Metaswitch Networks

Protection And Restoration In MPLS Networks

An examination of the methods for protecting MPLS LSPs against failures of network resources

Ed Harrison, eph@metaswitch.com
Ben Miller, bmm@metaswitch.com
Adrian Farrel, afarrel@movaz.com

First issued October 2001

Table of Contents

1. Introduction

Multi-Protocol Label Switching (MPLS) is growing in popularity as a set of protocols for provisioning and managing core networks. The networks may be data-centric like those of ISPs, voice-centric like those of traditional telecommunications companies, or one of the modern networks that combine voice and data. These networks are converging on a model that uses the Internet Protocol (IP) to transport data.

MPLS overlays an IP network to allow resources to be reserved and routes pre-determined. Effectively, MPLS superimposes a connection-oriented framework over the connectionless IP network. It provides virtual links or tunnels through the network to connect nodes that lie at the edge of the network.

A well-established requirement in telephone networks is that the network should display very high levels of reliability and availability. Subscribers should not have their calls dropped, and should always have access to their service. Downtime must consequently be kept to a minimum, and backup resources must be provided to take over when any component (link, switch, switch sub-component) fails.

The data world is increasingly demanding similar levels of service to those common in the arena of telephony. Individual customers expect to be able to obtain service at all times and expect reasonable levels of bandwidth. Corporate customers expect the same services, but may also have data streams that are sensitive to delays and disruption.

As voice and data networks merge they inherit the service requirements of their composite functions. Thus, modern integrated networks need to be provisioned using protocols, software and hardware that can guarantee high levels of availability.

High Availability (HA) is typically claimed by equipment vendors when their hardware achieves availability levels of at least 99.999% (five 9s). This may be achieved by provisioning backup copies of hardware and software. When a primary copy fails, processing is switched to the backup. This process, called failover, should result in minimal disruption to the data plane.

Network providers can supply the required levels of service to their customers by building their network from equipment that provides High Availability. This, on its own, is not enough, since network links are also prone to failure, and entire switches may fail. The network provider must also provide backup routes through the network so that data can travel between customer sites even if there is a failure at some point in the network.

This white paper examines the features inherent in MPLS networks that facilitate high availability and considers techniques to build resilient networks by utilizing MPLS. It also examines proposals in the Internet Engineering Task Force (IETF) to standardize methods of signaling and provisioning MPLS networks to achieve protection against failures.

Readers familiar with the concepts of MPLS, network components and network failure survival may want to turn straight to section 3.

2. Background

2.1 Introduction to MPLS

Multi-Protocol Label Switching (MPLS) is rapidly becoming a key technology for use in core networks, including converged data and voice networks. MPLS does not replace IP routing, but works alongside existing and future routing technologies to provide very high-speed data forwarding between Label-Switched Routers (LSRs) together with reservation of bandwidth for traffic flows with differing Quality of Service (QoS) requirements.

MPLS enhances the services that can be provided by IP networks, offering scope for Traffic Engineering, guaranteed QoS and Virtual Private Networks (VPNs).

The basic operation of an MPLS network is shown in the diagram below.

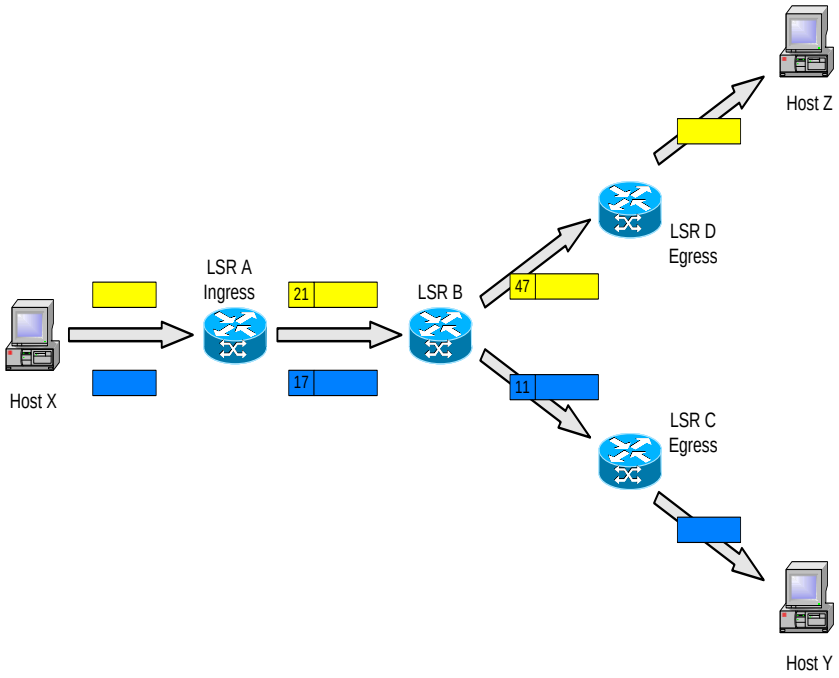

Figure 1: Two LSPs in an MPLS Network

MPLS uses a technique known as label switching to forward data through the network. A small, fixed-format label is inserted in front of each data packet on entry into the MPLS network. At each hop across the network, the packet is routed based on the value of the incoming interface and label, and dispatched to an outwards interface with a new label value.

The path that data follows through a network is defined by the transition in label values, as the label is swapped at each LSR. Since the mapping between labels is constant at each LSR, the path is determined by the initial label value. Such a path is called a Label Switched Path (LSP).

MPLS may also be applied to data switching technologies that are not packet based. The path followed by data through the network is still defined by the transition of switching labels and so is still legitimately called an LSP. However, these non-packet labels (such as wavelength identifiers or timeslots in optical networks) are only used to set up connections, known as cross-connects, at the LSRs. Once the cross-connect is in place all data can be routed without being inspected, so there is no need to place the label value in each packet. Viewed another way, the wavelength or timeslot is itself the label.

At the ingress to an MPLS network, each packet is examined to determine which LSP it should use and hence what label to assign to it. This decision is a local matter but is likely to be based on factors including the destination address, the quality of service requirements and the current state of the network. This flexibility is one of the key elements that make MPLS so useful.

The set of all packets that are forwarded in the same way is known as a Forwarding Equivalence Class (FEC). One or more FECs may be mapped to a single LSP.

Figure 1 shows two data flows from host X: one to Y, and one to Z. Two LSPs are shown.

- LSR A is the ingress point into the MPLS network for data from host X. When it receives packets from X, LSR A determines the FEC for each packet, deduces the LSP to use and adds a label to the packet. LSR A then forwards the packet on the appropriate interface for the LSP.

- LSR B is an intermediate LSR in the MPLS network. It simply takes each labeled packet and uses the pairing {incoming interface, label value} to decide the pairing {outgoing interface, label value} with which to forward the packet. This procedure can use a simple lookup table that can be implemented in hardware - together with the swapping of label value and forwarding of the packet. This allows MPLS networks to be built on existing label switching hardware such as ATM and Frame Relay. This way of forwarding data packets is potentially much faster than examining the full packet header to decide the next hop.

  In the example, each packet with label value 21 will be dispatched out of the interface towards LSR D, bearing label value 47. Packets with label value 17 will be re-labeled with value 11 and sent towards LSR C.

- LSR C and LSR D act as egress LSRs from the MPLS network. These LSRs perform the same lookup as the intermediate LSRs, but the {outgoing interface, label value} pair marks the packet as exiting the LSP. The egress LSRs strip the labels from the packets and forward them using layer 3 routing.

So, if LSR A identifies all packets for host Z with the upper LSP and labels them with value 21, they will be successfully forwarded through the network, emerging from the LSP at D, which then forwards the packets through normal IP to Z.

Note that the exact format of a label and how it is added to the packet depends on the layer 2 link technology used in the MPLS network. For example, a label could correspond to an ATM VPI/VCI, a Frame Relay DLCI, or a DWDM wavelength for optical networking. For other layer 2 types (such as Ethernet and PPP) the label is added to the data packet in an MPLS "shim" header, which is placed between the layer 2 and layer 3 headers. As mentioned above, if the LSP is set up through a network that is not packet switching (such as an optical network), there is no need to place the label in the data packet.

2.1.1 Label Distribution

In order that LSPs can be used, the forwarding tables at each LSR must be populated with the mappings from {incoming interface, label value} to {outgoing interface, label value}. This process is called LSP setup, or Label Distribution.

The MPLS architecture document [3] does not mandate a single protocol for the distribution of labels between LSRs. In fact it specifically allows multiple different label distribution protocols for use in different scenarios, including the following.

- LDP [11]
- CR-LDP [10]
- RSVP-TE [9]
- BGP4
- OSPF

A detailed review of how these protocols are used for label distribution is outside the scope of this white paper. For a comparative analysis of RSVP and CR-LDP, refer to the white paper *MPLS Traffic Engineering: A choice of Signaling Protocols* [1] from Metaswitch.

Several different approaches to label distribution can be used depending on the requirements of the hardware that forms the MPLS network, and the administrative policies used on the network. The underlying principles are that an LSP is set up either in response to a request from the ingress LSR (downstream-on-demand), or preemptively by LSRs in the network, including the egress LSR (downstream unsolicited). It is possible for both to take place at once and for the LSP setup to meet in the middle.

New ideas introduced by the IETF in the MPLS Generalized Signaling draft [13] also allow labels to be pushed from upstream to set up bi-directional LSPs.

Alternatively, LSPs may be configured as "static" or "permanent" LSPs by programming the label mappings at each LSR on the path using some form of management such as SNMP control of the MIBs.

2.1.2 Tunnels and Label Stacks

A key feature of MPLS, especially when considering VPNs, is that once the labels required for an LSP have been exchanged between the LSRs that support the LSP, intermediate LSRs transited by the LSP do not need to examine the content of the data packets flowing on the LSP. For this reason, LSPs are often considered to form tunnels across all or part of the backbone MPLS network. A tunnel carries opaque data between the tunnel ingress and tunnel egress LSRs.

