Inter-Domain Routing
Internet-Draft
Intended status: Standards Track
Expires: March 7, 2019

K. Talaulikar
C. Filsfils
K. Swamy
Cisco Systems
S. Zandi
G. Dawra
LinkedIn
M. Durrani
Equinix
September 3, 2018

BGP Link-State Extensions for BGP-only Fabric
draft-ketant-idr-bgp-ls-bgp-only-fabric-01

Abstract

BGP is used as the only routing protocol in some networks today. In such networks, it is useful to get a detailed view of the nodes and underlying links in the topology along with their attributes similar to one available when using link state routing protocols. Such a view of a BGP-only fabric enables use cases like traffic engineering and forwarding of services along paths other than the BGP best path selection.

This document defines extensions to the BGP Link-state address-family (BGP-LS) and the procedures for advertisement of the topology in a BGP-only network. It also describes a specific use-case for traffic engineering based on Segment Routing.

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].

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 March 7, 2019.

Copyright Notice

Table of Contents

1. Introduction

Network operators are going for a BGP-only routing protocol for certain networks like Massively Scaled Data Centers (MSDCs). BGP MSDC [RFC7938] describes the requirement, design and operational aspects for use of BGP as the only routing protocol in MSDCs. The underlying link and topology information between BGP routers is hidden or abstracted in this design from the underlay routing for improving scalability and stability in a large scale network. On the flip side, there is no detailed topology view similar to one available in form of the Traffic Engineering (TE) Database (TED) when running link state routing protocols like OSPF [RFC2328] with extensions specified in OSPF-TE [RFC3630].

BGP-LS [RFC7752] enables advertisement of a link state topology via BGP that can be consumed by a controller or in general any software component to get a complete topology view of the network. BGP-LS extensions for advertisement of a BGP topology for the Egress Peer Engineering (EPE) use-case SR Central EPE [I-D.ietf-spring-segment-routing-central-epe] are specified in BGP-LS EPE [I-D.ietf-idr-bgpls-segment-routing-epe]. This document leverages the BGP-LS TLVs defined for BGP-LS EPE and other BGP-LS documents and specifies the procedures for advertising the underlying topology in a more generic BGP-only fabric use-case.

This document specifies the operations and procedures when using the design involving EBGP single-hop sessions over direct point-to-point links connecting the network nodes (refer BGP MSDC [RFC7938] for details). Certain modifications and other considerations are required when using a different design using IBGP or EGBP multihop and these would be specified in a future version of this document. While a data-center design is used as a reference, the procedures for topology advertisement may also apply to other networks with BGP-only fabric or to BGP-only portions of a larger network topology.

2. Topology Collection Mechanism

BGP-LS [RFC7752] has been defined to allow BGP to convey topology information in the form of Link-State objects - node, link and prefix. The properties of each of these objects are encoded as BGP-LS attributes. Applications need a topological view and visibility even for networks where BGP is the only routing protocol. In such networks, each BGP router advertises its local information which includes its node, links and prefix attributes via BGP-LS.

Figure 1 describes a typical deployment scenario. Every BGP router in the network is enabled for BGP-LS and forms BGP-LS sessions with one or more centralized BGP speakers over which it conveys its local

topology information. Each BGP router MAY receive the topology information from all other BGP routers via the centralized BGP speakers. This way, any BGP router (as also the centralized BGP speakers) MAY obtain aggregated Link-State information for the entire BGP network. An external component (e.g. a controller) can obtain this information from the centralized BGP speakers or directly by doing BGP-LS peering to the BGP routers. An internal software component on any of the BGP routers (e.g. TE module) can also receive the entire BGP network topology information from its local BGP process.

```
                +------------+
                | Controller |
                +------------+
                      ^
                      |
                      v
             +-------------------+
             |    BGP Speaker    |       +------------+
             |  (Centralized)    |       | Controller |
             +-------------------+       +------------+
                   ^   ^   ^                   ^
                   |   |   |                   |
       +-----------+   |   +----------------+  |
       |               |                    |  |
       v               v                    v  v
  +-----------+    +-----------+         +-----------+     +-----------+
  |    BGP    |    |    BGP    |         |    BGP    | <-->| Local     |
  |  Router   |    |  Router   |  . . .  |  Router   |     | Consumer  |
  +-----------+    +-----------+         +-----------+     +-----------+
       ^                ^                    ^
       |                |                    |
   Local Info       Local Info           Local Info
  (node & links)  (node & links)        (node & links)
```

