SPRING
Internet-Draft
Intended status: Standards Track
Expires: July 11, 2021

S. Hegde
C. Bowers
Juniper Networks Inc.
X. Xu
Alibaba Inc.
A. Gulko
Refinitiv
A. Bogdanov
Google Inc.
J. Uttaro
ATT
L. Jalil
Verizon
M. Khaddam
Cox communications
A. Alston
Liquid Telecom

January 7, 2021

Seamless Segment Routing
draft-hegde-spring-mpls-seamless-sr-04

Abstract

In order to operate networks with large numbers of devices, network operators organize networks into multiple smaller network domains. Each network domain typically runs an IGP which has complete visibility within its own domain, but limited visibility outside of its domain. Seamless Segment Routing (Seamless SR) provides flexible, scalable and reliable end-to-end connectivity for services across independent network domains. Seamless SR accommodates domains using SR, LDP, and RSVP for MPLS label distribution as well as domains running IP without MPLS (IP-Fabric).It also provides seamless connectivity across domains having different IPv6 technologies such as SRv6 and SRm6.

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].

Status of This Memo

This Internet-Draft will expire on July 11, 2021.

Copyright Notice

Table of Contents

1. Introduction

Evolving wireless access technology and cloud applications are expected to place new requirements on the packet transport networks. These services are contributing to significantly higher bandwidth throughput which in turn leads to a growing number of transport network devices. As an example, 5G networks are expected to require up to 250Gbps in the fronthaul and up to 400Gbps in the backhaul. There is a desire to allow many network functions to be virtualized and cloud native. In order to support latency-sensitive cloud-native network functions, packet transport networks should be capable of providing low-latency paths end-to-end. Some services will require low-latency paths while others may require different QoS properties. The network should be able to differentiate between the services and provide corresponding SLA transport paths. In addition, as these applications become more sensitive and less loss tolerant, more and more emphasis is placed on overall service availability and reliability.

The Seamless SR architecture builds upon the Seamless MPLS architecture and caters to new requirements imposed by the 5G transport networks and the cloud applications. [I-D.ietf-mpls-seamless-mpls], contains a good description of the Seamless MPLS architecture. Although [I-D.ietf-mpls-seamless-mpls] has not been published as an RFC, it serves as a useful description of the Seamless MPLS architecture. [I-D.ietf-mpls-seamless-mpls] describes the Seamless MPLS architecture, which uses LDP and/or RSVP for intra-domain label distribution, and BGP-LU [RFC3107] for end-to-end label distribution. Seamless SR focuses on using segment routing for intra-domain label distribution. The mechansims described in this document are equally applicable to intra-domain tunneling mechanisms deployed using RSVP and/or LDP.

By using segment routing for intra-domain label distribution, Seamless SR is able to easily support both SR-MPLS on IPv4 and IPv6 networks. This overcomes a limitation of the classic Seamless MPLS architecture, which was limited to run MPLS on IPv4 networks in practice. Seamless SR (like Seamless MPLS) can use BGP-LU (RFC 3107) to stitch different domains. However, Seamless SR can also take advantage of BGP Prefix-SID [RFC8669] to provide predictable and deterministic labels for the inter-domain connectivity.

The basic functionality of the Seamless SR architecture does not require any enhancements to existing protocols. However, in order to support end-to-end service requirements across multiple domains, protocol extensions may be needed. This draft discusses use cases, requirements, and potential protocol enhancements.

Section Section 3 describes usecases and section Section 4 describes requirements arising out of the usecases. There may be alternative solutions available to solve the same usecases. This document does not exclude other possible solutions. Section Section 5 refers to possible alternative solutions and describes how the different archictures can co-exist in the same network or be deployed independently.

2. Terminology

This document uses the following terminology

- o 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 and Cell Site Routers in 5G networks. Access nodes have only limited MPLS functionalities in order to reduce complexity in the access network.

- o Pre-Aggregation Node (P-AGG): A pre-aggregation node (P-AGG) is a node which aggregates several access nodes (ANs).

- o Aggregation Node (AGG): A aggregation node (AGG) is a node which aggregates several pre-aggregation nodes (P-AGG).

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

- o Label Switch Router (LSR): Label Switch router are pure transit nodes. ideally have no customer or service state and are therefore decoupled from service creation.

- o Use Case: Describes a typical network including service creation points and distribution of remote node loopback prefixes.

Figure 1: Terminology

3. Use Cases

3.1. Service provider network

Service provider transport networks use multiple domains to support scalability. For this analysis, we consider a representative network design with four level of hierarchy: access domains, pre-aggregation domains, aggregation domains and a core. (See Figure 2). The 5G transport networks in particular are expected to scale to very large number of access nodes due to the shorter range of the 5G radio

technology. The networks are expected to scale up to one million nodes.

```
                 +-------+   +-------+   +------+   +------+
                 |       |   |       |   |      |   |      |
              +--+ P-AGG1+---+ AGG1  +---+ ABR1 +---+ LSR1 +--> to ABR
             /   |       |  /|       |   |      |   |      |
      +----+/    +-------+\/ +-------+   +------+  /+------+
      | AN |              /\                     \/
      +----+\    +-------+  \+-------+   +------+/\ +------+
             \   |       |   |       |   |      |  \|      |
              +--+ P-AGG2+---+ AGG2  +---+ ABR2 +---+ LSR2 +--> to ABR
                 |       |   |       |   |      |   |      |
                 +-------+   +-------+   +------+   +------+

      ISIS L1       ISIS L2                   ISIS L2

      |-Access-|--Aggregation Domain--|---------Core-----------------|
```

Figure 2: 5G network

Many network functions in a 5G network will be virtualized/ containerized and distributed across multiple data centers. Virtualized network functions are instantiated dynamically across different compute resources. This requires that the underlying transport network supports the stringent SLA on end-to-end paths.

5G networks support variety of service use cases that require end-to-end slicing. In certain cases the end-to-end connectivity requires differentiated forwarding capabilities. Seamless SR architecture should provide the ability to establish end-to-end paths that satisfy the required SLAs. For example, end user requirement could be to establish a low latency path end-to-end. The System Architecture for the 5G System [TS.23.501-3GPP] currently defines four standardized Slice/Service Types: Enhanced Mobile Broadband (eMBB), Ultra-Reliable Low Latency Communication (URLLC), massive Internet of Things (mIoT), Vehicle to everything (V2X). The Seamless SR should support end-to-end Service Level Objectives(SLO) to allow the creation of network slices with these four Slice/Service Types.

Many deployments consist of ring topologies in the access and aggregation networks. In the ring topologies, there are at most two forwarding paths for the traffic, whereas the core networks consist of nodes with more denser connectivity compared to ring topologies. Thus core networks may have a larger number of TE paths while access networks will have a smaller number of TE paths. The Seamless SR

architecture should support the ability to have more TE paths in one domain and lesser number of TE paths in another domain and provide the ability to effectively connect the domains end-to-end while satisfying end-to-end constraints.

3.2. Large scale WAN networks

As WAN networks grow beyond several thousand nodes, it is often useful to divide the network into multiple IGP domains, as illustrated in Section 3.2. Separate IGP domains increase service availability by establishing a constrained failure domain. Smaller IGP domains may also improve network performance and health by reducing the device scale profile (including protocol and FIB scale).

```
                 +-------+     +-------+     +-------+
                 |       |     |       |     |       |
                 |      ABR1  ABR2    ABR3   ABR4    |
                 |       |     |       |     |       |
              PE1+DOMAIN1+-----+DOMAIN2+-----+DOMAIN3+PE2
                 |       |     |       |     |       |
                 |      ABR11  ABR22  ABR33  ABR44   |
                 |       |     |       |     |       |
                 +-------+     +-------+     +-------+


                |-ISIS1-|      |-ISIS2-|     |-ISIS3-|
```

Figure 3: WAN Network

These Large WAN networks often cross national boundaries. In order to meet data sovereignty requirements, operators need to maintain strict control over end-to-end traffic-engineered(TE) paths. Segment Routing provides two main solutions to implement highly constrained TE paths. Flex-algo (defined in [I-D.ietf-lsr-flex-algo]) uses prefix-SIDs computed by all nodes in the IGP domain using the same pruned topology. Highly constrained TE paths for the data sovereignty use case can also be implemented using SR-TE policies ([I-D.ietf-spring-segment-routing-policy]) built using unprotected adjacency SIDs.

Both of these approaches work well for intra-domain TE paths. However, they both have limitations when one tries to extend them to the creation of highly constrained inter-domain TE paths. A goal of seamless SR is to be able to create highly constrained inter-domain TE paths in a scalable manner.

Some deployments may use a centralized controller to acquire the topologies of multiple domains and build end-to-end constrained paths. This can be scaled with hierarchical controllers. However, there is still significant risk of a loss of network connectivity to one or more controllers, which can result in a failure to satisfy the strict requirements of data sovereignty. The network should have pre-established TE paths end-to-end that don't rely on controllers in order to address these failure scenarios.