This means that the entire payload, including IP headers, may safely be encrypted without damaging the ability of the network to forward data.

In Figure 1, both LSPs are acting as tunnels. LSR B forwards the packets based only on the label attached to each packet. It does not inspect the contents of the packet or the encapsulated IP header.

An egress LSR may distribute labels for multiple FECs and set up multiple LSPs. Where these LSPs are parallel they can be directed, together, down a higher-level LSP tunnel between LSRs in the network. Labeled packets entering the higher-level LSP tunnel are given an additional label to see them through the network, and retain their first-level labels to distinguish them when they emerge from the higher-level tunnel. This process of placing multiple labels on a packet is known as label stacking and is shown in Figure 2.

Label stacks allow a finer granularity of traffic classification between tunnel ingress and egress nodes than is visible to the LSRs in the core of the network, which need only route data on the basis of the topmost label in the stack. This helps to reduce both the size of the forwarding tables that need to be maintained on the core LSRs and the complexity of managing data forwarding across the backbone.

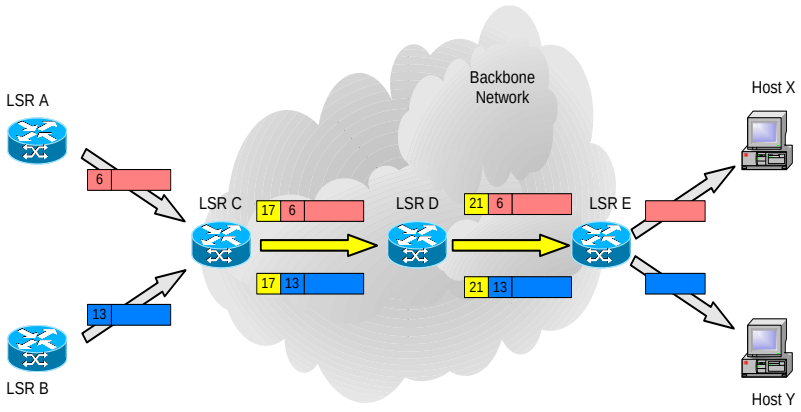

Figure 2: Label Stacks across the backbone

In Figure 2, two LSPs between LSR A and LSR E, and between LSR B and LSR E, shown as red and blue labels, are transparently tunneled across the backbone network in a single outer LSP between LSR C and LSR E.

At the ingress to the backbone network, LSR C routes both incoming LSPs down the LSP tunnel to LSR E, which is the egress from the backbone. To do this, it pushes an additional label onto the label stack of each packet (shown in yellow). LSRs within the backbone, such as LSR D, are aware only of the outer tunnel, shown by the yellow labels. Note that the inner labels are unchanged as LSRs C and D switch the traffic through the outer tunnel – only the outer label is swapped at LSR D.

At the egress of the outer tunnel, the top label is popped off the stack and the traffic is switched according to the inner label. In the example shown, LSR E also acts as the egress for the inner LSPs, so it pops the inner label too and routes the traffic to the appropriate host. The egress of the inner LSPs could be disjoint from E in the same way that LSR A and LSR B are separate from LSR C. Equally, an LSR can act as the ingress for both levels of LSP.

A label stack is arranged with the label for the outer tunnel at the top and the label for the inner LSP at the bottom. On the wire (or fiber) the topmost label is transmitted first and is the only label used for routing the packet until it is popped from the stack and the next highest label becomes the top label.

When a device allocates a label, it can allocate it either from a per platform label space (the Global Label Space) or from a per interface label space. In the first case, the label has global meaning within the device and therefore the outgoing interface and label for the LSP can be identified from this label only. In the second case, the incoming label can only be interpreted in the context of the incoming interface.

If each device allocating a label for the bottom label of a stack (the red and blue labels in Figure 2) allocates such labels from the Global Label Space, the outer tunnel can be re-routed transparently to the inner tunnels (provided that the ingress and egress of the re-routed tunnel are LSRs C and E, respectively). This is because when the packets arrive at LSR E, the outer label will be stripped and the inner label will be correctly interpreted from the Global Label Space. If the labels for the bottom of the stack are from per interface label spaces, this will not be possible. This is because although the re-routed LSP may terminate at the same LSR E, it may terminate on a different interface on LSR E. Once the outer label has been stripped, LSR E will interpret the inner labels as per-interface labels, but now on the wrong interface.

Lastly, it is worth noting that a device can use per interface label spaces for some interfaces, and the Global Label Space for others. Using the Global Label Space for all interfaces on the device gives maximum flexibility for re-routing, but reduces flexibility on allocation of labels.

For a description of the use of label stacking to support VPNs see the white paper *MPLS Virtual Private Networks: A review of the implementation options for MPLS VPNs including the ongoing standardization work in the IETF MPLS Working Group* [2] from Metaswitch.

2.2 Components of an MPLS Network

Before discussing how the elements of an MPLS network can discover, survive or recover from failures, it is important to identify the different components of the network. We can then examine which can fail, what the consequences of the failure are, and how the failures can be handled.

Figure 3 shows a simple piece of an MPLS network. One of the LSRs is shown "exploded" to reveal its internal components.

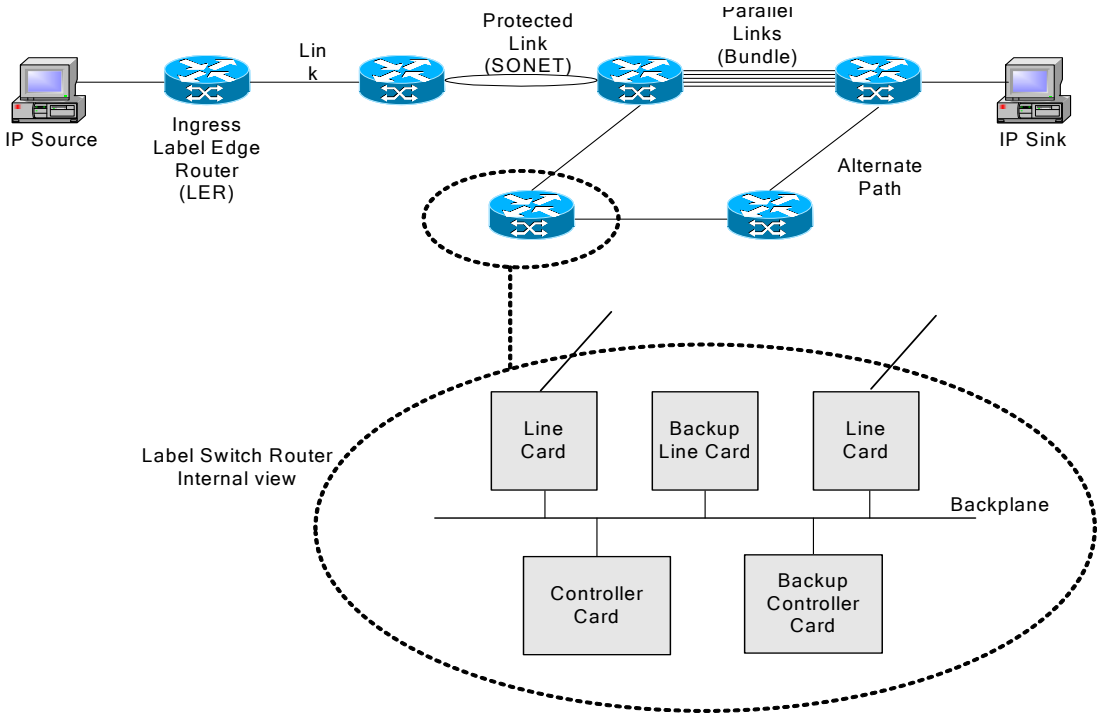

Figure 3: Components of an MPLS network

The diagram highlights some key terms that are expanded here so that they can be used freely throughout the rest of this

IP Source	This is the place from which IP data is sent into the network. Usually thought of as PC (or "host"), this may be any IP device, for example, an IP telephone. It may also be a gateway device that converts between a non-IP service and IP.
IP Sink	The target of the IP data transmission. A partner device to the IP source.
LSR	Label Switch Router. The key switching component of an MPLS network. Responsible for forwarding data according to the rules established by the MPLS signaling protocol.

LER	Label Edge Router. An LSR at the edge of the network that originates or terminates an LSP.
Ingress LER	The LER that receives IP data from the IP source, classifies it and injects it into an LSP for transmission across the network.
Egress LER	The partner of the ingress LER that terminates an LSP and forwards IP data to the IP source. Note that the role a device plays can be different for different LSPs. The same device can be an LSR, ingress LER and egress LER for different LSPs.
Cross-connect	The term used to describe the connection in the hardware between an {ingress interface, ingress label} and {egress interface, egress label}.
Link	A physical connection between two nodes in the network. This may be an electrical connection or an optical fiber.
Protected Link	A protected link is a physical link with some form of redundancy built in so that data transfer is not disrupted by a failure of one of the components of the link. A protected link appears to the MPLS control plane as a single point of connection within the network. There are many link protection schemes, but a popular one uses SONET/SDH protocols on an optical fiber loop.
Parallel (Bundled) Links	There may more than one link between a pair of nodes in a network. Unlike a protected link, these individual links do appear as separate points of connection within the network. They may be managed as distinct entities providing different (but parallel) routes within the network, or they can be managed as a “bundle” where the choice of component link is only available to the nodes that are connected by the link.
Alternate Path	An alternate path is precisely that: a different route through the network to travel between the same to end points. Parallel links provide the simplest alternate paths. More complicated alternate paths will involve traversing distinct links and transiting other nodes. The preferred route is usually calculated using Shortest Path First (SPF) algorithms, or specified at the ingress after performing Traffic

	Engineering (TE) calculations. Alternate routes may often be longer or less desirable.
Controller Card	The internals of switches and routers are usually organized such that the main processor is present on a controller card. This card usually runs the main software in the system and is responsible for coordinating the other components.
Line Card	Line cards manage the ends of the links - known as ports or interfaces. One card may have multiple ports and so service multiple links. There would typically be many line cards in any one switch. Line cards can be dumb and do nothing more than provide the hardware to terminate the links. Smart line cards also include a processor that may run part or all of the protocol software that signals to set up LSPs on the links.
Backplane	The backplane is like a LAN within the switch. It provides connectivity between the controller cards and line cards.
Backup Card	Resilience against faults within a switch is achieved by having backup cards. There can be backup controller cards and backup line cards. A backup card will run a backup copy of the software on the primary card and takes over processing in the event that the hardware or software on the primary card fails. A single backup card may be dedicated to backup a specific primary card, or may be shared by several primaries.
Disjoint Paths	Two paths through the network are said to be disjoint if they do not share any links or nodes, other than the ingress and egress notes.
Link Disjoint Paths	Similar to Disjoint Paths, although Link Disjoint Paths can share nodes, provided that they do not share links.

