

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: December 2, 2018

K. Patel  
Arrcus, Inc.  
A. Lindem  
Cisco Systems  
S. Zandi  
Linkedin  
W. Henderickx  
Nokia  
May 31, 2018

Shortest Path Routing Extensions for BGP Protocol  
draft-ietf-lsvr-bgp-spf-01.txt

Abstract

Many Massively Scaled Data Centers (MSDCs) have converged on simplified layer 3 routing. Furthermore, requirements for operational simplicity have lead many of these MSDCs to converge on BGP as their single routing protocol for both their fabric routing and their Data Center Interconnect (DCI) routing. This document describes a solution which leverages BGP Link-State distribution and the Shortest Path First (SPF) algorithm similar to Internal Gateway Protocols (IGPs) such as OSPF.

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 <http://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 December 2, 2018.

Copyright Notice

Copyright (c) 2018 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 (<http://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.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

#### Table of Contents

1.	Introduction . . . . .	3
1.1.	BGP Shortest Path First (SPF) Motivation . . . . .	4
1.2.	Requirements Language . . . . .	5
2.	BGP Peering Models . . . . .	5
2.1.	BGP Single-Hop Peering on Network Node Connections . . . . .	5
2.2.	BGP Peering Between Directly Connected Network Nodes . . . . .	5
2.3.	BGP Peering in Route-Reflector or Controller Topology . . . . .	6
3.	BGP-LS Shortest Path Routing (SPF) SAFI . . . . .	6
4.	Extensions to BGP-LS . . . . .	6
4.1.	Node NLRI Usage and Modifications . . . . .	7
4.2.	Link NLRI Usage . . . . .	7
4.3.	Prefix NLRI Usage . . . . .	8
4.4.	BGP-LS Attribute Sequence-Number TLV . . . . .	8
5.	Decision Process with SPF Algorithm . . . . .	9
5.1.	Phase-1 BGP NLRI Selection . . . . .	10
5.2.	Dual Stack Support . . . . .	10
5.3.	NEXT_HOP Manipulation . . . . .	11
5.4.	IPv4/IPv6 Unicast Address Family Interaction . . . . .	11
5.5.	NLRI Advertisement and Convergence . . . . .	11
5.6.	Error Handling . . . . .	12
6.	IANA Considerations . . . . .	12
7.	Security Considerations . . . . .	12
7.1.	Acknowledgements . . . . .	12
7.2.	Contributors . . . . .	12

8. References . . . . .	13
8.1. Normative References . . . . .	13
8.2. Information References . . . . .	14
Authors' Addresses . . . . .	15

## 1. Introduction

Many Massively Scaled Data Centers (MSDCs) have converged on simplified layer 3 routing. Furthermore, requirements for operational simplicity have lead many of these MSDCs to converge on BGP [RFC4271] as their single routing protocol for both their fabric routing and their Data Center Interconnect (DCI) routing. Requirements and procedures for using BGP are described in [RFC7938]. This document describes an alternative solution which leverages BGP-LS [RFC7752] and the Shortest Path First algorithm similar to Internal Gateway Protocols (IGPs) such as OSPF [RFC2328].

[RFC4271] defines the Decision Process that is used to select routes for subsequent advertisement by applying the policies in the local Policy Information Base (PIB) to the routes stored in its Adj-RIBs-In. The output of the Decision Process is the set of routes that are announced by a BGP speaker to its peers. These selected routes are stored by a BGP speaker in the speaker's Adj-RIBs-Out according to policy.

[RFC7752] describes a mechanism by which link-state and TE information can be collected from networks and shared with external components using BGP. This is achieved by defining NLRI advertised within the BGP-LS/BGP-LS-SPF AFI/SAFI. The BGP-LS extensions defined in [RFC7752] makes use of the Decision Process defined in [RFC4271].

This document augments [RFC7752] by replacing its use of the existing Decision Process. Rather than reusing the BGP-LS SAFI, the BGP-LS-SPF SAFI is introduced to insure backward compatibility. The Phase 1 and 2 decision functions of the Decision Process are replaced with the Shortest Path First (SPF) algorithm also known as the Dijkstra algorithm. The Phase 3 decision function is also simplified since it is no longer dependent on the previous phases. This solution avails the benefits of both BGP and SPF-based IGPs. These include TCP based flow-control, no periodic link-state refresh, and completely incremental NLRI advertisement. These advantages can reduce the overhead in MSDCs where there is a high degree of Equal Cost Multi-Path (ECMPs) and the topology is very stable. Additionally, using a SPF-based computation can support fast convergence and the computation of Loop-Free Alternatives (LFAs) [RFC5286] in the event of link failures. Furthermore, a BGP based solution lends itself to multiple peering models including those incorporating route-reflectors [RFC4456] or controllers.

Support for Multiple Topology Routing (MTR) as described in [RFC4915] is an area for further study dependent on deployment requirements.

#### 1.1. BGP Shortest Path First (SPF) Motivation

Given that [RFC7938] already describes how BGP could be used as the sole routing protocol in an MSDC, one might question the motivation for defining an alternate BGP deployment model when a mature solution exists. For both alternatives, BGP offers the operational benefits of a single routing protocol. However, BGP SPF offers some unique advantages above and beyond standard BGP distance-vector routing.

A primary advantage is that all BGP speakers in the BGP SPF routing domain will have a complete view of the topology. This will allow support for ECMP, IP fast-reroute (e.g., Loop-Free Alternatives), Shared Risk Link Groups (SRLGs), and other routing enhancements without advertisement of addition BGP paths or other extensions. In short, the advantages of an IGP such as OSPF [RFC2328] are availed in BGP.

With the simplified BGP decision process as defined in Section 5.1, NLRI changes can be disseminated throughout the BGP routing domain much more rapidly (equivalent to IGPs with the proper implementation).

Another primary advantage is a potential reduction in NLRI advertisement. With standard BGP distance-vector routing, a single link failure may impact 100s or 1000s prefixes and result in the withdrawal or re-advertisement of the attendant NLRI. With BGP SPF, only the BGP speakers corresponding to the link NLRI need withdraw the corresponding BGP-LS Link NLRI. This advantage will contribute to both faster convergence and better scaling.

With controller and route-reflector peering models, BGP SPF advertisement and distributed computation require a minimal number of sessions and copies of the NLRI since only the latest version of the NLRI from the originator is required. Given that verification of the adjacencies is done outside of BGP (see Section 2), each BGP speaker will only need as many sessions and copies of the NLRI as required for redundancy (e.g., one for the SPF computation and another for backup). Functions such as Optimized Route Reflection (ORR) are supported without extension by virtue of the primary advantages. Additionally, a controller could inject topology that is learned outside the BGP routing domain.

Given that controllers are already consuming BGP-LS NLRI [RFC7752], reusing for the BGP-LS SPF leverages the existing controller implementations.

Another potential advantage of BGP SPF is that both IPv6 and IPv4 can be supported in the same address family using the same topology. Although not described in this version of the document, multi-topology extensions can be used to support separate IPv4, IPv6, unicast, and multicast topologies while sharing the same NLRI.

Finally, the BGP SPF topology can be used as an underlay for other BGP address families (using the existing model) and realize all the above advantages. A simplified peering model using IPv6 link-local addresses as next-hops can be deployed similar to [RFC5549].

## 1.2. 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.

## 2. BGP Peering Models

Depending on the requirements, scaling, and capabilities of the BGP speakers, various peering models are supported. The only requirement is that all BGP speakers in the BGP SPF routing domain receive link-state NLRI on a timely basis, run an SPF calculation, and update their data plane appropriately. The content of the Link NLRI is described in Section 4.2.

### 2.1. BGP Single-Hop Peering on Network Node Connections

The simplest peering model is the one described in section 5.2.1 of [RFC7938]. In this model, EBGP single-hop sessions are established over direct point-to-point links interconnecting the SPF domain nodes. For the purposes of BGP SPF, Link NLRI is only advertised if a single-hop BGP session has been established and the Link-State/SPF address family capability has been exchanged [RFC4790] on the corresponding session. If the session goes down, the corresponding Link NLRI will be withdrawn.

### 2.2. BGP Peering Between Directly Connected Network Nodes

In this model, BGP speakers peer with all directly connected network nodes but the sessions may be multi-hop and the direct connection discovery and liveness detection for those connections are independent of the BGP protocol. How this is accomplished is outside the scope of this document. Consequently, there will be a single session even if there are multiple direct connections between BGP speakers. For the purposes of BGP SPF, Link NLRI is advertised as

long as a BGP session has been established, the Link-State/SPF address family capability has been exchanged [RFC4790] and the corresponding link is considered is up and considered operational.

### 2.3. BGP Peering in Route-Reflector or Controller Topology

In this model, BGP speakers peer solely with one or more Route Reflectors [RFC4456] or controllers. As in the previous model, direct connection discovery and liveness detection for those connections are done outside the BGP protocol. More specifically, the Liveness detection is done using BFD protocol described in [RFC5880]. For the purposes of BGP SPF, Link NLRI is advertised as long as the corresponding link is up and considered operational.

### 3. BGP-LS Shortest Path Routing (SPF) SAFI

In order to replace the Phase 1 and 2 decision functions of the existing Decision Process with an SPF-based Decision Process and streamline the Phase 3 decision functions in a backward compatible manner, this draft introduces the BGP-LS-SFP SAFI for BGP-LS SPF operation. The BGP-LS-SPF (AF 16388 / SAFI TBD1) [RFC4790] is allocated by IANA as specified in the Section 6. A BGP speaker using the BGP-LS SPF extensions described herein MUST exchange the AFI/SAFI using Multiprotocol Extensions Capability Code [RFC4760] with other BGP speakers in the SPF routing domain.

### 4. Extensions to BGP-LS

[RFC7752] describes a mechanism by which link-state and TE information can be collected from networks and shared with external components using BGP protocol. It describes both the definition of BGP-LS NLRI that describes links, nodes, and prefixes comprising IGP link-state information and the definition of a BGP path attribute (BGP-LS attribute) that carries link, node, and prefix properties and attributes, such as the link and prefix metric or auxiliary Router-IDs of nodes, etc.

The BGP protocol will be used in the Protocol-ID field specified in table 1 of [I-D.ietf-idr-bgpls-segment-routing-epe]. The local and remote node descriptors for all NLRI will be the BGP Router-ID (TLV 516) and either the AS Number (TLV 512) [RFC7752] or the BGP Confederation Member (TLV 517) [I-D.ietf-idr-bgpls-segment-routing-epe]. However, if the BGP Router-ID is known to be unique within the BGP Routing domain, it can be used as the sole descriptor.

#### 4.1. Node NLRI Usage and Modifications

The SPF capability is a new Node Attribute TLV that will be added to those defined in table 7 of [RFC7752]. The new attribute TLV will only be applicable when BGP is specified in the Node NLRI Protocol ID field. The TBD TLV type will be defined by IANA. The new Node Attribute TLV will contain a single-octet SPF algorithm as defined in [I-D.ietf-ospf-segment-routing-extensions].

```

      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
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |                                     |
|               Type                 |               Length                 |
+-----+-----+-----+-----+-----+-----+-----+-----+
| SPF Algorithm |
+-----+-----+-----+-----+-----+-----+

```

The SPF Algorithm may take the following values:

- 0 - Normal Shortest Path First (SPF) algorithm based on link metric. This is the standard shortest path algorithm as computed by the IGP protocol. Consistent with the deployed practice for link-state protocols, Algorithm 0 permits any node to overwrite the SPF path with a different path based on its local policy.
- 1 - Strict Shortest Path First (SPF) algorithm based on link metric. The algorithm is identical to Algorithm 0 but Algorithm 1 requires that all nodes along the path will honor the SPF routing decision. Local policy at the node claiming support for Algorithm 1 MUST NOT alter the SPF paths computed by Algorithm 1.

When computing the SPF for a given BGP routing domain, only BGP nodes advertising the SPF capability attribute will be included the Shortest Path Tree (SPT).

#### 4.2. Link NLRI Usage

The criteria for advertisement of Link NLRI are discussed in Section 2.

Link NLRI is advertised with local and remote node descriptors as described above and unique link identifiers dependent on the addressing. For IPv4 links, the links local IPv4 (TLV 259) and remote IPv4 (TLV 260) addresses will be used. For IPv6 links, the local IPv6 (TLV 261) and remote IPv6 (TLV 262) addresses will be used. For unnumbered links, the link local/remote identifiers (TLV 258) will be used. For links supporting having both IPv4 and IPv6

addresses, both sets of descriptors may be included in the same Link NLRI. The link identifiers are described in table 5 of [RFC7752].

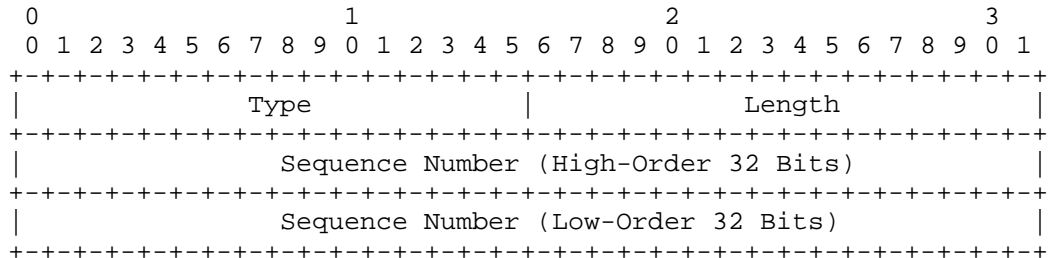
The link IGP metric attribute TLV (TLV 1095) as well as any others required for non-SPF purposes SHOULD be advertised. Algorithms such as setting the metric inversely to the link speed as done in the OSPF MIB [RFC4750] MAY be supported. However, this is beyond the scope of this document.

#### 4.3. Prefix NLRI Usage

Prefix NLRI is advertised with a local node descriptor as described above and the prefix and length used as the descriptors (TLV 265) as described in [RFC7752]. The prefix metric attribute TLV (TLV 1155) as well as any others required for non-SPF purposes SHOULD be advertised. For loopback prefixes, the metric should be 0. For non-loopback prefixes, the setting of the metric is a local matter and beyond the scope of this document.

#### 4.4. BGP-LS Attribute Sequence-Number TLV

A new BGP-LS Attribute TLV to BGP-LS NLRI types is defined to assure the most recent version of a given NLRI is used in the SPF computation. The TBD TLV type will be defined by IANA. The new BGP-LS Attribute TLV will contain an 8-octet sequence number. The usage of the Sequence Number TLV is described in Section 5.1.



#### Sequence Number

The 64-bit strictly increasing sequence number is incremented for every version of BGP-LS NLRI originated. BGP speakers implementing this specification MUST use available mechanisms to preserve the sequence number's strictly increasing property for the deployed life of the BGP speaker (including cold restarts). One mechanism for accomplishing this would be to use the high-order 32 bits of the sequence number as a wrap/boot count that is incremented anytime the BGP router loses its sequence number state or the low-order 32 bits wrap.

When incrementing the sequence number for each self-originated NLRI, the sequence number should be treated as an unsigned 64-bit value. If the lower-order 32-bit value wraps, the higher-order 32-bit value should be incremented and saved in non-volatile storage. If by some chance the BGP Speaker is deployed long enough that there is a possibility that the 64-bit sequence number may wrap or a BGP Speaker completely loses its sequence number state (e.g., the BGP speaker hardware is replaced or experiences a cold-start), the phase 1 decision function (see Section 5.1) rules will insure convergence, albeit, not immediately.

## 5. Decision Process with SPF Algorithm

The Decision Process described in [RFC4271] takes place in three distinct phases. The Phase 1 decision function of the Decision Process is responsible for calculating the degree of preference for each route received from a BGP speaker's peer. The Phase 2 decision function is invoked on completion of the Phase 1 decision function and is responsible for choosing the best route out of all those available for each distinct destination, and for installing each chosen route into the Loc-RIB. The combination of the Phase 1 and 2 decision functions is characterized as a Path Vector algorithm.

The SPF based Decision process replaces the BGP best-path Decision process described in [RFC4271]. This process starts with selecting only those Node NLRI whose SPF capability TLV matches with the local BGP speaker's SPF capability TLV value. Since Link-State NLRI always contains the local descriptor [RFC7752], it will only be originated by a single BGP speaker in the BGP routing domain. These selected Node NLRI and their Link/Prefix NLRI are used to build a directed graph during the SPF computation. The best paths for BGP prefixes are installed as a result of the SPF process.

When BGP-LS-SPF NLRI is received, all that is required is to determine whether it is the best-path by examining the Node-ID and sequence number as described in Section 5.1. If the received best-path NLRI had changed, it will be advertised to other BGP-LS-SPF peers. If the attributes have changed (other than the sequence number), a BGP SPF calculation will be scheduled. However, a changed NLRI MAY be advertised to other peers almost immediately and propagation of changes can approach IGP convergence times. To accomplish this, the MinRouteAdvertisementIntervalTimer and MinRouteAdvertisementIntervalTimer [RFC4271] are not applicable to the BGP-LS-SPF SAFI.

The Phase 3 decision function of the Decision Process [RFC4271] is also simplified since under normal SPF operation, a BGP speaker would advertise the NLRI selected for the SPF to all BGP peers with the

BGP-LS/BGP-LS-SPF AFI/SAFI. Application of policy would not be prevented however its usage to best-path process would be limited as the SPF relies solely on link metrics.

#### 5.1. Phase-1 BGP NLRI Selection

The rules for NLRI selection are greatly simplified from [RFC4271].

1. If the NLRI is received from the BGP speaker originating the NLRI (as determined by the comparing BGP Router ID in the NLRI Node identifiers with the BGP speaker Router ID), then it is preferred over the same NLRI from non-originators. This rule will assure that stale NLRI is updated even if a BGP-LS router loses its sequence number state due to a cold-start.
2. If the Sequence-Number TLV is present in the BGP-LS Attribute, then the NLRI with the most recent, i.e., highest sequence number is selected. BGP-LS NLRI with a Sequence-Number TLV will be considered more recent than NLRI without a BGP-LS Attribute or a BGP-LS Attribute that doesn't include the Sequence-Number TLV.
3. The final tie-breaker is the NLRI from the BGP Speaker with the numerically largest BGP Router ID.

The modified SPF Decision Process performs an SPF calculation rooted at the BGP speaker using the metrics from Link and Prefix NLRI Attribute TLVs [RFC7752]. As a result, any attributes that would influence the Decision process defined in [RFC4271] like ORIGIN, MULTI\_EXIT\_DISC, and LOCAL\_PREF attributes are ignored by the SPF algorithm. Furthermore, the NEXT\_HOP attribute value is preserved but otherwise ignored during the SPF or best-path.

#### 5.2. Dual Stack Support

The SPF-based decision process operates on Node, Link, and Prefix NLRI's that support both IPv4 and IPv6 addresses. Whether to run a single SPF instance or multiple SPF instances for separate AFs is a matter of a local implementation. Normally, IPv4 next-hops are calculated for IPv4 prefixes and IPv6 next-hops are calculated for IPv6 prefixes. However, an interesting use-case is deployment of [RFC5549] where IPv6 next-hops are calculated for both IPv4 and IPv6 prefixes. As stated in Section 1, support for Multiple Topology Routing (MTR) is an area for future study.

### 5.3. NEXT\_HOP Manipulation

A BGP speaker that supports SPF extensions MAY interact with peers that don't support SPF extensions. If the BGP-LS address family is advertised to a peer not supporting the SPF extensions described herein, then the BGP speaker MUST conform to the NEXT\_HOP rules specified in [RFC4271] when announcing the Link-State address family routes to those peers.

All BGP peers that support SPF extensions would locally compute the Loc-RIB next-hops as a result of the SPF process. Consequently, the NEXT\_HOP attribute is always ignored on receipt. However, BGP speakers SHOULD set the NEXT\_HOP address according to the NEXT\_HOP attribute rules specified in [RFC4271].

### 5.4. IPv4/IPv6 Unicast Address Family Interaction

While the BGP-LS SPF address family and the IPv4/IPv6 unicast address families install routes into the same device routing tables, they will operate independently much the same as OSPF and IS-IS would operate today (i.e., "Ships-in-the-Night" mode). There will be no implicit route redistribution between the BGP address families. However, implementation specific redistribution mechanisms SHOULD be made available with the restriction that redistribution of BGP-LS SPF routes into the IPv4 address family applies only to IPv4 routes and redistribution of BGP-LS SPF route into the IPv6 address family applies only to IPv6 routes.

Given the fact that SPF algorithms are based on the assumption that all routers in the routing domain calculate the precisely the same SPF tree and install the same set of routers, it is RECOMMENDED that BGP-LS SPF IPv4/IPv6 routes be given priority by default when installed into their respective RIBs. In common implementations the prioritization is governed by route preference or administrative distance with lower being more preferred.

### 5.5. NLRI Advertisement and Convergence

A local failure will prevent a link from being used in the SPF calculation due to the IGP bi-directional connectivity requirement. Consequently, local link failures should always be given priority over updates (e.g., withdrawing all routes learned on a session) in order to ensure the highest priority propagation and optimal convergence.

Delaying the withdrawal of non-local routes is an area for further study as more IGP-like mechanisms would be required to prevent usage of stale NLRI.

## 5.6. Error Handling

When a BGP speaker receives a BGP Update containing a malformed SPF Capability TLV in the Node NLRI BGP-LS Attribute [RFC7752], it MUST ignore the received TLV and the Node NLRI and not pass it to other BGP peers as specified in [RFC7606]. When discarding a Node NLRI with malformed TLV, a BGP speaker SHOULD log an error for further analysis.

## 6. IANA Considerations

This document defines an AFI/SAFI for BGP-LS SPF operation and requests IANA to assign the BGP-LS/BGP-LS-SPF (AFI 16388 / SAFI TBD1) as described in [RFC4750].

This document also defines two attribute TLV for BGP LS NLRI. We request IANA to assign TLVs for the SPF capability and the Sequence Number from the "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs" Registry.

## 7. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing [RFC4724] and [RFC4271].

### 7.1. Acknowledgements

The authors would like to thank Sue Hares, Jorge Rabadan, and Boris Hassanov for the review and comments.

### 7.2. Contributors

In addition to the authors listed on the front page, the following co-authors have contributed to the document.

Derek Yeung  
Arrcus, Inc.  
derek@arrcus.com

Gunter Van De Velde  
Nokia  
gunter.van\_de\_velde@nokia.com

Abhay Roy  
Cisco Systems  
akr@cisco.com

Venu Venugopal  
Cisco Systems  
venuv@cisco.com

## 8. References

### 8.1. Normative References

- [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-14 (work in progress), December 2017.
- [I-D.ietf-ospf-segment-routing-extensions]  
Psenak, P., Previdi, S., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPF Extensions for Segment Routing", draft-ietf-ospf-segment-routing-extensions-25 (work in progress), April 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>>.
- [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>>.

- [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>>.
- [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>>.
- [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>>.

## 8.2. Information References

- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <<https://www.rfc-editor.org/info/rfc2328>>.
- [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>>.
- [RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, DOI 10.17487/RFC4724, January 2007, <<https://www.rfc-editor.org/info/rfc4724>>.
- [RFC4750] Joyal, D., Ed., Galecki, P., Ed., Giacalone, S., Ed., Coltun, R., and F. Baker, "OSPF Version 2 Management Information Base", RFC 4750, DOI 10.17487/RFC4750, December 2006, <<https://www.rfc-editor.org/info/rfc4750>>.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <<https://www.rfc-editor.org/info/rfc4760>>.
- [RFC4790] Newman, C., Duerst, M., and A. Gulbrandsen, "Internet Application Protocol Collation Registry", RFC 4790, DOI 10.17487/RFC4790, March 2007, <<https://www.rfc-editor.org/info/rfc4790>>.

- [RFC4915] Psenak, P., Mirtorabi, S., Roy, A., Nguyen, L., and P. Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF", RFC 4915, DOI 10.17487/RFC4915, June 2007, <<https://www.rfc-editor.org/info/rfc4915>>.
- [RFC5286] Atlas, A., Ed. and A. Zinin, Ed., "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, DOI 10.17487/RFC5286, September 2008, <<https://www.rfc-editor.org/info/rfc5286>>.
- [RFC5549] Le Faucheur, F. and E. Rosen, "Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop", RFC 5549, DOI 10.17487/RFC5549, May 2009, <<https://www.rfc-editor.org/info/rfc5549>>.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <<https://www.rfc-editor.org/info/rfc5880>>.

## Authors' Addresses

Keyur Patel  
Arrcus, Inc.

Email: [keyur@arrcus.com](mailto:keyur@arrcus.com)

Acee Lindem  
Cisco Systems  
301 Midenhall Way  
Cary, NC 27513  
USA

Email: [acee@cisco.com](mailto:acee@cisco.com)

Shawn Zandi  
Linkedin  
222 2nd Street  
San Francisco, CA 94105  
USA

Email: [szandi@linkedin.com](mailto:szandi@linkedin.com)

Wim Henderickx  
Nokia  
Antwerp  
Belgium

Email: [wim.henderickx@nokia.com](mailto:wim.henderickx@nokia.com)

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: 19 August 2022

K. Patel  
Arrcus, Inc.  
A. Lindem  
Cisco Systems  
S. Zandi  
LinkedIn  
W. Henderickx  
Nokia  
15 February 2022

BGP Link-State Shortest Path First (SPF) Routing  
draft-ietf-lsvr-bgp-spf-16

Abstract

Many Massively Scaled Data Centers (MSDCs) have converged on simplified layer 3 routing. Furthermore, requirements for operational simplicity have led many of these MSDCs to converge on BGP as their single routing protocol for both their fabric routing and their Data Center Interconnect (DCI) routing. This document describes extensions to BGP to use BGP Link-State distribution and the Shortest Path First (SPF) algorithm used by Internal Gateway Protocols (IGPs) such as OSPF. In doing this, it allows BGP to be efficiently used as both the underlay protocol and the overlay protocol in MSDCs.

