MPLS Working Group N. Leymann, Ed.
Internet-Draft Deutsche Telekom AG
Intended status: Informational B. Decraene
Expires: December 30, 2014 Orange
C. Filsfils
M. Konstantynowicz, Ed.
Cisco Systems
D. Steinberg
Steinberg Consulting
June 28, 2014

Seamless MPLS Architecture
draft-ietf-mpls-seamless-mpls-07

Abstract

This documents describes an architecture which can be used to extend MPLS networks to integrate access and aggregation networks into a single MPLS domain ("Seamless MPLS"). The Seamless MPLS approach is based on existing and well known protocols. It provides a highly flexible and a scalable architecture and the possibility to integrate 100.000 of nodes. The separation of the service and transport plane is one of the key elements; Seamless MPLS provides end to end service independent transport. Therefore it removes the need for service specific configurations in network transport nodes (without end to end transport MPLS, some additional services nodes/configurations would be required to glue each transport domain). This draft defines a routing architecture using existing standardized protocols. It does not invent any new protocols or defines extensions to existing protocols.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

MPLS as a mature and well known technology is widely deployed in today's core and aggregation/metro area networks. Many metro area networks are already based on MPLS delivering Ethernet services to residential and business customers. Until now those deployments are usually done in different domains; e.g. core and metro area networks are handled as separate MPLS domains.

Seamless MPLS extends the core domain and integrates aggregation and access domains into a single MPLS domain ("Seamless MPLS"). This enables a very flexible deployment of an end to end service delivery. In order to obtain a highly scalable architecture Seamless MPLS takes

into account that typical access devices (DSLAMs, MSAN) are lacking some advanced MPLS features, and may have more scalability limitations. Hence access devices are kept as simple as possible.

Seamless MPLS is not a new protocol suite but describes an architecture by deploying existing protocols like BGP, LDP and ISIS. Multiple options are possible and this document aims at defining a single architecture for the main function in order to ease implementation prioritization and deployments in multi vendor networks. Yet the architecture should be flexible enough to allow some level of personalization, depending on use cases, existing deployed base and requirements. Currently, this document focus on end to end unicast LSP.

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

1.2. Terminology

This document uses the following terminology

- Access Node (AN): An access node is a node which processes customers frames or packets at Layer 2 or above. This includes but is not limited to DSLAMs or OLTs (in case of (G)PON deployments). Access nodes have only limited MPLS functionalities in order to reduce complexity in the access network.

- Aggregation Node (AGN): An aggregation node (AGN) is a node which aggregates several access nodes (ANs).

- Area Border Router (ABR): Router between aggregation and core domain.

- Deployment Scenario: Describes which an implementation of Seamless MPLS in order to fullfil the requirements derived from one or more use cases.

- Seamless MPLS Domain: A set of MPLS equipments which can set MPLS LSPs between them.

- Transport Node (TN): Transport nodes are used to connect access nodes to service nodes, and services nodes to services nodes. Transport nodes ideally have no customer or service state and are therefore decoupled from service creation.

- Seamless MPLS (S-MPLS): Used as a generic term to describe an architecture which integrates access, aggregation and core network in a single MPLS domain.

- Service Node (SN): A service node is used to create services for customers and is connected to one or more transport nodes. Typical examples include Broadband Network Gateways (BNGs), video servers

- Transport Pseudo Wire (T-PW): A transport pseudowire provides service independent transport mechanisms based on Pseudo-Wires within the Seamless MPLS architecture.

- Use Case: Describes a typical network including service creation points in order to describe the requirments, typical numbers etc. which need to be taken into account when applying the Seamless MPLS architecture.

2. Motivation

MPLS is deployed in core and aggregation network for several years and provides a mature and stable basis for large networks. In addition MPLS is already used in access networks, e.g. such as mobile or DSL backhaul. Today MPLS as technology is being used on two different layers:

- the Transport Layer and

- the Service Layer (e.g. for MPLS VPNs)

In both cases the protocols and the encapsulation are identical but the use of MPLS is different especially concerning the signalling, the control plane, the provisioning, the scalability and the frequency of updates. On the service layer only service specific information is exchanged; every service can potentially deploy it's own architecture and individual protocols. The services are running on top of the transport layer. Nevertheless those deployments are usually isolated, focussed on a single use case and not integrated into an end-to-end manner.

The motivation of Seamless MPLS is to provide an architecture which supports a wide variety of different services on a single MPLS platform fully integrating access, aggregation and core network. The architecture can be used for residential services, mobile backhaul, business services and supports fast reroute, redundancy and load balancing. Seamless MPLS provides the deployment of service creation points which can be virtually everywhere in the network. This enables network and service providers with a flexible service and

service creation. Service creation can be done based on the existing requirements without the needs for dedicated service creation areas on fixed locations. With the flexibility of Seamless MPLS the service creation can be done anywhere in the network and easily moved between different locations.

2.1. Why Seamless MPLS

Multiple Service Providers plan to deploy networks with 10k to 100k MPLS nodes, with varying levels of MPLS LSP connectivity between those nodes - sparse-mesh in access, partial-mesh in aggregation and full-mesh in core. This is typically at least one order of magnitude higher than current deployments and may require a new architecture. Multiple options are possible and it makes sense for the industry (both vendors and SP) to restrict the options in order to ease the first deployments (e.g. restrict the number of options to implement and/or scales for vendors, reduce interoperability and debugging issues for SP).

Many aggregation networks are already deploying MPLS but are limited to the use of MPLS per aggregation area. Those MPLS based aggregation domains are connected to a core network running MPLS as well. Nevertheless most of the services are not limited to an aggregation domain but running between several aggregation domains crossing the core network. In the past it was necessary to provide connectivity between the different domains and the core on a per service level and not based on MPLS (e.g. by deploying native IP-Routing or Ethernet based technologies between aggregation and core). In most cases service specific configurations on the border nodes between core and aggregation were required. New services led to additional configurations and changes in the provisioning tools (see Figure 1).

With Seamless MPLS there are no technology boundaries and no topology boundaries for the services. Network (or region) boundaries are for scaling and manageability, and do not affect the service layer, since the Transport Pseudowire that carries packets from the AN to the SN doesn't care whether it takes two hops or twenty, nor how many region boundaries it needs to cross. The network architecture is about network scaling, network resilience and network manageability; the service architecture is about optimal delivery: service scaling, service resilience (via replicated SNs) and service manageability. The two are decoupled: each can be managed separately and changed independently.

```
   +--------------+         +--------------+         +--------------+
   |  Aggregation |         |     Core     |         |  Aggregation |
   |   Domain #1  +---------+    Domain    +---------+   Domain #2  |
   |     MPLS     | ^       |     MPLS     |       ^ |     MPLS     |
   +--------------+ |       +--------------+       | +--------------+
                    |                              |
                    +------ service specific ------+
                             configuration
```

Figure 1: Service Specific Configurations

One of the main motivations of Seamless MPLS is to get rid of service specific configurations between the different MPLS islands. Seamless MPLS connects all MPLS domains on the MPLS transport layer providing a single transport layer for all services - independent of the service itself. The Seamless MPLS architecture therefore decuples the service and transport layer and integrates access, aggregation and core into a single platform. One of the big advantages is that problems on the transport layer only need to be solved once (and the solutions are available to all services). With Seamless MPLS it is not necessary to use service specific configurations on intermediate nodes; all services can be deployed in an end to end manner.