2.3 Potential Resource Failures

Any of the resources within a network might fail. To provide a proper high availability network the network provider must predict and plan for any of these failures.

The traditional error is a link failure caused by a man with a shovel or someone digging with a backhoe. Similar failures are caused by far less dramatic problems such as a cable being knocked out of its socket, or being unplugged by mistake. In optical networks component failure is also possible resulting in Loss Of Light (LOL) on a link as a laser ceases to function.

Typical problems also include router failure where a whole LSR goes down. This might be caused by a power loss or by failure of a key, non-replicated component of the LSR. When a router fails all of the links to and from that router also fail.

Hardware component failure within an LSR may be survivable. Dumb line card failure is similar to link failure. It may, however, be recovered internally (and even transparently to the network) by having a backup line card on the same link. In this way, a backup line card can be seen as an element of a protected link. One way this is achieved is by primary and backup line cards both being connected to the same link, using an optical splitter. Only the primary card is used to transmit and receive data, but on failure, the backup card is already connected to the link.

The failure of a smart line card or of a controller card can result in loss of software state. Controller card failure is similar to router failure, while smart line card failure is similar to the failure of the router on the subset of the links that the line card supported.

Software failure on a card in the LSR has the same effect as the failure of the card itself.

2.4 Objectives for Failure Survival

The key objective of failure survival in an MPLS network is to minimize the disruption to data traffic of any failure.

Where possible, established LSPs (which may be carrying data) should not be disturbed at all while the failure is recovered. This means that links and cross-connects should stay in place, and data packets should not be discarded.

In practice, many failures will require some disruption as new resources take over from the old. This disturbance should, however, be kept as small as possible. A figure of 60ms is often quoted in the telecommunications world as the largest disruption to voice traffic that can be managed by the human brain before the effect becomes noticeable as a break that interrupts understanding or flow. This means that for voice traffic any failure should, ideally, be detected, reported and repaired within a total of 60ms.

Even if the repair of an LSP takes longer than 60ms it is still important that the connection is restored automatically. Consider a telephone user – ideally they will not notice the fault at all, however, it is still better to hear a few clicks on the line and have to say, "Pardon; could you repeat that?" than to lose the connection and have to re-dial.

If there is disruption to the data flow, an important consideration is whether data is lost and if so, how much. Neither IP networks nor other networks such as ATM or Frame Relay attempt to provide reliable delivery of data, other than by using higher layer end-to-end protocols such as TCP over the network protocols. However, if a substantial amount of data is lost, such protocols may declare the connection failed, and require re-connection.

A slightly lower priority aim is that the signaling service should remain available. That is, that it should continue to be possible to establish new connections for data traffic after the failure. It may be that new connections cannot be signaled while the failure is being repaired. Although this is undesirable, it is generally acceptable for a user to retry a connection attempt (e.g. redial at a telephone) if the connection fails to establish the first time. Given the statistical likelihood of a new connection being attempted during a failure repair, it is often considered acceptable that signaling is temporarily suspended.

The process of repair in one part of the network should, of course, cause as little disruption as possible to other parts of the network. Broadcasting failure information around the network could seriously disrupt other signaling and data traffic. It is worth noting that the typical requirement is to survive a single network failure. Many network providers and device vendors are not attempting to provide solutions that survive multiple concurrent network failures. While this does reduce complexity, it implies that the recovery time of failed equipment must be low to ensure that the period of vulnerability is as short as possible.

All of the solutions to these requirements involve forms of redundancy whether within links, as extra links in the network, or through the provision of additional hardware components within a switch. The cost of these solutions imposes an additional requirement that redundant resources should be kept to a minimum and preferably shared between potential users. Not many people keep a spare car in the garage at home in case their everyday car breaks down – they prefer to share the cost with other people by relying on taxis or rental cars in the event of a failure of their own vehicle.

2.5 Detecting Errors

Fundamental to repairing network errors is their detection at adjacent nodes, and at more remote repair points. The different survival processes described in this white paper rely on various detection and reporting mechanisms, so they are all introduced here.

Loss of electrical connectivity	Typically, failure of an electrical link (such as Ethernet) is detected by the line card and reported to the device driver and so to the signaling code. Link failures are usually reported at both the upstream and downstream ends of a link, because such links are bi-directional.
Loss of Light (LOL)	Optical link failure can be detected downstream of the fault through loss of light. This process relies on some form of electrical termination of the link at the downstream node. In Opto-electrical switches (OEOs) this is simple, but in LSRs with photonic cross-connects (PXCs) there may be no scope for LOL detection until the electrical termination at the egress node. A solution provided in some PXCs is an optical tap that removes a few percent of the signal and terminates it electronically for the purpose of verifying connectivity. Note that none of the methods of detecting LOL result in the upstream LSR detecting the fault, and additional mechanisms are required to report the fault upstream (see LMP, below).
Link Management Protocol	The IETF has defined a new protocol LMP [4] for use in networks, and especially PXC MPLS networks, to discover and monitor links. Using a bi-directional, out-of-band, control channel, LMP allows downstream nodes to communicate upstream until the fault is isolated, and reported to the local management components.
Link layer protocol report	If the underlying network runs its own protocols, then link failures may be detected and reported to both upstream and downstream nodes. Such protocols require link bi-directionality. An example of such a scenario is SONET, or more subtly the case where LMP is being used to manage uni-directional links: the LMP protocol runs over an out-of-band control channel that is bi-directional.
Management intervention	Links may be controlled through management applications, and those applications can also be used to report link failures to the system. Similarly, internal hardware components may be subject to

	operator management, for example to fail processing over to a backup card before the primary card is removed from the rack.
Hardware manager	Most switches built using distributed components have a “Hardware Manager” that is responsible for monitoring the state of the controller and line cards, and the software running on them. When there is a failure, the Hardware Manager reports the problem to control software that can instigate recovery procedures.
Signaling hellos, keepalives	Many signaling protocols include “keepalive” processing where adjacent nodes poll each other periodically to check that the link is active and that the signaling software on the partner is active. RSVP includes Hello messages and LDP uses KeepAlive messages. However, the need to have multiple retries to allow for occasional data loss, and the general speed of these mechanisms, may mean that they detect failure much slower than hardware or lower layer protocols. This form of protocol exchange is particularly useful for detecting software failure or hardware card failure at an adjacent node since the physical link itself may be undisturbed by these faults.
IGP hellos	Interior Gateway Protocols (IGPs) disseminate routing and topology information within a network. IGPs also run hello message exchanges within their protocols. Such exchanges are typically relatively infrequent since routing table updates do not need to be rapid. This means that rapid recovery of data paths cannot rely on IGP hellos for error detection.
IGP topology updates	IGPs collect and distribute topology information. These updates will reflect the current state of the network as known by the IGP implementations on the nodes throughout the network, so link failure information is propagated in this way. As mentioned above, the IGP may not detect errors very fast, and does not distribute the topology very often.
Signaling error notifications	MPLS signaling protocols include ways of reporting failures to set up LSPs and of notifying upstream nodes, including the ingress (initiator), when an established LSP fails. Although this does not provide hardware failure detection or notification, much can be inferred from an LSP failure. Additionally, since the whole purpose of failure survival is to

	preserve or re-establish LSPs, LSP error notifications play an important part in recovery processing.
Notify messages	Generalized MPLS (GMPLS [13]) introduces a new Notify message to the signaling protocols so that LSP failures can be reported to the ingress or some other node responsible for error recovery. Notify messages may provide faster error reporting than the normal error notifications since they can contain information about multiple failed LSPs, and because they are sent direct to the consumer. Note that this function is initially only specified for RSVP-TE signaling and not CR-LDP.
Crankback	Crankback is a process of providing additional information about hardware faults and broken topologies on signaling error or notification messages. Rather than simply reporting the underlying cause of the problem in an error message, such message would also carry crankback information identifying the failed link or node. This information can provide rapid feedback into topology and routing tables within the network and allows LSPs to be correctly set up around the failed resource.

2.6 Overview of Approaches to Failure Survival

The rest of this white paper describes specific ways to survive failures in MPLS networks. This section just introduces the broad approaches that can be employed.

The MPLS network may rely on the underlying physical resources to protect it against failures. Lower layer protocols may run over resilient links (such as SONET), in which case MPLS is unaware of the failures and continues to operate as if nothing had happened.

Most MPLS networks will run IP routing protocols to distribute network topology information. This allows the MPLS signaling protocols to become aware of changes to the topology and to re-signal LSPs routed around network failures.

For rapid recovery after a failure, MPLS resources may be double provisioned so that each LSP has an alternate LSP set up to use a distinct path. When an LSP error is notified to the ingress the data may be “Protection Switched” to the backup LSP without any further signaling.

Various “Fast Re-route” schemes exist to reduce the need to propagate failure information across multiple hops. LSP failures are mended at the point of detection and data is immediately re-directed onto pre-reserved backup resources.

Hardware and software failures within an individual LSR may be repaired using “Fault Tolerance” schemes that involve duplication of hardware and software components. Configuration and state information is replicated from the primary to backup components so that the backups are ready to take over if the primaries fail.