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 19 August 2022.

Copyright Notice

Copyright (c) 2022 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 Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

## Table of Contents

1. Introduction . . . . .	3
1.1. Terminology . . . . .	4
1.2. BGP Shortest Path First (SPF) Motivation . . . . .	4
1.3. Document Overview . . . . .	6
1.4. Requirements Language . . . . .	6
2. Base BGP Protocol Relationship . . . . .	6
3. BGP Link-State (BGP-LS) Relationship . . . . .	7
4. BGP Peering Models . . . . .	8
4.1. BGP Single-Hop Peering on Network Node Connections . . . . .	8
4.2. BGP Peering Between Directly-Connected Nodes . . . . .	8
4.3. BGP Peering in Route-Reflector or Controller Topology . . . . .	9
5. BGP Shortest Path Routing (SPF) Protocol Extensions . . . . .	9
5.1. BGP-LS Shortest Path Routing (SPF) SAFI . . . . .	9
5.1.1. BGP-LS-SPF NLRI TLVs . . . . .	9
5.1.2. BGP-LS Attribute . . . . .	10
5.2. Extensions to BGP-LS . . . . .	11
5.2.1. Node NLRI Usage . . . . .	11
5.2.1.1. BGP-LS-SPF Node NLRI Attribute SPF Capability TLV . . . . .	11
5.2.1.2. BGP-LS-SPF Node NLRI Attribute SPF Status TLV . . . . .	12
5.2.2. Link NLRI Usage . . . . .	13
5.2.2.1. BGP-LS-SPF Link NLRI Attribute Prefix-Length TLVs . . . . .	14
5.2.2.2. BGP-LS-SPF Link NLRI Attribute SPF Status TLV . . . . .	15
5.2.3. IPv4/IPv6 Prefix NLRI Usage . . . . .	16
5.2.3.1. BGP-LS-SPF Prefix NLRI Attribute SPF Status TLV . . . . .	16
5.2.4. BGP-LS Attribute Sequence-Number TLV . . . . .	17
5.3. NEXT_HOP Manipulation . . . . .	18
6. Decision Process with SPF Algorithm . . . . .	18
6.1. BGP NLRI Selection . . . . .	19
6.1.1. BGP Self-Originated NLRI . . . . .	20
6.2. Dual Stack Support . . . . .	21
6.3. SPF Calculation based on BGP-LS-SPF NLRI . . . . .	21
6.4. IPv4/IPv6 Unicast Address Family Interaction . . . . .	26
6.5. NLRI Advertisement . . . . .	26
6.5.1. Link/Prefix Failure Convergence . . . . .	26

6.5.2. Node Failure Convergence . . . . .	27
7. Error Handling . . . . .	27
7.1. Processing of BGP-LS-SPF TLVs . . . . .	27
7.2. Processing of BGP-LS-SPF NLRIs . . . . .	28
7.3. Processing of BGP-LS Attribute . . . . .	29
8. IANA Considerations . . . . .	30
9. Security Considerations . . . . .	31
10. Management Considerations . . . . .	32
10.1. Configuration . . . . .	32
10.1.1. Link Metric Configuration . . . . .	32
10.1.2. backoff-config . . . . .	32
10.2. Operational Data . . . . .	33
11. Implementation Status . . . . .	33
12. Acknowledgements . . . . .	34
13. Contributors . . . . .	34
14. References . . . . .	34
14.1. Normative References . . . . .	34
14.2. Informational References . . . . .	36
Authors' Addresses . . . . .	38

## 1. Introduction

Many Massively Scaled Data Centers (MSDCs) have converged on simplified layer 3 routing. Furthermore, requirements for operational simplicity have led many of these MSDCs to converge on BGP [RFC4271] as their single routing protocol for both their fabric routing and their Data Center Interconnect (DCI) routing [RFC7938]. This document describes an alternative solution which leverages BGP-LS [RFC7752] and the Shortest Path First algorithm used by Internal Gateway Protocols (IGPs) such as OSPF [RFC2328].

This document leverages both the BGP protocol [RFC4271] and the BGP-LS [RFC7752] protocols. The relationship, as well as the scope of changes are described respectively in Section 2 and Section 3. The modifications to [RFC4271] for BGP SPF described herein only apply to IPv4 and IPv6 as underlay unicast Subsequent Address Families Identifiers (SAFIs). Operations for any other BGP SAFIs are outside the scope of this document.

This solution avails the benefits of both BGP and SPF-based IGPs. These include TCP based flow-control, no periodic link-state refresh, and completely incremental NLRI advertisement. These advantages can reduce the overhead in MSDCs where there is a high degree of Equal Cost Multi-Path (ECMPs) and the topology is very stable. Additionally, using an SPF-based computation can support fast convergence and the computation of Loop-Free Alternatives (LFAs). The SPF LFA extensions defined in [RFC5286] can be similarly applied to BGP SPF calculations. However, the details are a matter of

implementation detail. Furthermore, a BGP-based solution lends itself to multiple peering models including those incorporating route-reflectors [RFC4456] or controllers.

### 1.1. Terminology

This specification reuses terms defined in section 1.1 of [RFC4271] including BGP speaker, NLRI, and Route.

Additionally, this document introduces the following terms:

**BGP SPF Routing Domain:** A set of BGP routers that are under a single administrative domain and exchange link-state information using the BGP-LS-SPF SAFI and compute routes using BGP SPF as described herein.

**BGP-LS-SPF NLRI:** This refers to BGP-LS Network Layer Reachability Information (NLRI) that is being advertised in the BGP-LS-SPF SAFI (Section 5.1) and is being used for BGP SPF route computation.

**Dijkstra Algorithm:** An algorithm for computing the shortest path from a given node in a graph to every other node in the graph. At each iteration of the algorithm, there is a list of candidate vertices. Paths from the root to these vertices have been found, but not necessarily the shortest ones. However, the paths to the candidate vertex that is closest to the root are guaranteed to be shortest; this vertex is added to the shortest-path tree, removed from the candidate list, and its adjacent vertices are examined for possible addition to/modification of the candidate list. The algorithm then iterates again. It terminates when the candidate list becomes empty. [RFC2328]

### 1.2. BGP Shortest Path First (SPF) Motivation

Given that [RFC7938] already describes how BGP could be used as the sole routing protocol in an MSDC, one might question the motivation for defining an alternate BGP deployment model when a mature solution exists. For both alternatives, BGP offers the operational benefits of a single routing protocol as opposed to the combination of an IGP for the underlay and BGP as an overlay. However, BGP SPF offers some unique advantages above and beyond standard BGP distance-vector routing. With BGP SPF, the standard hop-by-hop peering model is relaxed.

A primary advantage is that all BGP SPF speakers in the BGP SPF routing domain will have a complete view of the topology. This will allow support for ECMP, IP fast-reroute (e.g., Loop-Free Alternatives), Shared Risk Link Groups (SRLGs), and other routing

enhancements without advertisement of additional BGP paths [RFC7911] or other extensions. In short, the advantages of an IGP such as OSPF [RFC2328] are availed in BGP.

With the simplified BGP decision process as defined in Section 6, NLRI changes can be disseminated throughout the BGP routing domain much more rapidly (equivalent to IGPs with the proper implementation). The added advantage of BGP using TCP for reliable transport leverages TCP's inherent flow-control and guaranteed in-order delivery.

Another primary advantage is a potential reduction in NLRI advertisement. With standard BGP distance-vector routing, a single link failure may impact 100s or 1000s prefixes and result in the withdrawal or re-advertisement of the attendant NLRI. With BGP SPF, only the BGP SPF speakers corresponding to the link NLRI need to withdraw the corresponding BGP-LS-SPF Link NLRI. Additionally, the changed NLRI will be advertised immediately as opposed to normal BGP where it is only advertised after the best route selection. These advantages will afford NLRI dissemination throughout the BGP SPF routing domain with efficiencies similar to link-state protocols.

With controller and route-reflector peering models, BGP SPF advertisement and distributed computation require a minimal number of sessions and copies of the NLRI since only the latest version of the NLRI from the originator is required. Given that verification of the adjacencies is done outside of BGP (see Section 4), each BGP SPF speaker will only need as many sessions and copies of the NLRI as required for redundancy (see Section 4). Additionally, a controller could inject topology that is learned outside the BGP SPF routing domain.

Given that controllers are already consuming BGP-LS NLRI [RFC7752], this functionality can be reused for BGP-LS-SPF NLRI.

Another advantage of BGP SPF is that both IPv6 and IPv4 can be supported using the BGP-LS-SPF SAFI with the same BGP-LS-SPF NRIs. In many MSDC fabrics, the IPv4 and IPv6 topologies are congruent, refer to Section 5.2.2 and Section 5.2.3. Although beyond the scope of this document, multi-topology extensions could be used to support separate IPv4, IPv6, unicast, and multicast topologies while sharing the same NLRI.

Finally, the BGP SPF topology can be used as an underlay for other BGP SAFIs (using the existing model) and realize all the above advantages.

### 1.3. Document Overview

The document begins with sections defining the precise relationship that BGP SPF has with both the base BGP protocol [RFC4271] (Section 2) and the BGP Link-State (BGP-LS) extensions [RFC7752] (Section 3). This is required to dispel the notion that BGP SPF is an independent protocol. The BGP peering models, as well as the their respective trade-offs are then discussed in Section 4. The remaining sections, which make up the bulk of the document, define the protocol enhancements necessary to support BGP SPF. The BGP-LS extensions to support BGP SPF are defined in Section 5. The replacement of the base BGP decision process with the SPF computation is specified in Section 6. Finally, BGP SPF error handling is defined in Section 7

### 1.4. 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.

## 2. Base BGP Protocol Relationship

With the exception of the decision process, the BGP SPF extensions leverage the BGP protocol [RFC4271] without change. This includes the BGP protocol Finite State Machine, BGP messages and their encodings, processing of BGP messages, BGP attributes and path attributes, BGP NLRI encodings, and any error handling defined in the [RFC4271] and [RFC7606].

Due to the changes to the decision process, there are mechanisms and encodings that are no longer applicable. While not necessarily required for computation, the ORIGIN, AS\_PATH, MULTI\_EXIT\_DISC, LOCAL\_PREF, and NEXT\_HOP path attributes are mandatory and will be validated. The ATOMIC\_AGGEGATE, and AGGREGATOR are not applicable within the context of BGP SPF and SHOULD NOT be advertised. However, if they are advertised, they will be accepted, validated, and propagated consistent with the BGP protocol.

Section 9 of [RFC4271] defines the decision process that is used to select routes for subsequent advertisement by applying the policies in the local Policy Information Base (PIB) to the routes stored in its Adj-RIBs-In. The output of the Decision Process is the set of routes that are announced by a BGP speaker to its peers. These selected routes are stored by a BGP speaker in the speaker's Adj-RIBs-Out according to policy.

The BGP SPF extension fundamentally changes the decision process, as described herein, to be more like a link-state protocol (e.g., OSPF [RFC2328]). Specifically:

1. BGP advertisements are readvertised to neighbors immediately without waiting or dependence on the route computation as specified in phase 3 of the base BGP decision process. Multiple peering models are supported as specified in Section 4.
2. Determining the degree of preference for BGP routes for the SPF calculation as described in phase 1 of the base BGP decision process is replaced with the mechanisms in Section 6.1.
3. Phase 2 of the base BGP protocol decision process is replaced with the Shortest Path First (SPF) algorithm, also known as the Dijkstra algorithm Section 1.1.

### 3. BGP Link-State (BGP-LS) Relationship

[RFC7752] describes a mechanism by which link-state and TE information can be collected from networks and shared with external entities using BGP. This is achieved by defining NLRI advertised using the BGP-LS AFI. The BGP-LS extensions defined in [RFC7752] make use of the decision process defined in [RFC4271]. This document reuses NLRI and TLVs defined in [RFC7752]. Rather than reusing the BGP-LS SAFI, the BGP-LS-SPF SAFI Section 5.1 is introduced to insure backward compatibility for the BGP-LS SAFI usage.

The BGP SPF extensions reuse the Node, Link, and Prefix NLRI defined in [RFC7752]. The usage of the BGP-LS NLRI, attributes, and attribute extensions is described in Section 5.2. The usage of others BGP-LS attributes is not precluded and is, in fact, expected. However, the details are beyond the scope of this document and will be specified in future documents.

Support for Multiple Topology Routing (MTR) similar to the OSPF MTR computation described in [RFC4915] is beyond the scope of this document. Consequently, the usage of the Multi-Topology TLV as described in section 3.2.1.5 of [RFC7752] is not specified.

The rules for setting the NLRI next-hop path attribute for the BGP-LS-SPF SAFI will follow the BGP-LS SAFI as specified in section 3.4 of [RFC7752].

#### 4. BGP Peering Models

Depending on the topology, scaling, capabilities of the BGP SPF speakers, and redundancy requirements, various peering models are supported. The only requirements are that all BGP SPF speakers in the BGP SPF routing domain exchange BGP-LS-SPF NLRI, run an SPF calculation, and update their routing table appropriately.

##### 4.1. BGP Single-Hop Peering on Network Node Connections

The simplest peering model is the one where EBGp single-hop sessions are established over direct point-to-point links interconnecting the nodes in the BGP SPF routing domain. Once the single-hop BGP session has been established and the BGP-LS-SPF AFI/SAFI capability has been exchanged [RFC4760] for the corresponding session, then the link is considered up from a BGP SPF perspective and the corresponding BGP-LS-SPF Link NLRI is advertised. If the session goes down, the corresponding Link NLRI will be withdrawn. Topologically, this would be equivalent to the peering model in [RFC7938] where there is a BGP session on every link in the data center switch fabric. The content of the Link NLRI is described in Section 5.2.2.

##### 4.2. BGP Peering Between Directly-Connected Nodes

In this model, BGP SPF speakers peer with all directly-connected nodes but the sessions may be between loopback addresses (i.e., two-hop sessions) and the direct connection discovery and liveliness detection for the interconnecting links are independent of the BGP protocol. For example, liveliness detection could be done using the BFD protocol [RFC5880]. Precisely how discovery and liveliness detection is accomplished is outside the scope of this document. Consequently, there will be a single BGP session even if there are multiple direct connections between BGP SPF speakers. BGP-LS-SPF Link NLRI is advertised as long as a BGP session has been established, the BGP-LS-SPF AFI/SAFI capability has been exchanged [RFC4760], and the link is operational as determined using liveliness detection mechanisms outside the scope of this document. This is much like the previous peering model only peering is between loopback addresses and the interconnecting links can be unnumbered. However, since there are BGP sessions between every directly-connected node in the BGP SPF routing domain, there is only a reduction in BGP sessions when there are parallel links between nodes.

#### 4.3. BGP Peering in Route-Reflector or Controller Topology

In this model, BGP SPF speakers peer solely with one or more Route Reflectors [RFC4456] or controllers. As in the previous model, direct connection discovery and liveness detection for those links in the BGP SPF routing domain are done outside of the BGP protocol. BGP-LS-SPF Link NLRI is advertised as long as the corresponding link is considered up as per the chosen liveness detection mechanism.

This peering model, known as sparse peering, allows for fewer BGP sessions and, consequently, fewer instances of the same NLRI received from multiple peers. Normally, the route-reflectors or controller BGP sessions would be on directly-connected links to avoid dependence on another routing protocol for session connectivity. However, multi-hop peering is not precluded. The number of BGP sessions is dependent on the redundancy requirements and the stability of the BGP sessions. This is discussed in greater detail in [I-D.ietf-lsvr-applicability].

### 5. BGP Shortest Path Routing (SPF) Protocol Extensions

#### 5.1. BGP-LS Shortest Path Routing (SPF) SAFI

In order to replace the existing BGP decision process with an SPF-based decision process in a backward compatible manner by not impacting the BGP-LS SAFI, this document introduces the BGP-LS-SPF SAFI. The BGP-LS-SPF (AFI 16388 / SAFI 80) [RFC4760] is allocated by IANA as specified in the Section 8. In order for two BGP SPF speakers to exchange BGP SPF NLRI, they MUST exchange the Multiprotocol Extensions Capability [RFC5492] [RFC4760] to ensure that they are both capable of properly processing such NLRI. This is done with AFI 16388 / SAFI 80 for BGP-LS-SPF advertised within the BGP SPF Routing Domain. The BGP-LS-SPF SAFI is used to carry IPv4 and IPv6 prefix information in a format facilitating an SPF-based decision process.

##### 5.1.1. BGP-LS-SPF NLRI TLVs

The NLRI format of BGP-LS-SPF SAFI uses exactly same format as the BGP-LS AFI [RFC7752]. In other words, all the TLVs used in BGP-LS AFI are applicable and used for the BGP-LS-SPF SAFI. These TLVs within BGP-LS-SPF NLRI advertise information that describes links, nodes, and prefixes comprising IGP link-state information.

In order to compare the NLRI efficiently, it is REQUIRED that all the TLVs within the given NLRI must be ordered in ascending order by the TLV type. For multiple TLVs of same type within a single NLRI, it is REQUIRED that these TLVs are ordered in ascending order by the TLV

value field. Comparison of the value fields is performed by treating the entire value field as a hexadecimal string. NLRI's having TLVs which do not follow the ordering rules MUST be considered as malformed and discarded with appropriate error logging.

[RFC7752] defines certain NLRI TLVs as a mandatory TLVs. These TLVs are considered mandatory for the BGP-LS-SPF SAFI as well. All the other TLVs are considered as an optional TLVs.

Given that there is a single BGP-LS Attribute for all the BGP-LS-SPF NLRI in a BGP Update, Section 3.3, [RFC7752], a BGP Update will normally contain a single BGP-LS-SPF NLRI since advertising multiple NLRI would imply identical attributes.

#### 5.1.2. BGP-LS Attribute

The BGP-LS attribute of the BGP-LS-SPF SAFI uses exactly same format of the BGP-LS AFI [RFC7752]. In other words, all the TLVs used in BGP-LS attribute of the BGP-LS AFI are applicable and used for the BGP-LS attribute of the BGP-LS-SPF SAFI. This attribute is an optional, non-transitive BGP attribute that is used to carry link, node, and prefix properties and attributes. The BGP-LS attribute is a set of TLVs.

The BGP-LS attribute may potentially grow large in size depending on the amount of link-state information associated with a single Link-State NLRI. The BGP specification [RFC4271] mandates a maximum BGP message size of 4096 octets. It is RECOMMENDED that an implementation support [RFC8654] in order to accommodate larger size of information within the BGP-LS Attribute. BGP SPF speakers MUST ensure that they limit the TLVs included in the BGP-LS Attribute to ensure that a BGP update message for a single Link-State NLRI does not cross the maximum limit for a BGP message. The determination of the types of TLVs to be included by the BGP SPF speaker originating the attribute is outside the scope of this document. When a BGP SPF speaker finds that it is exceeding the maximum BGP message size due to addition or update of some other BGP Attribute (e.g., AS\_PATH), it MUST consider the BGP-LS Attribute to be malformed and the attribute discard handling of [RFC7606] applies.

In order to compare the BGP-LS attribute efficiently, it is REQUIRED that all the TLVs within the given attribute must be ordered in ascending order by the TLV type. For multiple TLVs of same type within a single attribute, it is REQUIRED that these TLVs are ordered in ascending order by the TLV value field. Comparison of the value fields is performed by treating the entire value field as a hexadecimal string. Attributes having TLVs which do not follow the ordering rules MUST NOT be considered as malformed.

All TLVs within the BGP-LS Attribute are considered optional unless specified otherwise.

## 5.2. Extensions to BGP-LS

[RFC7752] describes a mechanism by which link-state and TE information can be collected from IGPs and shared with external components using the BGP protocol. It describes both the definition of the BGP-LS NLRI that advertise links, nodes, and prefixes comprising IGP link-state information and the definition of a BGP path attribute (BGP-LS attribute) that carries link, node, and prefix properties and attributes, such as the link and prefix metric or auxiliary Router-IDs of nodes, etc. This document extends the usage of BGP-LS NLRI for the purpose of BGP SPF calculation via advertisement in the BGP-LS-SPF SAFI.

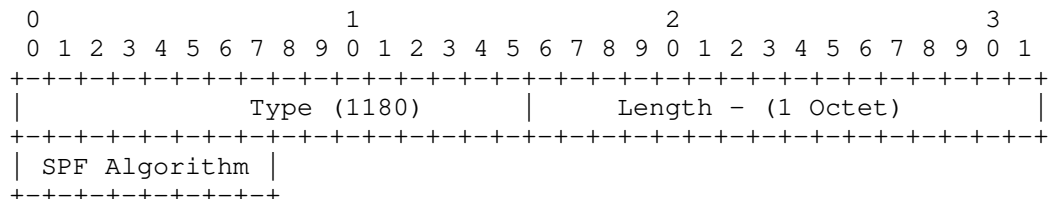
The protocol identifier specified in the Protocol-ID field [RFC7752] will represent the origin of the advertised NLRI. For Node NLRI and Link NLRI, this MUST be the direct protocol (4). Node or Link NLRI with a Protocol-ID other than direct will be considered malformed. For Prefix NLRI, the specified Protocol-ID MUST be the origin of the prefix. The local and remote node descriptors for all NLRI MUST include the BGP Identifier (TLV 516) and the AS Number (TLV 512) [RFC7752]. The BGP Confederation Member (TLV 517) [RFC7752] is not applicable and SHOULD not be included. If TLV 517 is included, it will be ignored.

### 5.2.1. Node NLRI Usage

The Node NLRI MUST be advertised unconditionally by all routers in the BGP SPF routing domain.

#### 5.2.1.1. BGP-LS-SPF Node NLRI Attribute SPF Capability TLV

The SPF capability is an additional Node Attribute TLV. This attribute TLV MUST be included with the BGP-LS-SPF SAFI and SHOULD NOT be used for other SAFIs. The TLV type 1180 will be assigned by IANA. The Node Attribute TLV will contain a single-octet SPF algorithm as defined in [RFC8665].



The SPF algorithm inherits the values from the IGP Algorithm Types registry [RFC8665]. Algorithm 0, (Shortest Path Algorithm (SPF) based on link metric, is supported and described in Section 6.3. Support for other algorithm types is beyond the scope of this specification.

When computing the SPF for a given BGP routing domain, only BGP nodes advertising the SPF capability TLV with same SPF algorithm will be included in the Shortest Path Tree (SPT) Section 6.3. An implementation MAY optionally log detection of a BGP node that has either not advertised the SPF capability TLV or is advertising the SPF capability TLV with an algorithm type other than 0.

#### 5.2.1.2. BGP-LS-SPF Node NLRI Attribute SPF Status TLV

A BGP-LS Attribute TLV of the BGP-LS-SPF Node NLRI is defined to indicate the status of the node with respect to the BGP SPF calculation. This will be used to rapidly take a node out of service Section 6.5.2 or to indicate the node is not to be used for transit (i.e., non-local) traffic Section 6.3. If the SPF Status TLV is not included with the Node NLRI, the node is considered to be up and is available for transit traffic. The SPF status is acted upon with the execution of the next SPF calculation Section 6.3. A single TLV type will be shared by the BGP-LS-SPF Node, Link, and Prefix NLRI. The TLV type 1184 will be assigned by IANA.

```

      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
+-----+-----+-----+-----+-----+-----+-----+-----+
|  Type (1184)  |          Length (1 Octet)          |
+-----+-----+-----+-----+-----+-----+-----+
|  SPF Status   |
+-----+-----+-----+-----+-----+-----+

```

BGP Status Values: 0 - Reserved  
                   1 - Node Unreachable with respect to BGP SPF  
                   2 - Node does not support transit with respect  
                       to BGP SPF  
                   3-254 - Undefined  
                   255 - Reserved

The BGP-LS-SPF Node Attribute SPF Status TLV, Link Attribute SPF Status TLV, and Prefix Attribute SPF Status TLV use the same TLV Type (1184). This implies that a BGP Update cannot contain multiple NLRI with differing status. If the BGP-LS-SPF Status TLV is advertised and the advertised value is not defined for all NLRI included in the BGP update, then the SPF Status TLV is ignored and not used in SPF computation but is still announced to other BGP SPF speakers. An implementation MAY log an error for further analysis.

If a BGP SPF speaker received the Node NLRI but the SPF Status TLV is not received, then any previously received information is considered as implicitly withdrawn and the update is propagated to other BGP SPF speakers. A BGP SPF speaker receiving a BGP Update containing a SPF Status TLV in the BGP-LS attribute [RFC7752] with a value that is outside the range of defined values SHOULD be processed and announced to other BGP SPF speakers. However, a BGP SPF speaker MUST NOT use the Status TLV in its SPF computation. An implementation MAY log this condition for further analysis.

#### 5.2.2. Link NLRI Usage

The criteria for advertisement of Link NLRI are discussed in Section 4.

