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

ernet Draft
ument: draft-osu-ipo-MPLS-issues-00.txt   July 2000
egory: Informational

IP over Optical Networks: A Summary of Issues

atus of this Memo

stract

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) has 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 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.

ventions 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.

tents:

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 an 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 optical channel trails, 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)
```

Fig.1: Layering Architectures Possible

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

The network model consists of IP routers attached to an optical core network. The optical network consists of multiple optical crossconnects (OXCs) interconnected by optical links. Each OXC is capable of switching a data stream using a switching function, controlled by appropriately configuring a crossconnect table. Thus, in this document, the term OXC is used to denote the hybrid node consisting of switching element referred to as optical layer crossconnect (OLXC) and a control plane. The switching within the OXC can be accomplished either in the electrical domain, or in the optical domain. In this network model, a switched lightpath is established between IP routers. Designing an IP-based control plane should include designing standard signaling and routing protocols for coherent end-to-end provisioning and restoration of lightpaths across multiple optical sub-networks, and determining IP reachability and seamless establishment of paths from one IP end-point to another over an optical core network.

Several standards organizations and interoperability forums have initiated work items to study the requirements and architectures for reconfigurable optical networks, under-scoring the importance of versatile networking capabilities in the optical domain. ITU-T recommendation G.872, for example, defines a functional architecture for an optical transport network (OTN) that supports the transport of digital client signals. It defines OTN as "a transport network bounded by optical channel access points". The architecture of G.872's OTN is based on a layered structure, which includes:
(a) an optical channel (OCh) layer network : The optical channel layer network supports end-to-end networking of optical channel trails providing functionalities like routing, monitoring, grooming, and protection and restoration of optical channels.

(b) an optical multiplex section (OMS) layer network : The optical multiplex section layer provides the transport of the optical channels. The information contained in this layer is a data stream comprising a set of n optical channels, which have a defined aggregate bandwidth.

(c) an optical transmission section (OTS) layer network : This provides functionality for transmission of the optical signals on optical media of different types.

To realize the functions in the OCh layer, an optical crossconnect with rearrangeable switch fabrics and a control plane will be critical. In the existing IP-centric data network domain, the functionalities of the OCh layer are performed by the MPLS traffic engineering control plane. Thus, there is a similarity between the IP/MPLS over WDM and the ITU recommendation.

In the following section, we stress on the relations that exist between the all-optical crossconnects of the optical networks and the label switch routers of the MPLS networks and identify how the control plane model of MPLS traffic engineering (TE) can be applied to that of optical transport networks. Before we propose a control plane model for the optical networks based on the MPLS control plane traffic engineering, we discuss how the similarities can help to expose the reusable software artifacts from the MPLS traffic engineering control plane model. Consider an IP-centric hybrid optical internetworking environment, which consist of both LSRs and OXCs. Let us assume that OXCs are programmable and support wavelength conversion.

Optical Switch Architecture

Multiprotocol 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). Each LSP has a set of criteria associated with it, which describes the traffic that traverses the LSP. This set of criteria groups the incoming traffic into classed called "Forwarding Equivalence Classes (FECs)." LSPs are setup using signaling protocols like RSVP or CR-LDP. A device that can classify traffic into FECs is called a label edge router (LER) while the devices which base their forwarding decision only on the basis of the incoming labels (and ports) are called Label Switched Routers (LSRs).

Here we consider a hybrid, IP-centric optical internetworking environment consisting of both label switching routers (LSRs) and OXCs. The OXCs are programmable and 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.

. Isomorphic Relations 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 an optical channel trail from an input port to an output port. An LSR performs label switching by establishing a relation between an <input port, input label> tuple and an <output port, output label> tuple. Similarly, OXC provisions optical channel trail by establishing a relation between an <input port, input optical channel> tuple and an <output port, output optical channel> tuple. The functions of the control plane of both LSRs and OXCs include resource discovery, distributed routing control, and connection management. 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 OTN, and to establish and maintain optical channel trails under various optical internetworking traffic engineering rules and policies [Awuduche99].

2 Distinctions Between OXCs And LSRs

Current generation of OXCs and LSRs differ in certain characteristics. While LSRs are datagram devices which can perform certain packet level operations in the data plane, OXCs cannot. It cannot perform packet level processing in the data plane. Conceptually another difference is there, which is that the forwarding information is carried explicitly in LSRs as part of the labels appended to the data packets, while in the OXCs switching information is not appended to the data packet, rather it is implied from the wavelength or the optical channel.

3 Isomorphic Relations Between Explicit LSPs And Optical Channel ails

Both the explicit LSPs and optical channel trails (OCTs) exhibit certain commonalties. For example, both of them are the abstractions of unidirectional, point-to-point virtual path connections. An explicit LSP provides a parameterized packet forwarding path (traffic-trunk) between an ingress LSR and an egress LSR, while an optical channel trail provides a optical channel between two endpoints for the transport of digital client signals [Awuduche99]. Another commonality is that the payload carried by both LSPs and optical trails are transparent along their respective paths. They can be parameterized to stipulate their performance, behavioral, and survivability requirements from the network. Paths that satisfy some demands and policy requirements subject to some constraints imposed by the operational environment can be selected using constraint-based routing scheme. There are certain similarities in the allocation of labels to LSPs and in the allocation of wavelengths to optical trails.