3.3. Data Center Interconnect (DCI) Networks

Data centers are playing an increasingly important role in providing access to information and applications. Geographically diverse data centers usually connect via a high speed, reliable and secure core network.

```
                 +-------+     +-------+     +-------+
                 |     ASBR1 ASBR2 ASBR3   ASBR4     |
                 |       |     |       |     |       |
              PE1+  DC1  +-----+  CORE +-----+  DC2  +PE2
                 |    ASBR11  ASBR22 ASBR33 ASBR44   |
                 |       |     |       |     |       |
                 +-------+     +-------+     +-------+


                 |-ISIS1-|      |-ISIS2-|    |-ISIS3-|
```

Figure 4: DCI Network

In many Data Center deployments, applications require end-to-end path diversity and/or end-to-end low latency paths. In certain cases it is desirable to have a uniform technology deployed in the core as well as in the Data Centers to create these SLA paths. Such uniformity simplifies the network to a great extent. In certain other cases, the datacenter environments may deploy SRv6 and the core network may be running MPLS. It is desirable for a solution to only require service-related configurations on the access end-points where services are attached, avoiding service-related configurations on the ABR/ASBR nodes.

3.4. Service Function Chaining

[RFC7665] defines service functions chaining as an ordered set of service functions and automated steering of traffic through these set of service functions. There could be a variety of service functions such as firewalls, parental control, CGNAT etc. In 5G networks these

functions may be completely virtualised or could be a mix of virtualized functions and physical appliances. It is required that the inter-domain solution caters to the service function chaining requirements.

3.5. Multicast Use cases

Multicast services such as IPTV and multicast VPN also need to be supported across a multi-domain service provider network.

```
                 +---------+---------+---------+
                 |         |         |         |
                 S1       ABR1      ABR2       R1
                 | Metro1  |  Core    |  Metro2 |
                 |         |         |         |
                 S2       ABR11     ABR22      R2
                 |         |         |         |
                 +---------+---------+---------+


                 |-ISIS1-|  |-ISIS2-|  |-ISIS3-|
```

Figure 5: Multicast usecases

Figure 5 shows a simplified multi-domain network supporting multicast. Multicast sources S1 and S2 lie in a different domain from the receivers R1 and R2. Using multiple IGP domains presents a problem for the establishment of multicast replication trees. Typically, a multicast receiver does a reverse path forwarding (RPF) lookup for a multicast source. One solution is to leak the routes for multicast sources across the IGP domains. However, this can compromise the scaling properties of the multi-domain architecture. SR-P2MP [I-D.voyer-pim-sr-p2mp-policy] offers a solution for both intra-domain and inter-domain multicast. However, it does not accommodate deployments using existing intra-domain multicast technology, such as mLDP [RFC6388] in some of the domains. A solution should accommodate a mixture of existing and newer technologies to better facilitate coexistence and migration.

4. Requirements

This section provides a summary of requirements derived from the use cases described in previous sections.

4.1. MPLS Transport

The architecture SHOULD provide MPLS transport between two service endpoints regardless of whether the two end-points are in the same IGP domain, different IGP domains, or in different autonomous systems.

The MPLS transport SHOULD be supported on IPv4, IPv6, and dual-stack networks.

4.2. SLA Guarantee

The architecture SHOULD allow the creation of paths that support end-to-end SLAs. The paths should for example obey constraints related to latency, diversity, bandwidth and availability.

The architecture SHOULD support end-to-end network slicing as described by 5G transport requirements [TS.23.501-3GPP].

4.3. Scalability

The architecture SHOULD be able to support up to 1 million nodes.

The architecture SHOULD facilitate the use of access nodes with low RIB/FIB and low CPU capabilities.

The architecture SHOULD facilitate the use of access nodes with low label stacking capability.

The architecture SHOULD allow for a scalable response to network events. An individual node SHOULD only need to respond to a limited subset of network events.

Service routes on the border nodes SHOULD be minimized.

4.4. Availability

Traffic SHOULD be Fast Reroute (FRR) protected against link, node, and SRLG failures within a domain.

Traffic SHOULD be Fast Reroute (FRR) protected against border node failures.

Traffic SHOULD be Fast Reroute (FRR) protected against egress node and egress link failures.

4.5. Operations

Each domain SHOULD be independent and SHOULD not depend on the transport technology in another domain. This allows for more flexible evolution of the network.

Basic MPLS OAM mechanisms described in [RFC8029] SHOULD be supported.

End-to-end mpls ping and traceroute procedures SHOULD be supported.

Ability to validate the path inside each domain SHOULD be supported.

Statistics for inter-domain paths on the ingress and egress PE nodes as well as border nodes SHOULD be supported.

4.6. Service Mapping

The architecture SHOULD support the automated steering of traffic on to transport paths based on communities carried in the service prefix advertisements.

The architecture SHOULD support the steering of traffic on to transport paths based on the DSCP value carried in IPv4/IPv6 packets.

Traffic steering based on EXP bits in the mpls header SHOULD be supported.

Traffic steering based on 5-tuple packet filter SHOULD be supported. Source address, destination address, source port, destination port and protocol fields should be allowed.

All traffic steering mechanisms SHOULD be supported for all kinds of service traffic including VPN traffic as well as global internet traffic.

The core domain is expected to have more traffic engineering constraints as compared to metros. The ability to map the services to appropriate transport tunnels at service attachment points SHOULD be supported.

5. Alternative Solutions

The usecases and requirements discussed in this document may be solved using alternative solutions. The solutions can be divided into two broad categories.

Centralized Solutions

Distributed Solutions

5.1. Centralized Solutions

A centralized solution uses one central entity or a set of central entities that have complete visibility into end-to-end paths. The nodes and links used to construct paths may be contained in a single topology database or a set of connected topology databases. A computing entity is also aware of the resource utilization and resource availability in this topology and makes informed computation decisions to construct paths. The solution described in "Interconnecting Millions of Endpoints with Segment Routing " ([RFC8604]) is an example of a centralized architecture. [I-D.saad-sr-fa-link] describes extensions that can be used to extend this architecture to brownfield networks and provides abstractions to scale the solution.

5.2. Distributed solutions

In a distributed solution, there is no central entity with complete visibility into the end-to-end paths. Each domain independently computes a portion of an end-to-end path, and these independent sub-paths are stitched together at the border nodes between domains. The current document describes Seamless SR, an example of a distributed solution, which uses BGP-based extensions to stitch together complete end-to-end paths that satisfy certain properties. The Seamless SR architecture uses BGP-LU (RFC 3107) and BGP-Prefix-SID (RFC 8669) for end-to-end best path and BGP-CT (draft-kaliraj-idr-bgp-classful-transport-planes) for multiple SLA paths. Seamless SR solution does not exclude possibility of future protocol extensions that adhere to the principles of the architecture to provide end-to-end paths.

5.3. Choice of Solution

The centralized and the distributed solutions can independently solve the usecases and the requirements discussed in previous sections. One architecture may be more suitable for certain usecases while the other may be more suitable for some other usecase. It is solely at the discretion of the operator to choose the solution that best solves the usecases one has.

The two type of solutions are complementary to each other and can co-exist together in the same network. A network operator can use both distributed and centralized solutions in the same network to handle traffic with different requirements. For example, a network operator may find it useful to use centralized solution for traffic that requires stringent latency-bounded paths across network domains under the complete control of the network operator. However, the same network operator may choose to deploy a distributed solution for traffic that crosses a co-operating transit domain, where a centralized solution is precluded.

6. Seamless Segment Routing architecture

6.1. Solution Concepts

The solution described below makes use of the following concepts. The definitions from the draft-kaliraj-idr-bgp-classful-transport-planes have been reproduced here for readability. In case of any conflicts, text from draft-kaliraj-idr-bgp-classful-transport-planes should be used.

- Transport Class (TC): A Transport Class is defined as a collection of end-to-end MPLS paths that satisfy a set of constraints or Service Level Agreements.

- BGP-Classful Transport (BGP-CT): A new BGP family used to establish Transport Class paths across different domains.

- Route Distinguisher (RD): The Route Distinguisher is defined in RFC4364. In BGP-CT, the RD is used in BGP advertisements to differentiate multiple paths to the same loopback address. It may be useful to automatically generate RDs in order to simplify configuration.

- Route Target (RT): The Route Target extended community is carried in BGP-CT advertisements. The RT represents the Transport Class of an advertised path. Note that the RT is only carried in the BGP-CT advertisements. No BGP-VPN related configuration or VPN family advertisements are needed when BGP-CT transport paths are used to carry non-VPN traffic.

- Mapping Community (MC): The Mapping Community is the BGP extended community as defined in RFC4360. In the Seamless SR architecture, an MC is carried by a BGP-CT route and/or a service route. The MC is used to identify the specific local policy used to map traffic for a service route to different Transport Class paths. When a mapping community is advertised in a BGP-CT route it identifies the specific local policy used to map the BGP-CT route to the intra-domain tunnels.The local policy can include additional traffic steering properties for placing traffic on different Transport Class paths. The values of the MCs and the corresponding local policies for service mapping are defined by the network operator.