2.2. Use Case #1

2.2.1. Description

In most cases at least residential and business services need to be supported by a network. This section describes a Seamless MPLS use case which supports such a scenario. The use case includes point to point services for business customers as well as typical service creation for residential customers.

```
                       +-------------+
                       |   Service   |
                       |  Creation   |
                       | Residential |
                       |  Customers  |
                       +------+------+
                              |
                              |
                              |
       PW1    +-------+   +---+---+
      #########################   |
      #    +--+ AGN11 +---+ AGN21 +  +------+
      #   /   |       |  /|       |\ |      |           +--------+
   +--#-+/    +-------+\/ +-------+ \|      |           | remote |
   | AN |              /\            + CORE +---......--+   AN   |
   +--#-+\    +-------+  \+-------+ /|      |         #######    |
      #   \   |       |   |       |/################### +--------+
      #    +--+ AGN12 +---+ AGN22 +##+------+  P2P Business Service
      ##############################
       PW2    +-------+   +-------+
```

Figure 2: Use Case #1: Service Creation

Figure 2 shows the different service creation points and the corresponding pseudowires between the access nodes and the service creation points. The use case does not show all PWs (e.g. not the PWs needed to support redundancy) in order to keep the figure simple. Node and link failures are handled by rerouting the PWs (based on standard mechanisms). End customers (either residential or business customers) are connected to the access nodes using a native technology like Ethernet. The access nodes terminates the PW(s) carrying the traffic for the end customers. The link between the access node (AN) and the aggregation node (AGN) is the first MPLS enabled link.

Residential Services: The service creation for all residential customers connected to the Access Nodes in an aggregation domain is located on an Service Node connected to the AGN2x. The PW (PW1) originated at the AN and terminates at the AGN2. A second PW is deployed in the case where redundancy is needed on the AN (the figure shows redundancy but this might not be the case for all ANs in this Use Case). Additonal PWs can be deployed as well in case more than a single service creation is needed for the residential service (e.g. one service creation point for Internet access and a second service creation point for IPTV services).

Business Sercvices: For business services the use cases shows point to point connections between two access nodes. PW2 originates at

the AN and terminates on the remote AN crossing two aggregation areas and the core network. If the access node needs connections to several remote ANs the corresponding number of PWs will be originated at the AN. Nevertheless taking the number of ports available and the number of business customers on a typical access node the number of PWs will be relatively small.

```
              +-------+   +-------+   +------+   +------+
              |       |   |       |   |      |   |      |
           +--+ AGN11 +---+ AGN21 +---+ ABR1 +---+ LSR1 +--> to AGN
          /   |       |  /|       |   |      |   |      |
   +----+/    +-------+\/ +-------+   +------+  /+------+
   | AN |              /\                    \/
   +----+\    +-------+  \+-------+   +------+/\ +------+
          \   |       |   |       |   |      |  \|      |
           +--+ AGN12 +---+ AGN22 +---+ ABR2 +---+ LSR2 +--> to AGN
              |       |   |       |   |      |   |      |
              +-------+   +-------+   +------+   +------+

   static route     ISIS L1 LDP             ISIS L2 LDP

   <-Access-><--Aggregation Domain--><---------Core--------->
```

Figure 3: Use Case #1: Redundancy

Figure 3 shows the redundancy at the access and aggregation network deploying a two stage aggregation network (AGN1x/AGN2x). Nevertheless redundancy is not a must in this use case. It is also possible to use non redundant connection between the ANs and AGN1 stage and/or between the AGN1 and AGN2 stages. The AGN2x stage is used to aggregate traffic from several AGN1x pairs. In this use case an aggregation domain is not limited to the use of a single pair of AGN2x; the deployment of several AGN2 pairs within the domain is also supported. As design goal for the scalability of the routing and forwarding within the Seamless MPLS architecture the following numbers are used:

- Number of Aggregation Domains: 100

- Number of Backbone Nodes: 1,000

- Number of Aggregation Nodes: 10,000

- Number of Access Nodes: 100,000

The access nodes (AN) are dual homed to two different aggregation nodes (AGN11 and AGN12) using static routing entries on the AN. The ANs are always source or sink nodes for MPLS traffic but not transit

nodes. This allows a light MPLS implementation in order to reduce the complexity in the AN. The aggregation network consists of two stages with redundant connections between the stages (AGN11 is connected to AGN21 and AGN22 as well as AGN12 to AGN21 and AGN22). The gateway between the aggregation and core network is realized using the Area Border Routers (ABR). From the perspective of the MPLS transport layer all systems are clearly identified using the loopback address of the system. An ingress node must be able to establish a service to an arbitrary egress system by using the corresponding MPLS transport label

2.2.2. Typical Numbers

Table 1 shows typical numbers which are expected for Use Case #1 (access node).

Parameter	Typical Value
IGP RIB Entries	2
IP FIB Entries	2
LDP LIB Entries	200
MPLS NHLFE Entries	200
MPLS ILM Entries	0
BGP RIB Entries	0
BGP FIB Entries	0

Table 1: Use Case #1: Typical Numbers for Access Node

2.3. Use Case #2

2.3.1. Description

In most cases, residential, wholesales and business services need to be supported by the network.

```
                         +-------------+
                         |   Service   |
                         |  platforms  |
                         |(VoIP, VoD..)|
                         | Residential |
                         |  Customers  |
                         +------+------+
                                |
                                |
           +---+    +-----+  +--+--+   +-----+
           |AN1|----+AGN11+--+AGN21+---+ ABR |
           +---+    +--+--+  +--+--+   +--+--+
                       |        |         |
           +---+    +--+--+     |         |          +----+
           |AN2|----+AGN12+     |         |        --+ PE |
           +---+    +--+--+     |         |          +----+
                       |        |         |
                       .        |         |
                       .        |         |
                       .        |         |
                       |        |         |
   +---+   +---+    +--+--+  +--+--+   +--+--+
   |AN4+---+AN3|----+AGN1x+--+AGN22+---+ ABR |
   +---+   +---+    +-----+  +-----+   +-----+

        <-Access-><--Aggregation Domain--><---------Core--------->
```

Figure 4: Use Case #2

The above topology (see Figure 4) is subject to evolutions, depending on AN types and capacities (in terms of number of customers and/or aggregated bandwidth). For examples, AGN1x connection toward AGN2y currently forms a ring but may latter evolve in a square or triangle topology; AGN2y nodes may not be present...

Most access nodes (AN) are single attached on one aggregation node using static routing entries on the AN and AGN. Some AN, are dual attached on two different AGN using static routes. Some AN are used as transit by some lower level AN. Static routes are expected to be used between those AN.

IPv4, IPv6 and MPLS interconnection between the aggregation and core network is realized using the Area Border Routers (ABR). Any ingress node must be able to establish IPv4, IPv6 and MPLS connections to any egress node in the seamless MPLS domain.

Regarding MPLS connectivity requirements, a full mesh of MPLS LSPs is required between the ANs of an aggregation area, at least for 6PE

purposes. Some additional LSPs are needed between ANs and some PE in the aggregation area or in the core area for access to services, wholesale and enterprises services. In short, a meshing of LSP is required between the AGN of the whole seamless MPLS domain. Finally, LSP between any node to any node should be possible.

From a scalability standpoint, the following numbers are the targets:

- Number of Aggregation Domains: 30

- Number of Backbone Nodes: 150

- Number of Aggregation Nodes: 1.500

- Number of Access Nodes: 40.000

2.3.2. Typical Numbers

Table 2 shows typical numbers which are expected for Use Case #2 for the purpose of establishing the transport LSPs. They do not take into account the services built in addition. (e.g. 6PE will require additional IPv6 routes).

Parameter	Typical Value
IGP RIB Entries	2
IP FIB Entries	2
LDP LIB Entries	1,400
MPLS NHLFE Entries	1,400
MPLS ILM Entries	1,400

Table 2: Use Case #2: Typical Numbers for Access Node

3. Requirements

The following section describes the overall requirements which need to be fulfilled by the Seamless MPLS architecture. Beside the general requirements of the architecture itself there are also certain requirements which are related to the different network nodes.

- End to End Transport LSP: MPLS based services (pseudowire based, L3-VPN or IP) SHALL be provided by the Seamless MPLS based infrastructure between any nodes.

- Scalability: The network SHALL be scalable to the minimum of 100.000 nodes.

- Fast convergence (sub second resilience) SHALL be supported. Fast reroute (LFA) SHOULD be supported.