4 Distinction between LSPs and Optical Channel Trails

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

5 General Requirements for the OXC Control Plane

This section has some of the requirements for the OXC control plane emphasizing on the routing components of these requirements. Some of the key aspects to these requirements are:(a) to expedite the capability to establish optical channel trails.(b) to support traffic engineering functions.(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 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. The next section gives a brief overview of MPLS traffic engineering.

5.1 Overview Of The MPLS Traffic Engineering Control

In this section, we discuss the components of the MPLS traffic engineering control plane model, which include the following modules [Awuduche99]:

(a) Resource discovery.

(b) State information dissemination to distribute relevant information concerning the state of the network. The state of the network includes topology and resource availability information. This can be accomplished by extending conventional interior gateway protocols (IGPs) to carry additional information in their link state advertisements.

(c) Path selection that is used to select an appropriate route through the MPLS network for explicit routing. It is implemented by introducing the concept of constraint-based routing which is used to compute paths that satisfy certain constraints, including constraints imposed by the operational environment.

(d) Path management, which includes label distribution, path placement, path maintenance, and path revocation. These functions are implemented through a signaling protocol, such as the RSVP extensions or through CR-LDP.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.

5.2 Required Enhancements To OXCs To Support MPLS Control Plane

This section discusses some of the enhancements to OXCs to support MPL(ambda)S.(a) 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.(b) An OXC must 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 optical channel trails within each fiber.(c) 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.

5.3 Required Enhancements To Current MPLS Control Plane

This section discusses the enhancements that are to be made in the MPLS control plane to support MPL(ambda)S [Basak99].

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 form 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 be 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.

There must be procedures to associate control channels to bearer channels between a pair of adjacent OXCs, if there exists multiple control channels and bearer channels between them. Procedures are required to demultiplex 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 and CR-LDP) need to be extended with objects that can provide sufficient details to establish reconfiguration parameters for OXC switch elements. IGP should be extended to carry information about the physical diversity of the fibers. IGP should be able to distribute information regarding the allocatable bandwidth granularity of any particular link.

6 Synthesizing The 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 [Awuduche99]. 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 an MPLS traffic engineering control plane would resemble a Label Switching Router, there are some important distinctions and limitations. 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 optical channel trails, each of which will have coarse discrete bandwidth granularities.

Routing in Optical Networks

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.

3.1 Models for IP-Optical Network Interaction

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 which model is whether there is a single/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 1) What is the interface defined between the two protocol instances? 2) What kind of information can be leaked from one protocol instance to the other? 3) Would one label switching protocol run on both domains? If that were to be the case then how would labels map to wavelengths? Also, how would IP QoS parameters be mapped into the optical domain?

..1 Overlay model

Under the overlay model, IP is more or less independent of the optical subnetwork. That is IP 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 divided into 2 parts:

a) Static overlay model: In this model path endpoints are specified through a network management system (NMS) though the paths may be laid out statically by the NMS or dynamically by the network elements. This would be similar to ATM permanent virtual circuits (PVCs) and ATM Soft PVCs (SPVCs).

b) Signaled overlay model: The path end-points are specified through signaling via a User to Network Interface (UNI). Paths must be laid out dynamically since they are specified by signaling. This is similar to ATM switched virtual circuits (SVCs). The Optical Domain Services Interoperability (ODSI) forum and Optical Internetworking Forum (OIF) are also defining similar standards for the Optical UNI. In these models user devices which reside on the edge of the optical network can signal and request bandwidth dynamically. These models use IP/optical layering. Endpoints are specified using a port number/IP address tuple. PPP is used for service discovery wherein a user device can discover whether it can use ODSI or OIF protocols to connect to an optical port. Unlike MPLS there are no labels to be setup. The resulting bandwidth connection will look like a leased line.

..2 The Integrated/ Augmented model

In the integrated model, the MPLS/IP layers act as peers of the optical transport network, such that a single routing protocol instance runs over both the IP/MPLS and optical domains. A common IGP like OSPF or IS-IS, with appropriate extensions will be used to distribute topology information. Also this model assumes a common address space for the optical and IP domain. 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 The peer model

The peer model is somewhat similar to the integrated model in that the IP reachability information might be passed around within the optical routing protocol but the actual flow will be terminated at the edge of the optical network and will only be reestablished upon reaching a non-peer capable node at the edge of the optical domain or at the edge of the domain, which implements both the peer and the overlay models.

2 The Lightpath routing solution

2.1 What is an IGP?

An IGP is a interior gateway routing protocol. Examples of IGPs would be OSPF and IS-IS. IGPs are used to exchange state information within a specified administrative domain and for topology discovery. This exchange of information inside the domain is done by advertising the Link state information periodically. Please refer to [OSPF] and [IS-IS] for more details.

2.2 How does MPLS fit into the picture?

While the idea of bandwidth-on-demand is certainly not new, existing networks do not support instantaneous service provisioning. Current provisioning of bandwidth is painstakingly static. Activation of large pipes of bandwidth takes anything from weeks to months. The imminent introduction of photonic switches in the transport networks opens new perspectives. Combining the bandwidth provisioning capabilities of photonic switches with the traffic engineering capabilities of MPLS, will allow routers and ATM switches to request bandwidth where and when they need it.

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

```
   +---------+     |     |TE       |     |    |IS-IS    |
   |         |     |     |Database |     |    |         |
   |  RSVP   |<--+       |         |     |    |TE EXT   |
   |         |           |         |<--+---   |OPT EXT  |
   +---------+           +---------+          +---------+

             Fig.2: Lightpath Selection
```

2.3 Lightpath Selection

The lightpath routing system is based on the MPLS Constraint based routing model. Fig.2 illustrates lightpath selection. These systems use CR-LDP or RSVP 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 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. To exchange information using IGP protocols like OSPF and IS-IS, certain extensions need to be made to both of these to support MPL(ambda) switching.

