BESS Working Group
Internet-Draft
Intended status: Standards Track
Expires: November 4, 2019

J. Drake
Juniper Networks
A. Farrel
Old Dog Consulting

May 3, 2019

BGP-LS Maps : A Framework for Network Slicing and Enhanced VPNs
draft-drake-bess-enhanced-vpn-00

Abstract

Future networks that support advanced services, such as those enabled by 5G mobile networks, envisions a set of overlay networks each with different performance and scaling properties. These overlays are known as network slices and are realized over a common underlay network.

In order to support network slicing, as well as to offer enhanced VPN services in general, it is necessary to define a mechanism by which specific resources (links and/or nodes) of an underlay network can be used by a specific network slice, VPN, or set of VPNs. This document sets out such a mechanism for use in Segment Routing networks.

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 4, 2019.

Table of Contents

1. Introduction

Network slicing is an approach to network operations that builds on the concept of network abstraction to provide programmability, flexibility, and modularity. Driven largely by needs surfacing from 5G, the concept of network slicing has gained traction, for example in [TS23501] and [TS28530]. Network slicing requires the underlying network to support partitioning the network resources to provide the client with dedicated (private) networking, computing, and storage resources drawn from a shared pool. The slices may be seen as (and operated as) virtual networks.

Advanced services drive a need to create virtual networks with enhanced characteristics. The tenant of such a virtual network can require a degree of isolation and performance that previously could only be satisfied by dedicated networks. Additionally, the tenant may ask for some level of control to their virtual networks, e.g., to customize the service forwarding paths in the underlying network.

The concepts of "enhanced VPNs" and "network slicing" are introduced in [I-D.ietf-teas-enhanced-vpn].

In order to support network slicing, as well as to offer enhanced VPN services in general, it is necessary to define a mechanism by which specific resources (links and/or nodes) of an underlay network can be used by a specific network slice, VPN, or set of VPNs. This document sets out such a mechanism for use in Segment Routing networks [RFC8402] and builds on the ideas introduced in [I-D.ietf-idr-segment-routing-te-policy].

Objectives/Notes...

- No per-VPN or per-flow state in P-routers

- Generalization of SR TE policy. Thus, degenerate case of this work is SR-TE policy

2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. Overview of Approach

The approach is based on the use of DSCP-based forwarding in the underlay network [RFC2474]. For each VPN or sets of VPNs that are to use a given underlay network, a central network controller assigns resources per {link, DSCP} pair based upon the {source, destination, DSCP} traffic matrix. That is, each VPN or set of VPNs gets a subset, either dedicated or shared, of the resources in the underlay network.

It should be noted that resources can be assigned at any of the following granularities:

- All PEs in a given VPN

- A set of PEs in a given VPN

- An individual PE in a given VPN.

Once the central controller has determined the resource assignments, it distributes this information to the PEs that participate in each VPN using the usual VPN information dissemination tools, e.g., route targets (RT), RT constraints, and route reflectors [RFC4364], [RFC4456].

One way to distribute this information to those PEs is to give them a customized but limited view of the underlay network. (Note that giving each PE a full view of the underlay network does not help the PEs to manipulate the resources assigned for use by a particular slice or VPN, but providing a customized and limited view of those resources as a "virtual network" allows the PE to direct traffic over the designated resources as necessary to best deliver the end-to-end services.)

The resource allocation information is encoded using BGP-LS. This approach is chosen for the following reasons:

- It is BGP-based so it integrates easily with the existing BGP-based VPN infrastructure [RFC4364] [RFC4684]

- It supports Segment Routing which is necessary to enforce the PEs' usage of the resources allocated to the VPN or set of VPNs

- It supports inter-AS connectivity which is a perquisite for supporting the existing BGP-based VPN infrastructure

- It is canonical, in that it can be used to advertise the resources of underlay networks that use either OSPF or IS-IS to advertise resources

It should be noted that this mechanism also follows the scalability model of the existing BGP-based VPN infrastructure, which is that the per-VPN information is restricted to only those PE routers that are supporting that VPN and that the P routers have no per-VPN state.

