LSPs, each of which is routed identically between a PSL and a PML. Thus, each LSP in a traffic group shares the same redundant routing between the PSL and the PML.

MPLS Protection Domain:

The set of LSRs over which a working path and its corresponding protection path are routed. The protection domain is denoted as: (working path, protection path).

Non-revertive:
A switching option in which streams are not automatically switched back from a protection path to its corresponding working path upon the restoration of the working path to a fault-free condition.

Opaque:
Used to denote a bearer channel characteristic where it is capable of being terminated.

Optical Channel Trail:
The elementary abstraction of optical layer connectivity between two end points is a unidirectional Optical Channel Trail. An Optical Channel Trail is a fixed bandwidth connection between two network elements established via the OLXCs. A bi-directional Optical Channel Trail consists of two associated Optical Channel Trails in opposite directions routed over the same set of nodes.

Optical layer crossconnect (OLXC):
A switching element which connects an optical channel from an input port to an output port. The switching fabric in an OLXC may be either electronic or optical.

Protected MPLS Traffic Group (PMTG):
An MPLS traffic group that requires protection.

Protected MPLS Traffic Portion:
The portion of the traffic on an individual LSP that requires protection. A single LSP may carry different classes of traffic, with different protection requirements. The protected portion of this traffic may be identified by its class, as for example, via the EXP bits in the MPLS shim header or via the priority bit in the ATM header.

Protection Merge LSR:
LSR that terminates both a working path and its corresponding protection path, and either merges their traffic into a single outgoing LSP, or, if it is itself the destination, passes the traffic on to the higher layer protocols.

Protection Switch LSR:
LSR that is the origin of both the working path and its corresponding protection path. Upon learning of a failure, either via the FIS or via its own detection mechanism, the protection switch LSR switches protected traffic from the working path to the corresponding backup path.

Protection or Backup LSP (or Protection or Backup Path):
A LSP established to carry the traffic of a working path (or paths) following a failure on the working path (or on one of the working

paths, if more than one) and a subsequent protection switch by the PSL. A protection LSP may protect either a segment of a working LSP (or a segment of a PMTG) or an entire working LSP (or PMTG). A protection path is denoted by the sequence of LSRs that it traverses.

Reverse Notification Tree:
A point-to-multipoint tree that is rooted at a PML and follows the exact reverse path of the multipoint-to-point tree formed by merging of working paths (due to label merging). The reverse notification tree allows the FIS to travel along its branches towards the PSLs, which perform the protection switch.

Revertive:
A switching option in which streams are automatically switched back from the protection path to the working path upon the restoration of the working path to a fault-free condition.

Soft state:
It has an associated time-to-live, and expires and may be discarded once that time is passed. To avoid expiration the state should be periodically refreshed. To reduce the overhead of the state maintenance, the expiration period may be increased exponentially over time to a predefined maximum. This way the longer a state has survived the fewer the number of refresh messages that are required.

Traffic Trunk:
An abstraction of traffic flow that follows the same path between two access points which allows some characteristics and attributes of the traffic to be parameterized.

Upstream node:
In a unidirectional lightpath, this is the node closer to the source.

Working or Active Path:
The portion of a working LSP that requires protection. (A working path can be a segment of an LSP (or a segment of a PMTG) or a complete LSP (or PMTG).) The working path is denoted by the sequence of LSRs that it tranverses.

References

[Awuduche] D. Awduche, Y. Rekhter, J. Drake, R. Coltun, "Multi-Protocol Lambda Switching: Combining MPLS Traffic Engineering Control With Optical Crossconnects," Internet Draft draft-awduche-mpls-te-optical-02.txt, Work in Progress, July 2000.

[BALA00] Bala rajagopalan, D.Saha, B.tang, "RSVP extensions for signaling optical paths," Internet Draft draft-saha-rsvp-optical-signalling-00.txt, Work in Progress, September 2000.

[Bell-Labs] B. Doshi, S. Dravida, P. Harshavardhana, et. al,"Optical Network Design and Restoration," Bell Labs Technical Journal, Jan-March, 1999.

[CRLDP] B. Jamoussi, et. al. "Constraint-Based LSP Setup using LDP," Internet Draft draft-ietf-mpls-cr-ldp-04.txt, Work in Progress, July 2000.

[ELLINAS] G. N. Ellinas, "Fault Restoration in Optical Networks: General Methodology and Implementation," PhD thesis, Columbia University.

[GERSTEL00] Gerstel and R. Ramaswami, "Optical layer survivability: A Services Perspective," *IEEE Communications,* March 2000, pp.104 - 113.

[GHANI01] N.Ghani et al.,"Architectural Framework for automatic protection provisioning in dynamic optical rings," Internet draft, draft-ghani-optical-rings-00.txt, Work in Progress, January 2000.