3 IS-IS/ OSPF enhancements

OSPF defined in [OSPF] and IS-IS defined in [IS-IS] are the commonly deployed routing protocols in large networks. OSPF/IS-IS have been extended to include traffic engineering capability [Katz99], [ISISTE]. There is a need to add the optical link state advertisement (LSA) to OSPF/IS-IS to support lightpath routing computation. The optical LSA would include a number of new elements, called TLVs (type-length-value) because of the way they are coded. The following sections describe some of the proposed TLVs.

3.1 Link Type

A network may have a link with many different characteristics. A link type TLV allows to identify a particular type of link. One way to describe the links would be [WANG00]:
a) Service transparent: Service transparent is a point-to-point
   physical link.
b) Service aware: A service aware link is a point-to-point logical
optical link.

Another way of classifying the links is based on the types of end nodes [Kompella00-a]. Nodes that can switch individual packets are called packet switch capable (PSC). Nodes that can transmit/receive SONET payloads are called time division multiplex (TDM) capable. Nodes that can switch individual wavelengths are called lambda switch capable (LSC). Finally, nodes that switch entire contents of one fiber into other are called fiber switch capable (FSC).

Links can be either physical (one hop) links or logical links consisting of multiple hop connections. Logical links are called "Forwarding Adjacencies (FAs)." This leads to the following types of links:

a) PSC links end (terminate or egress) on PSC nodes. Depending upon the hierarchy of LSPs tunneled within LSPs, several different types of PSC links can be defined.

b) TDM links end on TDM nodes and carry SONET/SDH payloads.

c) LSC links end on LSC nodes and consist of wavelengths.

d) FSC links end on FSC nodes and consist of fibers.

e) Forwarding Adjacency PSC (FA-PSC) links are forwarding adjacencies whose egress nodes are packet switching.

f) FA-TDM, FA-LSC, and FA-LSP are forwarding adjacencies whose egress nodes are TDM, LSC, and LSP capable, respectively.

3.2 Link Media Type (LMT)/ Link resource

A link may support a set of media types depending on resource availability and capacity of link. Such TLVs may have two fields of which the first one defines the media type, and the second field defines the lowest priority at which the media is available [Kompella00-a]. Link Media Types present a new constraint for LSP path computation. Specifically when a LSP is setup and it includes one or more subsequences of links which carry the LMT TLV then for all the links within each subsequence the encoding has to be the same and the bandwidth has to be at least the LSP's specified bandwidth. The total classified bandwidth available over one link can be classified using a resource component TLV [WANG00]. This TLV represents a group of lambdas with the same line encoding rate and total current available bandwidth over these lambdas. This TLV describes all lambdas that can be used on this link in this direction grouped by encoding protocol. There is one resource component per encoding type per fiber. If multiple fibers are used per link there will be a resource component per fiber to support fiber bundling.

3.3 Link ID

An identifier which identifies the optical link exactly as the point to point case for TE extensions.

3.4 Local Interface IP address

The interface address may be omitted in which case it defaults to the router address of the local node.

3.5 Remote Interface IP address

This address may be specified as an IP address on the remote node or the router address of the remote node .

3.6 Traffic Engineering (TE) Metric

This metric value can be assigned for path selection.

3.7 Path sub-TLV

When an LSP advertises a forwarding adjacency into an IGP, it may be desirable to carry the information about the path taken by this adjacency. This information may be used for path calculation by other LSRs.

3.8 Shared Risk Link group TLV

A set of links may constitute a 'shared risk link group' (SRLG) if they share a resource whose failure may affect all links in the set. An example would be two fibers in the same conduit. Also, a fiber may be part of more than one SRLG.

4 Control Channels, Data Channels, and IP Links

A pair of OXCs are said to be neighbors from the MPLS point of view if they are connected by one or more logical or physical channels. If several fibers share the same TE characteristic then a single control channel would suffice for all of them. From the IGP point of view this control channel along with all its fibers form a single IP link. Sometimes fibers may need to be divided into sets that share the same TE characteristic. Corresponding to each such set, there must be a logical control channel to form an IP link. All of the multiple logical control channels could be realized via one common control channel. When an adjacency is established over a logical control channel that is part of an IP link formed by the channel and a set of fibers, this link is announced into IS- IS/OSPF as a "normal" link; the fiber characteristics are represented as TE parameters of that link. If there are more than one fibers in the set, the set is announced using bundling techniques discussed in [Kompella00-b].

4.1 Excluding data traffic from control channels

The control channels between OXCs or between an OXC and a router are generally meant for low bandwidth control traffic. These control channels are advertised as normal IP links. However if regular traffic is forwarded on these links the channel capacity will soon be exhausted. To avoid this, if we assume that data traffic is sent over BGP destinations and control traffic is sent to IGP destinations. Ways to do this are discussed in [KOMPELLA00-a].

4.2 Forwarding adjacencies

An LSR at the head of an LSP may advertise this LSP as a link into a link state IGP. When this LSP is advertised into the same instance of the IGP as the one that determines the route taken this adjacency then we call such a link a "forwarding adjacency". Such an LSP is referred to as a "forwarding adjacency LSP" or just FA-LSP. Forwarding adjacencies may be statically provisioned or created dynamically. Forwarding adjacencies are by definition unidirectional.

