AALTO University

School of Electrical Engineering

Mikko Kinnunen

Signalling of Point to Multipoint Trees in Metro Ethernet and Core Networks

Master's Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science.

Espoo, October 17th 2011.

Supervisor Professori Raimo Kantola

Instructor Diplomi-insinööri Jorma Romo, TeliaSonera

AALTO YLIOPISTO Diplomityön tiivistelmä

Tekijä: Mikko Kinnunen

Työn nimi: Signalling of Point to Multipoint Trees in Metro Ethernet and Core Networks

Päivämäärä: 17.10.2011

Sivumäärä: viii + 75

Tiedekunta: Sähkötekniikan korkeakoulu
Professuuri: Tietoliikenne- ja tietoverkkotekniikan laitos

Työn valvoja: prof. Raimo Kantola
Työn ohjaaja: Jorma Romo, Diplomi-insinööri

Diplomityössä tutustuttiin IPTV-kanavien siirtoon Core-verkosta MetroEthernet-verkon asiakasta lähinnä olevalle laidalle. Tavoitteena oli kehittää nopeampi ratkaisu monilähetyspuiden konfigurointiin laitevalmistajan toteuttamilla protokollilla. Nykyinen ratkaisu, jossa käytetään Resource reSerVation Protocol:ia MultiProtocol Label Switching-tunneleiden signaloimiseen, Internet Group Management Protocol Snooping:ia halukkaiden vastaanottajien kartoittamiseen sekä Protocol Independent Multicast-Source Specific Multicast:ia runkoverkon monilähetykseen on liian työläs.

Uudet ratkaisut, joissa yhdistellään RSVP:tä, point-to-multipoint RSVP:tä, Fast ReRoutea ja PIM-SSM:ia testataan TeliaSoneran tietoverkkolaboratoriossa.

Tulosten perusteella ei voida sanoa paljoa varmasti, mutta FRR ME-verkossa vaikuttaa helppokäyttöiseltä ja toimivalta ratkaisulta. Lisäksi P2MP RSVP-TE herätti toiveita nopeammin vikatilanteista toipuvasta monilähetysratkaisusta runkoverkosta, kunhan ilmenneiden vikojen syyt saadaan selville.

Avainsanat: monilähetys, MPLS, point-to-multipoint, RSVP, IPTV

AALTO UNIVERSITY Abstract of the Master's Thesis

Author: Mikko Kinnunen

Name of the thesis: Signalling of Point to Multipoint Trees in Metro Ethernet and Core Networks

Date: 17.10.2011

Number of pages: viii + 75

Faculty: School of Electrical Engineering
Professorship: Communications and Networking

Supervisor: prof. Raimo Kantola
Instructor: Jorma Romo, M.Sc(Tech.)

This master's thesis studies the distribution of IPTV channels from a core network to the edges of a MetroEthernet network. The goal is to find a faster solution for configuring multicast trees using protocols implemented by vendors. The current solution which uses Resource reSerVation Protocol for signalling MultiProtocol Label Switched tunnels, Internet Group Management Protocol Snooping for mapping receivers and Protocol Independent Multicast-Source Specific Multicast for core multicast creates too much work.

The new solutions combine RSVP, point-to-multipoint RSVP, Fast ReRoute and PIM-SSM and they are tested in the TeliaSonera networking laboratory.

Based on test results there is not much certainty about many things but it can be said that FRR seems to be working well and it is easy to use. Furthermore, P2MP RSVP seemed promising for the core network with faster convergence times in failure cases than PIM-SSM. However, there are few problems to be solved before the protocol is ready for use in the production network.

Keywords: Multicast, MPLS, point-to-multipoint, RSVP, IPTV

Preface

I would like to thank Professor Raimo Kantola and Jorma Romo for all the instructions. I would also like to thank Pertti Mikkonen, Arto Hokkanen, Kari Nyman, Hannu Lundgren and Sami Minkkinen for all the help. And last but not least Ismo Eskel and my wife Laura for having patience.

17.10.2011, Espoo, Finland

Mikko Kinnunen

Signalling of Point to Multipoint Trees in Metro Ethernet and Core Networks

Table of Contents

Acronyms

ACK	ACKnowledgement
AS	Autonomous System
ASM	Any Source Multicast
ATT bit	ATTach bit
BCD	Binary Coded Decimal
BRAS	Broadband Remote Access Server
CIDR	Classless Inter-Domain Routing
CLNP	ConnectionLess Network Protocol
CSMA/CD	Carrier Sense Multiple Access with Collision Detection
CPU	Central Processing Unit
DM	Dense Mode
DSLAM	Digital Subscriber Line Access Multiplexer
EGP	Exterior Gateway Protocol
ERO	EXPLICIT_ROUTE Object
FDB	Forwarding DataBase
FE	Fast Ethernet
FEC	Forwarding Equivalency Class
FIB	Forwarding Information Table
FRR	Fast ReRoute
GigE	Gigabit Ethernet
GMQ	General Membership Query
GSQ	Group Specific Query
(S)HE	(Super) Head End
IEEE	Institute of Electrical and Electronics Engineers
IGMP	Internet Group Management Protocol
IGP	Interior Gateway Protocol
IIH	IS-IS Hello
IOM	Input/Output Module
IP	Internet Protocol
IS-IS	Intermediate System to Intermediate System
LAN	Local Area Network
LSDB	Link State DataBase
LSP	Label Switched Path
LSP	Link State PDU
MAC	Media Access Control
MDA	Media Dependent Adapter
MDT	Multicast Distribution Tree
MEN	MetroEthernet Network
(m)LDP	(multicast) Label Distribution Protocol
MPLS	MultiProtocol Label Switching
NET	Network Entity Title
NSEL	NET SELector
OSI	Open Systems Interconnection
(O)SPF	(Open) Shortest Path First
OUI	Organizationally Unique Identifier
P2MP	Point-to-MultiPoint
PDU	Packet Data Unit
PIM	Protocol Independent Multicast

PPP	Point-to-Point Protocol
QoS	Quality of Service
RFC	Request For Comments
RGW	Residential Gateway
RIB	Routing Information Base
RP	Rendezvous Point
RRO	Record Route Object
RSVP-TE	resource ReSerVation Protocol-Traffic Engineering
SM	Sparse Mode
SSM	Source Specific Multicast
S2L	Source-to-Leaf
SERO	P2MP_SECONDARY_EXPLICIT_ROUTE Object
SNMP	Simple Network Management Protocol
SPT	Shortest Path Tree
SRRO	P2MP Secondary Record Route Object
STB	Set-Top-Box
TCP	Transmission Control Protocol
TED	Traffic Engineering Database
TLV	Type Length Value
TSF	TeliaSonera Finland
UDP	User Datagram Protocol
VPLS	Virtual Private LAN Service

1 Introduction

1.1 Background

Today, there are multiple ways of sending traffic from a source to receiver(s) in packet data networks. The traditional way is called unicast where the traffic is sent from the source to only one specific receiver. Another one is to broadcast traffic from a source to all receivers in the domain. The broadcast traffic can be e.g. signalling traffic or conventional distribution of TV channels. The problem with broadcast is that not all receivers want to receive it and therefore bandwidth and other resources are wasted. One solution to this is multicast. In multicast communication there is a source and multiple receivers who want to receive the traffic from the source. Only one copy is sent from the source regardless of number of receivers. When the traffic comes to a point in the network where all the receivers can not be reached by that one copy, additional copies are created to reach each receiver. Multicast can save a significant amount of valuable bandwidth and therefore enables some bandwidth consuming applications, such as Internet Protocol TV (IPTV). IPTV would not scale well enough without multicast in the networks of today.

TeliaSonera Finland (TSF) has widely deployed multicast for distribution of cable TV and IPTV. Receivers join the multicast groups (one multicast group equals one TV channel) by Set-Top-Box (i.e. digibox) and Residential GateWay (RGW). The receiver selects channels by sending a Join message for a group. The first point where the receiver can join a multicast group is called a Digital Subscriber Line Access Multiplexer (DSLAM). The DSLAM aggregates all the receivers from a certain area to a trunk link to a MetroEthernet Network (MEN). If another receiver attached to the same DSLAM has already joined the same group, the DSLAM can add the new receiver to the group and replicate the channel. If the receiver is the first to request the channel, the DSLAM forwards the Join message towards the network i.e. upstream. The Join message is forwarded further upstream by other devices until the group is found. All the groups are brought to Broadband Remote Access Server (BRAS) gateway which means that it is the furthermost place to join a group. The BRAS is not the source of the groups, but is in the same network domain with the source. The source, Super Head End (SHE) has all the channels from different content providers. The SHE is connected to a core router.

All the paths from the SHE to the receivers form a multicast tree. The traffic traverses over the core network to the BRAS gateway and is the first part of the whole multicast tree. The data continues to traverse the TSF network over the MEN, which is the second part of the multicast tree. The first part of the tree is constant i.e. it does not depend on the amount of users. In MEN, the paths to the receivers are implemented after the receivers have signed an IPTV service contract.

ME switches are configured by network implementation staff and they also configure the multicast trees and backup paths statically. Because the MEN and number of receivers are constantly growing, updating the multicast trees, especially in case of topology changes of the underlying physical network, is time consuming and redundant compared to dynamic trees. Also in case of failures in MEN the multicast trees may recover in a sub-optimal way.

1.2 Goals and objectives

The goal of the thesis is to find a better solution for the multicast implementation at MEN and study new possibilities for implementing multicast in core and MEN combined. Today the ME switches do not support Protocol Independent Multicast (PIM) which is already implemented at the core network. Furthermore, studying different possible points for the Point to MultiPoint (P2MP) Label Switched Path (LSP) root is an important part of the thesis. Ideally the multicast trees would be dynamic i.e. adapting to changes in the network. Furthermore, the configuration of trees should be at least partially automatic and work only with one multicast protocol to make the implementation as easy as possible. However, if such option does not exist, other options should be examined. Theoretically feasible solutions are then tested in a laboratory environment and the results are evaluated. The test environment is a TSF laboratory where the TSF network is built in a smaller scale.

1.3 Scope

The thesis evaluates the results on the basis of technology, time and Quality of Service (QoS). Economical aspects of the solutions are not considered. The solutions are evaluated in regard to the ease of configuration and resiliency with the main emphasis on the ease of configuration. Alternative solutions which would require a new type of equipment or software from a new vendor are out of scope (e.g. Path Computation Element).

2 Network architecture

Networks have developed into a very complex entity and it may be hard to understand which part of the network does what. That is why functionality of the network is distributed into different layers and within layers into different protocols. This chapter introduces network devices, addressing and protocols.

2.1 Network components

2.1.1 Transmission Control Protocol/Internet Protocol

"Transmission Control Protocol/Internet Protocol (TCP/IP) is the protocol suite used for communication between hosts in most local networks and on the Internet. [1]" The purpose of TCP/IP is to create modularity into complex networks. TCP/IP is divided into four layers which can be seen in Table 1.

Layer	Purpose of the Layer
Application layer	Defines the applications used to process requests and the ports and sockets used
Transport layer	Defines the type of connection established between hosts and the way acknowledgements are sent
Internet layer	Defines the protocols used for addressing and routing the data packets
Link layer	Defines the ways the hosts connect to the network

Table 1: TCP/IP layers and their purpose

The Application layer is the highest level on the TCP/IP model. RFC 1122 divides Application layer protocols into two categories: "user protocols that provide service directly to users, and support protocols that provide common system functions [2]". The user protocols include e.g. Telnet and support protocols include e.g. SNMP (Simple Network Management Protocol).

"The Transport layer provides end-to-end communication services for applications [2]." Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) are the most widely used transport layer protocols, the former for reliable and the latter for faster but unreliable transport.

The most used protocol on the Internet layer is Internet Protocol (IP), which connects networks created with different technologies. It is designed to carry data from source to destination. IP is not a reliable protocol but the protocols on the layers above can provide reliability when needed.

The Link layer is also known as the media-access layer or network interface layer. The Link layer, as the name implies, defines the local network technologies and the links between two hosts.

The TCP/IP has gained popularity over Open Systems Interconnection (OSI) model which has 7 layers instead of 4. The OSI model was the basis of the networks for the last few decades even though TCP/IP model is older.

2.1.2 Network devices

The different parts of the TSF network consist of devices with different capabilities. As mentioned in Chapter 1 the network includes the core network which connects the MENs. The MEN belongs to the aggregation network. The DSLAM that was mentioned too is part of the TSF access network.

A router is an Internet layer device that provides connectionless data transfer. It receives and calculates information of a network. That information lies in a Routing Information Base (RIB). It has prefixes of sub-networks mapped to interfaces. The router runs a lookup to forward data packets to wanted sub-networks. This procedure is repeated in every router until the packet reaches the receiver. Nowadays routers are divided into two planes, control and forwarding plane. The control plane sends and receives prefixes from neighbouring routers and calculates the cost to reach any sub-network. The forwarding plane runs the lookup and sends the packets to the right interface. The routing procedures are examined in more detail in Sections 2.4, 2.5 and 2.6. The core network consists of routers.

A switch is a Link layer device. It receives frames from one interface and sends it out from another. It has an address table it has learned and does the switching according to that. Switches work only on a relatively small area because of switching table maintenance problems. The MEN devices are somewhere between switches and routers e.g. they support a routing protocol but it is not used for routing customer packets.

A DSLAM (Digital Subscriber Line Access Multiplexer) is an access network device which collects subscriber xDSL connections together and multiplexes them into a FastEthernet (FE) or GigabitEthernet (GigE) link. The FE/GigE link is connected to the MEN. In IPTV content distribution the DSLAM is responsible for the “last mile”.

A Super Head End (SHE) combines Cable TV (CTV) and IPTV to a single TV-service. The SHE is a localized, closed environment. Among other things it consists of packing and unpacking of channels, scrambling of Pay-TV channels, text and audio managing. It produces IP multicast (see 2.9) streams which are taken out of multicast gateways located between the SHE and the core network which is roughly presented in Figure 1. The core network is used to distribute multicast from the area of SHE to other areas.

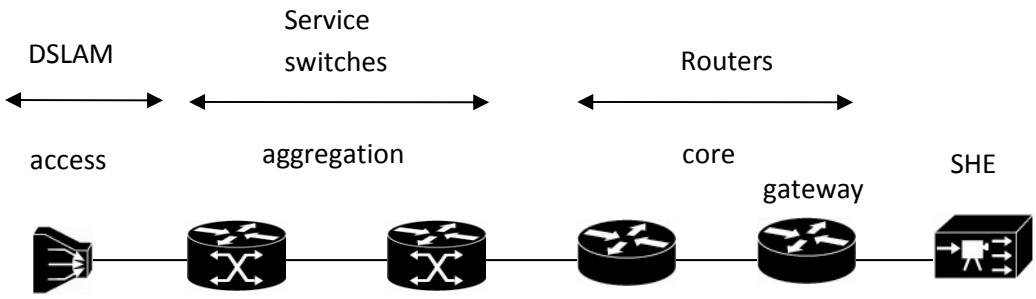


Figure 1: Operator network devices involved in IPTV distribution

An N2X emulator is used at later stages of the thesis in a laboratory environment to emulate SHE and DSLAM. When N2X functions as SHE, it produces data streams at a set packet rate. When functioning as DSLAM it sends multicast Join (2.10 & 2.11) messages and receives data streams. The operating system allows a user to see the sent packet rate, whether the stream flows end-to-end, packet loss etc.

2.2 Addressing

In telecommunications, every device must have an address so that the devices can be identified. The addresses must be unique within a domain. There are different address families for different layers.

2.2.1 Internet layer

There are two versions of IP, IPv4 and IPv6. In TSF network, mainly IPv4 is used so IPv6 is not discussed in this thesis.

An IP address is of a fixed length of 32 bits i.e. 4 bytes but usually the IP address is presented in a decimal notation for easier reading. IP addresses used to have a strict form which included a network part and a so called rest part. Nowadays the network and host part are separated by a network mask of variable length. This is called Classless Inter-Domain Routing (CIDR). The network mask is a string of 1's which is often written in a decimal notation e.g. 24 ones is 255.255.255.0 as in Figure 2. It can also be stated by a prefix length at the end of the IP address e.g. 192.145.211.135/24. This makes it possible to refer to a group of addresses with a single address and the prefix length. This is very practical feature as can be seen e.g. in IS-IS section discussing route advertising. The size of the group can be $2^{32 - \text{prefix length}}$.

Some address blocks have been reserved by Internet Assigned Numbers Authority (IANA) [3] for special use e.g. multicast block is 224.0.0.0/4. The multicast addresses do not use network masks but are unique 32-bit addresses that represent a group of receivers.

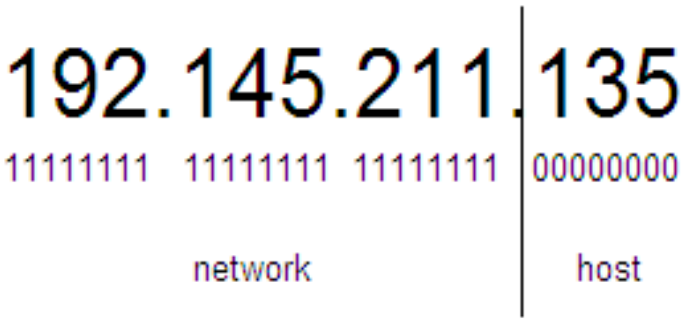


Figure 2: Hierarchy of an IP address

2.2.2 Data link layer

The Open Systems Interconnection (OSI) addressing behaves exactly in the same way as IP addressing when all the router interfaces are left unnumbered. A Network Entity Title (NET) is assigned to each router. NET is also known as the OSI address even though technically NET is a subset of the OSI address. NET consists of an Area-ID, a System-ID and a NET Selector (NSEL) as shown in Figure 3.

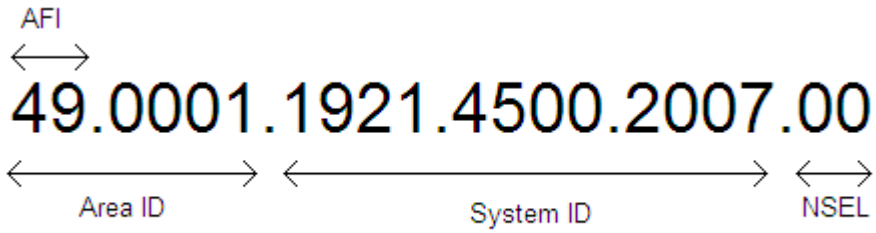


Figure 3: Explanations for the NET components

The Area-ID is a variable part of the NET and it ranges from 1 to 13 bytes in length. The first byte, Address Family Identifier (AFI), tells how to interpret the rest of the Area-ID. The System-ID is a unique identifier of a node. It is practically 6 bytes long (for some reason other lengths are theoretically also possible) and it is converted from an IP address. The most common conversion method is Binary Coded Decimal (BCD) encoding where an IP address e.g. 192.145.2.7 is encoded to 1921.4500.2007. The zero NSEL means “this system”. A more detailed explanation is given in Section 2.6 IS-IS.

Every device on an Ethernet has a locally unique Media Access Control (MAC) address which is a so called physical address. It is of a fixed length of 48-bits (6 bytes) e.g. 12-34-56-78-9A-BC. The first three bytes of the MAC address called Organizationally Unique Identifier (OUI) indicate the manufacturer (or is set to 01-00-5e in case of IP multicast) and the remaining bytes are called a Network Interface Controller (NIC). The NIC indicates the serial number of the device. A mapping between IP and MAC addresses is maintained in an Address Resolution Protocol (ARP) table.

Mapping IP multicast address to MAC address is problematic because the IPv4 address is 32 bits long and the variable part of the MAC address is only 24 bits long. Luckily the multicast IP address has always the same 4 most significant bits so that leaves only 28 bits to map. However, when mapping a multicast address, the first bit after the OUI is set to 0 to indicate multicast. So, only 23 bits remain to be used for multicast mapping.

This results in having 32 (28 - 23 = 5 and 2^5 = 32) IP addresses for every MAC address which is shown in Figure 4. Having 32 IP addresses for every MAC address is not the ideal situation but it can be managed by good planning. In addition, Source Specific Multicast (see 2.12) eases the problem or even solves it in practice.

Multicast MAC address in binary notation	00000001	00000000	01011110	0xxxxxxx	xxxxxxxx	xxxxxxxx
MultiCast MAC addr. in hexadecimal notation	01	00	5e	0 23 bits available for mapping		
Multicast IP address in binary notation			1110**1111**	**0**0000001	00000001	00000001
Multicast IP address			239	01	01	01
Mapped multicast address	00000001	00000000	01011110	00000001	00000001	00000001
Mapped multicast address in hexadecimal	01	00	5e	01	01	01

Figure 4: Mapping an IP multicast address to MAC address. The available bits for mapping are marked with x's on the top row. The second row shows the address in hexadecimal notation. The next two rows present the multicast address to be mapped in two notations. The final two rows present a mapped multicast address in binary and hexadecimal notations.

In Figure 4 the bits marked with x's on the top row are the 23 bits available for IP multicast mapping to MAC. The bits 5-9 (bolded 1111 0) on the third row are the 5 bits that are lost. The four bits in front of the lost bits are always the same so they can be ignored.

2.3 Ethernet encapsulation

2.3.1 802.3 frame

The most popular LAN and MAN solution is "employing Carrier Sense Multiple Access with Collision Detection (CSMA/CD) as a shared media access method and IEEE 802.3 (Ethernet) protocol and frame format for data communication" [4].

Preample (7 octets)
SFD (1 octet)
Destination address (6 octets)
Source address (6 octets)
Length/Type (2 octets)
Client data (46 to 1500 or 1504 or 1982 octets with padding)
Pad
FCS (4 octets)
Extension

Figure 5: 802.3 Frame inside 802.3 packet

The Ethernet/802.3 frame consists of destination and source address, 6 bytes each, length or type field which indicates the length of the frame or the type of the payload, client data and Frame Check Sequence (FCS) which is a cyclic redundancy check. The frame is inside a packet which encapsulates the Ethernet frame with preample, Start Frame Delimiter (SFD) and extension fields. The preample is used for synchronization,

SFD marks the starting point of the frame and the extension field is only used in case of 1Gb half duplex operation. This whole structure is presented in Figure 5.

2.3.2 Virtual LAN

Virtual LAN (VLAN) is specified by IEEE 802.1q. VLAN differentiates traffic on a LAN by tagging the 802.3 frames. There are three kinds of tagging. NULL tag means that tags are treated as user data and it is transparent to the network. Dot1q tag differentiates traffic on a LAN and only frames with equivalent tag are forwarded. Finally, QinQ tagging differentiates traffic with two tags and only frames with equivalent tag pair or outer tag are forwarded. [5, 6]

2.3.3 Virtual Private LAN Service

A Virtual Private LAN Service (VPLS) is a multipoint Data Link Layer Virtual Private Network (VPN) (see [7]) service which from users' perspective looks like a LAN which is used only by a given set of users. The users can be in different networks as long as the networks are connected. The VPLS is made private by adding a service ID to the encapsulated 802.3 frame. The tagged frames are then forwarded over an IP/MPLS (or Generic Routing Encapsulation, GRE [8]) network using a tunnel to another user. The IP/MPLS network is introduced later.

2.4 Routing

As mentioned in Section 2.1.2, routers make decisions based on the prefixes in the RIB. The prefixes can be inserted in to the RIB in two ways, manually or by routing protocol. The manual insertion is done by static routes which do not change even if there are changes in the network. This can cause e.g. routing loops. To avoid the problems caused by static routing, dynamic routing is used.