Standard VPNs do not receive this resource allocation information and continue to use CSPF-based Weighted ECMP (WECMP) in the underlay network. This means that resources used by enhanced VPNs are reserved and do not be part of the CSPF-based WECMP topology.

Additional to the programming of the PEs and its computation and assignment of resources for use by slices, VPN instances, or groups

of VPNs, the central controller also instructs the P routers to makes actual allocation of resources per-DSCP.

4. Detailed Protocol Operation

We define a BGP-LS Map to be a BGP-LS encoded description of a subset of the links and nodes in the underlay network. A BGP-LS Map defines the topology for a network slice or a set of one or more VPNs. The topology connects a set of one or more VPNs and which is be used by the PEs in those VPNs to send packets. A given map is tagged with the route targets of the VPNs whose PEs are to import the map. A BGP-LS map is pushed southbound to these PEs by a network controller and may provide more than one path between a given ingress/egress PE pair. It is assumed that the underlay network is enabled for segment routing. When an ingress PE needs to send a packet to an egress PE it selects a path to that egress PE from the topology defined by the BGP-LS maps it has imported, and the ingress PE specifies that path using a segment routing label stack.

To enable this function there is a need for a new attribute that is attached to a BGP-LS map update that contains a map ID, the version number, a map type (MP2MP, P2MP, or P2P), the total number of pieces in the map, and the specific piece number in hand. That is, it is assumed that a PE may import more than one BGP-LS map, that a given BGP-LS map may change over time, and that a given BGP-LS map may span multiple BGP updates. The map ID needs to be unique across the set of VPNs into which the BGP-LS map is to be imported.

A BGP-LS map that is imported by the PEs of more than one VPN must provide details of the connectivity between the PEs in each VPN into which that map is imported.

If a PE imports more than one BGP-LS map it may use the union of the links and nodes specified in each map when selecting a path. A PE should give precedence to BGP-LS maps of type P2MP and P2P when selecting a path. Routes targets specific to a given VPN/PE pair are needed for BGP-LS maps of type P2MP and P2P.

A given BGP-LS map may change in response to updates to the PE membership in a VPN to which the BGP-LS map applies or to updates to the underlay network. When this occurs, the network controller should push a new version of the affected BGP-LS maps. That is, it increments the version number of each BGP-LS map. This implies that the network controller needs to be connected to the route reflectors associated with the VPNs for which it is providing BGP-LS maps.

A BGP-LS map cannot be used by a PE until it is completely assembled. If the BGP-LS map that is being assembled is a newer version of a

BGP-LS map that the PE is currently using, the PE should continue to use its current version of the BGP-LS map until the newer version is completely assembled.

When selecting a path using one or more BGP-LS maps, an ingress PE can use a link or node only if it is active in the underlay network. If this precludes connectivity to the egress PE it may use links and nodes in the CSPF-based WECMP underlay network topology nominally allocated to non-enhanced VPN traffic.

Additionally, when there is a newly activated PE it will not be present in any of the BGP-LS maps used by the other PEs. Until a new BGP-LS map or maps that contain that PE has been distributed, other PEs will have to use these links and nodes to reach the newly activated PE and it will have to use these links and nodes to reach other PEs.

Notes:

- The approach is a generalization of SR TE policy, modulo the absence of per-path weights which we could probably add.
- I.e., SR TE policy is P2P and does not support VPNs while this is MP2MP and supports VPNs.
- The PE routers understand the topology of the multi-AS underlay network by participating in BGP-LS.
- The controller uses BGP-LS (southbound) in combination with the existing VPN infrastructure to provide the PEs with the subset of the underlay network resources that each VPN can use.
- I.e., the subset of underlay network resources that a given VPN can use is encoded in BGP-LS updates that are tagged with that VPN's route targets and this causes the PEs in that VPN to import the BGP-LS updates for that VPN
- these BGP-LS updates act as a filter on the underlay network topology.