Link NLRI is advertised with unique local and remote node descriptors dependent on the IP addressing. For IPv4 links, the link's local IPv4 (TLV 259) and remote IPv4 (TLV 260) addresses will be used. For IPv6 links, the local IPv6 (TLV 261) and remote IPv6 (TLV 262) addresses will be used. For unnumbered links, the link local/remote identifiers (TLV 258) will be used. For links supporting having both IPv4 and IPv6 addresses, both sets of descriptors MAY be included in the same Link NLRI. The link identifiers are described in table 5 of [RFC7752].

For a link to be used in Shortest Path Tree (SPT) for a given address family, i.e., IPv4 or IPv6, both routers connecting the link MUST have an address in the same subnet for that address family. However, an IPv4 or IPv6 prefix associated with the link MAY be installed without the corresponding address on the other side of link.

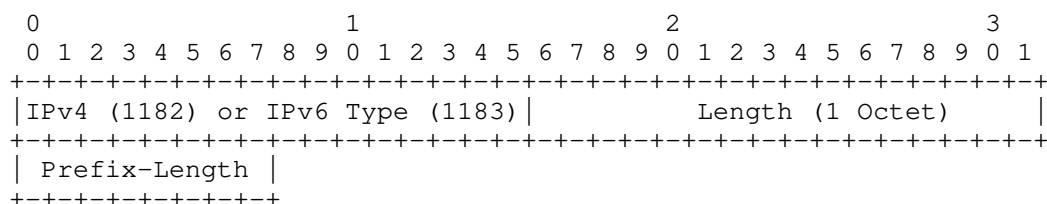
The link IGP metric attribute TLV (TLV 1095) MUST be advertised. If a BGP SPF speaker receives a Link NLRI without an IGP metric attribute TLV, then it SHOULD consider the received NLRI as a malformed and the receiving BGP SPF speaker MUST handle such malformed NLRI as 'Treat-as-withdraw' [RFC7606]. The BGP SPF metric length is 4 octets. Like OSPF [RFC2328], a cost is associated with the output side of each router interface. This cost is configurable by the system administrator. The lower the cost, the more likely the

interface is to be used to forward data traffic. One possible default for metric would be to give each interface a cost of 1 making it effectively a hop count. Algorithms such as setting the metric inversely to the link speed as supported in the OSPF MIB [RFC4750] MAY be supported. However, this is beyond the scope of this document. Refer to Section 10.1.1 for operational guidance.

The usage of other link attribute TLVs is beyond the scope of this document.

#### 5.2.2.1. BGP-LS-SPF Link NLRI Attribute Prefix-Length TLVs

Two BGP-LS Attribute TLVs of the BGP-LS-SPF Link NLRI are defined to advertise the prefix length associated with the IPv4 and IPv6 link prefixes derived from the link descriptor addresses. The prefix length is used for the optional installation of prefixes corresponding to Link NLRI as defined in Section 6.3.



Prefix-length - A one-octet length restricted to 1-32 for IPv4  
Link NLRI endpoint prefixes and 1-128 for IPv6  
Link NLRI endpoint prefixes.

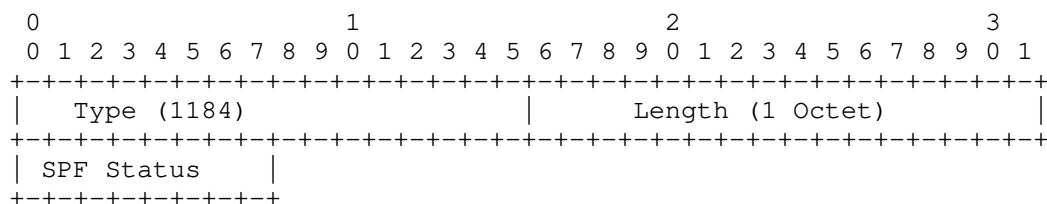
The Prefix-Length TLV is only relevant to Link NLRIs. The Prefix-Length TLVs MUST be discarded as an error and not passed to other BGP peers as specified in [RFC7606] when received with any NLRIs other than Link NLRIs. An implementation MAY log an error for further analysis.

The maximum prefix-length for IPv4 Prefix-Length TLV is 32 bits. A prefix-length field indicating a larger value than 32 bits MUST be discarded as an error and the received TLV is not passed to other BGP peers as specified in [RFC7606]. The corresponding Link NLRI is considered as malformed and MUST be handled as 'Treat-as-withdraw'. An implementation MAY log an error for further analysis.

The maximum prefix-length for IPv6 Prefix-Length Type is 128 bits. A prefix-length field indicating a larger value than 128 bits MUST be discarded as an error and the received TLV is not passed to other BGP peers as specified in [RFC7606]. The corresponding Link NLRI is considered as malformed and MUST be handled as 'Treat-as-withdraw'. An implementation MAY log an error for further analysis.

#### 5.2.2.2. BGP-LS-SPF Link NLRI Attribute SPF Status TLV

A BGP-LS Attribute TLV of the BGP-LS-SPF Link NLRI is defined to indicate the status of the link with respect to the BGP SPF calculation. This will be used to expedite convergence for link failures as discussed in Section 6.5.1. If the SPF Status TLV is not included with the Link NLRI, the link is considered up and available. The SPF status is acted upon with the execution of the next SPF calculation Section 6.3. A single TLV type will be shared by the Node, Link, and Prefix NLRI. The TLV type 1184 will be assigned by IANA.



BGP Status Values: 0 - Reserved  
 1 - Link Unreachable with respect to BGP SPF  
 2-254 - Undefined  
 255 - Reserved

The BGP-LS-SPF Node Attribute SPF Status TLV, Link Attribute SPF Status TLV, and Prefix Attribute SPF Status TLV use the same TLV Type (1184). This implies that a BGP Update cannot contain multiple NLRI with differing status. If the BGP-LS-SPF Status TLV is advertised and the advertised value is not defined for all NLRI included in the BGP update, then the SPF Status TLV is ignored and not used in SPF computation but is still announced to other BGP SPF speakers. An implementation MAY log an error for further analysis.

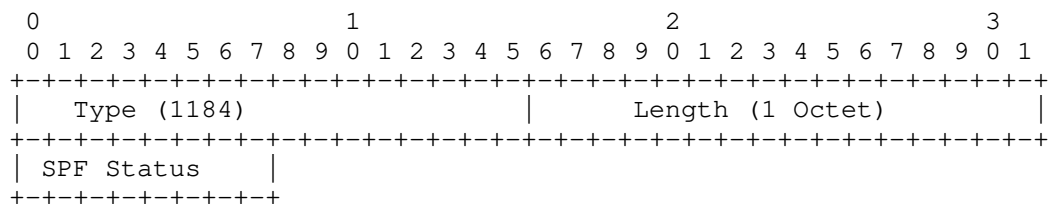
If a BGP SPF speaker received the Link NLRI but the SPF Status TLV is not received, then any previously received information is considered as implicitly withdrawn and the update is propagated to other BGP SPF speakers. A BGP SPF speaker receiving a BGP Update containing an SPF Status TLV in the BGP-LS attribute [RFC7752] with a value that is outside the range of defined values SHOULD be processed and announced to other BGP SPF speakers. However, a BGP SPF speaker MUST NOT use the Status TLV in its SPF computation. An implementation MAY log this information for further analysis.

### 5.2.3. IPv4/IPv6 Prefix NLRI Usage

IPv4/IPv6 Prefix NLRI is advertised with a Local Node Descriptor and the prefix and length. The Prefix Descriptors field includes the IP Reachability Information TLV (TLV 265) as described in [RFC7752]. The Prefix Metric attribute TLV (TLV 1155) MUST be advertised. The IGP Route Tag TLV (TLV 1153) MAY be advertised. The usage of other attribute TLVs is beyond the scope of this document. For loopback prefixes, the metric should be 0. For non-loopback prefixes, the setting of the metric is a local matter and beyond the scope of this document.

#### 5.2.3.1. BGP-LS-SPF Prefix NLRI Attribute SPF Status TLV

A BGP-LS Attribute TLV to BGP-LS-SPF Prefix NLRI is defined to indicate the status of the prefix with respect to the BGP SPF calculation. This will be used to expedite convergence for prefix unreachability as discussed in Section 6.5.1. If the SPF Status TLV is not included with the Prefix NLRI, the prefix is considered reachable. A single TLV type will be shared by the Node, Link, and Prefix NLRI. The TLV type 1184 will be assigned by IANA.



BGP Status Values: 0 - Reserved  
 1 - Prefix Unreachable with respect to SPF  
 2-254 - Undefined  
 255 - Reserved

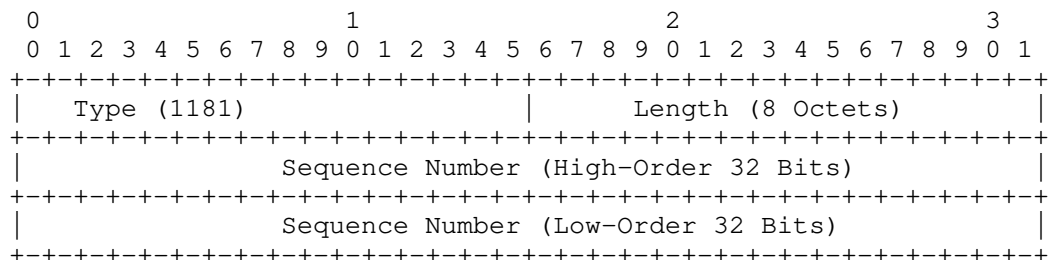
The BGP-LS-SPF Node Attribute SPF Status TLV, Link Attribute SPF Status TLV, and Prefix Attribute SPF Status TLV use the same TLV Type (1184). This implies that a BGP Update cannot contain multiple NLRI

with differing status. If the BGP-LS-SPF Status TLV is advertised and the advertised value is not defined for all NLRI included in the BGP update, then the SPF Status TLV is ignored and not used in SPF computation but is still announced to other BGP SPF speakers. An implementation MAY log an error for further analysis.

If a BGP SPF speaker received the Prefix NLRI but the SPF Status TLV is not received, then any previously received information is considered as implicitly withdrawn and the update is propagated to other BGP SPF speakers. A BGP SPF speaker receiving a BGP Update containing an SPF Status TLV in the BGP-LS attribute [RFC7752] with a value that is outside the range of defined values SHOULD be processed and announced to other BGP SPF speakers. However, a BGP SPF speaker MUST NOT use the Status TLV in its SPF computation. An implementation MAY log this information for further analysis.

#### 5.2.4. BGP-LS Attribute Sequence-Number TLV

A BGP-LS Attribute TLV of the BGP-LS-SPF NLRI types is defined to assure the most recent version of a given NLRI is used in the SPF computation. The Sequence-Number TLV is mandatory for BGP-LS-SPF NLRI. The TLV type 1181 has been assigned by IANA. The BGP-LS Attribute TLV will contain an 8-octet sequence number. The usage of the Sequence Number TLV is described in Section 6.1.



**Sequence Number** The 64-bit strictly-increasing sequence number MUST be incremented for every self-originated version of BGP-LS-SPF NLRI. BGP SPF speakers implementing this specification MUST use available mechanisms to preserve the sequence number's strictly increasing property for the deployed life of the BGP SPF speaker (including cold restarts). One mechanism for accomplishing this would be to use the high-order 32 bits of the sequence number as a wrap/boot count that is incremented any time the BGP router loses its sequence number state or the low-order 32 bits wrap.

When incrementing the sequence number for each self-originated NLRI, the sequence number should be treated as an unsigned 64-bit value. If the lower-order 32-bit value wraps, the higher-order 32-bit value

should be incremented and saved in non-volatile storage. If a BGP SPF speaker completely loses its sequence number state (e.g., the BGP SPF speaker hardware is replaced or experiences a cold-start), the BGP NLRI selection rules (see Section 6.1) will insure convergence, albeit not immediately.

The Sequence-Number TLV is mandatory for BGP-LS-SPF NLRI. If the Sequence-Number TLV is not received then the corresponding Link NLRI is considered as malformed and MUST be handled as 'Treat-as-withdraw'. An implementation MAY log an error for further analysis.

### 5.3. NEXT\_HOP Manipulation

All BGP peers that support SPF extensions would locally compute the LOC-RIB Next-Hop as a result of the SPF process. Consequently, the Next-Hop is always ignored on receipt. The Next-Hop address MUST be encoded as described in [RFC4760]. BGP SPF speakers MUST interpret the Next-Hop address of MP\_REACH\_NLRI attribute as an IPv4 address whenever the length of the Next-Hop address is 4 octets, and as a IPv6 address whenever the length of the Next-Hop address is 16 octets.

[RFC4760] modifies the rules of NEXT\_HOP attribute whenever the multiprotocol extensions for BGP-4 are enabled. BGP SPF speakers MUST set the NEXT\_HOP attribute according to the rules specified in [RFC4760] as the BGP-LS-SPF routing information is carried within the multiprotocol extensions for BGP-4.

## 6. Decision Process with SPF Algorithm

The Decision Process described in [RFC4271] takes place in three distinct phases. The Phase 1 decision function of the Decision Process is responsible for calculating the degree of preference for each route received from a BGP SPF speaker's peer. The Phase 2 decision function is invoked on completion of the Phase 1 decision function and is responsible for choosing the best route out of all those available for each distinct destination, and for installing each chosen route into the LOC-RIB. The combination of the Phase 1 and 2 decision functions is characterized as a Path Vector algorithm.

The SPF based Decision process replaces the BGP Decision process described in [RFC4271]. This process starts with selecting only those Node NLRI whose SPF capability TLV matches with the local BGP SPF speaker's SPF capability TLV value. Since Link-State NLRI always contains the local node descriptor Section 5.2, each NLRI is uniquely originated by a single BGP SPF speaker in the BGP SPF routing domain (the BGP node matching the NLRI's Node Descriptors). Instances of the same NLRI originated by multiple BGP SPF speakers would be

indicative of a configuration error or a masquerading attack (Section 9). These selected Node NLRI and their Link/Prefix NLRI are used to build a directed graph during the SPF computation as described below. The best routes for BGP prefixes are installed in the RIB as a result of the SPF process.

When BGP-LS-SPF NLRI is received, all that is required is to determine whether it is the most recent by examining the Node-ID and sequence number as described in Section 6.1. If the received NLRI has changed, it will be advertised to other BGP-LS-SPF peers. If the attributes have changed (other than the sequence number), a BGP SPF calculation will be triggered. However, a changed NLRI MAY be advertised immediately to other peers and prior to any SPF calculation. Note that the BGP MinRouteAdvertisementIntervalTimer and MinASOriginationIntervalTimer [RFC4271] timers are not applicable to the BGP-LS-SPF SAFI. The scheduling of the SPF calculation, as described in Section 6.3, is an implementation issue. Scheduling MAY be dampened consistent with the SPF back-off algorithm specified in [RFC8405].

The Phase 3 decision function of the Decision Process [RFC4271] is also simplified since under normal SPF operation, a BGP SPF speaker MUST advertise the changed NLRIs to all BGP peers with the BGP-LS-SPF AFI/SAFI and install the changed routes in the Global RIB. The only exception are unchanged NLRIs or stale NLRIs, i.e., NLRI received with a less recent (numerically smaller) sequence number.

#### 6.1. BGP NLRI Selection

The rules for all BGP-LS-SPF NLRIs selection for phase 1 of the BGP decision process, section 9.1.1 [RFC4271], no longer apply.

1. Routes originated by directly connected BGP SPF peers are preferred. This condition can be determined by comparing the BGP Identifiers in the received Local Node Descriptor and OPEN message. This rule will assure that stale NLRI is updated even if a BGP-LS router loses its sequence number state due to a cold-start.
2. The NLRI with the most recent Sequence Number TLV, i.e., highest sequence number is selected.
3. The route received from the BGP SPF speaker with the numerically larger BGP Identifier is preferred.

When a BGP SPF speaker completely loses its sequence number state, i.e., due to a cold start, or in the unlikely possibility that 64-bit sequence number wraps, the BGP routing domain will still converge.

This is due to the fact that BGP SPF speakers adjacent to the router will always accept self-originated NLRI from the associated speaker as more recent (rule # 1). When a BGP SPF speaker reestablishes a connection with its peers, any existing session will be taken down and stale NLRI will be replaced. The adjacent BGP SPF speaker will update their NLRI advertisements, hop by hop, until the BGP routing domain has converged.

The modified SPF Decision Process performs an SPF calculation rooted at the BGP SPF speaker using the metrics from the Link Attribute IGP Metric TLV (1095) and the Prefix Attribute Prefix Metric TLV (1155) [RFC7752]. As a result, any other BGP attributes that would influence the BGP decision process defined in [RFC4271] including ORIGIN, MULTI\_EXIT\_DISC, and LOCAL\_PREF attributes are ignored by the SPF algorithm. The NEXT\_HOP attribute is discussed in Section 5.3. The AS\_PATH and AS4\_PATH [RFC6793] attributes are preserved and used for loop detection [RFC4271]. They are ignored during the SPF computation for BGP-LS-SPF NLRI.

#### 6.1.1. BGP Self-Originated NLRI

Node, Link, or Prefix NLRI with Node Descriptors matching the local BGP SPF speaker are considered self-originated. When self-originated NLRI is received and it doesn't match the local node's NLRI content (including sequence number), special processing is required.

- \* If a self-originated NLRI is received and the sequence number is more recent (i.e., greater than the local node's sequence number for the NLRI), the NLRI sequence number will be advanced to one greater than the received sequence number and the NLRI will be readvertised to all peers.
- \* If self-originated NLRI is received and the sequence number is the same as the local node's sequence number but the attributes differ, the NLRI sequence number will be advanced to one greater than the received sequence number and the NLRI will be readvertised to all peers.
- \* If self-originated Link or Prefix NLRI is received and the Link or Prefix NLRI is no longer being advertised by the local node, the NLRI will be withdrawn.

The above actions are performed immediately when the first instance of a newer self-originated NLRI is received. In this case, the newer instance is considered to be a stale instance that was advertised by the local node prior to a restart where the NLRI state is lost. However, if subsequent newer self-originated NLRI is received for the same Node, Link, or Prefix NLRI, the readvertisement or withdrawal is delayed by 5 seconds since it is likely being advertised by a misconfigured or rogue BGP SPF speaker Section 9.

## 6.2. Dual Stack Support

The SPF-based decision process operates on Node, Link, and Prefix NLRI that support both IPv4 and IPv6 addresses. Whether to run a single SPF computation or multiple SPF computations for separate AFs is an implementation matter. Normally, IPv4 next-hops are calculated for IPv4 prefixes and IPv6 next-hops are calculated for IPv6 prefixes.

## 6.3. SPF Calculation based on BGP-LS-SPF NLRI

This section details the BGP-LS-SPF local routing information base (RIB) calculation. The router will use BGP-LS-SPF Node, Link, and Prefix NLRI to compute routes using the following algorithm. This calculation yields the set of routes associated with the BGP SPF Routing Domain. A router calculates the shortest-path tree using itself as the root. Optimizations to the BGP-LS-SPF algorithm are possible but MUST yield the same set of routes. The algorithm below supports Equal Cost Multi-Path (ECMP) routes. Weighted Unequal Cost Multi-Path routes are out of scope. The organization of this section owes heavily to section 16 of [RFC2328].

The following abstract data structures are defined in order to specify the algorithm.

- \* Local Route Information Base (LOC-RIB) - This routing table contains reachability information (i.e., next hops) for all prefixes (both IPv4 and IPv6) as well as BGP-LS-SPF node reachability. Implementations may choose to implement this with separate RIBs for each address family and/or Prefix versus Node reachability. It is synonymous with the Loc-RIB specified in [RFC4271].
- \* Global Routing Information Base (GLOBAL-RIB) - This is Routing Information Base (RIB) containing the current routes that are installed in the router's forwarding plane. This is commonly referred to in networking parlance as "the RIB".

- \* Link State NLRI Database (LSNDB) - Database of BGP-LS-SPF NLRI that facilitates access to all Node, Link, and Prefix NLRI.
- \* Candidate List (CAN-LIST) - This is a list of candidate Node NLRI's used during the BGP SPF calculation Section 6.3. The list is sorted by the cost to reach the Node NLRI with the Node NLRI with the lowest reachability cost at the head of the list. This facilitates execution of the Dijkstra algorithm Section 1.1 where the shortest paths between the local node and other nodes in graph area computed. The CAN-LIST is typically implemented as a heap but other data structures have been used.

The algorithm is comprised of the steps below:

1. The current LOC-RIB is invalidated, and the CAN-LIST is initialized to empty. The LOC-RIB is rebuilt during the course of the SPF computation. The existing routing entries are preserved for comparison to determine changes that need to be made to the GLOBAL-RIB in step 6.
2. The computing router's Node NLRI is updated in the LOC-RIB with a cost of 0 and the Node NLRI is also added to the CAN-LIST. The next-hop list is set to the internal loopback next-hop.
3. The Node NLRI with the lowest cost is removed from the candidate list for processing. If the BGP-LS Node attribute doesn't include an SPF Capability TLV (Section 5.2.1.1, the Node NLRI is ignored and the next lowest cost Node NLRI is selected from candidate list. If the BGP-LS Node attribute includes an SPF Status TLV (Section 5.2.1.1) indicating the node is unreachable, the Node NLRI is ignored and the next lowest cost Node NLRI is selected from candidate list. The Node corresponding to this NLRI will be referred to as the Current-Node. If the candidate list is empty, the SPF calculation has completed and the algorithm proceeds to step 6.
4. All the Prefix NLRI with the same Node Identifiers as the Current-Node will be considered for installation. The next-hop(s) for these Prefix NLRI are inherited from the Current-Node. The cost for each prefix is the metric advertised in the Prefix Attribute Prefix Metric TLV (1155) added to the cost to reach the Current-Node. The following will be done for each Prefix NLRI (referred to as the Current-Prefix):
  - \* If the BGP-LS Prefix attribute includes an SPF Status TLV indicating the prefix is unreachable, the Current-Prefix is considered unreachable and the next Prefix NLRI is examined in Step 4.

- \* If the Current-Prefix's corresponding prefix is in the LOC-RIB and the LOC-RIB cost is less than the Current-Prefix's metric, the Current-Prefix does not contribute to the route and the next Prefix NLRI is examined in Step 4.
  - \* If the Current-Prefix's corresponding prefix is not in the LOC-RIB, the prefix is installed with the Current-Node's next-hops installed as the LOC-RIB route's next-hops and the metric being updated. If the IGP Route Tag TLV (1153) is included in the Current-Prefix's NLRI Attribute, the tag(s) are installed in the current LOC-RIB route's tag(s).
  - \* If the Current-Prefix's corresponding prefix is in the LOC-RIB and the cost is less than the LOC-RIB route's metric, the prefix is installed with the Current-Node's next-hops replacing the LOC-RIB route's next-hops and the metric being updated and any route tags removed. If the IGP Route Tag TLV (1153) is included in the Current-Prefix's NLRI Attribute, the tag(s) are installed in the current LOC-RIB route's tag(s).
  - \* If the Current-Prefix's corresponding prefix is in the LOC-RIB and the cost is the same as the LOC-RIB route's metric, the Current-Node's next-hops will be merged with LOC-RIB route's next-hops. If the number of merged next-hops exceeds the Equal-Cost Multi-Path (ECMP) limit, the number of next-hops is reduced with next-hops on numbered links preferred over next-hops on unnumbered links. Among next-hops on numbered links, the next-hops with the highest IPv4 or IPv6 addresses are preferred. Among next-hops on unnumbered links, the next-hops with the highest Remote Identifiers are preferred [RFC5307]. If the IGP Route Tag TLV (1153) is included in the Current-Prefix's NLRI Attribute, the tag(s) are merged into the LOC-RIB route's current tags.
5. All the Link NLRI with the same Node Identifiers as the Current-Node will be considered for installation. Each link will be examined and will be referred to in the following text as the Current-Link. The cost of the Current-Link is the advertised IGP Metric TLV (1095) from the Link NLRI BGP-LS attribute added to the cost to reach the Current-Node. If the Current-Node is for the local BGP Router, the next-hop for the link will be a direct next-hop pointing to the corresponding local interface. For any other Current-Node, the next-hop(s) for the Current-Link will be inherited from the Current-Node. The following will be done for each link:

- a. The prefix(es) associated with the Current-Link are installed into the LOC-RIB using the same rules as were used for Prefix NLRI in the previous steps. Optionally, in deployments where BGP-SPF routers have limited routing table capacity, installation of these subnets can be suppressed. Suppression will have an operational impact as the IPv4/IPv6 link endpoint addresses will not be reachable and tools such as traceroute will display addresses that are not reachable.
- b. If the Current-Node NLRI attributes includes the SPF status TLV (Section 5.2.1.2) and the status indicates that the Node doesn't support transit, the next link for the Current-Node is processed in Step 5.
- c. If the Current-Link's NLRI attribute includes an SPF Status TLV indicating the link is down, the BGP-LS-SPF Link NLRI is considered down and the next link for the Current-Node is examined in Step 5.
- d. The Current-Link's Remote Node NLRI is accessed (i.e., the Node NLRI with the same Node identifiers as the Current-Link's Remote Node Descriptors). If it exists, it will be referred to as the Remote-Node and the algorithm will proceed as follows:
  - \* If the Remote-Node's NLRI attribute includes an SPF Status TLV indicating the node is unreachable, the next link for the Current-Node is examined in Step 5.
  - \* All the Link NLRI corresponding the Remote-Node will be searched for a Link NLRI pointing to the Current-Node. Each Link NLRI is examined for Remote Node Descriptors matching the Current-Node and Link Descriptors matching the Current-Link. For numbered links to match, the Link Descriptors MUST share a common IPv4 or IPv6 subnet. For unnumbered links to match, the Current Link's Local Identifier MUST match the Remote Node Link's Remote Identifier and the Current Link's Remote Identifier MUST the Remote Node Link's Local Identifier [RFC5307]. If these conditions are satisfied for one of the Remote-Node's links, the bi-directional connectivity check succeeds and the Remote-Node may be processed further. The Remote-Node's Link NLRI providing bi-directional connectivity will be referred to as the Remote-Link. If no Remote-Link is found, the next link for the Current-Node is examined in Step 5.