Looking through this list, we see that most survival mechanisms involve duplication of resources in some way. This is, of course, expensive for the network provider. Although these costs can be passed on to the user, who pays extra for the additional level of service, it is still desirable to reduce these costs as far as possible. The sections below not only describe the techniques for making an MPLS network capable of surviving failures, but also discuss possible approaches to reduce the costs by sharing the resources.

3. Protected Link Resources

Various link types have inherent redundancy built into them and can survive damage to the link without needing to report the problem to the LSRs that terminate the links. Data may be lost during the period immediately following the failure, but the link is self-repairing and data transfer is automatically resumed.

A SONET ring is an example of such a link. If part of the ring is damaged, data is automatically routed the other way around the ring.

SONET rings can be used to attach a pair of LSRs as a protected point-to-point link, or can connect more than two LSRs in what is effectively a protected multi-drop link.

Note that when a protected link is used, the MPLS components will not be advised of any failure. No switch reprogramming is required and no signaling takes place. MPLS networks providing particularly high levels of failure recovery service might choose to take a notification from the hardware to management components so that topology features can be updated. It is even possible to consider that the ingress might choose to re-route LSPs after such a failure so that the LSP runs entirely on protected links again.

When MPLS uses a link bundle, each LSR is responsible for choosing the component link to use between it and the downstream LSR. The required link can be encoded in the explicit route , but it is usually left entirely in the hands of the upstream LSR to choose on the basis of link availability and current distribution of LSPs. The choice of component link is signaled to the downstream LSR in the Interface_ID object as described in GMPLS [13].

If a component link from a bundle fails, the upstream LSR could choose to re-issue the LSP setup request to the downstream LSR indicating a different component link. Provided that the downstream LSR can handle this change of plan, this provides a simple and effective way of repairing the damaged LSP without recourse to re-routing or major re-signaling. This is effectively a form of local repair – see next section.

It is important to note that many link failures are caused by some catastrophic physical event such as a backhoe cutting through a cable or fiber. Link bundles often consist of a cluster of parallel fibers running close together. It is highly likely that if one link is severed then all of the links in the bundle will also be broken. Loop topologies are more robust as they usually use geographically distinct routes.

4. Local Repair

IP is a connectionless protocol designed to send data as datagrams through the network from source to destination, using the best available route. To provide the background for MPLS local repair, we look at raw IP networks first.

Routing protocols run within the network to collect a propagate topology information. This is processed using Shortest Path First (SPF) algorithms to produce a routing table at each node in the network that tells the node in which direction to forward an IP packet based on its destination address.

When there is a link or node failure within an IP network, the change in topology is distributed by the routing protocol and the routing tables are updated at each node. Initially, this may result in data packets being lost for one of three reasons.

- They are sent down the broken link because the local routing table hasn't been updated.
- They are discarded because no suitable route is known.
- They are dropped because the routing table shows the best route to be back in the direction from which the packet came.

Nevertheless, after a period of time (frequently measured in seconds), the routing protocol stabilizes and the routing tables either show that no route exists from source to destination (the network has fragmented) or a new route has been advertised and IP data flows again.

Some Service Providers achieve rapid healing of their networks and protection against failures simply by double-provisioning their entire network. Every link and every router has a shadow. Any single failure is rapidly repaired by the routing protocol to use the backup resource.

4.1 Local Repair in MPLS Networks

The communication of signaling information in MPLS uses IP. When a failure occurs in the network, an LSR upstream of the failure can attempt to re-signal the LSP. LSP signaling relies on IP routing, and therefore can take advantage of the fact that the routing table may be updated with new routes to the downstream nodes. Note, however, that this may take many seconds, and will not necessarily result in a new route being available.

Data is forwarded based on the {incoming interface, incoming label} to {outgoing interface, outgoing label} mappings – not information in the IP routing table. Therefore, updates to the IP routing table

do not affect the data flows. Data paths can only change once a new LSP has been signaled and devices on the LSP programmed with the new label mappings.

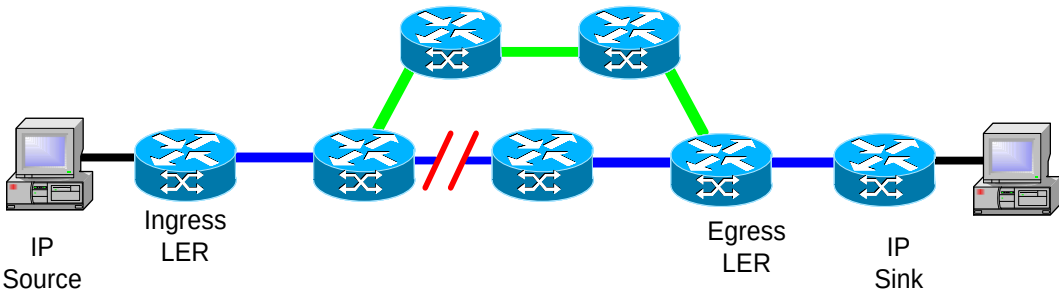

Figure 4: Re-routing around link failure

Figure 4 illustrates re-routing around a link failure in a simple network.

Because this re-signaling is time consuming and may in any case not result in successful re-establishment of the LSP, the signaling protocols impose some restrictions on the extent of local repair that is supported:

- CR-LDP does not include facilities for routing repair at the node that detects the fault. In fact, if a connection between two LSRs fails, CR-LDP mandates that the affected LSPs are torn down and an error notification is sent back to the ingress. An LSP can be re-signaled from the ingress and may merge with components of the old LSP downstream of the fault.
- RSVP-TE has its roots in RSVP (RFC2205 [8]), which was intended to keep resource allocations in line with IP microflows as they reacted to routing table changes. As an MPLS signaling protocol, RSVP-TE is more flexible than CR-LDP and allows LSPs to be re-routed according to changes to the IP topology. This re-routing is, however, usually restricted to the point of failure detection and the ingress – if each LSR on the path attempted to re-route and re-signal the LSP, but failed (e.g. due to inability to find a route that matched the requested constraints), it might take far too long for the error to finally propagate back to the ingress node.
- LDP is not a Traffic Engineering protocol and is more closely tied to the routing topology. In general, LDP will react to the new routes entered into the routing tables at each LSR by distributing new labels to allow label-based forwarding to operate.

Since network topologies are rarely full meshes, local repair might not succeed, and re-routing may need to be resolved at the ingress.

4.2 Traffic Engineering Implications

Traffic Engineered LSPs may have been established using constrained routes. In this case, it may be highly undesirable to re-route the LSP. Not only might this result in resources being used against the wishes of the Traffic Engineering (TE) application, but also since the re-route is local, the TE application may never find out that the resources have been taken by the redirected LSP. This can seriously break the TE model.

To counter this, TE LSPs can be "pinned" so that re-routing is not permitted once the LSP has been established. If there is a local failure, this will be propagated back to the ingress, where appropriate TE calculations can again be made.

4.3 Recovery Speed

It is usual to implement a delay within MPLS signaling protocols so that attempts to re-route LSPs are not made immediately the fault is reported to the signaling code. This gives the routing protocol time to react to the failure and to set up new entries in the routing table. This makes it harder to recover quickly, for example to hit the recovery goal of 60ms.

However, local repair relies on the speed of propagation of routing table updates. This can be slow (up to 30 seconds) which is unacceptable for many MPLS applications. Further, even if the routing table update is quick, this solution requires additional signaling at the time of failure which will further delay the restoration of a data path.

4.4 Crankback

Crankback is a process of reporting information about route failures back along the route towards the ingress. Recent drafts in the IETF [5] define extensions to the protocol messages that report LSP failures, to carry the details of which link has failed at which LSR.

Crankback can be used to augment the IGP link state databases, especially those that advise Traffic Engineering. Frequently, this will provide substantially faster feedback than routing protocol updates and can be used to help re-route the LSP, especially from the ingress. A crankback update could either be used to (a) re-route only the LSP to which it relates, or to (b) update the IGP link state databases, for routing all future LSP establishment requests and for affecting existing LSPs. The danger of using the crankback information for anything other than the LSP to which it relates is that there is currently no mechanism to tie together crankback information and other IGP link state information. Hence, it might be difficult to identify when information introduced from a crankback update should be discarded or updated due to IGP updates.

When an LSP crosses domain or area boundaries, crankback can be used to re-route the LSP from the domain boundary without needing to propagate the error all the way back to the ingress.

4.5 Return to Preferred Routes

When a broken link is restored, the routing protocol will advertise the link again and this will lead to the preferred (shorter) routes being restored in the routing tables.

It is an implementation choice whether LSPs should be re-routed to take advantage of the restored routes. There are obvious advantages to shortening the data path and to putting the resource reservations back on to the optimum path.

However, there are risks associated with changing the data path since the new path may not be stable. Further, as the repaired topology propagates through the network, multiple LSRs on the path of the LSP might decide to re-route the LSP back to the preferred route. This would lead to the MPLS equivalent of route flap, which should be avoided. This could be improved by only allowing the LSR that successfully signaled the backup path to re-route to the preferred path, although a route flap could still occur if the original failure was load dependent.

5. Protection Switching

5.1 Basic Operation

Protection Switching is a method of ensuring recovery from link or node failure with minimal disruption to the data traffic. Many references to this function include a target failover time of 60ms that is reputed to be the longest acceptable disruption to voice traffic.

In Protection Switching, data is switched from a failed LSP to a backup LSP at the repair point, which is not the point of failure and is conventionally the ingress, although may also be at other well-defined points on the LSP. The backup LSP is usually pre-provisioned.

In Figure 5 data is switched on the red and green primary LSPs. The blue backup LSP takes a less favorable path, is ready and set up, but does not carry any data. When an error in one of the primary LSPs is reported back to the ingress LER (perhaps using Notify messages, see later), data is immediately switched to the backup LSP. Note that the blue path shown could be a backup for both the red and the green paths simultaneously. See below for discussion of backup modes.

This can be considerably quicker than local repair since the backup LSP does not need to be signaled at the time of failure.

