SPRING Working Group
Internet-Draft
Intended status: Informational
Expires: November 20, 2021

A. Farrel
Old Dog Consulting
J. Drake
Juniper Networks

May 19, 2021

Interconnection of Segment Routing Sites - Problem Statement and Solution Landscape
draft-farrel-spring-sr-domain-interconnect-06

Abstract

Segment Routing (SR) is a forwarding paradigm for use in MPLS and IPv6 networks. It is intended to be deployed in discrete sites that may be data centers, access networks, or other networks that are under the control of a single operator and that can easily be upgraded to support this new technology.

Traffic originating in one SR site often terminates in another SR site, but must transit a backbone network that provides interconnection between those sites.

This document describes a mechanism for providing connectivity between SR sites to enable end-to-end or site-to-site traffic engineering.

The approach described allows connectivity between SR sites, utilizes traffic engineering mechanisms (such as RSVP-TE or Segment Routing) across the backbone network, makes heavy use of pre-existing technologies, and requires the specification of very few additional mechanisms.

This document provides some background and a problem statement, explains the solution mechanism, gives references to other documents that define protocol mechanisms, and provides examples. It does not define any new protocol mechanisms.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on November 20, 2021.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (https://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction 3
1.1. Terminology 4
2. Problem Statement 4
3. Solution Technologies 7
3.1. Characteristics of Solution Technologies 7
4. Decomposing the Problem 9
5. Solution Space 10
5.1. Global Optimization of the Paths 10
5.2. Figuring Out the GWs at a Destination Site for a Given Prefix 11
5.3. Figuring Out the Backbone Egress ASBRs 12
5.4. Making use of RSVP-TE LSPs Across the Backbone 12
5.5. Data Plane 13
5.6. Centralized and Distributed Controllers 15
6. BGP-LS Considerations 18
7. Worked Examples 21
8. Label Stack Depth Considerations 25
8.1. Worked Example 26
9. Gateway Considerations 27
9.1. Site Gateway Auto-Discovery 27
9.2. Relationship to BGP Link State and Egress Peer Engineering 28
9.3. Advertising a Site Route Externally 28
9.4. Encapsulations 29

10. Security Considerations 29
11. Management Considerations 30
12. IANA Considerations 30
13. Acknowledgements 30
14. Informative References 30
Authors' Addresses 34

1. Introduction

Data Centers are a growing market sector. They are being set up by new specialist companies, by enterprises for their own use, by legacy ISPs, and by the new wave of network operators. The networks inside Data Centers are currently well-planned, but the traffic loads can be unpredictable. There is a need to be able to direct traffic within a Data Center to follow a specific path.

Data Centers are attached to external ("backbone") networks to allow access by users and to facilitate communication among Data Centers. An individual Data Center may be attached to multiple backbone networks, and may have multiple points of attachment to each backbone network. Traffic to or from a Data Center may need to be directed to or from any of these points of attachment.

Segment Routing (SR) is a technology that places forwarding state into each packet as a stack of loose hops. SR is an option for building Data Centers, and is also seeing increasing traction in edge and access networks as well as in backbone networks. It is typically deployed in discrete sites that are under the control of a single operator and that can easily be upgraded to support this new technology.

Traffic originating in one SR site often terminates in another SR site, but must transit a backbone network that provides interconnection between those sites. This document describes an approach that builds on existing technologies to produce mechanisms that provide scalable and flexible interconnection of SR site, and that will be easy to operate.

The approach described allows end-to-end connectivity between SR sites across an MPLS backbone network, utilizes traffic engineering mechanisms (such as RSVP-TE or Segment Routing) across the backbone network, makes heavy use of pre-existing technologies, and requires the specification of very few additional mechanisms.

This document provides some background and a problem statement, explains the solution mechanism, gives references to other documents that define protocol mechanisms, and provides examples. It does not define any new protocol mechanisms.

1.1. Terminology

This document uses Segment Routing terminology from [RFC7855] and [RFC8402]. Particular abbreviations of note are:

- SID: a segment identifier
- SRGB: an SR Global Block

In the context of this document, the terms "optimal" and "optimality" refer to making the best possible use of network resources, and achieving network paths that best meet the objectives of the network operators and customers.

Further terms are defined in Section 2.

2. Problem Statement

Consider the network in Figure 1. Without loss of generality, this figure can be used to represent the architecture and problem space for steering traffic within and between SR edge sites. The figure shows a single destination for all traffic that we will consider.

In describing the problem space and the solution we use six terms as follows:

SR domain : This term is defined in [RFC8402]. It is the collection of all interconnected SR-capable network nodes that may be colocated in a site, distributed across multiple sites, present in SR-capable backbone networks, or located at key points within the backbone network.

SR site : In this document, an SR site is a collection of SR-capable nodes under the care of one administrator or protocol. This means that each SR site is attached to the backbone network through one or more gateways. Examples include, access networks, Data Center sites, backbone networks that run SR, and blessings of unicorns.

Host : A node within an SR site. It may be an end system or a transit node in the SR site.

Gateway (GW) : Provides access to or from an SR site. Examples are Customer Edge nodes (CEs), Autonomous System Border Routers (ASBRs), and Data Center gateways.

Provider Edge (PE) : Provides access to or from the backbone network.

Autonomous System Border Router (ASBR) : Provides access to one Autonomous System (AS) in the backbone network from another AS in the backbone network.

These terms can be seen in use in Figure 1, where the various sources and the destination are hosts. In this figure we distinguish between the PEs that provide access to the backbone network, and the Gateways that provide access to the SR sites: these may, in fact, be the same equipment and the PEs might be located at the site edges.