[GMPLS00] P. Ashwood-Smith *et al.*, "Generalized MPLS - Signaling Functional Description," Internet Draft draft-ietf-mpls-generalized-mpls-signaling-01.txt, Work in progress, November 2000.

[GMPLS-CONTROL] Y. Xu et al, "GMPLS Control Plane Architecture for Automatic Switched Transport Network," Nov 2000

[GSMP] A. Doria, et. al. "General Switch Management Protocol V3,"Internet Draft draft-ietf-gsmp-08.txt, Work in Progress, November 2000.

[HAHM00] Jin Ho hahm, K.Lee, "Bandwidth provisioning and restoration mechanism in Optical networks", Internet draft, draft-hahm-optical-restoration-01.txt, Work in Progress, December 2000.

[ISIS] ISO 10589, "Intermediate System to Intermediate System Intra-Domain Routing Exchange Protocol for use in Conjunction with the Protocol for Providing the Connectionless-mode Network Service."

[ISISTE] Henk Smit, Tony Li, "IS-IS extensions for Traffic Engineering," Internet Draft, draft-ietf-isis-traffic-02.txt, work in progress, March 2000.

[Johnson99] D. Johnson, N. Hayman and P. Veitch, "The Evolution of a Reliable Transport Network," *IEEE Communications* , August 1999, pp. 52-57.

[Kompella00-b] Kompella, K., Rekhter, Y., "Link Bundling in MPLS Traffic Engineering," draft-kompella-mpls-bundle-04.txt, Work in Progress, November 2000.

[Lang00] J.P. Lang, "Link Management Protocol (LMP)," Internet Draft draft-lang-mpls-lmp-02.txt, Work in Progress, July 2000.

[Luciani00] J. Luciani, B. Rajagopalan, D. Awuduche, B. Cain, Bilel Jamoussi, Debanjan Saha, "IP Over Optical Networks - A Framework," Internet Draft draft-many-ip-optical-framework-01.txt, Work in Progress, November 2000.

[MANCHESTER99] J. Manchester, P. Bonenfant and C. Newton, "The Evolution of Transport Network Survivability," *IEEE Communications*, August 1999, pp. 44-51.

[NHRP] Luciani, et. al. "NBMA Next Hop Resolution Protocol (NHRP)," RFC 2332, April 1998.

[ODSI00] G.Bernstein et. al., "Optical Domain Service Interconnect (ODSI) Functional Specification," ODSI Coalition, April 2000.

[OSPF] Moy, J., _"OSPF Version 2," RFC 1583, March 1994

[OWENS00] Ken Owens, Srinivas Makam, Vishal Sharma, Ben Mack-Crane, Changcheng Huan, "A Path Protection/Restoration mechanism for MPLS networks," Internet Draft draft-chang-mpls-path-protection-02.txt, Work in progress, November 2000

[Pendarakis00] D. Pendarakis, B. Rajagopalan, D. Saha, "Routing Information Exchange in Optical Networks," Internet Draft draft-prs-optical-routing-01.txt, Work in progress, November 2000.

[SAHA00] B. Rajagopalan, D.Saha, B. Tang, K. Bala , "Signaling framework for automated provisioning and restoration of paths in optical mesh networks," Internet Draft draft-rstb-optical-signaling-framework-01.txt

[SCTP] R.R. Stewart *et al.*, "Stream Control Transmission Protocol," RFC 2960, October 2000.

[SHARMA00] Vishal Sharma et al. "Framework for MPLS-based recovery," Internet Draft draf-ietf-mpls-recovery-frmwrk-01.txt, Work in progress, November 2000

[Suurballe] J. Suurballe, "Disjoint Paths in a Network," Networks, vol. 4, 1974.

[UNI00] O. S. Aboul-Magd *et al.*, "Signaling Requirements at the Optical UNI," Internet Draft draft-bala-mpls-optical-uni-signaling-01.txt, Work in Progress, November 2000.

[WU] T.H. Wu, "A Passive Protected Self Healing Mesh Network Architecture and Applications," *IEEE Transactions on Networking*, **Vol. 2**, No. 1, February 1994.

Author's Addresses

S. Seetharaman, A. Durresi, R. Jagannathan, N. Chandhok, K. Vinodkrishnan
Department of Computer and Information Science
The Ohio State University
2015 Neil Avenue, Columbus, OH 43210-1277, USA
Phone: (614)-292-3989
Email: {seethara, durresi, rjaganna, chandhok, vinodkri}@cis.ohio-state.edu

Raj Jain
Nayna Networks, Inc.
157 Topaz Street
Milpitas, CA 95035
Phone: (408)-956-8000X309
Email: raj@nayna.com

l Copyright Statement