Figure 1: Link State info collection

The design described above relies on the base BGP IPv4 or IPv6 routing underlay or any other mechanism for reachability for the BGP-LS session establishment with the centralized BGP speakers. Another alternate design would be to enable BGP-LS as well on the hop by hop EBGP sessions in the underlay. This approach results in the topology information being flooded via BGP-LS hop-by-hop along the BGP routers in the network. Other peering designs for BGP-LS sessions may also be possible and they are not precluded by this document and may be specified in a future version of this document.

3. Advertising BGP-only Network Topology

This sections specifies the BGP-LS TLVs and sub-TLVs and their use for advertising the topology of a BGP-only network in the form of BGP-LS Node, Link and Prefix NLRIs.

BGP-LS [RFC7752] defines the BGP-LS NLRI that can be a Node NLRI, a Link NLRI or a Prefix NLRI. BGP-LS EPE [I-D.ietf-idr-bgpls-segment-routing-epe] specifies the BGP Protocol ID to be used for signaling EPE information and the same is used for advertising of BGP topology as well.

BGP-LS TE [I-D.ietf-idr-te-lsp-distribution] defines the BGP-LS NLRI that can be used to advertise the RSVP-TE or Segment Routing (SR) policies instantiated on a BGP Router head-end along with their properties and state.

The corresponding BGP-LS attribute is a Node Attribute, a Link Attribute or a Prefix Attribute. BGP-LS [RFC7752] defines the TLVs that map link-state information to BGP-LS NLRI and the BGP-LS attribute.

3.1. Node Advertisements

BGP-LS [RFC7752] defines Node NLRI Type as follows and also defines the Node Descriptor TLVs:

```
  0                   1                   2                   3
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
 +-+-+-+-+-+-+-+-+
 |  Protocol-ID  |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |                           Identifier                          |
 |                            (64 bits)                          |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 //                Local Node Descriptors (variable)            //
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

BGP-LS EPE [I-D.ietf-idr-bgpls-segment-routing-epe] introduces additional Node Descriptor TLVs for BGP protocol for EPE use-case and the same are used by this document.

The following Node Descriptors TLVs MUST appear in the Node NLRI as Local Node Descriptors:

- o BGP Router-ID, which contains the BGP Identifier of the originating BGP router

o Autonomous System Number, which contains the advertising router ASN.

The following Node Attribute TLVs are defined in respective documents to signal the router properties and capabilities (Section 4.1 defines the procedures for their advertisements):

TLV Code Point	Description	Reference Document
1026	Node Name	[RFC7752]
1161	SID/Label	[I-D.ietf-idr-bgp-ls-segment-routing-ext]
1034	SRGB & Capabilities	[I-D.ietf-idr-bgp-ls-segment-routing-ext]
1036	SR Local Block	[I-D.ietf-idr-bgp-ls-segment-routing-ext]
266	Node MSD	[I-D.ietf-idr-bgp-ls-segment-routing-msd]

Table 1: Node Attribute TLVs

3.2. Link Advertisements

BGP-LS [RFC7752] defines Link NLRI Type as follows and also defines the Node and Link Descriptor TLVs used:

```
  0                   1                   2                   3
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
 +-+-+-+-+-+-+-+-+
 |  Protocol-ID  |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |                           Identifier                          |
 |                            (64 bits)                          |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 //               Local Node Descriptors (variable)             //
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 //               Remote Node Descriptors (variable)            //
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 //                  Link Descriptors (variable)                //
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The following Node Descriptors TLVs MUST appear in the Link NLRI as Local Node Descriptors:

- o BGP Router-ID, which contains the BGP Identifier of the originating BGP router

- o Autonomous System Number, which contains the advertising router ASN.