When a forwarding adjacency is statically provisioned, the parameters that can be configured are: the head-end address, the tail-end address, bandwidth, and resource color constraints. The path taken by the FA-LSP can be computed by the Constrained Shortest Path Formulation (CSPF) mechanism or MPLS TE or by explicit configuration. When a forwarding adjacency is created dynamically its parameters are inherited by the LSP which induced its creation. Note that the bandwidth of the FA-LSP must be at least as big as the LSP that induced it.

When an FA-LSP is advertised into IS-IS/OSPF, the link type associated with this LSP is the link type of the last link in the FA-LSP. Some of the attributes of this link can be derived from the FA-LSP but others need to be configured. Configuration of the attributes of statically provisioned FAs is straightforward, but for dynamically provisioned FAs a policy-based mechanism may be needed.

The link media type of the FA is the most restrictive of the link media types of the component links of the forwarding adjacency. FAs will not be used to establish peering relationships between routers at the end of the adjacencies but will only be used for CSPF computation.

4.3 Two way connectivity

CSPF shouldn't perform any two way connectivity check on links used by CSPF. This is because some of the links are unidirectional and may be associated with forwarding adjacencies.

4.4 Optical LSAs

There needs to be a way of controlling the protocol overhead introduced by optical LSAs. One way to do this is to make sure that a Link State Advertisement happens only when there is a significant change in the value of metrics since the last advertisement. A definition of significant change is when the difference between the currently available bandwidth and last advertised bandwidth crosses a threshold[WANG00]. The frequency of these updates can be decreased dramatically using event driven feedback.

5 Some Open Questions

Some issues that have not been resolved so far are: How to ensure that end-to-end information is propagated across as an optical network? How to accommodate proprietary optimizations within optical sub-networks for provisioning and restoration of lightpaths? Whether dynamic and precomputed information can be used and if so what is the interaction between them? What QOS related parameters need to be defined? How to ensure fault tolerant operation at protocol level when hardware does not support fault tolerance? How to address scalability issues? What additional modifications are required to support a network for routing control traffic?

Signaling & Control

Signaling means to intimate any particular element of certain characteristics or services. 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 Fig.3

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

Fig.3: OXC Architecture

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 as proposed in, or it may function purely as a controller for the optical network and carry no IP data traffic. Although the IP router performs all management and control functions, lightpaths may carry arbitrary types of traffic.

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

Fig.4: Control Plane Architecture

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. On each link within the network, one channel is assigned as the default routed (one hop) lightpath. The routed lightpath provides router to router connectivity over this link. These routed lightpaths reflect (and are thus identical to) the physical topology. The assignment of this default lightpath is by convention, e.g. the 'first' channel. All traffic using this lightpath is IP traffic and is forwarded by the router. All control messages are sent in-band on a routed lightpath as regular IP datagrams, potentially mixed with other data but with the highest forwarding priority. It is assumed multiple channels on each link, a fraction of which is reserved at any given time for restoration. The default-routed lightpath is restored on one of these channels. Therefore, we can assume that as long as the link is functional, there is a default routed lightpath on that link.

The IP router 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. Fig.4 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.

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.

4.2 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. This proposal uses the concept of softness of lightpaths. This implies that the lightpath expires unless refreshed periodically bu the source. The first-hop router must periodically resend the lightpath setup request. If the state of a lightpath expires at a particular node, the state is locally removed and all resources allocated to the lightpath are reclaimed.

3.1 Basic Path Setup Procedure

Techniques for link provisioning depend upon whether the OXCs do or do not have wavelength conversion. Both these cases are discussed below.

3.1.1 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) A sample format for the setup request has been defined in [Chaudhuri00]. 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.

3.1.2 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-linearities, chromatic dispersion, amplifier spontaneous emission and other factors together may limit the scalability of an all-optical network. Routing in such networks will 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 uni-directional 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.

8.2 CR-LDP Extensions for Path Setup

Label Distribution Protocol (LDP) is defined for distribution of labels inside one MPLS domain. CR-LDP is the contraint-based extesion 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.

Automated establishment of lightpaths involves setting up the crossconnect table entries in the appropriate OLXCs in a coordinated manner such that the desired physical path is realized. The request to establish an lightpath may arise either from a router (or some other device) connected to the OXCs or from a management system. Such a request must identify the ingress and the egress OXC as endpoints of the lightpath. In addition, it may also optionally specify the input and output ports, wavelengths, and TDM channels.

The request may also 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 CR-LDP to set up the path.

In optical networks, label mapping corresponds to the assignment of input or output ports for paths by optical switches and the communication of this information to the appropriate neighbors.

A Label Request message is used by an upstream LSR to request a label binding from the downstream LSR for a specified 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: This field specifies ports to be assigned for setting up the path. Such a "label" (wavelength) must be assigned in a coordinated manner by a pair of adjacent OXCs, since the "label" at one OXC is tied to a specific "label" at a neighboring OXC based on physical connectivity.

* Signaling Optical Switched Path Identifier: This field identifies the lightpath being established. This provides the flexibility of establishing LSPs on the top of an lightpath already setup.

* Signaling the two end points of the path being set up: These fields indicate the two end-points at the port level of the lightpath. The port selected for the egress node is propagated to the egress node.

* Signaling requirements for both span and path protection: This field signals the protection levels required for both span (or local) and path protection. Examples of span (or local) protection include SONET 1+1 and 1:N APS. Examples of path protection include various levels regarding how an alternate path is shared such as in a style of 1+1 or 1:N analogous to span protection.