There are two kinds of route calculation algorithms, link-state and distance vector. The distance vector algorithm is run in every router and each router has only a limited view of the network. Routers have information on their neighbours only. In contrast the link-state algorithms have a full view of the topology. This is of course a heavier solution but it does provide more efficient routes to destinations. In this thesis, only link-state routing protocols are considered. The reason for dividing the router into different planes is the strain that link-state algorithm puts on the router. In addition, this way a router is functional in case of a network convergence and traffic peak at the same time.

There are two kinds of routing protocols, Interior Gateway Protocols (IGPs) and Exterior Gateway Protocols (EGPs). IGP is responsible for routing inside an AS and the range of different IGPs is wide. This thesis however, only introduces IS-IS, because other routing protocols do not concern the topic of the thesis. EGPs take care of routing between ASs and it is often called internet routing. EGPs are a small group of protocols and at the moment Border Gateway Protocol (BGP) is practically the only one used. EGP is not used in the solutions in the thesis so it is not discussed further.

2.5 Shortest Path First (SPF) algorithm

The best known link-state algorithm is the Shortest Path First (SPF). It is also known as Dijkstra algorithm by its developer E.W. Dijkstra. SPF algorithm "find(s) the path of minimum total length between two given nodes (S) and (D). [9]" All nodes are connected by branches.

In the course of the solution the nodes are divided into three groups:

A. Nodes are added to this group in order of increasing minimum path length from node S.
B. All the nodes that are connected to at least one node in group A but are not part of the group A. The nodes in group A are from this group.
C. The remaining nodes

The branches are divided into three groups as well

i. The branches occurring in the minimal paths from S to all the nodes in A.
ii. The branches of group i are selected from this group. Only one branch in this group leads to each node in group B.
iii. The rejected or not yet considered branches.

In the beginning all the nodes and branches are in groups C and iii. First, the node S is moved to group A and the following steps are repeated until the shortest path to D is found.

Step 1.
Consider all the branches connected to node S which were just moved to the group A. If node T belongs to group B it is examined if branch t results a shorter connection from S to T than the current branch. If so, the current branch is rejected and the branch t is placed into group ii. If the node R is part of group C it is placed into group B and the branch t is placed into group ii. [9]

Step 2.
If groups i and ii are the only ones considered, there is only one way of connecting the nodes in B to node S. The node with minimum distance to S is added to group A and the corresponding branch is moved to group i. These steps are then repeated until node D is moved to group A. This algorithm works even if the length of a branch is different in different directions between two nodes. [9]

2.5.1 Constrained Shortest Path First

Constrained Shortest Path First (CSPF) is an advanced version of SPF. It is used for calculating shortest paths for Label Switched Paths (LSPs), which are introduced in Section 2.7, in a more detailed way by excluding paths according to constraints. CSPF uses Traffic Engineering Database (TED) information which is provided by IS-IS (2.6) extensions. The constraints or attributes include but are not limited to bandwidth requirements, hop limitations and reservable bandwidth of the links i.e. non-reserved bandwidth on a link.

The CSPF algorithm does the following procedures:

1) Collect attribute information on every link

2) Flood the attributes to other nodes using IS-IS extensions
3) The links are grouped according to attributes. Combine each node to the groups. Form a topology of each protection entity and related link attribute information.
4) Calculate constrained paths for the network

The algorithm uses two databases: PATHS and TENT. The TENT database has the tentative nodes that have been tried before finding the shortest path. The information is moved to the PATHS database only when the shortest path to a node has been found.

The calculation of the constrained paths is done in the following steps:

1) A node doing the calculation is put into the PATHS and TENT is pre-loaded from the local database.
2) All the neighbouring nodes from the node in PATHS are examined. If a node is in TENT and if the path is shorter than a path in PATHS the old one is replaced by the new one. Paths may also be equal length. If a neighbour node is not in TENT, all the links and nodes that do not match the constraints are deleted.
3) The nodes with the least cost are moved from TENT to PATHS.
4) When TENT is empty or the destination node is reached, the calculation is completed.
5) Equal cost paths are chosen based on a policy such as random selection, maximum remaining bandwidth rate of path first and minimum remaining bandwidth rate of path first. [10]

2.6 Intermediate System to Intermediate System

2.6.1 Introduction

In IS-IS language an intermediate system means a router so the protocol name could be translated router-to-router. It is a link-state Interior Gateway Protocol (IGP) i.e. protocol for intra-AS routing. It is the foundation of the network and of the topic of the thesis. It calculates all the paths the data has to traverse, keeps track of topology changes and distributes that information to its neighbours. Finally, it combines different parts of the network by providing a common language and rules.

Many people are more familiar with other IGPs, so it might be easier to explain IS-IS by looking at how IS-IS is different from other IGPs, such as Open Shortest Path First (OSPF). IS-IS runs natively on Link layer i.e. it does not need valid interface addressing to transmit a message. It is suitable for routing multiple protocols. It is totally agnostic about what kind of prefixes (e.g. IPv4, IPv6, CLNP) it transports in its message. It is an independent protocol and finally, it is usually used in large service provider networks such as the TSF network.

IS-IS only understands two kinds of interface types: point-to-point (P2P) and broadcast. The most common encapsulations for these are IEEE 802.3 and Point-to-Point Protocol (PPP) encapsulation.

2.6.2 Areas

Link-state protocols usually have worked with one set of routers which form a network. Every router has to have a complete picture of the network for path calculation

purposes. It is obvious that this is a major threat for scalability. The IS-IS protocol developers solved the problem by constructing the network from smaller areas and therefore topological horizon became smaller for IS-IS routers. Of course, the available IP prefixes or reachability information, which is an IS-IS term, needs to be injected into other areas at the area borders but the amount of routes and CPU demand is reduced. The reachability information is advertised in summary routes e.g. an area which consists of prefixes 172.16.1/24 – 172.16.4/24 sends a summary route 172.16/16 which is illustrated in Figure 6.

IS-IS uses Link State Packet data units (LSPs) to transport information between routers. LSPs are like envelopes that can be used for transportation of information such as IP reachability information, checksums and state of links. The most important elements of the LSP header are Lifetime, Sequence Number, LSP-ID and Checksum.

Lifetime indicates how many seconds the LSP is valid and therefore makes sure that the state information is fresh enough. The maximum lifetime is about 18 hours and that is the longest time an IS-IS router would have to wait to start the sequence numbering again from 1. The sequence number is used for identifying the newest LSP. That way contradicting LSPs can be ranked according to the freshness. The LSP-ID determines the LSP type. It consists of the System-ID (6 bytes) of the LSP originator, a pseudonode-ID (1 byte) which indicates if the node is a real router and not a so called pseudonode i.e. a router representing a LAN. Finally, there is the Fragment-ID which is used for fragmentation support.

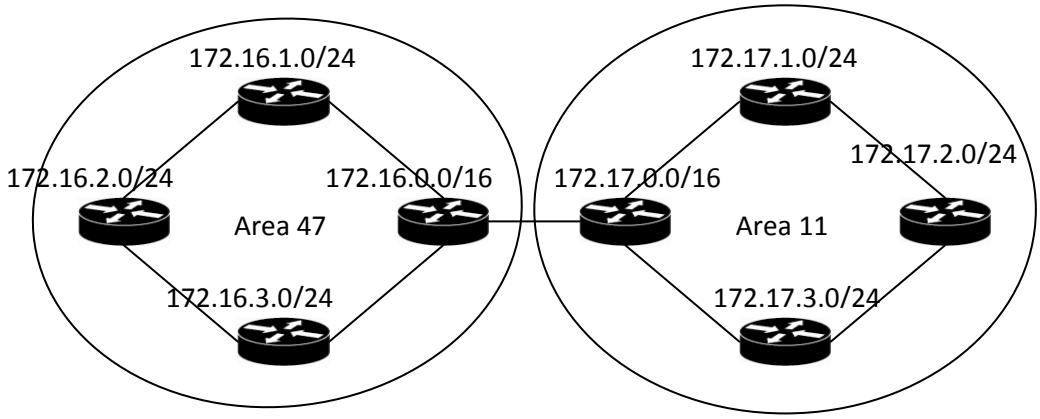


Figure 6: Summarizing prefixes at area border router. Adapted from [11].

The difference between IS-IS and OSPF is that in IS-IS the area border is between routers i.e. there is a distinction between area boundaries and routing hierarchy levels (see Figure 7). A level is a tool for creating routing hierarchy in IS-IS and that is discussed in the next section.

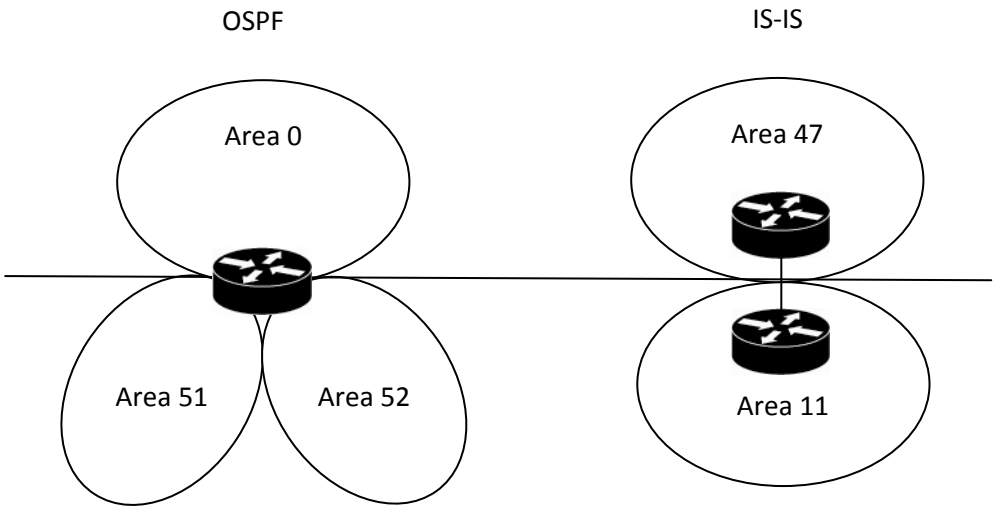


Figure 7: Differences of IS-IS and OSPF routers on area borders. Adapted from [11].

2.6.3 Levels

In IS-IS the levels are indicated by tags. As illustrated in Figure 8, each link is at Level-1 or Level-2 or both and the tags indicating this are L1, L2 or L1L2. The links do not have to have a matching Area-ID on both ends like in OSPF. However, all routers participating in L1 topology have to share an Area-ID, otherwise no adjacencies will form. For L2, the only constraints are that the L2 topology is contiguous and no L2 routers are isolated from any others.

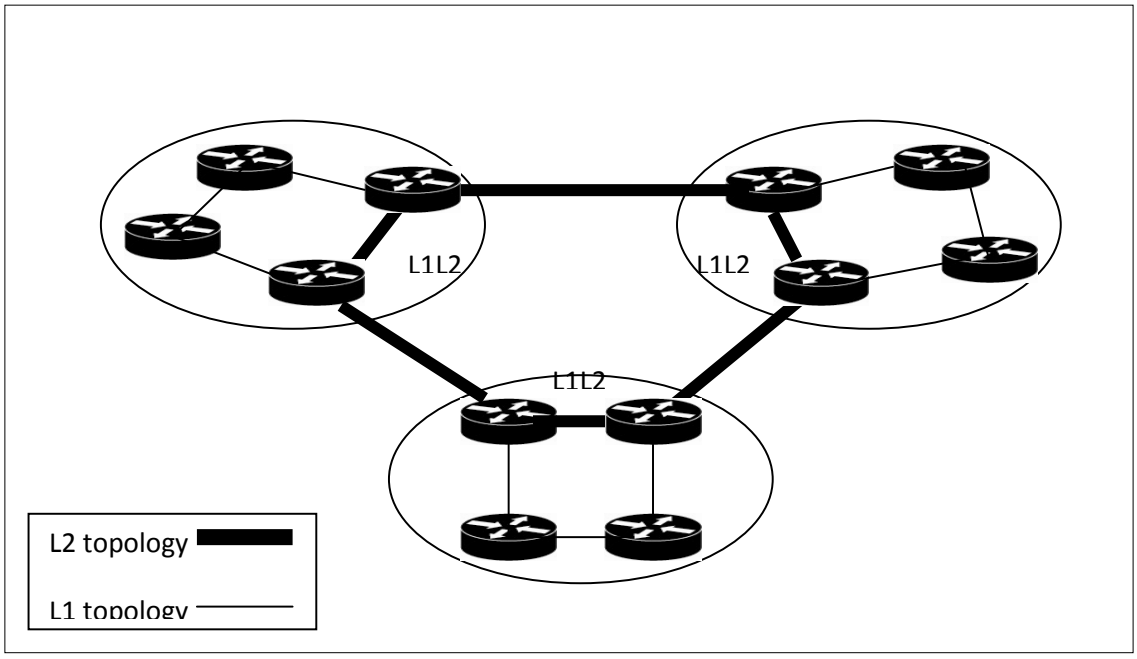


Figure 8: Levels L1, L2 and L1L2

IS-IS leaks information from L1 to L2 but not vice versa. Any router which is part of L2 topology sets an Attach (ATT) bit on its routing messages (Figure 9). The routers in areas calculate their shortest path to the closest router which has sent messages with the ATT bit and installs a default 0/0 route in its routing/forwarding table pointing to the closest L1L2 router.

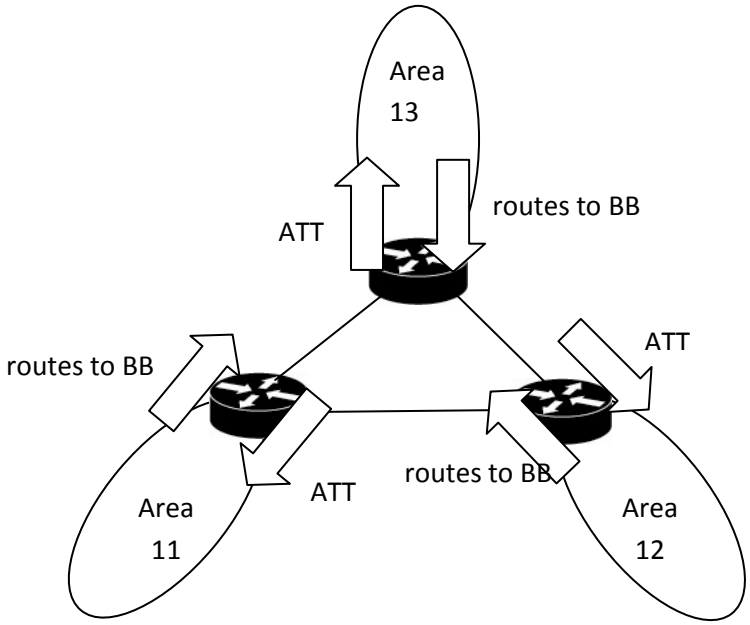


Figure 9: IS-IS information leaking. Routes are leaked to backbone (BB) i.e. from L1 to L2 but not vice versa. Adapted from [11].

Again, in contrast to OSPF, IS-IS router can be a part of multiple areas at the same time. This is necessary when merging, splitting or renumbering areas. In an IS-IS Hello (IIH) message, there is room for multiple Area-IDs and with this option, area migration is done without any maintenance window.

2.6.4 Local SPF computation

IS-IS follows a "distributed databases, local computation" principle. It means that routers decide which routers belong to a network and how the routers are connected to each other. Each router then receives the same network topology information and prefixes unaltered. All the information is stored in Link State DataBase (LSDB) with L1 and L2 having their own dedicated LSDBs. After receipt of all LSPs in a given IS-IS network the router calculates the shortest path with the SPF algorithm (2.5) for a given prefix through the network.

2.6.5 Shortest Path First and route calculation in IS-IS

An IS-IS router uses SPF algorithm to calculate loop-free paths in a network. In order to change traffic flow, the following steps need to be taken for each route SPF calculation, Route resolution and Prefix insertion.

The SPF calculation is executed based on information received in LSP. After that, in the route resolution part, the router needs to know about dependent routes. The dependent

routes are affected by topology change supplied by IS-IS. “In an Internet environment with full routing tables, finding the dependent routes is one of the most dominating factor in the total route-calculation period. [11]“ After determining the influenced dependent routes, the prefix insertion follows. It consists of deleting or changing the old prefixes and downloading the new ones to the line cards.

The SPF is based on LSDB and maintains three lists (UNKNOWN, TENTative, PATHs) to calculate the shortest path to every N node. First, all the nodes in the LSDB are moved to UNKNOWN list. The local router doing the SPF calculation puts itself into the TENT list with all the direct next-hop routers. The calculation takes at the most N loops. The loop has the following steps which form a simplified version of the SPF presented in Section 2.5:

1) Find the node with the lowest cost to the SPF calculating router and move it to the PATH list.
2) Find every next-hop from the destination node in the PATH list and move them into the TENT list but…
3) A two-way mutual connection must be verified or the adjacency is ignored.
4) Of all the adjacencies the lowest cost to the root (i.e. the SPF calculating router) is moved to the PATH list. Also the next-hop cost is saved because the forwarding engine works with the next-hop information.

2.6.6 Type Length Value (TLV) and Sub-TLV

Routing protocols must be extensible i.e. able to make developments without disturbing its function. IS-IS uses Type Length Values (TLVs) to encode necessary information for the IS-IS routing protocol. TLV has three fields, type, length and value, as the name implies. They inform a receiver what to read and how much. The type field is a 1- byte code which tells what to read. The length field of 1 byte tells how much to read and the value is the “payload” e.g. a prefix. The value can be 1-255 bytes in length.

Every new message element needs a dedicated TLV and this can exhaust the TLV space very quickly. That’s why sub-TLVs are used. It is an extra encapsulation inside the TLV.

2.6.7 IP reachability information

As stated earlier, IS-IS uses TLVs to distribute prefixes which are called IP reachability information in IS-IS language.

IETF defines two wide-metrics TLVs, Extended IS Reachability TLV #22 and Extended IP Reachability TLV #135. These are new TLVs and the reasons for defining them were limitations of old TLVs on metric space and information to an adjacency. Now in the Extended IS Reachability TLV #22 there is a 24-bit metric space and a possibility for additional information about the link with sub-TLVs even though IS-IS can only express an adjacency. The metric field expresses the preference of using that link and typically it is calculated according to an inverse bandwidth of the link. There is a value called the Reference Bandwidth which is divided by the interface bandwidth to yield the metric. The Reference Bandwidth should be set so that it is likely to be useful for the next ten years. It can not be too big either because most routing protocols have

limited Metric fields. There is also a possibility to configure the metrics statically in order to avoid suboptimal rerouting.

TLV Type		
TLV Length		
Metric		
U/D	Sub	Prefix Length
Prefix		
Optional all-subTLVs Length		
Optional subTLVs Type		
Optional subTLVs Length		
Optional subTLVs Value		
…		
Metric		
U/D	Sub	Prefix Length
…		

Figure 10: The Extended IP Reachability TLV #135 with two metric fields.

The Extended IP Reachability TLV #135 is a combination of old TLVs #128 and #130. The TLV #135 has a 32-bit metric space to stay compatible with other routing protocols. Among other things there are sub-TLV fields and Up/Down Bit. The whole TLV is described in Figure 10 which also shows how multiple metrics could be inserted into one TLV. After 1-247 bytes of optional subTLV value, a new metric can be inserted and the fields continue the same way as with the earlier metric. The idea of TLV #135 is to encode the useful information only. For example an IP address 172.16.64/19 has only 3 bytes of useful information and only those need to be stored. [11]

Routing software is updated rarely so the new TLVs must be compatible with the old ones. Vendors have their own compatibility solutions because RFC 3784 does not discuss the subject.

IS-IS was designed with scalability in mind and it can be seen e.g. in prefix leaking. Each Level 1 prefix is leaked to Level 2 but the other direction is blocked by default. Only a default route is leaked by L1L2 router. However, sometimes it is useful to trade some scalability for optimality of traffic flow. The previously mentioned Up/Down Bit is used as a Marker Bit to avoid looping the leaked prefixes. If Down bit is found in IP Reachability TLV it means that the prefix is leaked.

If L2 prefixes are wanted to leak to L1, it has to be specifically configured. There are two options for that, controlling the leaking via an extended access list or controlling the leaking via a route-map. For smaller networks, extended access list is enough. To the other direction, “L1 prefixes get leaked to L2 by default because the Extended IP Reachability TLV #135 has no notion of internal versus external prefixes. [11]”

2.7 MultiProtocol Label Switching

Even though a routing network is very capable and a fine piece of work it also has its downsides. Every packet finds its way from source to destination but every packet and

packet sending user is fighting for the same resources. In many cases that is not enough. If for example business users are considered, a corporation may need its data be transferred in a secure and assured way between offices. One widely used way of providing that kind of service is Virtual Private Networks (VPNs). VPN is, as the name indicates, a virtual network for private use on top of physical network. It is available only for those who have paid for it and it can be taylor made for customer's needs. Probably the most common tool for implementing VPN is MultiProtocol Label Switching (MPLS).

MPLS is a mechanism which combines the advantages of routing and switching i.e. the manageability of routing and fast and light forwarding of switching. It uses labels to switch packets quickly and in a controlled way. As the name suggests, MPLS can work in any network protocol, not just Internet Protocol (IP). Compared to other label based protocols, MPLS has a benefit of stacking labels which allows creation of VPNs, Traffic Engineering (TE), Fast ReRouting (FRR) and Quality of Service (QoS).

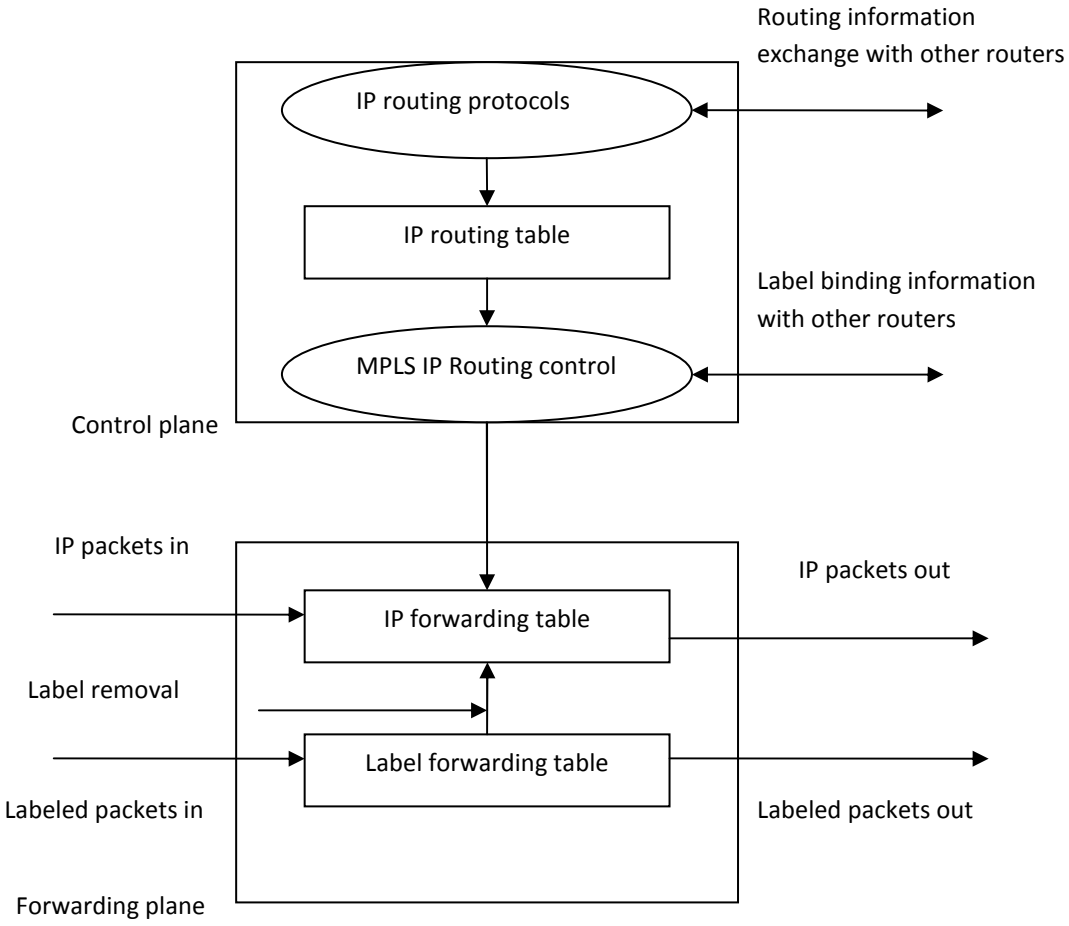


Figure 11: Information exchange between edge-LSR and no-edge-LSR (the latter is not shown in figure). Adapted from [12].

