RSVP-TE Point-to-Multipoint

An overview of the RSVP-TE P2MP extensions, their application to IPTV, and support for other (G)MPLS features

Nic Neate, nic.neate@metaswitch.com

First issued March 2009
The rise of broadcast IPTV has created requirements for efficient and reliable multicast transport, with guaranteed Quality of Service, through service provider core networks. RSVP-TE point-to-multipoint (P2MP) LSPs are a good functional fit to these requirements, and this emerging market presents a strong case for their deployment.

- The RSVP-TE P2MP extensions allow efficient multicast distribution, with traffic duplication occurring as close to the receivers as possible.
- Traffic Engineering support means that dedicated network bandwidth is reserved. Further, the P2MP LSPs can be explicitly routed for optimal usage of resources.
- Mechanisms for advanced function such as resilience to network failures are being developed.

However, there are competing technologies. IP multicast, for example, meets many of these requirements. It also benefits from being more widely deployed than multicast MPLS, and it too is being extended to provide some level of resilience.

While the protocol definition for P2MP support in RSVP-TE provides a broad range of function, it is also highly complex. This complexity may hinder interoperability. High quality, flexible and robust implementations will be required.

Further, there are still gaps in the standards which will need to be filled before fully interoperable implementations of some function, such as protection and restoration of data streams using redundant P2MP LSPs, are possible.
# Table of contents

1. Introduction ............................................................................................................................1
   1.1 What is RSVP-TE point-to-multipoint? ........................................................................ 2
   1.2 Protocol standardization .............................................................................................. 3

2. Requirements from IPTV ..........................................................................................................5
   2.1 End-to-end IPTV network architecture ........................................................................ 5
   2.2 Multicast transport requirements ................................................................................ 7
       2.2.1 High availability ................................................................................................. 7
       2.2.2 Quality of Service ............................................................................................. 8
       2.2.3 Fast channel change ........................................................................................ 8
       2.2.4 Efficiency ......................................................................................................... 9
       2.2.5 Manageability .................................................................................................. 10
       2.2.6 Summary ......................................................................................................... 11
   2.3 Transport network architecture ................................................................................... 11
       2.3.1 VPLS multicast ............................................................................................... 12
       2.3.2 L3VPN multicast ............................................................................................ 13

3. RSVP-TE Point-to-Multipoint LSPs ...........................................................................................14
   3.1 Overview ..................................................................................................................... 14
       3.1.1 Explicit routing ................................................................................................. 14
       3.1.2 Tree modifications and repairs ......................................................................... 14
       3.1.3 Back-compatibility ......................................................................................... 15
   3.2 Protocol mechanisms .................................................................................................... 16
       3.2.1 Source-to-Leaf (S2L) sub-LSPs ....................................................................... 16
       3.2.2 Sub-groups ..................................................................................................... 17
   3.3 Grafting and pruning ..................................................................................................... 19
       3.3.1 Grafting ......................................................................................................... 19
       3.3.2 Pruning ......................................................................................................... 19
   3.4 Re-optimization ........................................................................................................... 19
       3.4.1 Make-before-break ......................................................................................... 19
       3.4.2 Sub-group-based sub-tree re-optimization ...................................................... 20
       3.4.3 Dynamic reroute ............................................................................................. 20
   3.5 Re-merging ................................................................................................................... 20

4. Routing and Management ..................................................................................................... 22
   4.1 Routing ....................................................................................................................... 22
       4.1.1 Leaf discovery ................................................................................................. 22
       4.1.2 IGP extensions for P2MP capability information .......................................... 23
       4.1.3 Tree computation ............................................................................................ 23
       4.1.4 Path Computation Elements (PCEs) ............................................................... 24
   4.2 Configuration ............................................................................................................... 24
1. Introduction

This white paper explains the point-to-multipoint (P2MP) extensions to RSVP-TE and the applications of, and requirements on, P2MP TE LSPs.

- The remainder of chapter 1 provides a high level introduction, introducing some key terms used throughout the document and summarizing the current state of the protocol standards in this area.

- Chapter 2 discusses the requirements for reliable and efficient transport of broadcast video coming from the burgeoning IPTV industry, currently the key motivator behind the deployment of P2MP TE LSPs.

- Chapter 3 is a high level functional description of the base protocol as defined in RFC 4875.

- Chapter 4 discusses the key additional pieces of routing and management function needed to provide a complete P2MP solution, including fault diagnosis using LSP ping.

- Chapter 5 analyzes in more detail how a number of other existing features for P2P LSPs can be applied to P2MP LSPs in order to deploy them in an MPLS or GMPLS network.

- Chapters 6 describes the protocol mechanisms available (and not yet available) to provide high availability for P2MP LSPs through timely detection and recovery from faults and graceful restart – key requirements for IPTV transport.

- Chapter 7 summarizes the points raised and conclusions drawn by the document.

- Finally, chapter 8 provides a brief overview of Metaswitch.
1.1 What is RSVP-TE point-to-multipoint?

The original RSVP-TE protocol is used to signal point-to-point (P2P) label switched paths (LSPs) across MPLS and GMPLS networks.

P2P LSPs, illustrated in figure 1 below, provide a data forwarding path which is traffic-engineered. Resources are reserved at each LSR meaning that QoS can be guaranteed, and explicit routing allows optimal usage of network resources. However, such LSPs are restricted to unicast, one-to-one traffic.

RSVP-TE point-to-multipoint (P2MP) extensions complement the existing protocol with a means to signal efficient, traffic-engineered, one-to-many communication channels. P2MP LSPs deliver information from a single source to many receivers along an optimized point-to-multipoint tree, as illustrated in figure 2 below.

Each LSR in a P2MP LSP plays one of the following roles.

- **Source** node (LSR A) acts like a normal P2P LSP ingress (although it may also be a branch). There is always a single source node.
- Each **leaf** node (LSRs E and G) behaves like a normal P2P LSP egress.
- **Branch** nodes (LSR C) replicate data flows and forward them to multiple next-hop LSRs.
- **Bud** nodes (LSR F) are a combination of a leaf node and a branch/transit node. They replicate data traffic and both pass it up to local endpoints and forward it to next-hop LSRs.
- A **transit** node is any node which forwards traffic (just like a normal P2P LSP transit node).
Examples of P2MP LSPs in this white paper are all given in terms of general case mesh networks like the one above. However, it is worth noting that in practice, some MPLS and GMPLS networks will be less complicated. For example, they may be designed in a single ring topology.

P2MP LSPs are equally applicable to ring topologies. In that scenario, they will have no branch nodes – just a source, a series of transit or bud nodes going around the ring, and finally a leaf. No complex routing algorithms are required, and the data plane need only support bud (drop-and-continue) behavior, not full branch behavior.

### 1.2 Protocol standardization

The RSVP-TE protocol extensions for signaling P2MP LSPs have been published in RFC 4875. The IETF MPLS working group also have a number of other standards in the pipeline, applying existing MPLS features to P2MP LSPs. These drafts are all now reasonably stable.

- TE MIB extensions for the configuration and management of P2MP LSPs (draft-ietf-mpls-p2mp-te-mib).
- Facility fast reroute using P2MP bypass LSPs (draft-ietf-mpls-p2mp-te-bypass).
- LSP ping for P2MP LSPs (draft-ietf-mpls-p2mp-lsp-ping and draft-ietf-mpls-remote-lsp-ping).

However, there are still weaknesses and gaps in the standards. For example, the application of one-to-one (detour) fast reroute to P2MP LSPs is very weakly defined, and no standards are in place for GMPLS protection and restoration of P2MP LSPs (see sections 6.1 and 6.2). Standards work will be required before that function can be made interoperable.
Other working groups within the IETF are developing further protocols that support or utilize RSVP-TE P2MP LSPs, although it is still early days for most of this work.

- The PCE working group recently started work on protocol extensions to allow the use of Path Computation Elements in computing P2MP routes.
- The BFD working group are currently specifying use of the Bidirectional Forwarding Detection protocol for connectivity verification of both P2P and P2MP LSPs.
- The L2VPN and L3VPN working groups are both working on extensions for efficient multicasting of VPN data and some of these standards allow use of RSVP-TE P2MP LSPs in the provider network.

All of these standards are discussed in detail later in this document, and references to them are provided at the end.
2. Requirements from IPTV

Opportunities to generate revenue from the provision of IPTV are currently the primary motivation for service providers to consider deployment of RSVP-TE P2MP LSPs. However, it should be noted that there are also other potential applications for this technology such as multicast VPNs, video conferencing, and stock reporting.

IPTV video distribution services broadly fall into two categories.