* Recording the precise route of the path being established: This is done by letting each OXC insert its node ID and the both output and input port selected for the path in the Label Mapping message. The message received by the ingress OXC will have the complete route at the port level. This information is useful for network management functions.

3.3 Extensions to RSVP for Path Setup

Resource reSerVation Protocol (RSVP) 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 receipients. The receivers interested in communicating with the source send the "Resv" messages.

The following extensions to RSVP have been proposed to support path setup [Jonathan00]:

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

These extensions are described below.

3.3.1 Reduction Of Trail Establishment Latency

Currently due to receiver-oriented nature of RSVP, the internal configuration of an OXC in the downstream direction cannot be initiated until it receives the Resv message from the downstream node. The ability to begin configuring an OXC before receiving a Label Object in the Resv message can provide a significant reduction in the setup latency, especially in OXCs with non-negligible configuration time. To accomplish this, a new approach has been proposed in which an upstream OXC suggest a (fiber, lambda) label for the downstream node to use by including the suggested Label object in the Label Request object of the Path message. The Label object will contain the downstream node's Label for the bearer channel, which can be obtained through the Link Management Protocol (LMP). This will allow the upstream OXC to begin its internal configuration before receiving the Resv message from the downstream node.

3.3.2 Establishment Of Bi-directional Lightpaths

In the new approach that is proposed, a Label Object is added to the Path message in the downstream direction. In this way, the upstream direction of the bi-directional path is established on the first pass from the source to destination, reducing the latency of the reservation process. For bi-directional trails, if a label suggestion is also used, there will be two Labels in the Path message: the upstream Label in the Label object and the suggested Label in the Label_Request object.

3.3.3 Failure Notification

A new RSVP message, called the Notify message, can be used to notify RSVP nodes when failures occur. The Notify message will be transmitted with the router alert option turned off so that intermediate nodes will not process or modify the message, but only perform standard IP forwarding of the message.

3.3.4 Bundling of Notifications

Another extension to RSVP has been recently proposed to allow the use of bundle messages in order to reduce the overall message-handling load. An RSVP bundle message consists of a bundle header followed by a body consisting of a variable number of standard RSVP messages. Support for the bundle message is optional, and currently, bundle messages can only be sent to adjacent RSVP nodes. In order to effectively restore a network to a stable state, nodes that are running restoration algorithms should consider as many failed trails as possible before making restoration decisions. To improve performance and ensure that the nodes are provided with as many of the affected paths as possible, it is useful to include the entire set of Notify messages in a single bundle message and send it to the responsible RSVP node directly, without message processing by the intermediate RSVP nodes. This can be accomplished by addressing the bundle message to the source RSVP nodes and turning off the router alert option in the IP header. Intermediate RSVP nodes then should perform standard IP forwarding of this message.

4 Resource Discovery and Maintenance

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.

For optical networks, the following information need to be stored at each node and propagated throughout the network as OSPF link-state information:

- Representation of the current network topology and the link states (which reflect the wavelength availability). This can be achieved by associating with the link state,
    - total number of active channels
    - number of allocated non-preemptable channels
    - number of allocated preemptable channels
    - number of reserved protection channels

- Optional physical layer parameters for each link. These parameters are not expected to be required in a network with 3R signal regeneration, but may be used in all-optical networks.

All of the above information is obtained via OSPF updates, and is propagated throughout the network. In networks with OXCs without wavelength converters, decisions at the first-hop router are made without knowledge of wavelength availability. This is done to reduce the state information that needs to be propagated within the network.

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) is 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. There are six classes of GSMP request-response message:

- Connection Management
- Reservation Management
- Port Management
- State and Statistics
- Configuration, and
- Quality of Service

The switch may also generate asynchronous Event messages to inform the controller of asynchronous events.

5 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. NHRP functions are performed by two types of logical entities:

Next Hop Server (NHS) - implemented in routers
Next Hop Client (NHC) - implemented in routers or NBMA-attached hosts.

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 are 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 must 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).

ce a control channel is configured between two OXCs, a Hello protocol n 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 ilures 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 must 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 uni-directional, 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 must 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 must be used to rapidly locate the failure so that appropriate protection/restoration mechanisms can be initiated. This is discussed further in Section 6.10.

2 Optical Performance Monitoring (OPM)

Current-generation DWDM 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 DWDM 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 must 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 [Ceuppens00]. Optical cross talk, dispersion, 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 uses 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.

. 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 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 an 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 bidirectional line switched ring architecture, a bidirectional 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.

..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 micro-seconds restoration times in WDM Shared protection Rings. Multiwavelength systems add extra complexity to the restoration problem. Under these circumstances, a 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.

Optical protection techniques for mesh architectures have also been proposed. They operate on a line rather than path protection basis. The fundamental unit being protected is a transmission link rather than an end to end connection. The methodology is a generalization of that is used in SPRings. The system requires 100% redundancy. Fault recovery decisions are made locally in a distributed fashion and independent of the state of activity of the network. The implementation uses simple and reliable protection switches in each network node so that protection is accomplished without significant processing, transmission, and propagation delays.

Thus, the main advantage of lower layer mechanisms is the fast restoration. 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 layers.

2 MPLS in Protection

Although the current routing algorithms are very robust and survivable, the amount of time they take to recover from a failure can be significant, on the order of several seconds or minutes, causing serious disruption of service in the interim. This is unacceptable to many organizations that aim to provide a highly reliable service, and thus require recovery times on the order of tens of milliseconds.