Figure 6: Solution Concepts

6.2. BGP Classful Transport

```
                 ----IBGP------EBGP----IBGP------EBGP-----IBGP---
                |            |     |           |     |           |

                 +-----------+     +-----------+     +-----------+
                 |           |     |           |     |           |
                 |        ASBR1+--+ASBR2    ASBR3+--+ASBR4       |
              PE1+     D1    |  X  |     D2    |  X  |     D3    +PE2
                 |        ASBR5+--+ASBR6    ASBR7+--+ASBR8       |
                 |           |     |           |     |           |
                 +-----+-----+     +-----------+     +-----------+
                      PE3

                 |---ISIS1---|      |---ISIS2---|      |---ISIS3---|
```

Figure 7: WAN Network

The above diagram shows a WAN network divided into 3 different domains. Within each domain, BGP sessions are established between the PE nodes and the border nodes as well as between border nodes. BGP sessions are also established between border nodes across domains. The goal is for PE1 to have MPLS connectivity to PE2, satisfying specific characteristics. Multiple MPLS paths from PE1 to PE2 are required in order to satisfy different SLAs. [I-D.kaliraj-idr-bgp-classful-transport-planes] defines a new BGP family called BGP-Classful Transport. The NLRI for this new family consists of a prefix and a Route Distinguisher. The prefix corresponds to the loopback of the destination PE, and RD is used to distinguish different paths to the same PE loopback. The BGP-CT advertisement also carries a Route Target. The RT specifies the Transport Class to which the BGP-CT advertisement belongs. BGP-CT mechanisms are applicable to single ownership networks that are organized into multiple domains. It is also applicable to multiple ASes with different ownership but closely co-operating administration. BGP-CT mechansims are not expected to be applied on the internet peering or between domains that have completely independent administrations.

```
                 BGP-CT advertisements for red Transport Class

            Prefix:PE2    Prefix:PE2  Prefix:PE2   Prefix:PE2   Prefix:PE2
            RD:RD1        RD:RD1      RD:RD1       RD:RD1       RD:RD1
            RT:Red        RT:Red      RT:Red       RT:Red       RT:Red(100)
            nh:ASBR1      nh:ASBR2    nh:ASBR3     nh:ASBR4     nh:PE2
            Label:L1      Label:L2    Label:L3     Label:L4     Label:L5


        PE1-------ASBR1------ASBR2---------ASBR3-------ASBR4--------PE2

                                                              VPNa Prefix:
                                                              10.1.1.1/32
                                                              RD: RD50
                                                              RT: RT-VPNa
                                                              ext-community:
                                                              Red(100)
                                                              nh: PE2
                                                              Label: S1

            +------+              +------+                   +------+
            | IL71 |              | IL72 |                   | IL73 |
            +------+   +------+   +------+      +------+     +------+
            | L1   |   | L2   |   |  L3  |      | L4   |     |  L5  |
            +------+   +------+   +------+      +------+     +------+
            | S1   |   | S1   |   |  S1  |      | S1   |     |  S1  |
            +------+   +------+   +------+      +------+     +------+

                      Label stacks along end-to-end path
                      S1 is the end-to-end service label.
            IL71, IL72, and IL73 are intra-domain labels corresponding to
                            red intra-domain paths.
```

Figure 8: BGP-CT Advertisements and Label Stacks

```
                  BGP-CT advertisements for blue Transport Class

            Prefix:PE2    Prefix:PE2  Prefix:PE2   Prefix:PE2   Prefix:PE2
            RD:RD2        RD:RD2      RD:RD2       RD:RD2       RD:RD2
            RT:Blue       RT:Blue     RT:Blue      RT:Blue      RT:Blue(200)
            nh:ASBR1      nh:ASBR2    nh:ASBR3     nh:ASBR4     nh:PE2
            Label:L11     Label:L12   Label:L13    Label:L14    Label:L15


        PE1-------ASBR1----ASBR2----------ASBR3-------ASBR4--------PE2

                                                              VPNb Prefix:
                                                              10.1.1.1/32
                                                              RD: RD51
                                                              RT: RT-VPNb
                                                              ext-community:
                                                              Blue(200)
                                                              nh: PE2
                                                              Label: S2

            +------+              +------+                    +------+
            | IL81 |              | IL82 |                    | IL83 |
            +------+   +------+   +------+      +------+     +------+
            | L11  |   | L12  |   |  L13 |      | L14  |     |  L15 |
            +------+   +------+   +------+      +------+     +------+
            | S2   |   | S2   |   |  S2  |      | S2   |     |  S2  |
            +------+   +------+   +------+      +------+     +------+

                      Label stacks along end-to-end path
                      S2 is the end-to-end service label.
            IL81, IL82, and IL83 are intra-domain labels corresponding to
                            blue intra-domain paths.
```

Figure 9: BGP-CT Advertisements and Label Stacks

For example, consider the diagram in Figure 8 and Figure 9 . The diagram shows the BGP-CT advertisements corresponding to two different end-to-end paths between PE1 and PE2. The two different paths belong to two different Transport Classes, red and blue.

The inter-domain paths created by BGP-CT Transport Classes can be used by any traffic that can be steered using BGP next-hop resolution, including vanilla IPv4 and IPv6, L2VPN, L3VPN, and eVPN. In the example above, we show how traffic from two different L3VPNs (VPNa and VPNb) is mapped onto two different BGP-CT Transport Classes (Red and Blue). The L3VPN advertisements for VPNa and VPNb are originated by PE2 as usual. PE1 receives these L3VPN advertisements

and uses the next-hop in the L3VPN advertisements to determine the path to use. In the absence of any BGP-CT Transport Classes in the network, PE1 would likely resolve the L3VPN next-hop over BGP-LU routes corresponding to the BGP best path. However, when BGP-CT Transport Classes are used, PE1 will resolve the L3VPN next-hop over a BGP-CT route.

In the example above, PE2 originates BGP-CT advertisements for the Red and Blue Transport Classes. These BGP-CT advertisements propagate across the multiple domains, causing forwarding state for the two Transport Classes to be installed at ABRs along the way. In order to create unique NLRIs for the two advertisements, PE2 uses two different RDs. In the example above, the red BGP-CT advertisement has an RD of RD1 and the blue BGP-CT advertisement has an RD of RD2. Note that the RD values used in the BGP-CT advertisement are completely independent of the RD values used in the L3VPN advertisements. In both cases, the RD values are simply a mechanism to guarantee uniqueness of a prefix/RD pair.

The RT values used in the BGP-CT advertisements are unrelated to the RT values used on the L3VPN advertisements. The L3VPN RT values identify VPN membership, as usual. The BGP-CT RT values identify Transport Class membership. In order to be able to easily map VPN traffic into BGP-CT Transport classes, it can be useful however to make an association between BGP-CT RT values and color extended community values in the L3VPN advertisements. In the example above,the RT value carried in the BGP-CT advertisement originated from PE2 for the red Transport Class is configured to correspond to the color extended community advertised in the VPN advertisement for VPNa. Similarly, the RT value for the blue Transport Class corresponds to the color extended community for VPNb. In this way, traffic on PE1 for each VPN can be mapped to a tranport class path by associating the value of the color extended community carried in the VPN advertisement with an RT value carried in a BGP-CT advertisement.

The example above also shows the label stacks at different points along the end-to-end paths for the forwarding entries which are established by the two advertisements. Labels L1-L4 are red BGP-CT labels advertised by border nodes ASBR1,2,3,and 4, while label L5 is advertised by PE2 for the red Transport Class. Labels L11-L14 are blue BGP-CT labels advertised by border nodes ASBR1,2,3,and 4, while label L15 is advertised by PE2 for the blue Transport Class.

IL71, IL72, and IL73 represent tunnels internal to the domains 1, 2, and 3 which correspond to the red Transport Class. IL81, IL82, and IL83 represent tunnels internal to the domains 1, 2, and 3 which correspond to the blue Transport Class. In this example, we assume that the intra-domain tunnels correspond to SRTE policies having red

SRTE-policy-color and blue SRTE-policy-color. Service labels are represented by S1 and S2.

Note that this example focuses on how signalling originated by PE2 results in forwarding state used by PE1 to reach PE2 on a specific Transport Class path. The solution supports the establishment of forwarding state for an arbitrary number of PEs to reach PE2. For example, PE3 in Figure 8 can reach PE2 on a red Transport Class path established using the same BGP-CT signalling. The signalling and forwarding state from ASBR1 all the way to PE2 is common to the paths used by both PE1 and PE3. This merging of signalling and forwarding state is essentially to the good scaling properties of the Seamless SR architecture. Millions of end-to-end Transport Class paths can be established in a scalable manner.

6.3. Automatically Creating Transport Classes

In order to simplify the creation of inter-domain paths, it may be desirable to automatically advertise a BGP-CT Transport Class based on the existence of an intra-domain tunnel. The RT value used on the BGP-CT advertisement is automatically derived from a property of the intra-domain tunnel that triggered its creation. How the Transpor Class RT value is derived for different types of intra-domain tunnels is discussed below.

6.3.1. Automatically Creating Transport Classes for BGP-SR-TE Intra-domain Tunnels

When the intra-domain tunnel is a BGP-SR-TE policy [I-D.ietf-idr-segment-routing-te-policy], the value of the Transport Class RT in the corresponding BGP-CT advertisement is derived from the Policy Color contained in SR Policy NLRI. The 32-bit Policy Color is directly converted to a 32-bit Transport Class RT.