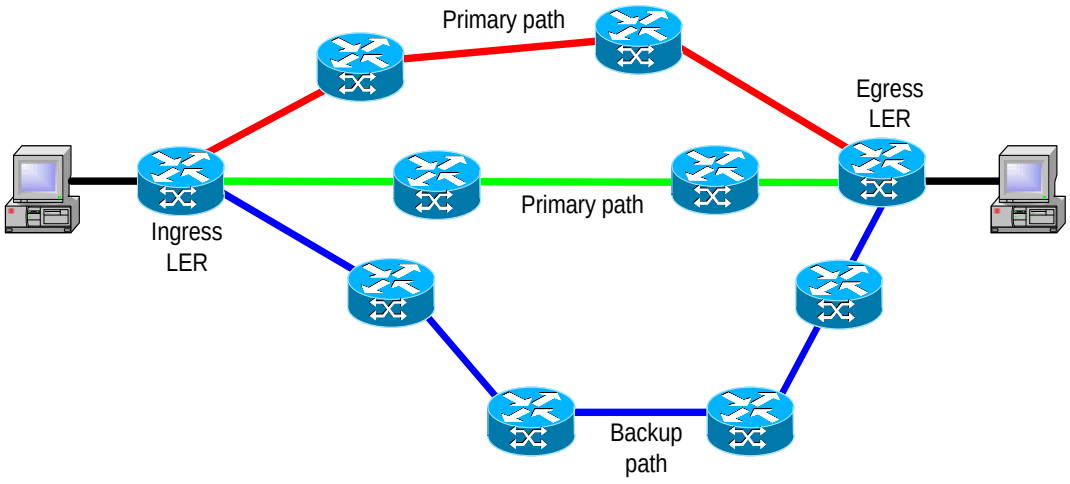

Figure 5: Protection switching

5.2 Backup Modes and Options

Backup LSPs can be pre-provisioned as in the description above.

However, it is also possible to consider a scenario where the backup LSP is configured in advance at the ingress, but is not signaled until the failure is reported. This has the advantage that network resources are not tied up by the backup LSP, but increases the failover time and is subject to the prospect of no resources being available when the backup is needed.

If the backup LSP is pre-provisioned, then there are several options.

- The backup LSP is ready for immediate use, the resources are fully dedicated and data is simultaneously sent down both the backup LSP and the primary LSP. Failover then only requires that the egress device read data from the backup LSP, rather than the primary. This changeover could be caused by a notification message coming from the point of failure or by the egress comparing the signal integrity on both primary and backup (e.g. in the case of LSPs in an optical network).

  This is referred to as 1+1 protection. It is the fastest protection switched recovery mechanism, but also the most expensive in terms of resources used.

- The backup LSP is idle and ready for immediate use but the data is not sent on the backup LSP. In this case, a notification to the ingress is required to switch the data flow. A notification could also flow to the egress, or it could simply be listening on both paths, ready to swap which path it receives from when it detects data on the backup.

  This is referred to as 1:1 protection (pronounced “one to one”). If this mechanism is used, the resources on the backup paths can be used as described in the next two options.

- The LSP is in use for low-priority traffic that can be thrown off when the primary fails. Such traffic may not even be MPLS traffic, but simply IP packets.

- The backup is pre-signaled and the resources have been reserved, but the resources are in use for other low priority LSPs that can be torn down when the backup LSP is needed for data. Note that these backup LSPs may not have to be literally torn down by signaling, but no data can be sent on them that may disturb the working of the backup path. Leaving the low priority LSPs in place reduces activity in the situation where the network will return to the primary path as soon as it is available.

Additionally, a pre-provisioned backup LSP (or the resources for that LSP) can provide support for more than one primary LSP so that it is used to carry data from the first LSP to fail. In Figure 5, the blue LSP provides a backup for the red and green LSPs. When one of the primary LSPs fails, data is switched to the backup and the other primary LSP becomes unprotected. This is often considered an

acceptable way of operating since the failure of a second primary LSP is assumed to be highly unlikely – in fact, after failover, the blue LSP is itself unprotected. This method reduces the cost of providing backup services.

5.3 Notifying Error Conditions

In order that the ingress point of an LSP can switch data from the failed primary to the backup LSP, the fault must be notified from the point of detection to the ingress.

The MPLS TE signaling protocols include such notification messages (PathErr in RSVP-TE and Withdraw in CR-LDP). These messages flow back upstream on the path of the LSP and are processed hop-by-hop. That is, they are sent using IP from one LSR to the LSR immediately upstream where they are passed to the signaling protocol code for processing.

In CR-LDP the Notification message causes the LSP to be torn down, so switch deprogramming is required. In base RSVP-TE, PathErr messages are simply forwarded upstream by the signaling code, with the expectation that a subsequent PathTear or ResvTear will be needed at each hop, but extensions to RSVP-TE in GMPLS [13] allow PathErr to remove state as well.

The notification messages eventually reach the ingress, which is able to switch over to the backup LSP.

GMPLS also introduces a new message for RSVP-TE to notify the LSP failure direct to the ingress. The Notify message has two performance advantages.

- It is addressed to a particular node on the LSP. It is not sent hop-by-hop and is therefore not processed at intervening LSRs. The ingress and egress LERs are the most likely devices to register for the Notify message, but others may as well. See the next section.
- The message is constructed to allow multiple failed LSPs to be indicated at once. This is important since a link failure is likely to affect very many LSPs and lots of these may need to be reported to a single ingress LER or other device requesting the Notify. Hence, the Notify message reduces the number of messages sent after a failure and therefore improves the speed of delivery of the error indication.

Note that some networks include data layer mechanisms to quickly detect failure and report failures. The OAM flows in ATM are an example of this, which provide fast error detection and notification. There are a number of drafts that have been put forward within the IETF discussing how similar flows could be used in MPLS [17, 18].

5.4 Alternate Repair Points

All the sections above are centered on protection switching repair at the ingress LSR. Most discussion of protection switching in the MPLS community has so far focused on this type of protection switching.

However, consideration is now being given to repair from other points. This would allow vulnerable parts of the network to be protected without allocating backup resources for the more robust links. It can also shorten the distance over which the errors must be notified resulting in shorter repair times.

Alternate repair points are also of interest in networks built up from autonomous systems or from rings of LSRs. In both cases, the topologies lend themselves to repairing broken LSPs within the network and not at the ingress.

Since the normal MPLS problem reporting is hop-by-hop, any upstream LSR on the LSP could choose to be a repair point and switch the data to a pre-established alternate LSP. However, in GMPLS for RSVP, a Path message specifies the desired recipient of a Notify message. As the Path progresses through the network, this request can be updated so that faults in different parts of the network can be reported by Notify messages sent to different repair points.

5.5 Recovery Speed

The main concern with protection switching is the speed of repair. The error must be detected and reported to the repair point. The backup LSP must be prepared, and finally the data must be transferred to the backup LSP.

Almost all of the options described above all require some amount of protocol signaling at the time of failover (the exception being 1+1 protection where the egress switches automatically to the backup path by detecting lack of data flow on the primary). This varies from simply propagating the error from the point of detection to the point of repair, to the full signaling of the backup LSP.

Obviously, the more signaling is required, the less likely the failover is to be timely. It is generally accepted that significant amounts of signaling (especially if more than one LSP has been broken by the failure) will not provide good enough failover times for most uses of LSPs. In fact, many people are suspicious that simply signaling the failure back to the repair point may compromise the failover.

The cost of providing backup facilities increases as the required speed of failover increases. Recovery systems that require substantial signaling at the time of failure also take less network resources and are therefore cheaper. Quick mechanisms require more dedicated resources and are therefore more expensive.

5.6 Sharing Resources

Another important concern with protection switching is that the resources must be pre-allocated to protect each primary LSP. This can be very expensive since resources used for the backups cannot be used to carry revenue-generating traffic.

Since the service level of a protected LSP is higher than that of a normal LSP, service providers can, of course, pass the costs on to their customers as increased charges. This creates two levels of service:

- bronze – no protection
- gold – protection switched.

Dividing the service levels in this way will reduce the pressure on resources, as not all customers will want to pay for protection switching. Nevertheless, service providers will want to search out ways of minimizing the need to reserve backup resources. If this creates an intermediate silver service level of "protected most of the time", this will be acceptable.

Several modes of operation have already been raised. The simplest has a single LSP providing the backup for more than one primary LSP. Since it is unlikely that both primaries will fail, this offers a good solution, but it does require that there is more than one primary LSP between ingress and egress – something that may often not be the case.

Another option that works when there are multiple data flows between ingress and egress is to use the backup LSP for low priority data. When the primary fails, the low priority data is dropped or reverts to best effort IP transfer. This is a nice model if MPLS is being used to handle DiffServ traffic as described in [12].

In a complex network, the issue may be wider than reducing the number of end-to-end backup LSPs. In this case, there is a need to reduce the amount of resources used on links in the core of the network.

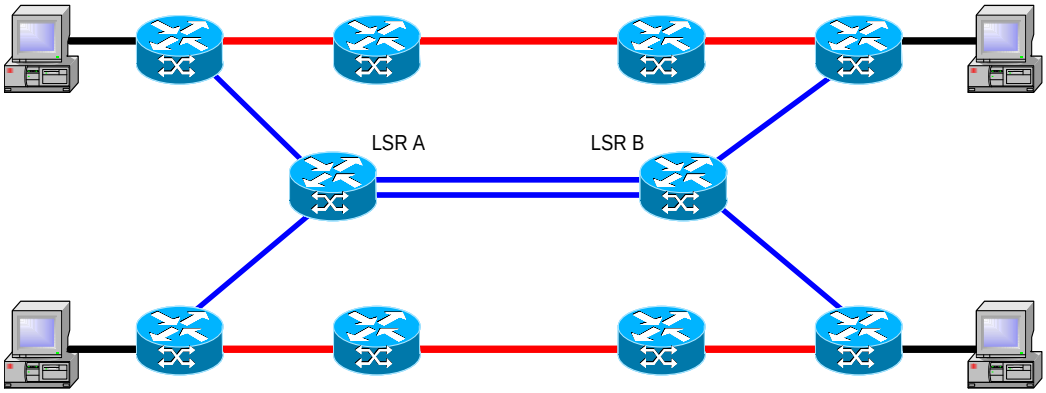

Figure 6: Protection switching

Figure 6 shows two entirely distinct primary LSPs in red. Protection switched backups (in blue) are signaled through the network, and for part of their routes they are coincident (between LSRs A and B). Now, the resource reservation load between LSRs A and B due to the backup LSPs could be very high, yet no data is actually passing down this route. At best this may impair the ability of the network to set up primary LSPs between A and B, and at worst it may mean that one of the backups cannot be established.