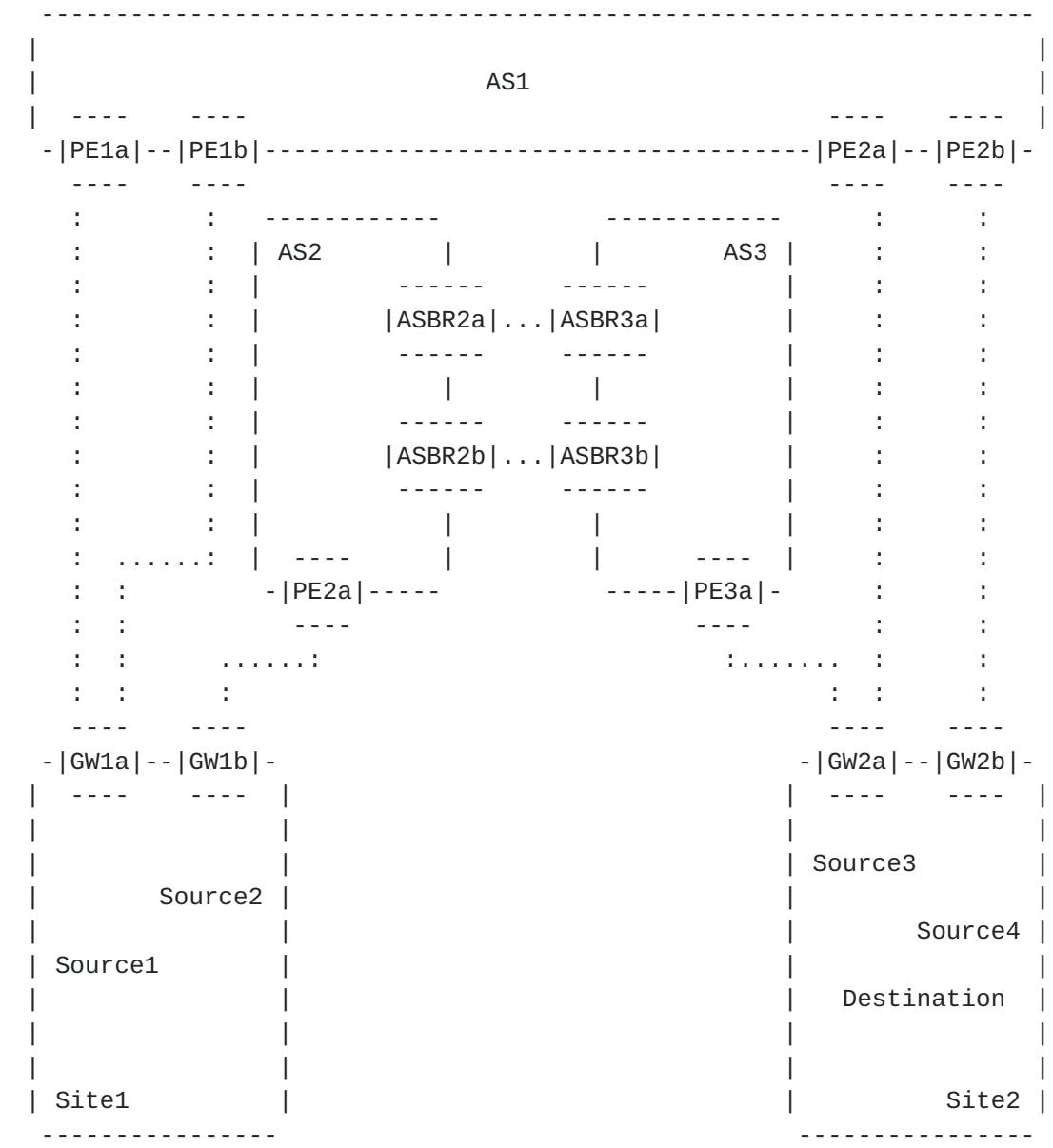


Figure 1: Reference Architecture for SR Site Interconnect

Traffic to the destination may originate from multiple sources within that site (we show two such sources: Source3 and Source4). Furthermore, traffic intended for the destination may arrive from outside the site through any of the points of attachment to the backbone networks (we show GW2a and GW2b). This traffic may need to be steered within the site to achieve load-balancing across network resources, to avoid degraded or out-of-service resources (including planned service outages), and to achieve different qualities of service. Of course, traffic in a remote source site may also need to be steered within that site. We class this problem as "Intra-Site Traffic Steering".

Traffic across the backbone networks may need to be steered to conform to common Traffic Engineering (TE) paradigms. That is, the path across any network (shown in the figure as an AS) or across any collection of networks may need to be chosen and may be different from the shortest path first (SPF) routing that would occur without TE. Furthermore, the points of inter-connection between networks may need to be selected and influence the path chosen for the data. We class this problem as "Inter-Site Traffic Steering".

The composite end-to-end path comprises steering in the source site, choice of source site exit point, steering across the backbone networks, choice of network interconnections, choice of destination site entry point, and steering in the destination site. These issues may be inter-dependent (for example, the best traffic steering in the source site may help select the best exit point from that site, but the connectivity options across the backbone network may drive the selection of a different exit point). We class this combination of problems as "End-to-End Site Interconnect Traffic Steering".

It should be noted that the solution to the End-to-End Site Interconnect Traffic Steering problem depends on a number of factors:

- What technology is deployed in the site.
- What technology is deployed in the backbone networks.
- How much information the sites are willing to share with each other.
- How much information the backbone network operators and the site operators are willing to share.

In some cases, the sites and backbone networks are all owned and operated by the same company (with the backbone network often being a private network). In other cases, the sites are operated by one company, with other companies operating the backbone.

3. Solution Technologies

Segment Routing (SR from the SPRING working group in the IETF [RFC7855] and [RFC8402]) introduces traffic steering capabilities into an MPLS network [RFC8660] by utilizing existing data plane capabilities (label pop and packet forwarding - "pop and go") in combination with additions to existing IGPs ([RFC8665] and [RFC8667]), BGP (as BGP-LU) [RFC8277], or a centralized controller to distribute "per-hop" labels. An MPLS label stack can be imposed on a packet to describe a sequence of links/nodes to be transited by the packet; as each hop is transited, the label that represents it is popped from the stack and the packet is forwarded. Thus, on a packet-by-packet basis, traffic can be steered within the SR domain.

This document broadens the problem space to consider interconnection of any type of site. These may be Data Center sites, but they may equally be access networks, VPN sites, or any other form of domain that includes packet sources and destinations. We particularly focus on "SR sites" being source or destination sites that utilize MPLS SR, but the sites could use other non-MPLS technologies (such as IP, VXLAN, and NVGRE) as described in Section 9.

Backbone networks are commonly based on MPLS-capable hardware. In these networks, a number of different options exist to establish TE paths. Among these options are static Label Switched Paths (LSPs), perhaps set up by an SDN controller, LSP tunnels established using a signaling protocol (such as RSVP-TE), and inter-site use of SR (as described above for intra-site steering). Where traffic steering (without resource reservation) is needed, SR may be adequate; where Traffic Engineering is needed (i.e., traffic steering with resource reservation) RSVP-TE or centralized SDN control are preferred. However, in a network that is fully managed and controlled through a centralized planning tool, resource reservation can be achieved and SR can be used for full Traffic Engineering. These solutions are already used in support of a number of edge-to-edge services such as L3VPN and L2VPN.

3.1. Characteristics of Solution Technologies

Each of the solution technologies mentioned in the previous section has certain characteristics, and the combined solution needs to recognize and address these characteristics in order to make a workable solution.

- When SR is used for traffic steering, the size of the MPLS label stack used in SR scales linearly with the length of the strict source route. This can cause issues with MPLS implementations that only support label stacks of a limited size. For example,

some MPLS implementations cannot push enough labels on the stack to represent an entire source route. Other implementations may be unable to do the proper "ECMP hashing" if the label stack is too long; they may be unable to read enough of the packet header to find an entropy label or to find the IP header of the payload. Increasing the packet header size also reduces the size of the payload that can be carried in an MPLS packet. There are techniques that can be used to reduce the size of the label stack. For example, a source route may be made less specific through the use of loose hops requiring fewer labels, or a single label (known as a "binding SID") can be used to represent a sequence of nodes; this label can be replaced with a set of labels when the packet reaches the first node in the sequence. It is also possible to combine SR with conventional RSVP-TE by using a binding SID in the label stack to represent an LSP tunnel set up by RSVP-TE.

- Most of the work on using SR for traffic steering assumes that traffic only needs to be steered within a single administrative domain. If the backbone consists of multiple ASes that are not part of a common administrative domain, the use of SR across the backbone may prove to be a challenge, and its use in the backbone may be limited to cases where private networks connect the sites, rather than cases where the sites are connected by third-party network operators or by the public Internet.

- RSVP-TE has been used to provide edge-to-edge tunnels through which flows to/from many endpoints can be routed, and this provides a reduction in state while still offering Traffic Engineering across the backbone network. However, this requires O(n2) connections and as the number of sites increases this becomes unsustainable.

- A centralized control system is capable of producing more efficient use of network resources and of allowing better coordination of network usage and of network diagnostics. However, such a system may present challenges in large and dynamic networks because it relies on all network state being held centrally, and it is difficult to make central control as robust and self-correcting as distributed control.

This document introduces an approach that blends the best points of each of these solution technologies to achieve a trade-off where RSVP-TE tunnels in the backbone network are stitched together using SR, and end-to-end SR paths can be created under the control of a central controller with routing devolved to the constituent networks where possible.

4. Decomposing the Problem

It is important to decompose the problem to take account of different regions spanned by the end-to-end path. These regions may use different technologies and may be under different administrative control. The separation of administrative control is particularly important because the operator of one region may be unwilling to share information about their networks, and may be resistant to allowing a third party to exert control over their network resources.

Using the reference model in Figure 1, we can consider how to get a packet from Source1 to the Destination. The following decisions must be made:

- In which site Destination lies.
- Which exit point from Site1 to use.
- Which entry point to Site2 to use.
- How to reach the exit point of Site1 from Source1.
- How to reach the entry point to Site2 from the exit point of Site1.
- How to reach Destination from the entry point to Site2.

As already mentioned, these decisions may be inter-related. This enables us to break down the problem into three steps:

1. Get the packet from Source1 to the exit point of Site1.
2. Get the packet from exit point of Site1 to entry point of Site2.
3. Get the packet from entry point of Site2 to Destination.

The solution needs to achieve this in a way that allows:

- Adequate discovery of preferred elements in the end-to-end path (such as the location of the destination, and the selection of the destination site entry point).
- Full control of the end-to-end path if all of the operators are willing.
- Re-use of existing techniques and technologies.