4.1. The BGP-LS Map Attribute

[RFC4271] defines the BGP Path attribute. This document introduces a new Optional Transitive Path attribute called the BGP-LS Map attribute with value TBD1 to be assigned by IANA. AFI 16388 and SAFI 71 are used because we are specifying links and nodes in the underlay network.

The first BGP-LS Map attribute MUST be processed and subsequent instances MUST be ignored.

The common fields of the BGP-LS Map attribute are set as follows:

- Optional bit is set to 1 to indicate that this is an optional attribute.
- The Transitive bit is set to 1 to indicate that this is a transitive attribute.
- The Extended Length bit is set according to the length of the BGP-LS Map attribute as defined in [RFC4271].
- The Attribute Type Code is set to TBD1.

The content of the BGP-LS Map attribute is a series of Type-Length-Value (TLV) constructs. Each TLV may include sub-TLVs. All TLVs and sub-TLVs have a common format that is:

- Type: A single octet indicating the type of the BGP-LS Map attribute TLV. Values are taken from the registry described in section XXXX.
- Length: A two octet field indicating the length of the data following the Length field counted in octets.
- Value: The contents of the TLV.

The formats of the TLVs defined in this document are shown in the following sections. The presence rules and meanings are as follows.

- The BGP-LS Map attribute MUST contain a Map TLV.
- The BGP-LS Map attribute MAY contain a DSCP List TLV.
- The BGP-LS Map attribute MAY contain a Color List TLV.

4.1.1. The Map TLV

The BGP-LS Map attribute MUST contain exactly one Map TLV. Its format is shown in Figure 1. Note that a given BGP-LS map may span multiple UPDATE messages and the Topology, Version Number, and the Number of Fragments fields in the BGP-LS Map attribute contained in each UPDATE message MUST be set to the same value or the BGP-LS map is unusable.

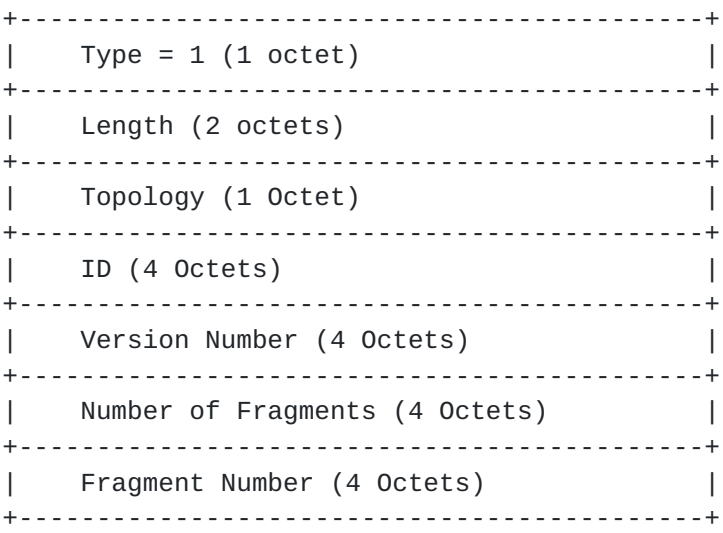

Figure 1: The Map TLV Format

The fields are as follows:

- Type is set to 1 to indicate a Map TLV.
- Length is set to 17 octets.
- Topology indicates whether this BGP-LS map is MP2MP, P2MP, P2P unidirectional, or P2P bidirectional.
- The ID of this BGP-LS map. This ID needs to be unique within the set of VPNs into which the BGP-LS map is to be imported.
- The Version Number of this BGP-LS map. I.e., the contents of a BGP-LS map with a given ID may change over time and this field indicates the latest version of that BGP-LS map.
- Number of Fragments indicates the number of BGP UPDATE messages defining this BGP-LS map.
- Fragment Number indicates ordinal position of this UPDATE message within the set of UPDATE messages defining this BGP-LS map. A BGP-LS map is not complete, i.e., usable, until all UPDATE messages have been received with Fragment Numbers in the range 1 <= Fragment Number <= Number of Fragments. An UPDATE message with a Fragment Number outside this range is to be ignored.