The ideal is for the resources of between A and B to be shared between the two backup LSPs. Since it is highly unlikely that both primary LSPs will fail at the same time, this is a good solution for a silver service.

Several issues with this resource sharing approach are still open for study at the time of writing this white paper.

- Firstly, how do LSRs A and B know that it is acceptable to share resources between the two LSPs? One possibility would be to mark the signaling requests as "backups". There are drafts proposed within the IETF [7, 19] that describe ways this can be done, for example by extending the Protection Information object defined in GMPLS [13].
- The second question is how should LSR A behave if both primary LSPs *do* fail and data starts to flow on both backup LSPs. A probable answer is that the backups are treated as first-come first-served so that the data on the second backup to be used is simply dropped at LSR A. This is hardly satisfactory, however, if both primary LSPs believe they are protected, and a better answer involves signaling to the ingress of the second primary that it is no longer protected. One way of doing this is discussed in [19] and relies on use of the Notify message in an RSVP network.
- This leads to a third question, which is how restoration of a failed primary can be achieved without disrupting data flow. One way of doing this is described as "bridge and roll",

- Lastly, care needs to be given to the process of data forwarding at LSR B. It is important that, however resource sharing is arranged between LSRs A and B, there are still distinct backup LSPs running in parallel. If this is not done, then packets arriving at LSR B will be indistinguishable and LSR B will not know to which egress LSR they should be forwarded.
- This system can lead to another level of complexity where links and nodes can have both primary and backup resources. Primary resources can be committed only once, but backup resources could be over-committed many times, leading to two separate resource spaces to be managed. This is something that the IGPs do not currently support.

5.7 Status of protection switching within IETF

There are many drafts proposed within the IETF that discuss the requirements and pitfalls of protection switching within different types of network (for example, packet networks and optical networks). Some of these drafts propose signaling enhancements, whilst some concentrate on providing solutions using existing MPLS or GMPLS signaling.

At time of writing there is no clear consensus over which drafts should be pushed forward, although it is expected that the MPLS working groups within the IETF will identify their preferred solutions soon.

6. Fast Re-route

The previous section highlights some of the concerns with protection switching. Errors need to be signaled through the network from the point of detection to the point of repair and this can significantly delay the repair time.

Fast re-route is a process where MPLS data can be directed around a link failure without the need to perform any signaling at the time that the failure is detected. Unlike protection switching, the repair point *is* the point of failure detection. Consequently there is no requirement to propagate the error to the repair point using the signaling protocol.

Most fast re-route protection schemes rely on pre-signaled backup resources. When the failure is reported to the repair point, it simply updates the programming of its switch so that data that was previously sent out of one interface with one label is sent out of a different interface with another label.

There are several fast re-route schemes currently under discussion in the IETF. The different approaches address different problems and vary in complexity. Some of the more established solutions are set out below.

6.1 Link Protection

The simplest form of fast re-route is called Link Protection. An LSP tunnel is set up through the network to provide a backup for a vulnerable physical link. The LSP provides a parallel *virtual* link.

When the link fails, the upstream node switches traffic from the physical link to the virtual link so that data continues to flow with a minimal disruption.

Figure 7 shows a tunnel (red) that has been set up to protect the link between LSRs A and B. When the link fails, the blue LSP is redirected down the red tunnel so that the data still flows from A to B.

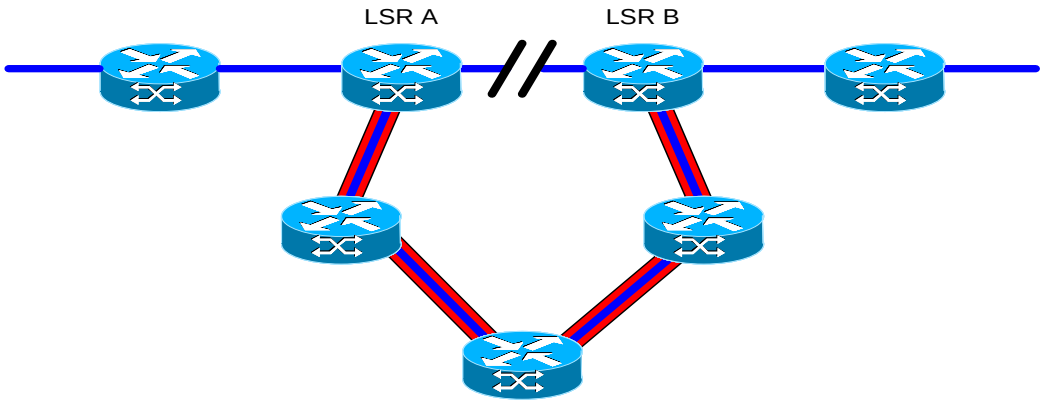

Figure 7: Fast re-route link protection

The capacity of the backup LSP should, of course, be sufficient to carry the protected LSPs. If all LSPs on a link are to be protected then the capacity should equal the bandwidth of the protected link. This can potentially lead to a huge amount of backup bandwidth being required, especially if multiple links must be protected in this way. Note that not all LSPs using a link need to be protected by the same backup LSP, or even at all. By leaving some LSPs over the link unprotected, the backup bandwidth requirement can be reduced.

Note that there are some very specific limits placed on the use of label spaces when this method of fast re-route is in use. The LSP that provides the backup virtual link is used as an LSP tunnel. That is, the data packets that would have been sent down the physical link have an additional label added and that top label is used to forward the packet along the backup LSP. When the other side of the broken link is reached (the egress of the backup LSP) the top label is stripped from the packet, and the data is forwarded according to the lower label – this is the label that would have been used to forward the packet down the broken physical link.

However, label switches provide a mapping from {ingress interface, ingress label} to {egress interface, egress label} and, although the ingress label has been preserved, the ingress interface will have changed. It will be reported either as the virtual interface that identifies the egress of the backup tunnel, or as the physical interface through which the backup tunnel arrives. There are two solutions.

- Map the virtual interface back to the original ingress physical interface. This may be possible, but it requires that the downstream node understands the reason why the backup LSP was established. This would probably involve configuration intervention.
- Use the Global Label Space for all LSPs 'protected' by the backup tunnel. This is easily achieved.

The main issues with link protection concern the increased complexity of configuration (each protected link must have a backup tunnel configured) and the amount of resources that must be reserved in the network.

6.2 Node Protection

Link protection only handles the case where a single link between two LSRs has failed. However, it is also possible that an entire LSR will fail.

Figure 8 shows a green tunnel running from LSR A to LSR C that protects against the failure of LSR B. When the failure is detected at LSR A, the blue LSP is re-routed down the tunnel and data continues to flow.

Label stacking is also used in this model, but the problems are increased because the ends of the protection tunnel are not adjacent LSRs. A packet sent down the green LSP from LSR A carries a higher-level label for navigating the protection LSP and a lower level label relating to the original blue LSP. However, the lower level label provided the switching information for LSR B. When the top label for the protection tunnel is removed at its egress (LSR C), the underlying label for the protected LSP will be unknown.

A proposed solution to this issue is discussed in an IETF draft [6]. Using relatively recent additions to RSVP-TE, the initial set up of the protected LSP reports the labels in use on each link as part of the Record Route object. This object contains a list of LSR Ids and labels describing each hop. It is passed upstream during LSP establishment (on the Resv message) so that every node on the path knows the labels used on every link.

Once this information has been passed back upstream, each LSR can determine the correct labels to use in the label stack when it re-routes an LSP after failure.

For example, in Figure 8 the label progression on the original LSP through LSRs A, B and C is 5, 23, 19, 7. The backup tunnel has labels 83, 12, 42. Using link protection would cause packets sent from LSR A to LSR D to be labeled (83, 23) which would mean that when the packets were received at LSR C the top label would be removed exposing a label of 23 which is unknown at LSR C.

However, if the label values are reported back in the Record Route, LSR A can know the correct label value to use (19) so that packets emerge from the protection tunnel at LSR C carrying the same label that they would have used if they had come on the original LSP.

The remaining question concerns the choice of label space. As with link protection, the simplest approach is to use the global label space.

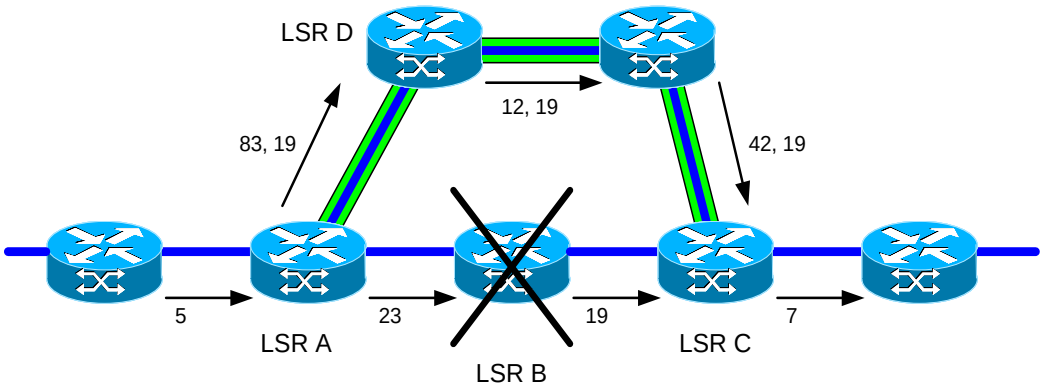

Figure 8: Fast re-route node protection

As with fast-reroute link protection, not all LSPs passing through a particular node need to be protected. In order to reduce the bandwidth required on the backup tunnel one may choose to leave low priority traffic unprotected – in this case the LSR at the start of the backup path must decide which LSPs to switch over to the backup tunnel and which to leave (with the consequence that data on the unprotected LSPs will be lost).