From a technology point of view we must support several functions and mixtures of those functions:

- If a site uses MPLS Segment Routing, the labels within the site may be populated by any means including BGP-LU [RFC8277], IGP [RFC8667] [RFC8665], and central control. Source routes within the site may be expressed as label stacks pushed by a controller or computed by a source router, or expressed as a single label and programmed into the site routers by a controller.

- If a site uses other (non-MPLS) forwarding, the site processing is specific to that technology. See Section 9 for details.

- If the sites use Segment Routing, the prefix-SIDs for the source and destination may be the same or different.

- The backbone network may be a single private network under the control of the owner of the sites and comprising one or more ASes, or may be a network operated by one or more third parties.

- The backbone network may utilize MPLS Traffic Engineering tunnels in conjunction with MPLS Segment Routing and the site-to-site source route may be provided by stitching TE LSPs.

- A single controller may be used to handle the source and destination site as well as the backbone network, or there may be a different controller for the backbone network separate from that that controls the two site, or there may be separate controllers for each network. The controllers may cooperate and share information to different degrees.

All of these different decompositions of the problem reflect different deployment choices and different commercial and operational practices, each with different functional trade-offs. For example, with separate controllers that do not share information and that only cooperate to a limited extent, it will be possible to achieve end-to-end connectivity with optimal routing at each step (site or backbone AS), but the end-to-end path that is achieved might not be optimal.

5. Solution Space

5.1. Global Optimization of the Paths

Global optimization of the path from one site to another requires either that the source controller has a complete view of the end-to-end topology or some form of cooperation between controllers (such as in Backward Recursive Path Computation (BRPC) in [RFC5441]).

BGP-LS [RFC7752] can be used to provide the "source" controller with a view of the topology of the backbone: that topology may be abstracted or partial. This requires some of the BGP speakers in each AS to have BGP-LS sessions to the controller. Other means of obtaining this view of the topology are of course possible.

5.2. Figuring Out the GWs at a Destination Site for a Given Prefix

Suppose GW2a and GW2b both advertise a route to prefix X, each setting itself as next hop. One might think that the GWs for X could be inferred from the routes' next hop fields, but typically only the "best" route (as selected by BGP) gets distributed across the backbone: the other route is discarded. But the best route according to the BGP selection process might not be the route via the GW that we want to use for traffic engineering purposes.

The obvious solution would be to use the ADD-PATH mechanism [RFC7911] to ensure that all routes to X get advertised. However, even if one does this, the identity of the GWs would get lost as soon as the routes got distributed through an ASBR that sets next hop self. And if there are multiple ASes in the backbone, not only will the next hop change several times, but the ADD-PATH mechanism will experience scaling issues. So this "obvious" solution only works within a single AS.

A better solution can be achieved using the Tunnel Encapsulation [RFC9012] attribute as follows.

We define a new tunnel type, "SR tunnel", and when the GWs to a given site advertise a route to a prefix X within the site, they each include a Tunnel Encapsulation attribute with multiple remote endpoint sub-TLVs each of which identifies a specific GW to the site.

In other words, each route advertised by any GW identifies all of the GWs to the same site (see Section 9 for a discussion of how GWs discover each other). Therefore, only one of the routes needs to be distributed to other ASes, and it doesn't matter how many times the next hop changes, the Tunnel Encapsulation attribute (and its remote endpoint sub-TLVs) remains unchanged and disclose the full list of GWs to the site.

Further, when a packet destined for prefix X is sent on a TE path to GW2a we want the packet to arrive at GW2a carrying, at the top of its label stack, GW2a's label for prefix X. To achieve this we place the SID/SRGB in a sub-TLV of the Tunnel Encapsulation attribute. We define the prefix-SID sub-TLV to be essentially identical in syntax to the prefix-SID attribute (see [RFC8669]), but the semantics are somewhat different.

We also define an "MPLS Label Stack" sub-TLV for the Tunnel Encapsulation attribute, and put this in the "SR tunnel" TLV. This allows the destination GW to specify a label stack that it wants packets destined for prefix X to have. This label stack represents a source route through the destination site.

5.3. Figuring Out the Backbone Egress ASBRs

We need to figure out the backbone egress ASBRs that are attached to a given GW at the destination site in order to properly engineer the path across the backbone.

The "cleanest" way to do this is to have the backbone egress ASBRs distribute the information to the source controller using the egress peer engineering (EPE) extensions of BGP-LS [I-D.ietf-idr-bgpls-segment-routing-epe]. The EPE extensions to BGP-LS allow a BGP speaker to say, "Here is a list of my EBGP neighbors, and here is a (locally significant) adjacency-SID for each one."

It may also be possible to consider utilizing cooperating PCEs or a Hierarchical PCE approach in [RFC6805]. But it should be observed that this question is dependent on the questions in Section 5.2. That is, it is not possible to even start the selection of egress ASBRs until it is known which GWs at the destination site provide access to a given prefix. Once that question has been answered, any number of PCE approaches can be used to select the right egress ASBR and, more generally, the ASBR path across the backbone.

5.4. Making use of RSVP-TE LSPs Across the Backbone

There are a number of ways to carry traffic across the backbone from one site to another. RSVP-TE is a popular mechanism for establishing tunnels across MPLS networks in similar scenarios (e.g., L3VPN) because it allows for reservation of resources as well as traffic steering.

A controller can cause an RSVP-TE LSP to be set up by talking to the LSP head end using PCEP extensions as described in [RFC8281]. That document specifies an "LSP Initiate" message (the PCInitiate message) that the controller uses to specify the RSVP-TE LSP endpoints, the explicit path, a "symbolic pathname", and other optional attributes (specified in the PCEP specification [RFC5440]) such as bandwidth.

When the head end receives a PCInitiate message, it sets up the RSVP-TE LSP, assigns it a "PLSP-id", and reports the PLSP-id back to the controller in a PCRpt message [RFC8231]. The PCRpt message also contains the symbolic name that the controller assigned to the LSP, as well as containing some information identifying the LSP-initiate

message from the controller, and details of exactly how the LSP was set up (RRO, bandwidth, etc.).

The head end can add a TE-PATH-BINDING TLV to the PCRpt message [I-D.ietf-pce-binding-label-sid]. This allows the head end to assign a "binding SID" to the LSP, and to report to the controller that a particular binding SID corresponds to a particular LSP. The binding SID is locally scoped to the head end.

The controller can make this label be part of the label stack that it tells the source (or the GW at the source site) to impose on the data packets being sent to prefix X. When the head end receives a packet with this label at the top of the stack it will send the packet onward on the LSP.

5.5. Data Plane

Consolidating all of the above, consider what happens when we want to move a data packet from Source1 to Destination in Figure 1via the following source route:

Source1---GW1b---PE2a---ASBR2a---ASBR3a---PE3a---GW2a---Destination

Further, assume that there is an RSVP-TE LSP from PE2a to ASBR2a and an RSVP-TE LSP from ASBR3a to PE3a both of which we want to use.

Let's suppose that the Source pushes a label stack as instructed by the controller (for example, using BGP-LU [RFC8277]). We won't worry for now about source routing through the sites themselves: that is, in practice there may be additional labels in the stack to cover the source route from Source1 to GW1b and from GW2a to the Destination, but we will focus only on the labels necessary to leave the source site, traverse the backbone, and enter the egress site. So we only care what the stack looks like when the packet gets to GW1b.

When the packet gets to GW1b, the stack should have six labels:

Top Label:

> Peer-SID or adjacency-SID identifying the link or links to PE2a. These SIDs are distributed from GW1b to the controller via the EPE extensions of BGP-LS. This label will get popped by GW1b, which will then send the packet to PE2a.

Second Label:

> Binding SID advertised by PE2a to the controller for the RSVP-TE LSP to ASBR2a. This binding SID is advertised via the PCEP

extensions discussed above. This label will get swapped by PE2a for the label that the LSP's next hop has assigned to the LSP.

Third Label:

Peer-SID or adjacency-SID identifying the link or links to ASBR3a, as advertised to the controller by ASBR2a using the BGP-LS EPE extensions. This label gets popped by ASBR2a, which then sends the packet to ASBR3a.