6.3.2. Automatically Creating Transport Classes for Flex-Algo Tunnels

When the intra-domain tunnel is created using Flex-Algo [I-D.ietf-lsr-flex-algo], the value of the Transport Class RT in the corresponding BGP-CT advertisement is derived from the 8-bit Algorithm value carried in SR-Algorithm sub-TLV (RFC8667). The conversion from 8-bit Algorithm value to 32-bit Transport Class RT is done by treating both as unsigned integers. Note that this definition allows for intra-domain tunnels created via standardized algorithm (0-127) as well as flex-algo (128-255).

6.3.3. Auto-deriving Transport Classes for PCEP

When the intra-domain tunnel is created using PCEP, the value of the Transport Class RT in the corresponding BGP-CT advertisement is derived from the Color of the SR Policy Identifiers TLV defined in [I-D.ietf-pce-segment-routing-policy-cp]. The 32-bit Color is directly converted to a 32-bit Transport Class RT.

6.4. Inter-domain flex-algo with BGP-CT

Flex-algo (defined in [I-D.ietf-lsr-flex-algo]) provides a mechanism to separate routing planes. Multiple algorithms are defined and prefix-SIDs are advertised for each algorithm. BGP-CT can be used to advertise these flex-algo SIDs in BGP-CT. BGP Prefix-SID (RFC 8669) is an attribute and can be carried in the BGP-CT NLRI. Multiple transport classes that correspond to each of the flex-algo in IGP domain are defined. These Transport Classes advertise the IGP flex-algo SIDs in the prefix-SIDs attribute in the BGP-CT NLRI.

6.5. Applicability to color-only policies

Color-only policies consist of (nullEndpont, color) as specified in [I-D.ietf-spring-segment-routing-policy]. Special steering mechanisms are defined with "CO" flags defined in the color extended community [I-D.ietf-idr-segment-routing-te-policy]. Color-only policies can be advertised in BGP-CT with the prefix being NULL (0.0.0.0/32 or 0::0/128). Seperate RD will be advertised for each NULL advertisement with different color. The Route target carries the Policy Color contained in SR Policy NLRI. The steering mechanisms defined in [I-D.ietf-spring-segment-routing-policy] MUST be honoured while resolving services prefixes on the BGP-CT advertisements.

6.6. Data sovereignty

```
              +-----------+     +-----------+     +-----------+
              |           |     |  +-+  AS2 |     |           |
              |           A1+--+A2 | |      A3+--+A4          |
           PE1+    AS1    |     |  |Z|      |     |     AS3   +PE3
              |           A5+--+A6 | |      A7+--+A8          |
              |           |     |  +-+      |     |           |
              +--A13--A15-+     +-A17--A19--+     +-----------+
                 |     |           |    |
                 |     |           |    |
                 |     |           |    |
              +--A14--A16-+     +-A18--A20--+
              |           |     |           |
              |          A9+--+A10          |
           PE4+   AS4     |     |   AS5     |
              |          A11+-+A12          |
              |           |     |           |
              +-----------+     +-----------+
```

Figure 10: Multi domain Network

Consider a WAN network with multiple ASes as shown in the diagram Figure 10. The ASes roughly correspond to the geographical location of the nodes. In this example, we assume that each AS corresponds to a continent. The data sovereignty requirement in this example is that certain traffic from PE1(in AS1) to PE3(in AS3) must not cross through country Z in AS2. As indicate by the location of country Z in the diagram, all paths that go directly from AS1 to AS3 through AS2 necessarily passes through country Z. Using BGP-LU to provide connectivity from PE1 to PE3 would generally result in a path that goes from AS1 to AS2 to AS3, which does not satisfy the data sovereignty requirement in this example. Instead, the solution using BGP-CT will go from AS1 to AS4 to AS5 to AS2 to AS3. BGP-CT will ensure that when the traffic passes through AS2, only intra-domain paths satisfying the data sovereignty requirement will be used.

Within AS2, there are several different intra-domain TE mechanisms that can be used to exclude links that pass through country Z. For example, RSVP-TE or flex-algo can be used to create intra-domain paths that satisfy the data sovereignty requirement. BGP-CT allows the constrained intra-domain paths to satisfy requirements for end-to-end inter-domain paths. LSPs created by RSVP-TE or Flex-algo that satisfy the "exclude country Z" constraint are associated with a color Green. A Green Transport Class is defined on border nodes in all ASes. This Green Transport Class is associated with a mapping community called Not-Z.

In AS2, the ASBRs are configured such that the presence of the mapping community Not-Z in BGP-CT routes results in a strict route resolution mechanism for those routes. A BGP-CT route carrying the color extended community Not-Z will only resolve on the Green Tranport Class. So it will only use Green intra-domain tunnels.

In AS1, AS3, AS4, and AS5, no links pass through country Z, so all intra-domain paths automatically satisfy the data sovereignty requirement. So there is no need for the creation of Green intra-domain tunnels. In these ASes, the presence of the mapping community Not-Z in BGP-CT routes results in resolution on best-effort paths. Even though the ASBRs in these ASes do not need to create Green intra-domain tunnels, they still need to allocate labels to identify traffic using the Green Transport Class. These labels will be used by the ASBRs in AS2 to put traffic on the Green intra-domain tunnels in AS2.

The requirement is that only a subset of traffic honor the data sovereignty requirement. The service prefixes from PE1 to PE2 that need to honor the data sovereignty requirement will be associated with Green extended color community in the service advertisements. This will result in PE1 using the BGP-CT labels corresponding to {PE2, Green} to forward the traffic. BGP-CT labels corresponding to {PE2, Green} will exist at every ASBR along the path. The traffic originating on PE1, will be associated with Green color community. The bottom-most label in the packet consists of a VPN label. Above the VPN label, BGP-CT label is imposed. Above BGP-CT label, the intra-domain transport label is imposed. Let us assume the traffic from PE1 needs to go to PE2 through AS1, AS4, AS5, AS2, and AS3. The BGP-CT label for {PE2, Green} will be swapped at the border nodes.

Note that end-to-end inter-domain data sovereignty can in principle be accomplished using BGP-LU with multiple loopbacks and associating those loopbacks to appropriate transport tunnels at every border node in every domain. This is very configuration intensive and require multiple loopbacks. BGP-CT builds on the basic mechanisms of BGP-LU while greatly simplifying such use cases.

6.7. Interconnecting IP Fabric Data Centers

```
            Prefix:TOR2   Prefix:TOR2 Prefix:TOR2  Prefix:TOR2  Prefix:TOR2
            RD:RD2        RD:RD2      RD:RD2       RD:RD2       RD:RD2
            RT:Blue       RT:Blue     RT:Blue      RT:Blue      RT:Blue
            nh:ASBR1      nh:ASBR2    nh:ASBR3     nh:ASBR4     nh:TOR2
            Label:L11     Label:L12   Label:L13    Label:L14    Label:L15

          +-----------+       +-----------+        +-----------+
          |           ASBR1  ASBR2     ASBR3      ASBR4        |
          |           |       |           |        |           |
      TOR1+  DC1      +-------+  CORE     +--------+  DC2      +TOR2
          |           ASBR11 ASBR22     ASBR33    ASBR44       |
          |           |       |           |        |           |
          +-----------+       +-----------+        +-----------+


            +------+              +------+                   +------+
            | UDP  |              | IL82 |                   |  UDP |
            +------+   +------+   +------+      +------+     +------+
            | L11  |   | L12  |   |  L13 |      | L14  |     |  L15 |
            +------+   +------+   +------+      +------+     +------+
            | S2   |   | S2   |   |  S2  |      | S2   |     |  S2  |
            +------+   +------+   +------+      +------+     +------+


              Label stacks along end-to-end path
                      S2 is the end-to-end service label.
            IL82, is intra-domain labels corresponding to
                            blue intra-domain paths.
```

Figure 11: Operation in IP fabric

Many data center networks consist of IP fabrics which do not have MPLS packet processing capability. A common requirement is that traffic originated from an IP Fabric data center needs to satisfy certain constraints in the MPLS-enable core, for example, only using a subset of links (blue links). It is useful for the traffic originating in an IP Fabric DC to carry information that allows the MPLS-enable core to treat it accordingly. MPLSoUDP, as defined in [RFC7510], is a mechanism where a UDP header is imposed on an MPLS packets on the border nodes. In Figure 11 above, the traffic needs to take blue paths in the core. The Blue Transport Class is defined on the ASBRs. In the core, Blue intra-domain tunnels are created. The BGP-CT advertisements for the Blue Transport Class are as shown in the diagram. The BGP-CT advertisements originate at TOR2 and propagate through all the ASBRs, until finally reaching TOR1. Within DC1, traffic is encapsulated with a UDP header. Traffic with the UDP header gets decapsulated at ASBR1. The traffic follows Blue paths in

the core. At ASBR4, the MPLS packet gets encapsulated with a UDP header. The UDP header is removed at TOR2, and the lookup will be done for the service label.

