Network Working Group
Internet Draft
Expiration date: August 2001

Peter Ashwood-Smith (Nortel Networks)
Daniel Awduche (Movaz Networks)
Ayan Banerjee (Calient Networks)
Debashis Basak (Accelight Networks)
Lou Berger (Movaz Networks)
Greg Bernstein (Ciena Corporation)
John Drake (Calient Networks)
Yanhe Fan (Axiowave Networks)
Don Fedyk (Nortel Networks)
Gert Grammel (Alcatel)
Kireeti Kompella (Juniper Networks)
Alan Kullberg (NetPlane Systems)
Jonathan P. Lang (Calient Networks)
Fong Liaw (Zaffire, Inc.)
Dimitri Papadimitriou (Alcatel)
Dimitrios Pendarakis (Tellium, Inc.)
Bala Rajagopalan (Tellium, Inc.)
Yakov Rekhter (Juniper Networks)
Debanjan Saha (Tellium, Inc.)
Hal Sandick (Nortel Networks)
Vishal Sharma (Jasmine Networks)
George Swallow (Cisco Systems)
Z. Bo Tang (Tellium, Inc.)
John Yu (Zaffire, Inc.)
Alex Zinin (Cisco Systems)

Eric Mannie (Ebone) - Editor

February 2001

Generalized Multi-Protocol Label Switching (GMPLS) Architecture

draft-many-gmpls-architecture-00.txt

Status of this Memo

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

1. Abstract

Future data and transmission networks will consist of elements such as routers, switches, DWDM systems, Add-Drop Multiplexors (ADMs), photonic cross-connects (PXCs) or optical cross-connects (OXCs), etc that will use Generalized MPLS (GMPLS) to dynamically provision resources and to provide network survivability using protection and restoration techniques.

This document describes the architecture of GMPLS. GMPLS extends MPLS to encompass time-division (e.g. SDH/SONET, PDH, G.709), wavelength (lambdas) and spatial switching (e.g. incoming port or fiber to outgoing port or fiber). The main focus of GMPLS is on the control plane of these various layers since each of them can use totally different data or forwarding planes. The intention is to cover both the signaling and the routing part of that control plane.

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [2].

3. Introduction

The architecture presented in this document covers the main building blocks needed to build a consistent control plane for multiple switching layers. It does not restrict the way that these layers work together. Different models can be applied: e.g. overlay, augmented or integrated. Moreover, each pair of contiguous layer may work jointly in a different way. It results that a number of combinations are possible, at the discretion of manufacturers and operators.

This document generalizes the MPLS architecture [MPLS-ARCH], and in some cases can differ slightly from that architecture since non packet-based forwarding planes are now considered. It is not the intention of this document to describe concepts already described in the current MPLS architecture. The goal is to describe specific concepts of GMPLS.

However, some of the concepts described hereafter are not described

in the current MPLS architecture and are applicable to both MPLS and GMPLS, i.e. link bundling, unnumbered links and LSP hierarchy. Since they raised from the GMPLS needs and since they are of paramount importance for an operational GMPLS network, they will be introduced here.

The following sections will first introduce GMPLS. Then the specific GMPLS building blocks will be presented and we will explain how they

can be combined together. Details about these separate building blocks can be found in the corresponding documents.

3.1. Acronyms & abbreviations

ABR	Area Border Router
AS	Autonomous System
ASBR	Autonomous System Boundary Router
BGP	Border Gateway Protocol
CR-LDP	Constraint-based Routing LDP
CSPF	Constraint-based Shortest Path First
DWDM	Dense Wavelength Division Multiplexing
FA	Forwarding Adjacency
GMPLS	Generalized Multi-Protocol Label Switching
IGP	Interior Gateway Protocol
LDP	Label Distribution Protocol
LMP	Link Management Protocol
LSA	Link State Advertisement
LSR	Label Switching Router
LSP	Label Switched Path
MIB	Management Information Base
MPLS	Multi-Protocol Label Switching
RSVP	ReSource reserVation Protocol
SDH	Synchronous Digital Hierarchy
STM(-N)	Synchronous Transport Module (-N)
STS(-N)	Synchronous Transport Signal-Level N (SONET)
TE	Traffic Engineering

3.2. Multiple Types of Switching and Forwarding Hierarchies

Generalized MPLS differs from traditional MPLS in that it supports multiple types of switching, i.e. the addition of support for TDM, lambda, and fiber (port) switching. The support for the additional types of switching has driven generalized MPLS to extend certain base functions of traditional MPLS and, in some cases, to add functionality. These changes and additions impact basic LSP properties, how labels are requested and communicated, the

unidirectional nature of LSPs, how errors are propagated, and information provided for synchronizing the ingress and egress LSRs.

The MPLS architecture [MPLS-ARCH] was defined to support the forwarding of data based on a label. In this architecture, Label Switching Routers (LSRs) were assumed to have a forwarding plane that is capable of (a) recognizing either packet or cell boundaries, and (b) being able to process either packet headers (for LSRs capable of recognizing packet boundaries) or cell headers (for LSRs capable of recognizing cell boundaries).

This original architecture is here extended to include LSRs whose forwarding plane recognizes neither packet, nor cell boundaries, and therefore, can't forward data based on the information carried in either packet or cell headers. Specifically, such LSRs include devices where the forwarding decision is based on time slots, wavelengths, or physical ports. So, the new set of LSRs, or more

precisely interfaces on these LSRs, can be subdivided into the following classes:

1. Packet-Switch Capable (PSC) interfaces:

Interfaces that recognize packet/cell boundaries and can forward data based on the content of the packet/cell header. Examples include interfaces on routers that forward data based on the content of the "shim" header, interfaces on ATM-LSRs that forward data based on the ATM VPI/VCI.

2. Time-Division Multiplex Capable (TDM) interfaces:

Interfaces that forward data based on the data's time slot in a repeating cycle. An example of such an interface is an interface on a SDH/SONET Cross-Connect (XC), Terminal Multiplexer (TM) or Add-Drop Multiplexer (ADM). Other examples are an interface implementing G.709 (the digital wrapper), or a PDH interface.

3. Lambda Switch Capable (LSC) interfaces:

Interfaces that forward data based on the wavelength on which the data is received. An example of such an interface is an interface on an Optical Cross-Connect that can operate at the level of an individual wavelength. Another example is an interface that can operate at the level of a group of wavelengths, i.e. a waveband.

4. Fiber-Switch Capable (FSC) interfaces:

Interfaces that forward data based on a position of the data in the

real world physical spaces. An example of such an interface is an interface on a photonic Cross-Connect that can operate at the level of a single (or multiple) fibers.

A circuit can be established only between, or through, interfaces of the same type. Depending on the particular technology being used for each interface, different circuit names can be used, e.g. SDH circuit, optical trail, light path etc. In the context of GMPLS, all these circuits are referenced by a common name: Label Switched Path (LSP).

The concept of nested LSP (LSP within LSP) already available in the traditional MPLS allows here to build a forwarding hierarchy, i.e. a hierarchy of LSPs. This hierarchy of LSPs can occur on the same interface, or between different interfaces.

It can occur on the same interface if this interface is capable of multiplexing several LSPs from the same technology (layer), e.g. a lower order SDH/SONET LSP (VC-12) nested in a higher order SDH/SONET LSP (VC-4). Several levels of signal (LSP) nesting are defined in the SDH/SONET multiplexing hierarchy.

The nesting can also occur between interfaces. At the top of the hierarchy are FSC interfaces, followed by LSC interfaces, followed

by TDM interfaces, followed by PSC interfaces. This way, an LSP that starts and ends on a PSC interface can be nested (together with other LSPs) into an LSP that starts and ends on a TDM interface. This LSP, in turn, can be nested (together with other LSPs) into an LSP that starts and ends on an LSC interface, which in turn can be nested (together with other LSPs) into an LSP that starts and ends on a FSC interface.

3.3. Extension of the MPLS Control Plane

The establishment of LSPs that span only Packet Switch Capable (PSC) interfaces is defined for the original MPLS and/or MPLS-TE control planes. GMPLS extends these control planes to support each of the four classes of interfaces (i.e. layers) defined in the previous section.

Note that the GMPLS control plane supports as well an overlay model, an augmented model or an integrated model. The benefits of using an augmented or integrated model will have to be clarified and evaluated in the future. In the mean time, GMPLS is very suitable for controlling each layer completely independently. This elegant approach will facilitate the future deployment of other models.