Fourth Label:

Binding SID advertised by ASBR3a for the RSVP-TE LSP to PE3a. This binding SID is advertised via the PCEP extensions discussed above. ASBR3a treats this label just like PE2a treated the second label above.

Fifth label:

Peer-SID or adjacency-SID identifying link or links to GW2a, as advertised to the controller by ASBR3a using the BGP-LS EPE extensions. ASBR3a pops this label and sends the packet to GW2a.

Sixth Label:

Prefix-SID or other label identifying the Destination advertised in a Tunnel Encapsulation attribute by GW2a. This can be omitted if GW2a is happy to accept IP packets, or prefers a VXLAN tunnel for example. That would be indicated through the Tunnel Encapsulation attribute of course.

Note that the size of the label stack is proportional to the number of RSVP-TE LSPs that get stitched together by SR.

See Section 7 for some detailed examples that show the concrete use of labels in a sample topology.

In the above example, all labels except the sixth are locally significant labels: peer-SIDs, binding SIDs, or adjacency-SIDs. Only the sixth label, a prefix-SID, has a value that is unique across the whole SR domain. To impose that label, the source needs to know the SRGB of GW2a. If all nodes have the same SRGB, this is not a problem. Otherwise, there are a number of different ways GW3a can advertise its SRGB. This can be done via the segment routing extensions of BGP-LS, or it can be done using the prefix-SID attribute or BGP-LU [RFC8277], or it can be done using the BGP Tunnel Encapsulation attribute. The technique to be used will depend on the details of the deployment scenario.

The reason the above example is primarily based on locally significant labels is that it creates a "strict source route", and it presupposes the EPE extensions of BGP-LS. In some scenarios, the EPE extension to BGP-LS might not be available (or BGP-LS might not be available at all). In other scenarios, it may be desirable to steer a packet through a "loose source route". In such scenarios, the label stack imposed by the source will be based upon a sequence of "node-SIDs" that are unique across the whole SR domain, where each represents one of the hops of source route. Each label has to be computed by adding the corresponding node-SID to the SRGB of the node that will act upon the label. One way to learn the node-SIDs and SRGBs is to use the segment routing extensions of BGP-LS. Another way is to use BGP-LU as follows:

Each node that may be part of a source route originates a BGP-LU route with one of its own loopback addresses as the prefix. The BGP prefix-SID attribute is attached to this route. The prefix-SID attribute contains a SID that is the SID corresponding to the node's loopback address and which is unique across the whole SR domain. The attribute also contains the node's SRGB.

While this technique is useful when BGP-LS is not available, there needs to be some other means for the source controller to discover the topology. In this document, we focus primarily on the scenario where BGP-LS, rather than BGP-LU, is used.

5.6. Centralized and Distributed Controllers

A controller or set of controllers is needed to collate topology and TE information from the constituent networks, to apply policies and service requirements to compute paths across those networks, to select an end-to-end path, and to program key nodes in the network to take the right forwarding actions (pushing label stacks, stitching LSPs, forwarding traffic).

- It is commonly understood that a fully optimal end-to-end path can only be computed with full knowledge of the end-to-end topology and available Traffic Engineering resources. Thus, one option is for all information about the site networks and backbone network to be collected by a central controller that makes all path computations and is responsible for issuing the necessary programming commands. Such a model works best when there is no commercial or administrative impediment (for example, where the sites and the backbone network are owned and operated by the same organization). There may, however, be some scaling concerns if the component networks are large.

In this mode of operation, each network may use BGP-LS to export Traffic Engineering and topology information to the central controller, and the controller may use PCEP to program the network behavior.

- A similar centralized control mechanism can be used with a scalability improvement that risks a reduction in optimality. In this case, the site networks can export to the controller just the feasibility of connectivity between data source/sink and gateway, perhaps enhancing this with some information about the Traffic Engineering metrics of the potential paths.

  This approach allows the central controller to understand the end-to-end path that it is selecting, but not to control it fully. The source route from data source to site egress gateway is left to the source host or a controller in the source site, while the source route from site ingress gateway to destination is left as a decision for the site ingress gateway or to a controller in the destination site and in both cases the traffic may be left to follow the IGP shortest path.

  This mode of operation still leaves overall control with a centralized server and that may not be considered suitable when there is separate commercial or administrative control of the networks.

- When there is separate commercial or administrative control of the networks, the site operator will not want the backbone operator to have control of the paths within the sites and may be reluctant to disclose any information about the topology or resource availability within the sites. Conversely, the backbone operator may be very unwilling to allow the site operator (a customer) any control over or knowledge about the backbone network.

  This "problem" has already been solved for Traffic Engineering in MPLS networks that span multiple administrative domains and leads to several potential solutions:

  * Per-domain path computation [RFC5152] can be seen as "best effort optimization". In this mode the controller for each domain is responsible for finding the best path to the next domain, but has no way of knowing which is the best exit point from the local domain. The resulting path may end up significantly sub-optimal or even blocked.

  * Backward recursive path computation (BRPC) [RFC5441] is a mechanism that allows controllers to cooperate across a small set of domains (such as ASes) to build a tree of possible paths

and so allow the controller for the ingress domain to select the optimal path. The details of the paths within each domain that might reveal confidential information can be hidden using Path Keys [RFC5520]. BRPC produces optimal paths, but scales poorly with an increase in domains and with an increase in connectivity between domains. It can also lead to slow computation times.

* Hierarchical PCE (H-PCE) [RFC6805] is a two-level cooperation process between PCEs. The child PCEs remain responsible for computing paths across their domains, and they coordinate with a parent PCE that stitches these paths together to form the end-to-end path. This approach has many similarities with BRPC but can scale better through the maintenance of "domain topology" that shows how the domains are interconnected, and through the ability to pipe-line computation requests to all of the child domains. It has the drawback that some party has to own and operate the parent PCE.

* An alternative approach is documented by the TEAS working group [RFC7926]. In this model each network advertises to controllers for adjacent networks (using BGP-LS) selected information about potential connectivity across the network. It does not have to show full topology and can make its own decisions about which paths it considers optimal for use by its different neighbors and customers. This approach is suitable for the End-to-End Domain Interconnect Traffic Steering problem where the backbone is under different control from the domains because it allows the overlay nature of the use of the backbone network to be treated as a peer network relationship by the controllers of the domains - the domains can be operated using a single controller or a separate controller for each domain.

It is also possible to operate domain interconnection when some or all domains do not have a controller. Segment Routing is capable of routing a packet toward the next hop based on the top label on the stack, and that label does not need to indicate an immediately adjacent node or link. In these cases, the packet may be forwarded untouched, or the forwarding router may impose a locally-determined additional set of labels that define the path to the next hop.

PCE can be used to instruct the source host or a transit node about what label stacks to add to packets. That is, a node that needs to impose labels (either to start routing the packet from the source host, or to advance the packet from a transit router toward the destination) can determine the label stack to use based on local function or can have that stack supplied by a PCE. The PCE Communication Protocol (PCEP) has been extended to allow the PCE to

supply a label stack for reaching a specific destination either in response to a request or in an unsolicited manner [RFC8664].

6. BGP-LS Considerations

This section gives an overview of the use of BGP-LS to export an abstraction (or summary) of the connectivity across the backbone network by means of two figures that show different views of a sample network.

Figure 2 shows a more complex reference architecture.

Figure 3 represents the minimum set of nodes and links that need to be advertised in BGP-LS with SR in order to perform Site Interconnect with traffic engineering across the backbone network: the PEs, ASBRs, and GWs, and the links between them. In particular, EPE [I-D.ietf-idr-bgpls-segment-routing-epe] and TE information with associated segment IDs is advertised in BGP-LS with SR.

Links that are advertised may be physical links, links realized by LSP tunnels or SR paths, or abstract links. It is assumed that intra-AS links are either real links, RSVP-TE LSPs with allocated bandwidth, or SR TE policies as described in [I-D.ietf-idr-segment-routing-te-policy]. Additional nodes internal to an AS and their links to PEs, ASBRs, and/or GWs may also be advertised (for example, to avoid full mesh problems).