6.8. Translating Transport Classes across Domains

```
                 Prefix:PE2        Prefix:PE2  Prefix:PE2
                 RD:RD2            RD:RD2      RD:RD2
                 RT:Red            RT:Blue     RT:Blue
                 nh:ASBR1          nh:ASBR2    nh:PE2
                 Label:L11         Label:L12   Label:L13

          +-----------+                +-----------+
          |           ASBR1           ASBR2        |
          |           |                |           |
       PE1+  AS1      +----------------+    AS2    +PE2
          |           ASBR11          ASBR22       |
          |           |                |           |
          +-----------+                +-----------+

            +------+              +------+
            | IL1  |              | IL2 |
            +------+   +------+   +------+      +------+
            | L11  |   | L12  |   |  L13 |      | L14  |
            +------+   +------+   +------+      +------+
            | S2   |   | S2   |   |  S2  |      | S2   |
            +------+   +------+   +------+      +------+

              Label stacks along end-to-end path
                      S2 is the end-to-end service label.
            IL1 and IL2 are intra-domain labels corresponding to
                            red  intra-domain path in AS1 and Blue intra-domain
                            path in AS2.
```

Figure 12: Translating Transport Classes across Domains

In certain scenarios, the TE intent represented by Transport Classes may differ from one domain to another. This could be the result of two independent organizations merging into one. It could also occur when two ASes are under different administration, but use BGP-CT to provide an end-to-end service. In both scenarios, the same color may represent different intent in each domain. When the traffic needs to satisfy certain TE characteristic, the colors need to be mapped correctly at the border. In the example in Figure 12, there are two ASes. The low latency TE intent is represented with the Red

Transport Class in AS1 and with the Blue Transport Class in AS2. PE2 advertises a BGP-CT prefix with RT of Blue. ASBR2 sets the nexthop to self and advertises a new label L12. On ASBR1, the Blue BGP-CT advertisement is imported into the Red Transport RIB and the advertisement from ASBR1 will carry a Red RT. This ensures that the BGP-CT prefix for PE2 resolves on a Red intra-domain path in AS1. The detailed protocol procedures for this usecase is described in section 10 of [I-D.kaliraj-idr-bgp-classful-transport-planes].

6.9. SLA Guarantee

6.9.1. Low latency

Many network functions are virtualized and distributed. Certain functions are time and latency sensitive. In inter-domain networks, End-to-End latency measurement is required. Inside a domain, latency measurement mechanisms such as TWAMP [RFC5357] are used and link latency is advertised in IGP using extensions described in [RFC8570]and [RFC7471] .

[I-D.ietf-idr-performance-routing] extends the BGP AIGP attribute [RFC7311] by adding a sub TLV to carry an accumulated latency metric. The BGP best path selection algorithm used for a Transport Class requiring low latency will consider the accumulated latency metric to choose the lowest latency path.

6.9.2. Traffic Engineering (TE) constraints

TE constraints generally include the ability to send traffic via certain nodes or links or avoid using certain nodes or links. In the Seamless SR architecture, the intra-domain transport technology is responsible for ensuring the TE constraints inside the domain, BGP-CT ensures that the end-to-end path is constructed from intra-domain paths and inter-AS links that individually satisfy the TE constraints.

For example, in order to construct a pair of diverse paths, we can define a red and a blue Transport Class. Within each domain, the red and blue Transport Class path are realized using intra-domain path diversity mechanisms. For example, in a domain using flex-algo, red and blue Transport Classes are realized using red and blue flex-algo definitions (FAD) which don't share any links. To maintain path diversity on inter-AS links, BGP policies are used to associate two inter-AS peers with the red Transport Class and another two inter-AS peers with the blue Transport Class.

6.9.3. Bandwidth constraints

The Seamless SR architecture does not natively support end-to-end bandwidth reservations. In this architecture, the bandwidth utilization characteristics of each domain are managed independently. The intra-domain bandwidth management can make use of a variety of tools.

Link bandwidth extended community as defined in [I-D.ietf-idr-link-bandwidth] allows for efficient weighted load-balancing of traffic on multiple BGP-CT paths that belong to the same Transport Class. For optimized path placement, a centralized TE system may be deployed with BGP policies/communities used for path placement.

6.10. Scalability

6.10.1. Access node scalability

The Seamless SR architecture needs to be able to accommodate very large numbers of access devices. These access devices are expected to be low-end devices with limited FIB capacity. The Seamless MPLS architecture, as described in [I-D.ietf-mpls-seamless-mpls], recommends the use of LDP DOD mode to limit the size of both the RIB and the FIB needed on the access devices. In the Seamless SR architecture, networks use IGP-based label distribution and do not have this selective label request mechanism. However, RIB scalability of access nodes has not been a problem for real seamless MPLS deployments. In cases where access devices are low on CPU and memory and unable to support large a RIB, BGP filtering policies can be applied at the ABR/ASBR routers to restrict the number of BGP-CT advertisements towards the access devices. The access devices will receive only the PE loopbacks that it needs to connect to.

6.10.1.1. Automating Filtering of BGP-CT Advertisements using Route Target Constraints

When access devices have CPU and memory constraints, it is useful to be able to filter BGP-CT advertisements using policies on border nodes so that only a subset of BGP-CT advertisements are sent to a given access device. While this filtering of BGP-CT advertisements could be done via explicit configuration, it is desirable to have an automated filtering mechanism.

When a service prefix advertisement is received on an access device, the protocol nexthop of the service prefix indicates the remote loopback address from which the service prefix is originated. An access device only needs to receive the subset of BGP-CT

advertisements corresponding to the originators of the service prefixes recieved by that access device. When an access node receives a service prefix with a particular remote loopback address as the protocol nexthop, it can selectively request the BGP-CT advertisement for this particular loopback address from the Route Reflector.

This mechanism is similar to how Route Target Constraints are used to selectively filter VPN advertisements. [RFC4684]. The Route Target Constraint defined in [RFC4684] currently allows for filtering based on Route Target information. Applying a similar mechanism to the filtering of BGP-CT advertisements based on individual loopback addresses requires an extension. The minor protocol enhancements required to achieve this are described in section 11 of [I-D.kaliraj-idr-bgp-classful-transport-planes]

6.10.2. Label stack depth

The ability for a device to push multiple MPLS labels on a packet depends on hardware capabilities. Access devices are expected to have limited label stack push capabilities. Assuming shortest path SR-MPLS in the access domain, the access domain transport will use a single label. Lightweight traffic-engineering and slicing could also be achieved with a single label as described in [I-D.ietf-lsr-flex-algo]. The Seamless SR architecture can provide cross-domain MPLS connectivity with a single label. Assuming the use of a service label, end-to-end connectivity is provided by pushing one service label, one BGP-CT label, and one intra-domain transport label (which could also be a Binding-SID). Therefore, access nodes will only need to be able to push 3 labels for most applications.

6.10.3. Label Resources

```
               -----IBGP----- -----IBGP----- -----IBGP------
              |              |              |              |

                                                         BGP-CT Advt:
                                                         Prefix: 2.2.2.2 (PE2
loopback)
                                                         RD:20000
                                                         RT: 128
                          Label:100       Label:100      Label:101
                          Next hop:ABR3   Next hop:ABR3  Next hop: PE2
        ----------------------------------------------------------------

                                          BGP-CT Advt:
                                          Prefix: 30.30.30.30 (ABR3 loopback)
                                          RD:30000
                                          RT:128
                        Label:2000        Label:2001
                        Nexthop:ABR1      Nexthop:ABR3

               +-----------+   +------------+  +-----------+
              /             \ /              \/             \
              |             ABR1            ABR3            |
              |              |               |              |
           PE1+    Metro1    +     Core      +    Metro2    +PE2
              |              |               |              |
              |             ABR2            ABR4            |
              \              /\             /\              /
               +------------+  +-----------+  +------------+


                 |-ISIS1-|      |-ISIS2-|       |-ISIS3-|

                 +------+        +------+        +------+
                 | 11111|        | 22222|        | 33333|    IGP-labels:
                 +------+        +------+        +------+    11111,22222,33333
                 | 2000 |        | 2001 |        | 101  |    BGP-CT label:
                 +------+        +------+        + -----+    For ABR3:
                 | 100  |        | 100  |        | VPN  |    2000,2001
                 +------+        +------+        +------+    For PE2:
                 | VPN  |        | VPN  |                    100, 101
                 +------+        +------+
```

Figure 13: Recursive Route Resolution

The label resources are an important consideration in MPLS networks. On access devices, labels are consumed by services as well as for transport loopbacks inside IGP domain where the access device resides. For example, in the above diagram PE1 would have to allocate label resources equal to the number of customers connecting (i.e. the number of L2/L3 VPNs). Based on the size of the IGP domain that PE1 resides in, it will also have to allocate labels for IGP loopbacks. This number is at most a few thousands. So overall a typical access device should have adequate label resources in Seamless SR architecture. The P routers need to allocate labels for IGP loopbacks. This number again is small. At most it will be a few thousand based on number of nodes in the largest IGP domains. The metro networks connect to the core network through ABRs. It is possible that a given ABR may end up having to maintain forwarding entries for a large subset of the transport loopback routes. There may be a large number of metro networks connecting to a given ABR, and in this case, the ABR will need forwarding entries for every access node in the directly connected metros. So, this ABR may have to maintain on the order of 100k routes. With BGP-CT each Transport Class will have to be separately allocated a label. So, in the above example, the ABR1 would have to use 300k labels if there were 3 Transport Classes. This large number of label forwarding entries could be problematic.