- \* If the Remote-Link NLRI attribute includes an SPF Status TLV indicating the link is down, the Remote-Link NLRI is considered down and the next link for the Current-Node is examined in Step 5.
  - \* If the Remote-Node is not on the CAN-LIST, it is inserted based on the cost. The Remote Node's cost is the cost of Current-Node added the Current-Link's IGP Metric TLV (1095). The next-hop(s) for the Remote-Node are inherited from the Current-Link.
  - \* If the Remote-Node NLRI is already on the CAN-LIST with a higher cost, it must be removed and reinserted with the Remote-Node cost based on the Current-Link (as calculated in the previous step). The next-hop(s) for the Remote-Node are inherited from the Current-Link.
  - \* If the Remote-Node NLRI is already on the CAN-LIST with the same cost, it need not be reinserted on the CAN-LIST. However, the Current-Link's next-hop(s) must be merged into the current set of next-hops for the Remote-Node.
  - \* If the Remote-Node NLRI is already on the CAN-LIST with a lower cost, it need not be reinserted on the CAN-LIST.
- e. Return to step 3 to process the next lowest cost Node NLRI on the CAN-LIST.
6. The LOC-RIB is examined and changes (adds, deletes, modifications) are installed into the GLOBAL-RIB. For each route in the LOC-RIB:
- \* If the route was added during the current BGP SPF computation, install the route into the GLOBAL-RIB.
  - \* If the route modified during the current BGP SPF computation (e.g., metric, tags, or next-hops), update the route in the GLOBAL-RIB.
  - \* If the route was not installed during the current BGP SPF computation, remove the route from both the GLOBAL-RIB and the LOC-RIB.

#### 6.4. IPv4/IPv6 Unicast Address Family Interaction

While the BGP-LS-SPF address family and the IPv4/IPv6 unicast address families MAY install routes into the same device routing tables, they will operate independently much the same as OSPF and IS-IS would operate today (i.e., "Ships-in-the-Night" mode). There is no implicit route redistribution between the BGP address families.

It is RECOMMENDED that BGP-LS-SPF IPv4/IPv6 route computation and installation be given scheduling priority by default over other BGP address families as these address families are considered as underlay SAFIs. Similarly, it is RECOMMENDED that the route preference or administrative distance give active route installation preference to BGP-LS-SPF IPv4/IPv6 routes over BGP routes from other AFI/SAFIs. However, this preference MAY be overridden by an operator-configured policy.

#### 6.5. NLRI Advertisement

##### 6.5.1. Link/Prefix Failure Convergence

A local failure will prevent a link from being used in the SPF calculation due to the IGP bi-directional connectivity requirement. Consequently, local link failures SHOULD always be given priority over updates (e.g., withdrawing all routes learned on a session) in order to ensure the highest priority propagation and optimal convergence.

An IGP such as OSPF [RFC2328] will stop using the link as soon as the Router-LSA for one side of the link is received. With a BGP advertisement, the link would continue to be used until the last copy of the BGP-LS-SPF Link NLRI is withdrawn. In order to avoid this delay, the originator of the Link NLRI SHOULD advertise a more recent version with an increased Sequence Number TLV for the BGP-LS-SPF Link NLRI including the SPF Status TLV (Section 5.2.2.2) indicating the link is down with respect to BGP SPF. The configurable LinkStatusDownAdvertise timer controls the interval that the BGP-LS-LINK NLRI is advertised with SPF Status indicating the link is down prior to withdrawal. If the link becomes available in that period, the originator of the BGP-LS-SPF LINK NLRI SHOULD advertise a more recent version of the BGP-LS-SPF Link NLRI without the SPF Status TLV in the BGP-LS Link Attributes. The suggested default value for the LinkStatusDownAdvertise timer is 2 seconds.

Similarly, when a prefix becomes unreachable, a more recent version of the BGP-LS-SPF Prefix NLRI SHOULD be advertised with the SPF Status TLV (Section 5.2.3.1) indicating the prefix is unreachable in the BGP-LS Prefix Attributes and the prefix will be considered

unreachable with respect to BGP SPF. The configurable PrefixStatusDownAdvertise timer controls the interval that the BGP-LS-Prefix NLRI is advertised with SPF Status indicating the prefix is unreachable prior to withdrawal. If the prefix becomes reachable in that period, the originator of the BGP-LS-SPF Prefix NLRI SHOULD advertise a more recent version of the BGP-LS-SPF Prefix NLRI without the SPF Status TLV in the BGP-LS Prefix Attributes. The suggested default value for the PrefixStatusDownAdvertise timer is 2 seconds.

#### 6.5.2. Node Failure Convergence

With BGP without graceful restart [RFC4724], all the NLRI advertised by a node are implicitly withdrawn when a session failure is detected. If fast failure detection such as BFD is utilized, and the node is on the fastest converging path, the most recent versions of BGP-LS-SPF NLRI may be withdrawn. This will result into an older version of the NLRI being used until the new versions arrive and, potentially, unnecessary route flaps. For the BGP-LS-SPF SAFI, NLRI SHOULD NOT be implicitly withdrawn immediately to prevent such unnecessary route flaps. The configurable NLRIImplicitWithdrawalDelay timer controls the interval that NLRI is retained prior to implicit withdrawal after a BGP SPF speaker has transitioned out of Established state. This will not delay convergence since the adjacent nodes will detect the link failure and advertise a more recent NLRI indicating the link is down with respect to BGP SPF (Section 6.5.1) and the BGP SPF calculation will fail the bi-directional connectivity check Section 6.3. The suggested default value for the NLRIImplicitWithdrawalDelay timer is 2 seconds.

### 7. Error Handling

This section describes the Error Handling actions, as described in [RFC7606], that are specific to SAFI BGP-LS-SPF BGP Update message processing.

#### 7.1. Processing of BGP-LS-SPF TLVs

When a BGP SPF speaker receives a BGP Update containing a malformed Node NLRI SPF Status TLV in the BGP-LS Attribute [RFC7752], it MUST ignore the received TLV and MUST NOT pass it to other BGP peers as specified in [RFC7606]. When discarding an associated Node NLRI with a malformed TLV, a BGP SPF speaker SHOULD log an error for further analysis.

When a BGP SPF speaker receives a BGP Update containing a malformed Link NLRI SPF Status TLV in the BGP-LS Attribute [RFC7752], it MUST ignore the received TLV and MUST NOT pass it to other BGP peers as specified in [RFC7606]. When discarding an associated Link NLRI with a malformed TLV, a BGP SPF speaker SHOULD log an error for further analysis.

When a BGP SPF speaker receives a BGP Update containing a malformed Prefix NLRI SPF Status TLV in the BGP-LS Attribute [RFC7752], it MUST ignore the received TLV and MUST NOT pass it to other BGP peers as specified in [RFC7606]. When discarding an associated Prefix NLRI with a malformed TLV, a BGP SPF speaker SHOULD log an error for further analysis.

When a BGP SPF speaker receives a BGP Update containing a malformed SPF Capability TLV in the Node NLRI BGP-LS Attribute [RFC7752], it MUST ignore the received TLV and the Node NLRI and MUST NOT pass it to other BGP peers as specified in [RFC7606]. When discarding a Node NLRI with a malformed TLV, a BGP SPF speaker SHOULD log an error for further analysis.

When a BGP SPF speaker receives a BGP Update containing a malformed IPv4 Prefix-Length TLV in the Link NLRI BGP-LS Attribute [RFC7752], it MUST ignore the received TLV and the Node NLRI and MUST NOT pass it to other BGP peers as specified in [RFC7606]. The corresponding Link NLRI is considered as malformed and MUST be handled as 'Treat-as-withdraw'. An implementation MAY log an error for further analysis.

When a BGP SPF speaker receives a BGP Update containing a malformed IPv6 Prefix-Length TLV in the Link NLRI BGP-LS Attribute [RFC7752], it MUST ignore the received TLV and the Node NLRI and MUST NOT pass it to other BGP peers as specified in [RFC7606]. The corresponding Link NLRI is considered as malformed and MUST be handled as 'Treat-as-withdraw'. An implementation MAY log an error for further analysis.

## 7.2. Processing of BGP-LS-SPF NLRIs

A Link-State NLRI MUST NOT be considered as malformed or invalid based on the inclusion/exclusion of TLVs or contents of the TLV fields (i.e., semantic errors), as described in Section 5.1 and Section 5.1.1.

A BGP-LS-SPF Speaker MUST perform the following syntactic validation of the BGP-LS-SPF NLRI to determine if it is malformed.

1. Does the sum of all TLVs found in the BGP MP\_REACH\_NLRI attribute correspond to the BGP MP\_REACH\_NLRI length?
2. Does the sum of all TLVs found in the BGP MP\_UNREACH\_NLRI attribute correspond to the BGP MP\_UNREACH\_NLRI length?
3. Does the sum of all TLVs found in a BGP-LS-SPF NLRI correspond to the Total NLRI Length field of all its Descriptors?
4. When an NLRI TLV is recognized, is the length of the TLV and its sub-TLVs valid?
5. Has the syntactic correctness of the NLRI fields been verified as per [RFC7606]?
6. Has the rule regarding ordering of TLVs been followed as described in Section 5.1.1?

When the error determined allows for the router to skip the malformed NLRI(s) and continue processing of the rest of the update message (e.g., when the TLV ordering rule is violated), then it **MUST** handle such malformed NLRIs as 'Treat-as-withdraw'. In other cases, where the error in the NLRI encoding results in the inability to process the BGP update message (e.g., length related encoding errors), then the router **SHOULD** handle such malformed NLRIs as 'AFI/SAFI disable' when other AFI/SAFI besides BGP-LS are being advertised over the same session. Alternately, the router **MUST** perform 'session reset' when the session is only being used for BGP-LS-SPF or when its 'AFI/SAFI disable' action is not possible.

### 7.3. Processing of BGP-LS Attribute

A BGP-LS Attribute **MUST NOT** be considered as malformed or invalid based on the inclusion/exclusion of TLVs or contents of the TLV fields (i.e., semantic errors), as described in Section 5.1 and Section 5.1.1.

A BGP-LS-SPF Speaker **MUST** perform the following syntactic validation of the BGP-LS Attribute to determine if it is malformed.

1. Does the sum of all TLVs found in the BGP-LS-SPF Attribute correspond to the BGP-LS Attribute length?
2. Has the syntactic correctness of the Attributes (including BGP-LS Attribute) been verified as per [RFC7606]?
3. Is the length of each TLV and, when the TLV is recognized then, its sub-TLVs in the BGP-LS Attribute valid?

When the detected error allows for the router to skip the malformed BGP-LS Attribute and continue processing of the rest of the update message (e.g., when the BGP-LS Attribute length and the total Path Attribute Length are correct but some TLV/sub-TLV length within the BGP-LS Attribute is invalid), then it MUST handle such malformed BGP-LS Attribute as 'Attribute Discard'. In other cases, when the error in the BGP-LS Attribute encoding results in the inability to process the BGP update message, then the handling is the same as described above for malformed NLRI.

Note that the 'Attribute Discard' action results in the loss of all TLVs in the BGP-LS Attribute and not the removal of a specific malformed TLV. The removal of specific malformed TLVs may give a wrong indication to a BGP SPF speaker that the specific information is being deleted or is not available.

When a BGP SPF speaker receives an update message with Link-State NLRI(s) in the MP\_REACH\_NLRI but without the BGP-LS-SPF Attribute, it is most likely an indication that a BGP SPF speaker preceding it has performed the 'Attribute Discard' fault handling. An implementation SHOULD preserve and propagate the Link-State NLRIs in such an update message so that the BGP SPF speaker can detect the loss of link-state information for that object and not assume its deletion/withdrawal. This also makes it possible for a network operator to trace back to the BGP SPF speaker which actually detected a problem with the BGP-LS Attribute.

An implementation SHOULD log an error for further analysis for problems detected during syntax validation.

When a BGP SPF speaker receives a BGP Update containing a malformed IGP metric TLV in the Link NLRI BGP-LS Attribute [RFC7752], it MUST ignore the received TLV and the Link NLRI and MUST NOT pass it to other BGP peers as specified in [RFC7606]. When discarding a Link NLRI with a malformed TLV, a BGP SPF speaker SHOULD log an error for further analysis.

## 8. IANA Considerations

This document defines the use of SAFI (80) for BGP SPF operation Section 5.1, and requests IANA to assign the value from the First Come First Serve (FCFS) range in the Subsequent Address Family Identifiers (SAFI) Parameters registry.

This document also defines five attribute TLVs of BGP-LS-SPF NLRI. We request IANA to assign types for the SPF capability TLV, Sequence Number TLV, IPv4 Link Prefix-Length TLV, IPv6 Link Prefix-Length TLV, and SPF Status TLV from the "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs" Registry.

Attribute TLV	Suggested Value	NLRI Applicability
SPF Capability	1180	Node
SPF Status	1184	Node, Link, Prefix
IPv4 Link Prefix Length	1182	Link
IPv6 Link Prefix Length	1183	Link
Sequence Number	1181	Node, Link, Prefix

Table 1: NLRI Attribute TLVs

## 9. Security Considerations

This document defines a BGP SAFI, i.e., the BGP-LS-SPF SAFI. This document does not change the underlying security issues inherent in the BGP protocol [RFC4271]. The Security Considerations discussed in [RFC4271] apply to the BGP SPF functionality as well. The analysis of the security issues for BGP mentioned in [RFC4272] and [RFC6952] also applies to this document. The analysis of Generic Threats to Routing Protocols done in [RFC4593] is also worth noting. As the modifications described in this document for BGP SPF apply to IPv4 Unicast and IPv6 Unicast as undelay SAFIs in a single BGP SPF Routing Domain, the BGP security solutions described in [RFC6811] and [RFC8205] are somewhat constricted as they are meant to apply for inter-domain BGP where multiple BGP Routing Domains are typically involved. The BGP-LS-SPF SAFI NLRI described in this document are typically advertised between EBGP or IBGP speakers under a single administrative domain.

In the context of the BGP peering associated with this document, a BGP speaker MUST NOT accept updates from a peer that is not within any administrative control of an operator. That is, a participating BGP speaker SHOULD be aware of the nature of its peering relationships. Such protection can be achieved by manual configuration of peers at the BGP speaker.

In order to mitigate the risk of peering with BGP speakers masquerading as legitimate authorized BGP speakers, it is recommended that the TCP Authentication Option (TCP-AO) [RFC5925] be used to authenticate BGP sessions. If an authorized BGP peer is compromised, that BGP peer could advertise modified Node, Link, or Prefix NLRI will result in misrouting, repeating origination of NLRI, and/or excessive SPF calculations. When a BGP speaker detects that its self-originated NLRI is being originated by another BGP speaker, an appropriate error should be logged so that the operator can take corrective action.

## 10. Management Considerations

This section includes unique management considerations for the BGP-LS-SPF address family.

### 10.1. Configuration

All routers in BGP SPF Routing Domain are under a single administrative domain allowing for consistent configuration.

#### 10.1.1. Link Metric Configuration

Within a BGP SPF Routing Domain, the IGP metrics for all advertised links SHOULD be configured or defaulted consistently. For example, if a default metric is used for one router's links, then a similar metric should be used for all router's links. Similarly, if the link cost is derived from using the inverse of the link bandwidth on one router, then this SHOULD be done for all routers and the same reference bandwidth should be used to derive the inversely proportional metric. Failure to do so will not result in correct routing based on link metric.

#### 10.1.2. backoff-config

In addition to configuration of the BGP-LS-SPF address family, implementations SHOULD support the "Shortest Path First (SPF) Back-Off Delay Algorithm for Link-State IGPs" [RFC8405]. If supported, configuration of the INITIAL\_SPF\_DELAY, SHORT\_SPF\_DELAY, LONG\_SPF\_DELAY, TIME\_TO\_LEARN, and HOLDDOWN\_INTERVAL MUST be supported [RFC8405]. Section 6 of [RFC8405] recommends consistent configuration of these values throughout the IGP routing domain and this also applies to the BGP SPF Routing Domain.

## 10.2. Operational Data

In order to troubleshoot SPF issues, implementations SHOULD support an SPF log including entries for previous SPF computations. Each SPF log entry would include the BGP-LS-SPF NLRI SPF triggering the SPF, SPF scheduled time, SPF start time, SPF end time, and SPF type if different types of SPF are supported. Since the size of the log will be finite, implementations SHOULD also maintain counters for the total number of SPF computations and the total number of SPF triggering events. Additionally, to troubleshoot SPF scheduling and back-off [RFC8405], the current SPF back-off state, remaining time-to-learn, remaining holddown, last trigger event time, last SPF time, and next SPF time should be available.

## 11. Implementation Status

Note RFC Editor: Please remove this section and the associated references prior to publication.

This section records the status of known implementations of the protocol defined by this specification at the time of posting of this Internet-Draft and is based on a proposal described in [RFC7942]. The description of implementations in this section is intended to assist the IETF in its decision processes in progressing drafts to RFCs. Please note that the listing of any individual implementation here does not imply endorsement by the IETF. Furthermore, no effort has been spent to verify the information presented here that was supplied by IETF contributors. This is not intended as, and must not be construed to be, a catalog of available implementations or their features. Readers are advised to note that other implementations may exist.

According to RFC 7942, "this will allow reviewers and working groups to assign due consideration to documents that have the benefit of running code, which may serve as evidence of valuable experimentation and feedback that have made the implemented protocols more mature. It is up to the individual working groups to use this information as they see fit".

The BGP-LS-SPF implementation status is documented in [I-D.psarkar-lsvr-bgp-spf-impl].

## 12. Acknowledgements

The authors would like to thank Sue Hares, Jorge Rabadan, Boris Hassanov, Dan Frost, Matt Anderson, Fred Baker, Lukas Krattiger, Yingzhen Qu, and Haibo Wang for their review and comments. Thanks to Pushpasis Sarkar for discussions on preventing a BGP SPF Router from being used for non-local traffic (i.e., transit traffic).

The authors extend special thanks to Eric Rosen for fruitful discussions on BGP-LS-SPF convergence as compared to IGPs.

## 13. Contributors

In addition to the authors listed on the front page, the following co-authors have contributed to the document.

Derek Yeung  
Arrcus, Inc.  
derek@arrcus.com

Gunter Van De Velde  
Nokia  
gunter.van\_de\_velde@nokia.com

Abhay Roy  
Arrcus, Inc.  
abhay@arrcus.com

Venu Venugopal  
Cisco Systems  
venuv@cisco.com

Chaitanya Yadlapalli  
AT&T  
cy098d@att.com

## 14. References

### 14.1. 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>>.

- [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>>.
- [RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <<https://www.rfc-editor.org/info/rfc4272>>.
- [RFC4593] Barbir, A., Murphy, S., and Y. Yang, "Generic Threats to Routing Protocols", RFC 4593, DOI 10.17487/RFC4593, October 2006, <<https://www.rfc-editor.org/info/rfc4593>>.
- [RFC4750] Joyal, D., Ed., Galecki, P., Ed., Giacalone, S., Ed., Coltun, R., and F. Baker, "OSPF Version 2 Management Information Base", RFC 4750, DOI 10.17487/RFC4750, December 2006, <<https://www.rfc-editor.org/info/rfc4750>>.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <<https://www.rfc-editor.org/info/rfc4760>>.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <<https://www.rfc-editor.org/info/rfc5492>>.
- [RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <<https://www.rfc-editor.org/info/rfc5925>>.
- [RFC6793] Vohra, Q. and E. Chen, "BGP Support for Four-Octet Autonomous System (AS) Number Space", RFC 6793, DOI 10.17487/RFC6793, December 2012, <<https://www.rfc-editor.org/info/rfc6793>>.
- [RFC6811] Mohapatra, P., Scudder, J., Ward, D., Bush, R., and R. Austein, "BGP Prefix Origin Validation", RFC 6811, DOI 10.17487/RFC6811, January 2013, <<https://www.rfc-editor.org/info/rfc6811>>.
- [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>>.

- [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>>.
- [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>>.
- [RFC8205] Lepinski, M., Ed. and K. Sriram, Ed., "BGPsec Protocol Specification", RFC 8205, DOI 10.17487/RFC8205, September 2017, <<https://www.rfc-editor.org/info/rfc8205>>.
- [RFC8405] Decraene, B., Litkowski, S., Gredler, H., Lindem, A., Francois, P., and C. Bowers, "Shortest Path First (SPF) Back-Off Delay Algorithm for Link-State IGPs", RFC 8405, DOI 10.17487/RFC8405, June 2018, <<https://www.rfc-editor.org/info/rfc8405>>.
- [RFC8654] Bush, R., Patel, K., and D. Ward, "Extended Message Support for BGP", RFC 8654, DOI 10.17487/RFC8654, October 2019, <<https://www.rfc-editor.org/info/rfc8654>>.
- [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>>.

#### 14.2. Informational References

- [I-D.ietf-lsvr-applicability]  
Patel, K., Lindem, A., Zandi, S., and G. Dawra, "Usage and Applicability of Link State Vector Routing in Data Centers", Work in Progress, Internet-Draft, draft-ietf-lsvr-applicability-05, 24 March 2020, <<http://www.ietf.org/internet-drafts/draft-ietf-lsvr-applicability-05.txt>>.
- [I-D.psarkar-lsvr-bgp-spf-impl]  
Sarkar, P., Patel, K., Pallagatti, S., and s. sajibasil@gmail.com, "BGP Shortest Path Routing Extension Implementation Report", Work in Progress, Internet-Draft, draft-psarkar-lsvr-bgp-spf-impl-00, 2 June 2020, <<http://www.ietf.org/internet-drafts/draft-psarkar-lsvr-bgp-spf-impl-00.txt>>.

- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <<https://www.rfc-editor.org/info/rfc2328>>.
- [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>>.
- [RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, DOI 10.17487/RFC4724, January 2007, <<https://www.rfc-editor.org/info/rfc4724>>.
- [RFC4915] Psenak, P., Mirtorabi, S., Roy, A., Nguyen, L., and P. Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF", RFC 4915, DOI 10.17487/RFC4915, June 2007, <<https://www.rfc-editor.org/info/rfc4915>>.
- [RFC5286] Atlas, A., Ed. and A. Zinin, Ed., "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, DOI 10.17487/RFC5286, September 2008, <<https://www.rfc-editor.org/info/rfc5286>>.
- [RFC5307] Kompella, K., Ed. and Y. Rekhter, Ed., "IS-IS Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 5307, DOI 10.17487/RFC5307, October 2008, <<https://www.rfc-editor.org/info/rfc5307>>.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <<https://www.rfc-editor.org/info/rfc5880>>.
- [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>>.
- [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>>.
- [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>>.

[RFC7942] Sheffer, Y. and A. Farrel, "Improving Awareness of Running Code: The Implementation Status Section", BCP 205, RFC 7942, DOI 10.17487/RFC7942, July 2016, <<https://www.rfc-editor.org/info/rfc7942>>.

## Authors' Addresses

Keyur Patel  
Arrcus, Inc.

Email: [keyur@arrcus.com](mailto:keyur@arrcus.com)

Acee Lindem  
Cisco Systems  
301 Midenhall Way  
Cary, NC 27513  
United States of America

Email: [acee@cisco.com](mailto:acee@cisco.com)

Shawn Zandi  
LinkedIn  
222 2nd Street  
San Francisco, CA 94105  
United States of America

Email: [szandi@linkedin.com](mailto:szandi@linkedin.com)

Wim Henderickx  
Nokia  
Antwerp  
Belgium

Email: [wim.henderickx@nokia.com](mailto:wim.henderickx@nokia.com)

LSVR  
Internet-Draft  
Intended status: Informational  
Expires: January 3, 2019

K. Patel  
Arrcus, Inc.  
A. Lindem  
Cisco Systems  
S. Zandi  
G. Dawra  
Linkedin  
July 2, 2018

Usage and Applicability of Link State Vector Routing in Data Centers  
draft-keyupate-lsvr-applicability-02.txt

Abstract

This document discusses the usage and applicability of Link State Vector Routing (LSVR) extensions in the CLOS architecture of Data Center Networks. The document is intended to provide a simplified guide for the deployment of LSVR extensions.

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 <http://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 January 3, 2019.

Copyright Notice

Copyright (c) 2018 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 (<http://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 . . . . .	2
2. Requirements Language . . . . .	2
3. Recommended Reading . . . . .	3
4. Common Deployment Scenario . . . . .	3
5. Justification for BGP SPF Extension . . . . .	4
6. LSVR Applicability to CLOS Networks . . . . .	4
6.1. Usage of BGP-LS SAFI . . . . .	5
6.1.1. Relationship to Other BGP AFI/SAFI Tuples . . . . .	5
6.2. Peering Models . . . . .	5
6.2.1. Bi-Connected Graph Heuristic . . . . .	6
6.3. BGP Peer Discovery . . . . .	6
6.3.1. BGP Peer Discovery Requirements . . . . .	6
6.3.2. BGP Peer Discovery Alternatives . . . . .	7
6.3.3. Data Center Interconnect (DCI) Applicability . . . . .	7
6.4. Non-CLOS/FAT Tree Topology Applicability . . . . .	8
7. IANA Considerations . . . . .	8
8. Security Considerations . . . . .	8
9. Acknowledgements . . . . .	8
10. References . . . . .	8
10.1. Normative References . . . . .	8
10.2. Informative References . . . . .	9
Authors' Addresses . . . . .	10

## 1. Introduction

This document complements [I-D.ietf-lsvr-bgp-spf] by discussing the applicability of the technology in a simple and fairly common deployment scenario, which is described in Section 4.