6.3 Generalizing node and link protection

The above description of node and link protection leads naturally to a generalization, whereby one can protect any vulnerable segment of a network. Hence, a protection LSP could be used to route around a number of nodes and links. This may be useful in protecting an autonomous system (AS) that is prone to failure by setting up a protection LSP through another system. This could require huge resources, as the protection LSP potentially has to match the resources of the protected AS, but may be reasonable if only some LSPs through the AS need to be protected.

6.4 Automatic Protection using Detours

Another scheme for fast-reroute within an RSVP network is described in [20].

This provides a mechanism for establishing detour paths on a per-LSP basis that can route data around downstream link and node failures. In order to simplify configuration requirements, the detour paths are set up automatically. Moreover they can adapt to the latest network topology without manual intervention.

In figure 9, the primary path, shown in blue, is protected by four separate detours in green. The detours are set up such that any single node or link failure on the primary path can be avoided. For example, if there is a failure on LSR B, then data can flow on detour path 1. A failure on the primary link between LSR D and the Egress can be avoided by taking detour path 4.

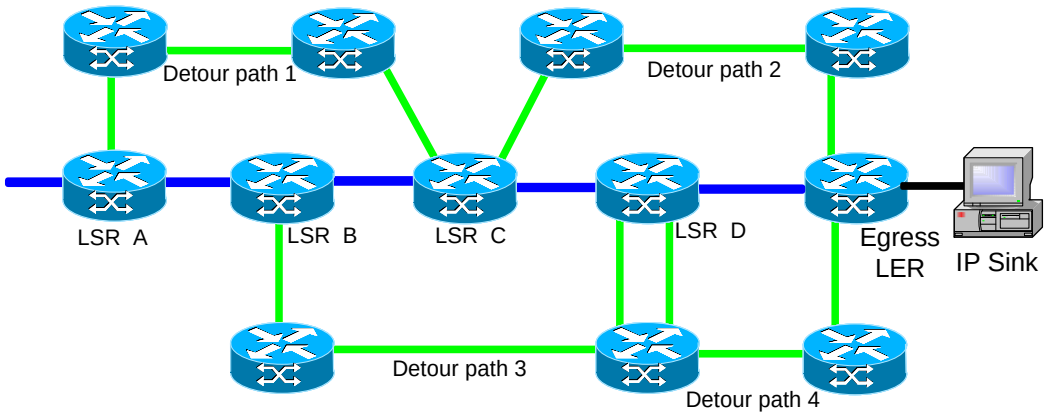

Figure 9: Fast re-route protection with Detours

The draft provides new RSVP Signaling messages that are used to request that detour paths are set up once the primary path is set-up. LSRs supporting this function can then automatically initiate computation of a detour path that protects against the next downstream node and link in the primary path. Note that LSRs adjacent to the Egress LSR can only compute a detour that protects against the link between itself and the egress.

In order to compute the detour paths, the LSR needs to know

- which downstream nodes the primary path goes through
- which outgoing link the primary path uses on the LSR
- which downstream nodes are to be protected against
- the traffic engineering requirement for the detour.

The first three items can be determined by using recorded route information, which is available during LSP establishment (i.e. on the Resv in RSVP). The traffic engineering requirements can be signaled on path creation within the new fast-reroute object. This allows detour paths to have different requirements to the main path. This might be useful in order to reduce backup bandwidth requirements – if any failure is expected to be short lived, it may be appropriate to reduce the bandwidth requirement over the backup path on the premise that a lower quality of service is better than no service.

Given this information the LSR determines the destination for the detour path as either the next but one LSR in the main path or, if the LSR is the penultimate hop, the egress. It then computes a route to the destination satisfying the following constraints.

- The detour path must originate from the current LSR.
- It should not traverse the immediate outgoing link.

- It should not traverse the next hop downstream node (unless this is the egress).
- It should satisfy the traffic engineering requirements specified.

The results of the detour computation allow the LSR to generate a detour path request. This will be an RSVP Path request containing an explicit route following the computed path, the traffic parameters specified within the original fast-reroute object and a new detour object that identifies the path as being a detour, providing an indication of the node at the start of the detour and the node that the detour is avoiding (if appropriate).

An LSR can recalculate its detour paths in order to take advantage of any favorable changes in network topology. If a subsequent calculation produces a different detour path to the one currently set up, the LSR can replace it.

The main feature of this approach is the fact that detour paths are set up dynamically without any operator intervention. However, as it stands, the current draft has a number of deficiencies.

- There is no discussion on label allocation for the detour LSPs. As described in section 6.2, some thought must be given to label allocation in order to ensure that data passed over the detour path is correctly forwarded upon arrival at the LSR at the end of the detour.
- The methods described are only suitable for unidirectional LSPs.
- There is no discussion regarding reuse of the bandwidth provisioned for the detour paths when there are no failures.

6.5 Current status of fast-reroute protection within IETF

The drafts described above [6, 20] present two methods of achieving fundamentally the same goals, that of enabling fast re-routing without any signaling overhead at time of failure. A further draft [21] discusses ways in which these methods can interoperate within a single network.

Ideally, though, the best features of each of these drafts would be combined to produce a single fast-reroute solution. The IETF has agreed that this approach should be investigated and a combined fast-reroute draft is expected soon.

7. Comparison of LSP Protection Techniques

The following table provides a summary of some of the main features of the LSP protection techniques described in the previous sections. It focuses on the main concerns:

- the amount of resource that must be pre-allocated
- the speed of restoration
- the increased complexity of configuration and signaling
- the change in the length of data paths.

Repair method	Resource Requirements	Speed of Repair	Configuration Complexity	Complexity of Signaling	Data Path Length
Local Repair	No pre-allocation. Repaired LSP uses same amount of resource.	Slow. Depends on routing table updates and additional signaling.	No additional configuration.	No change from existing signaling.	Repaired path might not be shortest available.
Re-route at ingress	As above.	As above with addition of error notification to the ingress.	No additional configuration.	No change from existing signaling.	Repaired path is shortest available.
Protection Switching	Backup LSP is pre-allocated, but may be shared among several primaries.	Speed may be limited by propagation of error to repair point.	Backup LSP must be configured at ingress.	No change in signaling technique, but two LSPs are signaled.	Protection path is chosen through configuration .
Fast Re-route Link Protection	Link backup pre-allocated. Need a backup for each protected link.	Rapid repair as soon as error is detected.	Link backups need to be configured.	No change, but may be limited to the global label space.	Data path extended by length of backup tunnel.

Fast Re-route Node Protection	Node backup pre-allocated. Need a backup for each protected node.	Rapid repair as soon as error is detected.	Node backups need to be configured.	Requires reporting of labels in Recorded Route. May be limited to the global label space.	Data path extended by length of backup tunnel.
Automatic Fast Re-route using Detour Paths	Detour paths pre-allocated. Need a detour path for each protected link and node.	Rapid repair as soon as error is detected.	Detour paths automatically set-up without additional configuration.	Requires signaling to request fast-reroute and distinguish detour path.	Data path extended by length of detour paths.

8. Summary

Building MPLS systems that can survive network failures is not simple.

This is partly due to the fact that MPLS is built on top of IP, which has less demanding recovery requirements, and its own ways of resolving routing changes. It also owes something to the origins of MPLS, which was not originally designed with much attention to rapid recovery from failures.

Protecting LSPs against network failure can use the inherent properties of IP and IP routing, the techniques of protection switching derived from experience in other signaling protocols, or fast re-route methods developed specifically for MPLS. Each has its own specific characteristics, advantages and disadvantages so that the method(s) used must be chosen carefully with regard to the requirements of the network and the users.

If very rapid repair is needed, for example for voice traffic, then fast re-route will probably provide the best solution. If quick repair with the possibility of sharing backup resources is desired then protection switching can be chosen. If repair time is not crucial and network resources are limited then local repair can be used.

9. Glossary

AS	Autonomous System. A part of the network under a single administration and usually running a single routing protocol for internal routing.
BGP	Border Gateway Protocol. The Exterior Gateway Protocol used for distributing routes over the Internet backbone.
CR-LDP	Constraint-based Routed Label Distribution Protocol. Extensions to LDP to set up Traffic Engineered LSPs, as defined in the Internet Draft "Constraint-based LSP Setup using LDP" [10].
DiffServ	Differentiated Services. A system of differentiating data packets for IP networks that is based on setting relative priorities and drop precedence for each DSCP. It is defined by the DiffServ Working Group.
DLCI	Data Link Circuit Identifier. The labels used in Frame Relay that are equivalent to MPLS labels.
ECMP	Equal Cost Multi-Path. A scheme where more than one path of the same "cost" can exist in the routing table. The choice between routes is made using some other principle such as bandwidth requirements.
FEC	Forwarding Equivalence Class. A logical aggregation of traffic that is forwarded in the same way by an LSR. A FEC can represent any aggregation that is convenient for the SP. FECs may be based on such things as destination address and VPN Id.
FT	Fault Tolerance. A scheme where an piece of hardware, such as an LSR, is built using duplicate hardware and software components such that the whole is resilient to failures of individual components and can provide a highly available system.
GMPLS	Generalized MPLS [13, 14, 15]. Extensions to the MPLS Traffic Engineering signaling protocols to support additional features such as optical networks, bi-directional LSPs, and source control of labels and link.