MPLS enabled nodes have two components, control plane and forwarding plane. The control plane interacts with neighbouring nodes to change routing and label binding information and maintains RIB and Label Information Base (LIB). On the forwarding plane there is Forwarding Information Base (FIB), which has all the labels assigned by this node and label mappings learned from other nodes. It also includes the Label

Forwarding Information Base (LFIB) which contains only the labels which are currently used for forwarding. Figure 11 shows how an edge-LSR interacts with other LSRs. A non-edge-LSR does not have an IP Forwarding table in the Data Plane but everything else is the same.

The MPLS nodes need to perform one or more of the following tasks: Push i.e. adding a label to a label stack, Pop i.e. removing a top label from the label stack and/or Swap i.e. popping the old label and pushing a new one to the stack.

The nodes also need to be able to classify the incoming packets. This can be done in many ways e.g. according to an address prefix or a port. Each group of packets which are treated the same way by the same node is called a Forwarding Equivalence Class (FEC). Every packet under the same FEC uses the same label on the same MPLS link.

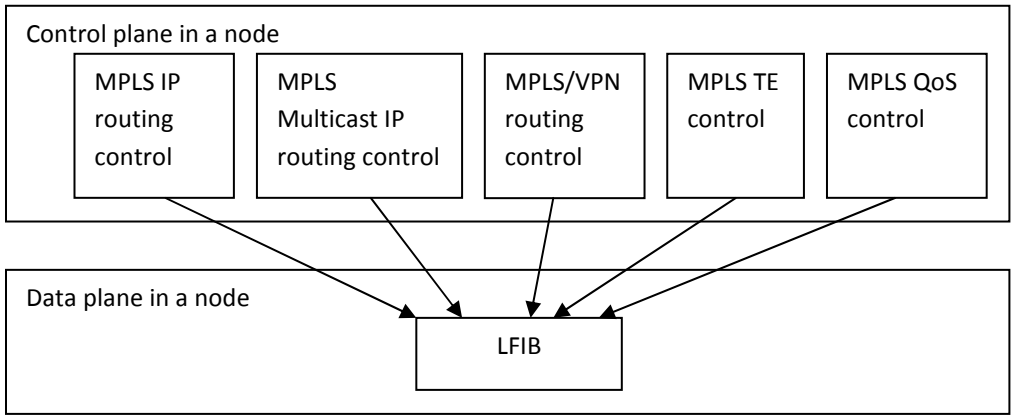


Figure 12: MPLS application interactions. Adapted from [12].

The MPLS enabled devices can be divided into several groups. Label Switching Router (LSR) is a router which is able to do label switching. Edge-LSR or Label Edge Router (eLSR/LER) is an LSR which sits at the edge of a network. If an LER is at the ingress of an MPLS network it pushes a label in front of a packet arriving to the network and sends it to the next LSR. If the LER is at the egress of the MPLS network it pops the outer label and forwards the underlying packet to a next-hop router. LSRs in between ingress and egress swap the outer label of the incoming packet. The LSR makes a lookup in the LFIB (Figure 12) and makes the label swap according to that. These steps are illustrated in Figure 14. The path taken by packets of a FEC is called a Label Switched Path (LSP). The propagation of the packet goes through a few steps. First, the Ingress Edge-LSR receives the packet and classifies it into a FEC. The packet is labelled accordingly and sent forward. Second, the core LSR receives the labelled packet and uses the label forwarding table to switch the label according to the FEC. Finally, the Egress Edge-LSR receives the packet and pops the label and performs a L3 lookup on the carried packet. [13]

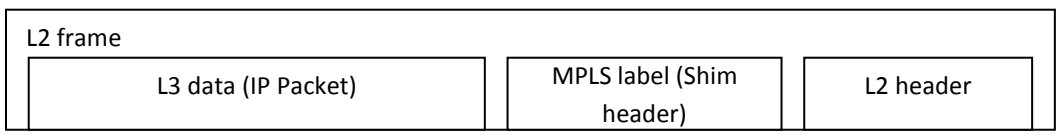


Figure 13: Labelled IP packet in L2 frame. Adapted from [12].

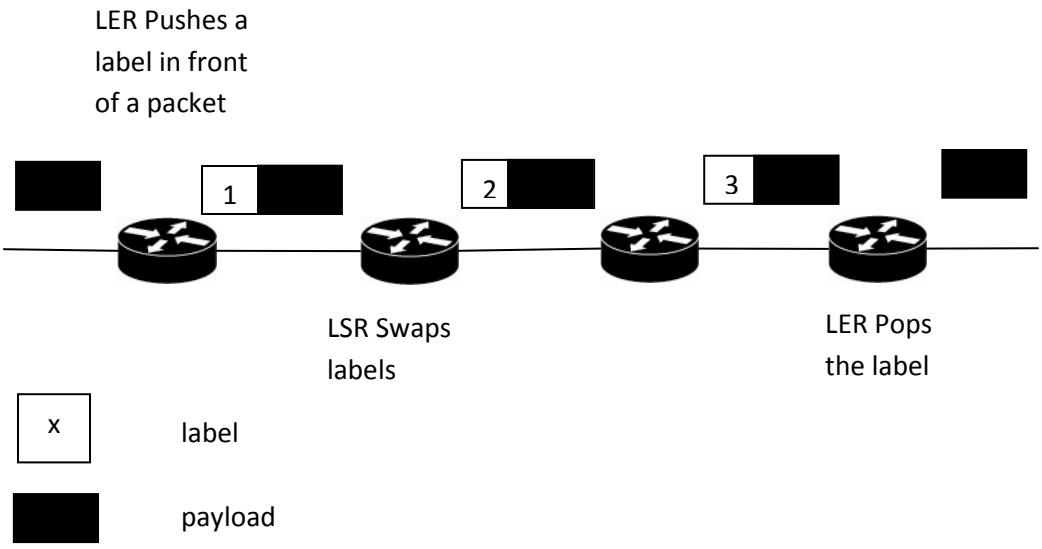


Figure 14: Push swap and pop functions while a packet traverses an MPLS network

The label itself is part of a MPLS label stack header (or shim header) which is inserted between the L2 header and L3 contents of the L2 frame (see Figure 13). The shim header (Figure 15) has a Bottom-of-Stack bit which indicates the bottom label and helps a LSR to function accordingly, an Exp-bit which is for QoS purposes and Time To Live (TTL) field for loop detection. The stacking of labels is done by inserting more than one of the shim headers between the L2 header and L3 content. It is necessary that the receiver of a labelled packet knows that the packet is labelled. Therefore new protocol types have been defined above L2. E.g. in LAN environment the labelled packets carrying unicast or multicast L3 packets use ethertype values 8847 hex and 8848 hex.

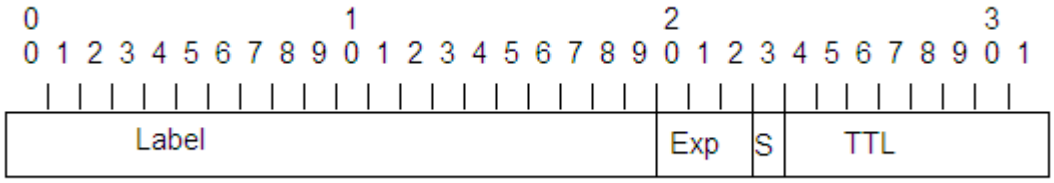


Figure 15: MPLS label stack header. Adapted from [14]

So far it has been assumed that the routers somehow know which label to assign to a packet when forwarding it. An MPLS-device uses a label distribution protocol "as a set of procedures by which one LSR informs another of the meaning of labels used to forward traffic between and through them [13]". The original label distribution protocol called Label Distribution Protocol (LDP) [15] was designed specifically for that purpose. Today, Resource reSerVation Protocol (RSVP) is also widely used. Even though it was originally designed for other purposes it has re-emerged for distributing MPLS-labels and it has numerous extensions e.g. Traffic Engineering (TE) and Point-to-MultiPoint (P2MP) features.

2.8 Resource Reservation Protocol- Traffic Engineering

2.8.1 Background

MPLS does not assume just one signalling protocol for forming LSPs. Even though Label Distribution Protocol (LDP) is designed just for that, Resource ReSerVation Protocol- Traffic Engineering (RSVP-TE) is also fit for that purpose. In many cases, it is also better e.g. if TE is wanted. In the TSF MEN, the P2MP trees are implemented with (P2P) RSVP-TE. The only option was to configure the P2MP trees manually and statically hop by hop because the TSF ME devices do not support any P2MP protocol. The backup LSPs are also configured the same way.

Nodes that support RSVP and MPLS can associate labels with FECs. Once the LSP is defined and the data through the path is defined by the label the path can be treated as a tunnel because it is tunnelling below normal IP routing and filtering mechanisms.

"The signalling protocol model uses downstream-on-demand label distribution. A request to bind labels to a specific LSP tunnel is initiated by an ingress node through the RSVP Path message [16]" The ingress LER sends a Path message downstream, i.e. towards the receiver, with a LABEL_REQUEST object. This causes label allocation. The egress LER responds with a Resv message upstream, i.e. towards the network, extended with a LABEL object. This is how the labels are distributed. The procedure is illustrated in Figure 18.

The signalling protocol model allows the LSP to be explicitly defined by an Explicit Route Object (ERO). The ERO is part of the Path message and defines the nodes the LSP must traverse. If the ERO is missing, the traversed nodes are defined by a routing protocol.

2.8.2 LSP tunnels and TE tunnels

Since RSVP has been used mainly for label distribution the earlier definition of RSVP session does not apply anymore. "When RSVP and MPLS are combined, a flow or session can be defined with greater flexibility and generality. [16]" An ingress node assigns a label to a FEC and defines a flow through the LSP. Such LSP is called an LSP tunnel because it is opaque to intermediate nodes. RSVP SESSION (see Figure 16), SENDER_TEMPLATE (see Figure 17) and FILTER_SPEC objects are defined to support the LSP tunnel feature. These objects are generically referred as LSP_TUNNEL. A tunnel ID and LSP ID are carried so they can be associated with LSP tunnels and possibly used for rerouting or TE. An Extended Tunnel ID field is used e.g. in rerouting for indicating the IP address of the LSP tunnel ingress node.

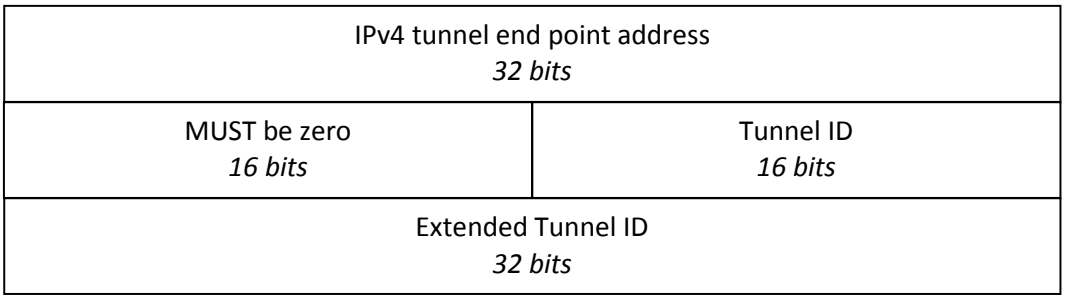


Figure 16: LSP_TUNNEL_IPv4 Session Object

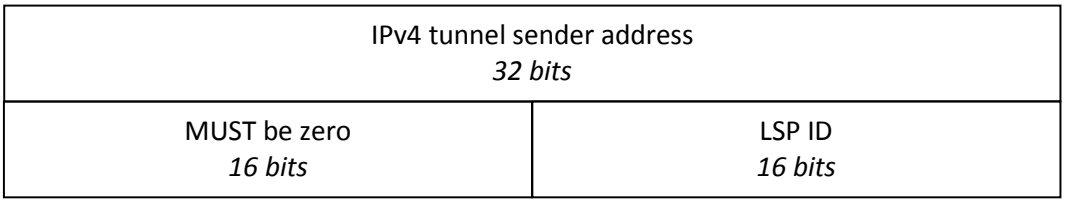


Figure 17: LSP_TUNNEL_IPv4 (&FILTER_SPEC) Sender Template Object

The LSP_TUNNEL_IPv4 Sender Template Object and FILTER_SPEC are identical.

2.8.3 Operation of LSP tunnels

The RSVP-TE has several capabilities in regard to operation of LSP tunnels. They include establishing LSP tunnels with or without QoS, dynamically rerouting established LSP tunnels, observing the actual route traversed by an established LSP tunnel, identifying and diagnosing LSP tunnels, pre-empting an established LSP tunnel under administrative policy control and performing downstream-on-demand label allocation, distribution and binding i.e. the ingress LER initiates the label binding to an LSP.

To create an LSP tunnel the ingress node must send downstream a Path message with a session type of LSP_TUNNEL_IPv4. A LABEL_REQUEST object is inserted into the Path message and it contains a request for a label binding for this LSP tunnel. The network layer protocol is also added because it can not be assumed to be IP. Furthermore, link layer header has MPLS as the higher layer protocol. If any node along the way to the egress is unable to make the label binding the ingress node is informed by a PathErr message.

If a specific route is wanted, the LSP tunnel can be configured by an Explicit Routing Object (ERO). It includes every node that must be traversed by the LSP tunnel. When the Path message with the ERO travels downstream, every node along the path records the ERO in its path state block. In case of an absent hop in the ERO or if the ERO itself is missing, IP routing protocol is used for defining the route. After the LSP tunnel is established, the ingress node might get new route information and the LSP tunnel could be dynamically rerouted by changing the ERO. If the actual route traversed needs to be known, a Record Route Object (RRO) is added to the Path message too. RRO can also be used for loop detection. Finally, a SESSION_ATTRIBUTE object can be added to the Path message for additional control information, session identification and

diagnostics. The additional control information, such as setup and holding priorities, can be used for verifying that the bandwidth exists at a particular priority along the entire path before pre-empting any lower priority reservations. [16]

The egress LER responds to the Path message with Resv message and includes a LABEL object in it. The Resv message travels the route in reverse order. Each node that receives the Resv message uses the label of the LABEL object for outgoing traffic associated with this LSP tunnel. The sender of the Resv message inserts a new label to the LABEL object if it needs to receive traffic with a specific label. When the Resv message propagates to the ingress LER, the LSP is established.

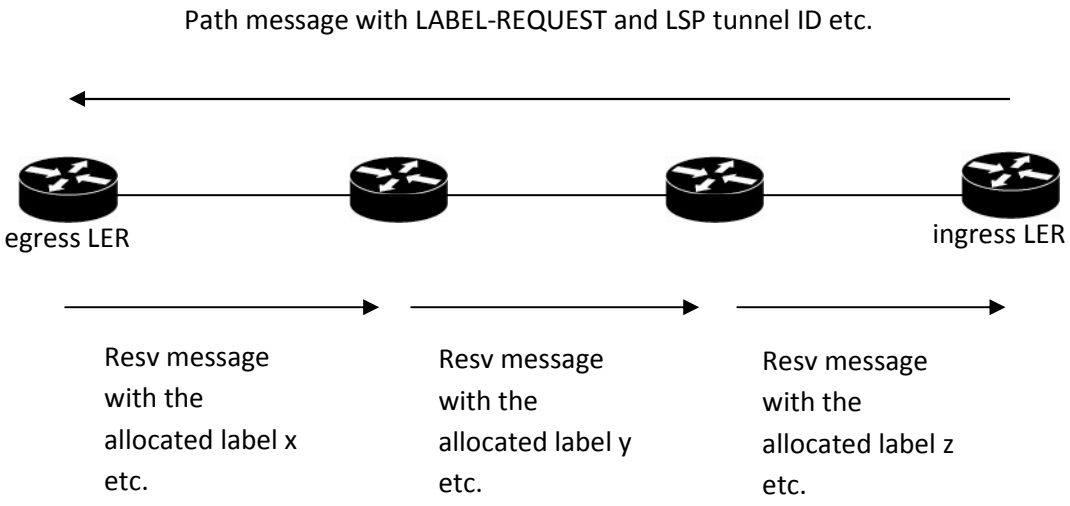


Figure 18: Path and Resv messages and label allocation

2.8.4 Reservation styles

The receiver node decides a reservation style for each session and each RSVP session must have a particular style. The sender node does not have any influence on the decision.

A Fixed Filter (FF) reservation style makes a reservation for each sender which is not shared with any other sender. This results in having a unique label assigned for each sender. The reservations produce point-to-point LSPs. [16]

A Wildcard Filter (WF) is a reservation style where a shared reservation is made for all the senders in a session. A single label is assigned for all the senders to a session and this creates a multipoint-to-point LSP or just a point-to-point LSP if only one sender is present. The WF reservation style is useful only if the senders are not talking at the same time. [16]

A Shared Explicit (SE) reservation style makes a reservation to a session for a group of senders but the receiver explicitly states by the Resv message which senders are included. These senders have a single label. If another group of senders with another label is wanted, a new LSP is created. The SE reservation style can be provided by point-to-point LSPs or a multipoint-to-point LSP. The multipoint-to-point case takes place if ERO is not used or if all the Path messages from senders have identical EROs. [16]

2.8.5 Rerouting TE tunnels

Traffic Engineering must support rerouting of established TE tunnels according to an administrative policy e.g. when a better route is detected traffic is moved to a new LSP tunnel. When traffic is moved from one LSP tunnel to another the old one must not be torn down before the new one is established. This type of rerouting is called make-before-break. It can cause a problem while making a bandwidth reservation because when the new LSP tunnel is being established there needs to be a double the amount of bandwidth than normally. Sometimes the old and the new LSP tunnels are competing for the same resources if they have common links. If there are not enough resources for both of them, Admission Control prevents the new LSP tunnel from being established.

A combination of LSP_TUNNEL SESSION object and SE reservation style accommodates a smooth transition in bandwidth and routing. The idea is that the old and the new LSP tunnel share resources where they have a common link. The LSP_TUNNEL SESSION object is used for identification of a particular TE tunnel. During the reroute or bandwidth operation the ingress node must appear in the session as two different senders. The ingress node forms a new LSP ID and a new SENDER_TEMPLATE and calculates a new ERO. Then the ingress node sends a new Path message with the old SESSION object, the new LSP ID, the new SENDER_TEMPLATE and the new ERO. The old LSP continues to be used and the Path message is refreshed. On the links which are not common, the new Path message is treated as a conventional new LSP tunnel setup. On common links the shared SESSION object and SE reservation style enable resource sharing among the new and the old LSP. Once the ingress node receives the Resv message it transitions the traffic to the new LSP tunnel and the old LSP tunnel can be torn down. [16]

If the ingress node tries to increase the bandwidth reservation, it creates a new LSP ID and sends it downstream in a new Path message while the old reservation is been refreshed. This is done in case the new reservation attempt fails and all reservation is lost.

2.9 Multicast

In networks today there are multiple ways of sending data from a source to a receiver or multiple receivers. The most common one is to use unicast which means one source sends data to one receiver using either Transmission Control Protocol (TCP) [17] or User Datagram Protocol (UDP) [18] depending on the requirements of data transfer. Another common way of sending data is broadcast which means that a single user sends it to every receiver in the network domain whether the receivers want it or not. Here, using TCP is impossible because of large number of ACK- messages would overload the source. Furthermore, broadcast is wasting bandwidth and other resources.

To solve this resource wasting problem, multicast is used. It is similar to broadcast with one sender and multiple receivers and using UDP but the data is sent only to receivers who wish to receive it. In multicast the paths from the source to the receivers form a distribution tree.

In Figure 19 the basic idea of multicast is presented. The black node at the top is the head end i.e. a source. The other black nodes are receiving data and the white nodes are not. It is clear that the amount of data is reduced compared to broadcast. It is reduced from 14 units to 9 units i.e. 36%. This is not, of course, an upper limit for the bandwidth and resource savings but it does depend heavily on the topology and the number of receivers wanting the data.

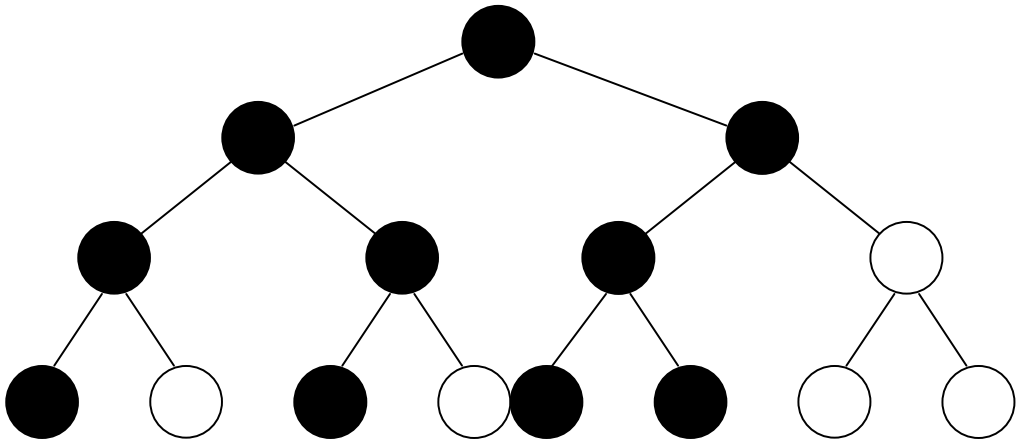


Figure 19: Example of data distribution by multicast where the black nodes receive the data

A multicast address is called a group address. The receivers state their interest in receiving data sent to a group by issuing a Join message. The Join message can be either of any source or a source specific type. The Any Source Multicast (ASM) Join is presented in form (*,G) where the '*' represents any source or a so called "wild card" and the 'G' is the group to be joined. In Source Specific Multicast (SSM) the Join message is in form (S,G) which, unlike any source Join message, has a specific source 'S' defined. The 'S' and the 'G' are IP addresses.

The routers keep track of different groups in the router. It is called multicast forwarding state. It has separate information lists for incoming and outgoing interfaces. The multicast forwarding state is set up from the receiver to the root of the tree which is the opposite of unicast destination based routing. Reverse Path Forwarding (RPF) check is used to determine the interface closest to the root of the tree. The RPF enables Join messages to find their way to the right place.