After describing the deployment scenario, Section 5 will describe the reasons for BGP modifications for such deployments.

Once the control plane routing protocol requirements are described, Section 6 will cover the LSVR protocol enhancements to BGP to meet these requirements and their applicability to Data Center CLOS networks.

## 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. Recommended Reading

This document assumes knowledge of existing data center networks and data center network topologies [CLOS]. This document also assumes knowledge of data center routing protocols like BGP [RFC4271], BGP-SPF [I-D.ietf-lsvr-bgp-spf], OSPF [RFC2328], as well as, data center OAM protocols like LLDP [RFC4957] and BFD [RFC5580].

### 4. Common Deployment Scenario

Within a Data Center, a common network design to interconnect servers is done using the CLOS topology [CLOS]. The CLOS topology is fully non-blocking and the topology is realized using Equal Cost Multipath (ECMP). In a CLOS topology, the minimum number of parallel paths between two servers is determined by the width of a tier-1 stage as shown in the figure 1.

The following example illustrates multistage CLOS topology.

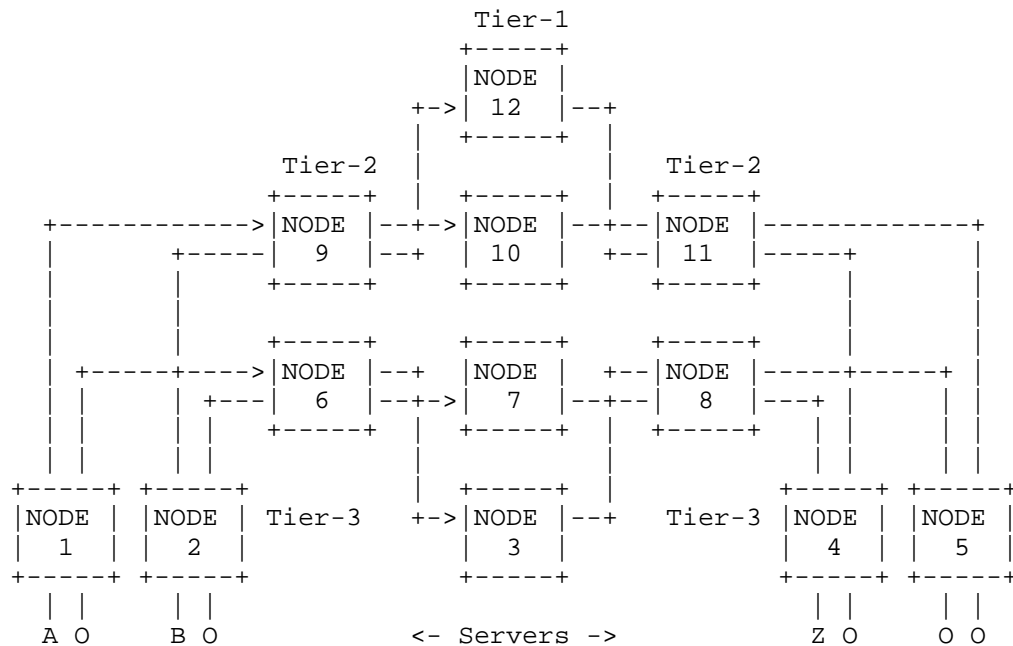


Figure 1: Illustration of the basic CLOS

## 5. Justification for BGP SPF Extension

Many data centers use BGP as a routing protocol to create an overlay as well as an underlay network for their CLOS Topologies to simplify layer-3 routing and operations [RFC7938]. However, BGP is a path-vector routing protocol. Since it does not create a fabric topology, it uses hop-by-hop EBGP peering to facilitate hop-by-hop routing to create the underlay network and to resolve any overlay next hops. The hop-by-hop BGP peering paradigm imposes several restrictions within a CLOS. It severely prohibits a deployment of Route Reflectors/Route Controllers as the EBGP sessions are inline with the data path. The BGP best path algorithm is prefix-based and it prevents announcements of prefixes to other BGP speakers until the best path decision process is performed for the prefix at each intermediate hop. These restrictions significantly delay the overall convergence of the underlay network within a CLOS.

The LSVR SPF modifications allow BGP to overcome these limitations. Furthermore, using the BGP-LS NLRI format [RFC7752] allows the LSVR data to be advertised for nodes, links, and prefixes in the BGP routing domain and used for SPF computations.

## 6. LSVR Applicability to CLOS Networks

With the BGP SPF extensions [I-D.ietf-lsvr-bgp-spf], the BGP best path computation and route computation are replaced with OSPF-like algorithms [RFC2328] both to determine whether an BGP-LS NLRI has changed and needs to be re-advertised and to compute the routing table. These modifications will significantly improve convergence of the underlay while affording the operational benefits of a single routing protocol [RFC7938].

Data center controllers typically require visibility to the BGP topology to compute traffic-engineered paths. These controllers learn the topology and other relevant information via the BGP-LS address family [RFC7752] which is totally independent of the underlay address families (usually IPv4/IPv6 unicast). Furthermore, in traditional BGP underlays, all the BGP routers will need to advertise their BGP-LS information independently. With the BGP SPF extensions, controllers can learn the topology using the same BGP advertisements used to compute the underlay routes. Furthermore, these data center controllers can avail the convergence advantages of the BGP SPF extensions. The placement of controllers can be outside of the forwarding path or within the forwarding path.

Alternatively, as each and every router in the BGP SPF domain will have a complete view of the topology, the operator can also choose to configure BGP sessions in hop-by-hop peering model described in

[RFC7938] along with BFD [RFC5580]. In doing so, while the hop-by-hop peering model lacks inherent benefits of the controller-based model, BGP updates need not be serialized by BGP best path algorithm in either of these models. This helps overall network convergence.

#### 6.1. Usage of BGP-LS SAFI

The BGP SPF extensions [I-D.ietf-lsvr-bgp-spf] define a new BGP-LS SAFI for announcement of BGP SPF link-state. The NLRI format and its associated attributes follow the format of BGP-LS for node, link, and prefix announcements. Whether the peering model within a CLOS follows hop-by-hop peering described in [RFC7938] or any controller-based or route-reflector peering, an operator can exchange BGP SPF SAFI routes over the BGP peering by simply configuring BGP SPF SAFI between the necessary BGP speakers.

The BGP-LS SPF SAFI can also co-exist with BGP IP Unicast SAFI which could exchange overlapping IP routes. The routes received by these SAFIs are evaluated, stored, and announced separately according to the rules of [RFC4760]. The tie-breaking of route installation is a matter of the local policies and preferences of the network operator.

Finally, as the BGP SPF peering is done following the procedures described in [RFC4271], all the existing transport security mechanisms including [RFC5925] are available for the BGP-LS SPF SAFI.

##### 6.1.1. Relationship to Other BGP AFI/SAFI Tuples

Normally, the BGP-LS AFI/SAFI is used solely to compute the underlay and is given preference over other AFI/SAFIs. Other BGP SAFIs, e.g., IPv6/IPv6 Unicast VPN would use the BGP-SPF computed routes for next hop resolution. However, if BGP-LS NLRI is also being advertised for controller consumption, there is no need to replicate the Node, Link, and Prefix NLRI in BGP-NLRI. Rather, additional NLRI attributes can be advertised in the BGP-LS SPF AFI/SAFI as required.

#### 6.2. Peering Models

As previously stated, BGP SPF can be deployed using the existing peering model where there is a single hop BGP session on each and every link in the data center fabric [RFC7938]. This provides for both the advertisement of routes and the determination of link and neighboring switch availability. With BGP SPF, the underlay will converge faster due to changes in the decision process which will allow NLRI changes to be advertised faster after detecting a change.

Alternately, BFD [RFC5580] can be used to swiftly determine the availability of links and the BGP peering model can be significantly

sparser than the data center fabric. BGP SPF sessions then only be established with enough peers to provide a bi-connected graph. If IEBGP is used, then the BGP routers at tier N-1 will act as route-reflectors for the routers at tier N.

#### 6.2.1. Bi-Connected Graph Heuristic

With this heuristic, discovery of BGP peers is assumed Section 6.3. Additionally, it assumed that the direction of the peering can be ascertained. In the context of a data center fabric, direction is either northbound (toward the spine), southbound (toward the Top-Of-Rack (TOR) switches) or east-west (same level in hierarchy). The determination of the direction is beyond the scope of this document. However, it would be reasonable to assume a technique where the TOR switches can be identified and the number of hops to the TOR is used to determine the direction.

In this heuristic, BGP speakers allow passive session establishment for southbound BGP sessions. For northbound sessions, BGP speakers will attempt to maintain two northbound BGP sessions with different switches (in data center fabrics there is normally a single layer-3 connection anyway). For east-west sessions, passive BGP session establishment is allowed. However, BGP speaker will never actively establish an east-west BGP session unless it can't establish two northbound BGP sessions.

#### 6.3. BGP Peer Discovery

##### 6.3.1. BGP Peer Discovery Requirements

The most basic requirement is to be able to discover the address of a single-hop peer without pre-configuration. This is being accomplished today with using IPv6 Router Advertisements (RA) [RFC4861] and assuming that a BGP sessions is desired with any discovered peer. Beyond the basic requirement, it is useful to have to following information relating to the BGP session:

- o Autonomous System (AS) and BGP Identifier of a potential peer. The latter can be used for debugging and to decrease the likelihood of BGP session establishment collisions.
- o Security capabilities supported and for cryptographic authentication, the security capabilities and possibly a key-chain [RFC8177] to be used.
- o Session Policy Identifier - A group number or name used to associate common session parameters with the peer. For example,

in a data center, BGP sessions with a Top of Rack (ToR) device could have parameters than BGP sessions between leaf and spine.

In a data center fabric, it is often useful to know whether a peer is southbound (towards the servers) or northbound (towards the spine or super-spine) Section 6.2.1. A potential requirement would also be to determine this dynamically. One mechanism, without specifying all the details, might be for the ToRs to be identified when installed and for the others switches in the fabric to determine their level based on the distance from the closest ToR.

If there are multiple links between BGP speakers or the links between BGP speakers are unnumbered, it is also useful to be able to establish multi-hop sessions using the loopback addresses. This will often require the discovery protocol to install route(s) toward the potential peer loopback addresses prior to BGP session establishment.

Finally, a simple BGP discovery protocol could also be used to establish a multi-hop session with one or more controllers by advertising connectivity to one or more controllers. However, once the multi-hop session actually traverses multiple nodes, it is bordering a distance-vector routing protocol and possibly this is not a good requirement for the discovery protocol.

#### 6.3.2. BGP Peer Discovery Alternatives

While BGP peer discovery is not part of [I-D.ietf-lsvr-bgp-spf], there are, at least, three proposals for BGP peer discovery. At least one of these mechanisms will be adopted and will be applicable to deployments other than the data center. It is strongly RECOMMENDED that the accepted mechanism be used in conjunction with BGP SPF in data centers. The BGP discovery mechanism should discover both peer addresses and endpoints for BFD discovery. Additionally, it would be great if there were a heuristic for determining whether the peer is at a tier above or below the discovering BGP speaker (refer to Section 6.2.1).

The BGP discovery mechanisms under consideration are [I-D.acee-idr-lldp-peer-discovery], [I-D.xu-idr-neighbor-autodiscovery], and [I-D.ymbk-lsvr-lsoe].

#### 6.3.3. Data Center Interconnect (DCI) Applicability

Since BGP SPF is to be used for the routing underlay and DCI gateway boxes typically have direct or very simple connectivity, BGP external sessions would typically not include the BGP SPF SAFI.

#### 6.4. Non-CLOS/FAT Tree Topology Applicability

The BGP SPF extensions [I-D.ietf-lsvr-bgp-spf] can be used in other topologies and avail the inherent convergence improvements. Additionally, sparse peering techniques may be utilized Section 6.2. However, determining whether or to establish a BGP session is more complex and the heuristic described in Section 6.2.1 cannot be used. In such topologies, other techniques such as those described in [I-D.li-dynamic-flooding] may be employed. One potential deployment would be the underlay for a Service Provider (SP) backbone where usage of a single protocol, i.e., BGP, is desired.

#### 7. IANA Considerations

No IANA updates are requested by this document.

#### 8. Security Considerations

This document introduces no new security considerations above and beyond those already specified in the [RFC4271] and [I-D.ietf-lsvr-bgp-spf].

#### 9. Acknowledgements

The authors would like to thank Alvaro Retana and Yan Filyurin for the review and comments.

#### 10. References

##### 10.1. Normative References

- [I-D.ietf-lsvr-bgp-spf]  
Patel, K., Lindem, A., Zandi, S., and W. Henderickx,  
"Shortest Path Routing Extensions for BGP Protocol",  
draft-ietf-lsvr-bgp-spf-01 (work in progress), May 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>>.
- [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>>.

## 10.2. Informative References

- [CLOS] "A Study of Non-Blocking Switching Networks", The Bell System Technical Journal, Vol. 32(2), DOI 10.1002/j.1538-7305.1953.tb01433.x, March 1953.
- [I-D.acee-idr-lldp-peer-discovery] Lindem, A., Patel, K., Zandi, S., Haas, J., and X. Xu, "BGP Logical Link Discovery Protocol (LLDP) Peer Discovery", draft-acee-idr-lldp-peer-discovery-03 (work in progress), June 2018.
- [I-D.li-dynamic-flooding] Li, T. and P. Psenak, "Dynamic Flooding on Dense Graphs", draft-li-dynamic-flooding-05 (work in progress), June 2018.
- [I-D.xu-idr-neighbor-autodiscovery] Xu, X., Bi, K., Tantsura, J., Triantafyllis, N., and K. Talaulikar, "BGP Neighbor Autodiscovery", draft-xu-idr-neighbor-autodiscovery-08 (work in progress), May 2018.
- [I-D.ymbk-lsvr-lsoe] Bush, R. and K. Patel, "Link State Over Ethernet", draft-ymbk-lsvr-lsoe-00 (work in progress), March 2018.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <<https://www.rfc-editor.org/info/rfc2328>>.
- [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>>.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <<https://www.rfc-editor.org/info/rfc4760>>.
- [RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, DOI 10.17487/RFC4861, September 2007, <<https://www.rfc-editor.org/info/rfc4861>>.

- [RFC4957] Krishnan, S., Ed., Montavont, N., Njedjou, E., Veerepalli, S., and A. Yegin, Ed., "Link-Layer Event Notifications for Detecting Network Attachments", RFC 4957, DOI 10.17487/RFC4957, August 2007, <<https://www.rfc-editor.org/info/rfc4957>>.
- [RFC5580] Tschofenig, H., Ed., Adrangi, F., Jones, M., Lior, A., and B. Aboba, "Carrying Location Objects in RADIUS and Diameter", RFC 5580, DOI 10.17487/RFC5580, August 2009, <<https://www.rfc-editor.org/info/rfc5580>>.
- [RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <<https://www.rfc-editor.org/info/rfc5925>>.
- [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>>.
- [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>>.
- [RFC8177] Lindem, A., Ed., Qu, Y., Yeung, D., Chen, I., and J. Zhang, "YANG Data Model for Key Chains", RFC 8177, DOI 10.17487/RFC8177, June 2017, <<https://www.rfc-editor.org/info/rfc8177>>.

## Authors' Addresses

Keyur Patel  
Arrcus, Inc.  
2077 Gateway Pl  
San Jose, CA 95110  
USA

Email: [keyur@arrcus.com](mailto:keyur@arrcus.com)

Acee Lindem  
Cisco Systems  
301 Midenhall Way  
Cary, NC 95110  
USA

Email: [acee@cisco.com](mailto:acee@cisco.com)

Shawn Zandi  
Linkedin  
222 2nd Street  
San Francisco, CA 94105  
USA

Email: [szandi@linkedin.com](mailto:szandi@linkedin.com)

Gaurav Dawra  
Linkedin  
222 2nd Street  
San Francisco, CA 94105  
USA

Email: [gdawra@linkedin.com](mailto:gdawra@linkedin.com)

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: November 16, 2018

X. Xu  
Alibaba Inc  
K. Bi  
Huawei  
J. Tantsura  
Nuage Networks  
N. Triantafyllis  
LinkedIn  
K. Talaulikar  
Cisco  
May 15, 2018

BGP Neighbor Autodiscovery  
draft-xu-idr-neighbor-autodiscovery-08

Abstract

BGP has been used as the underlay routing protocol in many hyper-scale data centers. This document proposes a BGP neighbor autodiscovery mechanism that greatly simplifies BGP deployments. This mechanism is very useful for those hyper-scale data centers where BGP is used as the underlay routing protocol.

Requirements Language

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

Status of This Memo

This Internet-Draft 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 16, 2018.

## Copyright Notice

Copyright (c) 2018 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 . . . . .	2
2. Terminology . . . . .	3
3. BGP Hello Message Format . . . . .	3
4. Hello Message Procedure . . . . .	10
5. Contributors . . . . .	11
6. Acknowledgements . . . . .	11
7. IANA Considerations . . . . .	12
7.1. BGP Hello Message . . . . .	12
7.2. TLVs of BGP Hello Message . . . . .	12
8. Security Considerations . . . . .	12
9. References . . . . .	13
9.1. Normative References . . . . .	13
9.2. Informative References . . . . .	13
Authors' Addresses . . . . .	14

## 1. Introduction

BGP has been used as the underlay routing protocol instead of IGP in many hyper-scale data centers [RFC7938]. Furthermore, there is an ongoing effort to leverage BGP link-state distribution mechanism to achieve BGP-SPF [I-D.keyupate-lsvr-bgp-spf]. However, BGP is not good as an IGP from the perspective of deployment automation and simplicity. For instance, the IP address and the Autonomous System Number (ASN) of each and every BGP neighbor have to be manually configured on BGP routers although these BGP peers are directly connected. Furthermore, for those BGP routers with multiple physical links being connected, it's usually not ideal to establish BGP sessions over their directly connected interface addresses because the BGP update volume would be unnecessarily increased, meanwhile, it may not be suitable to configure those links as a Link Aggregation Group (LAG) due to some reasons. As a result, it's more common that

loopback interface addresses of those directly connected BGP peers are used for BGP session establishment purpose. To make those loopback addresses of directly connected BGP peers reachable from one another, either static routes have to be configured or some kind of IGP has to be enabled. The former is not good from the network automation perspective while the latter is not good from the network simplification perspective (i.e., running less routing protocols).

This draft specifies a BGP neighbor autodiscovery mechanism by borrowing some ideas from the Label Distribution Protocol (LDP) [RFC5036]. More specifically, directly connected BGP routers could automatically discover each other through the exchange of the to-be-defined BGP Hello messages. The BGP session establishment process as defined in [RFC4271] could be triggered once directly connected BGP neighbors are discovered from one another. Note that the BGP session should be established over the discovered the peering address of the BGP neighbor and in most cases the peering address is a loopback address. In addition, to eliminate the need of configuring static routes or enabling IGP for the loopback addresses, a certain type of routes towards the BGP neighbor's loopback addresses as advertised as peering addresses are dynamically instantiated once the BGP neighbor has been discovered. The administrative distance of such type of routes MUST be smaller than their equivalents that are learnt by the regular BGP update messages. Otherwise, circular dependency would occur once these loopback addresses are advertised via the regular BGP updates.

## 2. Terminology

This memo makes use of the terms defined in [RFC4271].

## 3. BGP Hello Message Format

To automatically discover directly connected BGP neighbors, a BGP router periodically sends BGP HELLO messages out those interfaces on which BGP neighbor autodiscovery are enabled. The BGP HELLO message MUST sent as a UDP packet with a destination port of TBD (179 is the preferred port number value) addressed for the "all routers on this subnet" group multicast address (i.e., 224.0.0.2 in the IPv4 case and FF02::2 in the IPv6 case). The IP source address is set to the address of the interface over which the message is sent out.

The HELLO message contains the following fields:

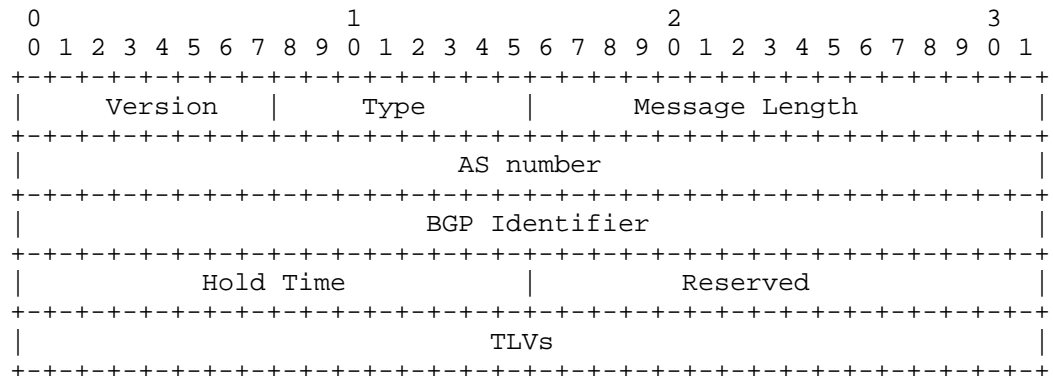


Figure 1: BGP Hello Message

**Version:** This 1-octet unsigned integer indicates the protocol version number of the message. The current BGP version number is 4.

**Type:** The type of BGP message (Hello - TBD value from BGP Message Types Registry)

**Message Length:** This 2-octet unsigned integer specifies the length in octets of the TLVs field.

**AS number:** AS Number of the Hello message sender.

**BGP Identifier:** BGP Identifier of the Hello message sender.

**Hold Time:** Hello hold timer in seconds. Hello Hold Time specifies the time the receiving BGP peer will maintain its record of Hellos from the sending BGP peer without receipt of another Hello. The RECOMMENDED default value is 15 seconds. A value of 0 means that the receiving BGP peer should maintain its record until the link is UP.

**Reserved:** SHOULD be set to 0 by sender and MUST be ignored by receiver.

**TLVs:** This field contains one or more TLVs as described below.

The Accepted ASN List TLV is an optional TLV that is used to signal the AS numbers from which the router would accept BGP sessions. When not signaled, it indicates that the router will accept BGP peering from any ASN from its neighbors. Only a single instance of this TLV is included and its format is shown below.

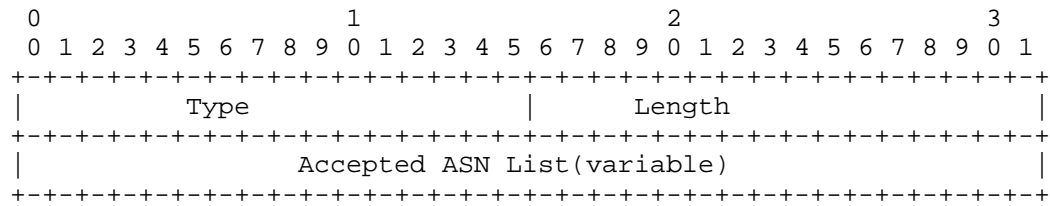


Figure 2: Accepted ASN List TLV

Type: TBD1

Length: Specifies the length of the Value field in octets.

Accepted ASN-List: This variable-length field contains one or more accepted 4-octet ASNs.

The Peering Address TLV is used to indicate to the neighbor the address to which they should establish BGP session. For each peering address, the router can specify its supported AFI/SAFI(s). When the AFI/SAFI values are specified as 0/0, then it indicates that the neighbor can attempt for negotiation of any AFI/SAFIs. The indication of AFI/SAFI(s) in the Peering Address TLV is not intended as an alternative for the MP capabilities negotiation mechanism.

The Peering Address TLV format is shown below and at least one instance of this TLV MUST be present.

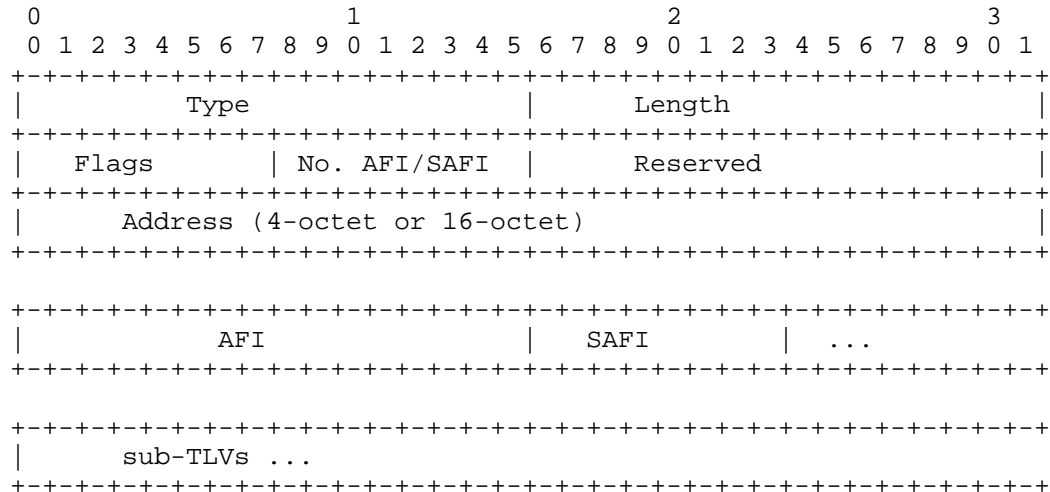


Figure 3: Peering Address TLV

Type: TBD2

Length: Specifies the length of the Value field in octets.