The following Node Descriptors TLVs MUST appear in the Link NLRI as Remote Node Descriptors:

- o BGP Router-ID, which contains the BGP Identifier of the peer BGP router

- o Autonomous System Number, which contains the peer ASN.

The following Link Descriptors TLVs MUST appear in the Link NLRI as Link Descriptors:

- o Link Local/Remote Identifiers containing the 4-octet Link Local Identifier followed by the 4-octet value 0 indicating the Link Remote Identifier is unknown

In addition, the following Link Descriptors TLVs SHOULD appear in the Link NLRI as Link Descriptors based on the address family used for setting up the BGP Peering:

- o IPv4 Interface Address contains the address of the local interface through which the BGP session is established using IPv4 address.

- o IPv6 Interface Address contains the address of the local interface through which the BGP session is established using IPv6 address.

- o IPv4 Neighbor Address contains the IPv4 address of the peer interface used by the BGP session establishment using IPv4 address.

- o IPv6 Neighbor Address contains the IPv6 address of the peer interface used by the BGP session establishment using IPv6 address.

The following Node Attribute TLVs are defined in respective documents which are used to signal the router's local links' properties and capabilities (Section 4.2 defines the procedures for their advertisements) :

TLV Code Point	Description	Reference Document
1088	Administrative group (color)	[RFC7752]
1089	Maximum link bandwidth	[RFC7752]
1092	TE Default Metric	[RFC7752]
1096	SRLG	[RFC7752]
1098	Link Name	[RFC7752]
267	Link MSD	[I-D.ietf-idr-bgp-ls-segment-routing-msd]
1172	L2 Bundle Member	[I-D.ietf-idr-bgp-ls-segment-routing-ext]
1104	Unidirectional link delay	[I-D.ietf-idr-te-pm-bgp]
1105	Min/Max Unidirectional link delay	[I-D.ietf-idr-te-pm-bgp]
1106	Min/Max Unidirectional link delay	[I-D.ietf-idr-te-pm-bgp]
1107	Unidirectional packet loss	[I-D.ietf-idr-te-pm-bgp]
1101	EPE Peer Node SID	[I-D.ietf-idr-bgpls-segment-routing-epe]
1102	EPE Peer Adj SID	[I-D.ietf-idr-bgpls-segment-routing-epe]
1103	EPE Peer Set SID	[I-D.ietf-idr-bgpls-segment-routing-epe]

Table 2: Link Attribute TLVs

3.3. Prefix Advertisements

BGP-LS [RFC7752] defines Prefix NLRI Type as follows and also defines the Node and Prefix Descriptor TLVs used:

```
     0                   1                   2                   3
     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
    +-+-+-+-+-+-+-+-+
    |  Protocol-ID  |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                           Identifier                          |
    |                            (64 bits)                          |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    //              Local Node Descriptors (variable)              //
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    //                Prefix Descriptors (variable)                //
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The following Node Descriptors TLVs MUST appear in the Node NLRI as Local Node Descriptors:

o BGP Router-ID, which contains the BGP Identifier of the originating BGP router

o Autonomous System Number, which contains the advertising router ASN.

The Prefix Descriptor MUST contain the IP Reachability information TLV to identify the prefix.

The following Prefix Attribute TLVs are defined in respective documents which are used to signal the router's own prefix properties and capabilities (Section 4.3 defines the procedures for their advertisements):

TLV Code Point	Description	Reference Document
1158	Prefix SID	[I-D.ietf-idr-bgp-ls-segment-routing-ext]

Table 3: Prefix Attribute TLVs

3.4. TE Policy Advertisements

BGP-LS TE [I-D.ietf-idr-te-lsp-distribution] defines TE Policy NLRI Type as follows and also defines the Headend Node and TE Policy Descriptor TLVs used:

```
     0                   1                   2                   3
     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
    +-+-+-+-+-+-+-+-+
    |  Protocol-ID  |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                           Identifier                          |
    |                           (64 bits)                           |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    //                Headend (Node Descriptors)                   //
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    //                TE Policy Descriptors (variable)             //
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The Node Descriptors TLVs are the same as specified in Section 3.1. The semantics for the TE Policy Descriptor TLVs and the its Attribute TLVs are used as specified in BGP-LS TE [I-D.ietf-idr-te-lsp-distribution].

4. Procedures

In a network where BGP is the only routing protocol, the BGP-LS session is used to advertise the necessary information about the local node properties, its local links' properties and where necessary the prefix's owned by the node. TE Policies, that are instantiated on the local node (i.e. when it is the head-end for the policy), along with their properties are also advertised via the BGP-LS session. This information, once collected across all BGP routers in the network, provides a complete topology view of the network. Many of these attributes are not part of the base BGP protocol operations and are either configured or provided by other components on the router. BGP-LS performs the role of collecting this information and propagating it across the BGP network.

The following sections describe the procedures for the propagation of the BGP-LS NLRIs on a BGP router into the BGP-LS session. These procedures for propagation of BGP topology information via BGP-LS SHOULD be applied only in deployments and use-cases where necessary and SHOULD NOT be applied in every BGP deployment when BGP-LS is enabled. Implementations MAY provide a configuration option to enable these procedures in required deployments.

4.1. Advertisement of Router's Node Attributes

Advertisement of the Node NLRI via BGP-LS by each BGP router in a BGP-only network enables the discovery of all the router nodes in the topology. The Node NLRI MUST be generated by a BGP router only for

itself and even when there are no attributes to be advertised along with it.

The Node attributes defined currently related to SEGMENT ROUTING [I-D.ietf-spring-segment-routing] have been described in Table 1 and are to be advertised when SR is enabled. This includes:

- o The Segment Routing Global Block (SRGB) provisioned on the router which is used by BGP for BGP SR [I-D.ietf-idr-bgp-prefix-sid] and other SR control plane protocols on the router MUST be advertised. The value for Flags field in the TLV is not defined for BGP protocol and MUST be set to 0 by the originator and ignored by receivers.

- o The Segment Routing Local Block (SRLB) provisioned on the router which may be used by BGP for BGP-LS EPE [I-D.ietf-idr-bgpls-segment-routing-epe] for BGP Peering SIDs SHOULD be advertised. The value for Flags field in the TLV is not defined for BGP protocol and MUST be set to 0 by the originator and ignored by receivers.

- o The Node level MSD provides the Node's capabilities for SR SID operations and SHOULD be advertised.

The Node Name Attribute SHOULD be advertised when available.

This document introduces some of the TE concepts into BGP-only networks. However, the advertisement and need for provisioning of a TE Router-ID is not required. The BGP Router-ID along with the ASN provides similar capability for uniquely identifying a BGP router in the network.

Other Node Attributes applicable to a BGP Router may also be included and this document does not describe the exhaustive list.

4.2. Advertisement of Router's Local Links Attributes

Each BGP router in a BGP-only network also advertises its local links using the Link NLRIs thru BGP-LS. The Link NLRI for a given link between two BGP routers is advertised as uni-directional logical "half-link" and its link descriptors allow the correlation between the two NLRIs "half-links" originated by the peering routers to describe the bi-directional logical link and its attributes on both routers.