4.1.2. The DSCP List TLV

The DSCP List TLV MAY be included in the BGP-LS Map attribute. If included, a packet whose DSCP matches a DSCP in the DSCP list is to be forwarded using the BGP-LS map defined by the containing BGP-LS Map attribute. The first DSCP List TLV MUST be processed and subsequent instances MUST be ignored. The format of the DSCP List TLV is shown in Figure 2.

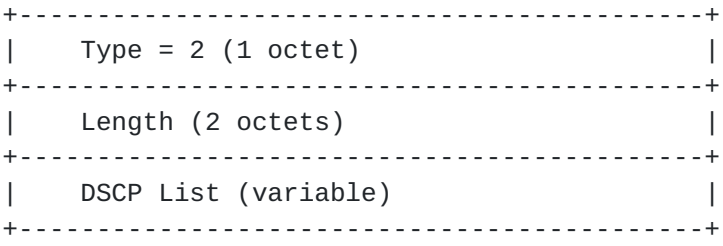

Figure 2: The DSCP List TLV Format

The fields are as follows:

- Type is set to 2 to indicate a DSCP List TLV.
- Length indicates the length in octets of the DSCP List.
- DSCP List contains a list of DSCPs, each one octet in length and encoded in the standard format.

4.1.3. The Color List TLV

The Color List TLV MAY be included in the BGP-LS Map attribute. If included, a packet whose Color, as defined by [RFC5512] matches a Color in the Color list is to be forwarded using the BGP-LS map defined by the containing BGP-LS Map attribute. The first Color List TLV MUST be processed and subsequent instances MUST be ignored. The format of the Color List TLV is shown in Figure 3.

Note that if both a DSCP List and a Color List TLV are included in a BGP-LS Map attribute, packets matching an entry in either list are to be forwarded using the containing BGP-LS Map attribute.

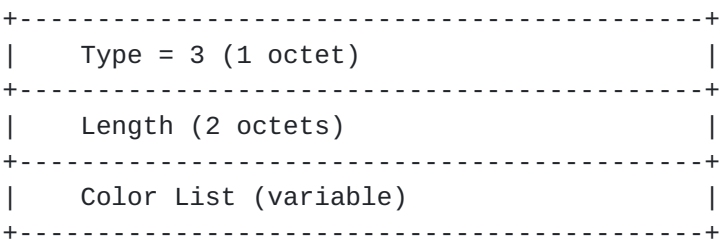

Figure 3: The Color List TLV Format

The fields are as follows:

- Type is set to 3 to indicate a Color List TLV.
- Length indicates the length in octets of the Color List.
- Color List contains a list of Colors, each four octets in length.

4.2. Error Handling

Section 6 of [RFC4271] describes the handling of malformed BGP attributes, or those that are in error in some way. [RFC7606] revises BGP error handling specifically for the for UPDATE message, provides guidelines for the authors of documents defining new attributes, and revises the error handling procedures for a number of existing attributes. This document introduces the BGP-LS Map attribute and so defines error handling as follows:

- When parsing a message, an unknown Attribute Type code or a length that suggests that the attribute is longer than the remaining message is treated as a malformed message and the "treat-as-withdraw" approach used as per [RFC7606].
- When parsing a message that contains an BGP-LS Map attribute, the following cases constitute errors:
    1. Optional bit is set to 0 in BGP-LS Map attribute.
    2. Transitive bit is set to 0 in BGP-LS Map attribute.
    3. The attribute does not contain a Map TLV or it contains more than one Map TLV.
    4. The TLV length indicates that the TLV extends beyond the end of the BGP-LS Map attribute.

5. There is an unknown TLV type field found in BGP-LS Map attribute.

- The errors listed above are treated as follows:

  1., 2., 3., 4.: The attribute MUST be treated as malformed and the "treat-as-withdraw" approach used as per [RFC7606].

  5.: Unknown TLVs SHOULD be ignored, and message processing SHOULD continue.

5. Comparison With ACTN

TBD

6. Examples

Figure 4shows a sample underlay topology. Six PEs (PE1 through PE6) are connected across a network of twelve P nodes (P1 through P12). Each PE is dual-homed, and the P nodes are variously connected so that there are multiple routes between PEs.