In highly scaled scenarios, it is therefore desirable to reduce the forwarding state on the ABRs. This reduction can be achieved with label stacking as a result of recursive route resolution. Figure 13 illustrates how the forwarding state on ABRs can be greatly reduced by removing forward state for PEs in remote domains from the ABRs. In this example, we assume that we are setting up end-to-end paths for a single Transport Class, for example red. PE2 advertises a BGP-CT prefix of 2.2.2.2 with nexthop of 2.2.2.2 and label 101. 2.2.2.2 is PE2's loopback. ABR3 advertises label 100 for BGP-CT prefix 2.2.2.2 and changes the nexthop to self. When ABR1 receives the BGP-CT advertisement for 2.2.2.2, it does not change the nexthop and advertises same label advertised by ABR3. When PE1 receives the BGP-CT advertisement for 2.2.2.2 with a nexthop of ABR3, it resolves the route using reachability to ABR3.

The reachability of ABR3 has been learned by PE1 as the result of a BGP-CT advertisement originated by ABR3. As shown in Figure 13, ABR3 advertises BGP-CT prefix 30.30.30.30 with label 2001. ABR1 advertises label 2000 for BGP-CT prefix 30.30.30.30 and sets nexthop to self. PE1 constructs the service data packet with a VPN label at the bottom followed by 2 BGP-CT labels 100 and 2000. The top most label 2000 is the transport label for the metro1 domain. Removing the forwarding state for PEs in remote domains on the ABRs comes at the expense of one additional BGP-CT label on the data packet.

Recursive route resolution provides significant forwarding state reduction on the ABRs. ABRs have to allocate label resources only for the PEs in their local domain. The number of PEs in the same domain as a given ABR is much lower than the total number of PEs in the network.

The examples in this draft generally show VPN routes resolving on BGP-CT prefixes. However, the mechanisms are equally applicable to non-VPN routes.

6.11. Availability

Transport layer availability is very important in latency and loss sensitive networks. Any link or node failure must be repaired with 50ms convergence time. 50 ms convergence time can be achieved with Fast ReRoute (FRR) mechanisms. The seamless SR architecture provides protection against intra-domain link and node failures, Protection against border node failures and the egress link and node failures are also provided. Details of the FRR techniques are described in the sections below.

6.11.1. Intra domain link and node protection

In the seamless SR architecture, protection against node and link failure is achieved with the relevant FRR techniques for the corresponding transport mechanism used inside the domain. In the case of an IP fabric, ECMP FRR or LFA can be used. In SR networks, TI-LFA [I-D.ietf-rtgwg-segment-routing-ti-lfa] provides link and node protection. For SR-TE transport ([I-D.ietf-spring-segment-routing-policy]), link and node protection can be achieved using TI-LFA, combined with mechanisms described in [I-D.hegde-spring-node-protection-for-sr-te-paths].

6.11.2. Egress link and node protection

[RFC8679] describes the mechanisms for providing protection for border nodes and PE devices where services are hosted. The mechanism can be further simplified operationally with anycast SIDs and anycast service labels, as described in [I-D.hegde-rtgwg-egress-protection-sr-networks].

6.11.3. Border Node protection

Border node protection is very important in a network consisting of multiple domains. Seamless SR architecture can achieve 50ms FRR protection in the event of node failure using anycast addresses for the ABR/ASBRs. The requires that a set of ABRs advertise the same

label for a given BGP-CT Prefix. The detailed mechanism is described in [I-D.hegde-rtgwg-egress-protection-sr-networks].

6.12. Operations

6.12.1. MPLS ping and Traceroute

The Seamless SR Architecture consists of 3 layers: the service layer, intra-domain transport, and BGP-CT transport. Within each layer, connectivity can be verified independently. Within the BGP-CT transport layer, end-to-end connectivity can be verified using a new OAM FEC for BGP-CT defined in draft [I-D.kaliraj-idr-bgp-classful-transport-planes]. The draft describes end-to-end connectivity verification as well as fault isolation. BGP-CT verification happens only on the BGP nodes. The intra-domain connectivity verification and fault isolation will be based on the technology deployed in that domain as defined in [RFC8029] and [RFC8287].

6.12.2. Counters and Statistics

Traffic accounting and the ability to build demand matrix for PE to PE traffic is very important. With BGP-CT, per-label transit counters should be supported on every transit router. Per-label transit counters provide details of total traffic towards a remote PE measured at every BGP transit router. Per-label egress counters should be supported on ingress PE router. Per-label egress counters provide total traffic from ingress PE to the specific remote PE.

6.13. Service Mapping

Service mapping is an important aspect of any architecture. It provides means to translate end users SLA requirements into operator's network configurations. Seamless SR architecture supports automatic steering with extended color community. The Transport Class and the route target carried by the BGP-CT advertisement directly map to the extended color community. Services that require specific SLA carry the extended color community which maps to the Transport Class to which the BGP-CT advertisement belongs.

Other types of traffic steering such as DSCP based forwarding is expressed with mapping-community. Mapping community is a standard BGP community and is completely generic and user defined. The mapping community will have a specific service mapping feature associated with it along with required fallback behaviour when the primary transport goes down. The below list provides a general guideline into the different service mapping features and fallback options an implementation should provide.

DSCP based mapping with each DSCP mapping to a Transport Class.

DSCP based mapping with default mapping to a best-effort transport

DSCP based mapping with fallback to best-effort when primary transport tunnel goes down.

Extended color community based mapping with fallback to best effort

Fallback options with specific protocol during migrations

Fallback options to a different Transport Class.

No Fallback permitted.

6.14. Migrations

Networks that migrate from Seamless MPLS architecture to Seamless SR architecture, require that all the border nodes and PE devices be upgraded and enabled with new family on the BGP session. In cases where legacy nodes that cannot be upgraded, exporting from BGP-LU into BGP-CT and vice versa SHOULD be supported. Once the entire network is migrated to support BGP-CT, there is no need to run BGP-LU family on the BGP sessions. BGP-CT itself can advertise a best effort Transport Class and BGP-LU family can be removed.

6.15. SRv6 interworking with MPLS domains

SRv6 defines the Segment Routing architecture for IPv6 data plane with a new extension header as described in [RFC8402]. As described in Section 3.3 of the current document, data center and access/ aggregation networks may deploy SRv6 and connect to the WAN networks. Since current WAN networks predominantly use MPLS, it is important to provide solutions that interconnect SRv6 and MPLS domains. The seamless SR architecture supports interconnecting domains that deploy SRv6 and MPLS.

The SRv6 Network Programming draft [I-D.ietf-spring-srv6-network-programming] defines an SRv6 SID as consisting of locator, function, and argument bits. The locator part of the SRv6 SID is routable, and the route leads to the node that instantiates the SID. The seamless SR architecture builds on this concept to enable interworking between SRv6 and other domains. In the Seamless SR architecture, different domains are loosely coupled, and prefixes are not leaked from the IGP in one domain into the IGP of another domain. BGP is used to stitch the different domains together and build an end-to-end path. In SRv6, a seperate locator

is allocated for each color. The service SIDs that need to use the particular colored path will be derived based on corresponding locator. Locators are IPv6 prefixes of length less than 128 bits. These locators are advertised in BGP in AFI 2/ SAFI 1 family (IPv6 unicast). BGP will install these locator routes on each border node, so each border node will have reachability for the SRv6 SIDs. In order to transparently traverse an MPLS domain, the SRv6 traffic is encapsulated with MPLS headers at the ingress MPLS border node and decapsulated at the egress MPLS border node. The association of the SRv6 locator with a particular color is also carried in the IPv6 unicast advertisement so that specific transport class paths can be used when desired. This is illustrated in the following example.

```
        Locator for Red Transport Class : 5:6::/96
        Locator for Blue Transport Class: 5:7::/96



          BGP AFI2/SAFI 1 advertisements for Red transport class

       Pfx:5:6::/96    Pfx:5:6::/96  Pfx:5:6::/96   Pfx:5:6::/96 Pfx:5:6::/96
       Ext-Com: Red    Ext-Com:Red   Ext-Com:Red    Ext-Com:Red  Ext-Com:Red
       nh:ASBR1        nh:ASBR2      nh:ASBR3       nh:ASBR4     nh:PE2


PE1------------ASBR1-----------ASBR2---------ASBR3-------ASBR4--------PE2
|               |                |             |            |           |
 ------SRv6------                 -----MPLS-----            ----SRv6-----

                                                              VPNa Prefix:
                                                              10.1.1.0/24
                                                              RD: RD50
                                                              RT: RT-VPNa
                                                              ext-community:
                                                              Red(100)
                                                              nh: PE2
                                                              END.DT4 SID:
5:6::16/128
                            +-----------+
                            |     IL1   |
                            +-----------+
                            |     IL2   |
+---------+  +------------+ +-----------+         +-----------+
|src:PE1  |  | src:PE1    | |src:PE1    |         |src:PE1    |
|dst:ASBR1|  | dst:5:6::16| |dst:5:6::16|         |dst:5:6::16|
|SRH: SL=1|  |SRH: SL = 0 | |SRH: SL=0  |         |SRH: SL=0  |
|5:6::16  |  |5:6::16     | |5:6::16    |         |5:6::16    |
+---------+  +------------+ +-----------+         +-----------+     +----------
+
| orig    |  | orig       | |  Orig     |         |   Orig    |     |    Orig
|
+---------+  +------------+ +-----------+         +-----------+     +----------
+

                      Packet format along end-to-end path
                      Orig is the original packet destined to 10.1.1.1
            IL1, IL2,  intra-domain labels corresponding to
                            red intra-domain paths in MPLS domain.
```