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.

2.1 Motivations

The need for MPLS layer protection and for open standards in protection arises because of the following:

1. Layer 3 or IP rerouting may be too slow for a core MPLS network that needs to support high reliability/availability.

2. Layer 0 (for example, optical layer) or Layer 1 (for example, SONET) mechanisms may be limited to ring topologies and may not include mesh protection.

3. Layer 0 or Layer 1 mechanisms may have no visibility into higher layer operations. Thus, while they may provide link protection for example, they cannot easily provide MPLS path protection.

4. Establishing interoperability of protection mechanisms between multi-vendor LSRs in core MPLS networks is urgently required to enable the adoption of MPLS as a viable core transport technology.

2.2 Goals

Based on our motivations, we can state some goals for MPLS based protection:

1. MPLS-based recovery mechanisms should facilitate fast (10Æs of ms) recovery times.

2. MPLS-based recovery techniques should be applicable for protection of traffic at various granularities. For example, it should be possible to specify MPLS- based recovery for a portion of the traffic on an individual path, for all traffic on an individual path, or for all traffic on a group of paths.

3. MPLS-based recovery techniques may be applicable for an entire end-to-end path or for segments of an end-to-end path.

4. MPLS-based recovery mechanisms should be able to take into consideration the recovery actions of other layers.

5. MPLS-based recovery actions should avoid network-layering violations. That is, defects in MPLS-based mechanisms should not trigger lower layer protection switching.

6. MPLS-based recovery mechanisms should minimize the loss of data and packet reordering during recovery operations.

7. MPLS-based recovery mechanisms should minimize the state overhead incurred for each recovery path maintained.

8. MPLS-based recovery mechanisms should be able to preserve the constraints on traffic after switchover, if desired. That is, if desired, the recovery path should meet the resource requirements of, and achieve the same performance characteristics, as the working path.

3 Protection options

3.1 Dynamic Protection

These protection mechanisms dynamically create protection entities 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.

3.2 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.

3.3 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.

3.4 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 path splitting LSR (PSL).

3.5 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 preconfigured, 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.

3.6 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.

3.7 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.

3.8 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.

3.9 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. In MPLS, this bandwidth may be considered wasted. Alternately, this bandwidth could be 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 path merge LSR (PML).

3.10 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.

3.11 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.

3.11.1 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.

3.11.2 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.

4. Failure detection

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 Loss of Function (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.

5 Failure Notification

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 fault 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 fault restoration 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).

5.1 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.

6. Timing

There are a number of timing parameters that need to be specified for proper restoration of the IP/WDM networks. Some of these are described below.

6.1 Protection Switching Interval Timer T1:

Controls the maximum duration within which a protection switch must be accomplished, following the detection of a failure.

6.2 Inter-FIS Packet Timer T2:

Interval at which successive FIS packets are transmitted by a LSR to its upstream neighbor.

6.3 Maximum FIS duration timer T3:

Max. time for which FIS packets are transmitted by an LSR to is upstream peer.

6.4 Protection switching dampening timer T4:

Time interval between receipt of a protection switch trigger and the initiation of the protection switch. The purpose of this timer is to minimize misordering of packets at a PML following a protection (restoration) switch from the working (backup) to the backup (working) path. This is because packets buffered on the working (backup) path may continue to arrive at the PML even as working traffic begins to arrive on the protection (working) path. Therefore, forcing the PSL to hold off the protection (or restoration) switching action, gives the buffers on the working (protection) path time to clear before data on the protection (working) path begins to arrive.

6.5 Liveness Message Send interval T5:

Interval at which successive Liveness messages are sent by an LSR to peer LSRs that have a working path (and RNT) through this LSR.

6.6 Failure Indication Hold-off Timer T6:

Interval between the detection of a failure at an LSR, and the generation of the first FIS message, to allow time for lower layer protection to take effect.

6.7 Lost Liveness Message Threshold K:

No. of Liveness messages that can be lost before an LSR will declare LF and generate the FIS.

For proper operation, it is required that T1 >> T2 > T3 and T1> T4

7. Signaling Requirements related to restoration

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 must 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 must be quick and reliable. It is especially important to minimize signaling delays during restoration.

4. Lightpaths are typically bidirectional. 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 must be well-defined to avoid false restoration attempts during path set-up or tear down.

8 RSVP/CR-LDP Support for Restoration

Special requirements for protecting and restoring lightpaths and the extensions to RSVP 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 or1: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.

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.

9 Fast restoration of MPLS LSPs

Fast recovery in MPLS is hampered by the fact that detecting an LSP failure at the ingress LSR can take a long time. After a break in an LSP hop, Notification messages are propagated along the LSP intermediate nodes back to the ingress LSR.

The fastest detection occurs at the local end of a link failure. Schemes that try to mend connections at the point of failure are known as "local repair" schemes.

A problem with single L2 link failure is that multiple LSPs can be affected and many (hundreds) ingress points must be informed. Just as a single L2 failure can affect multiple LSPs, a single L1 failure can affect multiple L2 links.

As noted earlier, L1 failure detection is fast due to physical methods (loss of light, loss of carrier signal). This is an attractive property. Further, in a TDM, optical mux (SONET), or optical cross connect network, when a link fails all of the paths (at that layer) which use the link go down.