Note that Figure 3 does not show full interconnectivity. For example, there is no possibility of connectivity between PE1a and PE1c (because there is no RSVP-TE LSP established across AS1 between these two nodes) and so no link is presented in the topology view. [RFC7926] contains further discussion of topological abstractions that may be useful in understanding this distinction.

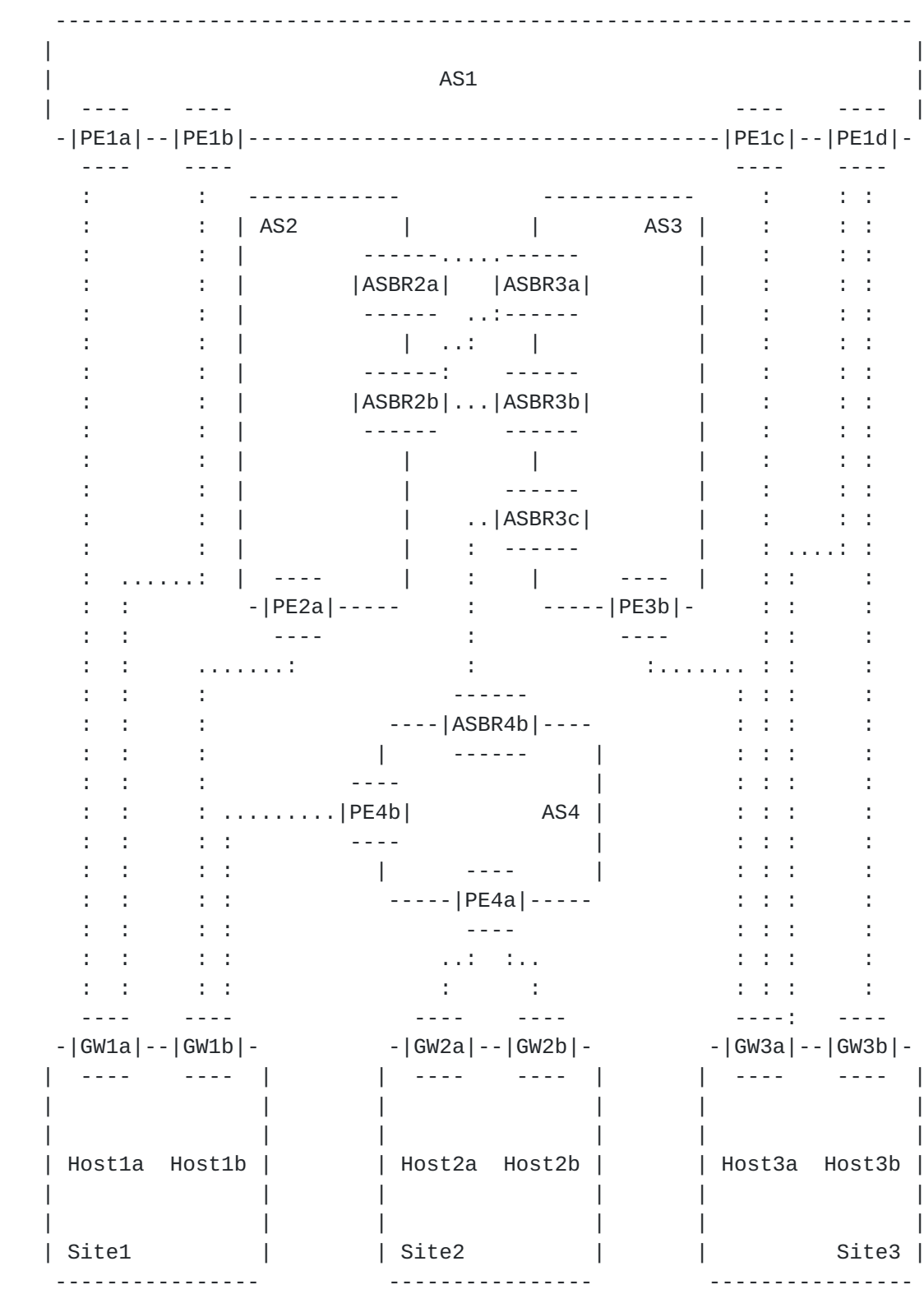


Figure 2: Network View of Example Configuration

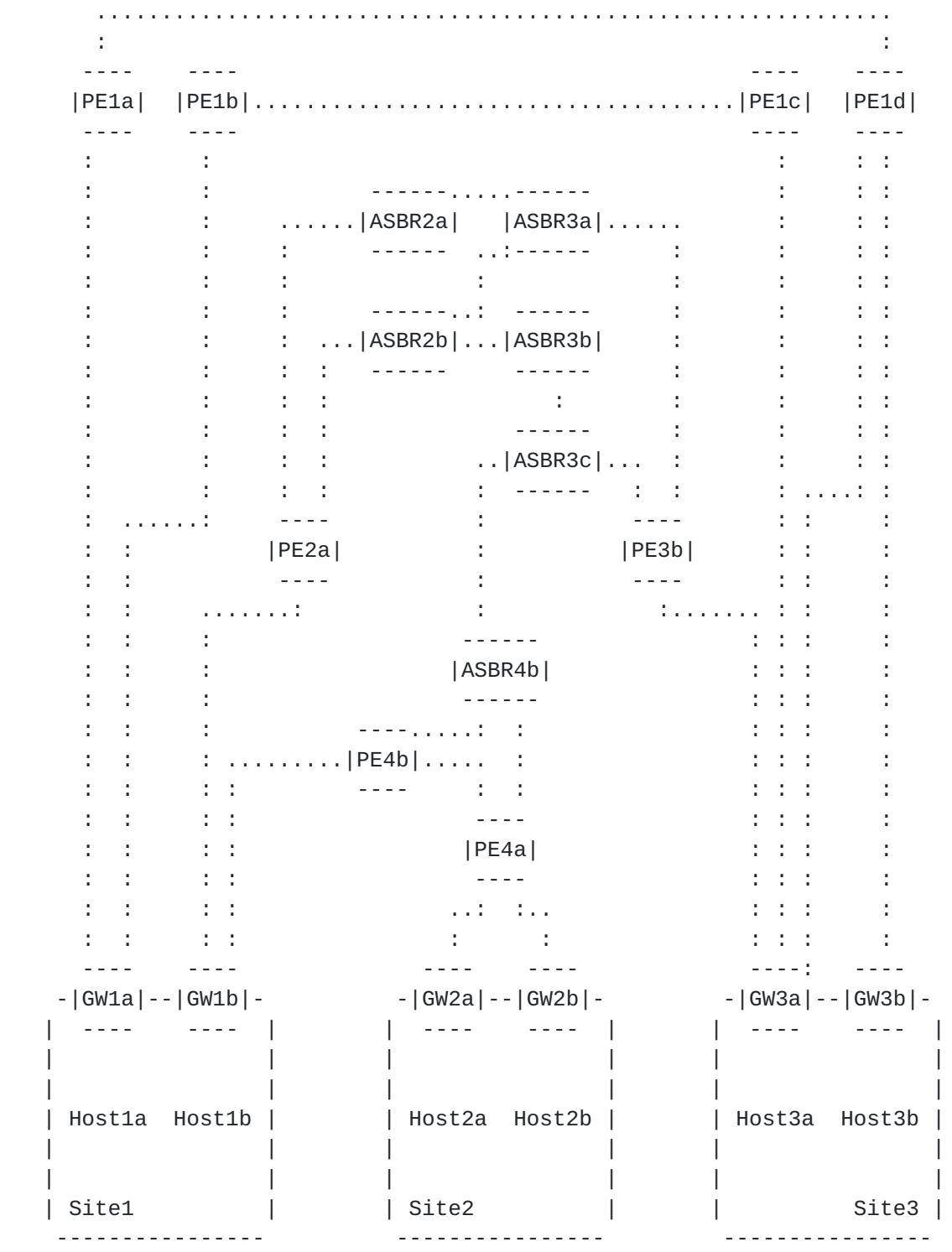


Figure 3: Topology View of Example Configuration

A node (a PCE, router, or host) that is computing a full or partial path correlates the topology information disseminated in BGP-LS with

the information advertised in BGP (with the Tunnel Encapsulation attributes) and uses this to compute that path and obtain the SIDs for the elements on that path. In order to allow a source host to compute exit points from its site, some subset of the above information needs to be disseminated within that site.