HA	High Availability. High Availability (HA) is typically claimed by equipment vendors when their hardware achieves availability levels of at least 99.999% (five 9s).
IETF	Internet Engineering Task Force. The worldwide grouping of individuals from the computing and networking industry and from academia that devises and standardizes protocols for communications within the Internet. Responsible for the development of MPLS.
IGP	Interior Gateway Protocol. Any routing protocol used for distributing routes within a single Autonomous System such as OSPF.
IP	Internet Protocol. A connectionless, packet-based protocol developed by the IETF and at the root of communications within the Internet.
Labels RSVP	See RSVP-TE.
LDP	Label Distribution Protocol. A protocol defined [11] by the IETF MPLS working group for distributing labels to set up MPLS LSPs.
LER	Label Edge Router. An LSR at the edge of the MPLS network. LERs typically form the ingress and egress points of LSP tunnels.
LMP	Link Management Protocol. A protocol under development by the IETF to discover and manage links, and to detect and isolate link failures.
LOL	Loss Of Light. The process of detecting the failure of an optical link by discovering that no signal is being received.
LSP	Label Switched Path. A data forwarding path determined by labels attached to each data packet where the data is forwarded at each hop according to the value of the labels.
LSP Tunnel	A Traffic Engineered LSP capable of carrying multiple data flows.
LSR	Label Switching Router. A component of an MPLS network that forwards data based on the labels associated with each data packet.
MPLS	MultiProtocol Label Switching. A standardized technology that provides connection-oriented switching based on IP routing protocols and labeling of data packets.

OEO	Opto-Electronic Switch. Short for Optical-Electronic-Optical, this switch terminates each optical connection converting the signal to electronics before forwarding packets to other optical links. Compare with PXC.
OSPF	Open Shortest Path First. A common routing protocol that provides IGP function.
PPP	Point-to-Point Protocol. A common access protocol for VPNs particularly important in providing connection from roaming workstations.
PXC	Photonic Cross-Connect. A type of switch that is capable of switching and forwarding optical signals without needing to convert the signals to electronics. Compare with OEO.
RSVP	Resource ReSerVation Protocol (RFC 2205) [8]. A setup protocol designed to reserve resources in an Integrated Services Internet. RSVP has been extended to form Labels RSVP.
RSVP-TE	Extensions to RSVP to set up Traffic Engineered LSPs [9]. Throughout this document, Labels RSVP or RSVP-TE is referred to simply as "RSVP".
SPF	Shortest Path First. An algorithm for selecting a route through a physical topology. "Shortest" may apply to the number of hops (i.e. nodes) in the route, but some hops may be weighted to reflect other "length" characteristics such as the absolute length of the physical link between nodes. See OSPF.
TCP	Transmission Control Protocol. A transport level protocol developed by the IETF for reliable data transfer over IP.
TE	Traffic Engineering. The process of balancing the load on a network by applying constraints to the routes which individual data flows may take.
VPI/VCI	Virtual Path Identifier / Virtual Channel Identifier. The labels used in ATM layer 2 networks that are equivalent to MPLS labels.
VPN	Virtual Private Network. A private network provided by securely sharing resources within a wider, common network.

10. References

The following documents are referenced within this white paper. All RFCs and Internet drafts are available from www.ietf.org URLs are provided for other references. Note that all Internet drafts are "work in progress" and may be subject to change, or may be withdrawn, without notice.

1	White paper from Metaswitch (www.metaswitch.com)	MPLS Traffic Engineering: A choice of Signaling Protocols
2	White paper from Metaswitch (www.metaswitch.com)	MPLS Virtual Private Networks: A review of the implementation options for MPLS VPNs including the ongoing standardization work in the IETF MPLS Working Group
3	RFC 3031	Multiprotocol Label Switching Architecture
4	draft-ietf-mpls-lmp	Link Management Protocol (LMP)
5	draft-iwata-mpls-crankback	Crankback Routing Extensions for MPLS Signaling
6	draft-swallow-rsvp-bypass-label	RSVP Label Allocation for Backup Tunnels
7	draft-li-shared-mesh-restoration	RSVP-TE Extensions For Shared-Mesh Restoration in Transport Networks
8	RFC 2205	Resource ReSerVation Protocol (RSVP)
9	draft-ietf-mpls-rsvp-lsp-tunnel	Extensions to RSVP for LSP Tunnels
10	draft-ietf-mpls-cr-ldp	Constraint-based Routed LSP Setup Using LDP
11	RFC 3036	LDP specification
12	draft-ietf-mpls-diff-ext	MPLS Support of Differentiated Services
13	draft-ietf-mpls-generalized-signaling	Generalized MPLS Signaling
14	draft-ietf-mpls-generalized-rsvp-te	Generalized MPLS Signaling extensions for RSVP-TE
15	draft-ietf-mpls-generalized-cr-ldp	Generalized MPLS Signaling extensions for CR-LDP

16	draft-ietf-rsvp-refresh-reduct	RSVP Refresh Overhead Reduction Extensions
17	draft-harrison-mpls-oam	OAM Functionality for MPLS Networks
18	draft-allan-mpls-oam-frmwk	A Framework for MPLS User Plane OAM
19	draft-lang-ccamp-recovery	Generalized MPLS Recovery Mechanisms
20	draft-gan-fast-reroute	A Method for MPLS LSP Fast-Reroute Using RSVP Detours
21	draft-atlas-rsvp-local-protect-interop	MPLS RSVP-TE Interoperability for Local Protection/Fast Reroute

In addition, the following is a number of other Internet drafts for MPLS which may be of interest.

draft-ietf-mpls-recovery-frmwrk	Framework for MPLS-based Recovery
draft-chang-mpls-path-protection	A Path Protection/Restoration Mechanism for MPLS Networks
draft-chang-mpls-rsvpte-path-protection-ext	Extensions to RSVP-TE for MPLS Path Protection
draft-owens-crldp-path-protection-ext	Extensions to CR-LDP for MPLS Path Protection
draft-shew-lsp-restoration	Fast Restoration of MPLS Label Switched Paths
draft-kini-restoration-shared-backup	Shared backup Label Switched Path restoration
draft-suraev-mpls-globl-recov-enhm	Global path recovery enhancement using Notify reverse LSP
draft-azad-mpls-oam-messaging	MPLS user-plane OAM messaging
draft-harrison-mpls-oam-req	Requirements for OAM in MPLS Networks
draft-many-optical-restoration	Restoration Mechanisms and Signaling in Optical Networks

11. About Metaswitch

Metaswitch is a privately owned technology company based in London, UK. We have US offices in Alameda, CA, Reston, VA, and Boxborough, MA.

Our Network Protocols Division is the leading developer and supplier of (G)MPLS, OSPF(-TE), ISIS(-TE), BGP, VPN, RIP, PIM, IGMP, MLD, ATM, MGCP, Megaco, SCTP, SIP, VoIP Conferencing, Messaging, Directory and SNA portable products. Customers include Alcatel, Cisco, Fujitsu, Hewlett-Packard, Hitachi, IBM Corp., Microsoft, Nortel and Sun.

Our company culture focuses on building software of consistently high quality, developed and supported by engineers who are with Metaswitch for the long term.

- Founded in 1981, we have over 450 employees, of whom 280 are engineers. The average length of service of engineers at Metaswitch is 8 years, and the annual attrition rate is 3%.
- Throughout this period, Metaswitch has been consistently profitable with profits exceeding 15% of revenue. 2007-2008 revenues were $118m with $22m profit.
- Over 90% of revenue is generated from exports and 80% is from customers in the US (so we are very used to working with American companies).
- The company is privately held by top-tier investment firms Francisco Partners and Sequoia Capital, as well as the Employee Benefit Trust (EBT). As part of this ownership structure, Metaswitch distributes a share of profit to all employees, equitably rewarding them for their contribution and encouraging long-term commitment.
- As a private company with an emphasis on long-term stability, we are not driven by the short-term requirements of quarterly profit statements. This means that we can concentrate on providing software as we would like – that is, developing high quality implementations of complex technologies.

The DC-MPLS product family provides OEMs with a flexible source code solution with the same high quality architecture and support for which Metaswitch's other communications software products are renowned. It runs within Metaswitch's existing high performance portable execution environment (the N-BASE). This provides extensive scalability and flexibility by enabling distribution of protocol components across a wide range of hardware configurations from DSPs to line cards to specialized signaling processors. It has fault tolerance designed in from the start, providing hot swap on failure or upgrade of hardware or software.

DC-MPLS is suitable for use in a wide range of IP switching and routing devices including Label Switch Routers (LSRs) and Label Edge Routers (LERs). Support is provided for a range of label distribution methods including Resource ReSerVation Protocol (RSVP), Constraint-based Routed Label Distribution Protocol (CR-LDP) and Label Distribution Protocol (LDP). The rich feature set gives DC-MPLS the performance, scalability and reliability required for the most demanding MPLS applications, including VPN solutions for massively scalable access devices.

DC-MPLS integrates seamlessly with Metaswitch's other protocol products, and uses the same proven N-BASE communications execution environment. The N-BASE has been ported to a large number of operating systems including VxWorks, Linux, OSE, pSOS, Chorus, Nucleus, Solaris and Windows NT, and has been used on many processors including x86, i960, Motorola 860, Sparc, IDT and MIPS.

All of the Metaswitch protocol implementations are built with scalability, distribution across multiple processors and fault tolerance architected in from the beginning. We have developed extremely consistent development processes that result in on-time delivery of highly robust and efficient software. This is backed up by an exceptionally responsive and expert support service, staffed by engineers with direct experience in developing the protocol solutions.

About the authors

Ed Harrison is a Development Manager for the DC-MPLS product family, and is a co-author of the GMPLS MIBs draft in the IETF.

Adrian Farrel was originally Architect and Development Manager for the DC-MPLS product family, and contributed to the GMPLS drafts in the IETF. He is now a Member of Technical Staff with Movaz Networks Inc.

Ben Miller is General Manager of Metaswitch's MPLS Group.

Metaswitch Networks
100 Church Street
Enfield
EN2 6BQ
England
+44 20 8366 1177
http://www.metaswitch.com