- Broadcast TV, in which content is multicast simultaneously to all subscribers which have requested it. This can be used as a direct replacement for existing broadcast TV services, and is also well suited to the transmission of sports events and other live television.

- Video on Demand (VoD), which allows subscribers to access a library of content, and have it streamed to them at their convenience. While this has a greater revenue potential, it is also unicast in nature and will place a much higher load on network resources.

A third category, near-VoD, involves broadcasting the same content repeatedly starting at intervals of, say, every 5 minutes. This offers much of the convenience of VoD, with some of the efficiency savings of multicast delivery.

Broadcast TV and near-VoD services, delivered over IP/MPLS networks, require reliable and efficient multicast transport, and are therefore a potential application for P2MP TE LSPs. This chapter discusses the requirements of that multicast transport and the applicability of P2MP TE LSPs as a solution.

2.1 End-to-end IPTV network architecture

IPTV will typically be just one service provided by an IP-based triple-play network which also supports

- transport services, such as High Speed Internet (HSI) for residential customers and VPNs for enterprises
- other application services, such as VoIP.

Figure 3 below illustrates a simplified architecture for such a network.
The **super headend** (SHE) is responsible for acquiring content from TV stations, encoding it and delivering it through the core network. It is distributed for redundancy.

One or more **video hub office** (VHO) is situated in each region / metro area to which TV services are being supplied. They insert regional content such as local news and advertisements then distribute the TV services to anything from 100,000 to 500,000 homes.

**Video serving offices** (VSOs), sometimes called video switching offices, house aggregation routers such as DSLAMs, each of which might provide triple-play services to hundreds or thousands of homes.

There are number of standards bodies working on detailed IPTV network architectures.

- The ITU-T IPTV Global Standards Initiative (GSI) is developing standards for all aspects of the end-to-end IPTV architecture, including various multicast transport options.

- The ATIS IPTV Interoperability Forum (IIF) have also defined an end-to-end high level architecture for IPTV.

- The Open IPTV Forum have produced a functional architecture for IPTV together with detailed standards for various aspects of the service.

While the output of all of these standards bodies is broadly consistent with the basic network architecture presented in figure 3 above, none of them mandate a particular technology for multicast transport through the core and aggregation networks.

The following section analyzes the requirements on and solutions for that multicast transport.
2.2 Multicast transport requirements

Requirements on IPTV services are primarily driven by the subscribers’ Quality of Experience (QoE). Television users have become accustomed to a high standard of service from existing broadcast, satellite and cable networks, and will only be prepared to pay for an IPTV offering that meets or exceeds it.

The Quality of Experience which users demand places tough requirements on the transport network.

- The service must be highly available. A failure of mainstream TV is simply unacceptable, and operators therefore need to provide at least 5 9s (99.999%) availability.

- Visible artifacts in digital television, such as blocking or noise, are extremely annoying. Providers typically need to offer assurances of less than one visible artifact per two hour movie. That equates to a requirement on the transport network to provide guaranteed Quality of Service (QoS) with a packet loss rate lower than $10^{-6}$.

- Users expect to be able to change channels just as fast as they can click the remote control. One factor in achieving a truly fast channel change is excellent performance for the delivery of a new multicast stream to a user.

In addition to the QoE demands coming from end users, service providers also have their own efficiency and manageability requirements for IPTV multicast transport networks.

- Video distribution is bandwidth intensive. The bit rate per channel currently ranges from 2 Mbit/s for standard definition to 12 Mbit/s for high definition encoded using H.264 (MPEG4). If the multicast transport makes optimal use of available network resources then more channels can be provided to more subscribers without increasing CAPEX.

- Network management and maintenance costs, and the efficiency with which new services can be deployed and problems can be resolved, must also be considered.

The following sub-sections discuss those five requirements on the multicast transport network in more detail and examine the applicability of RSVP-TE P2MP LSPs as a solution.

2.2.1 High availability

Any multicast transport network being used for broadcast IPTV services must be resilient to failures, and various techniques are available to achieve that.

- Individual nodes in the network can provide non-stop forwarding through mechanisms such as graceful restart, failover to redundant backup processors and hot software upgrade.

- The transport protocol may also provide fault tolerance by quickly diverting traffic around a failure or by delivering multiple redundant data streams in parallel.
RSVP-TE P2MP LSPs can help here, but current standardized protocol mechanisms alone do not provide a complete solution. Manual provisioning must be used to some extent to fill in the gaps.

Chapter 6 discusses the high availability mechanisms available for P2MP LSPs. In packet networks, facility fast reroute can reroute around failed links and nodes in less than 50ms. However, that will still result in the loss of some packets, the consequence of which will be visible artifacts that damage the viewer's QoE. GMPLS protection with 1+1 redundancy is more likely to achieve a seamless switchover, but has not yet been standardized for P2MP LSPs.

Both fast reroute and GMPLS protection will make the multicast transport resilient to failures of links or nodes within the core of the network. For an end-to-end highly available solution, resilience to source and leaf node failures is also required.

Leaf node failures can be handled simply by adding more redundant leaf nodes in the same VHO or VSO. Source failures are more difficult. RSVP-TE P2MP does not provide a mechanism for source (SHE) redundancy. Fast switchover to a backup SHE will require a proprietary solution in which a redundant P2MP LSP is pre-signaled from the backup SHE to all VHOs, and some mechanism external to the protocol is used to associate the primary and backup data streams.

2.2.2 Quality of Service

In order to deliver good picture and sound quality without artifacts, the multicast transport must guarantee that sufficient forwarding resources will be available to broadcast IPTV streams, with a very low packet loss rate.

RSVP-TE is an ideal solution here, as it allows providers to guarantee QoS by reserving dedicated bandwidth for each LSP throughout the network.

2.2.3 Fast channel change

Achieving a channel change speed that can compete with existing broadcast TV services is one of the biggest challenges for IPTV service providers, and also a major factor in the subscribers' QoE.

Currently, the majority of IPTV channel change time (up to half a second) is taken up by the lag in receiving an I-frame in the MPEG stream. Compressed video streams are made up of different types of packets, as illustrated in figure 4 below.

- I-frames encode a complete picture and can be rendered independently.
- P-frames encode differences from the previous I- or P-frame, and cannot be rendered unless that frame has also been received.
- B-frames encode differences from both the next and previous I- or P-frames, and cannot be rendered until both those frames have been received.
Figure 4: Frames making up a compressed MPEG stream

When a new channel is selected, the picture cannot be rendered until the next I-frame is received. Mechanisms are available to reduce the delay in receiving that I-frame. For example the RTCP Full Intra (I-frame) Request command, defined in RFC 5104, allows a set-top box to request a new I-frame immediately. Techniques like that will be crucial for achieving truly fast channel change in IPTV. However, this is purely a data plane problem and is beyond the scope of the multicast transport.

The time taken to start delivering new streams to users is also a key factor in channel change time that needs to be optimized. In order to minimize that delay, the channel must already be being delivered to a device as close to the user as possible – certainly to the VSO and ideally to the user’s DSLAM (or equivalent aggregation device). If that is done, then again the channel change time is not affected by the multicast transport protocol performance in the aggregation network (no signaling to graft a new leaf into the P2MP LSP will be required).

Broadcasting all channels to all VSOs in this way, regardless of whether a subscriber has requested them, will result in sub-optimal use of network resources. However, it is also necessary that a valid request for a new channel is guaranteed to succeed. The only alternative use for the bandwidth taken up by unwatched channels would therefore be for low priority traffic that could be preempted.

2.2.4 Efficiency

High definition video, even when compressed, requires around 12 Mbit/s of bandwidth. Multicast, rather than unicast, transport is therefore essential in order to avoid unnecessary duplication of data and provide good scaling of bandwidth usage with the number of subscribers or VSOs.

RSVP-TE P2MP LSPs allow explicit routing for truly optimal usage of the available network bandwidth. If different channels are mapped to different LSPs, they can be given different routes.
through the network for improved load balancing. Further, P2MP LSPs are not restricted to Shortest Path Trees (SPTs) – any optimization scheme is allowed. For example Steiner Trees, in which the total cost of the P2MP tree (rather than the individual source-to-leaf costs) is minimized.

The difference between SPTs and Steiner Trees is illustrated in a simple network in figure 5 below. Both trees provide routes from the source A to leaves D, E and F. In the SPT the maximum distance from the source to any leaf is two hops, and there are 5 hops in total. In the Steiner Tree, leaf F is three hops from the source, but there are only 4 hops in total. Total network bandwidth usage is therefore minimized.