What is advertised external to a given AS is controlled by policy at the ASes' PEs, ASBRs, and GWs. Central control of what each node should advertise, based upon analysis of the network as a whole, is an important additional function. This and the amount of policy involved may make the use of a Route Reflector an attractive option.

Local configuration at each node determines which links to other nodes are advertised in BGP-LS, and determines which characteristics of those links are advertised. Pairwise coordination between link end-points is required to ensure consistency.

Path Weighted ECMP (PWECMP) is a mechanism to load-balance traffic across parallel equal cost links or paths. In this approach an ingress node distributes the flows from it to a given egress node across the equal cost paths to the egress node in proportion to the lowest bandwidth link on each path. PWECMP can be used by a GW for a given source site to send all flows to a given destination site using all paths in the backbone network to that destination site in proportion to the minimum bandwidth on each path. PWECMP may also be used by hosts within a source site to send flows to that site's GWs.

7. Worked Examples

Figure 4 shows a view of the links, paths, and labels that can be assigned to part of the sample network shown in Figure 2 and Figure 3. The double-dash lines (===) indicate LSP tunnels across backbone ASes and dotted lines (...) are physical links.

A label may be assigned to each outgoing link at each node. This is shown in Figure 4. For example, at GW1a the label L201 is assigned to the link connecting GW1a to PE1a. At PE1c, the label L302 is assigned to the link connecting PE1c to GW3b. Labels ("binding SIDs") may also be assigned to RSVP-TE LSPs. For example, at PE1a, label L202 is assigned to the RSVP-TE LSP leading from PE1a to PE1c.

At the destination site, label L305 is a "node-SID"; it represents Host3b, rather than representing a particular link.

When a node processes a packet, the label at the top of the label stack indicates the link (or RSVP-TE LSP) on which that node is to transmit the packet. The node pops that label off the label stack before transmitting the packet on the link. However, if the top

label is a node-SID, the node processing the packet is expected to transmit the packet on whatever link it regards as the shortest path to the node represented by the label.

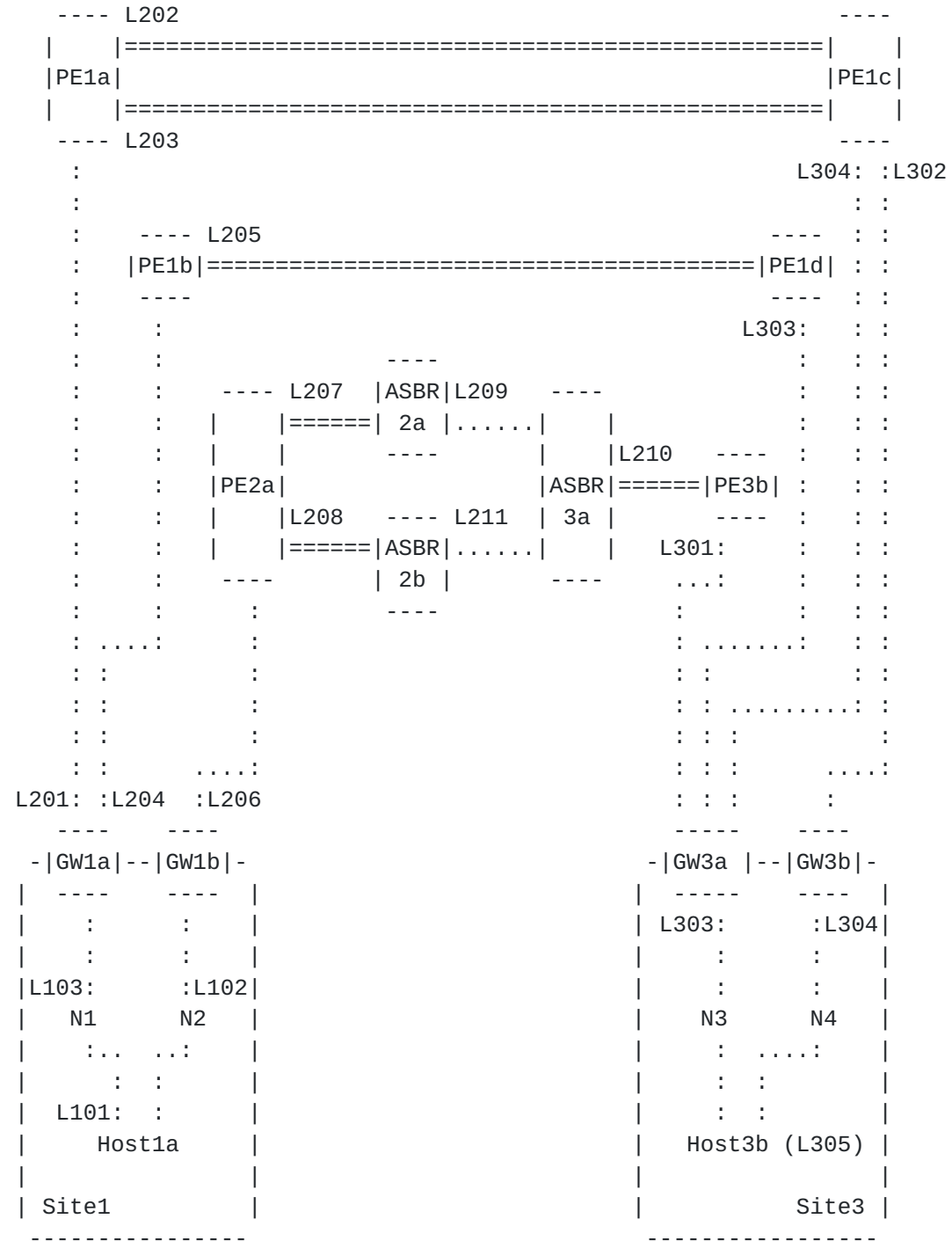


Figure 4: Tunnels and Labels in Example Configuration

Note that label spaces can overlap so that, for example, the figure shows two instances of L303 and L304. This is acceptable because of the separation between the sites, and because SIDs applied to outgoing interfaces are locally scoped.

Let's consider several different possible ways to direct a packet from Host1a in Site1 to Host3b in Site3.

a. Full source route imposed at source

In this case it is assumed that the entity responsible for determining an end-to-end path has access to the topologies of both the source and destination sites as well as of the backbone network. This might happen if all of the networks are owned by the same operator in which case the information can be shared into a single database for use by an offline tool, or the information can be distributed using routing protocols such that the source host can see enough to select the path. Alternatively, the end-to-end path could be produced through cooperation between computation entities each responsible for different sites and ASes along the path.

If the path is computed externally it is pushed to the source host. Otherwise, it is computed by the source host itself.

Suppose it is desired for a packet from Host1a to travel to Host3b via the following source route:

```
Host1a->N1->GW1a->PE1a->(RSVP-TE
LSP)->PE1c->GW3b->N4->Host3b
```

Host1a imposes the following label stack (with the first label representing the top of stack), and then sends the packet to N1:

```
L103, L201, L202, L302, L304, L305
```

N1 sees L103 at the top of the stack, so it pops the stack and forwards the packet to GW1a. GW1a sees L201 at the top of the stack, so it pops the stack and forwards the packet to PE1a. PE1a sees L202 at the top of the stack, so it pops the stack and forwards the packet over the RSVP-TE LSP to PE1c. As the packet travels over this LSP, its top label is an RSVP-TE signaled label representing the LSP. That is, PE1a imposes an additional label stack entry for the tunnel LSP.

At the end of the LSP tunnel, the MPLS tunnel label is popped, and PE1c sees L302 at the top of the stack. PE1c pops the

stack and forwards the packet to GW3b. GW3b sees L304 at the top of the stack, so it pops the stack and forwards the packet to N4. Finally, N4 sees L305 at the top of the stack, so it pops the stack and forwards the packet to Host3b.

b. It is possible that the source site does not have visibility into the destination site.