A Shortest Path Tree (SPT) has the root as a source as illustrated in Figure 20. "If a router learns that an interested listener for a group is on one of its directly connected interfaces, it tries to join the tree for the group. [19]" If the source is known, the RPF is used to determine the interface closest to the source and receiving the group data.

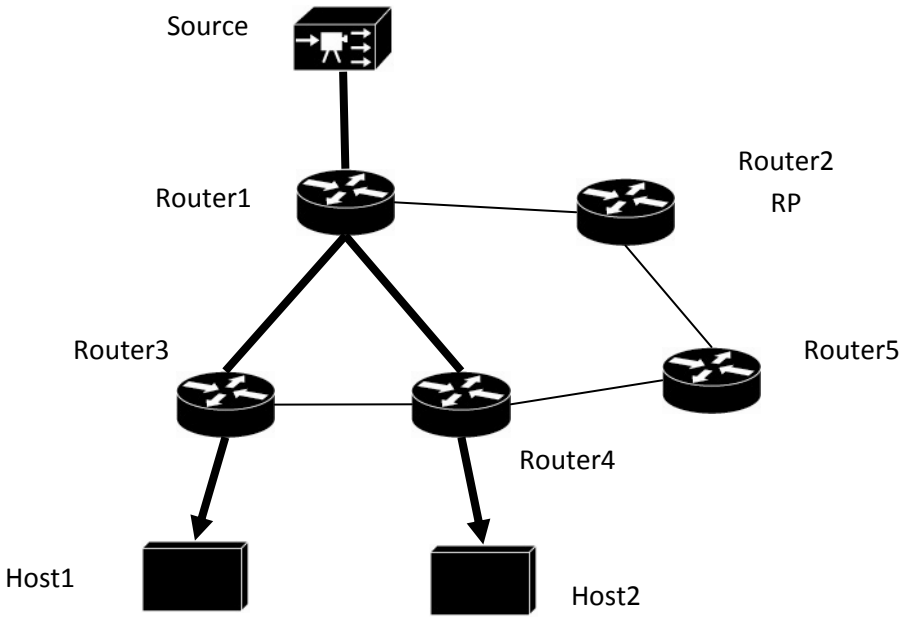


Figure 20: Shortest Path Tree from a source to receivers. Adapted from [19].

When the router receives a (S,G) Join message, it adds the interface where the Join message came from to the outgoing interface list. The RPF check is done again and the Join message is sent to an upstream router. This procedure is repeated until a router which has the group 'G' is found. Furthermore, if a router directly connected to the source 'S' is found, a new branch is created. When every router has a forwarding state for the source-group pair, the data can start to flow.

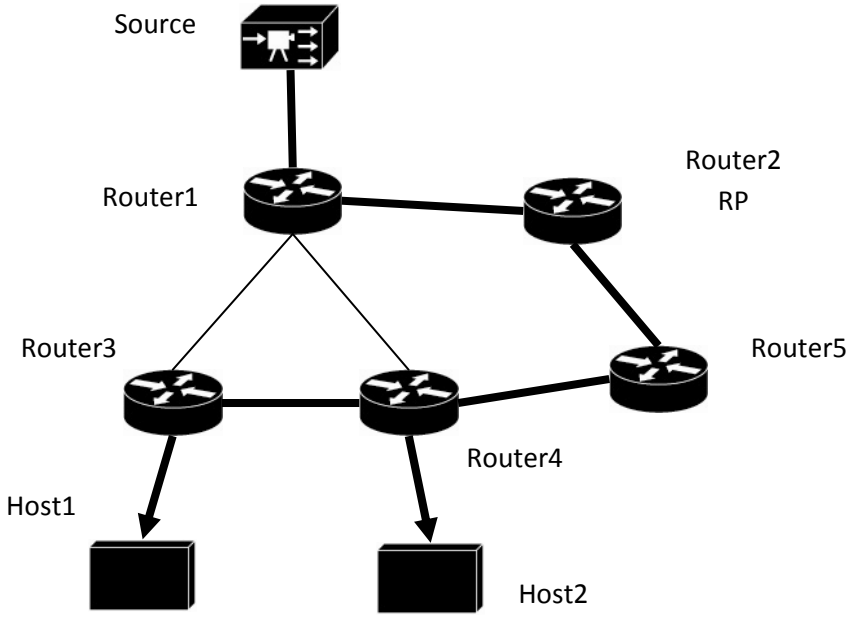


Figure 21: Shared Tree using a Rendezvous Point (Router 2). Adapted from [19].

When the source is unknown, a so called shared tree is used for connecting the source and the receivers. In the shared tree the root is a so called Rendezvous Point (RP). The receivers send Join messages towards the RP. The source sends data to the RP. This way the distribution tree is established. The Join messages are in (*,G) form because the source 'S' is known only by the RP. The shared tree and the SPT creation is the same except that the roots of the trees are in a different point of a network. In Figure 21 the root of the tree is at Router2 which is the RP.

2.10 Internet Group Management Protocol

In previous section multicast Join messages were discussed. They are sent by multicast protocols and in this thesis two of them are discussed, Internet Group Management Protocol (IGMP) and Protocol Independent Multicast (PIM, see Section 2.11). IGMP has three different versions. IGMPv2 is the latest Any Source Multicast (ASM) version and IGMPv3 is the latest version. IGMPv3 allows sending Source Specific Multicast (SSM) Join messages. IGMP gathers information about hosts wanting to receive multicast data. PIM collaborates with a routing protocol and creates a multicast routing table. These two interact at IGMP/PIM border.

2.10.1 IGMPv2

"IGMPv2 is used by IP hosts to report their multicast membership to any immediate neighbouring multicast routers. [21]" The multicast router keeps a multicast group membership list of the hosts. Each membership has a timer which is reset when a Report message is received. A Membership Query message, which can be either General or Group Specific Query, is sent at a certain interval to enquire which hosts belong to a multicast group. In a General Query message the address field is set to 0 in contrast to Group Specific Query message which has a specific group address set. In each network there is usually one querier which sends the Query messages. When a host receives a Query message, a timer for each group the host is a member of, is set to a random number. When the timer expires, a Membership Report message is multicast to the group. The Membership Report message is used for answering to the Query message. An interface where the host replies from is added to the membership list. If the Report message is received before the timer expires, the host suppresses its own message to avoid multiple Reports. When a host is the first to join a multicast group it has to send an unsolicited Membership Report message i.e. without the query. If the timer of the group in the router expires, it assumes that there are no receivers attached to its interfaces. If a host is the last one to send a Report message to a group it has to send a Leave message to the group when leaving. When the querier receives the Leave message, it sends multiple Group Specific Query messages to the group being left and if no response is received until the timer expires the querier assumes that there are no members left.

2.10.2 IGMPv3

For the most part IGMPv3 works the same way as IGMPv2 but the versions have their differences. IGMPv3 has three different types of Query messages, General Membership Query (GMQ), Group Specific Query (GSQ) and Group-and-Source-Specific Query. The GMQ is used for learning the full state of an interface. The GSQ queries the state of a specific group on an interface. Finally, the Group-and-Source-Specific Query is sent

to learn if any interface wants to receive data sent to a specific multicast group. The sources are also specified by a source list. In contrast to version 2, the version 3 host can announce the sources it wants to receive data from. The Membership Report is in either Include or Exclude mode. The sources specified in an Include mode Report message are the ones the host wants to receive data from. If the list is empty it acts like the leave message of IGMPv2. The Exclude Report message informs about the sources the host does not want to receive data from. The IGMPv3 Group States are also either in Include or Exclude mode. IGMPv3 is compatible with the previous versions. [22]

2.10.3 IGMP Snooping

Even though IGMP is an internet layer protocol, IGMP Snooping can be used by link layer switches. It is a method in which the switch can constrain multicast packets to only those ports that have requested the stream. The switch only inspects IGMP Host Membership Reports traversing the switch. A multicast group address is added to the forwarding table associated with the port the IGMP message was received from. Join and Leave procedures are done the same way. It is an alternative solution to manual Media Access Control (MAC) address configuration. It would be required, if there was no IGMP Snooping, to avoid flooding because switch MAC address learning is source based. The downside of IGMP Snooping is the extra work of snooping every multicast data and control packet. When frames flow to the internet layer there is a possibility to inspect only control information (IGMP) and reduce the used processing capacity. A switch multicast forwarding database (FDB) is based on multicast IP address and not MAC addresses like many other LAN switch implementations. [23]

2.11 Protocol Independent Multicast- Sparse Mode

Protocol Independent Multicast – Sparse Mode (PIM-SM) is a signalling protocol used between routers which enable creation and maintaining of Multicast Distribution Trees (MDT) which can be either Shared Path Trees or Source Path Trees. PIM-SM also enables routers to forward multicast packets, prevent loops and application of policy.

PIM-SM operation has two planes, control and data plane. The control plane is used for enabling multicast on interfaces connected to other routers and IGMP is enabled on any interface with a potential receiver. It is also used for neighbour establishment by PIM Hello's. After enabling PIM-SM on interfaces and finding neighbours, the router is ready to receive and forward multicast packets over multicast enabled interfaces. The data plane is used for MDT setup. A first-hop router registers to the Rendezvous Point (RP) and last-hop router notifies the RP when a receiver appears with PIM Join message. The source and the receiver meet at the RP, like in Figure 21 and a flow is established. The flow will pass the RP after the initial meeting which equals the situation in Figure 20.

The PIM-SM works by an "Explicit Join" model i.e. if a multicast node does not receive a specific request for the data, it should not send data to the network. There are three ways of receiving the Explicit Join: dynamic IGMP Host Membership Report which is based on receiver application activity, PIM Join message which is based on PIM protocol activity and static IGMP Host Membership Report which is based on manual configuration.

MDTs can be either Source Path Trees or Shared Path Trees. The Source Path Trees are formed after the source and the receiver have met at the RP. For a while traffic travels via two routes, through the RP and straight between the source and the receiver. After forming the Source Path Tree, the Shared Path Tree is torn down.

The PIM-SM uses Reverse Path Forwarding (RPF) check for loop detection. It also ensures that the PIM signalling traffic propagates through the network correctly. [24]

2.12 Protocol Independent Multicast- Source Specific Multicast

Protocol Independent Multicast- Source Specific Multicast (PIM-SSM) is a derivative of Sparse Mode PIM (PIM-SM). The biggest difference is that PIM-SSM does not use Shared Path Trees. The RP used by the PIM-SM is only necessary if a source is unknown. If the source is known, (S,G) Joins and Source Path Trees can be used which is the case with the PIM-SSM. The (S,G) tuple provides uniqueness to differentiate between SSM channels, i.e. (S1, G) is completely different from (S2, G) even though both S1 and S2, are part of the same multicast group G. SSM also solves the MAC multicast address mapping problem (Section 2.2) at least partially, "because the chance of two different multicast groups corresponding to the same MAC address appearing on the same LAN is relatively slim. [19]"

However, even if the source is known, IGMPv3 must be supported because the earlier versions of IGMP do not enable joining a specific (S,G) group. This is a greater limitation than knowing the source address. If IGMPv3 is not supported, the PIM-SSM translation must be configured at the first-hop router from the receiver which then propagates the Join- message to the source. [25]

2.13 Point-to-MultiPoint Resource Reservation Protocol-Traffic Engineering

In Section 2.8 the RSVP-TE provided a tool to create LSPs. It can only create LSPs between two points i.e. P2P, which is not practical when a Point-to-MultiPoint (P2MP) LSP is needed. P2MP RSVP-TE has been designed for this purpose.

"A P2MP Tunnel comprises one or more P2MP LSPs [26]." A P2MP SESSION object illustrated in Figure 22, identifies the tunnel by the P2MP Identifier (P2MP ID) which is the destination IP-address, a tunnel Identifier (Tunnel ID) and an extended tunnel identifier (Extended Tunnel ID) which is the sender IP-address. These IDs provide an ID for the P2MP TE Tunnel destinations. In addition to these, a P2MP LSP needs the tunnel sender address and LSP ID fields of the P2MP SENDER_TEMPLATE (Figure 23) for identification. The P2MP LSP state is managed with RSVP messages and because the P2MP LSP comprises multiple Source-to-Leaf (S2L) Sub-LSPs i.e. a P2P LSP from the sender to a destination, one IP packet may not be enough to represent the full state. Furthermore, the Sub-Group fields Sub-Group ID and Sub-Group Originator are added to the SENDER_TEMPLATE and FILTER_SPEC objects because it must be possible to efficiently add and remove endpoints and handle the "re-merge" problem. The S2L_SUB_LSP includes the destination address.

P2MP ID *a 32-bit destination address*	
MUST be zero	Tunnel ID *a 16-bit number that identifies the Tunnel*
Extended Tunnel ID *a 32-bit tunnel sender address*	

Figure 22: A P2MP SESSION object

IPv4 tunnel sender address *Sender address in IPv4 format*	
Reserved	LSP ID *a 16-bit identifier*
Sub-Group Originator ID *IPv4-address*	
Reserved	Sub-Group ID *a 16-bit identifier*

Figure 23: P2MP LSP Tunnel IPv4 SENDER_TEMPLATE Object. Only IPv4 tunnel sender address and LSP ID are used for initiating a P2MP LSP. Rest of the fields are used for editing the P2MP LSP.

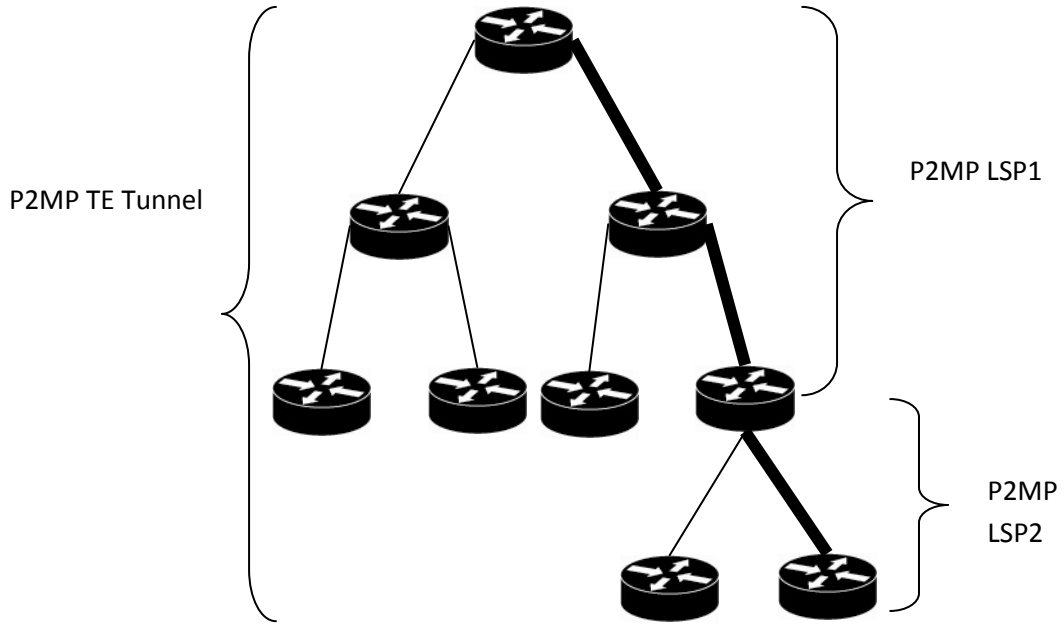


Figure 24: P2MP tunnel comprising two P2MP LSPs. The thick line is a Source to Leaf (S2L) sub-LSP

The whole tree is called the P2MP TE Tunnel which could equal a P2MP LSP. The P2MP LSP can however, consist of more than one LSPs like in the Figure 24 where the P2MP TE Tunnels comprise two P2MP LSPs. The thick black line marks a S2L sub-LSP.

An explicit route of a S2L sub-LSP is specified by the EXPLICIT_ROUTE Object (ERO) or optionally by the P2MP_SECONDARY_ROUTE Object (SERO). They

correspond to a particular S2L_SUB_LSP object which is just a destination IP-address. Path messages are used to signal P2MP LSPs and it may signal multiple S2L sub-LSPs. One Path message may not be big enough to contain the S2L sub-LSPs so multiple Path messages may be needed and a separate manipulation of sub-trees of P2MP LSPs is allowed. In this operation the Sub-Group Originator ID is changed but the Sub-Group ID remains the same.

The signalling of S2L sub-LSPs is done by Path, Resv and PathTear messages. The request is answered by Resv message which makes the reservation and distributes labels for the path. The PathTear message is used for pruning a S2L sub-LSP.

The Path message is initiated by an ingress node. It is sent to every egress node of the P2MP LSP to request a bandwidth reservation for a path expressed by ERO. Each S2L sub-LSP is tied to a particular P2MP LSP by P2MP SESSION Object and <Sender address, LSP ID> fields of the P2MP SENDER_TEMPLATE Object. This allows a P2MP LSP to be signalled by multiple Path messages or signalling multiple P2MP LSPs by one Path message. The S2L sub-LSP must be propagated according to the ERO if ERO is present. Otherwise a hop-by-hop routing is used for propagating the S2L sub-LSP towards the egress. If the ERO is present, the propagation mode is called a strict mode. If the hop-by-hop routing is used, the mode is called loose. The Path message may include a SERO which includes the rest of the S2L sub-LSPs. The first node of the SERO is the branch node. This is illustrated in Figure 25. [26]

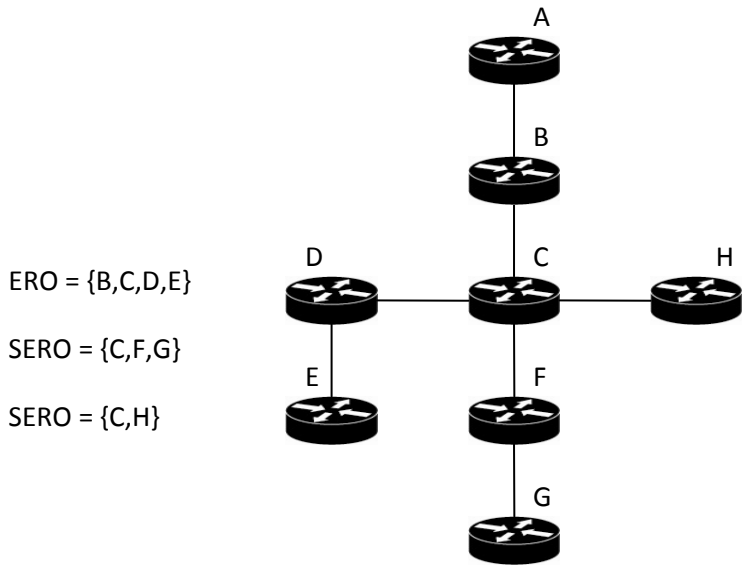


Figure 25: ERO and SERO

A Record Route Object (RRO) contains all the LSRs traversed by the Path message. Each LSR adds itself to the RRO and propagate it downstream. A branch LSP sends a new RRO so that every sub-LSP is included in the RROs.

The building of a P2MP LSP may face some problems and when this happens it cannot be built. In this case there are two options, the P2MP LSP can be built up to the point of failure or if the “LSP integrity” bit is set the whole P2MP LSP fails.

Furthermore, there has to be a way to add egress LSR(s) to a P2MP LSP. There are two ways to solve this problem. The first one is to add the LSR to an ERO of an existing Path message and then refresh the whole Path message. This may cause re-computation of the ERO compression encoding. The other way is to add the LSR to a new Path message and signal only the S2L sub-LSP which is affected by the LSR addition.

The Resv message is used to distribute labels for the P2MP LSP. The previously created RRO defines the path that needs the labels. Each <Sender address, LSP ID> pair has its own label and it is allocated by the egress LSR. The node upstream from the egress LSR must allocate its own label and send it upstream in the Resv message. Furthermore, a node must associate the label sent upstream, for a P2MP LSP, with all the labels received from downstream for that P2MP LSP. In order to keep a branch node from getting too much Resv messages at one time, a downstream node can combine Resv messages to one bigger message or delay the Resv message for a random time so that they do not arrive to the upstream node all at once.

The path reserved for the P2MP LSP is defined by the RECORD_ROUTE Object which represents the ERO and the P2MP_SECONDARY_RECORD_ROUTE Object which represents the SERO. Bandwidth reservations for the P2MP LSP are done by either Fixed Filter (FF) or Shared Explicit (SE) which were introduced in Section 2.8.4. If the reservations are done by FF, the resources or labels are not shared with the P2MP sub-LSP belonging to another P2MP LSP. If SE is used, then resources should be shared but the labels must not be shared.

A PathTear message is used for removing leaf nodes from P2MP LSPs at different points in time. The operation is called pruning. The pruning can be done either by the implicit or the explicit S2L Sub-LSP Teardown.

Implicit teardown uses standard RSVP message processing. A modified message is sent to either Path or Resv message, which ever was the last to advertise a S2L Sub-LSP, to inform which S2L Sub-LSPs are not torn down. The node processing the message must be certain that the S2L Sub-LSP is not in any other Path state associated with session.

The Explicit S2L Sub-LSP Teardown works with the PathTear message which corresponds to a Path message. The PathTear message is signalled with the SENDER_TEMPLATE and SESSION objects corresponding to the P2MP LSP and <Sub-group originator ID, Sub-group ID> tuple corresponding to the Path message. The PathTear message tears down all the S2L Sub-LSPs signalled by the Path message. [26]

2.14 Multicast Label Distribution Protocol

The P2MP RSVP-TE introduced in the previous section is not the only MPLS signalling protocol which has P2MP capabilities. This section introduces Multicast Label Distribution Protocol (mLDP), which is still in draft phase but is already implemented in some devices.

Setting up a P2MP LSP with mLDP starts from leaf nodes as does tearing down. The leaf nodes and a root node install forwarding state for mapping traffic into the P2MP LSP and out from the P2MP LSP. “Transit nodes install MPLS forwarding state and propagate P2MP LSP setup/tear-down messages toward the root. [27]” The devices which support the P2MP capability advertise it by sending a P2MP Capability Type Length Value (TLV) in a LDP Initialization Message.

The P2MP LSP setup uses a P2MP FEC Element which is presented in Figure 26.

<table>
<tr><td>P2MP Type
to be assigned by IANA</td><td colspan="2">Address Family
the Root LSR Address family e.g. IPv4</td><td>Address Length
4 or 16 octets for IPv4 or IPv6</td></tr>
<tr><td colspan="4">Root Node Address
a host address in format defined by the Address Family field</td></tr>
<tr><td colspan="2">Opaque Length
the length of opaque value in octets</td><td colspan="2">Opaque Value...
identifies the P2MP LSP …</td></tr>
<tr><td colspan="4">Opaque Value continues
…in context of…</td></tr>
<tr><td colspan="2">Opaque Value continues
… the Root Node</td><td colspan="2"></td></tr>
</table>

Figure 26: The P2MP FEC Element with fields explained.

The LDP Opaque Value Element inside the P2MP FEC Element carries information relevant to the Ingress and Leaf LSRs but not to the Transit LSRs. It has TLV fields.

The P2MP FEC Element, which includes the root address X and the opaque value Y, is associated with the forwarding state which has the following form: L’ -> {<I1, L1> <I1, L2> …<In, Ln>}. The L’ is an incoming label on a received packet. The packet is replicated n times and labels L1 – Ln are placed on the replicated packets and forwarded over interfaces I1 – In.