Unlike higher layers, the endpoints of those paths detect the failure quickly because the signaling of the failure is very fast (e.g., AIS signals in SONET) and because the signaling is sent to each channel of the failed link. So in L1 networks, the detection of a failed connection is fast and scales well for all connections on the failed link.

A key to the solution for fast detection is the alignment of L1, L2, and L3 capabilities into a single node. This architecture and its impacts on the ability to detect LSP failure are now described.

).1 L1/L2/L3 Integration

As was noted earlier, in MPLS LSRs, the alignment of the L3 and L2 topology brings some advantages in the speed at which the network can react to a link failure. This integration is extended to encompass L1 components in order to realize further speed advantages.

An L1/L2/L3 switch is defined as an LSR combined with an L1 cross connect switch. This could be a SONET Add/Drop Mux, an optical cross connect, or traditional TDM switch. The integrated switch is able to originate and terminate IP traffic from the L1 cross connects. Conceptually, this is done over dedicated L1 channels between the L1 cross connect and the pure IP router function of the integrated switch.

Two L1/L2/L3 nodes are connected by a physical L1 link. A channel in that link is used as a router-router IP link. For example, an OC-3 channel of an OC-48 link with PPP over SONET for the framing. This is analogous to the L2 control channel between two MPLS switches connected over an ATM interface.

A key difference between this type of network and L2/L3 networks, which are overlayed on L1 networks, is that the L1/L2/L3 network does not have any L1 paths, which act as router-router links. In an integrated network, the L3 routing protocol has a view of both the L2 and L1 topology since those layers are now aligned.

Here, in an L1/L2/L3 network, an L1 path has an LSR at every cross connect point. To use an L1 path, treat it as if were an LSP, or overlay an LSP onto this path. That is, consider the L1 path as a cut-through. When an incoming IP packet is matched to a Forwarding Equivalence Class associated with this L1 cut- through, the IP forwarding table entry points to the start of this L1 path. As with L2 cut-throughs, an L2 header is added. The packet is sent to this path and is then L1 switched until it reaches the end of the path. At the termination point, the packet could be L2 switched or L3 forwarded.

).2 An example

Using L1 cut-throughs in an L1/L2/L3 network enables fast detection of LSP failure. Consider two LSPs that are L1 cut-throughs:
   LSR1-LSR2-LSR3-LSR4 and
   LSR5-LSR2-LSR3-LSR6

If L1 link LSR2-LSR3 goes down, all nodes in both LSRs can detect the path failure based on L1 physical methods. For example, loss of light (Alarm Indication Signal in SONET) or carrier signal (TDM). In particular, the LSP endpoints can determine that the LSP is down much faster than the protocol based method in LDP of Notification messages which is processed at each LSR on the paths back to the ingress and egress. For example, propagation of the physical failure is about 5 microseconds per kilometer.

Not only is the failure detection fast, but it scales for all LSPs that are affected by a single L1 failure. In the example above, two LSPs are notified, but if there were 192 paths in an OC192 link, then all of their endpoints could detect the link failure within a short period of time (a few milliseconds).

When an LSP failure is detected, the LSR can reroute the traffic to a backup LSP. This backup LSP could be pre-defined to be link disjoint from the primary LSP, and could also be set up in advance. To avoid wasting dedicated bandwidth (i.e., a dedicated backup L1 cut-through), the backup LSP for the L1 cut-through could be an LSP created over L2 connections which share bandwidth (e.g., ATM UBR VC).

Assuming that a backup LSP is already set up, restoration of a failed LSP that is overlayed on an L1 cut-through could be implemented with similar performance to SONET Line and Ring restoration.

For LSR which provide L3 connectionless forwarding, traffic from the failed LSP could also be immediately handled by L3 forwarding if a backup path LSP is not provided.

.0 LMP's Fault localization mechanism

If bearer channels fail between two OXCs, the power monitoring system in all of the downstream nodes will detect loss of light (LOL) and indicate a failure. As part of the fault localization, a monitoring window can be used in each node to determine if a single bearer channel has failed or if multiple bearer channels have failed.

As part of the fault localization, a downstream node that detects bearer channel failures across a link will send a Channel_Fail message to its upstream neighbor (bundling together the notification of all of the failed bearer channels) and the node will put the ports associated with the failed bearer channels into the standby state. An upstream node that receives the Channel_Fail message will correlate the failure to see if there is a failure on the corresponding input and output ports for the optical trail(s). If there is also a failure on the input channel(s) of the upstream node, the node will return a Channel_Fail_Ack message to the downstream node (bundling together the notification of all the channels), indicating that it too has detected a failure. If, however, the fault is CLEAR in the upstream node (i.e., there is no LOL on the corresponding input channels), then the upstream node will have localized the failure and will return a Channel_Fail_Nack message to the downstream node, and initiate protection/restoration procedures. The protection channels may be preconfigured or they may be dynamically selected by the OXC on the transmit side.

In Fig. 5, a sample network is shown where four OXCs are connected in a linear array configuration. The control channels are bi-directional and are labeled with a "c". All optical trails are uni-directional going left to right. In the example there is a failure on a single bearer channel between OXC2 and OXC3. Both OXC3 and OXC4 will detect the failure and each node will send a Channel_Fail message to the corresponding upstream node (OXC3 will send a message to OXC2 and OXC4 will send a message to OXC3). When OXC3 receives the Channel_Fail message from OXC4, it will correlate the failure and return a Channel_Fail_Ack message back to OXC4. Upon receipt of the Channel_Fail_Ack message, OXC4 will move the associated ports into a standby state. When OXC2 receives the Channel_Fail message from OXC3, it will correlate the failure, verify that it is CLEAR, localize the failure to the bearer channel between OXC2 and OXC3, and send a Channel_Fail_Nack message back to OXC3.