The GMPLS control plane is made of several building blocks that will be described in more details in the following sections. These building blocks are indeed well-known IETF signaling and routing protocols that have been extended and/or modified. They use IPv4 and/or IPv6 addresses. Only one new specialized protocol was required to support the operations of GMPLS, a signaling protocol for link management [LMP].

GMPLS is indeed based on the Traffic Engineering (TE) extensions to MPLS, a.k.a. MPLS-TE. This is because most of the technologies that can be used below the PSC level require some traffic engineering. The placement of LSPs at these levels needs in general to take several constraints into consideration (such as bandwidth, protection capability, etc) and to bypass the legacy Shortest-Path First (SPF) algorithm. Note however that this is not mandatory and that in some cases an SPF routing could be applied.

In order for such a constrained-based SPF routing of LSPs to happen, the nodes performing LSP establishment need more information about the links in the network than standard intra-domain routing protocols provide. These TE attributes are distributed using the transport mechanisms already available in IGPs and are taken into consideration by the LSP routing algorithm. Optimization of the LSP trajectories may also require some external simulations using heuristics that serve as input for the actual path calculation and LSP establishment process.

Extensions to traditional routing protocols and algorithms are needed to uniformly encode and carry TE link information, and explicit routes (e.g. source routes) are required in the signaling. In addition, the signaling must now be capable of transporting the

required circuit (LSP) parameters such as the bandwidth, the type of signal, the desired protection, the position in a particular multiplex, etc. Most of these extensions have already been defined for PSC (IP) traffic engineering with MPLS. GMPLS mainly adds additional extensions for TDM, LSC and FSC traffic engineering, by staying as generic as possible. Only a very few elements are technology specific.

Thus, GMPLS extends the two signaling protocols defined for MPLS-TE signaling, i.e. RSVP-TE and CR-LDP. However, GMPLS does not specify which one of these two signaling protocols must be used. It is the role of manufacturers and operators to evaluate the two possible solutions for their own interest.

Since GMPLS is based on RSVP-TE and CR-LDP, it mandates a

downstream-on-demand label allocation and distribution, with an ingress initiated ordered control. A liberal label retention is normally used, but a conservative label retention mode could be used. There is no restriction on the label allocation strategy, it can be request driven (obvious for circuit switching technologies), traffic/data driven, or even topology driven. There is no restriction neither on the route selection, explicit routing is normally used (strict or loose) but an hop-by-hop routing could be used as well.

GMPLS extends also two traditional intra-domain routing protocols already extended for TE, i.e. OSPF-TE and IS-IS-TE. However, if explicit routing is used, the routing algorithms used by these protocols don't need to be standardized anymore since they are now used to compute explicit routes only, and are thus not used anymore for hop-by-hop routing. Extensions for inter-domain routing (e.g. BGP) are for further study.

The use of technologies like DWDM (Dense Wavelength Division Multiplexing) implies that we can now have a very large number of parallel links between two directly adjacent nodes (hundreds of wavelengths, or even thousands of wavelengths if multiple fibers are used). Such a large number of links was not originally considered for an IP or MPLS control plane. Some slight adaptations of that control plane are thus required if we want to reuse it in the GMPLS context.

For instance, the traditional IP routing model assumes the establishment of a routing adjacency over each link connecting two adjacent nodes. Having such a large number of adjacencies is not scalable at all. Each node needs to maintain each of its adjacencies one by one, and link state routing information must be flooded in the topology for each link.

To solve this issue the concept of bundling was introduced. Moreover, the manual configuration and control of these links, even if they are unnumbered, becomes totally impractical. The Link Management Protocol (LMP) was specified to solve these problems.

LMP runs between data-plane adjacent nodes and is used for both link provisioning and fault isolation. LMP was defined in the context of GMPLS, but was specified independently of the GMPLS signaling specification. It results that LMP can be reused in other contexts, with non-GMPLS signaling protocols as well.

A unique feature of LMP is that it is able to isolate faults in both

opaque and transparent networks, independent of the encoding scheme used for the data. LMP will be used to verify connectivity between nodes; and isolate link, fiber, or channel failures within the network.

The MPLS signaling and routing protocols require at least one bi-directional control channel to communicate even if two adjacent node are connected by unidirectional links. Several control channels can be used. LMP can be used to establish, maintain and manage these control channels.

GMPLS does not specify how these control channels must be implemented, but GMPLS requires IP to transport the signaling and routing protocols over them. Control channels can be either in-band or out-of-band, and several solutions can be used to carry IP. Note also that one type of LMP message is used in-band in the data plane and may not be transported over IP, but this is a particular case, needed to verify connectivity in the data plane.

3.4. Key Differences Between MPLS-TE and GMPLS

Some key differences between MPLS-TE and GMPLS are highlighted in the following. Some of them are key advantages of GMPLS to control non-PSC layers.

- In MPLS-TE, links traversed by an LSP can include an intermix of links with heterogeneous label encoding (e.g. links between routers, links between routers and ATM-LSRs, and links between ATM-LSRs. GMPLS extends this by including links where the label is encoded as a time slot, or a wavelength, or a position in the real world physical space.

- In MPLS-TE, an LSP that carries IP has to start and end on a router. GMPLS extends this by requiring an LSP to start and end on similar type of LSRs.

- The type of a payload that can be carried in GMPLS by an LSP is extended to allow such payloads as SONET/SDH, 1 or 10Gb Ethernet, etc.

- For non-PSC interfaces, bandwidth allocation for an LSP can be performed only in discrete units.

- It is expected to have (much) fewer labels on non-PSC links than on PSC links.

- The use of Forwarding Adjacencies (FA), provides a mechanism that may improve bandwidth utilization, when bandwidth allocation can be performed only in discrete units, as well as a mechanism to aggregate forwarding state, thus allowing the number of required labels to be reduced

- GMPLS allows for a label to be suggested by an upstream node to reduce the setup latency. This suggestion may be overridden by a downstream node but, in some cases, at the cost of higher LSP setup time.

- GMPLS extends on the notion of restricting the range of labels that may be selected by a downstream node. In GMPLS, an ingress or other upstream node may restrict the labels that may be used by an LSP along either a single hop or along the whole LSP path.

- While traditional TE-based (and even LDP-based) LSPs are unidirectional, GMPLS supports the establishment of bi-directional LSPs.

- GMPLS supports the termination of an LSP on a specific egress port, i.e. the port selection at the destination side.

- GMPLS with RSVP-TE supports an RSVP specific mechanism for rapid failure notification.

4. Routing and addressing model

GMPLS is based on the IP routing and addressing models. This assumes that IPv4 and/or IPv6 addresses are used to identify interfaces and that traditional (distributed) IP routing protocols are also reused. Indeed, the discovery of the topology and the resource state of all links in a routing domain is achieved via these routing protocols.

Since control and data planes are de-coupled in GMPLS, one cannot do anymore the assumption that control-plane neighbors (i.e. IGP-learnt neighbors) are data-plane neighbors, hence mechanisms like LMP are needed to associate TE links with neighboring nodes.

IP addresses are not used only to identify interfaces of IP hosts and routers, but more generally to identify any PSC and non-PSC interfaces. Similarly IP routing protocols are not used only to find routes for IP datagrams but also to find routes for non-PSC circuits by using a CSPF algorithm instead of legacy SPF.

However, some additional mechanisms are needed to increase the scalability of these models and to deal with specific traffic engineering requirements of non-PSC layers. These mechanisms will be introduced in the following.

Re-using existing IP routing protocols allows for non-PSC layers to take advantages of all the valuable developments that toke place

since years for IP routing, in particular in the context of link-state routing and policy routing.

Each particular non-PSC layer can be seen as a set of Autonomous Systems (ASs) interconnected in an arbitrary way. Similarly to the traditional IP routing, each AS is managed by a single administrative authority. For instance, an AS can be an SDH/SONET network operated by a given carrier. The set of interconnected ASs being an SDH/SONET Internetwork.

Exchange of routing information between ASs can be done via an inter-domain routing protocol like BGP-4. There is obviously a huge value of re-using well-known policy routing facilities provided by BGP in a non-PSC context. Extensions for BGP traffic engineering in the context of non-PSC layers are for further study.

Each AS can be subdivided in different routing domains, and each can run a different intra-domain routing protocol. In turn, each routing-domain can be divided in areas.