The forwarding state is created by a downstream node sending a Label Map <X, Y, L> to an upstream node which updates its forwarding state if the Label Map has new information. The updating could be a label withdrawal, a label addition or a change of the label. [27]

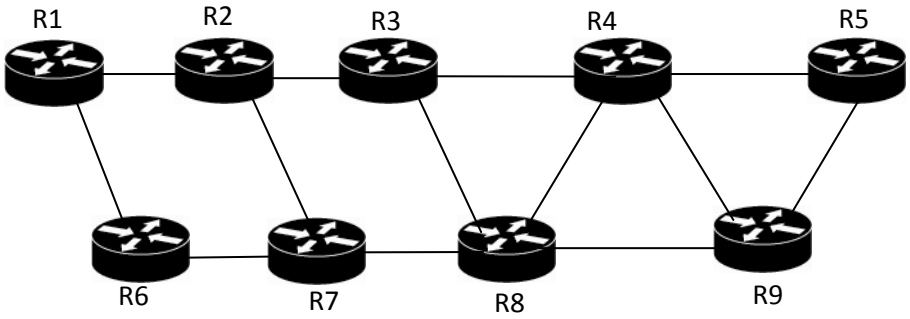
2.15 Fast ReRoute

When using LSPs for traffic transport, problems arise when they go down. LDP and RSVP-TE converge automatically around the failure point if possible. In that case RSVP-TE signalled paths must be in loose mode. This feature is however, too slow for many applications. Standby backup LSPs can be configured but that results in double the work load if all the LSPs have backup paths. To solve this problem, a fast and more sophisticated method has been developed.

The RSVP-TE has an extension called Fast ReRoute (FRR) which creates backup paths for the RSVP-TE signalled LSP tunnels. The backup paths are computed and signalled in advance so switching from the original path to the backup path takes only 10s of milliseconds. Traffic is re-directed as close to the failure point as possible. The FRR

applies only to explicitly-routed LSP tunnels i.e. dynamic LSPs cannot have a backup LSP.

There are two different local repair methods. In one-to-one method a Point of Local Repair (PLR) computes a detour LSP for each protected LSP. A facility backup method computes a single bypass tunnel for all the LSPs that are protected by the PLR. In the one-to-one backup method an LSP is established which intersects the protected LSP somewhere downstream of the point of failure.

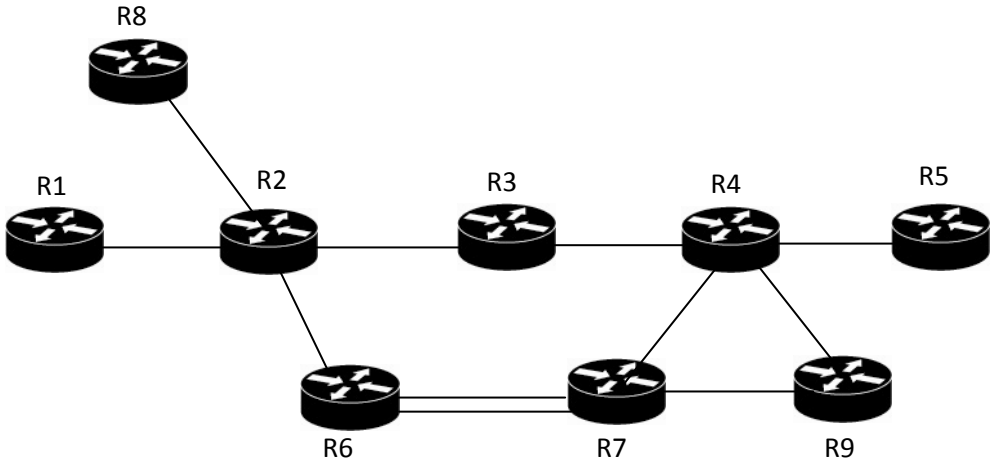


Protected LSP: [R1->R2->R3->R4->R5]
R1's Backup: [R1->R6->R7->R8->R3]
R2's Backup: [R2->R7->R8->R4]
R3's Backup: [R3->R8->R9->R5]
R4's Backup: [R4->R9->R5]

Figure 27: Detours provided by R1 – R4

As can be seen in Figure 27: the detour by a PLR provides protection for only one LSR i.e. the detour intersects the original LSP as soon as possible and the LSPs are merged. All the detours except the one provided by R4 are node protection paths. The one provided by R4 is a link protection path. When a failure occurs, the PLR switches traffic onto a local detour. E.g. in Figure 27 when there is a failure on link R2-R3 or on node R3, the node R2 switches the traffic received from the node R1 to the node R7. The node R2 sends the traffic to R7 with a label learned when R2 created the detour. Traffic moves through R8 to R4 which notices the label which was learned from R5 when the backup LSP was created. The label stack stays the same all the time but the labels being used change. [28]

The facility backup method uses the MPLS label stack to create a backup LSP for multiple LSPs. The backup LSP is called a bypass tunnel and it intersects the original LSP somewhere downstream of the PLR. The LSPs that have a common node somewhere downstream and the PLR as a common node and do not use the nodes of the bypass tunnel in normal situation can use the bypass tunnel as a backup.



Protected LSP 1: [R1->R2->R3->R4->R5]
Protected LSP 2: [R8->R2->R3->R4]
Protected LSP 3: [R2->R3->R4->R9]
Bypass LSP Tunnel: [R2->R6->R7->R4]

Figure 28: Facility backup i.e. bypass tunnel

As can be seen in Figure 28, the bypass tunnel provides protection for LSPs 1-3 but it only protects from failures on link R2-R3 or node R3. E.g. if the link R2-R3 is down, R2 switches traffic from R1 to the bypass tunnel and the label is switched to notify R4 that the traffic is coming through the bypass tunnel. Furthermore, the label of the bypass tunnel is pushed on top of the stack. If penultimate-hop-popping is used R4 receives the original packet without the label of the bypass tunnel.

For both protection methods, if full protection is wanted, (N – 1) bypass/detour tunnels are needed to protect the LSP of N nodes. However, each bypass tunnel can protect several LSPs.

3 IPTV Service Provisioning

This chapter discusses different aspects of bringing the previously discussed protocols and technologies together. When combining two different protocols or protocol implementations of two different vendors, interoperability issues may arise. Furthermore, as in the case of this thesis, every multicast protocols and P2MP LSP signalling protocol may or may not be supported in devices of different vendors.

P2MP RSVP-TE allows the source of multicast traffic to forward packets to one or many multicast receivers over a network without requiring a multicast protocol, such as PIM, to be configured in the network core routers. In theory, using P2MP RSVP-TE would be significantly easier solution for forming the MEN P2MP trees but the problem is that MEN devices do not support the protocol at the moment. Other problems may occur as well with P2MP RSVP-TE, such as the protocol signalling the tree through unwanted nodes. Fortunately, a compromise can be made between easy configuration and control over tree formation. As stated in the Section 2.13 trees can be formed using either the strict or loose mode. When using the loose mode, the path taken by the P2MP tree is decided completely by IGP. This means that an efficient IS-IS metrics setup enables efficient P2MP tree formation. However, this may not be enough but fortunately the tree can be forced to take specific hops in the network by stating the wanted nodes in the strict hop message.

The need for adding new branches comes from customers. When a customer signs an IPTV service contract, a new branch needs to be created for the customer. If there are already other customers in the area, they can use the same branch. Adding new branches to an existing tree or removing a tail end is a very easy procedure with P2P or P2MP RSVP-TE. If P2MP RSVP-TE is available, the implementation stuff just adds or removes a tail end from a destination list and the P2MP tree is updated. If the tree is re-signalled from head end to tail end, it is clear that it causes break in data distribution. If instead, only the changed part is signalled (see Section 2.13), data stream does not experience break to the same extent. That way the already signalled part of the tree remains untouched and data flows without interruptions. Furthermore, if the P2MP tree is signalled with RSVP-TE, only the new part of the tree must be added or removed. Figure 29 shows that it is a very simple addition to the P2MP tree even if there is only P2P signalling protocol is available.

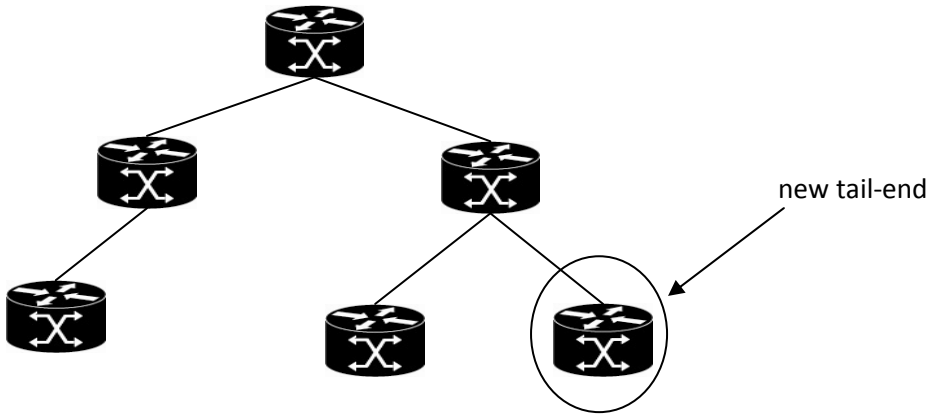


Figure 29: Adding a new tail end to a P2MP-tree and the rest of the tree remains untouched

Problems occur when the network implementation staff has to make changes to the topology of the tree in MEN (see Figure 30). If the P2MP tree is made of one-hop LSPs, it is very slow to tear down the tree and build it up again because every tear-down and build-up has to be done for each one-hop LSP. But, if P2MP RSVP-TE was used instead, the tear-down and build-up would, at least in theory, be significantly faster. That is because only one Path message would have to be sent for tearing down and one updated Path message for building up the tree. In practice, with P2MP RSVP-TE it makes no difference how many tail ends there is, the signalling time is practically always the same. Of course, if all the delays experienced by customers were added up, the total delay would be significant but if just the convergence time was considered, there would be no significant difference. Furthermore, work load of the implementer would include just updating the destination list and optionally the strict hop list.

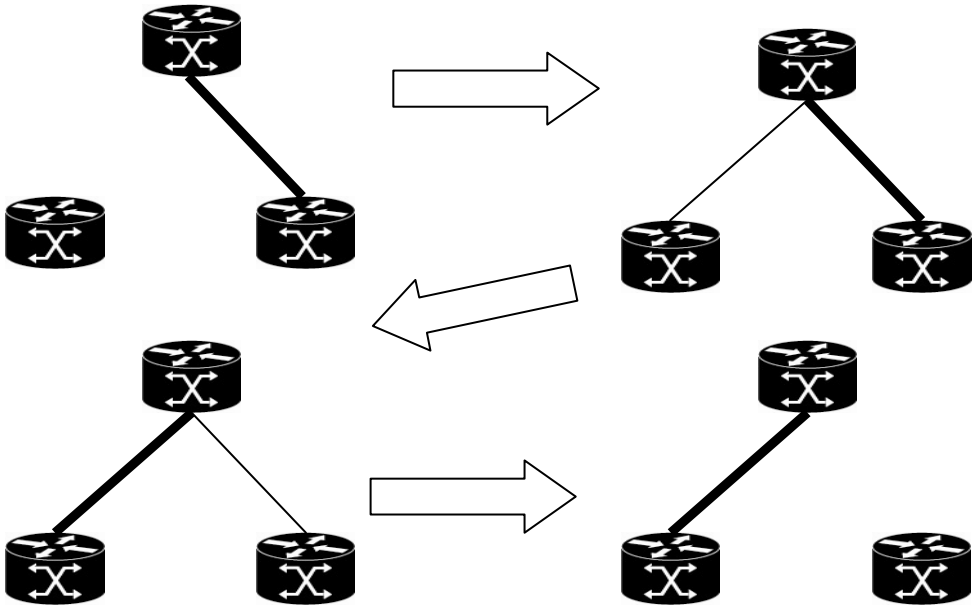


Figure 30: Creating a new LSP next to an old one, moving data flow to the new LSP and deleting the old one. The thick line represents data flow and the thin represents configured LSP without data flow

Another problem could be an interoperability issue between Any Source Multicast (ASM) and SSM. Some domains do not necessarily support SSM so when Join messages are sent in ASM form they need to be mapped or converted to SSM form if the other domain uses SSM.

Overall Join messages play a crucial part in this thesis. The Join messages enable the customer to change channels. Residental GateWays (RGWs) initiate a (*,G) Join message which is received by a DSLAM. If the DSLAM already has the wanted group, it can join the receiver to the group and the Join message is not sent forward to MEN. If the DSLAM does not have the group, the Join message is sent up the tree to MEN and only when a node that has the group is found the message is no longer sent forward.

All in all, what a functioning multicast solution needs is a routing protocol to distribute information about locations of devices, multicast protocol to keep track of users who want to receive data flows and a solution for distributing data to specific receivers at the broadcast domain.

The IGP is IS-IS in every solution. It spans from the core network to the MEN. At the moment the MEN does not support PIM but IGMP Snooping instead. In the future the MEN could support also PIM which would ease things significantly. As sections IGMP (2.10) and PIM (2.11) say, the difference between IGMP and PIM is that PIM is a multicast protocol and IGMP is just a multicast group membership discovery protocol. It means that PIM associates with IS-IS or some other IGP and constructs multicast distribution trees and multicast routing tables and IGMP does not. Because IGMP does not provide the multicast distribution tree, the multicast data is broadcast in MEN by default. That is why, as stated earlier, IGMP Snooping is used. It snoops the membership report messages and forms a MAC address table which has information about the multicast group memberships. This way the group data can be switched out of the right interface. In ME devices IGMP Snooping is only configurable in a service context. It means that the port must be associated to Service Access Point (SAP), Virtual Private LAN Service (VPLS) or Service Distribution Point (SDP). Unlike in other LAN switch implementations, the switch multicast FDB is based on L3 multicast IP address. Translated multicast MAC addresses are not visible to the administrator either.

3.1 ME device service entities

ME devices have different service entities that need to be taken account when creating services. This thesis only concentrates on the IPTV content distribution and for that purpose a VPLS service is used. Any other services are not considered.

A subscriber who uses the service needs a customer account. Creating the account is a very simple procedure, basically all the customer information that is required is a customer ID. It is also very wise to include other information such as name and a description to help the maintenance.

The customer is connected to the service via Service Access Point (SAP) which is associated with a device port on the access side. The SAP is defined by an Ethernet port, encapsulation type and encapsulation ID. It is then connected to the service which has a record of all the customers using the service. On the network side a Service

Distribution Point (SDP) distributes the service to a network and is a link between SAPs. The SDP must have an SDP ID to identify the end-points participating in the service. It also needs encapsulation which is either General Routing Encapsulation, which is good for best-effort traffic because of its low overhead, or MPLS encapsulation, which is well suited for more complex services. Over all the SAPs and SDPs are used for managing services. They bind together different information about the service, such as port type, encapsulation type and different kinds of IDs. How the SAPs and SDPs are located in a network is illustrated in Figure 31.

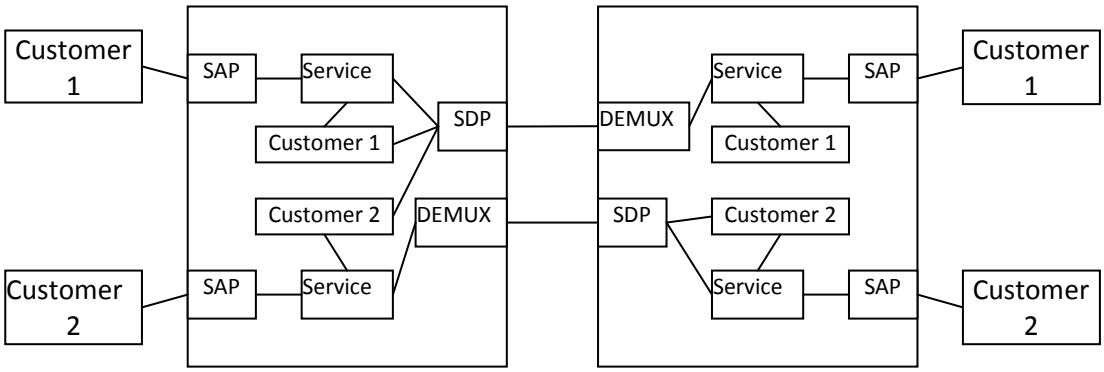


Figure 31: ME device Service Entities

3.2 ME: IGMP Snooping & RSVP-TE, Core: PIM-SSM

The current P2MP solution in the TSF network is a combination of IGMP Snooping and RSVP-TE in MEN and PIM-SSM in the core network. This is not ideal because the more different protocols the network has the more complex and difficult the network is to manage. Furthermore, because the RSVP-TE is a P2P protocol, it is obviously very time consuming to signal a P2MP tree with it (see Figure 32). The reason for this solution is that the equipment in ME domain does not support PIM so the devices cannot send PIM Join messages to the SHE. In addition, the ME switches do not support any other P2MP protocol e.g. P2MP RSVP-TE. Thus there are not many alternatives for the network designing staff. Therefore the P2MP trees in the MEN are built of P2P RSVP-TE LSPs so that every hop is a separate LSP and every LSP needs one manually configured ERO (see Section 2.8). If the P2MP tree has a branch node, more than one LSP are created at the egress of the branch node. It is almost impossible to make longer LSPs with one ERO because a vast majority of ME devices are on the edge of the ME domain and have, or can have in the future, customer connections.

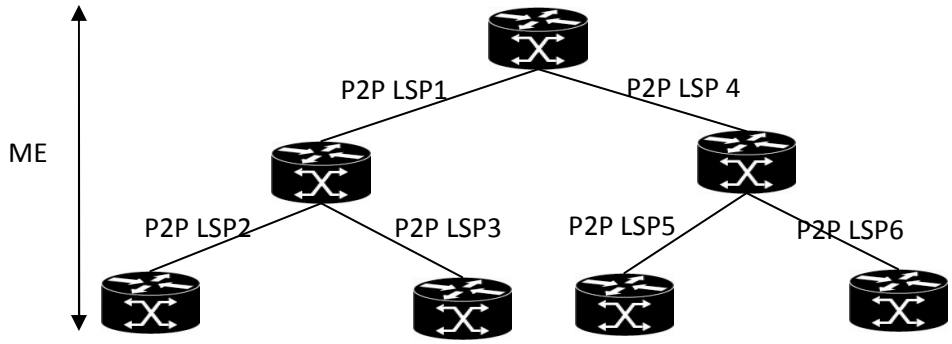


Figure 32: Visualization of signalling P2MP MPLS tree with RSVP-TE

Backup LSPs are also created manually even though FRR extension of RSVP-TE is supported at MEN. A solution with FRR is one of the test cases introduced in a later section. Figure 33 represents the whole network and which protocols are used in ME and core networks.

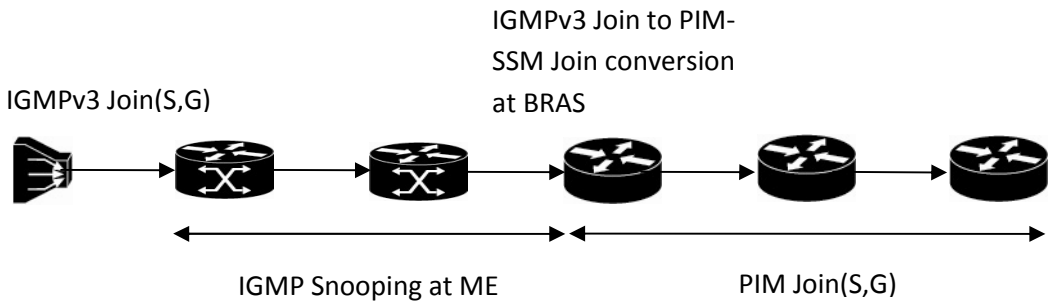


Figure 33: Multicast Join messages traversing the network. IGMPv3 Join messages are converted at BRAS to PIM-SSM Join messages

The signalling of multicast trees starts with a Join message (Section 2.10.1). A customer who wishes to receive a channel (a multicast group) sends a signal with STB to a RGW which initiates an IGMP Join message. The DSLAM is the first possible joining branch to the group i.e. if the DSLAM already has the channel, it just replicates the stream to the requesting RGW. If the channel is not present, the Join message is sent up the P2MP tree to the MEN and from there it continues towards the BRAS until a node, which has the channel, is found. In the junction of two network domains, MEN and core network, the IGMP Join is converted into a PIM- SSM Join message. If the BRAS does not have the wanted group, the Join message starts to move upstream again, but this time the P2MP tree of the core, until it reaches a node that has the group.

The major problem with this solution is the time spent making changes to the ME topology. First, every Service Access Point (SAP) has to be torn down which may take a while when using a generic network configuration tool. After that new SAPs have to be created which again takes a long time.

The upside is that the network designer can control precisely the route of the data stream.

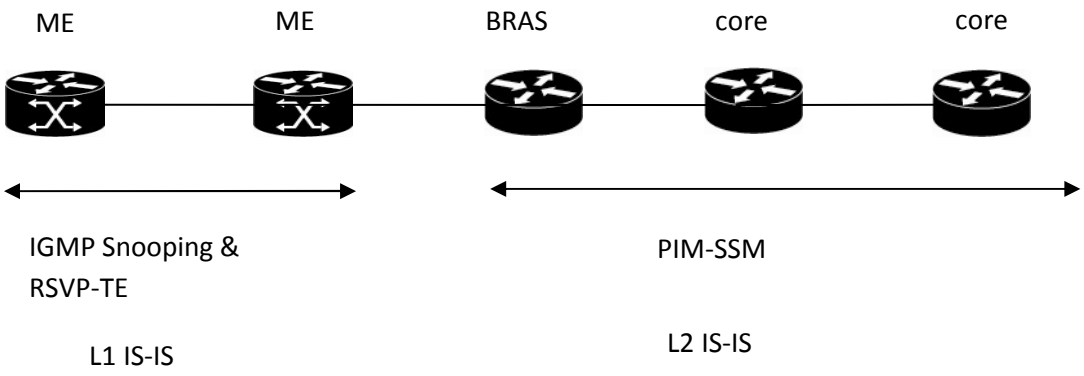


Figure 34: Multicast protocols by domain and IS-IS level

This network works as a reference solution for the other networks. The criteria of the evaluation are stated in Section 1.2.

3.2.1 ME device MPLS and LSP configuration outline

The ME device does not have any prerequisites regarding MPLS tunnels. Although, it is assumed that basic configuration, such as interface IP addressing is done. When configuring strict LSPs, first, MPLS is enabled in every router participating in MPLS tunnel, every network interface and system interface. Second, RSVP-TE is enabled as the MPLS signaling protocol in each router, network interface and system interface where MPLS is enabled. Both MPLS and RSVP instances are set to <no shutdown> state.

After the protocols have been enabled, a strict hop MPLS path, which is used by LSP, is configured. The configuration includes setting an MPLS path name and defining a strict path hop to the next device by system IP address and hop number. Defining hops is repeated until the destination device is reached. After that the path is set to <no shutdown> state.

The MPLS paths can be used by more than one LSP initiated at the same device.
Configuring an LSP requires defining the name of the LSP and the destination IP address. Second, the primary and secondary MPLS paths are defined. Then the LSP is set to <no shutdown> state. The secondary MPLS path is configured in the same way as the primary but it has a different route to the destination. Both MPLS tunnel and LSP need to be created to both directions in order to have two-way traffic.

3.2.2 ME device service configuration outline

Devices included in the VPLS service need the customer ID for defining a customer, SAP at every service access point interface and SDP at every far-end (i.e. ingress and egress) node participating in the service.