![Figure 5: Shortest Path Trees and Steiner Trees](image)

### 2.2.5 Manageability

Finally, service providers need a network with low OPEX costs that is easy to maintain and manage.

The multicast distribution trees for IPTV, going between the SHE and VHOs or between a VHO and VSOs, are likely to be very static. Addition and removal of VHOs and VSOs will not be frequent operations. Static configuration, rather than dynamic signaling, of P2MP LSPs may therefore be sufficient. Note, however, that some protocol features such as sophisticated protection and restoration are only available for dynamically signaled LSPs.

Tools for diagnosing faults in the multicast transport will also be required if the network is to be maintainable. For MPLS transport networks, Bidirectional Forwarding Detection (BFD) and LSP ping can be used to detect and isolate failures in the data plane, using the P2MP extensions described in section 4.3.
2.2.6 Summary

RSVP-TE P2MP LSPs are a good fit to the requirements for broadcast IPTV transport in both core and aggregation networks. All of the key requirements for high availability, QoS, fast channel change, efficiency and manageability are met.

Although there are a number of other competing technologies for IPTV multicast transport, all of them have shortcomings that mean they will only be appropriate in networks that are able to relax one or more of the above requirements.

- Pure Ethernet networks provide efficient multicasting of data, but they cannot provide TE to guarantee QoS or fast recovery from failures. Scalability problems also mean that network diameter is limited to around 7 hops.
- PBB-TE adds Traffic Engineering to Metro Ethernet networks. However, it does not yet provide multicast capabilities that match P2MP TE LSPs.
- IP multicast is a strong contender. The basic multicast function it provides has recently been augmented with protocol extensions that provide some level of high availability. For example, fast convergence allows sub-second recovery from network failures and multi-topology routing can provide 1+1 redundancy for the data stream. However the speed and flexibility of these solutions is not on a par with MPLS. Further, IP multicast still cannot reserve dedicated bandwidth and guarantee QoS. It is therefore only sufficient in networks where traffic patterns can be accurately predicted and there is sufficient over-provisioning to ensure the bandwidth required for broadcast IPTV will always be available.

2.3 Transport network architecture

As noted earlier, multicast VPNs are another potential application for P2MP LSPs. In fact, there is overlap here with IPTV. This section looks at how a multicast VPN architecture may be used for the provision of IPTV services.

In this model, the transport network is providing a multicast VPN to the IPTV provider. Broadcast IPTV data is multicast from VHOs to VSOs over this VPN, and P2MP LSPs are used within the transport network to carry the multicast data.

Both the L3VPN and L2VPN working groups at the IETF are working on multicast extensions to their VPN services that will provide standardized mechanisms to achieve that. These are discussed in the following sub-sections.
2.3.1 VPLS multicast

VPLS multicast extensions, allowing the efficient transport of multicast data between VPN sites using P2MP LSPs, are defined in draft-ietf-l2vpn-vpls-mcast. An application of this network architecture to IPTV broadcast from the VHO to VSOs through the aggregation network is shown in figure 6 below.

Figure 6: VPLS multicast for IPTV transport

The key elements of this network are as follows.

- The aggregation network is providing a VPLS instance to the IPTV provider. This gives the illusion of direct Ethernet connectivity between the router in the VHO and the DSLAMs in the VSOs.

- The VSOs forward subscriber channel requests up through the VPLS to the VHO. The VHO then streams multicast data over the VPLS to the VSOs. (The “channel request” is actually an IGMP Report message and the DSLAMS act as IGMP Proxies to aggregate requests from users before forwarding them up through the network.)

- The PE router for the VHO is configured to map IPTV channels to P2MP LSPs. Typically those P2MP LSPs will be pre-provisioned to every VSO in order to allow fast channel change (as discussed in section 2.2.3). Alternatively, the VPLS multicast extensions also allow auto-discovery and dynamic setup of P2MP LSPs to the PE routers that are serving subscribers that have requested a particular channel.

- Each VSO PE router may also be connected to multiple DSLAMs. In that case it can use IGMP to learn which DSLAMs have requested a particular channel, and only forward the multicast data to those DSLAMs. However, doing so may again be detrimental to channel change times (see section 2.2.3).
2.3.2 L3VPN multicast

The correct way to do multicast in Layer 3 VPNs has been a contentious issue within the IETF, with several competing solutions. The details of these and the trade offs between them are complex.

- draft-ietf-l3vpn-2547bis-mcast and draft-ietf-l3vpn-mvpn-considerations are being standardized by the IETF L3VPN working group. These do support the use of P2MP TE LSPs to carry customer multicast data between VPN sites, but also allow alternatives such as IP multicast.

- draft-rosen-vpn-mcast was never adopted by the IETF, although much of its function was absorbed into draft-ietf-l3vpn-2547bis-mcast. It has a number of deployments, and was recently refreshed with ‘informational’ status. It does not define procedures for use of P2MP LSPs in the provider network, but hints that an implementation could choose to use them.

- draft-mnapierala-mvpn-rev, also unadopted and now expired, suggested another alternative approach. It offered some flexibility in the tunneling protocol used in the provider network, but RSVP-TE P2MP LSPs were not supported.

It remains to be seen whether support for TE of customer multicast data will gain traction among vendors and service providers.
RFC 4875 defines P2MP extensions to RSVP-TE for both MPLS and GMPLS (packet and optical) networks. This chapter provides a description of that base protocol. Some more advanced pieces of function covered by the RFC (fast reroute and LSP hierarchy) are left to later chapters for a more detailed analysis.

### 3.1 Overview

Point-to-multipoint LSPs are initiated by the source (ingress) LSR and signaled to all leaf (egress) LSRs. As with P2P LSPs, the identity of the leaves must be known by the source in advance. This is very different to IP multicast, where the tree is set up in the opposite direction: leaves signal their intention to join to the source. Section 4.1.1 presents various ways in which a source might discover which LSRs wish to be leaves of a P2MP LSP.

The source LSR is allowed to compute a complete or partial route for the P2MP LSP. RSVP Path messages are then sent downstream until one arrives at each leaf LSR, and each leaf responds with a Resv which is forwarded upstream to the source. Every LSR along the path allocates a single label for the LSP and communicates it upstream in the Resv so that a complete P2MP LSP is set up. Branch nodes must be capable of duplicating flows in the data plane and forwarding them to all of their downstream LSRs with the appropriate label for each.

#### 3.1.1 Explicit routing

As with P2P LSPs, it is possible for the source LSR to specify an explicit route for a complete P2MP LSP, or just some sub-trees, using both strict and loose hops. Any kind of tree optimization is allowed, as discussed in section 4.1.3. The rules for hop specification are the same for standard P2P LSPs. An LSP can also be signaled without an explicit route, in which case it will be routed hop-by-hop.

#### 3.1.2 Tree modifications and repairs

Once a P2MP LSP has been setup and is carrying data, it can then be modified in various ways without disrupting the data stream.
- New leaves can be grafted in and start receiving data branched off from some point in the existing tree. Similarly, existing leaves can be pruned and removed from the tree. The mechanisms for doing this are described in section 3.3 below.

- The whole P2MP LSP, or just some sub-tree, can be rerouted according to the procedures described in section 3.4. This can be done without data loss by setting up the new route before tearing down the old one. However, the switchover is still likely to cause some jitter. Such rerouting may be desirable to re-optimize the route of the LSP or to divert data around a planned network outage. Mechanisms that provide high availability in the case of a network failure are described in chapter 6.

### 3.1.3 Back-compatibility

P2MP RSVP-TE signaling is not back-compatible with legacy LSRs that don’t support the protocol extensions in RFC 4875. All P2MP signaling messages will be rejected by legacy nodes.

However, it is possible to use LSP tunneling to allow P2MP LSPs to traverse legacy LSRs. Consider the diagram below in which LSR C does not support RFC 4875.

![Figure 7: P2MP LSP traversing non-P2MP capable LSR C](image)

By setting up P2P LSPs A-C-E and A-C-F which act as forwarding adjacencies (virtual interfaces) A-E and A-F, it is possible to make LSR A the branch node of the P2MP LSP with data traffic tunneled through LSR C.

In a more complex network, an LSR might expand a loose hop in a P2MP LSP and find that it needs to traverse other LSRs which do not support P2MP signaling. LSP setup can still proceed if it then dynamically signals P2P LSPs to the next LSRs which do support P2MP signaling, and uses those tunnels as outgoing interfaces for the P2MP LSP.
In an optical network, where label stacking is not possible, similar function can be achieved through use of LSP stitching (RFC 5150) instead of tunneling.