A routing domain is made of GMPLS nodes. These nodes can be either edge nodes (i.e. hosts, ingress LSRs or egress LSRs), or internal LSRs. An example of non-PSC host is an SDH/SONET Terminal Multiplexer (TM). Another example, is an SDH/SONET interface card within an IP router or ATM switch.

Note that traffic engineering in the intra-domain requires the use of link-state routing protocols like OSPF or IS-IS.

GMPLS defines extensions to these protocols. These extensions are needed to disseminate specific non-PSC static and dynamic characteristics related to nodes and links. The current focus is on intra-area traffic engineering. However, inter-area traffic engineering is also under investigation.

4.1 Addressing of PSC and non-PSC layers

The fact that IPv4 and/or IPv6 addresses are used doesn't imply at all that they should be allocated in the same addressing space than public IPv4 and/or IPv6 addresses used for the Internet. Each layer could have a different addressing authority responsible for address allocation and re-using the full addressing space, completely independently.

Private IP addresses can be used if they don't require to be exchanged with any other operator, public IP addresses are otherwise required. Of course, if an integrated model is used, two layers

could share the same addressing space.

Note that there is a benefit of using public IPv4 and/or IPv6 Internet addresses for non-PSC layers if an integrated model with the IP layer is foreseen.

If we consider the scalability enhancements proposed in the next section, the IPv4 (32 bits) and the IPv6 (128 bits) addressing spaces are both more than sufficient to accommodate any non-PSC

layer. We can reasonably expect to have much less non-PSC devices (e.g. SDH/SONET nodes) than we have today IP hosts and routers.

Other kinds of addressing schemes (e.g. NSAP) are not considered here since this would imply a modification of the already existing signaling and routing protocols that uses IPv4 and/or IPv6 addresses. This would be incompatible to our objectives of re-using existing IP protocols.

4.2 GMPLS scalability enhancements

Non-PSC layers introduce new constraints on the IP addressing and routing models since several hundreds of parallel physical links (e.g. wavelengths) can now connect two nodes. Most of the carriers already have today several tenths of wavelengths per fiber between two nodes. New generation of DWDM systems will allow several hundreds of wavelengths.

It becomes rather impractical to associate an IP address to each end of each physical link, to represent each link as a separate routing adjacency, and to advertise link states for each of these links. For that purpose, GMPLS enhances the MPLS routing and addressing models to increase their scalability.

Two optional mechanisms can be used to increase the scalability of the addressing and the routing: unnumbered links and link bundling. These two mechanisms can also be combined. They require extensions to signaling (RSVP-TE and CR-LDP) and routing (OSPF-TE and IS-IS-TE) protocols.

4.3 Extensions to IP TE routing protocols

Traditionally, a TE link is advertised as an adjunct to a "regular" OSPF or IS-IS link, i.e., an adjacency is brought up on the link, and when the link is up, both the regular IGP properties of the link (basically, the SPF metric) and the TE properties of the link are then advertised.

However, GMPLS challenges this notion in three ways:

- first, links that are non-PSC may yet have TE properties; however, an OSPF adjacency cannot be brought up directly on such links.

- second, an LSP can be advertised as a point-to-point TE link in the routing protocol, i.e. as a Forwarding Adjacency (FA); thus, an advertised TE link need no longer be between two OSPF neighbors. Forwarding Adjacencies (FA) are further described in a separate section.

- third, a number of links may be advertised as a single TE link (e.g. for improved scalability), so again, there is no longer a one-to-one association of a regular adjacency and a TE link.

Thus we have a more general notion of a TE link. A TE link is a logical link that has TE properties, some of which may be configured on the advertising LSR, others which may be obtained from other LSRs by means of some protocol, and yet others which may be deduced from the component(s) of the TE link.

An important TE property of a TE link is related to the bandwidth accounting for that link. GMPLS will define different accounting rules for different non-PSC layers. Generic bandwidth attributes are however defined by the TE routing extensions and by GMPLS, such as the unreserved bandwidth, the maximum reservable bandwidth, the maximum LSP bandwidth.

It is expected in a dynamic environment to have frequent changes of bandwidth accounting information. A flexible policy for triggering link state updates based on bandwidth thresholds and link dampening mechanism can be implemented.

TE properties associated with a link should also capture protection and restoration related characteristics. For instance, shared protection can be elegantly combined with bundling. Protection and restoration are mainly generic mechanisms also applicable to MPLS. It is expected that they will first be developed for MPLS and later on generalized to GMPLS.

A TE link between a pair of LSRs doesn't imply the existence of an IGP adjacency between these LSRs. A TE link must also have some means by which the advertising LSR can know of its liveness (e.g. by using LMP hellos). When an LSR knows that a TE link is up, and can determine the TE link's TE properties, the LSR may then advertise that link to its GMPLS enhanced OSPF or IS-IS neighbors using the TE

objects/TLVs. We call the interfaces over which GMPLS enhanced OSPF or ISIS adjacencies are established "control channels".

5. Unnumbered links

Unnumbered links (or interfaces) are links(or interfaces) that do not have IP addresses. Using such links involves two capabilities: (a) the ability to carry (TE) information about unnumbered links in IGP TE extensions (ISIS or OSPF), and (b) the ability to specify unnumbered links in MPLS TE signaling.

The former is covered in ISIS-TE and OSPF-TE. The later requires extensions to RSVP-TE and CR-LDP since MPLS-TE signaling doesn't provide support for unnumbered links. GMPLS defines simple extensions to indicate an unnumbered link in the Explicit Route and Record Route Objects/TLVs of these protocols, using a new Interface ID object/TLV.

Since unnumbered links are not identified by an IP address, then for the purpose of MPLS TE each end need some other identifier, local to the LSR to which the link belongs. Note that links are directed, i.e., a link l is from some LSR A to some other LSR B. LSR A chooses the interface identifier for link l, we call this the "outgoing

interface identifier from LSR A's point of view". If there is a reverse link from LSR B to LSR A, B chooses the outgoing interface identifier for the reverse link. There is no a priori relationship between the two interface identifiers. Both ends must also agree on each of these identifiers.

5.1 Unnumbered Forwarding Adjacencies

If an LSR that originates an LSP advertises this LSP as an unnumbered FA in IS-IS or OSPF, the LSR must allocate an Interface ID to that FA. If the LSP is bi-directional, the tail-end LSR advertises the reverse LSP as an unnumbered FA, the tail-end LSR must allocate an Interface ID to the reverse FA.

Signaling has been enhanced to carry the Interface IDs. When an LSP is created which will be advertised as an FA, the head-end LSR assigns an Interface ID and includes it in the signaling request. The tail-end LSR responds by assigning and including an Interface ID in the signaling response.

6. Link bundling

When a pair of LSRs is connected by multiple links, it is possible to advertise several (or all) of these links as a single link into

OSPF and/or IS-IS. This process is called link bundling, or just bundling. The resulting logical link is called a bundled link as its physical links are called component links.

The purpose of link bundling is to improve routing scalability by reducing the amount of information that has to be handled by OSPF and/or IS-IS. This reduction is accomplished by performing information aggregation/abstraction. As with any other information aggregation/abstraction, this results in losing some of the information. To limit the amount of losses one need to restrict the type of the information that can be aggregated/abstracted.

6.1 Restrictions on bundling

The following restrictions are required for GMPLS. All component links in a bundle must begin and end on the same pair of LSRs, and share some common characteristics: they must have the same type (e.g. point-to-point), the same TE metric, the same set of resource classes, and the same multiplexing capabilities. An FA may be a component link; in fact, a bundle can consist of a mix of point-to-point links and FAs.

6.2 Routing considerations for bundling

A bundled link is just another kind of TE link such as those defined by OSPF-TE or IS-IS-TE. The liveness of the bundled link is determined by the liveness of each of the component links within the bundled link. The liveness of a component link can be determined by any of several means: IS-IS or OSPF hellos over the component link,

or RSVP Hello, or LMP hellos, or from layer 1 or layer 2 indications.

Once a bundled link is determined to be alive, it can be advertised as a TE link and the TE information can be flooded. If IS-IS/OSPF hellos are run over the component links, IS-IS/OSPF flooding can be restricted to just one of the component links.

Note that advertising a (bundled) TE link between a pair of LSRs doesn't imply that there is an IGP adjacency between these LSRs that is associated with just that link. In fact, in certain cases a TE link between a pair of LSRs could be advertised even if there is no IGP adjacency at all between the LSR (e.g. when the TE link is an FA).