A Link NLRI MUST be generated by a BGP router for each of its local link over which it is establishing a BGP session to its neighbors. The Link NLRI MUST be generated even when there are no link

attributes to be advertised for it. The Link NLRI represents a peering adjacency between BGP routers and its association with the underlying Layer 3 link. When the underlying Layer 3 link or the BGP session on top of it goes down, the Link NLRI MUST be withdrawn by the BGP router. Advertisement of the Link NLRIs via BGP-LS by each BGP router in a BGP-only network enables the discovery of all the active links in the topology.

The monitoring of links, detecting of their failures and notification to BGP may be performed using mechanisms like BFD. When failures are detected and the BGP session over it goes down, then the corresponding Link NLRI is also withdrawn. This enables faster detection of failures and verification of the underlying links.

The discovery of all the links in the BGP-only network relies on the design that uses EBGP sessions over each interconnecting link using the link IP addresses (refer BGP MSDC [RFC7938]). When doing EBGP multi-hop sessions between directly connected BGP routers, the underlying link information would need to learn by some discovery protocol or provisioning entity. The mechanisms to learn the underlying link information for BGP-LS advertisements are outside the scope of this document. However, to provide a true link topology picture, the advertisement of underlying links is RECOMMENDED for most use-cases instead of a single EBGP peering representation of a link between the routers.

TE attributes for links have been traditionally associated with Link State Routing protocols. However, with the ability to discover the link topology via BGP-LS as specified in this document, the TE attributes and their principles can also be applied to a network running BGP alone. The TE attributes for a link have been described in Table 2 and are to be advertised when TE use-cases are enabled. This includes:

o The maximum bandwidth of a link is its protocol independent attribute and SHOULD be advertised.

o TE concepts of Administrative Groups (also known as affinities) and Shared Risk Link Groups (SRLGs) MAY be provisioned locally on links and then MUST be advertised.

o The BGP base protocol does not operate with link metrics, however, a TE metric concept can be introduced in a BGP only network as well for TE use-cases. Implementations MAY provide the ability to provision TE metric value for a link for BGP use including a different default value for it. The TE metric attribute SHOULD be advertised for each link when configured and its default value is

taken as 1. When not advertised for a link, implementations who intend to use the TE metric MUST assume the value to be 1.

o The delay and loss TE metrics for links are measured via MPLS PerfMon [RFC6374] and their measurement mechanism over a link are independent of the routing protocol. The same mechanism MAY be enabled in BGP-only networks and their values advertised via BGP-LS.

The Link attributes defined currently related to the Segment Routing feature BGP-LS EPE [I-D.ietf-idr-bgpls-segment-routing-epe] have been described in Table 2 and are to be advertised when SR use-cases are enabled. This includes:

o The BGP Peering SIDs provide a functionality similar to Adjacency-SID (refer SEGMENT ROUTING [I-D.ietf-spring-segment-routing]) in BGP-only networks. Implementations SHOULD allocate the BGP Peer-Adjacency SID for all its links and the BGP Peer-Node SID for all its peer routers. Implementations MAY allocate the BGP Peer-Set SID based on local configuration.

o The Link level MSD provides the per link capabilities for SR SID operations and SHOULD be advertised when the router links have differing capabilities.

The use of Layer 3 bundle links which comprise of multiple layer 2 member links are often used in BGP networks. When BGP session is configured over such a layer 3 link, the link attributes of the underlying layer 2 links MAY be advertised individually using the L2 Bundle Member TLV. The applicable attributes for the L2 links are described in BGP-LS SR [I-D.ietf-idr-bgp-ls-segment-routing-ext] .

The Link Name Attribute MAY be advertised when available.

Other Link Attributes applicable to a BGP Router may also be included and this document does not describe the exhaustive list.

4.3. Advertisement of Router's Prefix Attributes

Advertisement of the Prefix NLRI via BGP-LS is required only in specific use-cases. Since the base BGP protocol along with its extensions already signals Prefix reachability via different NLRIs, there is no necessity to always duplicate the information via BGP-LS session. However, for specific use-cases related to SR Traffic Engineering (SR-TE), it is required for each router to advertise it's Prefix SID(s) (refer SEGMENT ROUTING [I-D.ietf-spring-segment-routing]) that can be used to direct traffic via specific BGP routers. Advertising such BGP Prefix SID for every

BGP router provides this key attribute via BGP-LS and avoids the requirement for the consumer of the topology information (e.g. a controller or local TE process) to tap into other BGP NLRI information.