### 3.2 Protocol mechanisms

Two key techniques are used to extend RSVP-TE P2P LSP signaling for support of P2MP LSP setup.

- The P2MP LSP is modeled as a set of P2P LSPs, each of which stretches from the source to a single leaf. These are called source-to-leaf (S2L) sub-LSPs and are discussed further in section 3.2.1 below.

- Individual S2L sub-LSPs can either be signaled in separate Path and Resv messages, or, so that control plane bandwidth usage is minimized, combined into a small number of messages. This is done by grouping S2L sub-LSPs into sub-groups as described in section 3.2.2.

#### 3.2.1 Source-to-Leaf (S2L) sub-LSPs

A P2MP LSP is described in the control plane as a set of P2P LSPs called source-to-leaf (S2L) sub-LSPs, each stretching from the source LSR to a single leaf or bud LSR.

For example, the P2MP LSP pictured in figure 8 below is made up of three S2L sub-LSPs: A-C-E, A-C-F-G, and A-C-F.

![Figure 8: S2L sub-LSPs](image)

Although S2L sub-LSPs can be signaled in separate Path and Resv messages, they all belong to the same P2MP LSP. LSRs such as C in figure 8 above, handling multiple incoming S2L sub-LSPs on the
same interface, will therefore allocate a single label and advertise it to LSR A, meaning that there is no unnecessary duplication of state or traffic in the data plane.

### 3.2.2 Sub-groups

S2L sub-LSPs can be grouped into sub-groups by any node along the course of a P2MP LSP. Each sub-group must be signaled in a separate Path message.

Figure 9 below illustrates a possible division of a P2MP LSP into sub-groups. Each sub-group is identified by the sub-group originator ID (the LSR which defined the sub-group) and the sub-group ID (assigned by that LSR).

![Figure 9: Sub-group division](image)

An LSR might want to restrict the size of its sub-groups for two reasons.

- In order to limit the size of the Path messages being sent downstream and avoid IP fragmentation (reducing the volume of retries required each time a single IP packet is dropped). In the example above, LSR C knows that the link C-F has a low MTU, and therefore defines sub-groups so that the S2L sub-LSPs on that link can be signaled in separate Path messages.

- According to local policy or configuration, for example because it knows a particular set of leaf nodes are likely to share the same fate and putting them in a sub-group allows it to teardown or re-optimize the sub-tree for those leaf nodes independently from the rest of the P2MP LSP (see section 3.4.2).
Also, an LSR may be constrained in how it signals S2L sub-LSPs by interoperability requirements. For example, if other devices in the network will only support one S2L sub-LSP per Path message, then the LSR may have to signal accordingly.

Although two S2L sub-LSPs traversing the same link can be put in different sub-groups and therefore signaled in different Path messages, they still share a single label in the data plane. Managing this asymmetry between the control plane and data plane state is one of the principle challenges for an RSVP-TE P2MP implementation.

When a sub-group containing multiple S2L sub-LSPs is signaled in a single Path message, a technique called “explicit route compression” is also used to reduce the message size. The explicit route of a P2MP LSP is defined as a series of hops for each S2L sub-LSP. When multiple S2L sub-LSPs are included in a single Path message their explicit routes are likely to share hops. The explicit route is therefore compressed by not duplicating hops that are shared by several S2L sub-LSPs. Instead, the list of hops for each only starts at the point where it branches from another S2L sub-LSP.

Leaf grafting and pruning and loose hop expansion can further complicate explicit route compression. For example, expanding a single AS abstract node into a P2MP tree through the AS might replace one branch point (the AS abstract node) with several, requiring the explicit route compression to be rewritten. This new explicit route processing is a significant extension to standard P2P RSVP-TE.

There are also a number of additional complications in the use of sub-groups in other protocol messages.

- The mapping of S2L sub-LSPs to Path messages is not the same as their mapping to Resv messages. Multiple sub-groups can be signaled in a single Resv, or a single sub-group can be split between multiple Resvs.

- Notifies, which would normally be sent end-to-end, must instead be intercepted by LSRs which have redefined the sub-group identifiers for a P2MP LSP so that those identifiers can be rewritten in the Notify messages as well. Branch LSRs are also required to intercept downstream Notify messages so that they can be duplicated and sent to LSRs on all branches which have requested them.

- Similarly, ResvConf messages cannot be sent end-to-end, but must be intercepted by LSRs that have redefined the sub-groups so that their sub-group identifiers can be rewritten.

Consequently, an LSR may wish to simplify its implementation by always putting every S2L sub-LSP in its own sub-group, and signaling them in different Path messages. Provided upstream LSRs do the same, this means sub-groups will never need to be redefined and explicit route compression is also not required. The downside is that control plane bandwidth usage will not be optimal.
3.3 Grafting and pruning

3.3.1 Grafting
New leaf LSRs can be grafted into (added to) a P2MP LSP only by explicit signaling from the source (just as at initial LSP setup time knowledge of all leaf LSRs is required at the source). The source LSR will graft in a set of leaves by sending a Path message including the new S2L sub-LSPs either in a new sub-group or added to an existing one.

3.3.2 Pruning
The source can also prune (remove) individual, or sets of, leaf nodes from a P2MP LSP.

- If one or more complete sub-groups are being removed, a PathTear message is sent for each sub-group. This technique, called explicit teardown, also applies to complete P2MP LSP teardown.

- If just a subset of the S2L sub-LSPs in a sub-group are being removed, a Path message is sent containing all the S2L sub-LSPs for the sub-group except those that are being removed. This is called implicit teardown, and means that a Path message is being used to tear down state – new function in RSVP-TE.

Nodes other than the source LSR can also prune leaves (or equivalently, preempt individual S2L sub-LSPs) by following the same procedure as described here for the source and also sending a PathErr back towards the source including the pruned S2L sub-LSPs and setting the PathStateRemoved flag.

3.4 Re-optimization
Following the grafting and pruning of a number of leaf nodes, a fast reroute, or some other change in the network, the path used by a P2MP LSP (or some subset of its S2L sub-LSPs) may cease to be optimal. LSRs may wish to monitor the paths that P2MP LSPs take and, when they become sub-optimal, reroute them. There are three mechanisms available to do that: make-before-break, sub-group-based sub-tree re-optimization, and dynamic reroute.

3.4.1 Make-before-break
A sub-optimal or broken P2MP LSP can be repaired using make-before-break in exactly the same way as a P2P LSP: by setting up a completely new instance of the LSP then tearing down the old one. The Shared Explicit reservation style can also be used to share data plane resources between the new and old LSPs, preventing double-allocation at common LSRs.

This has the disadvantage that it requires a complete new P2MP LSP to be signaled from the source to every leaf. That is an unnecessary overhead simply to reroute just a small subset of the S2L sub-LSPs. Further, it can only be initiated by the source LSR. It won’t allow a transit LSR at the border of
a new AS to re-optimize the path through that AS. Sub-group-based sub-tree re-optimization is therefore offered as an alternative.

### 3.4.2 Sub-group-based sub-tree re-optimization

This technique can be used to re-optimize a sub-tree (that is a subset of the S2L sub-LSPs from a given transit or branch LSR onwards) without disturbing the remainder of the P2MP LSP.

An LSR that wishes to perform such a re-optimization defines a new sub-group (or set of sub-groups) for the new sub-tree and sends Path message(s) downstream for them. Downstream LSRs treat those requests as being for new protocol state, because the sub-group identifiers have changed. The new sub-tree is therefore signaled as normal, and state for the old (sub-optimal) sub-tree is also left intact. Because the new sub-tree is part of the same P2MP LSP as the old one, data plane resources are also shared, as with make-before-break.

Once the LSR that initiated the re-optimization has received Resvs for all S2L sub-LSPs in the new sub-tree it can switch over data traffic and then tear down the old sub-tree using the pruning techniques described above.

### 3.4.3 Dynamic reroute

Finally, a complete P2MP LSP or sub-tree can be re-optimized simply by using dynamic reroute. This is not discussed in RFC 4875 (except for a brief passing reference) so standard RSVP-TE procedures apply: any LSR can issue a new Path message for an existing LSP out of a different interface (or with a changed explicit route) and let the old path state time out.

However, when a downstream LSR receives the new Path message, its sub-group identifiers may no longer match those for its existing path state (they may have been changed by an intermediate LSR). The downstream LSR must therefore assume that this is sub-group-based sub-tree re-optimization rather than dynamic reroute, and maintain both the old and new path state (as described above) until it times out. That inefficiency means that sub-group-based re-optimization should be preferred to dynamic reroute for P2MP LSPs.