Bandwidth accounting must be clearly defined since an abstraction is done. Bandwidth information is an important part of a bundle

advertisement. Some attributes can be sums of component characteristics such as the unreserved bandwidth and the maximum reservable bandwidth. A GMPLS node with bundled links must apply admission control on a per-component link basis.

6.3 Signaling considerations

Typically, an LSP's explicit route (contained in an ERO) will choose the bundled link to be used for the LSP, but not the component link(s), since information about the bundled link is flooded, but information about the component links is kept local to the LSR.

The choice of the component link to use is always made by an upstream node. If the LSP is bidirectional, the upstream node chooses a component link in each direction.

Three mechanisms for indicating this choice to the downstream node are possible.

- Mechanism 1: Implicit Indication

This mechanism requires that each component link has a dedicated signaling channel. The upstream node tells the receiver which component link to use by sending the message over the chosen component link's dedicated signaling channel.

- Mechanism 2: Explicit Indication by IP Address

This mechanism requires that each component link has a unique remote IP address. The upstream node can either send messages addressed to the remote IP address for the component link or encapsulate messages in an IP header whose destination address is the remote IP address. This mechanism does not require each component link to have its own control channel. In fact, it doesn't even require the whole (bundled) link to have its own control channel.

- Mechanism 3: Explicit Indication by Component Interface ID

With this mechanism, each component link in unnumbered and is assigned a unique Interface Identifier. These identifiers are exchanged by the two LSRs at each end of the bundled link. The choice of a component link is indicated by an upstream node by including the corresponding identifier in signaling messages. Discovering Interface Identifiers at bootstrap may be accomplished by configuration, by means of a protocol such as LMP (preferred solution), or by means of RSVP/CR-LDP (especially in the case where a component link is a Forwarding Adjacency). New objects are needed

to indicate Interface Identifiers in signaling, GMPLS defines one Upstream Interface ID object/TLV and one Downstream Interface ID object/TLV.

6.4 Unnumbered Bundled Link

A bundled link may itself be numbered or unnumbered independent of whether the component links are numbered or not. This affects how the bundled link is advertised in IS-IS/OSPF, and the format of LSP EROs that traverse the bundled link. Furthermore, unnumbered Interface Identifiers for all unnumbered outgoing links of a given LSR (whether component links, Forwarding Adjacencies or bundled links) MUST be unique in the context of that LSR.

7. UNI and NNI

The interface between an edge GMPLS node and a GMPLS LSR on the network side may be referred to as a User to Network Interface (UNI), while the interface between two network side LSRs may be referred to as a Network to Network Interface (NNI).

GMPLS does not specify separately a UNI and an NNI. Edge nodes are connected to LSRs on the network side, and these LSRs are in turn connected between them. Of course, the behavior of an edge node is not exactly the same as the behavior of an LSR on the network side. Note also, that an edge node may run a routing protocol, however it is expected that in most of the cases it will not (see also section 7.2 and the section about signaling with an explicit route).

Conceptually, a difference between UNI and NNI make sense either if both interface uses completely different protocols, or if they use the same protocols but with some outstanding differences. In the first case, separate protocols are often defined successively, with more or less success.

The GMPLS approach consisted in building a consistent model from day one, considering both the UNI and NNI interfaces at the same time. For that purpose a very few specific UNI particularities have been ignored in a first time. GMPLS is being enhanced to support such particularities at the UNI by some other standardization bodies, like the OIF.

7.1 OIF UNI versus GMPLS

The current OIF UNI specification [OIF-UNI] defines an interface between a client SDH/SONET equipment and an SDH/SONET network, each belonging to a distinct administrative authority. The OIF UNI

defines additional mechanisms on the top of GMPLS for the UNI

For instance, the OIF service discovery procedure is a precursor to obtaining UNI services. Service discovery allows a client to determine the static parameters of the interconnection with the network, including the UNI signaling protocols, the transparency levels as well as the protection level supported by the network.

Moreover, the only additional mechanism covered by the OIF UNI is the address allocation process. The corresponding mechanism is tightly related to the link bundle mechanism as described in LMP using LinkSummary (and include an address allocation request) and LinkSummaryAck (and include an address allocation response) messages.

Since the current OIF UNI interface does not cover photonic networks, G.709 Digital Wrapper, etc, it is a sub-set of the GMPLS Architecture.

7.2 Routing at the UNI

This section discusses the selection of an explicit route by an edge node. The selection of the first LSR by an edge node connected to multiple LSRs is part of that problem.

An edge node (host or LSR) can participate more or less deeply in the GMPLS routing. Four different routing models can be supported at the UNI: configuration based, partial peering, silent listening and full peering.

- Configuration based: this routing model requires the manual or automatic configuration of an edge node with a list of neighbor LSRs sorted by preference order. Automatic configuration can be achieved using DHCP for instance. No routing information is exchanged at the UNI, except maybe the ordered list of LSRs. The only routing information used by the edge node is that list. The edge node sends by default an LSP request to the preferred LSR. ICMP redirects could be send by this LSR to redirect some LSP requests to another LSR connected to the edge node. GMPLS does not preclude that model.

- Partial peering: limited routing information (mainly reachability) can be exchanged across the UNI using some extensions in the signaling plane. The reachability information exchanged at the UNI may be used to initiate edge node specific routing decision over the network. GMPLS does not have any capability to support this model today.

- Silent listening: the edge node can silently listen to routing protocols and take routing decisions based on the information obtained. An edge node receives the full routing information, including traffic engineering extensions. One LSR should forward

transparently all routing pdus to the edge node. An edge node can now compute a complete explicit route taking into consideration all the end-to-end routing information. GMPLS does not preclude this model.

- Full peering: In addition to silent listening, the edge node participates within the routing, establish adjacencies with its neighbors and advertises LSAs. This is useful only if there are benefits for edge nodes to advertise themselves traffic engineering information. GMPLS does not preclude this model.

8. Link Management

In the context of GMPLS, a pair of nodes (e.g., a photonic switch) may be connected by tenths of fibers, and each fiber may be used to transmit hundreds of wavelengths if DWDM is used. Furthermore, multiple fibers and/or multiple wavelengths may be combined into one or more bundled links as explained previously.

Dealing with hundreds or thousands of individual or bundled links between two nodes requires the help of some signaling tools. In addition, at least one control channel must be established and maintained between a node pair, possibly, using some of these links.

Link management is a collection of useful functionality between adjacent nodes that provide different local services such as control channel management, link connectivity verification, link property correlation, and fault isolation. A Link Management Protocol (LMP) has been defined to fulfill these operations. LMP was initiated in the context of GMPLS but is indeed a generic toolbox that can be also used in other contexts.

Control channel management and link connectivity verification are mandatory mechanisms of LMP. Link property correlation and fault isolation are optional.

8.1 Control channel

Control plane communications between neighboring nodes need a bi-directional control channel. The control channel can be used to exchange MPLS control-plane information such as signaling, routing and management information.

In GMPLS, the control channel(s) between two adjacent nodes is no longer required to use the same physical medium as the data-bearing links between those nodes. For example, a control channel could use a separate wavelength or fiber, an Ethernet link, or an IP tunnel through a separate management network. A consequence of allowing the

control channel(s) between two nodes to be physically diverse from the associated data-bearing links is that the health of a control channel does not necessarily correlate to the health of the data-bearing links, and vice-versa. Therefore, new mechanisms must be developed to manage links, both in terms of link provisioning and fault isolation.

It is essential that a control channel is always available, 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 node.

If a primary control channel cannot be established, then an alternate control channel should be tried. Of course, alternate control channels should be pre-configured, however, coordinating the switchover of the control channel to an alternate channel is still an important issue.

Specifically, if the control channel fails but the node is still operational (i.e., the data-bearing links are still passing user data), then both the local and remote nodes should switch to an alternate control channel. If the bi-directional control channel is implemented using two separate unidirectional channels, and only one direction of the control channel has failed, both the local and remote nodes need to understand that the control channel has failed so that they can coordinate a switchover. LMP provides a graceful switchover from one control channel to the other.

8.2 Control channel management

Once a control channel is configured between two neighboring nodes, a Hello protocol will be used to establish and maintain connectivity between the nodes and to detect failures. The Hello protocol of LMP is intended to be a lightweight keep-alive mechanism that will react to control channel failures rapidly so that IGP Hellos are not lost and the associated link-state adjacencies are not removed unnecessarily.