Figure 14: SRv6 and MPLS interworking

In the diagram above Figure 14 describes an example where the core is

MPLS domain and the datacenters deploy SRv6. In the example above, an end-to-end path is built for the Red transport class. The SRv6 domains in this example use best effort paths. On PE2, locator

5:6::/96 represents the Red transport class. PE2 would like for traffic for service prefix 10.1.1.0/24 to use a Red tranport class path. To accomplish this PE2 creates two BGP advertisements, a VPN advertisement and an IPv6 unicast advertisement.

PE2 creates a VPN advertisement using an END.DT4 SID derived from its Red locator 5:6::/96.(END.DT4 SID = 5:6::16/128 in this example.) The VPN advertisement also associates the Red extended color community with the service prefix 10.1.1.0/24.

PE2 also creates a IPv6 unicast BGP advertisement that associates the IPv6 prefix of the Red locator (5:6::/96) with the Red extended community. This advertisement allows PE1 as well as the ASBRs to have routes for 5:6::/96, and to associate those routes with the Red transport class where appropriate.

The routes that make up the end-to-end path from PE1 to PE2 are described below. On PE1, the VPN prefix 10.1.1.0/24. will resolve on the locator prefix 5:6::/96. The prefix 5:6::/96 will then resolve on an SRv6/IPv6 tunnel to ASBR1. ASBR1 will have a normal IPv6 route for 5:6::/96 installed by BGP to reach ASBR2. On ASBR2, the prefix 5:6::/96 will resolve on an MPLS tunnel belonging to Red transport class terminating on ASBR3. The route for 5:6::/96 from ASBR3 to ASBR4 is again a simple IPv6 route installed by BGP.On ASBR4, both BGP and the IGP will provide a route for 5:6::/96. In general, the active route will be derived from the IGP which will normally be preferred. In cases where a traffic engineered path is needed in the last SRv6 domain, the preference needs to be set appropriately by the administrator.

Below is a description of packet forwarding operations along the end-to-end path. On PE1, the original packet destined to 10.1.1.1 will get encapsulated in IPv6 header with one segment END.DT4SID. The destination address is set to ASBR1. On ASBR1, segment left is decremented and the END.DT4 sid 5:6::16 is copied into destination address. On ASBR1, forwarding will be based on the locator route programmed by BGP. Between ASBR1 and ASBR2, it is normal ipv6 forwarding. On ASBR2, an MPLS header corresponding to Red transport Class is pushed on the packet. The MPLS header gets removed when packet reaches ASBR3 and normal ipv6 forwarding based on the locator route is performed. On ASBR4, since best effort path for locator 5:6::/96 is used which is created by IGP, normal IPv6 forwarding is used. The packet reaches PE2 with 5:6::16 as the destination which is present in MyLocalSID table. IPv6 header is decapsulated and lookup for 10.1.1.1 is performed in the VPN table.

The example described above has complete domain seperation where SRv6 operations end on one border nodeand MPLS header operations are

performed on next border node. There may be cases where the a single border node needs to perform both SRv6 and MPLS operations. A goal for the Seamless SR architecture is to avoid service routes on border nodes and provide seamless end-to-end connectivity for the services. In order to satisfy this goal for the single border node use case, a new SID type is defined. The END.DTM SID decapsulates the IPv6 header and pushes an MPLS SID List. It is used to determine the MPLS labels for traffic flowing from a SRv6 domain to an MPLS domain. [draft-bonica-spring-srv6-end-dtm] provides details of this new SID and its operation in detail.

6.16. Service Function Chaining

Service Function Chaining involves steering traffic through an ordered set of service functions. Virtualized service functions may be deployed in a single Data Center location or across multiple Data Centers which are geographically separated. There are several different service function chaining solutions available. One set of solutions uses the source routing paradigm as described in [I-D.ietf-spring-sr-service-programming]. The source routing based solution may use SR-MPLS or SRv6 as described in above draft. Another set of solutions uses stitched tunnels to achieve the traffic steering through service functions. The tunneling technology can be MPLS tunneling or IP tunnelling. This set of solutions is described in [draft-hegde-spring-service-chaining-stitched-tunnel]. When a network deploys Seamless SR-based inter-domain solutions, it can deploy either of these solutions for service chaining. This section describes how service chaining is applied in a network that uses Seamless SR for inter-domain connectivity. For simplicity, the example below assumes service functions deployed in a single Data Center. The procedures are equally applicable when the service functions are spread across multiple geographically separated Data Centers.

```
           -----------------------------   --------------------
           |   ---                      | |                   |
           |  | S1|    TOR1             | |                   |  Z
           |   ---             SP1    DCGW1                   | /
           |   ---                      | |       WAN        PE2
           |  | S2|    TOR2             | |                   |
           |   ---             SP2    DCGW2                   |
           |   ---                      | |                   |
           |  | S3|    TOR3             | |                   |
           |   ---                      | |                   |
           |-----------------------------  -------------------
                           BGP-CT             BGP-CT
                       |----------------|---------------------|
```

Figure 15: SFC in a seamless SR based network

Figure 15 shows a Data Center (DC) network connected to a WAN network. We assume the traffic is originating at S1 in the DC network and destined for Z in the WAN network. The traffic should go through service functions deployed on S2 and S3. The DCGW1 and DCGW2 are the border nodes between the DC domain and WAN domain. BGP-CT is deployed to provide seamless end-to-end connectivity. We also assume that DC network deploys a pure IP underlay, and that the WAN uses an MPLS underlay. BGP-CT is deployed on the Top-of-Rack switches (TORs), and BGP-CT sessions are running from the TORs to the DCGWs, and from the DCGWs to PE2. All the BGP-CT speakers will have an SLA-specific forwarding entries to reach PE2.

When source routed SFC is used [I-D.ietf-spring-sr-service-programming], a packet originating at S1 will use an SR-MPLS or SRv6 SID-list to achieve service function chaining. In this example, the packet will have a SID-list corresponding to the service functions on S2 and S3. The SFC SID-list gets removed by the time the packets leaves S3. The packet arrives at TOR3 with its original IP header exposed. On TOR3 a lookup is done for destination Z. The packet follows SLA-specific BGP-CT paths in both the DC and the WAN.

When the stitched tunnel mechanism is used for service chaining [draft-hegde-spring-service-chaining-stitched-tunnel], it is typical for an an overlay orchestrator to build the tunnels in the DC fabric for the S1->S2 and S2->S3. The overlay orchestrator also provisions the appropriate firewall filters to steer the traffic across these stitched tunnels. When the packet arrives at S3, all service functions have been applied and a lookup on the original IP header is done. In the case, the packet also follows SLA-specific BGP-CT paths in both the DC and the WAN.

6.17. BGP based Multicast

BGP based multicast as described in draft [I-D.zzhang-bess-bgp-multicast] serves two main purposes. It can replace PIM/ mLDP inside a domain to natively do a BGP based multicast. It can also serve as an overlay stitching protocol to stitch multiple P2MP LSPs across the domain. This gives the ability to easily transition each domain independently from one technology to the other. BGP based multicast defines a new SAFI for carrying the MULTICAST TREE SAFI. Different route types are defined to support the various usecases. section 1.2.6 of [I-D.zzhang-bess-bgp-multicast] describes the use of new SAFI for stitching the multicast tunnels across different domains.

7. Backward Compatibility

8. Security Considerations

TBD

9. IANA Considerations

10. Acknowledgements

Many thanks to Kireeti Kompella, Ron Bonica, Krzysztof Szarcowitz, Srihari Sangli,Julian Lucek, Ram Santhanakrishnan for discussions and inputs. Thanks to Joel Halpern for review and comments.

11. Contributors

1.Kaliraj Vairavakkalai

Juniper Networks

kaliraj@juniper.net

2. Jeffrey Zhang

Juniper Networks

zzhang@juniper.net

12. References

12.1. Normative References

[I-D.hegde-rtgwg-egress-protection-sr-networks]
Hegde, S., Lin, W., and S. Peng, "Egress Protection for Segment Routing (SR) networks", draft-hegde-rtgwg-egress-protection-sr-networks-01 (work in progress), November 2020.