- Flexibility: The Seamless MPLS architecture SHALL be applied to a wide variety of existing MPLS deployments. It SHALL use a flexible approach deploying building blocks with the possiblity to use certain features only if those features are needed (e.g. dual homing ANs or fast reroute mechanisms).

- Service independence: Service and transport layer SHALL be decoupled. The architecture SHALL remove the need for service specific configurations on intermediate nodes.

- Native Multicast support: P2MP MPLS LSPs SHOULD be supported by the Seamless MPLS architecture.

- Interoperable end to end OAM mechanisms SHALL be implemented

3.1. Overall

3.1.1. Access

In respect of MPLS functionality the access network should be kept as simple as possible. Compared to the aggregation and/or core network within Seamless MPLS a typical access node is less powerful. The control plane and the forwarding should be as simple as possible. To reduce the complexity and the costs of an access node not the full MPLS functionality need to be supported (control and data plane). The use of an IGP should be avoided. Static routing should be sufficient. Required functionality to reach the required scalability should be moved out of the access node. The number of access nodes can be very high. The support of load balancing for layer 2 services should be implemented.

3.1.2. Aggregation

The aggregation network aggregates traffic from access nodes. The aggregation Node must have functionalities that enlarge the scalability of the simple access nodes that are connected. The IGP must be link state based. Each aggregation area must be a separated area. All routes that are interarea should use an EGP to keep the IGP small. The aggregation node must have the full scalability concerning control plane and forwarding. The support of load balancing for layer 2 services must be implemented.

3.1.3. Core

The core connects the aggregation areas. The core network elements must have the full scalability concerning control plane and forwarding. The IGP must be link state based. The core area must not include routes from aggregation areas. All routes that are interarea should use an EGP to keep the IGP small. Each area of the link state based IGP should have less than 2000 routes. The support of load balancing for layer 2 services must be implemented.

3.2. Multicast

Compared with unicast connectivity Multicast is more dynamic. User generated messages - like joining or leaving multicast groups - are interacting directly with network components in the access and aggregation network (in order to build the corresponding forwarding states). This leads to the need for a highly dynamic handling of messages on access and aggregation nodes. Nevertheless the core network SHOULD be stable and state changes triggered by user generated messages SHOULD be minimized. This rises the need for an hierarchy for the P2MP support in Seamless MPLS hiding the dynamic behaviour of the access and aggregation nodes

o mLDP

o P2MP RSVP-TE

3.3. Availability

All network elements should be high available (99.999% availability). Outage times should be as low as possible. A repair time of 50 milliseconds or less should be guarantied at all nodes and lines in the network that are redundant. Fast convergence features SHOULD be used in all control plane protocols. Local Repair functions SHOULD be used wherever possible. Full redundancy is required at all equipment that is shared in a network element.

o Power Supply

o Switch Fabric

o Routing Processor

A change from an active component to a standby component SHOULD happen without effecting customers traffic. The Influence of customer traffic MUST be as low as possible.

3.4. Scalability

The network must be highly scalable. Based on the use cases described in Sections 2.2 and 2.3, as a minimum requirement the following scalability figures should be met:

- Number of aggregation domains: 100

- Number of backbone nodes: 1,000

- Number of aggregation nodes: 10,000

- Number of access nodes: 100,000

3.5. Stability

- The platform should be stable under certain circumstances (e.g. missconfiguration within one area should not cause instability in other areas).

- Differentiate between "All Loopbacks and Link addresses should be ping able from every where." Vs. "Link addresses are not necessary ping able from everywhere".

4. Architecture

4.1. Overall

One of the key questions that emerge when designing an architecture for a seamless MPLS network is how to handle the sheer size of the necessary routing and MPLS label information control plane and forwarding plane state resulting from the stated scalability goals especially with respect to the total number of access nodes. This needs to be done without overwhelming the technical scaling limits of any of the involved nodes in the network (access, aggregation and core) and without introducing too much complexity in the design of the network while at the same time still maintaining good convergence properties to allow for quick MPLS transport and service restoration in case of network failures.

4.2. Multi-Domain MPLS networks

The key design paradigm that leads to a sound and scalable solution is the divide and conquer approach, whereby the large problem is decomposed into many smaller problems for which the solution can be found using well-known standard architectures.

In the specific case of seamless MPLS the overall MPLS network SHOULD be decomposed into multiple MPLS domains, each well within the scaling limits of well-known architectures and network node implementations. From an organizational and operational point of view it MAY make sense to define the boundaries of such domains along the pre-existing boundaries of aggregation networks and the core network.

Examples of how networks can be decomposed include using IGP areas as well as using multiple BGP autonomous systems.

4.3. Hierarchy

These MPLS domains SHOULD then be then be connected into an MPLS multi-domain network in a hierarchical fashion that enables the seamless exchange of loopback addresses and MPLS label bindings for transport LSPs across the entire MPLS internetwork while at the same time preventing the flooding of unnecessary routing and label binding information into domains or parts of the network that do not need them. Such a hierarchical routing and forwarding concept allows a scalability in different dimensions and allows to hide the complexity and size of the aggregation and access networks.

4.4. Intra-Domain Routing

The intra-domain routing within each of the MPLS domains (i.e. aggregation domains and core) SHOULD utilize standard IGP protocols like OSPF or ISIS. By definition, each of these domains is small enough so that there are no relevant scaling limits within each IGP domain, given well-known state-of-the-art IGP design principles and recent router technology.

The intra-domain MPLS LSP setup and label distribution SHOULD utilize standard protocols like LDP or RSVP.

Note that this document describes the design based on LDP, LDP Downstream-on-Demand and labeled BGP due to the higher degree of out-of-the-box automation and operational simplicity as well as compatibility with the existing backbone and backhaul designs & deployments which use LDP and not RSVP-TE. It also assumes relatively simple MPLS implementations on access nodes. The protocol choices for the design described in this document have been driven by the actual SP deployments. Design based on the hierarchy of RSVP-TE LSPs may be an alternative, but has not been considered in this document.

4.5. Inter-Domain Routing

The inter-domain routing is responsible for establishing connectivity between and across all MPLS domains. The inter-domain routing SHOULD establish a routing and forwarding hierarchy in order to achieve the scaling goals of seamless MPLS. Note that the IP aggregation usually performed between region (IGP areas/AS) in IP routing does not work for MPLS as MPLS is not capable of aggregating FEC (because MPLS forwarding use an exact match lookup, while IP uses longest match).

Therefore it is RECOMMENDED to utilize protocols that support indirect next-hops (e.g. using BGP to carry MPLS label information [RFC3107]). The mechanism for the LSP forwarding hierarchy is described in Section 5.3.

4.6. Access

Compared to the aggregation and core parts of the Seamless MPLS network the access part is special in two respects:

- The number of nodes in the access is at least one order of magnitude higher than in any other part of the network.

- Because of the large quantity of access nodes, the cost of these nodes is extremely relevant for the overall costs of the entire network, i.e. acess nodes are very cost sensitive.

This makes it desirable to design the architecture such that the AN functionality can be kept as simple as possible. This should always be kept in mind when evaluating different seamless MPLS architectures. The goal is to limit both the number of different protocols needed on the AN as well as the scale to which each protocol must perform to the absolute minimum.

4.7. Signalling and Label Distribution

Following figures show IP/MPLS signaling and label distribution for an LSP from AN2 to AN1 (192.0.2.1), detailing the signalling from AN1 to AN2 (left to right) and packet forwarding from AN2 to AN1 (right to left). It is assumed that Penultimate Hop Popping is used.