The Hello protocol consists of two phases: a negotiation phase and a keep-alive phase. The negotiation phase allows negotiation of some basic Hello protocol parameters, like the Hello frequency. The keep-alive phase consists of a fast lightweight Hello message exchange.

The failure of a control channel can also be detected by lower layers (e.g., SONET/SDH) since control channels are electrically terminated at each node.

8.3 Control channel interfaces

LMP functions to maintain logical control channels between a pair of nodes via control channel interfaces. Each control channel interface hides a set of control channels and which of these is actually used to transport the messages and how this is achieved. This isolate signaling, routing and management from the actual control channel management.

LMP does not specify how control channels are implemented, however it states that messages transported over a control channel must be IP encoded. Furthermore, since the messages are IP encoded, the link level encoding is not part of LMP.

LMP associates (possibly multiple) link bundles with a control channel. Multiple control channels may then be configured and associated with a control channel interface. One control channel is actually used while the others are backup control channels sorted by preference order. The control channel interface is announced into the IGP domain so that messages can be routed to that interface. The associations between the control channels and the control channel interface are purely a local matter.

The control channel of a link bundle can be either explicitly configured or automatically selected, however, GMPLS currently assume that the control channel is explicitly configured. Once a link bundle is associated with a control channel, it follows the failover of that control channel. The association of the control channel to the control channel interface is configured or automatically bootstrapped and is a local issue.

Between any two adjacent nodes (from the perspective of link bundles) there may be multiple active control channel interfaces, and these control channel interfaces are used for LMP, routing, and signaling messages. For purposes of flooding routing messages, LMP messages, and signaling messages, any of the active control channel interfaces may be used.

8.4 Link property correlation

A link property exchange mechanism allows to dynamically change some link characteristics. It allows for instance to add data-bearing links to a link bundle, change a link's protection mechanism, change port identifiers, or change component identifiers in a bundle. This mechanism is supported by an exchange of link summary messages.

8.5 Link connectivity verification

Link connectivity verification is an optional procedure that may be used to verify the physical connectivity of data-bearing links (e.g. component links of a bundle)as well as to exchange the link identifiers that will be further used in the RSVP-TE and CR-LDP signaling.

The use of this procedure is negotiated as part of the configuration exchange that take place during the negotiation phase of the Hello protocol. If enabled, the procedure is done initially when a link bundle is first established, and subsequently, on a periodic basis for all free component links of a link bundle.

Ping-type Test messages are exchanged over each of the data-bearing links specified in the bundled link. It should be noted that all LMP 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 data-bearing link verification process. The Test message is sent over the data-bearing link that is being verified. Data-bearing links are tested in the transmit direction as

they are uni-directional, and as such, it may be possible for both nodes to exchange the Test messages simultaneously.

Before exchanging these test messages, the node that initiates the verification indicates to the adjacent node that it will begin sending test messages across the data-bearing links of a particular bundled link. It indicates also the number of data-bearing links that are to be verified; the interval at which the test messages will be sent; the encoding scheme, the transport mechanism that are supported, and data rate for Test messages; and, in the case where the data-bearing links correspond to fibers, the wavelength over which the Test messages will be transmitted. The transport mechanism is negotiated between the two nodes. Furthermore, the local and remote bundle identifiers are transmitted at this time to perform the data-bearing link association with the bundle identifiers.

A unique characteristic of photonic switches (all-optical) is that the data being transmitted over a data-bearing link is not terminated at the switch, but instead passes through transparently. This characteristic of PXCs poses a challenge for validating the connectivity of the data-bearing links.

Therefore, to ensure proper verification of data-bearing link connectivity in that case, we require that until the links are allocated, it must be possible to terminate them locally. There is no requirement that all data-bearing links be terminated

simultaneously, but at a minimum, the data-bearing links must be able to be terminated one at a time. Furthermore, we assume that the nodal architecture is designed so that messages can be sent and received over any data-bearing link. Note that this requirement is trivial for a digital switch since each data-bearing link is received electronically before being forwarded to the next switch. This is an additional requirement for photonic switches.

8.6 Fault localization

Fault localization or isolation is an important requirement from the operational point of view. When a failure occurs an operator needs to know where exactly it happened. It can also be used to support some specific local protection/restoration mechanisms. Logically, fault localization can occur only after a fault is detected.

Fault 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 and are for further study. However, it should be clear that the mechanism used to locate the failure is independent of the mechanism used to detect the failure, but simply relies on the fact that a failure is detected.

In new technologies such as transparent photonic switching currently no method is defined to locate a fault, and the mechanism by which

the fault information is propagated must be sent ôout of bandö (via the control plane).

Fault localization is an optional LMP procedure that is used to rapidly locate link failures. The use of this procedure is also negotiated as part of the configuration exchange that take place during the negotiation phase of the Hello protocol. As before, we assume each link has a bi-directional control channel that is always available for inter-node communication and that the control channel spans a single hop between two neighboring nodes.

The mechanism used to rapidly isolate link failures is designed to work for unidirectional LSPs, and can be easily extended to work for bi-directional LSPs.

If data-bearing links fail between two photonic switches, the power monitoring system in all of the downstream nodes will detect LOL and indicate a failure. To correlate multiple failures between a pair of nodes, a monitoring window can be used in each node to determine if

a single data-bearing link has failed or if multiple data-bearing links have failed. As part of the fault localization, a downstream node that detects data-bearing link failures will send a channel fail message to its upstream neighbor (bundling together the notification of all of the failed data-bearing links).

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 LSP()s using this/these link(s). If there is also a failure on the input port(s) of the upstream node, the node will return a message to the downstream node (bundling together the notification of all the data-bearing links), indicating that it too has detected a failure. If, however, the fault is clear in the upstream node (e.g., there is no LOL on the corresponding input channels), then the upstream node will have localized the failure and will return a specific message to the downstream node. Once the failure has been localized, the signaling protocols can be used to initiate span or path protection/restoration procedures.

9. Generalized Signaling

The GMPLS signaling extends certain base functions of the RSVP-TE and CR-LDP signaling and, in some cases, add functionality. These changes and additions impact basic LSP properties, how labels are requested and communicated, the unidirectional nature of LSPs, how errors are propagated, and information provided for synchronizing the ingress and egress.

The GMPLS signaling specification is available in three parts:

1. A signaling functional description [GMPLS-SIG].
2. RSVP-TE extensions [GMPLS-RSVP-TE].
3. CR-LDP extensions [GMPLS-CR-LDP].

The following MPLS profile applies to GMPLS:

- Downstream-on-demand label allocation and distribution.
- Ingress initiated ordered control.
- Liberal (typical), or conservative (could) label retention mode.
- Request, traffic/data, or topology driven label allocation strategy.
- Explicit routing (typical), or hop-by-hop routing (could).

The GMPLS signaling defines the following new building blocks on the top of MPLS-TE:

1. A new label request format to encompass non-PSC characteristics.
2. Labels for non-PSC interfaces, generically known as Generalized Label.
3. Waveband switching support.
4. Label suggestion by the upstream for optimization purposes (e.g. latency).
5. Label restriction by the upstream to support some optical constraints.
6. Bi-directional LSP establishment with contention resolution.
7. Rapid failure notification to ingress node.
8. Explicit routing with explicit label control for a fine degree of control.

These building blocks will be described in mode details in the following. A complete specification can be found in the corresponding documents.

Note that GMPLS is highly generic and optional. Only building blocks 1 and 2 are mandatory, and only within the specific format that is needed. Typically building blocks 6 and 8 should be implemented. Building blocks 3, 4, 5 and 7 are optional.

A typical SDH/SONET switching network would implement building blocks: 1 (but the SDH/SONET format), 2 (the SDH/SONET label), 6 and 8. It could implement another format of label in case of link bundling. Building block 7 is optional since the protection/restoration can be achieved using SDH/SONET overhead bytes.

A typical wavelength switching network would implement building blocks: 1 (but the wavelength label), 2 (the generic format), 4, 5, 6, 7 and 8. Building block 3 is only needed in the particular case of waveband switching.

A typical fiber switching network would implement building blocks: 1 (but the port label), 2 (the generic format), 6, 7 and 8.