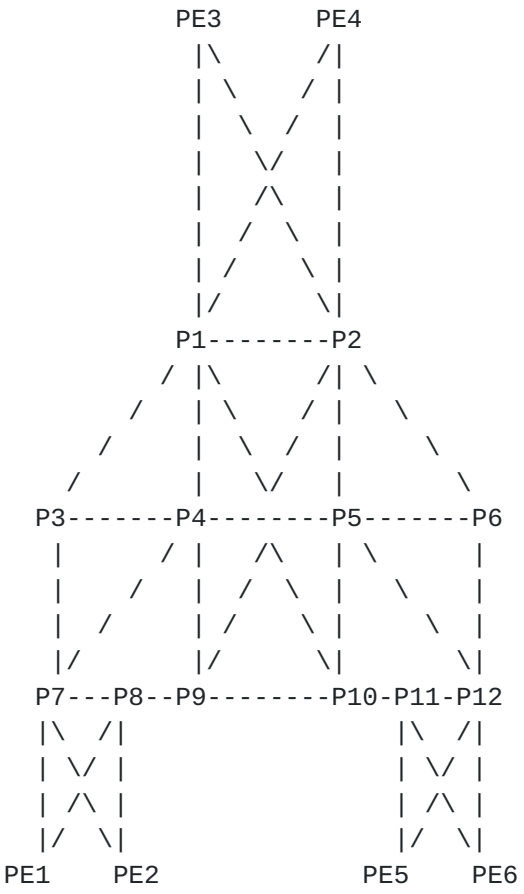

Figure 4: Underlay Network Topology

6.1. MP2MP Connectivity

Figure 5 shows how a Multi-point-to-multipoint (MP2MP) service that connects PE1, PE3, and PE6 can be installed over the underlay network. Path have been computed so that, for example, PE1 is connected to both PE3 and PE6 via a pair of redundant paths. Similarly, PE3 is connected to PE1 and PE6, and PE6 is connected to PE1 and PE3.

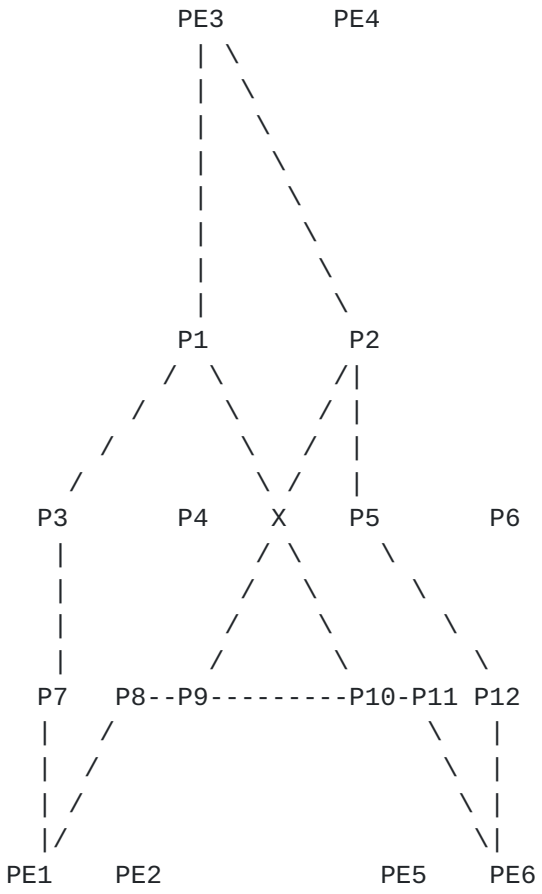

Figure 5: An MP2MP Service Installed at PE1, PE3, and PE6

6.2. P2MP Unidirectional Connectivity

Figure 6 shows the provision of a Point-to-Multipoint (P2MP) rooted at PE3 and connected to PE1 and PE6. As in the previous example, a redundant pair of paths is established between PE3 and each of PE1 and PE6. Thus, the two paths from PE3 to PE1 are PE3-P1-P4-P7-PE1 and PE3-P2-P9-P8-PE1.