This occurs if the destination site does not export its topology, but does export basic reachability information so that the source host or the path computation entity will know:

+ The GWs through which the destination can be reached.

+ The SID to use for the destination prefix.

Suppose we want a packet to follow the source route:

Host1a->N1->GW1a->PE1a->(RSVP-TE
LSP)->PE1c->GW3b->...->Host3b

The ellipsis indicates a part of the path that is not explicitly specified. Thus, the label stack imposed at the source host is:

L103, L201, L202, L302, L305

Processing is as per case a., but when the packet reaches the GW of the destination site (GW3b) it can either simply forward the packet along the shortest path to Host3b, or it can insert additional labels to direct the path to the destination.

c. Site1 only has reachability information for the backbone and destination networks

The source site (or the path computation entity) may be further restricted in its view of the network. It is possible that it knows the location of the destination in the destination site, and knows the GWs to the destination site that provide reachability to the destination, but that it has no view of the backbone network. This leads to the packet being forwarded in a manner similar to 'per-domain path computation' described in Section 5.6.

At the source host a simple label stack is imposed navigating the site and indicating the destination GW and the destination host.

L103, L302, L305

As the packet leaves the source site, the source GW (GW1a) determines the PE to use to enter the backbone using nothing more than the BGP preferred route to the destination GW (it could be PE1a or PE1b).

When the packet reaches the first PE it has a label stack just identifying the destination GW and the host (L302, L305). The PE uses information it has about the backbone network topology and available LSPs to select an LSP tunnel, impose the tunnel label, and forward the packet.

When the packet reaches the end of the LSP tunnel, it is processed as described in case b.

d. Stitched LSPs across the backbone

A variant of all these cases arises when the packet is sent using a path that spans multiple ASes. For example, one that crosses AS2 and AS3 as shown in Figure 2.

In this case, basing the example on case a., the source host imposes the label stack:

L102, L206, L207, L209, L210, L301, L303, L305

It then sends the packet to N2.

When the packet reaches PE2a, as previously described, the top label (L207) indicates an LSP tunnel that leads to ASBR2a. At the end of that LSP tunnel the next label (L209) routes the packet from ASBR2a to ASBR3a, where the next label (L210) identifies the next LSP tunnel to use. Thus, SR has been used to stitch together LSPs to make a longer path segment. As the packet emerges from the final LSP tunnel, forwarding continues as previously described.

8. Label Stack Depth Considerations

As described in Section 3.1, one of the issues with a Segment Routing approach is that the label stack can get large, for example when the source route becomes long. A mechanism to mitigate this problem is needed if the solution is to be fully applicable in all environments.

[I-D.ietf-idr-segment-routing-te-policy] introduces the concept of hierarchical source routes as a way to compress source route headers. It functions by having the egress node for a set of source routes

advertise those source routes along with an explicit request that each node that is an ingress node for one or more of those source routes should advertise a binding SID for the set of source routes for which it is the ingress. It should be noted that the set of source routes can either be advertised by the egress node as described here, or advertised by a controller on behalf of the egress node.

Such an ingress node advertises its set of source routes and a binding SID as an adjacency in BGP-LS as described in Section 6. These source routes represent the weighted ECMP paths between the ingress node and the egress node. Note also that the binding SID may be supplied by the node that advertises the source routes (the egress or the controller), or may be chosen by the ingress.

A remote node that wishes to reach the egress node constructs a source route consisting of the segment IDs necessary to reach one of the ingress nodes for the path it wishes to use along with the binding SID that the ingress node advertised to identify the set of paths. When the selected ingress node receives a packet with a binding SID it has advertised, it replaces the binding SID with the labels for one of its source routes to the egress node (it will choose one of the source routes in the set according to its own weighting algorithms and policy).

8.1. Worked Example

Consider the topology in Figure 4. Suppose that it is desired to construct full segment routed paths from ingress to egress, but that the resulting label stack (segment route) is too large. In this case the gateways to Site3 (GW3a and GW3b) can advertise all of the source routes from the gateways to Site1 (GW1a and GW1b). The gateways to Site1 then assign binding SIDs to those source routes and advertise those SIDs into BGP-LS.

Thus, GW3b advertises the two source routes (L201, L202, L302 and L201, L203, L302), and GW1a advertises into BGP-LS its adjacency to GW3b along with a binding SID. Should Host1a wish to send a packet via GW1a and GW3b, it can include L103 and this binding SID in the source route. GW1a is free to choose which source route to use between itself and GW3b using its weighted ECMP algorithm.

Similarly, GW3a can advertise the following set of source routes:

- L201, L202, L304

- L201, L203, L304

- L204, L205, L303
- L206, L207, L209, L210, L301
- L206, L208, L211, L210, L301

GW1a advertises a binding SID for the first three, and GW1b advertises a binding SID for the other two.

9. Gateway Considerations

As described in Section 5.2, [I-D.ietf-bess-datacenter-gateway] defines a new tunnel type, "SR tunnel", and when the GWs to a given site advertise a route to a prefix X within the site, they will each include a Tunnel Encapsulation attribute with multiple tunnel instances each of type "SR tunnel", one for each GW and each containing a Remote Endpoint sub-TLV with that GW's address.

In other words, each route advertised by any GW identifies all of the GWs to the same site.

Therefore, even if only one of the routes is distributed to other ASes, it will not matter how many times the next hop changes, as the Tunnel Encapsulation attribute (and its remote endpoint sub-TLVs) will remain unchanged.

9.1. Site Gateway Auto-Discovery

To allow a given site's GWs to auto-discover each other and to coordinate their operations, the following procedures are implemented as described in [I-D.ietf-bess-datacenter-gateway]:

- Each GW is configured with an identifier of the site that is common across all GWs to the site and unique across all sites that are connected.
- A route target [RFC4360] is attached to each GW's auto-discovery route and has its value set to the site identifier.
- Each GW constructs an import filtering rule to import any route that carries a route target with the same site identifier that the GW itself uses. This means that only these GWs will import those routes and that all GWs to the same site will import each other's routes and will learn (auto-discover) the current set of active GWs for the site.
- The auto-discovery route each GW advertises consists of the following:

* An IPv4 or IPv6 NLRI containing one of the GW's loopback addresses (that is, with AFI/SAFI that is one of 1/1, 2/1, 1/4, 2/4).

* A Tunnel Encapsulation attribute containing the GW's encapsulation information, which at a minimum consists of an SR tunnel TLV with a Remote Endpoint sub-TLV [RFC9012].

To avoid the side effect of applying the Tunnel Encapsulation attribute to any packet that is addressed to the GW, the GW should use a different loopback address in the advertisement from that used to reach the GW itself.

Each GW will include a Tunnel Encapsulation attribute for each GW that is active for the site (including itself), and will include these in every route advertised by each GW to peers outside the site. As the current set of active GWs changes (due to the addition of a new GW or the failure/removal of an existing GW) each externally advertised route will be re-advertised with the set of SR tunnel instances reflecting the current set of active GWs.

9.2. Relationship to BGP Link State and Egress Peer Engineering

When a remote GW receives a route to a prefix X it can use the SR tunnel instances within the contained Tunnel Encapsulation attribute to identify the GWs through which X can be reached. It uses this information to compute SR TE paths across the backbone network looking at the information advertised to it in SR BGP Link State (BGP-LS) [I-D.ietf-idr-bgp-ls-segment-routing-ext] and correlated using the site identity. SR Egress Peer Engineering (EPE) [I-D.ietf-idr-bgpls-segment-routing-epe] can be used to supplement the information advertised in BGP-LS.

9.3. Advertising a Site Route Externally

When a packet destined for prefix X is sent on an SR TE path to a GW for the site containing X, it needs to carry the receiving GW's label for X such that this label rises to the top of the stack before the GW completes its processing of the packet. To achieve this we place a prefix-SID sub-TLV for X in each SR tunnel instance in the Tunnel Encapsulation attribute in the externally advertised route for X.

Alternatively, if the GWs for a given site are configured to allow remote GWs to perform SR TE through that site for prefix X, then each GW computes an SR TE path through that site to X from each of the current active GWs and places each in an MPLS label stack sub-TLV [RFC9012] in the SR tunnel instance for that GW.