A typical MPLS-IP network would not implement any of these building blocks, since the absence of building block 1 would indicate regular

MPLS-IP. Note however that building block 1 can be used to signal MPLS-IP as well. In that case, the MPLS-IP network can benefit from the link protection type (not available in CR-LDP, some very basic form being available in RSVP-TE). Building block 2 is here a regular MPLS label and no new label format is required.

GMPLS does not specify any profile for RSVP-TE and CR-LDP implementations that have to support GMPLS - except for what is directly related to GMPLS procedures. It is to the manufacturer to decide which are the optional elements and procedures of RSVP-TE and CR-LDP that need to be implemented. Some optional MPLS-TE elements can be useful for non-PSC layers, for instance the setup and holding priorities that are inherited from MPLS-TE.

9.1. Overview: How to Request an LSP

A non-PSC LSP is established by sending a PATH/Label Request message downstream to the destination. This message contains a Generalized Label Request with the type of LSP (i.e. the layer concerned), its payload type and the requested local protection per link. An Explicit Route (ERO) is also normally added to the message, but this can be added and/or completed by the first/default LSR.

The requested bandwidth is encoded in the RSVP-TE SENDER_TSPEC and FLOWSPEC objects, or in the CR-LDP Traffic Parameters TLV. The end-to-end protection type is for further study. In case of SDH/SONET concatenation, the requested bandwidth is the total bandwidth and a field in the Generalized Label Request allows to know the number of components.

Specific parameters for a given technology are given in the Generalized Label Request, such as the type of concatenation and/or transparency for a SDH/SONET LSP.

If the LSP is a bi-directional LSP, an Upstream Label is also specified in the Path/Label request message. This label will be the one to use in the downstream to upstream direction.

Additionally, a Suggested Label, a Label Set and a Waveband Label can also be included in the message. Other operations are defined in MPLS-TE.

The downstream node will send back a Resv/Label Mapping message including one Generalized Label object/TLV that can contain several Generalized Labels. For instance, if a concatenated SDH/SONET signal is requested, several labels can be returned.

In case of SDH/SONET virtual concatenation, a list of labels is returned. Each label identifying one element of the virtual concatenated signal. This limits virtual concatenation to remain within a single (component) link.

In case of any type of SDH/SONET contiguous concatenation, only one label is returned. That label is the lowest signal of the contiguous concatenated signal (given an order specified in [GMPLS-SIG].

In case of SDH/SONET bundling, i.e. co-routing of circuits of the same type but without concatenation, the explicit list of all signals that take part in the bundling is returned.

9.2. Generalized Label Request

The Generalized Label Request is a new object/TLV to be added in an RSVP-TE Path message instead of the regular Label Request, or in a CR-LDP Request message in addition to the already existing TLVs. Only one label request can be used per message, so a single LSP can be requested at a time per signaling message.

The Generalized Label Request gives some major characteristics (parameters) required to support the LSP being requested, such as the LSP encoding type, the LSP payload type, the desired link protection.

GMPLS defines a generic Generalized Label Request, and in addition it can define specialized Generalized Label Requests, if and only if there are specific characteristics that cannot be signaled by the generic request, i.e. specific characteristics.

Currently, only one specific Generalized Label Request is defined, for SDH/SONET. The SDH/SONET Generalized Label Request indicates the same generic characteristics as the generic request but includes in addition the requested SDH/SONET concatenation and transparency (if needed).

Note that it is expected than a specific Generalized Label Request will be defined in the future for photonic (all optical) switching.

The characteristics described hereafter are generic to all technologies:

- The LSP encoding type.
- The LSP payload type.
- The link protection type.

The LSP encoding type indicates the type of technology (e.g. Ethernet, SDH, SONET, fiber, etc) to which this requested LSP corresponds. It represents the nature of the LSP, and not the nature of the links that the LSP traverses. A link may support a set of encoding formats, where support means that a link is able to carry and switch a signal of one or more of these encoding formats depending on the resource availability and capacity of the link.

For example, consider an LSP signaled with "photonic" encoding. It is expected that such an LSP would be supported with no electrical conversion and no knowledge of the modulation and speed by the

transit nodes. Some other formats (electrical) require other knowledge such as the bandwidth.

The LSP payload type identifies the payload carried by an LSP, i.e. the client layer of that LSP. This must be interpreted according to the technology encoding type of the LSP and is used by the nodes at the endpoints of the LSP to know to which client layer a request is destined.

The link protection type indicates the desired local link protection for each link of an LSP. If a particular protection type, i.e., 1+1, or 1:N, is requested, then a connection request is processed only if the desired protection type can be honored. Note that GMPLS advertises the protection capabilities of a link in the routing protocols. Path computation algorithms may take this information into account when computing paths for setting up LSPs.

9.3. Generalized Label

The Generalized Label extends the traditional MPLS label by allowing the representation of not only labels which identify and travel in-band with associated data packets, but also (virtual) labels which identify time-slots, wavelengths, or space division multiplexed positions.

For example, the Generalized Label may identify (a) a single fiber in a bundle, (b) a single waveband within fiber, (c) a single wavelength within a waveband (or fiber), or (d) a time-slot within a wavelength (or fiber). It may also be a generic MPLS label, a Frame Relay label, or an ATM label (VCI/VPI). The format of a label can be as simple as an integer value such as a wavelength label or can be more elaborated such as an SDH/SONET label.

SDH and SONET define each a multiplexing structure. These multiplexing structures will be used as naming trees to create unique labels. Such a label will identify the type of a particular signal (time-slot) and its exact position in a multiplexing structure (both are related). Since the SONET multiplexing structure may be seen as a subset of the SDH multiplexing structure, the same format of label is used for SDH and SONET.

Since the nodes sending and receiving the Generalized Label know what kinds of link they are using, the Generalized Label does not

identify its type, instead the nodes are expected to know from the context what type of label to expect.

A Generalized Label only carries a single level of label, i.e., it is non-hierarchical. When nested LSPs are used, each LSP must be established separately and has its own label at each local interface between two nodes at its level.

9.4. Waveband Switching

A special case of wavelength switching is waveband switching. A waveband represents a set of contiguous wavelengths which can be switched together to a new waveband. For optimization reasons it may be desirable for an photonic cross-connect to optically switch multiple wavelengths as a unit. This may reduce the distortion on the individual wavelengths and may allow tighter separation of the individual wavelengths. A Waveband label is defined to support this special case.

Waveband switching naturally introduces another level of label hierarchy and as such the waveband is treated the same way all other upper layer labels are treated. As far as the MPLS protocols are concerned there is little difference between a waveband label and a wavelength label except that semantically the waveband can be subdivided into wavelengths whereas the wavelength can only be subdivided into time or statistically multiplexed labels.

9.5. Label Suggestion by the Upstream

GMPLS allows for a label to be suggested by an upstream node. This suggestion may be overridden by a downstream node but, in some cases, at the cost of higher LSP setup time. The suggested label is valuable when establishing LSPs through certain kinds of optical equipment where there may be a lengthy (in electrical terms) delay in configuring the switching fabric. For example micro mirrors may have to be elevated or moved, and this physical motion and subsequent damping takes time. If the labels and hence switching fabric are configured in the reverse direction (the norm) the MAPPING/Resv message may need to be delayed by 10's of milliseconds per hop in order to establish a usable forwarding path. It can also be important for restoration purposes where alternate LSPs may need to be rapidly established as a result of network failures.

9.6. Label Restriction by the Upstream

An upstream node can optionally restrict (limit) the choice of label

of a downstream node to a set of acceptable labels. This restriction is done by giving a list of inclusive (acceptable) or exclusive (unacceptable) labels in a Label Set. If not applied, all labels from the valid label range may be used. There are four cases where a label restriction is useful in the "optical" domain.

The first case is where the end equipment is only capable of transmitting and receiving on a small specific set of wavelengths/bands.

The second case is where there is a sequence of interfaces which cannot support wavelength conversion and require the same wavelength be used end-to-end over a sequence of hops, or even an entire path.

The third case is where it is desirable to limit the amount of wavelength conversion being performed to reduce the distortion on the optical signals.

The last case is where two ends of a link support different sets of wavelengths.

The receiver of a Label Set must restrict its choice of labels to one which is in the Label Set. A Label Set may be present across multiple hops. In this case each node generates it's own outgoing Label Set, possibly based on the incoming Label Set and the node's hardware capabilities. This case is expected to be the norm for nodes with conversion incapable interfaces.