### 3.5 Re-merging

It is possible for two or more S2L sub-LSPs in a P2MP LSP to branch and then later re-merge due to routing inconsistencies. For example, LSRs in the network may have out-of-sync RIBs or use different routing algorithms, or the P2MP LSP may traverse multiple domains so that no one LSR has a complete network picture. The resulting unnecessary duplication of data traffic is inefficient and such re-merge scenarios should be avoided where possible.

Ideally, a re-merge LSR would detect the condition and then signal the error to the upstream LSR which caused it so that it could be corrected. RFC 4875 describes a procedure to achieve that but unfortunately the logic in the RFC is flawed – a more intelligent implementation is required.
In fact, there is no protocol mechanism by which a re-merge LSR can distinguish a permanent re-merge scenario from a valid, transient, condition experienced when sub-group-based re-optimization is being used to reroute a sub-tree. In order to avoid sabotaging such a re-optimization, the re-merge LSR must therefore allow the re-merge condition to persist by maintaining both sets of upstream state, as described in section 3.4.2 above. To avoid forwarding duplicate data traffic downstream, it must drop data received on all but one in segment.

It falls instead to upstream LSRs to correctly detect and correct permanent re-merge scenarios which they themselves have caused. Provided they have sufficient information in the RROs they receive in Resv messages to detect re-merges, they can then correct them using sub-group-based re-optimization.
4. Routing and Management

All network elements providing RSVP-TE P2MP support are likely to require certain key features to complement the base protocol support described in chapter 3. This chapter covers mechanisms that are available for the following.

- Ways for the source to discover which LSRs need to be leaves of a P2MP LSP, and then compute an optimal route to them.
- The MIB configuration and management interface for P2MP LSPs.
- Operations, Administration and Maintenance (OAM) tools for monitoring and troubleshooting P2MP LSPs: LSP ping and Bidirectional Forwarding Detection (BFD) for fault detection and isolation.

4.1 Routing

4.1.1 Leaf discovery

The first step of P2MP LSP set up, before a route can be calculated, is the discovery of which leaf LSRs need to be routed to. The RSVP-TE P2MP base protocol does not solve this problem. Unlike PIM or IGMP, there is no mechanism for leaves to advertise to the source that they wish to be sent a particular multicast stream.

IPTV transport networks will typically have long-lived, stable P2MP LSPs – the addition and removal of VHOs or VSOs will be rare. Consequently, it will often be acceptable to simply use manual configuration to program P2MP LSPs to a well-known set of leaves.

Various attempts have, nonetheless, been made to standardize protocols for auto-discovery of leaf LSRs wishing to be included in a particular P2MP LSP. The following two drafts aimed at the IETF MPLS working group were never adopted and have now expired.

- draft-leroux-mpls-p2mp-te-autoleaf defined a mechanism for leaf LSRs to advertise their requirement to join a particular P2MP LSP through the IGP. However, flooding that information throughout the routing domain is not efficient.
- draft-zheng-mpls-p2mp-topology-agent-reqs suggested a new network element, the Multicast Membership Agent, which would co-ordinate the membership of P2MP LSPs.

The IETF is working on standards for the use of P2MP LSPs for multicast transport in Layer 2 and Layer 3 VPNs. These provide interworking between PIM and IGMP, for endpoints to join multicast
groups in the customer network, and RSVP-TE P2MP signaling in the provider network. They are also applicable to IPTV transport networks, as discussed in section 2.3.

As new applications of P2MP LSPs are developed, further application-specific solutions to the leaf discovery problem may well also emerge.

### 4.1.2 IGP extensions for P2MP capability information

Some large networks may contain a diverse range of LSRs, some of which don't support P2MP. In such cases, route calculation will need to take the capabilities of each LSR into account.

- Which LSRs support RSVP-TE P2MP signaling? (It is possible to tunnel through legacy LSRs that do not have this support, as described in section 3.1.3.)
- Which LSRs are branch-capable in the data plane?
- Which LSRs are bud-capable in the data plane?

RFC 5073 defines OSPF and IS-IS extensions that allow this information to be distributed by the IGP so that it can be considered in route calculations.

### 4.1.3 Tree computation

Once the leaves of a P2MP LSP and the LSR capabilities have been established, the P2MP route can be calculated. This tree calculation is not a concern for interoperability between different LSRs (with the proviso that for loosely routed P2MP LSPs it is sensible to use the same algorithm throughout the network) and the algorithm is not specified by the protocol. It therefore provides an opportunity for vendors to differentiate themselves with the optimality, flexibility and performance of their algorithms.

As discussed in section 2.2.4, various types of tree optimization are possible.

- Shortest-path trees (SPTs) minimize the individual costs of the P2P routes from the source to each leaf (similar to a PIM multicast tree).
- Steiner trees minimize the total cost of the whole P2MP tree.

As well as applying some cost-based optimization, a routing algorithm may want to take additional constraints such as the following into account:

- The latency of the data path to each leaf (which may be affected by the hop count and the number of times data needs to be replicated at branch points).
- The need to be resource-disjoint from some other data path in order to provide protection for another LSP.
- Resource affinities, by which different LSPs are restricted to different subsets of the links in the network.
4.1.4 Path Computation Elements (PCEs)

The IETF PCE working group is also developing standards for the use of PCEs in P2MP LSP routing. Offloading P2MP route computation to an external PCE network element provides similar advantages to the use of PCEs for P2P LSPs, and may also allow more advanced routing algorithms to be deployed on specialist platforms. However, PCEs are by no means a pre-requisite for the deployment of P2MP LSPs.

There are three active drafts in this area, but it is still early days for this work.

- draft-ietf-pce-p2mp-app details the applicability of PCEs to P2MP routing.
- draft-ietf-pce-p2mp-req describes requirements on PCE protocol extensions to support P2MP routing. These are restricted to routing and don’t cover leaf discovery.
- draft-ietf-pce-pcep-p2mp-extensions defines the extensions to the PCE Communication Protocol for P2MP route calculation.

4.2 Configuration

Extensions to the TE MIB for the management of RSVP-TE P2MP LSPs are defined in draft-ietf-mpls-p2mp-te-mib, which is now stable. MIB objects are defined that allow

- creation, control and monitoring of P2MP LSPs, including per-branch performance statistics
- notifications of events such as activation, failure and rerouting of a P2MP LSP or S2L sub-LSP
- route management including explicit hop configuration, P2MP tree calculation and actual route monitoring.

4.3 LSP ping

LSP ping, defined in RFC 4379, provides a mechanism by which P2P LSP failures can be detected and isolated using MPLS echo requests. It supports two modes of operation.

- In ping mode, MPLS echo requests are sent as labeled data down the length of the LSP, to check that there is end-to-end connectivity.
- In traceroute mode, successive MPLS echo requests are sent with T TLs of 1, 2, 3, etc. so that each LSR along the length of the LSP receives them in turn and the point of any failures can be determined.

Note that neither MPLS echo requests nor replies are sent or received by the RSVP-TE protocol stack in the control plane. This is primarily data plane function. Requests are sent as labeled data along the LSP, and replies are returned as UDP packets (which may in turn be tunneled over an LSP).
4.3.1 P2MP LSP ping

LSP ping function for P2MP LSPs is defined in draft-ietf-mpls-p2mp-lsp-ping, which again supports both ping mode, to check for end-to-end connectivity, and traceroute mode, for point of failure determination.

Because MPLS echo requests are sent as labeled data through the P2MP LSP, they will always be replicated at branch points and delivered to every leaf. A P2MP LSP might scale to hundreds or even thousands of leaves, and if they all respond to an MPLS Echo request simultaneously, that will equate to a Denial of Service attack on the source. A new Jitter TLV has therefore been defined that the source LSR should include in the request so that leaves introduce a random delay before sending their responses.

The source LSR can also use ping or traceroute to analyze the path to an individual leaf by including a Responder Identifier TLV in the request. Although the request will still be sent (as labeled data) to all leaves, only LSRs on the path to the responder identified will send a response.

4.3.2 Proxy LSP ping

To further improve the scalability of LSP ping for P2MP LSPs, Proxy LSP ping, defined in draft-ietf-mpls-remote-lsp-ping, allows MPLS Echo requests to be sent on just a sub-tree rather than flooded throughout the whole LSP.

Consider the example in figure 10 below. Here, the source LSR has determined that there has been a failure that affects just one sub-tree of the P2MP LSP. Rather than sending an MPLS Echo request itself, the source node first determines the root of the failed sub-tree, perhaps by looking at the P2MP route recorded by RSVP-TE signaling. It then sends a Proxy MPLS Echo request, out-of-band, to that root node – the Proxy LSR. The Proxy LSR generates an MPLS Echo request for the P2MP LSP, and if there is connectivity, the leaves respond directly to the source LSR.