Flags : Current defined bits are as follows. All other bits SHOULD be cleared by sender and MUST be ignored by receiver.

Bit 0x1 - address is IPv6 when set and IPv4 when clear

Number of AFI/SAFI: indicates the number of AFI/SAFI pairs that the router supports on the given peering address.

Reserved: sender SHOULD set to 0 and receiver MUST ignore.

Address: This 4 or 16 octet field indicates the IPv4 or IPv6 address which is used for establishing BGP sessions.

AFI/SAFI : one or more pairs of these values that indicate the supported capabilities on the peering address.

Sub-TLVs : currently none defined

When the Peering Address used is not the directly connected interface address (e.g. when it is a loopback address) then local prefix(es) that cover the peering address(es) MUST be signaled by the router. This allows the neighbor to learn these local prefix(es) and to program routes for them over the directly connected interfaces over which they are being signalled. The Local Prefixes TLV is used to only signal prefixes that are locally configured on the router and its format is as shown below.

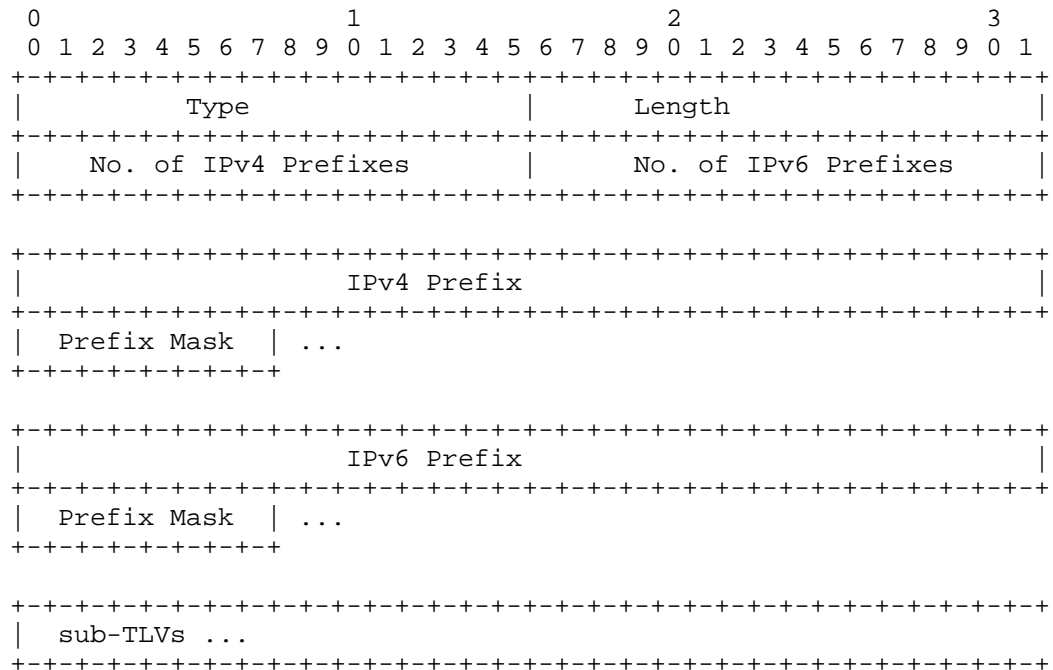


Figure 4: Local Prefixes TLV

Type: TBD3

Length: Specifies the length of the Value field in octets

No. of IPv4 Prefixes : specifies the number of IPv4 prefixes.  
When value is 0, then it indicates no IPv4 Prefixes are present.

No. of IPv6 Prefixes : specifies the number of IPv6 prefixes.  
When value is 0, then it indicates no IPv6 Prefixes are present.

IPv4 Prefix Address & Prefix Mask: Zero or more pairs of IPv4 prefix address and their mask.

IPv6 Prefix Address & Prefix Mask: Zero or more pairs of IPv6 prefix address and their mask.

Sub-TLVs : currently none defined

The Link Attributes TLV is a mandatory TLV that signals to the neighbor the link attributes of the interface on the local router. A single instance of this TLV MUST be present in the message. The Link Attributes TLV is as shown below.

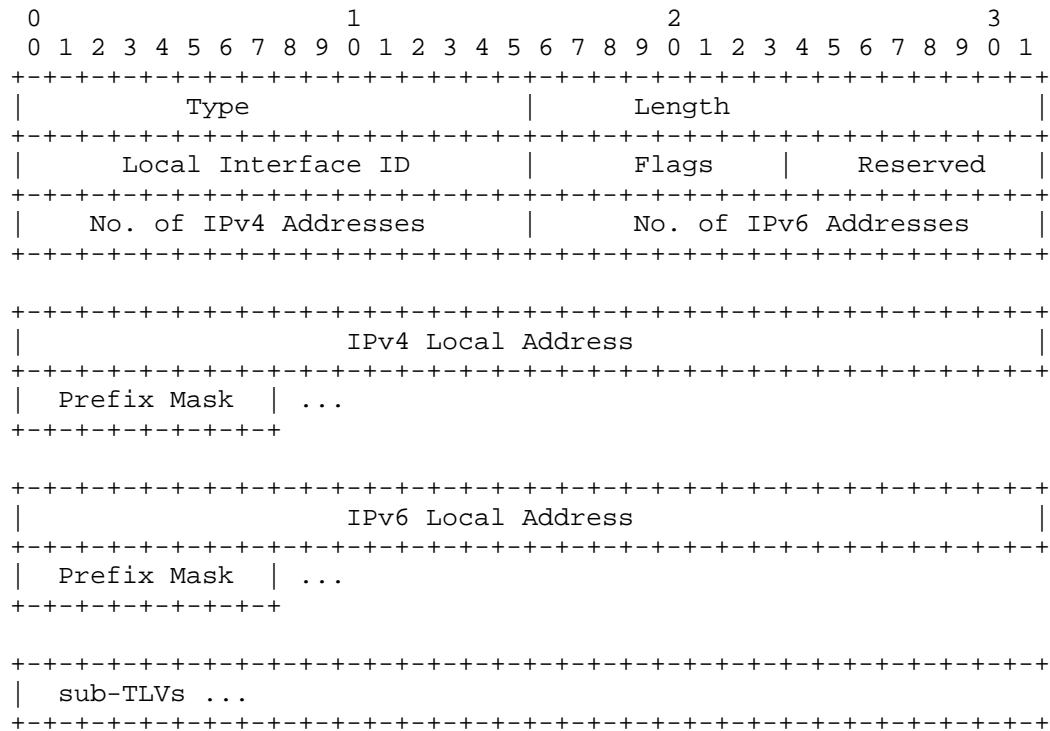


Figure 5: Link Attributes TLV

Type: TBD4

Length: Specifies the length of the Value field in octets

Local Interface ID : the local interface ID of the interface (e.g. the MIB-2 ifIndex)

Flags : Currently defined bits are as follows. Other bits SHOULD be cleared by sender and MUST be ignored by receiver.

Bit 0x1 - indicates link is enabled for IPv4

Bit 0x2 - indicates link is enabled for IPv6

Reserved: SHOULD be set to 0 by sender and MUST be ignored by receiver.

No. of IPv4 Addresses : specifies the number of IPv4 local addresses on the interface. When value is 0, then it indicates no IPv4 Prefixes are present or the interface is IP unnumbered.

No. of IPv6 Addresses : specifies the number of IPv6 Global addresses on the interface. When value is 0, then it indicates no IPv6 Global Prefixes are present or the interface is only configured with IPv6 link-local addresses

IPv4 Address & Mask: Zero or more pairs of IPv4 address and their mask.

IPv6 Address & Mask: Zero or more pairs of IPv6 address and their mask.

Sub-TLVs : currently none defined

The Neighbor TLV is used by a BGP router to indicate the peering address and information about the neighbors that have been discovered by the router on the specific link and their status. The BGP session establishment process begins when both the neighbors accept each other over at least one underlying inter-connecting link between them. The Neighbor TLV format is as shown below.

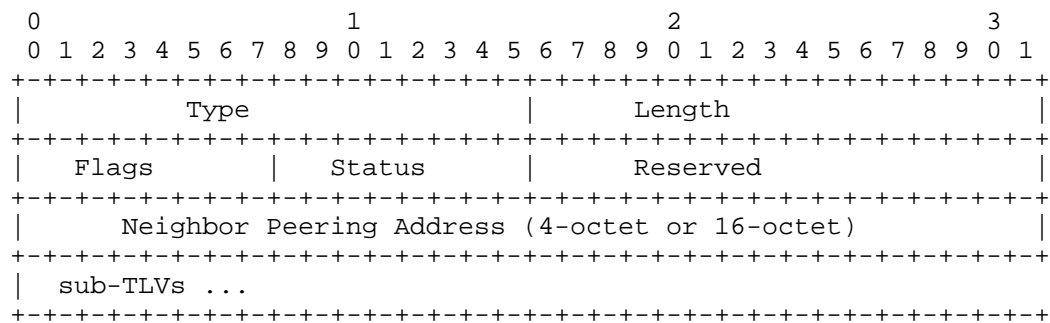


Figure 6: Neighbor TLV

Type: TBD5

Length: Specifies the length of the Value field in octets

Flags : Currently defined 0x1 bit is clear when Peering Address is IPv4 and set when IPv6. Other bits SHOULD be clear by sender and MUST be ignored by receiver.

Status : Indicates the status code of the peering for the particular session over this link. The following codes are currently defined

0 - Indicates 1-way detection of the peer

1 - Indicates rejection of the peer due to local policy reasons (i.e. local router would not be initiating or accepting session to this neighbor)

2 - Indicates 2-way detection of the peering by both neighbors

3 - Indicates that the BGP peering session has been established between the neighbors and that this link would be utilized for forwarding to the peer BGP nexthop

Reserved: SHOULD be set to 0 by sender and MUST be ignored by receiver.

Neighbor Peering Address: This 4 or 16 octet field indicates the IPv4 or IPv6 peering address of the neighbor for which peering status is being reported.

Sub-TLVs : currently none defined

#### 4. Hello Message Procedure

A BGP peer receiving Hellos from another peer maintains a Hello adjacency corresponding to the Hellos. The peer maintains a hold timer with the Hello adjacency, which it restarts whenever it receives a Hello that matches the Hello adjacency. If the hold timer for a Hello adjacency expires the peer discards the Hello adjacency.

We recommend that the interval between Hello transmissions be at most one third of the Hello hold time.

A BGP session with a peer has one or more Hello adjacencies.

A BGP session has multiple Hello adjacencies when a pair of BGP peers is connected by multiple links that have the same connection address (e.g., multiple point-to-point links between a pair of routers). In this situation, the Hellos a BGP peer sends on each such link carry the same Peering Address. In addition, to eliminate the need of configuring static routes or enabling IGP for advertising the loopback addresses, a certain type of routes towards the BGP neighbor's loopback addresses (i.e. carried in the Local Prefixes TLV) could be dynamically created once the BGP neighbor has been discovered. The administrative distance of such type of routes MUST be smaller than their equivalents which are learnt via the normal BGP update messages. Otherwise, circular dependency problem would occur once these loopback addresses are advertised via the normal BGP update messages as well.

BGP uses the regular receipt of BGP Hellos to indicate a peer's intent to keep BGP session identified by the Hello. A BGP peer maintains a hold timer with each Hello adjacency that it restarts when it receives a Hello that matches the adjacency. If the timer expires without receipt of a matching Hello from the peer, BGP concludes that the peer no longer wishes to keep BGP session for that link or that the peer has failed. The BGP peer then deletes the Hello adjacency. The route towards the BGP neighbor's loopback address that had been dynamically created due to that BGP Hello adjacency SHOULD be deleted accordingly. When the last Hello adjacency for an BGP session is deleted, the BGP peer terminates the BGP session and closing the transport connection.

## 5. Contributors

Satya Mohanty  
Cisco  
Email: satyamoh@cisco.com

Shunwan Zhuang  
Huawei  
Email: zhuangshunwan@huawei.com

Chao Huang  
Alibaba Inc  
Email: jingtang.hc@alibaba-inc.com

Guixin Bao  
Alibaba Inc  
Email: guixin.bgx@alibaba-inc.com

Jinghui Liu  
Ruijie Networks  
Email: liujh@ruijie.com.cn

Zhichun Jiang  
Tencent  
Email: zcjiang@tencent.com

## 6. Acknowledgements

The authors would like to thank Enke Chen for his valuable comments and suggestions on this document.

## 7. IANA Considerations

### 7.1. BGP Hello Message

This document requests IANA to allocate a new UDP port (179 is the preferred number ) and a BGP message type code for BGP Hello message.

Value	TLV Name	Reference
-----	-----	-----
	Service Name: BGP-HELLO	
	Transport Protocol(s): UDP	
	Assignee: IESG <iesg@ietf.org>	
	Contact: IETF Chair <chair@ietf.org>.	
	Description: BGP Hello Message.	
	Reference: This document -- draft-xu-idr-neighbor-autodiscovery.	
	Port Number: TBD1 (179 is the preferred value) -- To be assigned by IANA.	

### 7.2. TLVs of BGP Hello Message

This document requests IANA to create a new registry "TLVs of BGP Hello Message" with the following registration procedure:

Registry Name: TLVs of BGP Hello Message.

Value	TLV Name	Reference
-----	-----	-----
0	Reserved	This document
1	Accepted ASN List	This document
2	Peering Address	This document
3	Local Prefixes	This document
4	Link Attributes	This document
5	Neighbor	This document
6-65500	Unassigned	
65501-65534	Experimental	This document
65535	Reserved	This document

## 8. Security Considerations

For security purposes, BGP speakers usually only accept TCP connection attempts to port 179 from the specified BGP peers or those within the configured address range. With the BGP neighbor auto-discovery mechanism, it's configurable to enable or disable sending/receiving BGP hello messages on the per-interface basis and BGP hello messages are only exchanged between physically connected peers that are trustworthy. Therefore, the BGP neighbor auto-discovery mechanism doesn't introduce additional security risks associated with BGP.

In addition, for the BGP sessions with the automatically discovered peers via the BGP hello messages, the TTL of the TCP/BGP messages (dest port=179) MUST be set to 255. Any received TCP/BGP message with TTL being less than 254 MUST be dropped according to [RFC5082].

## 9. References

### 9.1. 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>>.
- [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>>.
- [RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "LDP Specification", RFC 5036, DOI 10.17487/RFC5036, October 2007, <<https://www.rfc-editor.org/info/rfc5036>>.
- [RFC5082] Gill, V., Heasley, J., Meyer, D., Savola, P., Ed., and C. Pignataro, "The Generalized TTL Security Mechanism (GTSM)", RFC 5082, DOI 10.17487/RFC5082, October 2007, <<https://www.rfc-editor.org/info/rfc5082>>.
- [RFC8279] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Przygienda, T., and S. Aldrin, "Multicast Using Bit Index Explicit Replication (BIER)", RFC 8279, DOI 10.17487/RFC8279, November 2017, <<https://www.rfc-editor.org/info/rfc8279>>.

### 9.2. Informative References

- [I-D.keyupate-lsvr-bgp-spf] Patel, K., Lindem, A., Zandi, S., and W. Henderickx, "Shortest Path Routing Extensions for BGP Protocol", draft-keyupate-lsvr-bgp-spf-00 (work in progress), March 2018.
- [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>>.

Authors' Addresses

Xiaohu Xu  
Alibaba Inc

Email: xiaohu.xxh@alibaba-inc.com

Kunyang Bi  
Huawei

Email: bikunyang@huawei.com

Jeff Tantsura  
Nuage Networks

Email: jefftant.ietf@gmail.com

Nikos Triantafyllis  
LinkedIn

Email: nikos@linkedin.com

Ketan Talaulikar  
Cisco

Email: ketant@cisco.com

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: May 29, 2020

X. Xu  
Alibaba Inc  
K. Talaulikar  
Cisco Systems  
K. Bi  
Huawei  
J. Tantsura  
Apstra  
N. Triantafyllis  
Amazon Web Services  
November 26, 2019

BGP Neighbor Discovery  
draft-xu-idr-neighbor-autodiscovery-12

Abstract

BGP is being used as the underlay routing protocol in some large-scaled data centers (DCs). Most popular design followed is to do hop-by-hop external BGP (EBGP) session configurations between neighboring routers on a per link basis. The provisioning of BGP neighbors in routers across such a DC brings its own operational complexity.

This document introduces a BGP neighbor discovery mechanism that greatly simplifies BGP operations in such DC and other networks by automatic setup of BGP sessions between neighbor routers using this mechanism.

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 May 29, 2020.

## Copyright Notice

Copyright (c) 2019 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. Requirements Language . . . . .	4
2. Terminology . . . . .	4
3. Applicability . . . . .	4
4. Requirements . . . . .	4
5. Overview . . . . .	6
6. UDP Message Header . . . . .	7
7. Hello Message Format . . . . .	8
8. Hello Message TLVs . . . . .	10
8.1. Accepted ASN List TLV . . . . .	10
8.2. Peering Address TLV . . . . .	11
8.3. Local Prefix TLV . . . . .	13
8.4. Link Attributes TLV . . . . .	14
8.5. Neighbor TLV . . . . .	17
8.6. Cryptographic Authentication TLV . . . . .	18
9. Neighbor Discovery Procedure . . . . .	20
9.1. Interface Procedures . . . . .	20
9.2. Adjacency State Machine . . . . .	21
9.2.1. Down State . . . . .	22
9.2.2. Initial State . . . . .	22
9.2.3. 1-Way State . . . . .	22
9.2.4. 2-Way State . . . . .	23
9.2.5. Adj-Reject State . . . . .	23
9.2.6. Adj-OK State . . . . .	24
9.2.7. Accepted State . . . . .	24
9.3. Adjacency Route . . . . .	25
10. Interactions with Base BGP Protocol . . . . .	26
11. Security Considerations . . . . .	27
12. Manageability Considerations . . . . .	28
12.1. Operational Considerations . . . . .	28
12.2. Management Considerations . . . . .	29

13. IANA Considerations . . . . .	29
13.1. BGP Hello Message . . . . .	30
13.2. TLVs of BGP Hello Message . . . . .	30
14. Acknowledgements . . . . .	30
15. Contributors . . . . .	30
16. References . . . . .	31
16.1. Normative References . . . . .	31
16.2. Informative References . . . . .	32
Authors' Addresses . . . . .	33

## 1. Introduction

BGP is being used as the underlay routing protocol instead of link-state routing protocols like IS-IS and OSPF in some large-scale data centers (DCs). [RFC7938] describes the design, configuration and operational aspects of using BGP in such networks. The most popular design scheme involves the setup of external BGP (EBGP) sessions over individual links between directly connected routers using their interface addresses. Such BGP neighbor provisioning requires configuration of the neighbor IP address and Autonomous System (AS) Number (ASN) for BGP neighbor on each and every link of every BGP router. As a DC fabric comprising of topology described in [RFC7938] grows with addition of new leafs, spines, and links between them, the BGP provisioning needs to be carefully updated. Unlike with the link-state protocols, in the case of BGP, there is no automatic discovery of neighbors and route exchange between them by simply adding links and nodes of the fabric into the routing protocol operation.

In some DC designs with BGP, multiple links are added between a leaf and spine to add additional bandwidth. Use of link-aggregation at Layer 2 level may not be always desirable in such cases due to the risk of flow polarization on account of a mix of ECMP at Layer 2 and Layer 3 levels. In such cases, one option is for EBGP sessions to be setup between two BGP neighbors over each of the links between them. In such a case, the BGP session scale and the resultant increase in update processing may pose scalability challenges. A second option is for a single EBGP session to be setup between the loopback IP addresses between the neighbor and then configure some static routes for loopback reachability over the underlying links. This option introduces an additional provisioning task for the static routes.

Furthermore, there is also a need for BGP to be able to describe its links and its neighbors on its directly connected links and export this information via BGP-LS [RFC7752] to provide a detail link-level topology view of a data center running BGP. The ability of BGP in discovering its neighbors over its links, monitoring their liveness and learning the link attributes (such as addresses) is required for

the conveying the link-state topology in such a BGP network. This information can be leveraged by the BGP-SPF proposal [I-D.ietf-lsvr-bgp-spf] which introduces link-state routing capabilities in BGP. This information can also be leveraged to convey the link-state topology in a network running traditional BGP routing using BGP-LS as described in [I-D.ketant-idr-bgp-ls-bgp-only-fabric] and to enabled end to end traffic engineering use-cases spanning across DCs and the core/access networks.

### 1.1. 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.

## 2. Terminology

This document makes use of the terms defined in [RFC4271] and [RFC7938] .

## 3. Applicability

The applicability of the BGP Neighbor Discovery mechanism described in this document is limited to deployments where BGP is used as routing protocol between directly connected routers and when there is a requirement for automatic setup of BGP peering between them.

- o In DC networks where BGP is used as a hop-by-hop routing protocol [RFC7938].
- o In metro networks where access aggregation topologies are architected as a CLOS topology (or similar other networks) and BGP is used as a hop-by-hop routing protocol.

While this document uses EBGp examples, the mechanism is equally applicable in designs that use IBGP similarly for hop-by-hop routing.

The applicability of the BGP Neighbor Discovery mechanism to any other BGP protocol deployment is outside the scope of this document.

## 4. Requirements

This section describe the requirements for the BGP hop-by-hop routing deployments that were considered for the definition of the BGP Neighbor Discovery extensions proposed in this document..

Following are the key requirements related for the BGP neighbor discovery process:

1. It should perform discovery of directly connected BGP routers. Mechanism should support either IPv4 or IPv6 or a dual stack design and it should be generic for any link-layer.
2. It should include exchange of BGP peering addresses (IPv4 or IPv6 or both) that routers can use to automatically setup BGP TCP peering between themselves. The mechanism should leverage the existing capability negotiation process performed as part of the BGP TCP session establishment.
3. When BGP peering is desired to be performed over loopback addresses of the routers, then the mechanism should automatically setup reachability to the loopback over one or more underlying directly connected links between them. In this scenario, the mechanism should also provide resolution for the BGP next-hop address (i.e. the loopback address) for the BGP routes exchanged over these sessions between the loopback addresses.
4. Mechanism should enable exchange of link-level information such as IP addresses and link attributes between the directly connected BGP routers. It should be extensible to include other information in the future.
5. Mechanism should be limited to link scope for security and use link-local addressing only. Cryptographic mechanisms should be also provided for additional security.
6. Mechanism should support capabilities for performing optional validation of parameters to detect misconfiguration (e.g. link address subnet mismatch, peering between incorrect AS, etc.) in an extensible manner before going on to use the link and the setup of the BGP TCP peering session over it.
7. The mechanism should not affect or change the BGP TCP session establishment procedures and the BGP routing exchange over the TCP session other than the interactions for triggering the setup/removal of peer session that is based on discovery mechanism.
8. The mechanism should leverage existing fast-detection techniques for failures that are used currently for EBGp sessions over directly connected links like fast-external-failover and BFD.
9. The mechanism should focus on the discovery process and exchange of status as a control plane procedure and be sufficiently loosely coupled with the base BGP operations to enable

implementations to ensure scalability of BGP operations when using the discovery procedures.

## 5. Overview

At a high level, this specification introduces the use of UDP based BGP Hello messages to be exchanged between directly connected BGP routers for neighbor discovery.

1. Information is exchanged between BGP routers on a per link basis leading to discovery of each others peering address and other information.
2. The TCP session establishment for the BGP protocol operation and the BGP routing exchange over these sessions can then follow without any change/modification from the existing BGP protocol operations as specified in [RFC4271].
3. As part of the neighbor information exchange the route to a neighbor's peering address is also automatically setup pointing over the links over which the neighbor is discovered.
4. This route is used for both the BGP TCP session establishment as well as for resolution of the BGP next-hop (NH) for the routes learnt via the neighbor instead of an underlying IGP or static route.

This document prefers the use of an extension to BGP protocol since the deployments and use-cases targeted (i.e. large-scale DCs) are already running BGP as their routing protocol. Extending BGP with neighbor discovery capabilities is operationally and implementation wise a simpler approach than requiring a new or an additional protocol to be first extended to do this functionality (to exchange BGP-specific parameters) and then also integrated its operations with BGP protocol operations.

The BGP Neighbor discovery mechanism is a control plane mechanism intended to discovery and maintain the BGP router's adjacencies with its neighbors over directly connected links. Maintaining an adjacency also involves detecting any changes in parameters using periodic messages and triggering corresponding actions based on the change. Such actions also include removal of the BGP TCP peering for an auto discovered peering session based on the neighbor discovery. However, the mechanism is not intended for a fast liveness detection of neighbor and existing mechanisms for this purpose such as BFD [RFC5880] may be leveraged.