9.7. Bi-directional LSP

GMPLS allows establishment of bi-directional LSPs. A bi-directional LSP has the same traffic engineering requirements including fate sharing, protection and restoration, LSRs, and resource requirements (e.g., latency and jitter) in each direction. In the remainder of this section, the term "initiator" is used to refer to a node that starts the establishment of an LSP and the term "terminator" is used to refer to the node that is the target of the LSP. For a bi-directional LSPs, there is only one initiator and one terminator.

Normally to establish a bi-directional LSP when using [RSVP-TE] or [CR-LDP] two unidirectional paths must be independently established. This approach has the following disadvantages:

1. The latency to establish the bi-directional LSP is equal to one round trip signaling time plus one initiator-terminator signaling transit delay. This not only extends the setup latency for successful LSP establishment, but it extends the worst-case latency

for discovering an unsuccessful LSP to as much as two times the initiator-terminator transit delay. These delays are particularly significant for LSPs that are established for restoration purposes.

2. The control overhead is twice that of a unidirectional LSP. This is because separate control messages (e.g. Path and Resv) must be generated for both segments of the bi-directional LSP.

3. Because the resources are established in separate segments, route selection is complicated. There is also additional potential race for conditions in assignment of resources, which decreases the overall probability of successfully establishing the bi-directional connection.

4. It is more difficult to provide a clean interface for SDH/SONET equipment that may rely on bi-directional hop-by-hop paths for protection switching. Note that existing SDH/SONET gear transmits the control information in-band with the data.

5. Bi-directional optical LSPs (or lightpaths) are seen as a requirement for many optical networking service providers.

With bi-directional LSPs both the downstream and upstream data paths, i.e. from initiator to terminator and terminator to initiator, are established using a single set of signaling messages.

This reduces the setup latency to essentially one initiator-terminator round trip time plus processing time, and limits the control overhead to the same number of messages as a unidirectional LSP.

For bi-directional LSPs, two labels must be allocated. Bi-directional LSP setup is indicated by the presence of an Upstream Label in the appropriate signaling message.

9.8. Bi-directional LSP Contention Resolution

Contention for labels may occur between two bi-directional LSP setup requests traveling in opposite directions. This contention occurs when both sides allocate the same resources (ports) at effectively the same time. The GMPLS signaling defines a procedure to resolve that contention, basically the node with the higher node ID will win the contention. To reduce the probability of contention, some mechanisms are also suggested.

9.9. Rapid Notification of Failure

GMPLS defines three signaling extensions for RSVP-TE that enable

expedited notification of failures and other events to nodes responsible for restoring failed LSPs, and modify error handling. For CR-LDP there is not currently a similar mechanism.

The first extension, identifies where event notifications are to be sent. The second, provides for general expedited event notification. Such extensions can be used by fast restoration mechanisms.

The final extension is an RSVP optimization to allow the faster removal of intermediate states in some cases.

9.10. Explicit Routing and Explicit Label Control

The path taken by an LSP can be controlled more or less precisely by using an explicit route. Typically, the node at the head-end of an LSP finds a more or less precise explicit route and builds an Explicit Route Object (ERO) that contains that route. Possibly, the edge node don't build any ERO, and just transmit a signaling request to a default neighbor LSR (as IP hosts today). For instance, an explicit route could be added to a signaling message by the first switching node, on behalf of the edge node. Note also that an explicit route is altered by intermediate LSRs during its progression towards the destination.

The ERO is originally defined by MPLS-TE as a list of abstract nodes (i.e. groups of nodes) along the explicit route. Each abstract node can be an IPv4 address prefix, an IPv6 address prefix, or an AS number. This capability allows the generator of the explicit route to have imperfect information about the details of the path. In the simplest case, an abstract node can be a full IP address that identify a specific node (called a simple abstract node).

MPLS-TE allows strict and loose abstract nodes. The path between a strict node and its preceding node must include only network nodes from the strict node and its preceding abstract node. The path between a loose node and its preceding node may include other network nodes that are not part of the strict node or its preceding abstract node.

This ERO was extended to include interface numbers as abstract nodes to support unnumbered interfaces; and further extended by GMPLS to include labels as abstract nodes. Having labels in an explicit route is an important feature that allows to control the placement of an LSP with a very fine granularity. This is more likely to be used for non-PSC links.

In particular, the explicit label control in the ERO allows to

terminate an LSP on a particular outgoing port to an egress node.

This can also be used when it is desirable to "splice" two LSPs together, i.e. where the tail of the first LSP would be "spliced" into the head of the second LSP.

Another use is when an optimization algorithm is used for an SDH/SONET network. This algorithm can provide very detailed explicit routes, including the label (time-slot) to use on a link, in order to minimize the external fragmentation of the SDH/SONET multiplex on the corresponding interface.

Another use is when the label indicates a particular component in a bundle in order to stay diverse with other components of that bundle, i.e. to control the usage of components in a bundle for different LSPs.

9.11 LSP modification and LSP re-routing

LSP modification and re-routing are two features already available in MPLS-TE. GMPLS does not add anything new. Elegant re-routing is possible with the concept of "make-before-break" whereby an old path is still used while a new path is set up by avoiding double reservation of resources. Then, the node performing the re-routing can swap on the new path and close the old path. This feature is supported with RSVP-TE (using shared explicit filters) and CR-LDP (using the action indicator flag).

LSP modification consists in changing some LSP parameters, but normally without changing the route. It is supported using the same mechanism as re-routing. However, the semantic of LSP modification will differ from one technology to the other. For instance, further studies are required to understand the impact of dynamically changing some SDH/SONET circuit characteristics such as the bandwidth, the protection type, the transparency, the concatenation, etc.

9.12. Route recording

In order to improve the reliability and the manageability of the LSP being established, the concept of the route recording was introduced in RSVP-TE to function as:

- First, a loop detection mechanism to discover L3 routing loops, or loops inherent in the explicit route (this mechanism is strictly exclusive with the use of explicit routing objects).

- Second, a route recording mechanism collects up-to-date detailed path information on a hop-by-hop basis during the LSP setup process. This mechanism provides valuable information to the source and destination nodes. Any intermediate routing change at setup time, in case of loose explicit routing, will be reported.

- Third, a recorded route can be used as input for an explicit route. This is useful if a source node receives the recorded route from a destination node and applies it as an explicit route in order to "pin down the path".

Within the GMPLS architecture only the second and third functions are mainly applicable for non-PSC layers.

10. Forwarding Adjacencies (FA)

To improve scalability of MPLS TE (and thus GMPLS) it may be useful to aggregate multiple LSPs inside a bigger LSP. Intermediate nodes see the external LSP only, they don't have to maintain forwarding states for each internal LSP, less signaling messages need to be exchanged and the external LSP can be somehow protected instead (or in addition) to the internal LSPs. This can considerably increase the scalability of the signaling.

The aggregation is accomplished by (a) an LSR creating a TE LSP, (b) the LSR forming a forwarding adjacency out of that LSP (advertising this LSP as a link into ISIS/OSPF), (c) allowing other LSRs to use forwarding adjacencies for their path computation, and (d) nesting of LSPs originated by other LSRs into that LSP (e.g. by using the label stack construct in the case of IP).

An LSR may (under its local configuration control) announce an LSP as a link into ISIS/OSPF. When this link is advertised into the same instance of ISIS/OSPF as the one that determines the route taken by the LSP, we call such a link a "forwarding adjacency" (FA). We refer to the LSP as the "forwarding adjacency LSP", or just FA-LSP. Note that since the advertised entity is a link in ISIS/OSPF, both the end point LSRs of the FA-LSP must belong to the same ISIS level/OSPF area.

In general, creation/termination of a FA and its FA-LSP could be driven either by mechanisms outside of MPLS (e.g., via configuration control on the LSR at the head-end of the adjacency), or by mechanisms within MPLS (e.g., as a result of the LSR at the head-end of the adjacency receiving LSP setup requests originated by some other LSRs).

ISIS/OSPF floods the information about FAs just as it floods the information about any other links. As a result of this flooding, an LSR has in its link state database the information about not just conventional links, but FAs as well.

An LSR, when performing path computation, uses not just conventional links, but FAs as well. Once a path is computed, the LSR uses RSVP-TE/CR-LDP for establishing label binding along the path. FAs needs simple extensions to signaling and routing protocols.

Forwarding adjacencies may be represented as either unnumbered or numbered links. A FA can also be a bundle of LSPs between two nodes.