Terminology used in the figures:

- LDP DoD: LDP Downstream on Demand

- LDP DU: LDP Downstream Unsolicited

- BGP LU: BGP Label Unicast (RFC 3107)

- BGP NH: BGP Next Hop Label LYZi: L=Label, Y= node advertising the FEC (G for AGN, B for ABR, A for AN), Z= protocol advertising it (B for BGP, L for LDP)

```
   ---------------------------------------------------------------------

   AN1----AGN1----AGN2----ABR1----LSR1----ABR2----AGN3----AGN4----AN2

   LDP DoD ->    BGP LU     ->           BGP LU            ->  LDP DoD
                     Next Hop Self

   192.0.2.1   192.0.2.1                192.0.2.1             192.0.2.1
   Labels: LAL1       LAB1                     LAB2             LAL2
               BGP NH: AGN1             BGP NH: ABR1
```

Figure 5: MPLS signalling for the AN / edge layer

```
   ---------------------------------------------------------------------

   AN1----AGN1----AGN2----ABR1----LSR1----ABR2----AGN3----AGN4----AN2

           | IS-IS level 2 | IS-IS level 1 -> IS-IS level 2 |
           | LDPDU - LDPDU - LDPDU - LDPDU - LDPDU - LDPDU  |

   FEC:           AGN1           ABR1            ABR1
   Label:     LGL1    LGL2    LBL1     LBL2   LBL3     LBL4
```

Figure 6: MPLS signalling and IP routing for the core layer

```
   ---------------------------------------------------------------------

    Static  | IS-IS level 2 | IS-IS level 1 | IS-IS level 2 |   Static
    Routing                                                     Routing

   192.0.2.1 -> 192.0.2/24  ->  192.0.2/20   -> 192.2/20         0/0
                AGN1           ABR1            ABR1
```

Figure 7: IP routing for the AN / edge layer

```
   ---------------------------------------------------------------------

   AN1----AGN1----AGN2----ABR1----LSR1----ABR2----AGN3----AGN4----AN2

   FEC AN1:   LAL1     LAB1   LAB2     LAB2    LAB2    LAB2     LAL2
   FEC AGN1:           LGL2   LBL1     LBL2    LBL3    LBL4
      /ABR1
```

Figure 8: Forwarding Plane

----

5. Deployment Scenarios

This section describes the deployment scenarios based on the use cases and the generic architecture above.

5.1. Deployment Scenario #1

Section describing the Seamless MPLS implementation of a large european ISP.

5.1.1. Overview

This deployment scenario describes one way to implement a seamless MPLS architecture. Specific to this implementation is the choice of intra- and inter-domain routing and label distribution protocols, as well as the details of the interworking of these protocols to achieve the overall scalable hierarchical architecture.

5.1.2. General Network Topology

There are multiple aggregation domains (in the order of up to 100) connected to the core in a star topology, i.e. aggregation domains are never connected among themselves, but only to the core. The core has its own domain.

```
           +-------+   +-------+   +------+   +------+
           |       |   |       |   |      |   |      |
        +--+ AGN11 +---+ AGN21 +---+ ABR1 +---+ LSR1 +--> to AGN
       /   |       |  /|       |   |      |   |      |
+----+/    +-------+\/ +-------+   +------+  /+------+
| AN |              /\                     \/     |
+----+\    +-------+  \+-------+   +------+/\ +------+
       \   |       |   |       |   |      |  \|      |
        +--+ AGN12 +---+ AGN22 +---+ ABR2 +---+ LSR2 +--> to AGN
           |       |   |       |   |      |   |      |
           +-------+   +-------+   +------+   +------+

static route     ISIS L1 LDP DU      ISIS L2 LDP DU

<-Access-><--Aggregation Domain--><---------Core--------->
```

Figure 9: Deployment Scenario #1

As shown in Figure 9, the access nodes (AN) are connected to the aggregation network via aggregation nodes called AGN1x, either to a

single AGN1x or redundantly to two AGN1x. Each AGN1x has redundant uplinks to a pair of second-level aggregation nodes called AGN2x.

Each aggregation domain is connected to the core via exactly two border routers (ABR) on the core side. There can be multiple AGN2 pairs per aggregation domain, but only one ABR pair for each aggregation domain. Each of the AGN2 in an AGN2 pair connects to one of the ABRs in the ABR pair responsible for that aggregation domain.

The ABRs on the core side have redundant connections to a pair of LSR routers.

The LSR pair is also connected via a direct link.

The core LSR are connected to other core LSR in a partly meshed topology so that there are disjunct, redundant paths from each LSR to each other LSR.

5.1.3. Hierarchy based on recursive BGP labeled route lookup

Inline with the explanation in section 4.5, LSP hierarchy is key to a scalable seamless MPLS architecture.

The LSP hierarchy in this design is achieved by:

- Forming separate MPLS domains for aggregation and core areas.

- Intra-domain LSP connectivity provided by combination of IS-IS (as the intra-domain link-state routing protocol) and LDP DU (used for MPLS label distribution for intra-domain LSPs).

- Inter-domain LSP connectivity provided by labeled BGP [RFC3107] (used for MPLS label distribution for inter-domain LSP FECs) and relying on IS-IS and LDP DU for intra-domain LSP connectivity between the LSR labeled BGP speakers (AGNs and ABRs). Note that the MPLS core notes are not carrying the labeled BGP routes.

The aggregation and core MPLS domains are mapped to IS-IS areas as follows: Aggregation domains are mapped to IS-IS L1 areas. The core is configured as IS-IS L2. The border routers connecting aggregation and core are IS-IS L1L2 and are referred to as ABRs. From a technical and operational point of view these ABRs are part of the core, although they also belong to the respective aggregation domain purely from a routing protocol point of view.

5.1.4. Intra-Area Routing

5.1.4.1. Core

The core uses ISIS L2 to distribute routing information for the loopback addresses of all core nodes. The border routers (ABR) that connect to the aggregation domains are also part of the respective aggregation ISIS L1 area and hence ISIS L1L2.

LDP DU is used to distribute MPLS label binding information for the loopback addresses of all core nodes.

5.1.4.2. Aggregation

The aggregation domains uses ISIS L1 as intra-domain routing protocol. All AGN loopback addresses are carried in ISIS.

As in the core, the aggregation also uses LDP DU to distribute MPLS label bindings for the loopback addresses.

5.1.5. Access

Access nodes do not have their own domain or IGP area. Instead, they directly connect to the AGN1 nodes in the aggregation domain. To keep access devices as simple as possible, ANs do not participate in ISIS.

Instead, each AN has two static default routes pointing to each of the AGN1 it is connected to. Appropriate techniques SHOULD be deployed to make sure that a given default route is invalidated when the link to an AGN1 or that node itself fails. Examples of such techniques include monitoring the pysical link state for loss of light/loss of frame, or using Ethernet link OAM or BFD [RFC5881].

The AGN1 MUST have a configured static route to the loopback address of each of the ANs it is connected to, because it cannot learn the AN loopback address in any other way. These static routes have to be monitored and invalidated if necessary using the same techniques as described above for the static default routes on the AN.

The AGN1 redistributes these routes into ISIS for intra-domain reachability of all AN loopback addresses.

LDP DU is used for MPLS label distribution between AGN1 and AN. In order to keep the AN control plane as lightweight as possible, and to avoid the necessity for the AN to store 100.000 MPLS label bindings for each upstream AGN1 peer, LDP is deployed in downstream-on-demand (DoD) mode, described below.

To allow the label bindings received via LDP DoD to be installed into the LFIB on the AN without having the specific host route to the destination loopback address, but only a default route, use of the LDP Extension for Inter-Area Label Switched Paths [RFC5283] is made.

5.1.5.1. LDP Downstream-on-Demand (DoD)

LDP downstream-on-demand mode is specified in [RFC5036]. In this mode the upstream LSR will explicitly ask the downstream LSR for a label binding for a particular FEC when needed.

The assumption is that a given AN will only have a limited number of services configured to an even more limited number of destinations, or egress LER. Instead of learning and storing all label bindings for all possible loopback addresses within the entire Seamless MPLS network, the AN will use LDP DoD to only request the label bindings for the FECs corresponding to the loopback addresses of those egress nodes to which it has services configured.

More detailed description of LDP DoD use cases for MPLS access and list of required LDP DoD procedures in the context of Seamless MPLS design is included in [RFC7032].

5.1.6. Inter-Area Routing

The inter-domain MPLS connectivity from the aggregation domains to and across the core domain is realized primarily using BGP with MPLS labels ("labled BGP/SAFI4" [RFC3107]). A very limited amount of route leaking from ISIS L2 into L1 is also used.