The SDP is defined by a locally unique SDP ID number, SDP encapsulation type (MPLS in this case) and originating and far-end node system IP addresses. The SDP is then configured to <no shutdown> state

After the SDP and the customer are defined, the VPLS service is configured. It includes defining VPLS as the type of the service and a service ID assigned to it, defining SAP port with VLAN ID and SDP port with pseudowire ID. In addition enabling IGMP Snooping is needed to have multicast group state information in the service. The VPLS state is setup to <no shutdown>. There are also numerous optional configurations, such as QoS, but they are out of scope.

3.2.3 PIM-SSM configurations at core network routers

Cisco and Juniper routers require the following configurations for PIM-SSM. Enabling PIM-SSM on routers and defining SSM address range if default range (232/8) is not used. Enabling PIM-SM on every interface and enabling IGMPv3 on interfaces that have hosts connected to them. If IGMPv3 is not available on a host, SSM mapping translates IGMPv1 and IGMPv2 membership reports to IGMPv3 reports. [29], [30]

3.3 ME: IGMP Snooping & RSVP-TE with FRR, Core: PIM-SSM

This solution is the same as the reference solution except for the included FRR. This reduces the network designers and implementers work because they do not have to consider the backup paths hop by hop. The FRR with CSPF provides the back up routes dynamically for a given set of links or nodes (see Section 2.15). When the FRR is configured in devices TE and CSPF are enabled on every router in MEN. In addition, FRR is enabled on an ingress and egress LER.

3.4 ME: IGMP Snooping & RSVP-TE, Core: P2MP RSVP-TE

The MEN part is the same as in Section 3.2 but the core network part is implemented with P2MP RSVP-TE. The P2MP RSVP-TE is an interesting choice for the core network because of P2MP RSVP-TE extensions such as FRR and it does not need multicast protocol to replicate data at branches either [2.13]. In theory that would be better for the core network because the PIM-SSM does not have any protection methods and the convergence time in case of failure is as slow as IS-IS.

3.4.1 MPLS P2MP-TE prerequisites and restrictions in Cisco routers

The Cisco MPLS P2MP-TE implementation has some prerequisites and restrictions. The prerequisites include that RSVP and TE must be configured on every router participating in the P2MP tree and Graceful Restart must be enabled on every router and interface participating in the P2MP tree. Naturally, every router must be from the series that supports MPLS P2MP-TE. [31]

The restrictions include that the scope of the P2MP tree is intra-AS. Penultimate hop popping is not available for P2MP MPLS. Protection is restricted to link protection i.e. node and path protection is not supported. Furthermore, only facility FRR (see 2.15) method is supported. All the destinations of the P2MP tree must be configured at the

head end of the tree and they must be manually added and removed. PIM-SM is not supported, only PIM-SSM but it can be fixed with mapping too. Finally, MPLS LSP Ping and MPLS Operations Administration and Maintenance (OAM) features are not supported. [31]

3.4.2 Ingress Provider Edge router configuration in Cisco routers

When Cisco router is the head end of a P2MP MPLS tunnel it requires a bit more configuration than an egress router. The configurations require that MPLS TE is enabled globally and on interfaces. Furthermore, multicast routing is enabled globally. Final global configurations create a destination list which has the node addresses i.e. the loopback addresses defined. Then a tunnel interface is created. It has its own configurations which include enabling P2MP MPLS TE on the tunnel, stating the destination list to be used, when creating the P2MP tunnel. One important thing to notice is that group membership is added on the interface statically. Finally, passive PIM is enabled i.e. it only passes or forwards IGMP traffic, not PIM control plane traffic. [31]

3.4.3 Provider router configuration in Cisco routers

The P routers i.e. the transit routers require no specific configuration. However, the router software must support P2MP signalling and implementation. All multicast traffic is label switched and no multicast routing or PIM configurations are needed. The P routers do however, need <mls ip multicast replication-mode egress> command. If the MLS (MultiLayer Switching) replication mode is ingress the traffic is not forwarded.

3.4.4 Egress Provider Edge router configuration in Cisco routers

The tail end routers remove the MPLS labels from the IP multicast packets and send the packets to the MFIB for regular multicast forwarding processing. The <ip mroute> command must be issued to configure a static route back to the head end router, thus enabling RPF checks.

The interface configurations that are needed for P2MP tunnels, are partly the same as in Section 3.4.2. Multicast routing and MPLS P2MP TE are enabled. PIM-SM is enabled on host facing interfaces. As said earlier, RPF checks need to be enabled to the head end. [31]

3.4.5 Ingress Provider Edge router configuration in Juniper routers

When configuring P2MP LSPs with Juniper routers, it is mainly done at ingress router. On transit and egress routers, RSVP and MPLS need to be enabled. Overall the whole procedure is very simple and takes only a few steps. The configuration starts by enabling RSVP and MPLS on interfaces involved in a P2MP LSP. Creating branch-LSPs (Juniper term) i.e. S2L sub-LSPs (see 2.13) to every egress router is equivalent to creating a destination list in Section 3.4.2. After defining the branch-LSPs they need to be associated with a primary P2MP LSP. The P2MP LSP does need a static path for SSM-traffic i.e. default SSM addresses have the P2MP LSP as next hop. The reason for this is that P2MP tunnels are not shown in the RIB, unlike P2P tunnels. After these

configurations it is possible to add other settings such as link protection or optimize timer. [32]

If dynamic routing is chosen, the CSPF handles the routing from ingress to egresses and no further configuration is needed. If instead, static routing is chosen the configuration is more complex and includes more steps but that solution is not covered in this thesis because the initial problem was too heavy implementation work load.

The IPTV service is very sensitive for interruptions in data stream so some kind of protection would be preferable. In Juniper routers for P2MP LSPs only link protection is provided. Enabling it is also very easy procedure. Only a <link-protection> configuration command under each [protocols mpls label-switched-path *lsp-name*] and on each interface which is wanted to be protected.

3.4.6 Egress PE router configuration in Juniper routers

Even though configuring P2MP LSPs requires only enabling RSVP and MPLS on the egress routers there are other things to consider too which are connected to P2MP LSPs. They include enabling PIM Sparse mode and IGMPv3 on host facing interfaces and defining SSM-groups under multicast routing-options if other address space than 232/8 is used.

3.5 ME: P2MP RSVP-TE, Core: P2MP RSVP-TE

This multicast solution for IPTV distribution is the main case of this thesis. This section introduces the procedures taken to create P2MP tunnels across the whole network. There are very few manuals or books about inter-vendor solutions and no instructions were found about how ME devices works with devices from other vendors. Followed by that, only a separate-P2MP tunnel case is introduced i.e. there are separate P2MP tunnels for the MEN and core network. The ideal solution would be having just one P2MP tunnel across the whole network which would be initiated and controlled from the core network. However, ME nodes will drop aggregated RSVP messages on the receive side if originated by another vendor's implementation. That means that at MEN each S2L LSP is signalled with one Path message.

3.5.1 ME device P2MP MPLS and LSP configuration outline

ME devices have a chassis mode which must be set right to get the wanted features. At first, the chassis mode has to be set right (3 or 4) on all MPLS nodes along the P2MP LSP. Second, an IGP is needed to distribute routes and if FRR is used, CSPF needs to be enabled.

After these prerequisites are met, the P2MP LSP is configured. Customers and services are configured the same way as in Section 3.2. A loose MPLS path is configured which follows the least cost path of IGP. An LSP is created with a keyword "p2mp-lsp" and a name is assigned to it. The LSP is set to <no shutdown> state. FRR is enabled if needed. A primary P2MP instance is configured with a keyword "primary-p2mp-instance" and a name is assigned to it. The primary P2MP instance is set to <no shutdown> state. Finally, within P2MP instance, loose S2L paths are defined to every receiver.

3.5.2 Cisco and Juniper P2MP MPLS configurations at core network

The core network P2MP tunnel configurations are exactly the same as in the case of Section 3.4.

4 Solution testing in laboratory environment

This chapter describes the laboratory topology and configurations made to devices. It is basically putting the solutions of the previous chapter into action. The different sections describe the individual laboratory cases in detail. Some sections have example configurations but the whole configurations related to creating the solutions, are presented in Appendix 1 and Appendix 2.

4.1 Laboratory settings

For identification purposes a loopback IP address, i.e. a node address, is assigned to each router and ME device. The addresses are from private IP address block and they are presented in Table 2 and Table 3. The MEN has 192.168.41.0/24 block and the core network has 192.168.40.0/24 block. Each router and ME device have an OSI address which is derived from the loopback IP addresses as introduced in 2.2. These addresses are not shown in this thesis but they are there for IS-IS.

Router name	IP address
c01	192.168.40.1
e01	192.168.40.2
e05	192.168.40.6
e06	192.168.40.248
e02	192.168.40.3

Table 2: Core routers and their loopback IP addresses

Device name	IP address
ME1	192.168.41.1
ME2	192.168.41.2
ME3	192.168.41.3

Table 3: ME devices and their loopback IP addresses

The test network topology is presented in Figure 35. The N2Xs (see 2.1.2) are the emulation devices which represent DSLAMs (see 2.1.2) and SHE (see 2.1.2). N2X/10 runs as SHE and it is connected to core router c01 port ge-0/2/2. N2X/7 connected to ME1 represents a DSLAM. The other N2Xs are connected to three edge routers e01, e02 and e06 and they act as check points when the tests are done. They provide extra information about what happens in failure cases e.g. where the data is rerouted.

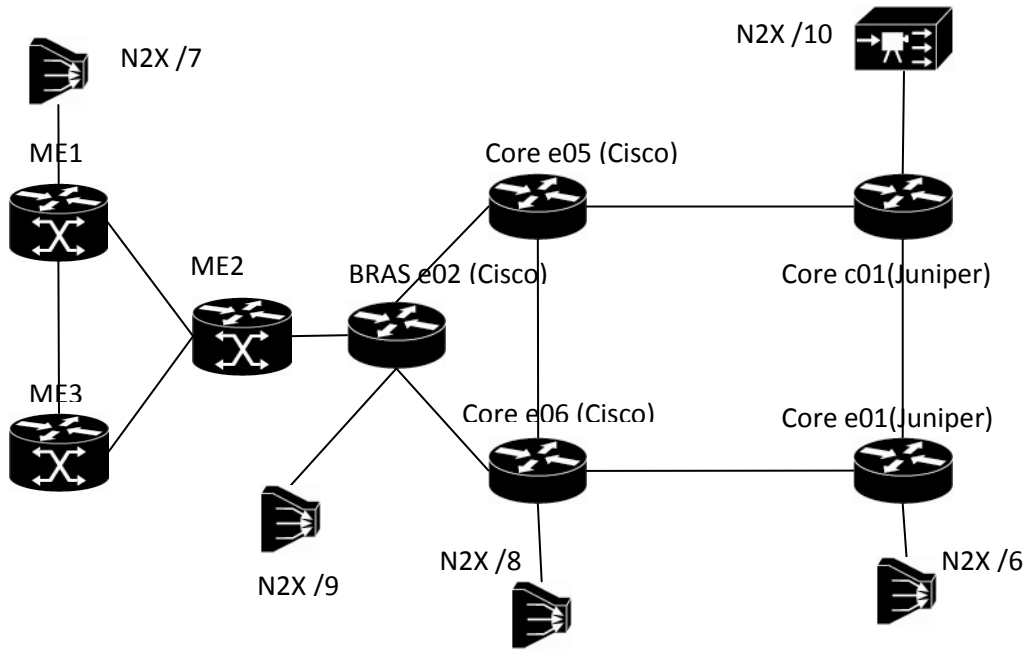


Figure 35: Laboratory topology

Figure 36 illustrates the interfaces used in the routers and ME devices. The "g" and "ge" are GigE interfaces and the "te" is a TenGigabit interface. The numbers after the interface types represent *line card/module/port* or *line card/port* of the devices. Figure 37 and Figure 38 illustrate the interface IP addresses in MEN and core networks.

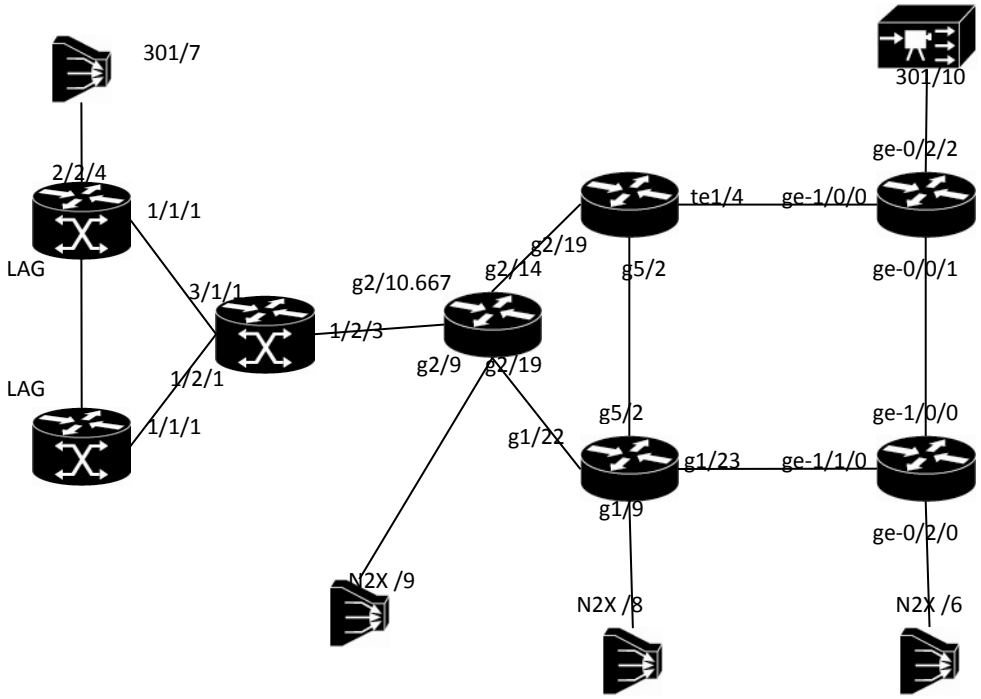


Figure 36: Interfaces

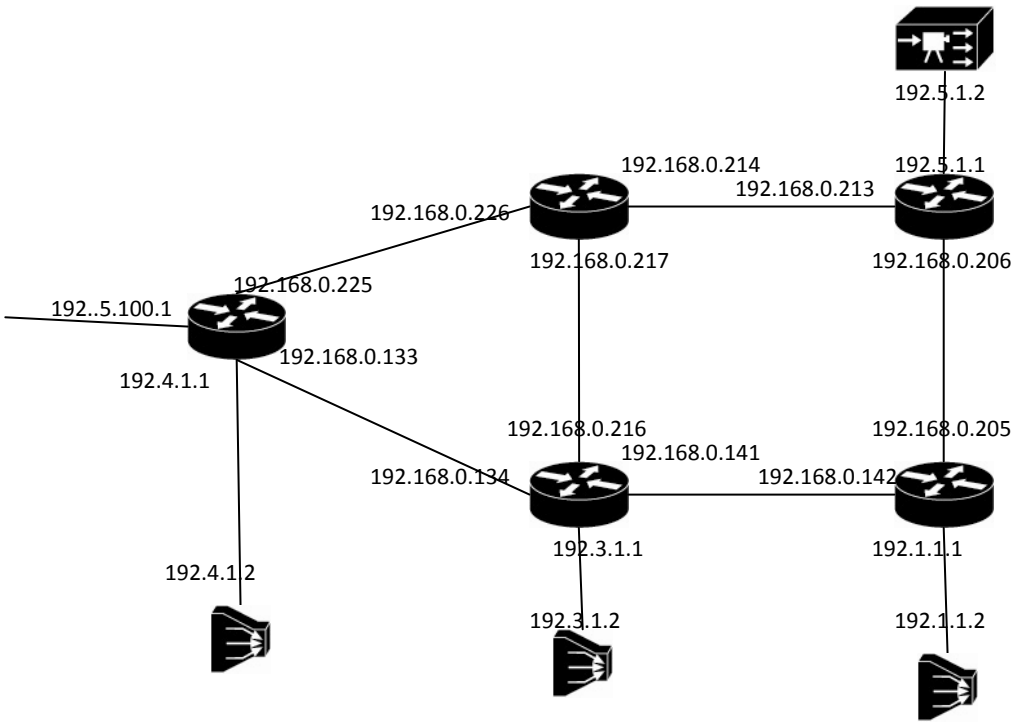


Figure 37: Interface IP addresses at core network

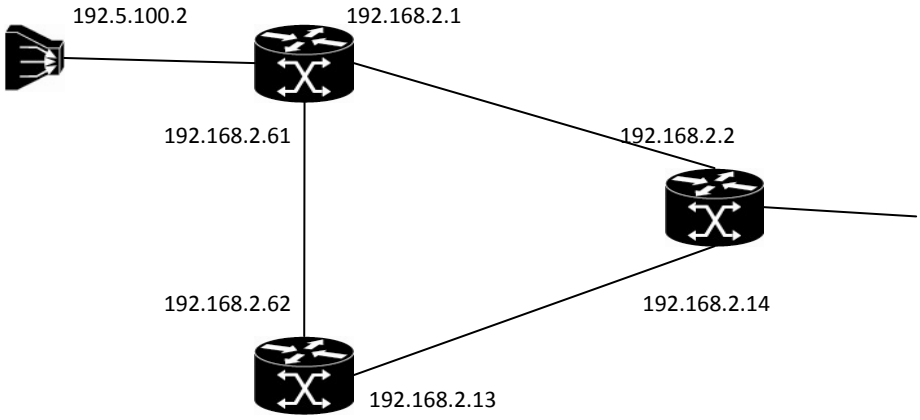


Figure 38: Network interface IP addresses at MEN

4.2 ME: IGMP Snooping & RSVP-TE, Core: PIM-SSM

This section introduces examples of the device configurations made while creating the test network.

4.2.1 MEN configurations

These example port configurations show e.g. port type, port mode and encapsulation type which is Ethernet. In both examples the port type is Ethernet and the port mode is access i.e. a service is connected to the port. The encapsulation type is dot1q i.e. a VLAN tag is included. The description gives information about the other end of the link. ME1 port 2/2/4 is the port facing the host end device and ME2 port 1/2/3 is facing the core network router.

ME1

```
port 2/2/4
        description "N2X_301/7"
        ethernet
            mode access
            encap-type dot1q
        exit
        no shutdown
    exit
```

ME2

```
port 1/2/3
        description "hkplabedge02 2/10"
        ethernet
            mode access
            encap-type dot1q
        exit
        no shutdown
    exit
```

4.2.2 Core configurations

The core configurations for PIM-SSM are very simple. In Juniper a PIM-SM group has been defined and that is set for each interface that is wanted to participate in PIM-SSM. The same is done for Cisco router interfaces. Cisco routers also need a SSM mapping if SSM is not supported by every host. Here is the configuration of e02 router interface ge-2/10.667:

```
interface GigabitEthernet2/10.667
    description p2mp_testi_mikko
    encapsulation dot1Q 667
    ip address 192.5.100.1 255.255.255.0
    ip pim sparse-mode
!
```

It can be seen that the interface is a sub-interface and it has VLAN tag 667 configured. This interface is connected to the ME2 port 1/2/3. The VLAN starts from e02 ge-2/10.667 and goes through to the ME1 port 2/2/4. Furthermore it can be seen that configuring PIM on an interface does not require much effort.

4.3 ME: IGMP Snooping & RSVP-TE with FRR, Core: PIM-SSM

4.3.1 Testing outline

The test result gathering is done in following steps. For each test case the same tests are done but with few exceptions. Those are explained later. The tests include link failure, node failure and routing engine switch over.

4.3.2 Link failure

Link failure is simulated by configuring an interface down and counting packet loss caused by traffic moving from the failed link to a functioning link. After the traffic has moved to the new link the original is configured up again and the packet loss is counted. Not every device or interface is tested, only relevant ones. Each interface test is repeated three times. In some cases a break of few minutes is required for the system to recover and before an interface can be configured up or down. Otherwise the traffic does not move successfully to another path. The reason for this kind of activity is not known but it does not matter in the scope of the thesis.
The link failure case is tested with and without link protection. That is also known as FRR facility back up method which is the only kind of protection provided for P2MP MPLS.

The link failure test is performed on the following devices and their interfaces: c01 (ge-1/0/0, ge-0/0/1), e05 (g5/2), e06 (g5/2), e02 (g2/14) and ME1 (1/1/1). They are illustrated with red link colour in Figure 39.

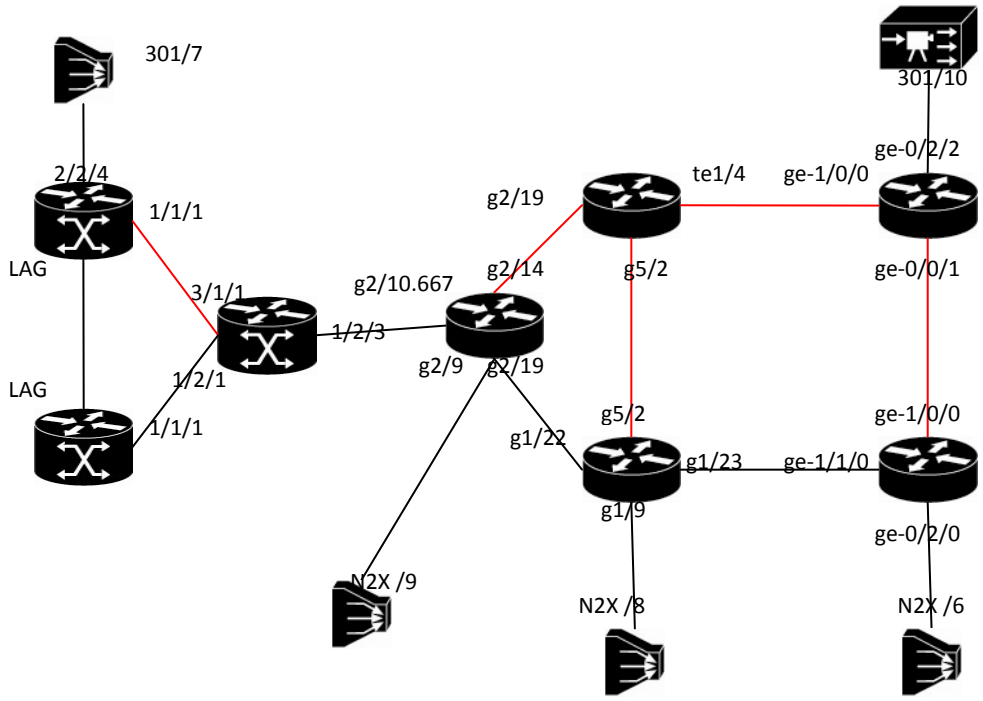


Figure 39: Link failure tested interfaces marked with red colour.

4.3.3 Node failure

In the node failure test the failure is simulated by booting the tested router. As in link failure test, the packet loss, caused by the booting, is measured. This test is done only once because it affects many other connections and tests being run in the laboratory. As in the link failure tests, in some cases the system needs some time to recover before a new test.

The node failure test is done to Juniper router e01**.** All the incoming traffic is routed through interface ge-0/0/1.