9.4. Encapsulations

If the GWs for a given site are configured to allow remote GWs to send them packets in that site's native encapsulation, then each GW will also include multiple instances of a tunnel TLV for that native encapsulation in the externally advertised routes: one for each GW, and each containing a remote endpoint sub-TLV with that GW's address. A remote GW may then encapsulate a packet according to the rules defined via the sub-TLVs included in each of the tunnel TLV instances.

10. Security Considerations

There are several security domains and associated threats in this architecture. SR is itself a data transmission encapsulation that provides no additional security, so security in this architecture relies on higher layer mechanisms (for example, end-to-end encryption of pay-load data), security of protocols used to establish connectivity and distribute network information, and access control so that control plane and data plane packets are not admitted to the network from outside.

This architecture utilizes a number of control plane protocols within sites, within the backbone, and north-south between controllers and sites. Only minor modifications are made to BGP as described in [I-D.ietf-bess-datacenter-gateway], otherwise this architecture uses existing protocols and extensions so no new security risks are introduced.

Special care should, however, be taken when routing protocols export or import information from or to domains that might have a security model based on secure boundaries and internal mutual trust. This is notable when:

- BGP-LS is used to export topology information from within a domain to a controller that is sited outside the domain.

- A southbound protocol such as BGP-LU or Netconf is used to install state in the network from a controller that may be sited outside the domain.

In these cases protocol security mechanisms should be used to protect the information in transit entering or leaving the domain, and to authenticate the out-of-domain nodes (the controller) to ensure that confidential/private information is not lost and that data or configuration is not falsified.

In this context, a domain may be considered to be a site, an AS, or the whole SR domain.

11. Management Considerations

Configuration elements for the approaches described in this document are minor but crucial.

Each GW to a site is configured with the same identifier of the site, and that identifier is unique across all sites that are connected. This requires some coordination both within a site, and between cooperating sites. There are no requirements for how this configuration and coordination is achieved, but it is assumed that management systems are involved.

Policy determines what topology information is shared by a BGP-LS speaker (see Section 6). This applies both to the advertisement of interdomain links and their characteristics, and to the advertisement of summarized domain topology or connectivity. This policy is a local (i.e., domain-scoped) configuration dependent on the objectives and business imperatives of the domain operator.

Domain boundaries are usually configured to limit the control and interaction from other domains (for example, to not allow end-to-end TE paths to be set up across AS boundaries). As noted in Section 9.3, the GWs for a given site can be configured to allow remote GWs to perform SR TE through that site for a given prefix, a set of prefixes, or all reachable prefixes.

Similarly, (as described in Section 9.4 the GWs for a given site can be configured to allow remote GWs to send them packets in that site's native encapsulation.

12. IANA Considerations

This document makes no requests for IANA action.

13. Acknowledgements

Thanks to Jeffery Zhang for his careful review.

14. Informative References

[I-D.ietf-bess-datacenter-gateway]
Farrel, A., Drake, J., Rosen, E., Patel, K., and L. Jalil, "Gateway Auto-Discovery and Route Advertisement for Segment Routing Enabled Domain Interconnection", draft-ietf-bess-datacenter-gateway-10 (work in progress), April 2021.

[I-D.ietf-idr-bgp-ls-segment-routing-ext]
Previdi, S., Talaulikar, K., Filsfils, C., Gredler, H., and M. Chen, "BGP Link-State extensions for Segment Routing", draft-ietf-idr-bgp-ls-segment-routing-ext-18 (work in progress), April 2021.

[I-D.ietf-idr-bgpls-segment-routing-epe]
Previdi, S., Talaulikar, K., Filsfils, C., Patel, K., Ray, S., and J. Dong, "BGP-LS extensions for Segment Routing BGP Egress Peer Engineering", draft-ietf-idr-bgpls-segment-routing-epe-19 (work in progress), May 2019.

[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-12 (work in progress), May 2021.

[I-D.ietf-pce-binding-label-sid]
Sivabalan, S., Filsfils, C., Tantsura, J., Previdi, S., and C. Li, "Carrying Binding Label/Segment Identifier in PCE-based Networks.", draft-ietf-pce-binding-label-sid-08 (work in progress), April 2021.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <https://www.rfc-editor.org/info/rfc4360>.

[RFC5152] Vasseur, JP., Ed., Ayyangar, A., Ed., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, DOI 10.17487/RFC5152, February 2008, <https://www.rfc-editor.org/info/rfc5152>.

[RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <https://www.rfc-editor.org/info/rfc5440>.

[RFC5441] Vasseur, JP., Ed., Zhang, R., Bitar, N., and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", RFC 5441, DOI 10.17487/RFC5441, April 2009, <https://www.rfc-editor.org/info/rfc5441>.

[RFC5520] Bradford, R., Ed., Vasseur, JP., and A. Farrel, "Preserving Topology Confidentiality in Inter-Domain Path Computation Using a Path-Key-Based Mechanism", RFC 5520, DOI 10.17487/RFC5520, April 2009, <https://www.rfc-editor.org/info/rfc5520>.

[RFC6805] King, D., Ed. and A. Farrel, Ed., "The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS", RFC 6805, DOI 10.17487/RFC6805, November 2012, <https://www.rfc-editor.org/info/rfc6805>.

[RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <https://www.rfc-editor.org/info/rfc7752>.

[RFC7855] Previdi, S., Ed., Filsfils, C., Ed., Decraene, B., Litkowski, S., Horneffer, M., and R. Shakir, "Source Packet Routing in Networking (SPRING) Problem Statement and Requirements", RFC 7855, DOI 10.17487/RFC7855, May 2016, <https://www.rfc-editor.org/info/rfc7855>.

[RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", RFC 7911, DOI 10.17487/RFC7911, July 2016, <https://www.rfc-editor.org/info/rfc7911>.

[RFC7926] Farrel, A., Ed., Drake, J., Bitar, N., Swallow, G., Ceccarelli, D., and X. Zhang, "Problem Statement and Architecture for Information Exchange between Interconnected Traffic-Engineered Networks", BCP 206, RFC 7926, DOI 10.17487/RFC7926, July 2016, <https://www.rfc-editor.org/info/rfc7926>.

[RFC8231] Crabbe, E., Minei, I., Medved, J., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for Stateful PCE", RFC 8231, DOI 10.17487/RFC8231, September 2017, <https://www.rfc-editor.org/info/rfc8231>.

[RFC8277] Rosen, E., "Using BGP to Bind MPLS Labels to Address Prefixes", RFC 8277, DOI 10.17487/RFC8277, October 2017, <https://www.rfc-editor.org/info/rfc8277>.

[RFC8281] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for PCE-Initiated LSP Setup in a Stateful PCE Model", RFC 8281, DOI 10.17487/RFC8281, December 2017, <https://www.rfc-editor.org/info/rfc8281>.

[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>.

[RFC8660] Bashandy, A., Ed., Filsfils, C., Ed., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing with the MPLS Data Plane", RFC 8660, DOI 10.17487/RFC8660, December 2019, <https://www.rfc-editor.org/info/rfc8660>.

[RFC8664] Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing", RFC 8664, DOI 10.17487/RFC8664, December 2019, <https://www.rfc-editor.org/info/rfc8664>.

[RFC8665] Psenak, P., Ed., Previdi, S., Ed., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPF Extensions for Segment Routing", RFC 8665, DOI 10.17487/RFC8665, December 2019, <https://www.rfc-editor.org/info/rfc8665>.

[RFC8667] Previdi, S., Ed., Ginsberg, L., Ed., Filsfils, C., Bashandy, A., Gredler, H., and B. Decraene, "IS-IS Extensions for Segment Routing", RFC 8667, DOI 10.17487/RFC8667, December 2019, <https://www.rfc-editor.org/info/rfc8667>.

[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>.

[RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <https://www.rfc-editor.org/info/rfc9012>.

Authors' Addresses

Adrian Farrel
Old Dog Consulting

Email: adrian@olddog.co.uk

John Drake
Juniper Networks

Email: jdrake@juniper.net