All ABR and PE nodes in the core are part of the labeled iBGP mesh, which can be either full mesh or based on route reflectors. These nodes advertise their respective loopback addresses (which are also carried in ISIS L2) into labeled BGP.

Each ABR node has labeled iBGP sessions with all AGN1 nodes inside the aggregation domain that they connect to the core. Since there are two ABR nodes per aggregation domain, this leads to each AGN1 node having an iBGP sessions with each of the two ABR. Note that the use of iBGP implies that the entire seamless MPLS internetwork is just a single AS to which all core and aggregation nodes belong. The AGN1 nodes advertise their own loopback addresses into labeled BGP, in addition to these loopbacks also being in ISIS L1.

Additionally the AGN1 nodes also redistribute all the statically configured routes to the AN loopback addresses into labeled BGP. Note that as stated obove, the AGN1 MUST ask the AN for label

bindings for the AN loopback FECs via LDP DoD in order to have a valid labeled route with a non-null label.

This architecture results in carrying all loopbacks of all nodes except pure P nodes (AN, AGN, ABR and core PE) in labeled BGP, e.g. there will be in the order of 100.000 routes in labeled BGP when approaching the stated scalability goal. Note that this only affects the BGP RIB size and does not necessarily imply that any node needs to actually have active forwarding state (LFIB) in the same order of magnitude. In fact, as will be discussed in the scalability analysis, no single node needs to install all labeled BGP routes into the LFIB, but each node only needs a small percentage of the RIB as active forwarding state in the LFIB. And from a RIB point of view, BGP is known to scale to hundreds of thousands of routes.

5.1.7. Labeled iBGP next-hop handling

The ABR nodes run labeled iBGP both to the core mesh as well as to the AGN1 nodes of their respective aggregation domains. Therefore they operate as iBGP route reflectors, reflecting labeled routes from the aggregation into the core and vice versa.

When reflecting routes from the core into the aggregation domain, the ABR SHOULD NOT change the BGP NEXT-HOP addresses (next-hop-unchanged). This is the usual behaviour for iBGP route reflection. In order to make these routes resolvable to the AGN1 nodes inside the aggregation domain, the ABR MUST leak all other ABR and core PE loopback addresses from ISIS L2 into ISIS L1 of the aggregation domain. Note that the number of leaked addresses is limited so that the overall scalability of the seamless MPLS architecture is not impacted. In the worst case all core loopback addresses COULD be leaked into ISIS L1, but even that would not be a scalability problem.

When reflecting routes from the aggregation into the core, the ABR MUST set then BGP NEXT-HOP to its own loopback addresses (next-hop-self). This is not the default behaviour for iBGP route reflection, but requires special configuration on the ABR. Note that this also implies that the ABR MUST allocate a new local MPLS label for each labeled iBGP FEC that it reflects from the aggregation into the core. This special next-hop handling is essential for the scalability of the overall seamless MPLS architecture since it creates the required hierarchy and enables the hiding of all aggregation and access addresses behind the ABRs from an IGP point of view. Leaking of aggregation ISIS L1 loopback addresses into ISIS L2 is not necessary and MUST NOT be allowed.

The resulting hierarchical inter-domain MPLS routing structure is similar to the one described in [RFC4364] section 10c, only that we use one AS with route reflection instead of using multiple ASes.

5.1.8. Network Availability

The seamless mpls architecture guarantees a sub-second loss of connectivity upon any link or node failures. Furthermore, in the vast majority of cases, the loss of connectivity is limited to sub-50msec.

These network availability properties are provided without any degradation on scale and simplicity. This is a key achievement of the design.

In the remainder of this section, we first introduce the different network availability technologies and then review their applicability for each possible failure scenario.

5.1.8.1. IGP Convergence

IGP convergence can be modelled as a linear process with an initial delay and a linear FIB update [ACM01].

The initial delay could conservatively be assumed to be 260msec: 50msec to detect failures with BFD (most failures would be detected faster with loss of light for example or with faster BFD timers), 50msec to throttle the LSP generation, 150msec to throttle the SPF computation (making sure than all the required LSP's are received even in case of SRLG failures) and 10msec for shortest-path-first tree computation.

Assuming 250usec per update (conservative), this allows for (1000-260)/0.250= 2960 prefixes update within a second following the outage. More precisely, this allows for 2960 important IGP prefixes updates. Important prefixes are automatically classified by the router implementation through simple heuristic (/32 is more important than non-/32).

The number of IGP important routes (loopbacks) in deployment case study 1 is much smaller than 2960, and hence sub-second IGP convergence is conservative.

IGP convergence is a simple technology for the operator provided that the router vendor optimizes the default IGP behavior (no need to tune any arcane knob).

5.1.8.2. Per-Prefix LFA FRR

A per-prefix LFA for a destination D is a precomputed backup IGP nexthop for that destination. This backup IGP nexthop can be link protecting or node protecting [RFC5286].

The analysis of the applicability of Per-Prefix LFA in the deployment model 1 of Seamless MPLS architecture is straightforward thanks to [RFC6571].

In deployment model 1, each aggregation network either follows the triangle or full-mesh topology. Further more, the backbone region implements a dual-plane. As a consequence, the failure of any link or node within an aggregation domain is protected by LFA FRR (sub-50msec) for all impacted IGP prefixes, whether intra-area or inter-area. No uloop may form as a result of these failures [RFC6571].

Per-Prefix LFA FRR is generally assessed as a simple technology for the operator [RFC6571]. It certainly is in the context of deployment case study 1 as the designer enforced triangle and full-mesh topologies in the aggregation network as well as a dual-plane core network.

5.1.8.3. Hierarchical Dataplane and BGP Prefix Independent Convergence

In a hierarchical dataplane, the FIB used by the packet processing engine reflects recursions between the routes. For example, a BGP route B recursing on IGP route I whose best path is via interface O is encoded as a hierarchy of FIB entry B pointing to a FIB entry I pointing to a FIB entry 0.

BGP Prefix Independent Convergence [I-D.rtgwg-bgp-pic] extends the hierarchical dataplane with the concept of a BGP Path-List. A BGP path-list may be abstracted as a set of primary multipath nhops and a backup nhop. When the primary set is empty, packets destined to the BGP destinations are rerouted via the backup nhop.

For complete description of BGP-PIC technology and its applicability please refer to [I-D.rtgwg-bgp-pic] and [ABR-FRR].

Hierarchical data plane and BGP-PIC are very simple technologies to operate. Their applicability to any topology, any routing policy and any BGP unicast address family allows router vendors to enable this behavior by default.

5.1.8.4. BGP Egress Node FRR

BGP egress node FRR is a Fast ReRoute solution and hence relies on local protection and the precomputation and preinstallation of the backup path in the FIB. BGP egress node FRR relies on a transit LSR (Point of Local Repair, PLR) adjacent to the failed protected BGP router to detect the failure and re-route the traffic to the backup BGP router. Number of BGP egress node FRR schemes are being investigated: [PE-FRR], [ABR-FRR], [I-D.minto-2547-egress-node-fast-protection], [I-D.bashandy-bgp-edge-node-frr], [I-D.bashandy-idr-bgp-repair-label], [I-D.bashandy-mpls-ldp-bgp-frr], [I-D.bashandy-bgp-frr-mirror-table], [I-D.bashandy-bgp-frr-vector-label], [I-D.bashandy-isis-bgp-edge-node-frr].

Differences between these schemes relate to the way backup and protected BGP routers get associated, how the protected router's BGP state is signalled to the backup BGP router(s) and if any other state is required on protected, backup and PLR routers. The schemes also differ in compatibility with IP-FRR and TE-FRR schemes to enable PLR to switch traffic towards the backup BGP router in case of protected BGP router failure.

In the Seamless MPLS design, BGP egress node FRR schemes can protect against the failures of PE, AGN and ABR nodes with no requirements on ingress routers.