In fact, the initiator of the Proxy MPLS Echo request needn’t be the source LSR. This technique could also be used, for example, for a leaf LSR to tell an upstream LSR to ping it, and step by step trace the route of its S2L sub-LSP right back to the source.
4.3.3 MPLS BFD and Multicast Connectivity Verification

The control plane processing overhead of LSP ping makes it unsuitable for providing rapid LSP failure detection. This problem is further exacerbated for P2MP LSPs because of the additional risk of swamping the source LSR with MPLS echo replies from 100s or 1000s of leaves.

Bidirectional Forwarding Detection (BFD) is being designed by the IETF as an extremely light-weight means of data plane connectivity verification. It can be implemented in hardware or firmware and is therefore suitable for timely detection of failures.

The application of BFD to P2P LSPs, overcoming the scalability problems of LSP ping, is defined in draft-ietf-bfd-mpls.

Further extensions to apply BFD to P2MP LSPs are being tackled by two drafts: draft-swallow-mpls-mcast-cv and draft-katz-ward-bfd-multipoint. The latter has now been adopted by the IETF BFD working group, but is still not a stable standard.
A large number of standards have been developed adding advanced function to P2P MPLS and GMPLS LSPs. Many of them can be applied equally to P2MP LSPs without any further standardization being necessary.

This chapter discusses a selection of features commonly used for P2P LSPs in packet and/or optical networks for which application to P2MP LSPs is tricky or undefined. In some cases, standards development will be needed to make them interoperable.

5.1 LSP hierarchy

LSP hierarchy, defined for P2P LSPs in RFC 4206, allows multiple higher layer LSPs to be tunneled through a lower layer hierarchical LSP (H-LSP). This can be used to improve MPLS and GMPLS network scalability. LSRs in the core of the network only need to manage state for the small number of H-LSPs through which many higher layer LSPs are tunneled. Tunneling in this way is also a prerequisite for facility fast reroute.

H-LSPs can also be advertised into the IGP as forwarding adjacencies so that other nodes in the network can use them in route computation.

The application of LSP hierarchy to P2MP LSPs is discussed in RFC 4875, but only for tunneling through P2P LSPs. However, more recent drafts have defined mechanisms that make tunneling through P2MP H-LSPs possible, as illustrated in figure 11 below.
Here, two P2MP LSPs are tunneled through a single P2MP hierarchical LSP, shown in green. LSR D only has state for the H-LSP. It is not on the control plane path for the other two LSPs, and data is forwarded opaquely using label stacking. In order to achieve this, upstream-assigned labels must be used, as described in section 5.1.1 below.

Note also that not every leaf of the H-LSP needs to be on the path of the P2MP LSPs that are being tunneled through it. All H-LSP leaves will receive data for a given tunneled LSP, but those not on its path will simply drop it. While this is inefficient use of forwarding resources, it allows a single H-LSP to support a wider range of tunneled LSPs and therefore further reduce state in the core of the network.

No standards have been defined for advertisement of P2MP LSPs as forwarding adjacencies in the IGP. Without that function, source LSRs are not able to discover and take advantage of P2MP H-LSPs in the core network, other than through explicit configuration. However, specification of P2MP unidirectional links in IGPs and their inclusion in routing algorithms would be highly complex, so explicit configuration may be more practical.

### 5.1.1 Upstream-assigned labels

In figure 11 above, data for P2MP LSP 1 is forwarded down the H-LSP and arrives at LSRs E, F and G. Those LSRs pop the outer H-LSP label, and then forward the data based on the inner label for P2MP LSP 1. LSR G will drop it because it does not recognize the inner label. LSRs E and F must use the same incoming label for P2MP LSP 1, because there is no way to deliver data to them with different inner labels.

![Upstream-assign labels for P2MP tunneling](image)

Figure 12: Upstream-assign labels for P2MP tunneling
Upstream-assigned labels, defined in RFC 5331, allow the upstream node, LSR B, to assign the same incoming label to P2MP LSP 1 at both LSRs E and F, and thus make tunneling through P2MP LSP possible. This is illustrated in figure 12 above, a close up of LSRs E and F from figure 11. The signaling extensions defined in draft-ietf-mpls-rsvp-upstream are also needed for upstream-assigned label support in RSVP-TE P2MP LSPs.

5.1.2 Context-specific label spaces

Because upstream-assigned labels are assigned by an upstream LSR, they are only unique within the context of that LSR. At LSR F in the example above, another upstream LSR might have assigned the same inner label to another P2MP LSP being tunneled through a different H-LSP. Upstream-assigned labels are therefore taken from a new context-specific label space.

In the case of tunneling through P2MP LSPs, upstream-assigned inner labels must be looked up within the context of the tunnel they were received on. That tunnel is identified using the outer label, and penultimate-hop-popping (PHP) cannot therefore be used on H-LSPs.

Protocol extensions in draft-ietf-mpls-rsvp-te-no-php-oob-mapping allow H-LSPs to be configured not to use PHP, enabling that condition to be met.

5.2 Graceful deactivation

GMPLS includes an ADMIN_STATUS protocol object that can be included in Path messages to gracefully deactivate P2P LSPs. When LSRs along the LSP receive a Path message setting the state to “admin down” and “deletion in progress”, they disable loss of light (or other) alarms for the LSP. A subsequent PathTear message is then used to actually remove state for the LSP.

This procedure can also be applied to the graceful deactivation of a whole P2MP LSP. However, there is no way to signal the admin status for just a single S2L sub-LSP, and therefore gracefully prune a leaf from the P2MP LSP. If that function is required then appropriate standards will need to be developed.

---

1 Upstream-assigned labels can also be used when a P2MP LSP has multiple next hops on a LAN segment to avoid duplication of broadcast data traffic on the LAN. In that case there is no outer tunnel that can be used by the receiver to determine that the label is upstream-assigned and belongs to a per-neighbor label space. Instead, RFC 5332 defines data link layer codepoints used to indicate that the top label of the encapsulated packet is upstream-assigned.
5.3 **Explicit label control**

Some GMPLS networks use explicit label control, defined in RFC 3473, to control the labels used along LSP paths as well as the LSRs they traverse. This is often necessary when some or all LSRs in the network are not capable of wavelength conversion.

The presence of explicit labels complicates the presentation and compression of P2MP explicit routes in ways which are not discussed in RFC 4875. For example, some parallel S2L sub-LSP segments may traverse identical LSRs but use different labels, meaning the parallel hops must be repeated in the P2MP explicit route.

5.4 **Bidirectional LSPs**

GMPLS includes support for bidirectional LSPs. However, RFC 4875 explicitly states that P2MP LSPs are unidirectional. Data flow from leaves to the source is not supported.

That is not a problem for applications like IPTV where the data flow is unidirectional. Other applications may require bidirectional, or even multipoint-to-multipoint TE LSPs, but those standards have not yet been defined.
6. High Availability

Individual LSRs can be made highly available using mechanisms like fault tolerance and hot software upgrade. This chapter, however, looks at three protocol techniques that can make P2MP LSPs resilient to failures of links and nodes in the network.

- **Fast reroute**, which allows data to be quickly diverted around failures along pre-programmed backup LSPs in packet networks.
- **GMPLS protection and restoration**, which offer a range of recovery techniques that can be applied to both packet and optical networks.
- **Graceful restart**, which allows a node whose control plane has restarted to reconstruct its signaling state based on refresh messages from neighbors and its preserved data plane.

As discussed in chapter 2, reliable transmission of multicast data is a key requirement for an IPTV transport network if it is to meet subscribers’ Quality of Experience expectations. One or more of these high availability techniques is therefore likely to be needed by an IPTV offering based on P2MP TE LSPs.

### 6.1 Fast reroute

Fast reroute procedures allowing P2P packet LSPs to be quickly rerouted around failures using pre-programmed backup paths are specified in RFC 4090. Although they enable sub-50ms recovery from data plane failures, even that level of reliability may not be sufficient for IPTV where any packet loss is likely to cause visible picture defects, and loss of a single I-frame could result in an outage of half a second (see section 2.2.3). Nonetheless, fast reroute extensions for P2MP LSPs have been defined, and may still be applicable to IPTV in combination with other resilience mechanisms.

There are two basic flavors of fast reroute which, although theoretically interoperable, are not generally expected to co-exist in the same network.