Advertisement of the Prefix NLRI via BGP-LS MUST be done only for its locally configured prefixes (e.g. its loopback interface address) and when BGP is advertising the BGP Prefix SID (BGP SR [I-D.ietf-idr-bgp-prefix-sid]) for it.

The Prefix attributes defined currently related to SEGMENT ROUTING [I-D.ietf-spring-segment-routing] have been described in Table 3 and are to be advertised when SR is enabled. This includes:

- The BGP Prefix SID provisioned on the router for the prefix MUST be advertised. The SID MUST be advertised as the index to be consistent with the Label-Index TLV of BGP Prefix SID attribute. The algorithm is not defined for BGP and MUST be set to 0 by the originator and ignored by the receiver. The flags are defined as the most significant 8 bits of the 16 bit field defined for Label-Index TLV in BGP SR [I-D.ietf-idr-bgp-prefix-sid].

Other Prefix Attributes applicable may also be included and this document does not describe the exhaustive list.

4.4. Advertisement of Router's TE Policy Attributes

TE Policies that are setup using RSVP-TE or SR-TE mechanisms MAY be instantiated on a BGP router. One use-case that results in such SR Policy instantiation on a BGP router is described later in this document in Section 5.2. Advertising such TE Policies instantiated for every BGP router as head-end via BGP-LS provides the consumer of the topology information (e.g. a controller or local TE process) a policy view of the BGP fabric as well.

The procedures for advertisement of the TE Policy NLRI via BGP-LS MUST be done only for its locally instantiated TE Policies and as specified in BGP-LS TE [I-D.ietf-idr-te-lsp-distribution]). Implementation MAY provide configuration options to control the specific set of TE Policies that are to be advertised from the local node.

5. Usage of BGP Topology

This section describes some of the use-cases for the building of the BGP topology information as specified in this document and leveraging it for enabling new functionality.

5.1. Topology View for Monitoring

The BGP-LS advertisement of the BGP topology as specified in this document provides a live topology view of the BGP network for an application or controller that is monitoring the network. The topology view is from the BGP protocol perspective and includes the underlying links as well that aids in network monitoring as well as diagnostics use-cases. BGP-LS is the de-facto protocol for northbound propagation of network topology related information for most IGP networks and extending this capability for BGP-only networks allows existing controllers and applications to consume the information with some incremental BGP protocol awareness.

5.2. SR-TE in BGP Networks

The SR-TE use-case for BGP builds on top of the BGP SR [I-D.ietf-idr-bgp-prefix-sid] functionality and also described in BGP SR MSDC [I-D.ietf-spring-segment-routing-msdc].The BGP SR Prefix SID signaled provides the basic connectivity between all BGP routers using their loopback addresses. This provides the basic best-effort paths in the network using the base BGP decision process that is unchanged. BGP and other overlay routes and services recurse on top of these loopback addresses of the egress nodes and the forwarding is done via the BGP SR Prefix SID labels in the underlay. While this version of the document focuses on the examples with MPLS dataplane instantiation for SR, the same is applicable for the IPv6 dataplane instantiation (SRv6) as well.

SR-TE for BGP provides underlay paths through the network for the overlay routes and services with specific SLA requirements and use-cases like path disjointness, low latency paths, inclusion or exclusion and other TE considerations.