The BGP Neighbor discovery mechanism is scoped to a link and works using link-local addressing. In a BGP DC network that is using IPv6 in the fabric underlay, it is possible that no IPv6 global addresses are assigned to the interfaces between the nodes and the IPv6 Global address(es) are assigned only to the loopback interfaces of these nodes. The Neighbor discovery mechanism enables the setup of BGP peering using the IPv6 Global addresses on the loopback interfaces and hop by hop routing with just IPv6 link-local addresses on the interfaces. Such a design eases introduction of nodes in the fabric and links between them from a provisioning aspect. In a deployment with IPv4 addressing, IP unnumbered could be similarly used for all the links between the nodes using the IPv4 address assigned to the loopback interfaces on those nodes.

The BGP neighbor discovery mechanism defined in this document borrows ideas from the Label Distribution Protocol (LDP) [RFC5036]. However, most importantly, only the concept of link-local signaling based neighbor discovery is borrowed while the discovery aspect for targeted LDP sessions does not apply to this BGP neighbor discovery mechanism.

The further sections in this document first describe the newly introduced message formats and TLVs and then go on to describe the procedures of BGP neighbor discovery and its integration with the base BGP protocol mechanism as specified in [RFC4271].

The operational and management aspects of the BGP neighbor discovery mechanism are described in Section 12.

## 6. UDP Message Header

The BGP neighbor discovery mechanism will operate using UDP messages. The UDP port of TBD (179 is the preferred port number to be assigned as specified in Section 13) is used which is same as the TCP port 179 used by BGP. The BGP UDP message common header format is specified as follows:

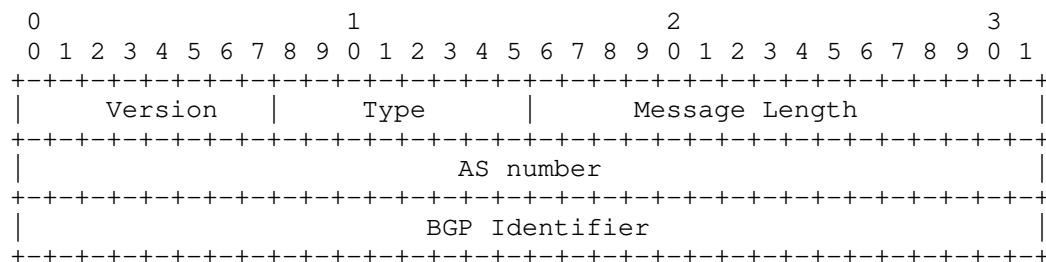


Figure 1: BGP UDP Message Header

**Version:** This 1-octet unsigned integer indicates the protocol version number of the message. The current BGP version number is 4.

**Type:** The type of BGP message

**Message Length:** This 2-octet unsigned integer specifies the length in octets of the entire BGP UDP message including the header.

**AS number:** AS Number of the UDP message sender.

**BGP Identifier:** BGP Identifier of the UDP message sender.

BGP UDP messages can be sent using either IPv4 or IPv6 depending on the address used for session establishment and provisioned on the interfaces over which these messages are sent.

## 7. Hello Message Format

A BGP router uses UDP based Hello messages to discover directly connected BGP neighbors over those interfaces enabled for Neighbor Discovery. The BGP Hello messages for the Neighbor Discovery procedure are used for link-locally signaling and hence MUST be addressed to the "all routers on this subnet" group multicast address (i.e., 224.0.0.2 in the IPv4 case and FF02::2 in the IPv6 case) and the TTL for the IP packets SHOULD be set to 1. The IP source address MUST be set to the address of the interface over which the message is sent out which would be the primary interface address or unnumbered address in the IPv4 case and the IPv6 link-local address on the interface in the IPv6 case.

The Hello message format is as follows:

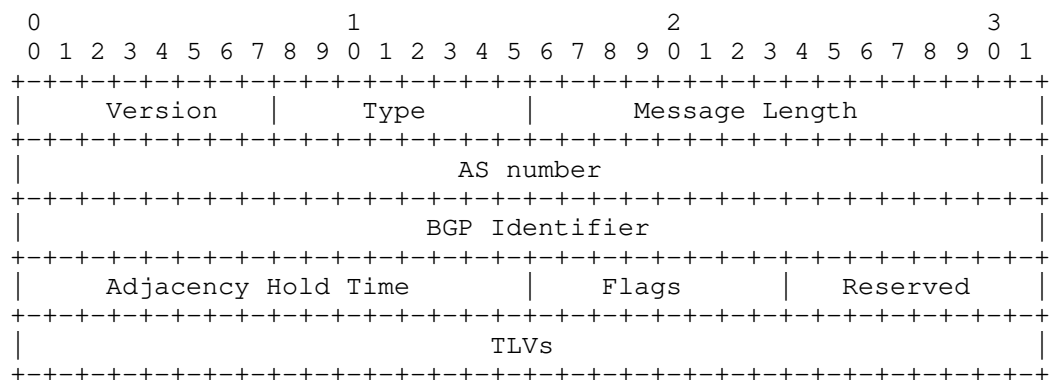


Figure 2: BGP Hello Message

**Version:** This 1-octet unsigned integer indicates the protocol version number of the message. The current BGP version number is 4.

**Type:** The type of BGP message (Hello - TBD value from BGP Message Types Registry)

**Message Length:** This 2-octet unsigned integer specifies the length in octets of the TLVs field.

**AS number:** AS Number of the BGP router sending the Hello message.

**BGP Identifier:** BGP Identifier of the BGP router sending the Hello message.

**Adjacency Hold Time:** Hello adjacency hold timer in seconds. Adjacency Hold Time specifies the time, for which the receiving BGP neighbor router SHOULD maintain adjacency state for it, without receipt of another Hello. A value of 0 means that the receiving BGP peer should immediately mark that the adjacency to the sender is going down.

**Flags :** Current defined bits are as follows. All other bits SHOULD be cleared by sender and MUST be ignored by receiver.

```

 0 1 2 3 4 5 6 7
+---+---+---+---+
|S|           |
+---+---+---+---+

```

where:

S bit - indicates that this is a State Change Hello message when SET and normal periodic Hello message when CLEAR

**Reserved:** SHOULD be set to 0 by sender and MUST be ignored by receiver.

**TLVs:** This field contains one or more TLVs as described below.

BGP HELLO messages can be sent using either IPv4 or IPv6 addresses depending on the addressing used for session establishment and provisioned on the interfaces over which these messages are sent. When both IPv4 and IPv6 is enabled on the interface, then IPv6 address SHOULD be used. Implementations MAY provide an option to override the choice of address family to be used. The choice of address family to be used MUST be consistent on all BGP routers on a given link for neighbor discovery.

Based on the setting of the S flag, there are two variants of the Hello message:

1. State Change Hello Message : these Hello messages include TLVs which convey the state and parameters of the local interface and adjacency to other routers on the link. They are generated only when there is a change in state of the adjacency or some parameter at the interface level.
2. Periodic Hello Message : these are the normal periodic Hello messages which do not include TLVs and are used to maintain the adjacency on the link during steady state conditions.

These Hello message variants are intended to limit the exchange of information and state via TLVs to only those periods where necessary while using lightweight Hello messages during steady state. This simplifies the Hello message processing and improves scalability of the discovery mechanism.

The neighbor discovery procedure using the Hello message is described in Section 9 and its relation with the BGP Keepalives and Hold Timer for the TCP session is described in Section 10.

## 8. Hello Message TLVs

The BGP Hello message carries TLVs as described in this section that enable exchange of information on a per interface basis between directly connected BGP neighbors. These messages enable the neighbor discovery process.

### 8.1. Accepted ASN List TLV

The Accepted ASN List TLV is an optional TLV that is used to signal an unordered list of AS numbers from which the BGP router would accept BGP sessions. When not signaled, it indicates that the router will accept BGP peering from any ASN from its neighbors. Indicating the list of ASNs, helps avoid the neighbor discovery process getting stuck in a 1-way state where one side keeps attempting to setup adjacency while the other does not accept it due to incorrect ASN.

The operational and management aspects of this ASN based policy control for BGP neighbor discovery are described further in Section 12.

This TLV SHOULD NOT be included in a Hello message with the S bit CLEAR. More than a single instance of this TLV MUST NOT be included in a Hello message. If a router receives multiple instances of this

TLV then it should only consider the first instance in the sequence and ignore the rest.

The format of this TLV is shown below

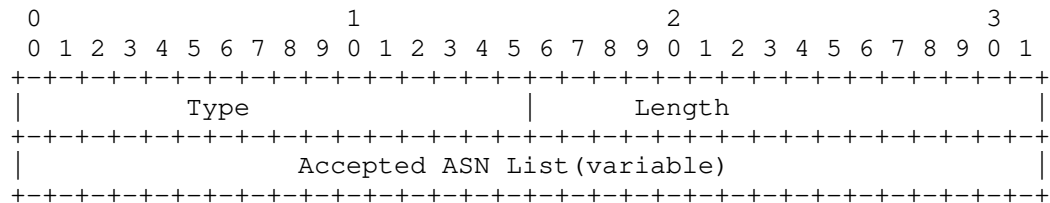


Figure 3: Accepted ASN List TLV

Type: TBD1

Length: Specifies the length of the Value field in octets (in multiple of 4)

Accepted ASN-List: This variable-length field contains one or more accepted 4-octet ASNs.

## 8.2. Peering Address TLV

The Peering Address TLV is used to indicate to the neighbor the address to be used for setting up the BGP TCP session. Along with the peering address, the router can specify its supported AFI/SAFI(s). When the AFI/SAFI values are specified as 0/0, then it indicates that the neighbor can attempt for negotiation of any AFI/SAFIs. The indication of AFI/SAFI(s) in the Peering Address TLV is not intended as an alternative for the MP capabilities negotiation mechanism done as part of the BGP TCP session establishment.

Multiple instances of this TLV MAY be included in the Hello message, one for each peering address (e.g. IPv4 and IPv6 or multiple IPv4 addresses for different AFI/SAFI sessions). When multiple peering addresses are provisioned, then the indication helps the router select the appropriate peer address of the neighbor based on its local peering address profile by matching the supported AFI/SAFIs.

This TLV is essential for the setting up of the TCP peering between BGP neighbors using the neighbor discovery mechanism. When a BGP router stops including a Peer Address in its State Change Hello messages, then it is no longer accepting TCP peering sessions to that address and the neighbor SHOULD clean up any peering session that was setup to that address via the discovery mechanism.

Implementations SHOULD support the signaling of an interface IP address in the Peering Address TLV and perform the BGP TCP session establishment using interface addresses (i.e. the neighbor discovery mechanism is not limited to the use of loopback addresses for the peering session establishment). Implementations MAY support the signaling of IPv6 Link Local addresses using the Peering Address TLV and using the same for the BGP TCP session setup.

This TLV SHOULD NOT be included in a Hello message with the S bit CLEAR.

The Peering Address TLV format is shown below.

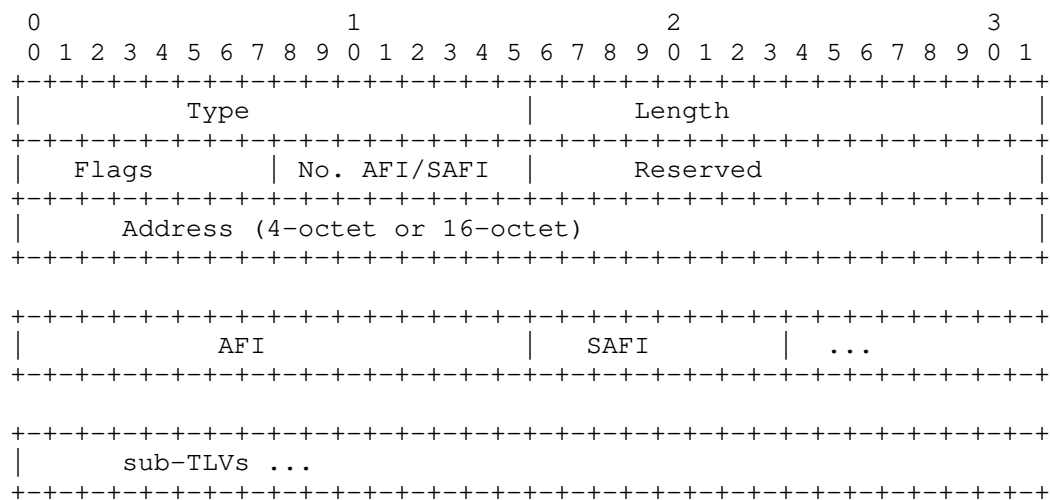
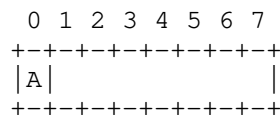


Figure 4: Peering Address TLV

Type: TBD2

Length: Specifies the length of the Value field in octets.

Flags : Current defined bits are as follows. All other bits SHOULD be cleared by sender and MUST be ignored by receiver.



where:

A bit - address is IPv6 when SET and IPv4 when CLEAR

Number of AFI/SAFI: indicates the number of AFI/SAFI pairs that the router supports on the given peering address.

Reserved: sender SHOULD set to 0 and receiver MUST ignore.

Address: This 4 or 16 octet field indicates the IPv4 or IPv6 address which is used for establishing BGP sessions.

AFI/SAFI : one or more pairs of these values that indicate the supported capabilities on the peering address.

Sub-TLVs : optional and currently none defined

### 8.3. Local Prefix TLV

BGP neighbor discovery mechanism, in certain scenarios, requires a BGP router to program a route in its local routing table for a prefix belonging to its neighbor router. On such scenario is when the BGP TCP peering is to be setup between the loopback addresses on the neighboring routers. This requires that the routers have reachability to their each other's loopback addresses before the TCP session can be brought up.

The Local Prefix TLV is an optional TLV which enables a BGP router to explicitly signal its local prefix to its neighbor for setting up of such a local routing entry pointing over the underlying link over which it is being signaled. This enables the BGP router to have control over the specific links over which its neighbor that may reach it for the specific local prefix. The details of the procedure for programming of the route corresponding to the prefix signaled using the Local Prefix TLV is described in Section 9.3..

Multiple instances of the Local Prefix TLV MAY be included in the Hello message with each carrying a specific prefix in it. This TLV SHOULD NOT be included in a Hello message with the S bit CLEAR.

The Local Prefix TLV format is as shown below.

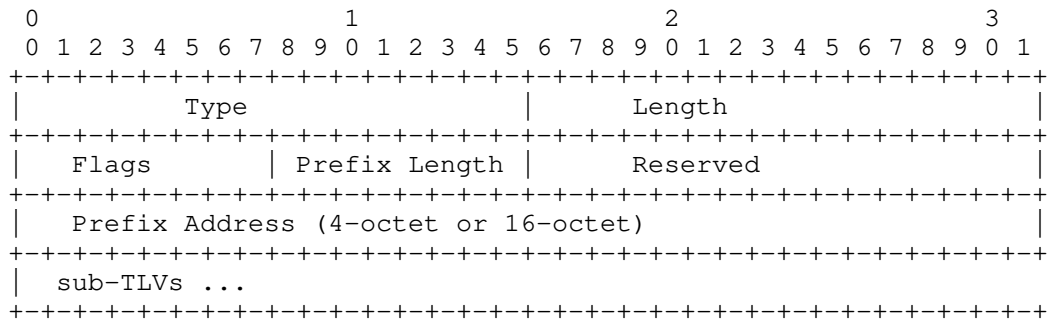
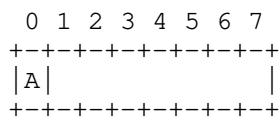


Figure 5: Local Prefix TLV

Type: TBD3

Length: Specifies the length of the Value field in octets

Flags : Current defined bits are as follows. All other bits SHOULD be cleared by sender and MUST be ignored by receiver.



where:

A bit - address is IPv6 when SET and IPv4 when CLEAR

Prefix Length: specifies the Prefix length

Reserved: sender SHOULD set to 0 and receiver MUST ignore.

Prefix Address: This 4 or 16 octet field indicates the IPv4 or IPv6 prefix address.

Sub-TLVs : optional and currently none defined

#### 8.4. Link Attributes TLV

The Link Attributes TLV is a mandatory TLV in a State Change Hello message that signals to the neighbor the link attributes of the interface on the local router. One and only one instance of this TLV MUST be included in the State Change Hello message. A State Change Hello message without this TLV included MUST be discarded and an error logged for the same.

This TLV enables a BGP router to learn all its neighbors IP addresses on the specific link as well as it's link identifier. When the interface is IPv4 enabled, all the IPv4 addresses configured on it are included in this TLV. IPv4 unnumbered address is not included in this TLV and no IPv4 address would be included for the interface in such cases. When the interface is IPv6 enabled, all the IPv6 global addresses configured on the interface are included in this TLV. IPv6 link-local addresses are not included in this TLV. In case of an interface running dual stack, both IPv4 and IPv6 addresses are included in this TLV irrespective of the address family that is used for UDP message exchange.

Additional sub-TLVs may be defined in the future to exchange other link attributes between BGP neighbors. This TLV SHOULD NOT be included in a Hello message with the S bit CLEAR.

The Link Attributes TLV format is as shown below.

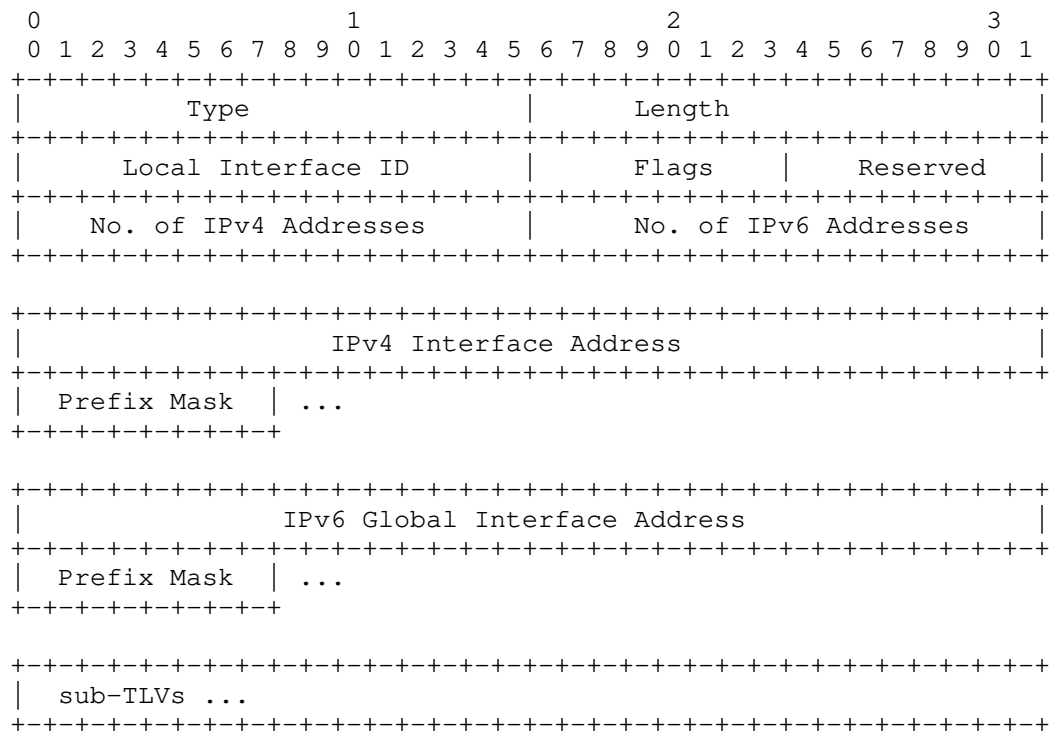


Figure 6: Link Attributes TLV

Type: TBD4

Length: Specifies the length of the Value field in octets

Local Interface ID : the local interface ID of the interface (refer unnumbered link section of [RFC2104] e.g. the MIB-2 ifIndex). This helps uniquely identify the link even when there are multiple links between two neighbors using IPv4 unnumbered address or only having IPv6 link-local addresses.

Flags : Currently defined bits are as follows. Other bits SHOULD be cleared by sender and MUST be ignored by receiver.

```

 0 1 2 3 4 5 6 7
+---+---+---+---+---+---+
| I | V | B |           |
+---+---+---+---+---+---+

```

where:

I bit - indicates link is enabled for IPv4

V bit - indicates link is enabled for IPv6

B bit - indicates support for BFD monitoring [RFC5880] over the link

Reserved: SHOULD be set to 0 by sender and MUST be ignored by receiver.

No. of IPv4 Addresses : specifies the number of IPv4 addresses on the interface. When value is 0, then it indicates no IPv4 Prefixes are present or the interface is IPv4 unnumbered if it is enabled for IPv4

No. of IPv6 Addresses : specifies the number of IPv6 global addresses on the interface. When value is 0, then it indicates no IPv6 Global Prefixes are present and the interface is only configured with IPv6 link-local addresses if it is enabled for IPv6.

IPv4 Address & Mask: Zero or more pairs of IPv4 address and their mask.

IPv6 Address & Mask: Zero or more pairs of IPv6 address and their mask.

Sub-TLVs : optional and currently none defined

## 8.5. Neighbor TLV

The Neighbor TLV is used by a BGP router to indicate its Hello adjacency state with its neighboring router(s) on the specific link. The neighbor is identified by its AS Number and BGP Identifier. The router MUST include the Neighbor TLV for each of its discovered neighbors on that link irrespective of its status.

The usage of the Neighbor TLV is described in detail in Section 9. This TLV SHOULD NOT be included in a Hello message with the S bit CLEAR.

The Neighbor TLV format is as shown below.

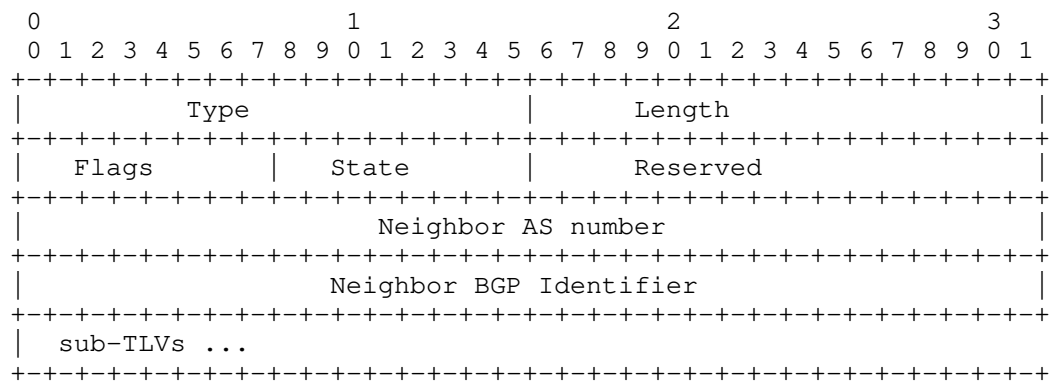
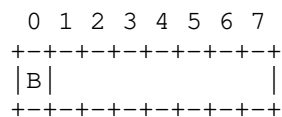


Figure 7: Neighbor TLV

Type: TBD5

Length: Specifies the length of the Value field in octets

Flags : Current defined bits are as follows. All other bits SHOULD be cleared by sender and MUST be ignored by receiver.



where:

B bit - When SET with the adjacency state not in Accepted state indicates that the adjacency is not accepted due to BFD down.

State : Indicates the state code of the adjacency state machine (refer to Section 9.2 for details) for the neighbor over this link. The following codes are currently defined

- 0 - Down (not to be used as state in this TLV)
- 1 - Initial (not to be used as state in this TLV)
- 2 - 1-way
- 3 - 2-way
- 4 - Adj-Reject
- 5 - Adj-OK
- 6 - Accepted

Reserved: SHOULD be set to 0 by sender and MUST be ignored by receiver.

Neighbor AS number: AS Number of the neighbor BGP router as signaled in its Hello message.

Neighbor BGP Identifier: BGP Identifier of the neighbor BGP router as signaled in its Hello message.

Sub-TLVs : currently none defined

#### 8.6. Cryptographic Authentication TLV

The Cryptographic Authentication TLV is an optional TLV that is used as part of an authentication mechanism for BGP Hello message by securing against spoofing attacks. It also introduces a cryptographic sequence number carried in the Hello messages that can be used to protect against replay attacks. Using this Cryptographic Authentication TLV, one or more secret keys (with corresponding Security Association (SA) IDs) are configured on each BGP router. For each BGP Hello message, the key is used to generate and verify an HMAC Hash that is stored in the Cryptographic Authentication TLV. For the cryptographic hash function, this document proposes to use SHA-1, SHA-256, SHA-384, and SHA-512 defined in US NIST Secure Hash Standard (SHS) [FIPS-180-4]. The HMAC authentication mode defined in [RFC2104] is used. Of the above, implementations MUST include support for at least HMAC-SHA-256, SHOULD include support for HMAC-SHA-1, and MAY include support for HMAC-SHA-384 and HMAC-SHA-512.

Further details for ensuring the security of the BGP Hello UDP messages are described in Section 11.

The Cryptographic Authentication TLV format is as shown below.