5.1.8.5. Assessing loss of connectivity upon any failure

We select two typical traffic flows and analyze the loss of connectivity (LoC) upon each possible failure in the Seamless MPLS design in the deployment scenario #1.

- Flow F1 starts from an AN1 in a left aggregation region and ends on an AN2 in a right aggregation region. Each AN is dual-homed to two AGN's.

- Flow F2 starts from a CE1 homed on L3VPN PE1 connected to the core LSRs and ends at CE2 dual-homed to L3VPN PE2 and PE3, both connected to the core LSRs.

Note that due to the symmetric network topology in case study 1, uni-directional flows F1' and F2', associated with F1 and F2 and forwarded in the reversed direction (AN2 to AN1 right-to-left and PE2 to PE1, respectively), take advantage of the same failure restoration mechanisms as F1 and F2.

5.1.8.5.1. AN1-AGN link failure or AGN node failure

F1 is impacted but LoC <50msec is possible assuming fast BFD detection and fast-switchover implementation on the AN. F2 is not impacted.

5.1.8.5.2. Link or node failure within the left aggregation region

F1 is impacted but LoC <50msec thanks to LFA FRR. No uloop will occur during the IGP convergence following the LFA protection. Note: if LFA is not available (other topology then case study one) or if LFA is not enabled, then the LoC would be < second as the number of impacted important IGP route in a seamless architecture is much smaller than 2960.

F2 is not impacted.

5.1.8.5.3. ABR node failure between left region and the core

F1 is impacted but LoC <50msec thanks to LFA FRR. No uloop will occur during the IGP convergence following the LFA protection.

Note: This case is also called "Local ABR failure" as the ABR which fails is the one connected to the aggregation region at the source of flow F1.

Note: remember that the left region receives the routes to all the remote ABR's and that the labelled BGP routes are reflected from the core to the left region with next-hop unchanged. This ensures that the loss of the (local) ABR between the left region and the core is seen as an IGP route impact and hence can be addressed by LFA.

Note: if LFA is not available (other topology then case study one) or if LFA is not enabled, then the LoC would be < second as the number of impacted important IGP routes in a seamless architecture is much smaller than 2960 routes.

F2 is not impacted.

5.1.8.5.4. Link or node failure within the core region

F1 and F2 are impacted but LoC <50msec thanks to LFA FRR.

This is specific to the particular core topology used in deployment case study 1. The core topology has been optimized [RFC6571] for LFA applicability.

As explained in [RFC6571], another alternative to provide <50msec in this case consists in using an MPLS-TE full-mesh and MPLS-TE FRR. This is required when the designer is not able or does not want to optimize the topology for LFA applicability and he wants to achieve <50msec protection.

Alternatively, simple IGP convergence would ensure a LoC < second as the number of impacted important IGP routes in a seamless architecture is much smaller than 2960 routes.

5.1.8.5.5. PE2 failure

F1 is not impacted.

F2 is impacted and the LoC is sub-300msec thanks to IGP convergence and BGP PIC.

The detection of the primary nhop failure (PE2 down) is performed by a single-area IGP convergence.

In this specific case, the convergence should be much faster than <sec as very few prefixes are impacted upon an edge node failure. Reusing the introduction on IGP convergence presented in an earlier section and assuming 2 important impacted prefixes (two loopbacks per edge node), one would expect that PE2's failure is detected in 260msec + 2*0.250msec.

If BGP PIC is used on the ingress PE (PE1) then the LoC is the same as for IGP convergence. The LoC for BGP/L3VPN traffic upon PE2 failure is thus expected to be <300msec.

Provided that all the deployment considerations have been met, LoC is sub-50msec with BGP egress node FRR.

5.1.8.5.6. PE2's PE-CE link failure

F1 is not impacted.

F2 is impacted and the LoC is sub-50msec thanks to local interface failure detection and local forwarding to the backup PE. Forwarding to the backup PE is achieved with hierarchical data plane and local-repair of BGP egress link providing fast re-route to the backup BGP nhop PE.

5.1.8.5.7. ABR node failure between right region and the core

F2 is not impacted.

F1 is impacted. We analyze the LoC for F1 for both BGP PIC and BGP egress node FRR.

LoC is sub-600msec thanks to BGP PIC.

The detection of the primary nhop failure (ABR down) is performed by a multi-area IGP convergence.

First, the two (local) ABR's between the left and core regions must complete the core IGP convergence. The analysis is similar to the loss of PE2. We would thus expect that the core convergence completes in ~260msec.

Second, the IGP convergence in the left region will cause all AGN1 routers to detect the loss of the remote ABR. This second IGP convergence is very similar to the first one (2 important prefixes to remove) and hence should also complete in ~260msec.

Once an AGN1 has detected the loss of the remote ABR, thanks to the BGP PIC, in-place modification of shared BGP path-list and pre-computation of BGP backup nhop, the AGN1 reroutes flow F1 via the alternate remote ABR in a few msec's [##BGP-PIC].

As a consequence, the LoC for F1 upon remote ABR failure is thus expected to be <600msec.

Provided that all the deployment considerations have been met, LoC is sub-50msec with BGP egress node FRR.

5.1.8.5.8. Link or node failure within the right aggregation region

F1 is impacted but LoC <50msec thanks to LFA FRR. No uloop will occur during the IGP convergence following the LFA protection.

Note: if LFA is not available (other topology then case study one) or if LFA is not enabled, then the LoC would be < second as the number of impacted important IGP route in a seamless architecture is much smaller than 2960.

F2 is not impacted.

5.1.8.5.9. AGN (connected to AN2) node failure

F1 is impacted but LoC <50msec thanks to LFA FRR. No uloop will occur during the IGP convergence following the LFA protection.

Note: remember that AGN redistributes the static routes to ANs within ISIS. The loss of an AGN on the IGP path to AN2 is thus seen as an IGP route impact and hence LFA FRR is applicable.

Note: if LFA is not available (other topology then case study one) or if LFA is not enabled, then the LoC would be < second as the number of impacted important IGP route in a seamless architecture is much smaller than 2960.

F2 is not impacted.

5.1.8.5.10. AGN-AN2 link failure

F2 is not impacted.

F1 is impacted.

LoC is sub-300msec with IGP convergence as only one prefix needs to be updated.

Sub-50msec could be guaranteed provided that the LFA implementation supports a redistributed static as a native IGP route.

5.1.8.5.11. AN2 failure

F1 is impacted and the LoC lasts until the AN is recovered.

F2 is not impacted.

5.1.8.5.12. Summary - Loss of connectivity upon any failure

The Seamless MPLS architecture illustrated in deployment case study 1 guarantees sub-50msec upon any link or node failures.

5.1.8.6. Network Resiliency and Simplicity

A fundamental aspect of the Seamless MPLS architecture is the requirement for operational simplicity.

In a network with 10k of IGP/BGP nodes and 100k of MPLS-enabled nodes, it is extremely important to provide a simple operational process.

LFA FRR plays a key role in providing simplicity as it is an automated behavior which does not require any configuration or interoperability testing.

More specifically, [RFC6571] plays a key role in the Seamless MPLS architecture as it describes simple design guidelines which determiniscally ensure LFA coverage for any link and node in the aggregation regions of the network. This is key as it provides for a simple <50msec protection for the vast majority of the node and link failures (>90% of the IGP/BGP3107 footprint at least).

If the guidelines cannot be met, then either the designer will rely on (1) augmenting native LFA coverage with remote LFA [I-D.ietf-rtgwg-remote-lfa], or (2) augmenting native LFA coverage with RSVP, or (3) a full-mesh TE FRR model, or (4) IGP convergence. The first option provides an automatic and fairly simple sub-50msec protection as LFA without introducing any additional protocols. The second option provides the same sub-50msec protection as LFA, but introduces additional RSVP LSPs. The thrid option optimizes for sub-50msec protection, but implies a more complex operational model. The fourth option optimizes for simple operation but only provides <1 sec protection. Up to each designer to arbitrate between these three options versus the possibility to engineer the topology for native LFA protection.