```
    +-------+        +-------+        +-------+        +-------+
    + OXC 1 +        + OXC 2 +        + OXC 3 +        + OXC 4 +
    +       +-- c ---+       +-- c ---+       +-- c ---+       +
----+-------+--------+---\   +        +       +        +       +
    +       +        +    \--+---##---+--\    +        +       +
    +       +        +       +   (A)  +   \   +        +       +
```

```
----+--\    +        +       +        +    \--+--------+-------+--->
    +   \   +        +    /--+--------+---\   +        +       +
    +    \--+--------+---/   +        +    \--+--------+-------+--->
    +       +        +       +        +       +        +       +
    +-------+        +-------+        +-------+        +-------+
```

Fig.5 : Example - We show one type of bearer channel failures (indicated by ## in the figure): a single bearer channel fails between two OXCs

Security Considerations

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

Acronyms

CR-LPD - Constraint-Based Routing Setup using LDP
CSPF - Constraint Shortest Path First
FIS - Failure Indication Signal
FRS - Failure Recovery Signal
GSMP - General Switch Management Protocol
LDP - Label Distribution Protocol
LF   - Link Failure
LMP - Link Management Protocol
LOP - Loss of Packets
LOS - Loss Of Signal
LP   - Lightpath
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
OPM - Opical Performance Monitoring
OSPF - Open Shortest Path First
OTN - Optical Transport Network
OXC - Optical Crossconnect
PML - Protection Merge LSR
PMTG - Protected MPLS Traffic Group
PMTP - Protected MPLS Traffic Portion
PSL - Protection Switch LSR
PXC - Photonic Crossconnect
RNT - Reverse Notification Tree
RSVP - Resource reSerVation Protocol
SPRing - Shared Protection ring
TLV - Type Length Value

Terminology

Channel:
A channel is a uni-directional 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 an uni-directional 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.

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 bidirectional 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 detectedat 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 trafficgroup 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 uni-directional Optical Channel Trail. A 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 theexact reverse path of the multipoint-to-point tree formed by merging ofworking 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 backfrom 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 must 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 an uni-directional 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

[Awduche99] 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-01.txt, Work in Progress, November 1999.

[Basak99] Debashis Basak, D. Awduche, J. Drake, Y. Rekhter, "Multi-Protocol Lambda Switching: Issues in Combining MPLS Traffic Engineering Control With Optical Crossconnects", Internet Draft draft-basak-mpls-oxc-issues-01.txt, Work in Progress, February 2000.

[Ceuppens00] L. Ceuppens, et. al. "Performance Monitoring in Photonic Networks in support of MPL(ambda)S," Internet Draft draft-ceuppens-mpls-optical-00.txt, March 2000.

[Chaudhuri00] S. Chaudhuri, et. al. "Control of Lightpaths in an Optical Network," Internet Draft draft-chaudhuri-ip-olxc-control-00.txt, February 2000.

[CRLDP] B. Jamoussi, et. al. "Constraint-Based LSP Setup using LDP," Internet Draft draft-ietf-mpls-cr-ldp-03.txt, February 1999.

[GSMP] A. Doria, et. al. "General Switch Management Protocol V3,"Internet Draft draft-ietf-gsmp-05.txt, April 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-01.txt, work in progress, May 1999

[Jonathan00] J.P. Lang, K. Mitra, J. Drake,, "Extensions to RSVP for Optical Networking", Internet Draft draft-lang-mpls-rsvp-oxc-00.txt, Work in Progress, March 2000.

[Jamoussi99] L. Andersson, A. Fredette, B. Jamoussi, et al., "Constraint- Based LSP Setup using LDP", Internet Draft draft-tang-crldp-optical-00.txt, Work in Progress, January 1999.

[Katz99] Katz, D. and Yeung, D., "Traffic Engineering Extensions to OSPF," Internet Draft dtaft-katz-yeung-traffic-01.txt, Work in progress, October 1999.

[KOMPELLA00-a] K. Kompella et. al., "Extensions to IS-IS/OSPF and RSVP in support of MPL(ambda)S", Internet Draft draft-kompella-mpls-optical-00.txt, Work in Progress, February 2000 .

[Kompella00-b] Kompella, K., Rekhter, Y., "Link Bundling in MPLS Traffic Engineering", draft-kompella-mpls-bundle-00.txt, February 2000.

[Lang00] J.P. Lang, "Link Management Protocol (LMP)," Internet Draft draft-lang- mpls-lmp-00.txt, March 2000.

[Luciani00] J. Luciani, B. Rajagopalan, D. Awuduche, B. Cain, Bilel Jamoussi, "IP Over Optical Networks - A Framework", Internet Draft draft-ip-optical-framework-00.txt, Work in Progress, March 2000.

[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

[Tang00] Z.B. Tang et. al. "Extensions to CR-LDP for Path Establishment in Optical Networks," Internet Draft draft-tang-crldp-optical-00.txt, March 2000

[WANG] G.Wang et. al., "Extensions to OSPF/IS-IS for Optical Routing", Internet Draft draft-wang-ospf-isis-lambda-te-routing-00.txt, Work in Progress, March 2000.

Author's Addresses

N. Chandhok, A. Durresi, R. Jagannathan, S. Seetharaman, 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: {chandhok, durresi, rjaganna, seethara, 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