4.3.4 Routing engine switch over

The routing engine switch over test is done to Juniper router e01. All the incoming traffic is routed through interface ge-0/0/1. This test gives information about packet loss when routing engine is switched. This test is done twice, once without and once with an overload bit. The overload bit informs adjacent nodes to ignore the node that has the overload bit. That is used for situations like the routing engine switch over. This way traffic is automatically moved to another link for the time the overload bit is set and packet loss is avoided. After the switch over the overload bit is removed and hopefully the traffic returns to the initial link.

4.4 ME: IGMP Snooping & RSVP-TE, Core: P2MP RSVP-TE

4.4.1 N2X configurations

N2X end device emulators are connected to ME1, BRAS, e06, e01 and c01. The N2X port attached to c01 works as a source and the rest of the N2X ports act as receivers. They send SSM Join messages to groups 239.0.0.0 – 239.0.0.9 and have the senders IP address as the source address (see Section 2.10.2). The source sends 10000 packets per second which allows easy down time calculation. The received packets are calculated live per stream e.g. core01 – edge01. The operating system displays stream information such as transported and received test packets, packet loss, delay and so on.

4.4.2 MEN configurations

MEN devices ME1, ME2 and ME3 have IS-IS, RSVP, MPLS and LDP configured on. They have interfaces towards the network and ME2 has an interface towards the core network, too. Each ME device has a Service Access Point (SAP) as a service end point in MEN. ME1 and ME2 have Service Distribution Point (SDP) for connecting SAPs, too. The connected SAPs form a service, a VPLS in this case. The VPLS uses LSPs for forwarding packets. The LSPs are strict hop LSPs and configured in both ways. They are signalled by RSVP-TE. ME2 is the head end device in MEN and the LSP from ME2 to ME1, LSP21, is protected by the FRR. The FRR is configured at both ends of the LSP and also requires CSPF. However, there are not any constraints configured to CSPF. In case of link failure (node protection is not possible because the network is so small that there are not any detours available to bypass the node ME2) the FRR finds a detour via ME3 and switches the traffic to the detour. ME3 does not require any

additional configurations to provide the detour. IGMP Snooping is enabled in VPLS context to guide the multicast traffic to the right receivers.

4.4.3 Core configurations

On head end router "optimize-timer" is set to 10 seconds under [protocols mpls label-switched-path *branch-to-edgex*], where x is 1, 2 and 6. The MPLS tunnel configuration of c01 router i.e. the tunnel head end is simple:

```
mpls {
      traffic-engineering bgp-igp-both-ribs;
      no-propagate-ttl;
      optimize-timer 10;
      label-switched-path branch_to_edge2 {
      to 192.168.40.3;
      optimize-timer 10;
      link-protection;
      p2mp p2mp_test_from_core1;
      }
      label-switched-path branch_to_edge6 {
      to 192.168.40.248;
      optimize-timer 10;
      link-protection;
      p2mp p2mp_test_from_core1;
      }
      label-switched-path branch_to_edge1 {
      to 192.168.40.2;
      optimize-timer 10;
      link-protection;
      p2mp p2mp_test_from_core1;
      }
      path p2mp_test_from_core1;
}
```

As stated in 3.4.5 the configuration includes defining a path which here is *p2mp_test_form_core1*. Then *label-switched-path*s are created to each egress router and they are named. Each *label-switched-path* uses the *p2mp_test_form_core1* path.

5 Test results and evaluation

This chapter introduces results of tests made according to Chapter 0. The first case is the current solution for distributing data from the SHE to the customers. The data is first brought to the BRAS and from there it is distributed over the MEN to the customers. The second one is the new solution with P2MP RSVP-TE at core network. In the last case a new function in MEN, FRR, is tested.

5.1 ME: IGMP Snooping & RSVP-TE, Core: PIM-SSM

Initially traffic is routed through the network via multiple paths. Those paths are shown in Figure 40. They are controlled with IS-IS link metrics which are also shown in the same figure.

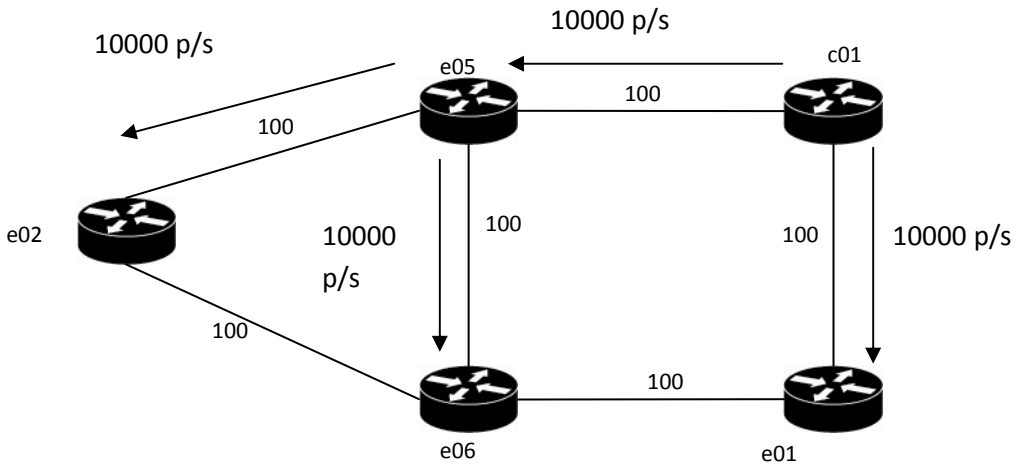


Figure 40: Initial traffic routing in the PIM-SSM tests

5.1.1 Link failure

Link failure was tested by configuring interfaces down one at a time and counting packet loss experienced by N2X emulation device. In tables showing the test results, each <shutdown> represents deactivating an interface and each <no shutdown> represents the interface activating again.

Table 4 shows that an interface going down causes a packet loss of ca. 5500 or less which equals half a second interruption or less. When the interface comes up again, the packet loss is almost 20000 packets i.e. almost two seconds. The paths taken by the traffic after the interface had gone down and came up again are shown in Figure 41 and Figure 40.

Event	Packet loss
shutdown	4113
no shutdown	18170
shutdown	5454
no shutdown	18063
shutdown	2256
no shutdown	18579

Table 4 : Packet loss on N2X port 301/6 when interface ge-0/0/1 at c01 was down

When interface ge-0/0/1 was down traffic took the paths shown in Figure 41.

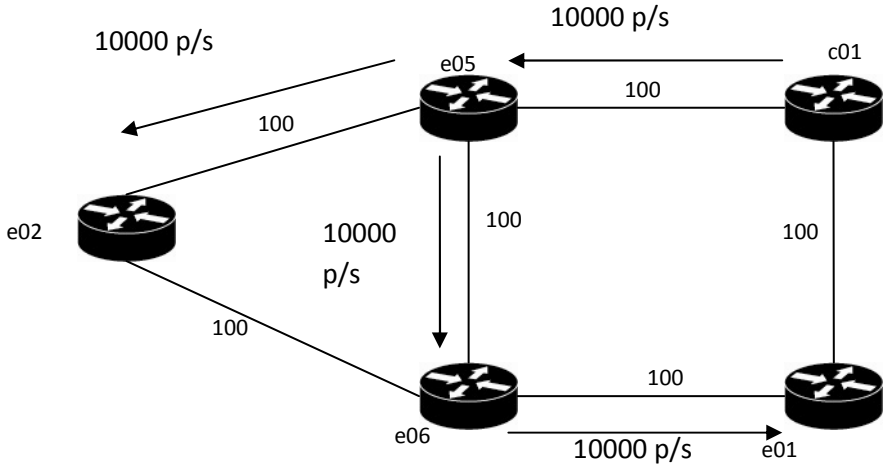


Figure 41: Paths taken by traffic when interface ge-0/0/1 was down

Table 5 shows packet loss on effected N2X ports when deactivating and activating interface ge-1/0/0 at c01. After deactivating the interface the path taken by traffic is the one in Figure 42. After activation, traffic returned to the paths shown in Figure 40.

Event	Packet loss
shutdown	~2000
no shutdown	~550
shutdown	1990
no shutdown	23203
shutdown	1552
no shutdown	24189

Table 5: Packet loss on N2X ports 301/7, 301/8 and 301/9 when interface ge-1/0/0 at c01 was down

These results are otherwise practically the same as in Table 4 but in the first test when the interface came up, it was much faster. It is an indication that the forwarding state was not fully changed and because of that the interface was up so fast. The result is excluded as an outlier.

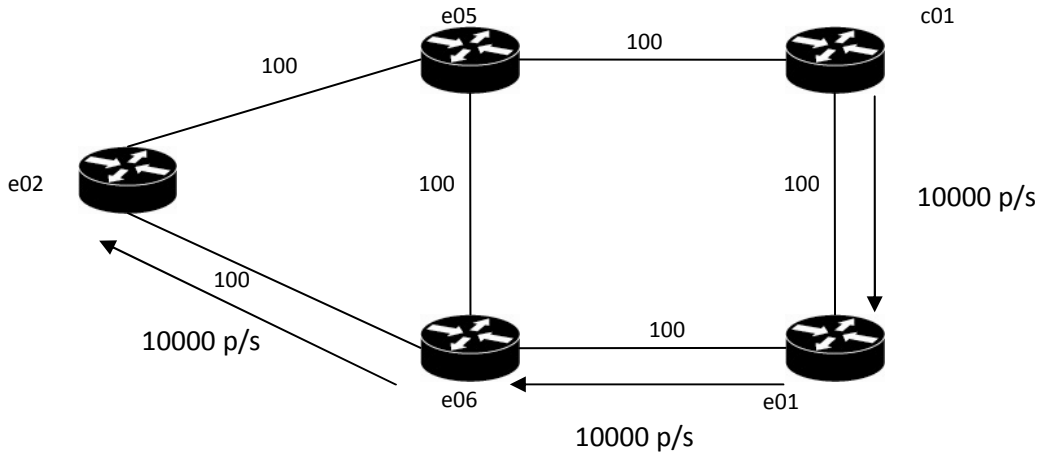


Figure 42: Path taken when c01 interface ge-1/0/0 was down

When testing the effect of link failure in interfaces ge-1/0/0 of e01 and te1/4 of e05 the results were practically the same as when testing with interfaces on the other end of the links which was expected.

The link failure tests were also made to e05 interface g5/2 and e06 interface g5/2. Traffic naturally moved in both cases to a new path which is presented in Figure 43.
The test results are shown in Table 6 and Table 7.

Event	Packet loss
shutdown	2054
no shutdown	19425
shutdown	2151
no shutdown	1139
shutdown	2205
no shutdown	1093

Table 6: Interface g5/2 down at e05 and packet loss caused by it at N2X port 301/8

Event	Packet loss
shutdown	2330
no shutdown	1231
shutdown	2384
no shutdown	1098
shutdown	2430
no shutdown	1428

Table 7: Interface g5/2 down at e06 and packet loss at N2x port 301/8

Configuring the interface down on either end of the link causes approximately the same packet loss. There can be seen an exception in Table 6: Interface g5/2 down at e05 and packet loss caused by it at N2X port 301/8 with the first <no shutdown> but that result can be ruled out because it is significantly different from the other results.

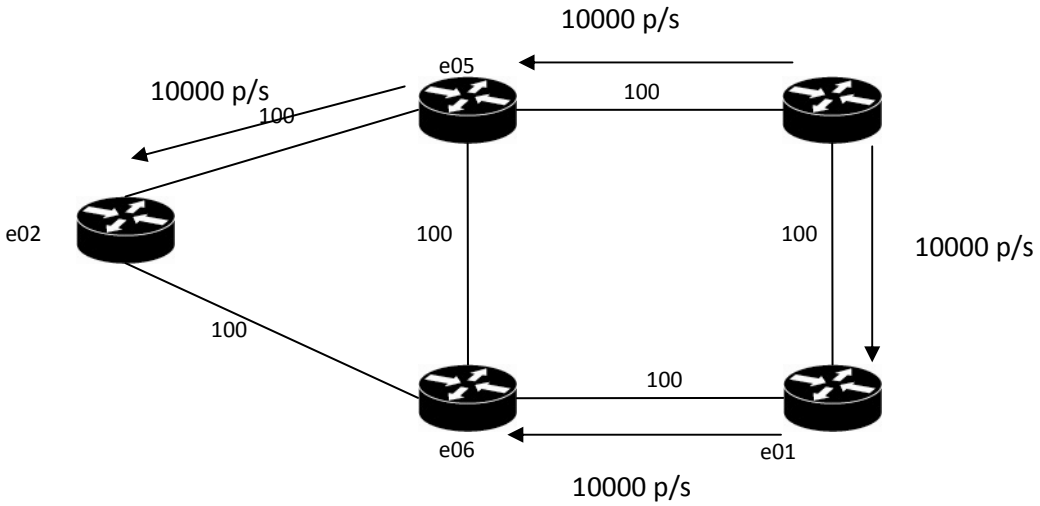


Figure 43: Traffic going around link 192.168.0.216/30 (e05 – e06)

The last link failure test was done to the link 192.168.0.224/30 by shutting down interface g2/14 of e02. The test results are shown in Table 8.

When the link failure tests were made there was a problem in getting the traffic back to the initial paths. If shutdown/ no shutdown commands were made too quickly, the traffic stopped flowing. The problem was solved by letting the network elements to recover fully after a change in the topology. This was done by waiting for about 10 minutes after every change in the interface state. This problem did not occur with any other link failure test and it remains unknown why the problem occurred with this link. However, when considering real production networks and link failure in them, it is unlikely that a network will face such fast changes that the problem would also occur there.

Event	Packet loss
shutdown	2083
no shutdown	20328
shutdown	2050
no shutdown	22574
shutdown	2115
no shutdown	20978

Table 8: Packet loss of N2X ports 301/7 and 301/8 when interface g2/14 of e02 was down

5.1.2 Node failure

The node failure test was done by rebooting a node and packet loss was counted. The node reboot was done to two Juniper routers c01 and e01. The results are presented in Table 9. The test was done only once.

Rebooting c01 obviously causes stopping of data flow in every receiver because c01 is a common point for every sub-LSP and there is no way around it. Furthermore, the reboot procedure is time consuming taking over six minutes for a router to become functional again. If the two second brake in the data stream in the earlier cases was somewhat annoying but also acceptable, this could be an extremely bad situation. The worst case scenario could be that the node would be rebooted for some reason during a final of a popular TV-show or a sports event. Fortunately, this kind of maneuver can be done during night time.

N2X port	Packet loss
301/6	3980669
301/7	1
301/8	1
301/9	1

Table 9: Packet loss at different N2X ports when rebooting e01

When rebooting some other node such as e01, the effect is much less drastic because at least some of the data can be rerouted. The port 301/6 is obviously affected because the only route to the port is via e01.

5.1.3 Routing engine switch over

The routing engine switch over test was done to e01 and packet loss was counted. The test results are shown in Table 10. As with node reboot test, this test was done only once too.

N2X port	Packet loss
301/6	1605451
301/7	1
301/8	1
301/9	1

Table 10: Routing engine switch over at e01 and its effects on N2X ports

The routing engine switch over affects the network and customers the same way as the node reboot but the node is up and running in ca. half the time of rebooting. Even though it is faster than the reboot, it is still a very long time for the customer to wait for three to four minutes to continue with the program.

5.2 ME: IGMP Snooping & RSVP-TE, Core: P2MP RSVP-TE

The initial setup is presented in Figure 44. Every link has a metric of 100 and the route chosen is the one with the least amount of hops from the source to the destination. Each

link that has traffic carries 10000 packets/s. Later, another setup is introduced, where the traffic is routed through another path in order to get multiple results with one test.

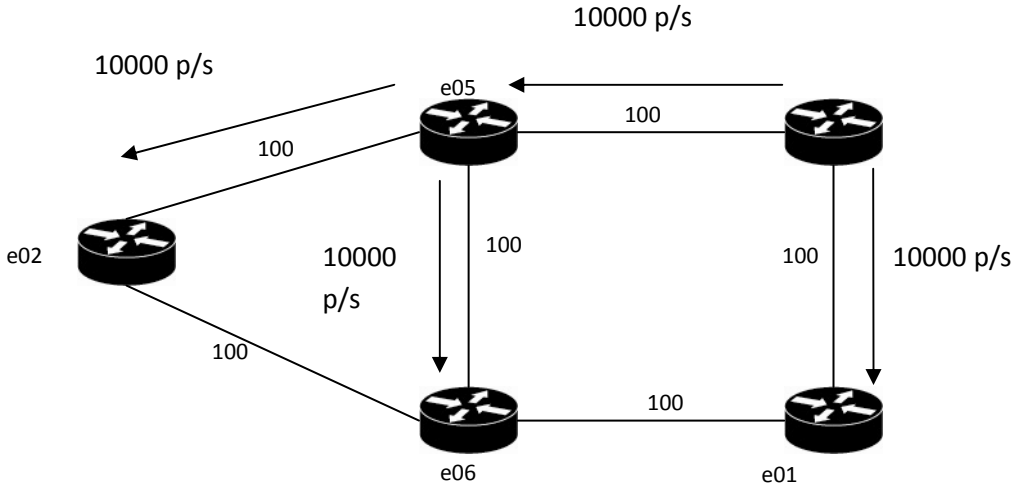


Figure 44: Setup with every link with metric of 100

5.2.1 Link failure

The test results are presented in Table 11-Table 14. Each interface shutdown causes practically the same packet loss. In Table 11 there is a clear outlier (the second <shutdown>) but other than that the test results are within ca. one second. Compared to results in the earlier section, the positive thing with these results is that the interface returning up again causes zero packet loss.

Event	Packet loss
shutdown	10878
no shutdown	0
shutdown	21057
no shutdown	0
shutdown	11346
no shutdown	0

Table 11: e02 interface g2/14

Event	Packet loss
shutdown	10194
no shutdown	0
shutdown	10517
no shutdown	0
shutdown	10601
no shutdown	0

Table 12: e05 interface g2/19

Event	Packet loss
shutdown	10567
no shutdown	0
shutdown	10640
no shutdown	0
shutdown	10617
no shutdown	0

Table 13: e05 interface te1/4

Event	Packet loss
shutdown	12070
no shutdown	0
shutdown	12204
no shutdown	0
shutdown	12410
no shutdown	0

Table 14: c01 interface ge-1/0/0

5.2.2 Node failure and routing engine switch over

The c01 was booted with command <request system reboot> and the routing engine switched from re1 to re0. It takes ca. 6 minutes for the router to come up again.

After the boot from re0, the traffic does not return and the device remains at re0. The traffic has to be stopped and the device must be booted again to return the traffic to flow end-to-end. Overall this test does not give much information. It just tells how long it takes for a router to come up after a boot. However, it also tells how the router and routing engines work when booting the router. In this case the bad thing is that booting from re0 the traffic does not start flowing without the stated procedures.

When e01 is tested, c01 interface ge-1/0/0 metric increased to 300 so that the traffic flows only via interface ge-0/0/1. Also the metric of the interface te1/4 at the opposite end is increased to 300. This setup is illustrated in Figure 45.

When giving command <request system reboot> at e01 re0, N2X port 301/7 does not receive packets for 63 seconds. After that the traffic moves to c01 interface ge-1/0/0. N2X port 301/6 obviously comes up again after ca. 6 minutes and 20 seconds and re0 is still the master routing engine. After that 10000 packets/s return to the initial path but another 10000 packets/s stay on the new path.

Doing the routing engine switch over instead of booting creates very similar results but N2X port 301/7 does not receive traffic for ca. 6 seconds so it recovers significantly faster.

If the boot is done at routing engine1, the recovery times are approximately the same as with the other routing engine but in this case the routing engine switches. Furthermore, the traffic returns fully to the initial path when the node is up again.

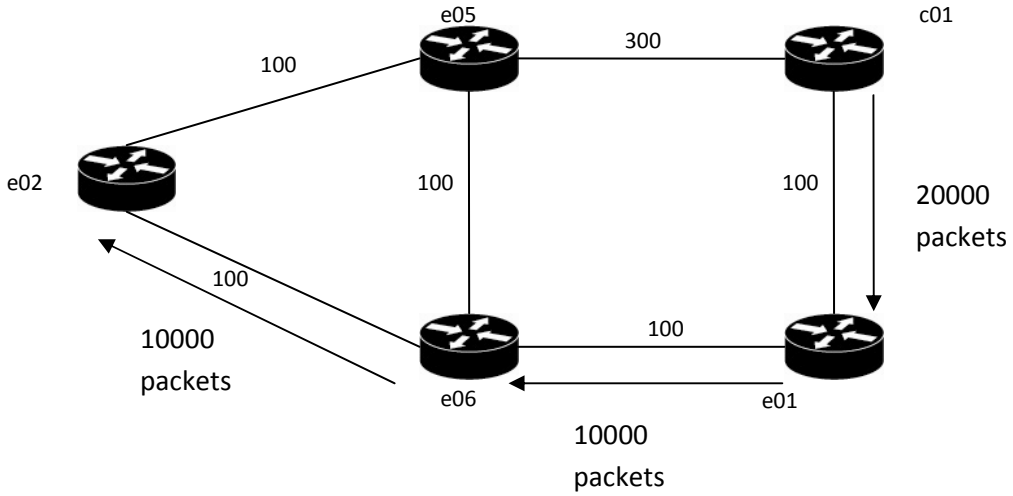


Figure 45: Setup where traffic is routed through only one path

5.2.3 Node reboot and routing engine switch over with overload bit

Here, the same procedures are done as in previous tests with the exception that the node to be booted or its routing engine being switched has an overload bit set. The overload bit tells the neighbouring nodes to ignore it as a neighbour. This way, hopefully, the node reboot or routing engine switch over does minimum interruption to the traffic flow.

The traffic was configured to traverse only one path i.e. as in Figure 45. Because the traffic takes this path, the only sensible choice for testing the overload bit is e01. The other Juniper node is the head end and it is obvious that bringing that node down would seize all the traffic with or without the overload bit. The remaining nodes on the path are Cisco routers which do not support the overload bit so the only option is the e01 router.

The first reboot test with the overload bit was made with the routing engine at e01 being re1. When the overload bit was configured, half of the traffic immediately moved to the interface ge-1/0/0 as in Figure 44. When the node e01 went down the interface ge-0/0/1 had no traffic on it. After 5 minutes and 30 seconds the node was up again and the traffic was on again as in Figure 44. At this point, N2X-ports 301/7, 301/8 and 301/9 experienced no packet loss. When the overload bit was removed from the e01 traffic returned back to the initial path i.e. as in Figure 45. After the reboot the node had re0 as the primary routing engine.

When the same test was made with the routing engine being re0 the exact same steps took place even though it took 6 minutes and 4 seconds for e01 to reboot and the routing engine did not switch but remained at re0. The main difference between these two tests was that when the overload bit removed from e01, the traffic did not return to the initial form but stayed as in Figure 44.

Also the routing engine switch over with the overload bit was tested with both routing engines. First, switching from re0 to re1 was tested. As in node reboot tests, traffic

moved immediately from path of Figure 45 to paths of Figure 44. It took 2 minutes and 21 seconds for the node to switch the routing engine and while switching there was no packet loss on N2X ports except on port 301/6. When the overload bit was removed from e01 traffic did not return to the initial path but remained on paths of Figure 44. Traffic was returned to only one path and the test was repeated with re1 to be switched to re0. This time it took 3 minutes and 20 seconds to switch over. When the overload bit was set, N2X ports 301/7, 301/8, 301/9 lost 526 packets, which equals a 52ms of interruption. After removing the overload bit the traffic returned to the path of Figure 45.

5.2.4 Link protection

The link protection tests were made by protecting each P2MP branch LSP and each RSVP interface involved in P2MP MPLS.