A similar choice involves protection against ABR node failure and L3VPN PE node failure. The designer can either use BGP PIC or BGP egress node FRR. Up to each designer to asssess the trade-off between the valuation of sub-50msec instead of sub-1sec versus additional operational considerations related to BGP egress node FRR.

5.1.8.7. Conclusion

The Seamless MPLS architecture illustrated in deployment case study 1 guarantees sub-50msec for majority of link and node failures by using LFA FRR, except ABR and L3PE node failures, and PE-CE link failure.

L3VPN PE-CE link failure can be protected with sub-50msec restoration, by using hierarchical data plane and local-repair fast-reroute to the backup BGP nhop PE.

ABR and L3PE node failure can be protected with sub-50msec restoration, by using BGP egress node FRR.

Alternatively, ABR and L3PE node failure can be protected with sub-1sec restoration using BGP PIC.

5.1.9. BGP Next-Hop Redundancy

An aggregation domain is connected to the core network using two redundant area boarder routers, and MPLS hierarchy is applied on these ABRs. MPLS hierarchy helps scale the FIB but introduces additional complexity for the rerouting in case of ABR failure. Indeed ABR failure requires a BGP converge to update the inner MPLS hierarchy, in addition to the IGP converge to update the outer MPLS hierarchy. This is also expected to take more time as BGP convergence is performed after the IGP convergence and because the number of prefixes to update in the FIB can be significant. This is clearly a drawback, but the architecture allows for two "local protection" solutions which restore the traffic before the BGP convergence takes place.

BGP PIC would be required on all edge LSR involved in the inner (BGP) MPLS hierarchy. Namely all routers except the AN which are not involved in the inner MPLS hierarchy. It involves pre-computing and pre-installing in the FIB the BGP backup path. Such back up path are activated when the IGP advertise the failure of the primary path. For specification see [BGP-PIC1, 2##].

BGP egress node FRR would be required on the egress LSR involved in the inner (BGP) MPLS hierarchy, namely AGN, ABR and L3VPN PEs. For specification see [PE-FRR], [ABR-FRR], [BGP-edge-FRR##].

Both approaches have their pros and cons, and the choice is left to each Service Provider or deployment based on the different requirements. The key point is that the seamless MPLS architecture can handle fast restoration time, even for ABR failures.

5.2. Scalability Analysis

5.2.1. Control and Data Plane State for Deployment Scenarios

5.2.1.1. Introduction

Let's call:

- #AN the number of Access Node (AN) in the seamless MPLS domain

- #AGN the number of AGgregation Node (AGN) in the seamless MPLS domain

- #Core the number of Core (Core) in the core network

- #Area the number of aggregation routing domains.

Let's take the following assumptions:

- Aggregation equipments are equally spread across aggregation routing domains
- the number of IGP links is three times the number of IGP nodes
- the number of IGP prefixes is five times the number of IGP nodes (links prefixes + 2 loopbacks)
- Each Access Node needs to have up to 1,000 (1k) LSPs. LSP scale requirement is driven by expected AN access line capacity and the sum of LSPs required for connectivity to PE routers providing edge services as well as a remote ANs. Number and type of services accessed by the AN has also an impact. Hence it is very service specific. Note that if the number of remote PE/LSPs is higher than the capacity of the AN, some route aggregation scheme can be enabled at the service layer, e.g. using RFC7024 for IP VPN. It is assumed that 100 LSPs per AN (10% of total) are FECs that are outside of their routing domain. Those 100 remote FEC are the same for all Access Nodes of a given AGN.

The following sections roughly evaluate the scalability, both in absolute numbers and relatively with the number of Access Node which is the biggest scalability factor.

5.2.1.2. Core Domain

The IGP & LDP core domain are not affected by the number of access nodes:

IGP:

```
      node : #Core ~ o(1)

      links : 3*#Core ~ o(1)

      IP prefixes : 5*#Core + #Area ~ o(1)
```

LDP FEC:

```
      #Core ~ o(1)
```

Core TN FIBs grows linearly with the number of node in the core domain. In other word, they are not affected by AGN and AN nodes:

Core TN:

IP FIB : 5*#Core + #Area ~ o(1)

MPLS ILM (LFIB) : #Core ~ o(1)

BGP carries all AN routes which is significant. However, all AN routes are only needed in the control plane, possibly in a dedicated BGP Route Reflector (just like for BGP/MPLS VPNs) and not in the forwarding plane. The number of routes (100k) is smaller than the number of number of routes in the Internet (300k and rising) or in major VPN SP (>500k and rising) so the target can be handled with current implementations. In addition, AN routes are internal routes whose churn and instability is smaller and more under control than external routes.

BGP Route Reflector (RR)

NLRI : #AN ~ o(n)

path : 2*#AN ~ o(2n)

ABR handles both the core and aggregations routes. They do not depend on the total number of AN nodes, but only on the number of AN in their aggregation domain.

ABR:

IP FIB : 5*#Core + (5*#AGN + #AN) / #Area + #Area ~ o(#AN/#Area)

MPLS ILM (LFIB) : #Core + (#AGN + #AN) / #Area ~ o(#AN / #Area)

5.2.1.3. Aggregation Domain

In the aggregation domain, IGP & LDP are not affected by the number of access nodes outside of their domain. They are not affected by the total number of AN nodes:

IGP:

node : #AGN / #Area ~ o(1)

links : 3*#AGN / #Area ~ o(1)

IP prefixes : #Core + #Area + (5*#AGN + #AN) / #Area ~ o(#AN *5/ #Area)

+ plus one aggregate per area (because the number of IP prefixes is equal to 1 loopback per core node), plus 5 prefixes per AGN in the area, plus 1 prefix per AN in the area.

LDP FEC:

Core + (#AGN + #AN) / #Area ~ o(#AN / #Area)

+ plus one aggregate per area (because the number of IP prefixes is equal to 1 loopback per core node), plus 5 prefixes per AGN in the area, plus 1 prefix per AN in the area.

AGN FIBs grows with the number of node in the core area, in their aggregation area, plus the number of inter domain LSP required by the AN attached to them. They do not depend on the total number of AN nodes. In the BGP control plane, AGN also needs to handle all the AN routes.

AGN:

IP FIB : #Core + #Area + (5*#AGN + #AN) / #Area ~ o(#AN *5/ #Area)

MPLS ILM (LFIB) : #Core + (#AGN + #AN) / #Area + 100 ~ o(#AN / #Area)

AN FIBs grow with the connectivity requirement of the services. They do not depend on the number of AN, AGN, SN or any others nodes.

AN:

IP RIB : 1 ~ o(1)

MPLS LIB : 1k ~ o(1)

IP FIB : 1 ~ o(1)

MPLS ILM : 1k ~ o(1) if the AN is used in transit by other ANs and hence is an LSR (use case #2 having chained AN)

MPLS ILM : 0 ~ o(1) if the AN is not providing transit to others ANs (use case #1 where AN are final leaf)

MPLS NHLFE : 1k ~ o(1)

5.2.1.4. Summary

AN requirements are kept to a minimum. BGP is not required on ANs and the size of their FIB is driven only by their own connectivity requirements. In the FIB scale analysis described in sections 5.2.1.x, it was assumed that any single AN will need no more than 1,000 LSPs. This assumption is based on the expected AN access line capacity and LSPs required for connectivity to PE routers providing edge services as well as a sparse mesh of connectivity between ANs.

In the core area, IGP and LDP are not affected by the nodes in the aggregation domains. In particular they do not grow with the number of AGNs or ANs.

In the aggregation areas, IGP and LDP are affected by the number of core nodes and the number of AGNs and ANs in their area. They are not affected by the total number of AGNs or ANs in the seamless MPLS domain.

No FIB of any node is required to handle the total number of AGNs or ANs in the Seamless MPLS domain. In other words, the number of AGNs and ANs in the Seamless MPLS domain is not a limiting factor, and the design can be scaled by growing the number of areas. The main limitation is the MPLS connectivity requirements on the AN, i.e. mainly the number of LSP needed per AN. Another limitation may be the number of different LSPs required by ANs attached to specific AGN. However, given the foreseen deployments and current AGN capabilities, this is not expected to be a limiting factor.