When a FA is created dynamically, its TE attributes are inherited from the TE 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, but may be bigger if only discrete bandwidths are available for the FA-LSP. In general, for dynamically provisioned forwarding adjacencies, a policy-based mechanism may be needed to associate attributes to forwarding adjacencies.

10.1 Routing and Forwarding Adjacencies

A FA advertisement could contain the information about the path taken by the FA-LSP associated with that FA. This information may be used for path calculation by other LSRs. This information is carried in a new OSPF and IS-IS TLV called the Path TLV.

It is possible that the underlying path information might change over time, via configuration updates, or dynamic route modifications, resulting in the change of that TLV.

If forwarding adjacencies are bundled (via link bundling), and if the resulting bundled link carries a Path TLV, the underlying path followed by each of the FA-LSPs that form the component links must be the same.

It is expected that forwarding adjacencies will not be used for establishing ISIS/OSPF peering relation between the routers at the ends of the adjacency.

10.2. Signaling aspects

For the purpose of processing the ERO in a Path/Request message of an LSP that is to be tunneled over a forwarding adjacency, an LSR at the head-end of the FA-LSP views the LSR at the tail of that FA-LSP as adjacent (one IP hop away).

10.3 Cascading of Forwarding Adjacencies

With an integrated model several layers are controlled using the same routing and signaling protocols. A network may then have links

example, a node may be able to multiplex/demultiplex individual packets on a given link, and may be able to multiplex/demultiplex channels within a SONET payload on other links.

A new OSPF and IS-IS TLV has been defined to advertise the multiplexing capability of each interface: PSC, TDM, LSC or FSC. The information carried in this TLV is used to construct LSP regions, and determine regions' boundaries.

Path computation may take into account region boundaries when computing a path for an LSP. For example, path computation may restrict the path taken by an LSP to only the links whose multiplexing/demultiplexing capability is PSC. When an LSP need to cross a region boundary, it can trigger the establishment of an FA at the underlying layer. This can trigger a cascading of FAs between layers with the following obvious order: TDM, then LSC, and then finally FSC.

11. Security considerations

GMPLS introduces no new security considerations to the current MPLS-TE signaling (RSVP-TE, CR-LDP) and routing protocols (OSPF-TE, IS-IS-TE).

12. Acknowledgements

This draft is the work of numerous authors and consists of a composition of a number of previous drafts in this area.

Many thanks to Ben Mack-Crane (Tellabs) for all the useful SDH/SONET discussions that we had together. Thanks also to Pedro Falcao (Ebone) and Michael Moelants (Ebone) for their SDH/SONET and optical technical advice and support. Finally, many thanks also to Krishna Mitra (Calient) and Curtis Villamizar (Avici).

A list of the drafts from which material and ideas were incorporated follows:

1. draft-ietf-mpls-generalized-signaling-01.txt
Generalized MPLS - Signaling Functional Description

2. draft-ietf-mpls-generalized-rsvp-te-00.txt
Generalized MPLS Signaling - RSVP-TE Extensions

3. draft-ietf-mpls-generalized-cr-ldp-00.txt
Generalized MPLS Signaling - CR-LDP Extensions

4. draft-ietf-mpls-lmp-01.txt
Link Management Protocol (LMP)

5. draft-ietf-mpls-lsp-hierarchy-01.txt
LSP Hierarchy with MPLS TE

6. draft-ietf-mpls-rsvp-unnum-00.txt

draft-many-gmpls-architecture-00.txt Feb 2001

Signalling Unnumbered Links in RSVP-TE

7. draft-ietf-mpls-crldp-unnum-00.txt
Signalling Unnumbered Links in CR-LDP

8. draft-kompella-mpls-bundle-04.txt
Link Bundling in MPLS Traffic Engineering

9. draft-kompella-ospf-gmpls-extensions-00.txt
OSPF Extensions in Support of Generalized MPLS

10. draft-ietf-isis-gmpls-extensions-01.txt
IS-IS Extensions in Support of Generalized MPLS

13. References

TBD

14. Author's Addresses

Peter Ashwood-Smith
Nortel Networks Corp.
P.O. Box 3511 Station C,
Ottawa, ON K1Y 4H7
Canada
Phone: +1 613 763 4534
Email:
petera@nortelnetworks.com

Fong Liaw
Zaffire Inc.
2630 Orchard Parkway
San Jose, CA 95134
USA
Email: fliaw@zaffire.com

Daniel O. Awduche
Movaz Networks
7296 Jones Branch Drive
Suite 615
McLean, VA 22102
USA
Phone: +1 703 847-7350
Email: awduche@movaz.com

Eric Mannie (editor)
Ebone (GTS)
Terhulpsesteenweg 6A
1560 Hoeilaart
Belgium
Phone: +32 2 658 56 52
Email: eric.mannie@gts.com

Ayan Banerjee

Dimitri Papadimitriou

Calient Networks
5853 Rue Ferrari
San Jose, CA 95138
USA
Phone: +1 408 972-3645
Email: abanerjee@calient.net

Alcatel - IPO NSG
Francis Wellesplein, 1
B-2018 Antwerpen
Belgium
Phone: +32 3 240-84-91
Email:
dimitri.papadimitriou@alcatel.be

Debashis Basak
Accelight Networks
70 Abele Road, Bldg.1200
Bridgeville, PA 15017
USA
Phone: +1 412 220-2102 (ext115)
email: dbasak@accelight.com

Dimitrios Pendarakis
Tellium, Inc.
2 Crescent Place
P.O. Box 901
Oceanport, NJ 07757-0901
USA
Email: DPendarakis@tellium.com

Lou Berger
Movaz Networks, Inc.
7926 Jones Branch Drive
Suite 615
MCLean VA, 22102
USA
Phone: +1 703 847-1801
Email: lberger@movaz.com

Bala Rajagopalan
Tellium, Inc.
2 Crescent Place
P.O. Box 901
Oceanport, NJ 07757-0901
USA
Phone: +1 732 923 4237
Email: braja@tellium.com

Greg Bernstein
Ciena Corporation
10480 Ridgeview Court
Cupertino, CA 94014
USA
Phone: +1 408 366 4713
Email: greg@ciena.com

Yakov Rekhter
Juniper
Email: yakov@juniper.net

John Drake
Calient Networks
5853 Rue Ferrari
San Jose, CA 95138
USA
Phone: +1 408 972 3720
Email: jdrake@calient.net

Hal Sandick
Nortel Networks
Email:
hsandick@nortelnetworks.com

Yanhe Fan
Axiowave Networks, Inc.
100 Nickerson Road
Marlborough, MA 01752
USA
Phone: +1 508 460 6969 Ext. 627
Email: yfan@axiowave.com

Debanjan Saha
Tellium Optical Systems
2 Crescent Place
Oceanport, NJ 07757-0901
USA
Phone: +1 732 923 4264
Email: dsaha@tellium.com

Don Fedyk
Nortel Networks Corp.
600 Technology Park Drive
Billerica, MA 01821
USA
Phone: +1-978-288-4506
Email:
dwfedyk@nortelnetworks.com

Vishal Sharma
Jasmine Networks, Inc.
3061 Zanker Road, Suite B
San Jose, CA 95134
USA
Phone: +1 408 895 5030
Email:
vsharma@jasminenetworks.com

Gert Grammel
Alcatel
Italy
Email:
gert.grammel@netit.alcatel.it

George Swallow
Cisco Systems, Inc.
250 Apollo Drive
Chelmsford, MA 01824
USA
Phone: +1 978 244 8143
Email: swallow@cisco.com

Kireeti Kompella
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
USA
Email: kireeti@juniper.net

Z. Bo Tang
Tellium, Inc.
2 Crescent Place
P.O. Box 901
Oceanport, NJ 07757-0901
USA
Phone: +1 732 923 4231
Email: btang@tellium.com

Alan Kullberg
NetPlane Systems, Inc.
888 Washington
St.Dedham, MA 02026
USA
Phone: +1 781 251-5319
Email: akullber@netplane.com

John Yu
Zaffire Inc.
2630 Orchard Parkway
San Jose, CA 95134
USA
Email: jzyu@zaffire.com

Jonathan P. Lang
Calient Networks
25 Castilian
Goleta, CA 93117
Email: jplang@calient.net

Alex Zinin
Cisco Systems
150 W. Tasman Dr.
San Jose, CA 95134
Email: azinin@cisco.com

Full Copyright Statement