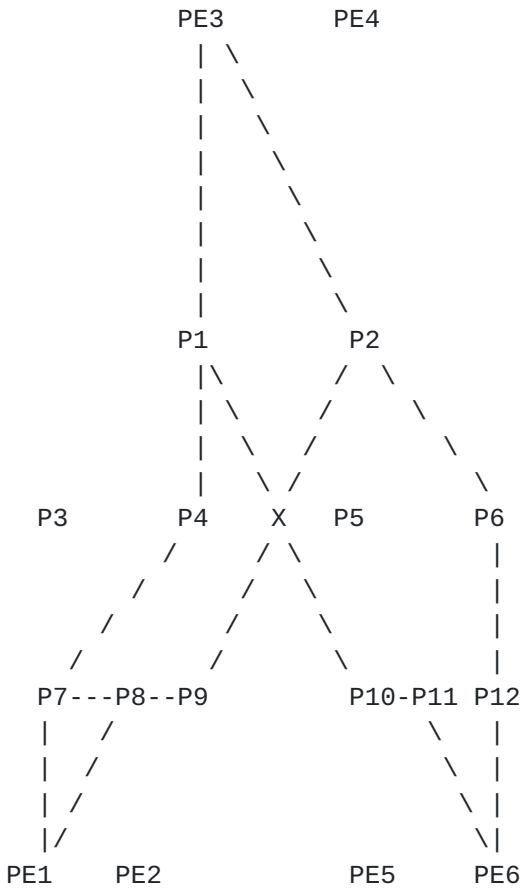

Figure 6: A P2MP Unidirectional Service Installed at PE3

6.3. P2P Unidirectional Connectivity

Figure 7 shows a Point-to-Point (P2P) service rooted at PE1 and connected to PE3. This is equivalent to a Segment Routing Traffic Engineering (SR TE) Policy [I-D.ietf-idr-segment-routing-te-policy] installed at PE1.

As in the previous examples, a pair of redundant paths are computed.

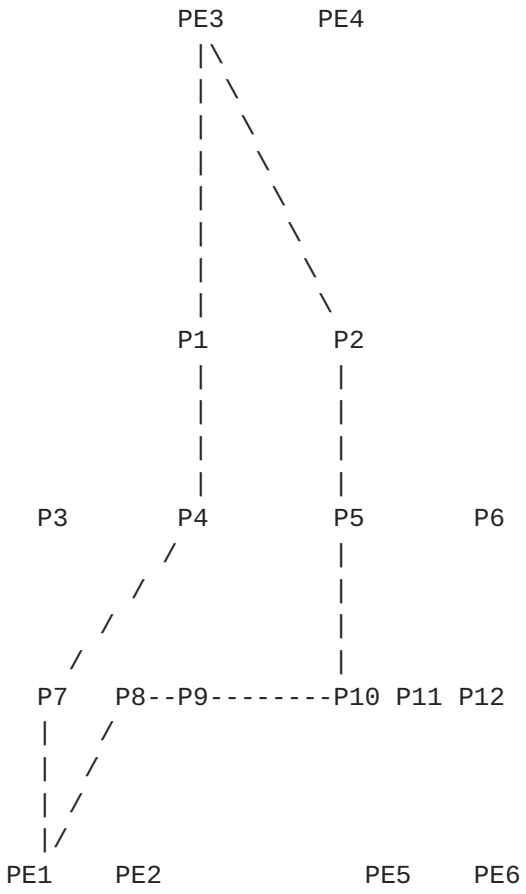

Figure 7: A P2P Unidirectional Service (SR TE Policy) Installed at PE1

6.4. P2P Bidirectional Connectivity

Figure 8 show a bidirectional P2P service connecting PE1 and PE6. This requires SR TE policy to be installed at PE1 and PE6.