In the control plane, BGP will typically handle all AN routes. This is expected to be substantial, but again the target deployment scale are well within the capabilities of current equipment . In addition, if required, additional techniques could be used to improve the scalability, based on the experience gained with scaling BGP/MPLS VPN (e.g. route partitioning between RR planes, route filtering (static or dynamic with ORF or route refresh) between ANs and on AGN to improve AGN scalability.

5.2.1.5. Numerical application for use case #1

As a recap, targets for deployment scenario 1 are:

o Number of Aggregation Domains 100

o Number of Backbone Nodes 1,000

o Number of AGgregation Nodes 10,000

- Number of Access Nodes 100,000

This gives the following scaling numbers for each category of nodes:

- AN IP FIB 1
- AN MPLS ILM 0
- AN MPLS NHLFE 1,000
- AGN IP FIB 2,600
- AGN MPLS ILM (LFIB) 2,200
- ABR IP FIB 6,600
- ABR MPLS ILM (LFIB) 2,100
- TN IP FIB 5,100
- TN MPLS ILM (LFIB) 1,000
- RR BGP NLRI 100,000
- RR BGP paths 200,000

5.2.1.6. Numerical application for use case #2

As a recap, targets for deployment scenario 1 are:

- Number of Aggregation Domains 30
- Number of Backbone Nodes 150
- Number of AGgregation Nodes 1,500
- Number of Access Nodes 40,000

This gives the following scaling numbers for each category of nodes:

- AN IP FIB 1
- AN MPLS ILM 1,000
- AN MPLS NHLFE 1,000
- AGN IP FIB 1,763

- AGN MPLS ILM 1,633
- ABR IP FIB 2,363
- ABR MPLS ILM 1,533
- TN IP FIB 780
- TN MPLS ILM 150
- RR BGP NLRI 40,000
- RR BGP paths 80,000

6. Acknowledgements

Many people contributed to this document. The authors would like to thank Wim Henderickx, Robert Raszuk, Thomas Beckhaus, Wilfried Maas, Roger Wenner, George Swallow, Kireeti Kompella, Yakov Rekhter, Mark Tinka, Curtis Villamizar and Simon DeLord for their suggestions and review.

7. IANA Considerations

This memo does not include any requests to IANA.

8. Security Considerations

The Seamless MPLS Architecture is subject to similar security threats as any MPLS LDP deployment. It is recommended that baseline security measures are considered as described in Security Framework for MPLS and GMPLS networks [RFC5920], in the LDP specification RFC5036 [RFC5036] and in [RFC6952]including ensuring authenticity and integrity of LDP messages, as well as protection against spoofing and Denial of Service attacks. Some deployments may require increased measures of network security if a subset of Access Nodes are placed in locations with lower levels of physical security e.g. street cabinets (common practice for VDSL access). In such cases it is the responsibility of the system designer to take into account the physical security measures (environmental design, mechanical or electronic access control, intrusion detection), as well as monitoring and auditing measures (configuration and Operating System changes, reloads, routes advertisements).

Security aspects specific to the MPLS access network based on LDP DoD in the context of Seamless MPLS design are described in the security section of [RFC7032].

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

9.2. Informative References

[ABR-FRR] Rekhter, Y., "Local Protection for LSP tail-end node failure, MPLS World Congress 2009", .

[ACM01] Francois, P., Filsfils, C., Evans, J., and O. Bonaventure, "Archiving sub-second IGP convergence in large IP networks, ACM SIGCOMM Computer Communication Review, v.35 n.3", July 2005.

[I-D.bashandy-bgp-edge-node-frr]
Bashandy, A., Pithawala, B., and K. Patel, "Scalable BGP FRR Protection against Edge Node Failure", draft-bashandy-bgp-edge-node-frr-03 (work in progress), July 2012.

[I-D.bashandy-bgp-frr-mirror-table]
Bashandy, A., Konstantynowicz, M., and N. Kumar, "BGP FRR Protection against Edge Node Failure Using Table Mirroring with Context Labels", draft-bashandy-bgp-frr-mirror-table-00 (work in progress), October 2012.

[I-D.bashandy-bgp-frr-vector-label]
Bashandy, A., Kumar, N., and M. Konstantynowicz, "BGP FRR Protection against Edge Node Failure Using Vector Labels", draft-bashandy-bgp-frr-vector-label-00 (work in progress), July 2012.

[I-D.bashandy-idr-bgp-repair-label]
Bashandy, A., Pithawala, B., and J. Heitz, "Scalable, Loop-Free BGP FRR using Repair Label", draft-bashandy-idr-bgp-repair-label-04 (work in progress), May 2012.

[I-D.bashandy-isis-bgp-edge-node-frr]
Bashandy, A., "IS-IS Extension for BGP FRR Protection against Edge Node Failure", draft-bashandy-isis-bgp-edge-node-frr-01 (work in progress), September 2012.

[I-D.bashandy-mpls-ldp-bgp-frr]
Bashandy, A. and K. Raza, "LDP Extension for FRR Edge Node Protection in BGP-Free LDP Core", draft-bashandy-mpls-ldp-bgp-frr-00 (work in progress), March 2012.

[I-D.ietf-rtgwg-remote-lfa]
Bryant, S., Filsfils, C., Previdi, S., Shand, M., and S. Ning, "Remote LFA FRR", draft-ietf-rtgwg-remote-lfa-06 (work in progress), May 2014.

[I-D.minto-2547-egress-node-fast-protection]
Jeganathan, J., Gredler, H., and B. Decraene, "2547 egress PE Fast Failure Protection", draft-minto-2547-egress-node-fast-protection-02 (work in progress), July 2013.

[I-D.rtgwg-bgp-pic]
Bashandy, A., Filsfils, C., and P. Mohapatra, "Abstract", draft-rtgwg-bgp-pic-02 (work in progress), October 2013.

[PE-FRR] Le Roux, J., Decraene, B., and Z. Ahmad, "Fast Reroute in MPLS L3VPN Networks - Towards CE-to-CE Protection, MPLS 2006 Conference", .

[RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, May 2001.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

[RFC5283] Decraene, B., Le Roux, JL., and I. Minei, "LDP Extension for Inter-Area Label Switched Paths (LSPs)", RFC 5283, July 2008.

[RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.

[RFC5881] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)", RFC 5881, June 2010.

[RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.

[RFC6571] Filsfils, C., Francois, P., Shand, M., Decraene, B., Uttaro, J., Leymann, N., and M. Horneffer, "Loop-Free Alternate (LFA) Applicability in Service Provider (SP) Networks", RFC 6571, June 2012.

[RFC6952] Jethanandani, M., Patel, K., and L. Zheng, "Analysis of BGP, LDP, PCEP, and MSDP Issues According to the Keying and Authentication for Routing Protocols (KARP) Design Guide", RFC 6952, May 2013.

[RFC7032] Beckhaus, T., Decraene, B., Tiruveedhula, K., Konstantynowicz, M., and L. Martini, "LDP Downstream-on-Demand in Seamless MPLS", RFC 7032, October 2013.

Authors' Addresses

Nicolai Leymann (editor)
Deutsche Telekom AG
Winterfeldtstrasse 21
Berlin 10781
DE

Phone: +49 30 8353-92761
Email: n.leymann@telekom.de

Bruno Decraene
Orange
38-40 rue du General Leclerc
Issy Moulineaux cedex 9 92794
FR

Email: bruno.decraene@orange.com

Clarence Filsfils
Cisco Systems
Brussels
Belgium

Email: cfilsfil@cisco.com

Maciek Konstantynowicz (editor)
Cisco Systems
London
United Kingdom

Email: maciek@cisco.com

Dirk Steinberg
Steinberg Consulting
Ringstrasse 2
Buchholz 53567
DE

Email: dws@steinbergnet.net