Event	Packet loss
shutdown	412
no shutdown	0
shutdown	382
no shutdown	0

Table 15: Packet loss on N2X ports

Test results of link protected branch LSPs are shown in Table 15. Each <shutdown> is a deactivation of the interface ge-0/0/1 at c01 and each <no shutdown> is an activation of the same interface. Every port (301/6, /7, /8, /9) had the same packet loss that is presented in the second column.

When the interface was activated the traffic flowed through one LSP, as in Figure 45. When the interface was deactivated the traffic moved to interface ge-1/0/0 of c01. The link protection worked as it was supposed to i.e. another route was calculated from c01 to e01. The link protected route can be seen in Figure 46. When the interface ge-0/0/1 was up again the traffic returned to flow through that interface and one LSP.

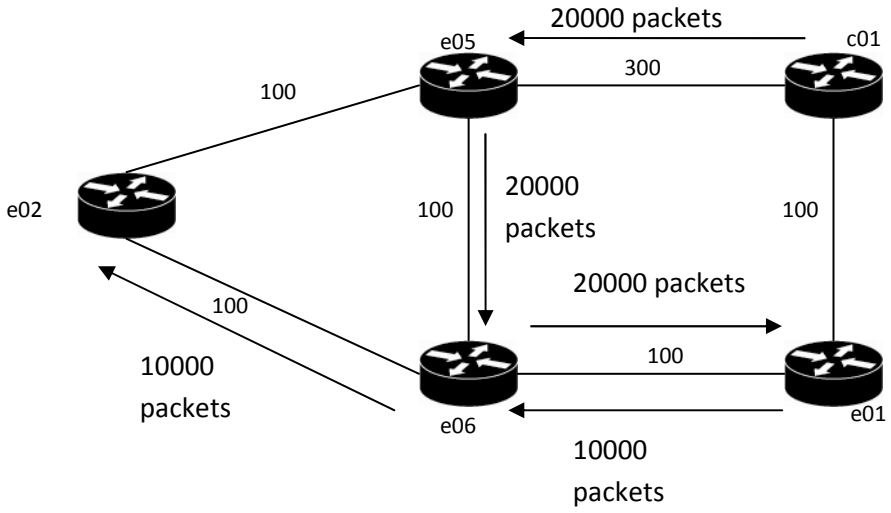


Figure 46: Link protected LSP between c01 and e01

After the second repetition of the test, IS-IS started to flap i.e. IS-IS kept sending new routes about every 45 seconds and caused 1 to 5 second interruptions. The configuration that caused this activity was found to be the link protection configuration under IS-IS. The problem occurred only with one interface i.e. ge-0/0/1. This problem could not be solved and the tests had to be stopped.

The results this far had however, been promising because the link protection had provided the ~50ms recovery time that the vendor and standard had promised. Also, many other interfaces had the same configuration and had no problems with it. If the cause of the problem is solved and the tests can be made again, this could be a good feature to use.

5.3 ME: IGMP Snooping & RSVP-TE with FRR, Core: PIM-SSM

This test case is practically just a ME FRR test. It does not really matter what protocols are used in the core network. ME MPLS signalling protocol could also be LDP but because the original idea of the thesis was to test RSVP-TE and its extensions (mainly the P2MP-feature) in the MEN, RSVP-TE is chosen.

5.3.1 Link failure

Link failure test is the only test done with FRR because there was not enough test equipment to test other kinds of failures.

In Table 16 packet loss is shown in the second column. The packet rate of the stream is still the same 10000 p/s so the loss caused by the link failure is less than 50 ms which is what the vendor had promised.

Event	Packet loss
shutdown	295
no shutdown	0
shutdown	185
no shutdown	0
shutdown	297
no shutdown	0

Table 16: Packet loss in MEN with FRR

6 Conclusions

In this section the results from previous section and the initial goal settings from Section 1.2 are brought together and discussed.

During the journey from examining the initial problem, which was to create a better solution for configuring MEN P2MP trees, to creating test cases for laboratory and testing them it became apparent that the current solution was used for a good reason. Even though the configuration of the P2MP MPLS tree with RSVP-TE is slow and requires lots of work when topology changes are needed it appears to be the only option available. The MEN devices not supporting the newer line cards which have the P2MP RSVP-TE protocol was apparently a good reason to implement the signalling the current way. Furthermore, the new MEN devices which support the new line cards require a different chassis mode to have the new features enabled. When changing the chassis mode to support the new cards, the old ones stopped working. The old cards have lots of configured services and it was too much work to move each service to the new card within given time limits.

Even though there was no solution to the initial problem, the laboratory tests showed promising results when testing FRR in MEN (5.3). Using FRR in MEN MPLS tunnels would reduce the workload almost 50 percent because the secondary paths would be signalled dynamically. This could seem odd that it is possible to create the secondary paths dynamically but primary paths would have to be signalled manually. The reason is that there are customers attached to almost every device on the primary path. If a loose mode primary path was made, it would be a path from a source to a receiver and customers could be attached to it at the receiver end only.

The MEN network not supporting P2MP RSVP-TE caused another setback. The interoperation test for different vendor implementations of P2MP RSVP-TE could not be made. This would have been interesting because if the results were good the P2MP paths could have been created over IS-IS area borders. That would have enabled P2MP tree controlling from one source i.e. from the core network.

The initial problem remains unsolved but on the core network side there was interesting tests to be made too. The core network routers support P2MP RSVP-TE and the tests that could not be made on the MEN side could be made on the core side. P2MP RSVP-TE proved to have its pros and cons too. When finally the laboratory problems were solved and the P2MP tunnels were up and running, even though the topology had to be

shortened, the P2MP MPLS tree proved to have some good features. Compared to PIM-SSM it took practically the same time to configure them up and in addition the P2MP tunnels recovered faster in failure cases. In some devices there were difficulties getting the tree up after a node reboot and the reason for that remained unknown. Furthermore, many extensions such as link protection and Loop Free Alternates (LFA [33]) did not work with P2MP tunnels. The link protection feature showed promising results with sub-50 millisecond recovery times but when it was configured on it also made the IS-IS flap. If that problem is solved then P2MP RSVP-TE would be better than PIM-SSM which does not have any protection features. However, the tests were made in a laboratory network which is significantly smaller than the production network. Furthermore, the devices and software versions were different than in production network so the results gained in this thesis are not directly applicable to the production.

At the moment it seems that the current solution is the best available excluding the MEN secondary path configuration. If FRR would be used it would make the configuration faster and the protocol works well too. However, it was tested in the laboratory and implementation in the production network could be different. If at some point in the future the new chassis are purchased it would be very interesting to test the P2MP feature in MEN. Basically, it seems to be a lot faster to configure and easier to maintain from the implementation engineer's perspective. On the other hand that would possibly increase the MEN designers' work load because they would have to consider the routes the P2MP tree would choose based on IS-IS metric. An IPTV distribution tree changing its form when ever it is needed could possibly be quite a thing to handle. It takes so much bandwidth that if the tree moved to a smaller link it could block every other service. Fortunately P2MP RSVP-TE, as the name implies, has TE abilities. By using simple bandwidth constraints the service blocking situations could be avoided. This is naturally just speculation and it is likely that this is not relevant in the near future.

Also on the core network side the current solution with PIM-SSM seems at this point to be reliable enough and easy to implement. However, things have to evolve and based on the tests it seems likely that with a few updates to P2MP RSVP-TE, which would make it more reliable in failure cases (see 5.2.2) and more compatible with protection methods such as FRR and LFA. In addition, P2MP RSVP-TE was easy to use and if functioning link protection is provided it recovers very quickly from failure too. However, it is not important how easy the configuration is because the size of the network is quite small if the amount of devices is considered. On the other hand the reliability and fast recovery from failures is very important on the core side because it affects a large area such as a MEN area. For this purpose, P2MP RSVP-TE would fit well if the link protection features work.

Overall, the laboratory work and test results provided as many questions as it did answers but the positive thing is that in the past and present assets were well used.

7 References

[1] Jeffrey S. Beasley. 2009. Networking 2nd Edition

[2] R. Braden. 1989. Requirements for Internet Hosts -- Communication Layers. RFC 1122

[3] IANA. 2003. IPv4 Address Space Registry. http://www.iana.org/assignments/ipv4-address-space/ipv4-address-space.xml

[4] IEEE Standard for Information Technology. 2008. Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications.

[5] L. Andersson, T. Madsen. 2005. Provider Provisioned Virtual Private Network (VPN) Terminology. RFC 4026

[6]Girish Chiruvolu, Bill Krogfoss, Andrew Ge. 2004. Encapsulation schemes to extend Ethernet to Metropolitan Area Networks, A comprehensive analysis of popular and evolving, encapsulation schemes for Metro-Ethernet. White paper

[7] VPN Consortium. 2008. VPN Technologies: Definitions and Requirements. http://www.vpnc.org/vpn-technologies.html

[8] S. Hanks, T. Li, D. Farinacci, P. Traina.1994. Generic Routing Encapsulation. RFC 1701

[9] E.W. Dijkstra. A Note on Two Problems in Connexion with Graphs. In: Numerische Mathematik 1. 1959, pages 269-271

[10] draft: Manayya KB. 2009. Constrained Shortest Path First. Internet-Draft, draft-manayya-constrained-shortest-path-first-00, Internet Engineering Task Force. Work in Progress

[11] Hannes Gredler, Walter Goralski. 2005. The Complete IS-IS Routing Protocol

[12] Ivan Pepelnjak, Jim Guichard. MPLS and VPN architectures

[13] E. Rosen, A Viswanathan, R. Callon. 2001. Multiprotocol Label Switching Architecture. RFC 3031

[14] E. Rosen, D. Tappan, G.Fedorkow. 2001. MPLS Lable Stack Encoding. RFC 3032

[15] L. Andersson, I. Minei, B. Thomas. 2007. LDP Specification. RFC 3036

[16] D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, G. Swallow. 2001. RSVP-TE: Extension to RSVP for LSP Tunnels. RFC 3209

[17] J. Postel. 1981. Transmission Control Protocol. RFC 793

[18] J. Postel. 1980. Used Datagram Protocol. RFC 768

[19] Brian M. Edwards, Leonard A. Giuliano, Brian R. Wright. 2002. Interdomain Multicast Routing: Practical Juniper Networks and Cisco Systems Solutions

[20] S. Deering. 1989. Host Extensions for IP Multicasting. RFC 1112

[21] W. Fenner. 1997. Internet Group Management Protocol, Version 2. RFC 2236

[22] B. Cain, S. Deering, I. Kouvelas, B. Fenner, A. Thyagarajan. 2002. Internet Group Management Protocol, Version 3. RFC 3376

[23] M. Christensen, K. Kimball, F. Solensky. 2006. Considerations for Internet group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches. RFC 4541

[24] B. Fenner, M. Handley, H. Holbrook, I Kouvelas. 2006. Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised). RFC 4601

[25] S. Bhattacharyya. 2003. An Overview of Source-Specific Multicast (SSM). RFC 3569

[26] R. Aggrawal, D. Papadimitriou. 2007. Extensions to Resource Reservation Protocol – Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs). RFC4875

[27] draft: I. Minei, K. Kompella, I. Wijnands, B. Thomas. 2010. Label Distribution Protocol Extension for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths. Internet-Draft, draft-ietf-mpls-ldp-p2mp-11, Internet Engineering Task Force. Work in Progress

[28] P. Pan, G. Swallow, A. Atlas. 2005. Fast Reroute Extensions to RSVP-TE for LSP Tunnels. RFC 4090

[29] Cisco IOS Configuration Guide: Configuring Source Specific Multicast. http://www.cisco.com/en/US/docs/ios/12_2/ip/configuration/guide/1cfssm.pdf

[30] Juniper JunOS 9.6 Configuration Guide Multicast: Multicast PIM SSM http://www.juniper.fr/techpubs/en_US/junos9.6/information-products/topic-collections/config-guide-multicast/multicast-pim-ssm.html

[31] Cisco IOS Configuration Guide: MPLS Point-to-Multipoint Traffic Engineering http://www.cisco.com/en/US/docs/ios/mpls/configuration/guide/mp_te_p2mp.html

[32] Juniper JunOS 9.3Configuration Guide: MPLS Applications Configuration guide http://www.juniper.net/techpubs/software/junos/junos93/swconfig-mpls-apps/frameset.html

[33] A. Atlas, A. Zinin. 2008. Basic Specification for IP Fast Reroute: Loop-Free Alternates. RFC 5286

Appendix 1: Point to Multipoint configurations

c01 P2MP Configurations

```
protocols mpls {
    traffic-engineering bgp-igp-both-ribs;
    no-propagate-ttl;
    optimize-timer 10;
    label-switched-path branch_to_edge2 {
        to 192.168.40.3;
        optimize-timer 10;
        p2mp p2mp_test_from_core1;
    }
    label-switched-path branch_to_edge6 {
        to 192.168.40.248;
        optimize-timer 10;
        p2mp p2mp_test_from_core1;
    }
    label-switched-path branch_to_edge1 {
        to 192.168.40.2;
        optimize-timer 10;
        p2mp p2mp_test_from_core1;
    }
    path p2mp_test_from_core1;

    interface ge-0/0/1
    interface ge-1/0/0
}
```

The other Juniper router e01 has only MPLS enabled interfaces configured because it is a tail end of the P2MP tree.

e02 P2MP Configurations

```
interface GigabitEthernet2/10.667
 description p2mp_testi_mikko
 encapsulation dot1Q 667
 ip address 192.5.100.1 255.255.255.0
 ip pim sparse-mode
!

interface GigabitEthernet2/9
 description to_301/9
```

```
 ip address 192.4.1.1 255.255.255.0
 ip pim sparse-mode
!

interface GigabitEthernet2/14
 description to-hkplabedger05
 bandwidth 1000000
 ip address 192.168.0.225 255.255.255.252
 ip router isis TSF-core-igp
 mpls ip
 mpls traffic-eng tunnels
 isis circuit-type level-2-only
 isis network point-to-point
 isis metric 100 level-2
 ip rsvp bandwidth percent 100
 ip rsvp signalling hello graceful-restart
!

interface GigabitEthernet2/19
 description hkplabedger06-gi1/22
 bandwidth 1000000
 ip address 192.168.0.133 255.255.255.252
 ip router isis TSF-core-igp
 mpls ip
 mpls label protocol ldp
 mpls traffic-eng tunnels
 isis circuit-type level-2-only
 isis network point-to-point
 isis metric 100 level-2
 ip rsvp bandwidth percent 100
 ip rsvp signalling hello graceful-restart
!

mls ip multicast replication-mode egress
ip multicast routing
mpls traffic-eng tunnels
ip multicast mpls traffic-eng
ip pim ssm default
ip mroute 192.5.1.0 255.255.255.0 192.168.40.1

router isis
 metric-style wide
 mpls traffic-eng router-id Loopback0
 mpls traffic-eng level-2
!
```

The other tail end routers are configured in a similar way. Interface g2/10.667 is the only exception because it faces the MEN.

Appendix 2: Protocol Independent Multicast configurations

c01 PIM Configurations

```
protocols pim {
  interface ge-0/2/2.0 {
    apply-groups PIM-SM;
  }
  interface ge-1/0/0.0 {
    apply-groups PIM-SM;
  }
  interface ge-0/0/1.0 {
    apply-groups PIM-SM;
  }
  join-load-balance;
}

routing-options multicast
ssm-groups [239.0.0.0/8 ];
```

e01 PIM Configurations

```
interface ge-0/2/0.0 {
    apply-groups PIM-SM;
    mode sparse;
    version 2;
    hello-interval 1;
    bfd-liveness-detection {
        minimum-interval 100;
        multiplier 3;
    }
}

interface ge-1/1/0.0 {
    apply-groups PIM-SM;
    hello-interval 1;
    bfd-liveness-detection {
        minimum-interval 100;
        multiplier 3;
    }
}

interface ge-1/0/0.0 {
    apply-groups PIM-SM;
```

```
}

routing-options multicast
ssm-groups [239.0.0.0/8 ];


protocols igmp
interface ge-0/2/0.0 {
    version 3;
    ssm-map ssm-map;
}
```

e06 PIM Configurations

```
Current configuration : 154 bytes
!
interface GigabitEthernet1/9
 description to_301/8
 ip address 100.3.25.1 255.255.255.0
 ip pim sparse-mode
 ip igmp version 3
 hold-queue 1000 in
end

hkplabedger06#show run int g1/23
Building configuration...

Current configuration : 1133 bytes
!
interface GigabitEthernet1/23
 dampening 15
 mtu 9176
 bandwidth 1000000
 ip address 192.168.0.141 255.255.255.252
 ip router isis lab-core
 ip pim sparse-mode
 mpls ip
 mpls label protocol ldp
 mpls traffic-eng tunnels
 isis circuit-type level-2-only
 isis network point-to-point
 isis metric 200 level-1
 isis metric 100 level-2
 isis authentication mode md5
 isis authentication key-chain ISIS-interface
 isis csnp-interval 10
end
```

ME1 with fast reroute

```
A:me-hkpak26lab-s01# configure service vpls 1
A:me-hkpak26lab-s01>config>service>vpls# info
----------------------------------------------
            description "p2mp_testi_vpls1"
            stp
                shutdown
            exit
```

```
            igmp-snooping
                no shutdown
            exit
            sap 2/2/4:667 create
            exit
            spoke-sdp 1:99 create
                description "to_s02"
            exit
            no shutdown
-----------------------------------------------



#--------------------------------------------------
echo "MPLS LSP Configuration"
#--------------------------------------------------
        mpls
            path "s01_s02"
                hop 1 192.168.2.2 strict
                no shutdown
            exit
            path "s01_s03"
                hop 1 192.168.2.62 strict
                no shutdown
            exit
            lsp "LSP_s01_s02"
                to 192.168.41.2
                from 192.168.41.1
                fast-reroute one-to-one
                exit
                primary "s01_s02"
                exit
                no shutdown
            exit
            lsp "LSP_s01_s03"
                to 192.168.41.3
                primary "s01_s03"
                exit
                no shutdown
            exit
            no shutdown
        exit


#--------------------------------------------------
echo "MPLS LSP Configuration"
#--------------------------------------------------
        mpls
            path "s01_s02"
                hop 1 192.168.2.2 strict
                no shutdown
            exit
            path "s01_s03"
                hop 1 192.168.2.62 strict
                no shutdown
            exit
            lsp "LSP_s01_s02"
                to 192.168.41.2
                from 192.168.41.1
```

```
                fast-reroute one-to-one
                exit
                primary "s01_s02"
                exit
                no shutdown
            exit
            lsp "LSP_s01_s03"
                to 192.168.41.3
                primary "s01_s03"
                exit
                no shutdown
            exit
            no shutdown
        exit


customer 99 create
            description "p2mp_testi_mikko"
            contact "elimaenkatu 2e"
            phone "555 12345"
        exit


vpls 1 customer 99 create
            description "p2mp_testi_vpls1"
            stp
                shutdown
            exit
            igmp-snooping
                no shutdown
            exit
            sap 2/2/4:667 create
            exit
            spoke-sdp 1:99 create
                description "to_s02"
            exit
            no shutdown
        exit




A:me-hkpak26lab-s01# configure router mpls lsp "LSP_s01_s02"
A:me-hkpak26lab-s01>config>router>mpls>lsp# info detail
----------------------------------------------
                to 192.168.41.2
                from 192.168.41.1
                rsvp-resv-style se
                adaptive
                no auto-bandwidth
                no cspf
                no include
                no exclude
                no adspec
                fast-reroute one-to-one
                    no hop-limit
                    node-protect
                exit
                hop-limit 255
```

```
                retry-limit 0
                retry-timer 30
                no least-fill
                metric 0
                ldp-over-rsvp include
                vprn-auto-bind include
                igp-shortcut
                class-type 0
                main-ct-retry-limit 0
                primary "s01_s02"
                    no hop-limit
                    no adaptive
                    no include
                    no exclude
                    record
                    record-label
                    no bandwidth
                    priority 7 0
                    no class-type
                    no backup-class-type
                    no shutdown
                exit
                no shutdown
-----------------------------------------------
A:me-hkpak26lab-s01>config>router>mpls>lsp#
```

ME2 Configurations with fast reroute

```
#--------------------------------------------------
echo "MPLS LSP Configuration"
#--------------------------------------------------
        mpls
            path "s02_s03"
                shutdown
                hop 1 192.168.2.13 strict
            exit
            lsp "LSP_s02_s01"
                to 192.168.41.1
                from 192.168.41.2
                cspf
                fast-reroute one-to-one
                exit
                primary "s02_s01"
                exit
                no shutdown
            exit
            no shutdown
        exit


 customer 99 create
            description "p2mp_testi_mikko"
            contact "elimaenkatu 2e"
            phone "555 12345"


       sdp 1 mpls create
            description "p2mp_testi_sdp1"
```

```
            far-end 192.168.41.1
            lsp "LSP_s02_s01"
            keep-alive
                shutdown
            exit
            no shutdown
        exit
        sdp 2 mpls create
            far-end 192.168.41.3
            lsp "s02_s03"
            keep-alive
                shutdown
            exit
            no shutdown
        exit



       vpls 1 customer 99 create
            description "p2mp_testi_mikko"
            stp
                shutdown
            exit
            igmp-snooping
                no shutdown
            exit
            sap 1/2/3:667 create
                description "edge02_2/10"
            exit
            spoke-sdp 1:99 create
            exit
            no shutdown
        exit



A:me-hkpak26lab-s02# configure router mpls lsp "LSP_s02_s01"
A:me-hkpak26lab-s02>config>router>mpls>lsp# info detail
----------------------------------------------
                to 192.168.41.1
                from 192.168.41.2
                rsvp-resv-style se
                adaptive
                no auto-bandwidth
                cspf
                no include
                no exclude
                no adspec
                fast-reroute one-to-one
                    no hop-limit
                    node-protect
                exit
                hop-limit 255
                retry-limit 0
                retry-timer 30
                no least-fill
                metric 0
                ldp-over-rsvp include
                vprn-auto-bind include
```

```
                igp-shortcut
                class-type 0
                main-ct-retry-limit 0
                primary "s02_s01"
                    no hop-limit
                    no adaptive
                    no include
                    no exclude
                    record
                    record-label
                    no bandwidth
                    priority 7 0
                    no class-type
                    no backup-class-type
                    no shutdown
                exit
                no shutdown
-------------------------------------------
A:me-hkpak26lab-s02>config>router>mpls>lsp#
```

ME3 Configurations

```
#--------------------------------------------------
echo "MPLS LSP Configuration"
#--------------------------------------------------
        mpls
            path "s03_s01_mikko"
                shutdown
                hop 1 192.168.2.61 strict
            exit
            path "s03_s02"
                shutdown
                hop 1 192.168.2.14 strict
            exit
            exit
            lsp "s03_s01_mikko"
                to 192.168.41.1
                from 192.168.41.3
                primary "s03_s01_mikko"
                exit
                no shutdown
            exit
            no shutdown
        exit


        customer 99 create
            description "p2mp testi"
            contact "elimaenkatu2e"
            phone "555 12345"
        exit


        sdp 3 mpls create
            far-end 192.168.41.1
            lsp "s03_s01_mikko"
            keep-alive
                shutdown
```

```
    exit
    no shutdown
exit


sdp 26 mpls create
    description "SDP_s03_s02"
    far-end 192.168.41.2
    ldp
    path-mtu 4462
    keep-alive
        shutdown
    exit
    no shutdown
exit
```