- **Facility** (or bypass) fast reroute tunnels data through bypass LSPs in the event of a failure on the protected path. The bypass LSP goes between two nodes on the protected LSP: the point of local repair (PLR) and the merge point. The PLR is responsible for diverting traffic down the bypass in the event of a failure of the protected LSP. A single bypass LSP can be shared by multiple protected LSPs.

- **One-to-one** (or detour) fast reroute sets up dedicated detour LSPs at the same time as the protected LSP. Each node along the protected LSP (except the egress) tries to make itself a
PLR by programming a detour along a route between itself and some downstream merge point which avoids the next hop on the protected LSP. Detour LSPs can also merge with each other before they re-merge with the protected LSP.

6.1.1 Facility (bypass) fast reroute

A basic application of fast reroute procedures to P2MP LSPs is defined in RFC 4875. It uses P2P bypass LSPs to protect individual S2L sub-LSPs. A single PLR may therefore need multiple bypass LSPs to provide full protection, each with a downstream merge point on a different sub-tree, but sometimes sharing the same initial hops. The resulting duplication of traffic in the data plane following a failure is very inefficient (exactly the problem that P2MP LSPs are meant to solve). This issue is addressed by draft-ietf-mpls-p2mp-te-bypass which proposes the use of P2MP bypass LSPs, as pictured below.

![Figure 13: P2MP bypass protection](image)

There are three LSPs in this picture.

- The P2MP protected LSP, running from source A to leaves F, G and H.
- The P2MP bypass LSP which reroutes around LSR C, delivering traffic from A to D and E.
- The backup LSP, which is the specific backup for the protected LSP, tunneled through the bypass LSP.

When LSR A detects a failure on the protected LSP it reroutes data down the backup LSP. That traffic is tunneled through the bypass LSP using label stacking.

Tunneling the backup LSP through a P2MP bypass LSP in this way requires the use of upstream-assigned labels as described in section 5.1.1. This is a significant difference to P2P facility fast reroute because the upstream-assigned labels used by the backup LSP will need to be signaled prior to failure and reroute of the protected LSP. (P2P facility fast reroute uses global labels at merge...
points and PHP on bypass LSPs so that data tunneled through the bypass LSP is automatically treated in the same way as data arriving on the protected LSP.)

### 6.1.2 One-to-one (detour) fast reroute

One-to-one fast reroute procedures are discussed only briefly in RFC 4875, with precise specification of how to handle arbitrary topologies left open to future documents.

The basic principles of one-to-one fast reroute for P2MP LSPs are as follows.

- P2P detours are signaled for each S2L sub-LSP by sending detour Path messages specifying the link/node to avoid and the S2L sub-LSP.

- A PLR must try and set up a detour for each S2L sub-LSP passing through it. That may mean that a single PLR has to set up several detours that re-merge with the protected LSP in different sub-trees.

- Detours can merge with each other before re-merging with the protected LSP. The merge rules are unchanged from RFC 4090, implying that detours for different S2L sub-LSPs cannot merge with each other.

- A single detour can protect multiple S2L sub-LSPs by specifying a list of them in the detour Path message. It is unclear whether that also means that P2MP detours are allowed, although the logical implication is that they are.

Merge behavior for P2MP detour LSPs could potentially be very complicated. For example, a single LSR could be a merge point for some S2L sub-LSPs included in a detour Path message, a branch point for others, and a cross-over point for still others. However, merge rules are not yet discussed in any internet draft. Some standardization work will be required before one-to-one fast reroute for P2MP LSPs can be implemented in an interoperable way.

### 6.2 GMPLS protection and restoration

Protection and restoration mechanisms for P2P GMPLS LSPs are defined in RFC 4872 RSVP-TE Extensions in Support of End-to-End GMPLS Recovery and RFC 4873 GMPLS Segment Recovery. These allow traffic to be delivered on a recovery LSP following a failure of the working LSP in a packet or optical network.

RFC 4875 doesn't make any attempt to define procedures for protection and restoration of P2MP GMPLS LSPs, and application of these techniques to P2MP LSPs would therefore require standardization. This section discusses the applicability of the different recovery schemes to P2MP LSPs, and issues that will need to be resolved if and when they are standardized.
6.2.1 End-to-end and segment recovery

For end-to-end (e2e) recovery, the recovery LSP spans the complete length of the working LSP, and is used in the event of a failure at any point along the working LSP. For segment recovery, each recovery LSP spans just a sub-section of the working LSP, and multiple recovery LSPs are required to protect the whole working LSP.

Segment recovery involves more complex topologies, and can be difficult to achieve in an optical network where not all LSRs are capable of wavelength conversion. However, it has the advantages that end-to-end disjoint routes are not required, and recovery LSPs can also be defined by non-ingress LSRs (such as the border node on entry to a new AS).

Both end-to-end and segment recovery offer the same basic recovery types, as detailed below, and so they are dealt with together in this document.

- **1+1 unidirectional protection** uses a dedicated recovery LSP, fully signaled and cross-connected, for every protected LSP. Data is transmitted down both in parallel, and the endpoints are responsible for selecting the best data stream.

- **1+1 bidirectional protection** is similar to 1+1 unidirectional protection, but uses Notifies to co-ordinate simultaneous switchover of traffic in both directions between the endpoints. Because P2MP LSPs are always unidirectional, this scheme is not applicable to them.

- **1:N protection** allows a single recovery LSP to protect multiple working LSPs. When any one working LSP fails, its data is diverted onto the recovery LSP. The recovery LSP may also be used for extra traffic when no working LSPs have failed.

- **Rerouting without extra traffic**, a restoration scheme, uses recovery LSPs which are signaled but not cross-connected prior to the failure. Their resources can therefore be used by other lower priority LSPs, which will be preempted when the working LSP fails and the recovery LSP is needed to divert traffic. Shared mesh restoration is a special case of rerouting without extra traffic which allows the resources for multiple recovery LSPs to be shared in the data plane. When one of the working LSPs fails, its recovery LSP is cross-connected and the shared resources committed to it.

- **Full LSP rerouting**, another restoration scheme, does not even signal the recovery LSP until the working LSP has failed.

Of these techniques, only full 1+1 protection will (if there is sufficient intelligence and buffering at endpoints) be able to provide seamless recovery from failures without packet loss. It is therefore the most relevant mechanism to P2MP LSPs providing broadcast IPTV transport.
6.2.2 1+1 unidirectional protection

In this scheme, a dedicated recovery LSP is fully set up at the same time as each working LSP. The two LSPs are resource disjoint. In the P2P case they can be either unidirectional or bidirectional. However, P2MP LSPs are always unidirectional, so only unidirectional working and recovery LSPs are considered here.

Identical data is sent simultaneously down both LSPs. The egress LSR can choose which stream to use based on its own quality metrics, and switch over and switch back between them without any signaling to the ingress.

An obvious application of end-to-end 1+1 protection to P2MP LSPs is pictured in figure 14 below, where a P2MP working LSP is protected by a resource-disjoint P2MP recovery LSP, and data is sent simultaneously down each.

![Diagram of P2MP 1+1 unidirectional end-to-end protection](image)

**Figure 14: P2MP 1+1 unidirectional end-to-end protection**

There is no reason why the P2P 1+1 unidirectional protection scheme cannot be directly applied to P2MP LSPs in this way. However, it is not yet discussed in any RFCs or internet drafts and would not be guaranteed to be interoperable.

6.2.3 Other recovery schemes

All other protection and restoration schemes require traffic to be diverted from the working to the recovery LSP following a failure, and in some cases the recovery LSP also needs to be signaled first. This makes them less applicable to IPTV, where the resulting packet loss will impact the viewers’ Quality of Experience. However, they may still be valuable in IPTV networks in combination with other high availability mechanisms, or for other applications of P2MP LSPs.

If these schemes are applied to P2MP LSPs, the following issues will need to be overcome.

---

2 Note that for 1+1 protection, working and recovery P2MP LSPs actually only need to be resource-disjoint on a per-S2L sub-LSP basis. Provided the S2L sub-LSPs for a given leaf in each P2MP LSP are resource-disjoint, traffic will always be delivered to that leaf via one of the P2MP LSPs following a resource failure, even if the complete P2MP trees share resources.
1:N protection uses Notify messages to co-ordinate switchover from the working to the recovery LSP. If the working LSP failure was caused by a node which is required to intercept and forward Notify messages (as described in section 3.2.2), then those Notify messages will fail.

Shared mesh restoration uses a PRIMARY_PATH_ROUTE object to allow recovery LSPs to describe the path of the LSPs that they protect, and therefore decide whether they can share resources. Use of shared mesh restoration for P2MP working LSPs will require extensions to that object to allow it to describe P2MP paths.