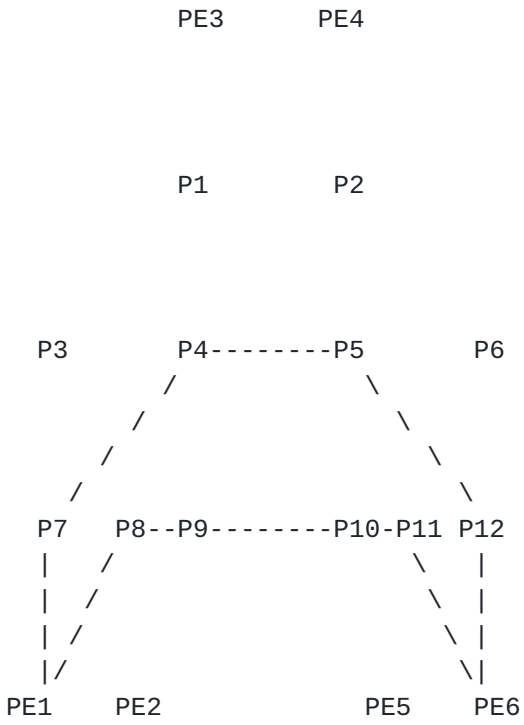

Figure 8: A P2P Bidirectional Service Installed at PE1 and PE6

7. Security Considerations

TBD

8. IANA Considerations

8.1. New BGP Path Attribute

IANA maintains a registry of "Border Gateway Protocol (BGP) Parameters" with a subregistry of "BGP Path Attributes". IANA is requested to assign a new Path attribute called "BGP-LS Map attribute" (TBD1 in this document) with this document as a reference.

8.2. New BGP-LS Map attribute TLVs Type Registry

IANA maintains a registry of "Border Gateway Protocol (BGP) Parameters". IANA is request to create a new subregistry called the "BGP-LS Map attribute TLVs" registry.

Valid values are in the range 0 to 255.

o Values 0 and 255 are to be marked "Reserved, not to be allocated".

- Values 1 through 254 are to be assigned according to the "First Come First Served" policy [RFC8126]

This document should be given as a reference for this registry. The new registry should track:

- Type
- Name
- Reference Document or Contact
- Registration Date

The registry should initially be populated as follows:

Type	Name	Reference	Date
1	Map TLV	[This.I-D]	Date-to-be-set
2	DSCP List TLV	[This.I-D]	Date-to-be-set
3	Color List TLV	[This.I-D]	Date-to-be-set

9. Acknowledgements

The authors are grateful to all those who contributed to the discussions that led to this work: TBD.

10. Contributors

The following people contributed text to this document:

A N Other
Email: another@foocorp.doc

11. References

11.1. Normative References

[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>.

[RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, DOI 10.17487/RFC2474, December 1998, <https://www.rfc-editor.org/info/rfc2474>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[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>.

[RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, DOI 10.17487/RFC5512, April 2009, <https://www.rfc-editor.org/info/rfc5512>.

[RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015, <https://www.rfc-editor.org/info/rfc7606>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

11.2. Informative References

[I-D.ietf-idr-segment-routing-te-policy]
Previdi, S., Filsfils, C., Jain, D., Mattes, P., Rosen, E., and S. Lin, "Advertising Segment Routing Policies in BGP", draft-ietf-idr-segment-routing-te-policy-05 (work in progress), November 2018.

[I-D.ietf-teas-enhanced-vpn]
Dong, J., Bryant, S., Li, Z., Miyasaka, T., and Y. Lee, "A Framework for Enhanced Virtual Private Networks (VPN+) Service", draft-ietf-teas-enhanced-vpn-01 (work in progress), February 2019.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <https://www.rfc-editor.org/info/rfc4456>.

[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>.

[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>.

[TS23501] 3GPP, "System architecture for the 5G System (5GS) - 3GPP TS23.501", 2016, <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144>.

[TS28530] 3GPP, "Management and orchestration; Concepts, use cases and requirements - 3GPP TS28.530", 2016, <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144>.

Authors' Addresses

John Drake
Juniper Networks

Email: jdrake@juniper.net

Adrian Farrel
Old Dog Consulting

Email: adrian@olddog.co.uk