SR-TE [I-D.ietf-spring-segment-routing-policy] specifies the SR-TE architecture and the SR Policy construct. The BGP SR-TE [I-D.ietf-idr-segment-routing-te-policy] describes the extensions to BGP for signaling of SR Policies from a controller to the SR-TE headend BGP router. BGP-LS has been extended to allow signaling of the SR Policies from SR-TE head-end to controllers via BGP-LS SR-TE [I-D.ietf-idr-te-lsp-distribution] which allows the controllers to learn the state of SR Policies instantiated on routers in the network. This document completes the missing piece that is related to getting the BGP topology information from all the routers to a controller as well the local SRTE process on each router for their path computation requirements.

The signaling of SR Polices from controller to SR-TE headend and reporting of the state back to the controller can also be done using

PCEP (PCEP SR [I-D.ietf-pce-segment-routing], PCE Initiated [RFC8281], PCEP Stateful [RFC8231]). However, the BGP topology learning via BGP-LS which is specified in this document is also required for the deployments that uses PCEP in the BGP-only network.

The topology collected via BGP-LS in a BGP-only fabric in a Segment Routing deployment comprise of:

- o The properties of every BGP router node and the Prefix SID to reach that node.

- o The properties of all the links between the BGP routers and the Peer-Adjacency-SIDs (and other EPE SIDs) corresponding to them that allow directing traffic over specific links and/or to specific neighbors.

- o The properties and state of the SR Policies instantiatied on each of the BGP routers along with their end points, their properties and most importantly the Binding SID to steer traffic into the SR Policies.

This topology information allows a computation node to build SR Policies for services over the BGP fabric for a given traffic engineering objective at any given node.

The topology of the BGP fabric is distributed to a centralized controller or application for use-cases that need a centralized computation of SR Policy which can then be signaled to the SR-TE head-end node via PCEP or BGP-SRTE. The topology is also available at any node in the BGP fabric to be used by its local SR-TE process to perform path computation for its own SR Policies for use-cases that are addressed by local computation.

A high level summary of the key topology information advertised via BGP-LS by BGP routers can be used for TE computations as follows

- o The BGP SR Prefix SIDs and the BGP EPE Peering Adjacency SIDs provide the equivalent of the IGP Prefix and Adjacency SIDs and can be used to direct traffic to a specific BGP router and over a specific BGP peer session or link respectively. Traffic for the BGP SR Prefix SIDs follow the path computed by the BGP decision process.

- o The TE administrative group (also known as affinities) and SRLG attributes can be configured over links to enable computation of paths with inclusion and exclusion of specific links or paths that are mutually disjoint.

- o The enabling of link delay and loss measurements and their advertisements can help monitoring the link quality and carve out paths based on latency and other SLA requirements.

- o The signaling of the Node and Link MSD allows controllers to instantiate SR Policies based on the capability of the routers.

This section attempts to highlight and describe at a high level some of the possible SR-TE solutions and use-cases in a BGP-only network. Further details on these use-cases can be found in SR-TE [I-D.ietf-spring-segment-routing-policy].

6. IANA Considerations

None

7. Manageability Considerations

This section is structured as recommended in [RFC5706].

7.1. Operational Considerations

7.1.1. Operations

Existing BGP and BGP-LS operational procedures apply. No additional operation procedures are defined in this document.

8. Security Considerations

Procedures and protocol extensions defined in this document do not affect the BGP security model. See the 'Security Considerations' section of [RFC4271] for a discussion of BGP security. Also refer to [RFC4272] and [RFC6952] for analysis of security issues for BGP.

9. Acknowledgements

10. References

10.1. Normative References

[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-08 (work in progress), May 2018.

[I-D.ietf-idr-bgp-ls-segment-routing-msd]
Tantsura, J., Chunduri, U., Mirsky, G., and S. Sivabalan, "Signaling MSD (Maximum SID Depth) using Border Gateway Protocol Link-State", draft-ietf-idr-bgp-ls-segment-routing-msd-02 (work in progress), August 2018.

[I-D.ietf-idr-bgp-prefix-sid]
Previdi, S., Filsfils, C., Lindem, A., Sreekantiah, A., and H. Gredler, "Segment Routing Prefix SID extensions for BGP", draft-ietf-idr-bgp-prefix-sid-27 (work in progress), June 2018.