[I-D.ietf-idr-performance-routing]
Xu, X., Hegde, S., Talaulikar, K., Boucadair, M., and C. Jacquenet, "Performance-based BGP Routing Mechanism", draft-ietf-idr-performance-routing-03 (work in progress), December 2020.

[I-D.kaliraj-idr-bgp-classful-transport-planes]
Vairavakkalai, K., Venkataraman, N., Rajagopalan, B., Mishra, G., Khaddam, M., and X. Xu, "BGP Classful Transport Planes", draft-kaliraj-idr-bgp-classful-transport-planes-06 (work in progress), January 2021.

[I-D.zzhang-bess-bgp-multicast]
Zhang, Z., Giuliano, L., Patel, K., Wijnands, I., mishra, m., and A. Gulko, "BGP Based Multicast", draft-zzhang-bess-bgp-multicast-03 (work in progress), October 2019.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, DOI 10.17487/RFC3107, May 2001, <https://www.rfc-editor.org/info/rfc3107>.

[RFC8669] Previdi, S., Filsfils, C., Lindem, A., Ed., Sreekantiah, A., and H. Gredler, "Segment Routing Prefix Segment Identifier Extensions for BGP", RFC 8669, DOI 10.17487/RFC8669, December 2019, <https://www.rfc-editor.org/info/rfc8669>.

12.2. Informative References

[I-D.hegde-spring-node-protection-for-sr-te-paths]
Hegde, S., Bowers, C., Litkowski, S., Xu, X., and F. Xu, "Node Protection for SR-TE Paths", draft-hegde-spring-node-protection-for-sr-te-paths-07 (work in progress), July 2020.

[I-D.ietf-idr-link-bandwidth]
Mohapatra, P. and R. Fernando, "BGP Link Bandwidth Extended Community", draft-ietf-idr-link-bandwidth-07 (work in progress), March 2018.

[I-D.ietf-idr-segment-routing-te-policy]
Previdi, S., Filsfils, C., Talaulikar, K., Mattes, P., Rosen, E., Jain, D., and S. Lin, "Advertising Segment Routing Policies in BGP", draft-ietf-idr-segment-routing-te-policy-11 (work in progress), November 2020.

[I-D.ietf-idr-tunnel-encaps]
Patel, K., Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-20 (work in progress), November 2020.

[I-D.ietf-lsr-flex-algo]
Psenak, P., Hegde, S., Filsfils, C., Talaulikar, K., and A. Gulko, "IGP Flexible Algorithm", draft-ietf-lsr-flex-algo-13 (work in progress), October 2020.

[I-D.ietf-mpls-seamless-mpls]
Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz, M., and D. Steinberg, "Seamless MPLS Architecture", draft-ietf-mpls-seamless-mpls-07 (work in progress), June 2014.

[I-D.ietf-pce-segment-routing-policy-cp]
Koldychev, M., Sivabalan, S., Barth, C., Peng, S., and H. Bidgoli, "PCEP extension to support Segment Routing Policy Candidate Paths", draft-ietf-pce-segment-routing-policy-cp-01 (work in progress), October 2020.

[I-D.ietf-rtgwg-segment-routing-ti-lfa]
Litkowski, S., Bashandy, A., Filsfils, C., Decraene, B., and D. Voyer, "Topology Independent Fast Reroute using Segment Routing", draft-ietf-rtgwg-segment-routing-ti-lfa-05 (work in progress), November 2020.

[I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", draft-ietf-spring-segment-routing-policy-09 (work in progress), November 2020.

[I-D.ietf-spring-sr-service-programming]
Clad, F., Xu, X., Filsfils, C., daniel.bernier@bell.ca, d., Li, C., Decraene, B., Ma, S., Yadlapalli, C., Henderickx, W., and S. Salsano, "Service Programming with Segment Routing", draft-ietf-spring-sr-service-programming-03 (work in progress), September 2020.

[I-D.ietf-spring-srv6-network-programming]
Filsfils, C., Camarillo, P., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "SRv6 Network Programming", draft-ietf-spring-srv6-network-programming-28 (work in progress), December 2020.

[I-D.saad-sr-fa-link]
Saad, T., Beeram, V., Barth, C., and S. Sivabalan, "Segment-Routing over Forwarding Adjacency Links", draft-saad-sr-fa-link-02 (work in progress), July 2020.

[I-D.voyer-pim-sr-p2mp-policy]
Voyer, D., Filsfils, C., Parekh, R., Bidgoli, H., and Z. Zhang, "Segment Routing Point-to-Multipoint Policy", draft-voyer-pim-sr-p2mp-policy-02 (work in progress), July 2020.

[RFC1997] Chandra, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, DOI 10.17487/RFC1997, August 1996, <https://www.rfc-editor.org/info/rfc1997>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <https://www.rfc-editor.org/info/rfc4684>.

[RFC5357] Hedayat, K., Krzanowski, R., Morton, A., Yum, K., and J. Babiarz, "A Two-Way Active Measurement Protocol (TWAMP)", RFC 5357, DOI 10.17487/RFC5357, October 2008, <https://www.rfc-editor.org/info/rfc5357>.

[RFC6388] Wijnands, IJ., Ed., Minei, I., Ed., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, DOI 10.17487/RFC6388, November 2011, <https://www.rfc-editor.org/info/rfc6388>.

[RFC7311] Mohapatra, P., Fernando, R., Rosen, E., and J. Uttaro, "The Accumulated IGP Metric Attribute for BGP", RFC 7311, DOI 10.17487/RFC7311, August 2014, <https://www.rfc-editor.org/info/rfc7311>.

[RFC7471] Giacalone, S., Ward, D., Drake, J., Atlas, A., and S. Previdi, "OSPF Traffic Engineering (TE) Metric Extensions", RFC 7471, DOI 10.17487/RFC7471, March 2015, <https://www.rfc-editor.org/info/rfc7471>.

[RFC7510] Xu, X., Sheth, N., Yong, L., Callon, R., and D. Black, "Encapsulating MPLS in UDP", RFC 7510, DOI 10.17487/RFC7510, April 2015, <https://www.rfc-editor.org/info/rfc7510>.

[RFC7665] Halpern, J., Ed. and C. Pignataro, Ed., "Service Function Chaining (SFC) Architecture", RFC 7665, DOI 10.17487/RFC7665, October 2015, <https://www.rfc-editor.org/info/rfc7665>.

[RFC8029] Kompella, K., Swallow, G., Pignataro, C., Ed., Kumar, N., Aldrin, S., and M. Chen, "Detecting Multiprotocol Label Switched (MPLS) Data-Plane Failures", RFC 8029, DOI 10.17487/RFC8029, March 2017, <https://www.rfc-editor.org/info/rfc8029>.

[RFC8287] Kumar, N., Ed., Pignataro, C., Ed., Swallow, G., Akiya, N., Kini, S., and M. Chen, "Label Switched Path (LSP) Ping/Traceroute for Segment Routing (SR) IGP-Prefix and IGP-Adjacency Segment Identifiers (SIDs) with MPLS Data Planes", RFC 8287, DOI 10.17487/RFC8287, December 2017, <https://www.rfc-editor.org/info/rfc8287>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

[RFC8570] Ginsberg, L., Ed., Previdi, S., Ed., Giacalone, S., Ward, D., Drake, J., and Q. Wu, "IS-IS Traffic Engineering (TE) Metric Extensions", RFC 8570, DOI 10.17487/RFC8570, March 2019, <https://www.rfc-editor.org/info/rfc8570>.

[RFC8604] Filsfils, C., Ed., Previdi, S., Dawra, G., Ed., Henderickx, W., and D. Cooper, "Interconnecting Millions of Endpoints with Segment Routing", RFC 8604, DOI 10.17487/RFC8604, June 2019, <https://www.rfc-editor.org/info/rfc8604>.

[RFC8679] Shen, Y., Jeganathan, M., Decraene, B., Gredler, H., Michel, C., and H. Chen, "MPLS Egress Protection Framework", RFC 8679, DOI 10.17487/RFC8679, December 2019, <https://www.rfc-editor.org/info/rfc8679>.

[TS.23.501-3GPP] 3rd Generation Partnership Project (3GPP), "System Architecture for 5G System; Stage 2, 3GPP TS 23.501 v16.4.0", March 2020.

Authors' Addresses

Shraddha Hegde
Juniper Networks Inc.
Exora Business Park
Bangalore, KA 560103
India

Email: shraddha@juniper.net

Chris Bowers
Juniper Networks Inc.

Email: cbowers@juniper.net

Xiaohu Xu
Alibaba Inc.
Beijing
China

Email: xiaohu.xxh@alibaba-inc.com

Arkadiy Gulko
Refinitiv

Email: arkadiy.gulko@refinitiv.com

Alex Bogdanov
Google Inc.

Email: bogdanov@google.com

James Uttaro
ATT

Email: ju1738@att.com

Luay Jalil
Verizon

Email: luay.jalil@verizon.com

Mazen Khaddam
Cox communications

Email: mazen.khaddam@cox.com

Andrew Alston
Liquid Telecom

Email: andrew.alston@liquidtelecom.com