Fast reroute allows the use of P2MP bypass LSPs whose leaves are not all on the path of the protected LSP. Similar function for GMPLS protection and restoration may be valuable, allowing P2MP recovery LSPs to have leaves that are not on the path of the working LSP, giving increased flexibility in network planning. Again, protocol extensions to allow that would need to be standardized.

6.3 Graceful restart

RFC 3473 defines graceful restart procedures that allow an LSR to reconstruct its control plane state following a restart, provided the data plane (and MPLS forwarding table) remain intact. This is achieved with the help of upstream LSRs who send Path refresh messages that also include the label they use to send data to the restarted LSR, allowing the restarted LSR to reconstruct its own path state and match it to its preserved data plane.

On their own, these procedures are not sufficient to reconstruct P2MP path state. The restarted LSR would also need to be able to reliably regenerate the sub-group identifiers it previously used in downstream Path messages, and correctly map S2L sub-LSPs to preserved out-segments.

However, the graceful restart extensions defined in RFC 5063 solve that problem. These update RFC 3473 graceful restart procedures to allow restarted LSRs to also relearn their path state from downstream LSRs which send them RecoveryPath messages, playing back the Path messages that the downstream LSRs previously received. Those would include the S2L sub-LSP to sub-group mappings needed to fully reconstruct path state at the restarted node.
7. Conclusions

RSVP-TE P2MP LSPs are a good fit to the multicast transport requirements of broadcast IPTV services. Building on the strengths of P2P TE LSPs, they provide:

- high availability, using features such as fast reroute and graceful restart
- guaranteed QoS through resource reservation
- explicit routing, for optimal use of core network bandwidth
- fault diagnosis using LSP ping.

There are other competing technologies. Notably, IP multicast provides a significant subset of this function. However, it cannot perform resource reservation or provide QoS guarantees.

The RSVP-TE protocol extensions for P2MP support are far from being a simple addition to the existing P2P function. In particular, areas like sub-group mapping and ERO compression represent significant challenges for a robust implementation. Complex new function is also needed in the data plane such as incoming data stream selection at re-merge LSRs, and context-specific label spaces.

Finally, there are gaps in the current standards that will need to be resolved before fully interoperable deployments of some important features are possible.

- GMPLS end-to-end and segment recovery procedures for P2MP LSPs have not been defined.
- Standards for the use of the P2MP LSPs to tunnel L2 and L3 VPN data through the provider network are not yet stable.
- Work on PCE extensions to support P2MP routing only began recently.
8. 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.

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.
## 9. Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BFD</td>
<td>Bidirectional Forwarding Detection</td>
</tr>
<tr>
<td>DSLAM</td>
<td>DSL Access Multiplexer</td>
</tr>
<tr>
<td>GMPLS</td>
<td>Generalized MPLS</td>
</tr>
<tr>
<td>IPTV</td>
<td>IP Television</td>
</tr>
<tr>
<td>LSP</td>
<td>Label Switched Path</td>
</tr>
<tr>
<td>LSR</td>
<td>Label Switching Router</td>
</tr>
<tr>
<td>MIB</td>
<td>Management Information Base</td>
</tr>
<tr>
<td>MPLS</td>
<td>Multi-Protocol Label Switching</td>
</tr>
<tr>
<td>P2P</td>
<td>Point-to-Point</td>
</tr>
<tr>
<td>P2MP</td>
<td>Point-to-Multipoint</td>
</tr>
<tr>
<td>QoS</td>
<td>Quality of Service</td>
</tr>
<tr>
<td>QoE</td>
<td>Quality of Experience</td>
</tr>
<tr>
<td>RIB</td>
<td>Routing Information Base</td>
</tr>
<tr>
<td>RSVP</td>
<td>Resource reSerVation Protocol</td>
</tr>
<tr>
<td>S2L sub-LSP</td>
<td>Source-to-Leaf sub-LSP</td>
</tr>
<tr>
<td>SHE</td>
<td>Super Headend</td>
</tr>
<tr>
<td>TE</td>
<td>Traffic Engineering</td>
</tr>
<tr>
<td>VHO</td>
<td>Video Hub Office</td>
</tr>
<tr>
<td>VoD</td>
<td>Video On Demand</td>
</tr>
<tr>
<td>VPLS</td>
<td>Virtual Private LAN Service</td>
</tr>
<tr>
<td>VPN</td>
<td>Virtual Private Network</td>
</tr>
<tr>
<td>VSO</td>
<td>Video Serving Office</td>
</tr>
</tbody>
</table>
10. References

10.1 IETF MPLS P2MP protocol extensions

RFC 4461  
Signaling Requirements for Point-to-Multipoint TE MPLS LSPs

RFC 4875  
Extensions to RSVP-TE for Point-to-Multipoint TE LSPs

RFC 5331  
MPLS Upstream Label Assignment and Context-Specific Label Space

RFC 5332  
MPLS Multicast Encapsulation

draft-ietf-mpls-p2mp-te-mib  
Point-to-Multipoint MPLS TE MIB module

draft-ietf-mpls-p2mp-te-bypass  
P2MP MPLS-TE Fast Reroute with P2MP Bypass Tunnels

draft-ietf-mpls-p2mp-lsp-ping  
Detecting Data Plane Failures in P2MP MPLS - Extensions to LSP Ping

draft-ietf-mpls-remote-lsp-ping  
Proxy LSP Ping

draft-ietf-mpls-rsvp-upstream  
MPLS Upstream Label Assignment for RSVP-TE

draft-ietf-mpls-rsvp-te-no-php-oob-mapping  
Non PHP Behavior and out-of-band mapping for RSVP-TE LSPs

10.2 Other IETF (G)MPLS specifications

RFC 2205  
Resource Reservation Protocol (RSVP)

RFC 3209  
RSVP-TE: Extensions to RSVP for LSP Tunnels

RFC 3473  
GMPLS Signaling RSVP-TE Extensions

RFC 4090  
Fast Reroute Extensions to RSVP-TE for LSP Tunnels

RFC 4206  
LSP Hierarchy with GMPLS TE

Copyright © Metaswitch Networks. Confidential | Page 41
10.3 Related IETF standards

10.3.1 Bidirectional Failure Detection for MPLS

draft-ietf-bfd-mpls  BFD for MPLS LSPs

draft-swallow-mpls-mcast-cv  Connectivity Verification for Multicast LSPs

draft-katz-ward-bfd-multipoint  BFD for Multipoint Networks

10.3.2 VPLS multicast

draft-ietf-l2vpn-vpls-mcast  Multicast in VPLS

draft-raggarwa-l2vpn-vpls-mcast-ctrl  Propagation of VPLS IP Multicast Group Membership Information

10.3.3 L3VPN multicast

draft-ietf-l3vpn-2547bis-mcast  Multicast in MPLS/BGP IP VPNs

draft-ietf-l3vpn-2547bis-mcast-bgp  BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs

draft-ietf-l3vpn-mvpn-considerations  Mandatory Features in a Layer 3 Multicast BGP/MPLS VPN Solution

draft-rosen-vpn-mcast  Multicast in MPLS/BGP IP VPNs

draft-mnapierala-mvpn-rev  Segmented Multicast MPLS/BGP VPNs
10.3.4 Expired drafts for P2MP leaf discovery

draft-leroux-mpls-p2mp-te-autoleaf  IGP Routing extensions for discovery of P2MP TE LSP Leaf LSRs

draft-zheng-mpls-p2mp-topology-agent-reqs  MPLS P2MP Topology Agent Requirements

10.3.5 PCE extensions for P2MP

draft-ietf-pce-p2mp-app  Application of the PCE to P2MP MPLS and GMPLS TE

draft-ietf-pce-p2mp-req  PCC-PCE Communication Requirements for P2MP MPLS TE

draft-ietf-pce-pcep-p2mp-extensions  Extensions to PCEP for P2MP TE LSPs

10.4 IPTV network architecture standards

FG IPTV-DOC-0189  IPTV Network Control Aspects
ITU-T IPTV Focus Group document

FG IPTV-DOC-0190  IPTV Multicast Frameworks
ITU-T IPTV Focus Group document

ATIS-0800007  IPTV High Level Architecture
ATIS IPTV Interoperability Forum

TR-101  Migration to Ethernet-Based DSL Aggregation
DSL Forum Technical Report

Open IPTV Forum Functional Architecture v1.1

10.5 Miscellaneous

Handling Broadcast IPTV Content  Light Reading Report

Multicast in MPLS/VPLS Networks  IP/MPLS Forum Tutorial

VPN Technologies – A Comparison  Metaswitch White Paper

IP Multicast Explained  Metaswitch White Paper