[I-D.ietf-idr-bgpls-segment-routing-epe]
Previdi, S., 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-15 (work in progress), March 2018.

[I-D.ietf-idr-te-lsp-distribution]
Previdi, S., Talaulikar, K., Dong, J., Chen, M., Gredler, H., and J. Tantsura, "Distribution of Traffic Engineering (TE) Policies and State using BGP-LS", draft-ietf-idr-te-lsp-distribution-09 (work in progress), June 2018.

[I-D.ietf-idr-te-pm-bgp]
Ginsberg, L., Previdi, S., Wu, Q., Tantsura, J., and C. Filsfils, "BGP-LS Advertisement of IGP Traffic Engineering Performance Metric Extensions", draft-ietf-idr-te-pm-bgp-10 (work in progress), March 2018.

[I-D.ietf-spring-segment-routing]
Filsfils, C., Previdi, S., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", draft-ietf-spring-segment-routing-15 (work in progress), January 2018.

[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>.

[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>.

[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>.

10.2. 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-04 (work in progress), July 2018.

[I-D.ietf-pce-segment-routing]
Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "PCEP Extensions for Segment Routing", draft-ietf-pce-segment-routing-12 (work in progress), June 2018.

[I-D.ietf-spring-segment-routing-central-epe]
Filsfils, C., Previdi, S., Dawra, G., Aries, E., and D. Afanasiev, "Segment Routing Centralized BGP Egress Peer Engineering", draft-ietf-spring-segment-routing-central-epe-10 (work in progress), December 2017.

[I-D.ietf-spring-segment-routing-msdc]
Filsfils, C., Previdi, S., Dawra, G., Aries, E., and P. Lapukhov, "BGP-Prefix Segment in large-scale data centers", draft-ietf-spring-segment-routing-msdc-09 (work in progress), May 2018.

[I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Sivabalan, S., daniel.voyer@bell.ca, d., bogdanov@google.com, b., and P. Mattes, "Segment Routing Policy Architecture", draft-ietf-spring-segment-routing-policy-01 (work in progress), June 2018.

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <https://www.rfc-editor.org/info/rfc2328>.

[RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, DOI 10.17487/RFC3630, September 2003, <https://www.rfc-editor.org/info/rfc3630>.

[RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <https://www.rfc-editor.org/info/rfc4272>.

[RFC5706] Harrington, D., "Guidelines for Considering Operations and Management of New Protocols and Protocol Extensions", RFC 5706, DOI 10.17487/RFC5706, November 2009, <https://www.rfc-editor.org/info/rfc5706>.

[RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, DOI 10.17487/RFC6374, September 2011, <https://www.rfc-editor.org/info/rfc6374>.

[RFC6952] Jethanandani, M., Patel, K., and L. Zheng, "Analysis of BGP, LDP, PCEP, and MSDP Issues According to the Keying and Authentication for Routing Protocols (KARP) Design Guide", RFC 6952, DOI 10.17487/RFC6952, May 2013, <https://www.rfc-editor.org/info/rfc6952>.

[RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, DOI 10.17487/RFC7938, August 2016, <https://www.rfc-editor.org/info/rfc7938>.

[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>.

[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>.

Authors' Addresses

Ketan Talaulikar
Cisco Systems
Pune 411057
India

Email: ketant@cisco.com

Clarence Filsfils
Cisco Systems
Brussels
Belgium

Email: cfilsfil@cisco.com

Krishna Swamy
Cisco Systems
San Jose
USA

Email: kriswamy@cisco.com

Shawn Zandi
LinkedIn
USA

Email: szandi@linkedin.com

Gaurav Dawra
LinkedIn
USA

Email: gdawra.ietf@gmail.com

Muhammad Durrani
Equinix
USA

Email: mdurrani@equinix.com