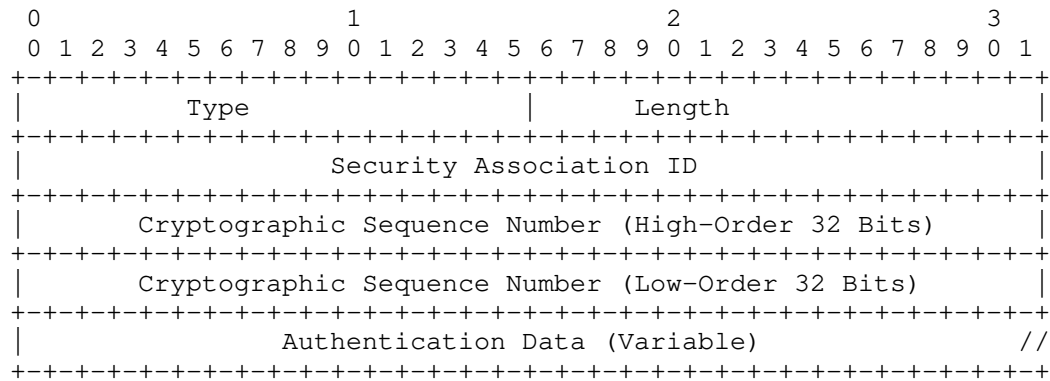


Figure 8: Cryptographic Authentication TLV

Type: TBD6

Length: Specifies the length of the Value field in octets

Security Association ID: The 32-bit field that maps to the authentication algorithm and the secret key used to create the message digest carried in Hello message payload.

Cryptographic Sequence Number: The 64-bit, strictly increasing sequence number that is used to guard against replay attacks. The 64-bit sequence number MUST be incremented for every BGP Hello message sent by the BGP router. Upon reception, the sequence number MUST be greater than the sequence number in the last BGP Hello message accepted from the sending BGP neighbor. Otherwise, the BGP hello message is considered a replayed packet and is dropped. The Cryptographic Sequence Number is a single space per BGP router.

Authentication Data: This field carries the digest computed by the Cryptographic Authentication algorithm in use. The length of the Authentication Data varies based on the cryptographic algorithm in use, which is shown below:

HMAC-SHA1 20 bytes

HMAC-SHA-256 32 bytes

HMAC-SHA-384 48 bytes

HMAC-SHA-512 64 bytes

## 9. Neighbor Discovery Procedure

The neighbor discovery mechanism in BGP is implemented with the introduction of an Interface state in BGP and an Adjacency Finite State Machine (FSM). This section describes the states, FSM and procedures involved.

### 9.1. Interface Procedures

In order to perform neighbor discovery, BGP needs to maintain state for the subset of its connected interfaces over which neighbor discovery is enabled. For these interfaces, BGP sends its Hello messages, including the TLVs described in Section 8, as long as its link is UP. The Neighbor TLV described in Section 8.5 is, included once a neighbor is discovered as described in Section 9.2 .

The Hello messages MUST be originated periodically at an interval which is less than or equal to one third of the Adjacency Hold Time indicated by the router in its Hello message. The RECOMMENDED default value for the Adjacency Hold Time is 45 seconds which makes the hello message interval to be 15 seconds. Periodic Hello messages ensure robustness of the neighbor discovery mechanism against transient loss of hello messages that are sent over unreliable UDP messaging channel and also enable detection of neighbor down events over specific links. Periodic Hello messages that do not convey any change in state SHOULD exclude TLVs that signal the local interface or adjacency state and have the S bit CLEAR as specified in Section 7.

A State Change Hello message MUST be triggered, without waiting for the periodic timer expiry, whenever there is a change in the router's Hello TLVs' content that needs to be signaled to its neighbor over the specific link. A State Change Hello message MUST also be triggered when a new neighbor's Hello message is first received or change is detected in the neighbor's Hello TLV's that results in change in its adjacency state. Once a State Change Hello message is triggered on a specific interface, the router MUST continue to generate State Change Hello messages on it with the necessary TLVs included at periodic hello message intervals for a period of time that is at least equal to the Adjacency Hold Time. This ensures that messages carrying the updated information and local state changes are not lost. The router can switch back to Periodic Hello messages

after it has transmitted State Change Hello messages with the latest TLV contents for the Adjacency Hold Time period.

When a router receives a Hello message from its neighbor, it MUST restart the Adjacency Hold timer that it is maintaining for the neighbor adjacency using the value indicated in the Hello message. When the message is of type State Change (i.e. with S bit SET), it additionally needs to process all the TLVs included and verify the signaled state against what was conveyed in the previous State Change Hello message from the same neighbor. Any changed identified would trigger the adjacency FSM change as described in Section 9.2.

When a router does not receive a Hello message from its neighbor for a period equal to Adjacency Hold Time, then it MUST treat this as an adjacency down event and clean up its adjacency state to this neighbor as described in Section 9.2.

Before the interface is shut or the neighbor discovery mechanism is disabled on it, the router SHOULD attempt to send out immediate Hello messages, with the S bit CLEAR (i.e. not including state related TLVs) and with Adjacency Hold Time set to 0, to trigger the adjacency down event on its neighbors. It MUST then clean up its own adjacency states on that specific link.

When either the BGP Identifier or the AS number are modified, then the router MUST send out a triggered Hello message, with the S bit CLEAR and with Adjacency Hold Time set to 0 using the old BGP Identifier and AS number values, over all the links enabled for BGP neighbor discovery.

A router receiving a Hello message with Adjacency Hold Time set to 0 MUST treat this event as if the adjacency hold timer has expired for the specific neighbor and proceed to bring down the adjacency.

An interface going down (e.g. due to link failure or loss of signal) MUST immediately trigger the adjacency down event for all adjacencies over it as if the adjacency hold timer expired for all neighbors on that link.

## 9.2. Adjacency State Machine

On a per interface basis, BGP needs to maintain an adjacency state for each neighbor that it discovers. The adjacency state is maintained as a FSM and it has states as described in the following sections.

### 9.2.1. Down State

This is the transient terminal state after which an adjacency is deleted.

When transitioning to the Down state from Accepted, the router removes the path corresponding to this adjacency from any Adjacency Route that it had setup to the neighbor's prefixes. If no other adjacency exists in Accepted state to the neighbor, then it also deletes the BGP TCP peering session(s) setup to the neighbor based on the neighbor discovery mechanism.

### 9.2.2. Initial State

This is the transient initial state from which an adjacency starts, when the router detects a hello message from a new neighbor on the link, and immediately transitions to the 1-way state.

### 9.2.3. 1-Way State

While in the 1-way state (or when entering it), the adjacency transitions from 1-way to 2-way state when the router detects a Neighbor TLV corresponding to itself in the neighbor's Hello message. If the state does not immediately transition on to 2-way after entering 1-way, the the router MUST immediately trigger a State Change Hello message with the inclusion of the neighbor in a Neighbor TLV with the state set to 1-way.

When transitioning to the 1-way state from Accepted, the router removes the path corresponding to this adjacency from any Adjacency Route that it had setup to the neighbor's prefixes. If no other adjacency exists in Accepted state to the neighbor, then it also deletes the BGP TCP peering session(s) setup to the neighbor based on the neighbor discovery mechanism.

Adjacency transitions to Down state for any of the following events:

- o Link goes down operationally or is administratively shut
- o Adjacency Hold Timer expires
- o Router receives a Hello message from its neighbor with Adjacency Hold Time value set to 0
- o Neighbor discovery is disabled on the link
- o Change in BGP Identifier or AS number on the local router

#### 9.2.4. 2-Way State

Upon transitioning into this state, the router triggers a State Change Hello message with the neighbor's status set to 2-way in the Neighbor TLV. At this stage, both neighbors have received each other's Hello messages and thus discovered each other.

When the router, in this adjacency state, detects that the neighbor's state for itself is 2-way or higher, then it performs the validation checks based on local policy and information exchanged in the Hello TLVs. Following are some of the validation checks that may be performed on the adjacency:

- o Verify subnet matching between the local and remote interface addresses.
- o Verify AS numbers based on local policy as well as against the Allowed ASN TLV when one is being exchanged.
- o Verify that BFD monitoring (when enabled) is indicating UP state.

When the adjacency passes the validation checks, it transitions to the Adj-OK state and transitions to the Adj-Reject state otherwise.

The adjacency transitions to Down state for any of the adjacency down events described in Section 9.2.3 .

The adjacency transitions to 1-way state when the router stops seeing itself in a Neighbor TLV of its Neighbor's State Change Hello messages.

#### 9.2.5. Adj-Reject State

Upon transitioning into this state, the router triggers a State Change Hello message with the neighbor's status set to Adj-Reject in the Neighbor TLV.

The adjacency remains in the Adj-Reject state as long as the parameters being exchanged via the State Change Hello messages do not pass validation checks. The neighbors continue to include each other in their respective State Change Hello messages.

The adjacency transitions to the Adj-OK state once the validation checks pass (e.g. due to update in any parameters or local policy).

The adjacency transitions to Down state for any of the adjacency down events described in Section 9.2.3 .

The adjacency transitions to 1-way state when the router stops seeing itself in a Neighbor TLV of its Neighbor's State Change Hello messages.

When transitioning to an Adj-Reject state from Accepted state, the router removes the path corresponding to this adjacency from any Adjacency Route that it had setup to the neighbor's prefixes. If no other adjacency exists in Accepted state to the neighbor, then it also deletes the BGP TCP peering session(s) setup to the neighbor based on the neighbor discovery mechanism.

#### 9.2.6. Adj-OK State

Upon transitioning into this state, the router triggers a State Change Hello message with the neighbor's status set to Adj-OK in the Neighbor TLV.

The adjacency transition to Adj-OK state indicates that the router has accepted its neighbor. However, it is possible that the neighbor has not accept it and is signaling Adj-Reject state for the adjacency from it's end.

The adjacency transitions to the Accepted state from Adj-OK once it detects that its neighbor is also signaling the Adj-OK or Accepted state for it.

The adjacency transitions to Down state for any of the adjacency down events described in Section 9.2.3 .

The adjacency transitions to 1-way state when the router stops seeing itself in a Neighbor TLV of its Neighbor's State Change Hello messages.

The adjacency transitions to Adj-Reject state when any of the validation checks listed in Section 9.2.4 fail.

When transitioning to an Adj-OK state from Accepted state, the router removes the path corresponding to this adjacency from any Adjacency Route that it had setup to the neighbor's prefixes. If no other adjacency exists in Accepted state to the neighbor, then it also deletes the BGP TCP peering session(s) setup to the neighbor based on the neighbor discovery mechanism.

#### 9.2.7. Accepted State

The adjacency transition to Accepted state indicates that both the neighboring routers have accepted the adjacency to each other.

On this transition, the router triggers a State Change Hello message with the neighbor's status set to Accepted in the Neighbor TLV. It then installs the Adjacency Route(s) for the Prefix(es) signaled by the neighbor via the Local Prefix TLV via this adjacency link using the neighbor's address on that link. If this is the first Accepted adjacency to the neighbor then the Adjacency Route gets added to the local routing table, otherwise an additional path corresponding to this adjacency link and neighbor address on it gets added to the existing Adjacency Route. The details are described in Section 9.3.

When this is the first Accepted adjacency to the neighbor, then the setup of the BGP TCP session to the Peering Address(es) signaled by the neighbor is also triggered.

The adjacency transitions to Down state for any of the adjacency down events described in Section 9.2.3.

The adjacency transitions to 1-way state when the router stops seeing itself in a Neighbor TLV of its Neighbor's State Change Hello messages.

The adjacency transitions to Adj-Reject state when any of the validation checks listed in Section 9.2.4 fail.

### 9.3. Adjacency Route

The Adjacency Route programming is an optional part of the BGP Neighbor Discovery mechanism for setting up reachability for the neighbor's prefixes signaled via the Local Prefix TLV corresponding to adjacencies in Accepted state.

Adjacency Routes establish reachability between local prefixes on directly connected BGP routers. They enable reachability between the Peering Addresses (generally loopbacks) of the two neighbors so that the BGP TCP session may come up between them. Then, for the BGP routes learnt over the TCP session, where the next-hop is the neighbor, they also provide the BGP NH resolution.

Unlike other BGP routes, these are not recursive routes as in they point to the neighbor's interface and IP address. These routes that are setup as part of the neighbor discovery procedure are hence different from the regular IBGP and EBGP routes. These routes also MUST have a better administrative distance as compared to the IBGP and EBGP routes to ensure that they do not get displaced from the forwarding by BGP routes learnt over the very session(s) established using these peering routes.

The Adjacency Routes SHOULD NOT be stored in any of BGP RIBs [RFC4271] since they are not computed based on the BGP decision process. It is RECOMMENDED that these routes be managed in a separate routing table within the BGP Neighbor Discovery function to ensure that none of the processing and validation for BGP RIB affects them and in turn they do not influence the BGP decision process and route calculation.

When there are multiple interconnecting links between two BGP neighbors, a single BGP TCP session may be setup between them over which routes are then exchanged. However, in the forwarding, the Adjacency route will have multiple paths - one for each of these interconnecting links. So the BGP routes learnt over the session actually end up getting resolved over this Adjacency route and in turn gets the ECMP load balancing even with a single BGP session.

#### 10. Interactions with Base BGP Protocol

The BGP Finite State Machine (FSM) as specified in [RFC4271] is unchanged and the BGP TCP session establishment, route updates and processing continues to follow the BGP protocol specifications.

BGP peering addresses along with their respective ASNs have traditionally been explicitly provisioned on both BGP neighbors. The difference that neighbor discovery mechanism brings about is in elimination of this configuration as these parameters are learnt via the neighbor discovery procedure. Once BGP router learns its neighbor's peering address and ASN, then it initializes the BGP Peer FSM for this neighbor in the Idle State - just as if this neighbor was configured. From thereon, the BGP Peer FSM actions follows.

The BGP Keepalives and Hold Timer for the session over TCP apply unchanged and they govern the operations of the BGP TCP session. While the BGP Keepalive works at the TCP session level, the BGP Adjacency Hold Timer monitors one or more underlying interconnecting link adjacencies between the neighbors. The reachability for the BGP TCP session may also be over the some BGP routes learnt via routing updates over the sessions setup via neighbor discovery. It is likely that even after all the underlying interconnecting link adjacencies between two neighbors are down that the neighbor's peering address is reachable via BGP routing over some other path in the network. In order to avoid this, it is RECOMMENDED that the BGP TCP sessions setup via neighbor discovery mechanism use TTL set to 1 to ensure they are setup only over directly attached links to the neighbors.

Since the BGP TCP session setup via neighbor discovery was meant for hop-by-hop routing, it would be necessary to bring down the session even while its BGP Hold Timer has not expired for faster convergence.

Therefore, when all the underlying link adjacencies between two BGP neighbors move out of the Accepted state (or go down), then the BGP TCP peering session that was setup using BGP Neighbor Discovery mechanism between these two neighbors is also deleted as if it was un-configured.

Since the BGP neighbor discovery mechanism runs over a UDP socket, it is isolated from the core BGP protocol working which is TCP based. Implementations SHOULD ensure that the hello processing does not affect the base BGP operations and scalability. One option may be to run the BGP neighbor discovery mechanism in a separate thread from the rest of BGP processing. These implementation details, however, are outside the scope of this document.

It is not generally expected that BGP sessions are explicitly provisioned along with the neighbor discovery mechanism. However, in such an event, the neighbor discovery mechanism MUST NOT affect or result in any changes to provisioned BGP neighbors and their operations. Specifically, BGP peering to auto-discovered neighbors MUST NOT be instantiated using the procedures described in this document when the same BGP neighbor is already provisioned. The configured BGP neighbor parameters take precedence and the auto-discovered values and parameters are not used for such configured BGP sessions.

## 11. Security Considerations

BGP routers accept TCP connection attempts to port 179 only from the provisioned BGP neighbors or, in some implementations, those from within a configured address range. With the BGP neighbor auto-discovery mechanism, it is now possible for BGP to automatically learn neighbors and initiate/receive TCP connections from them. This introduces the need for specific considerations to be taken care of to ensure security of the BGP protocol operations.

This document introduces UDP messages in BGP for the neighbor discovery mechanism using the BGP Hello messages. For security purposes, implementations MUST exchange the Hello messages only on interfaces specifically enabled for neighbor discovery. Hello messages MUST NOT be accepted on other than the 224.0.0.2 or FF02::2 addresses. Optionally, implementations MAY set TTL to 255 when originating the Hello messages and receivers check specifically for the TLV to be 254 and discard the packet when this is not the case. This ensures that the Hello packets signaling happens between directly connected BGP routers only.

The BGP neighbor discovery mechanism is expected to be run typically in DCs and between physically connected routers that are trustworthy.

The Cryptographic Authentication TLV (as described in Section 8.6) SHOULD be used in deployments where this assumption of trustworthiness is not valid. This mechanism is similar to one defined for LDP Hello messages that are also UDP based as specified in [RFC7349]. An updated future version of this document will describe similar procedures for BGP hello in more details.

Once the BGP hello messages and the neighbor discovery mechanism is secured, then the security considerations for BGP protocol operations apply for the auto-discovered neighbor sessions.

## 12. Manageability Considerations

This section is structured as recommended in [RFC5706].

### 12.1. Operational Considerations

The BGP neighbor discovery mechanism introduced by this document is not applicable to general BGP deployments as discussed in Section 3. The mechanism is specifically meant for networks where BGP is used as a hop-by-hop routing protocol E.g. as described in [RFC7938]. The neighbor discovery mechanism hence SHOULD NOT be enabled by default in BGP.

Implementations SHOULD provide configuration methods that allow enablement of BGP neighbor discovery on specific local interfaces. In a DC network, it is expected that the operator selects the appropriate links on which to enable this e.g. on a Tier 2 node it is enabled on all links towards the Tier 1 and Tier 3 nodes while on a Tier 1 node, it may be only enabled on the links towards the Tier 2 node. The details of this enablement are outside the scope of this document since it varies based on the DC design and may be implementation specific.

Implementations SHOULD provide configuration methods that enable the setup of BGP neighbor templates that enables operator to setup BGP neighbor discovery parameters on the BGP router. Some of the aspects to be considered in such a template are:

- o Local address to be used for the BGP TCP session peering along with the local ASN and the AFI/SAFI enabled for the auto-discovered sessions
- o BGP policies to be enabled for the auto-discovered sessions
- o Optionally specify the list of ASNs with which auto-discovered sessions should be brought up. This is to ensure that when links between different Tier nodes are not used by BGP when they get

connected wrongly due to accidents (e.g. say a Tier 3 node is connected to a Tier 1 node).

- o Authentication methods that are need to be enabled in an environment which is not secure
- o Local interfaces over which the specific template needs to be applied for BGP neighbor discovery
- o Other parameters like the Adjacency Hold Timer value to be used or other optional features

This mechanism does not impose any restrictions on the way ASNs or addresses are assigned to the nodes. Various automatic provisioning, auto-configuration or zero-touch-provisioning mechanisms may be used.

Implementations SHOULD report the state of the BGP operations over each link enabled for neighbor discovery including the status of all adjacencies learnt over it. Implementations SHOULD also report the operations of the auto-discovered BGP TCP peering sessions similar to the provisioned BGP neighbors.

Implementations SHOULD support logging of events like discovery of an adjacency using neighbor discovery including peering route updates and events like triggering of BGP TCP session establishment for them. Errors and alarms related to loss of adjacencies and tear down of BGP TCP peering sessions SHOULD also be generated so they could be monitored.

## 12.2. Management Considerations

This document introduces UDP based messaging in BGP protocol and therefore the necessary fault management mechanisms are required to be implemented for the same. Implementations MUST discard unsupported message types or version types other than 4 received over a UDP session. Such messages MUST NOT affect the neighbor discovery mechanism in operation using the Hello messages. Unknown TLVs received via the Hello messages MUST be ignored and the rest of the Hello message MUST be processed. Implementations SHOULD discard Hello messages with malformed TLVs and this should be logged as an error.

## 13. IANA Considerations

This documents requests IANA for updates to the BGP Parameters registry as described in this section.

## 13.1. BGP Hello Message

This document requests IANA to allocate a new UDP port (179 is the preferred number ) and a BGP message type code for BGP Hello message.

Value	TLV Name	Reference
-----	-----	-----
	Service Name: BGP-HELLO	
	Transport Protocol(s): UDP	
	Assignee: IESG <iesg@ietf.org>	
	Contact: IETF Chair <chair@ietf.org>.	
	Description: BGP Hello Message.	
	Reference: This document -- draft-xu-idr-neighbor-autodiscovery.	
	Port Number: 179 (preferred value) -- To be assigned by IANA.	

## 13.2. TLVs of BGP Hello Message

This document requests IANA to create a new registry "TLVs of BGP Hello Message" with the following registration procedure:

Registry Name: TLVs of BGP Hello Message.

Value	TLV Name	Reference
-----	-----	-----
0	Reserved	This document
1	Accepted ASN List	This document
2	Peering Address	This document
3	Local Prefix	This document
4	Link Attributes	This document
5	Neighbor	This document
6	Cryptographic Authentication	This document
7-65500	Unassigned	
65501-65534	Experimental	This document
65535	Reserved	This document

## 14. Acknowledgements

The authors would like to thank Enke Chen, Krishna Swamy and Ramesh Yakkala for their valuable comments and suggestions on this document.

## 15. Contributors

Satya Mohanty  
Cisco  
Email: satyamoh@cisco.com

Shunwan Zhuang  
Huawei  
Email: zhuangshunwan@huawei.com

Chao Huang  
Alibaba Inc  
Email: jingtang.hc@alibaba-inc.com

Guixin Bao  
Alibaba Inc  
Email: guixin.bgx@alibaba-inc.com

Jinghui Liu  
Ruijie Networks  
Email: liujh@ruijie.com.cn

Zhichun Jiang  
Tencent  
Email: zcjiang@tencent.com

Shaowen Ma  
Mellanox  
mashaowen@gmail.com

## 16. References

### 16.1. 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>>.
- [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>>.
- [RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "LDP Specification", RFC 5036, DOI 10.17487/RFC5036, October 2007, <<https://www.rfc-editor.org/info/rfc5036>>.

- [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>>.

## 16.2. Informative References

- [FIPS-180-4] Technology, N. I. O. S. A., "Secure Hash Standard (SHS), FIPS PUB 180-4", March 2012.
- [I-D.ietf-lsvr-bgp-spf] Patel, K., Lindem, A., Zandi, S., and W. Henderickx, "Shortest Path Routing Extensions for BGP Protocol", draft-ietf-lsvr-bgp-spf-06 (work in progress), September 2019.
- [I-D.ketant-idr-bgp-ls-bgp-only-fabric] Talaulikar, K., Filsfils, C., ananthamurthy, k., Zandi, S., Dawra, G., and M. Durrani, "BGP Link-State Extensions for BGP-only Fabric", draft-ketant-idr-bgp-ls-bgp-only-fabric-03 (work in progress), September 2019.
- [RFC2104] Krawczyk, H., Bellare, M., and R. Canetti, "HMAC: Keyed-Hashing for Message Authentication", RFC 2104, DOI 10.17487/RFC2104, February 1997, <<https://www.rfc-editor.org/info/rfc2104>>.
- [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>>.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <<https://www.rfc-editor.org/info/rfc5880>>.
- [RFC7349] Zheng, L., Chen, M., and M. Bhatia, "LDP Hello Cryptographic Authentication", RFC 7349, DOI 10.17487/RFC7349, August 2014, <<https://www.rfc-editor.org/info/rfc7349>>.
- [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>>.

[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>>.

Authors' Addresses

Xiaohu Xu  
Alibaba Inc  
China

Email: [xiaohu.xxh@alibaba-inc.com](mailto:xiaohu.xxh@alibaba-inc.com)

Ketan Talaulikar  
Cisco Systems  
India

Email: [ketant@cisco.com](mailto:ketant@cisco.com)

Kunyang Bi  
Huawei  
China

Email: [bikunyang@huawei.com](mailto:bikunyang@huawei.com)

Jeff Tantsura  
Apstra  
USA

Email: [jefftant.ietf@gmail.com](mailto:jefftant.ietf@gmail.com)

Nikos Triantafyllis  
Amazon Web Services  
USA

Email: [ntriantafyllis@gmail.com](mailto:ntriantafyllis@gmail.com)

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 3, 2019

R. Bush  
Arrcus & IIJ  
K. Patel  
Arrcus  
July 2, 2018

Link State Over Ethernet  
draft-ymbk-lsvr-lsoe-01

Abstract

Used in a Massive Data Center (MDC), BGP-LS and BGP-SPF need link neighbor discovery, liveness, and addressability data. Link State Over Ethernet protocols provide link discovery, exchange AFI/SAFIs, and discover addresses over raw Ethernet. These data are pushed directly to BGP-LS/SPF, obviating the need for centralized controller architectures. This protocol is more widely applicable, and has been designed to support a wide range of routing and similar protocols which need link discovery and characterisation.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described in [RFC2119] only when they appear in all upper case. They may also appear in lower or mixed case as English words, without normative meaning.

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 January 3, 2019.

## Copyright Notice

Copyright (c) 2018 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
2. Terminology . . . . .	3
3. Background . . . . .	4
4. Top Level Overview . . . . .	4
5. Ethernet to Ethernet Protocols . . . . .	5
5.1. Inter-Link Ether Protocol Overview . . . . .	6
5.2. PDUs and Frames . . . . .	7
5.2.1. Frame TLV . . . . .	7
5.2.2. Link Hello / KeepAlive . . . . .	10
5.2.3. Capability Exchange . . . . .	10
5.2.4. Timer Negotiation . . . . .	11
5.3. The AFI/SAFI Exchanges . . . . .	11
5.3.1. AFI/SAFI Capability Exchange . . . . .	12
5.3.2. The AFI/SAFI PDU Skeleton . . . . .	12
5.3.3. AFI/SAFI ACK . . . . .	13
5.3.4. Add/Drop/Prim . . . . .	13
5.3.5. IPv4 Announce / Withdraw . . . . .	13
5.3.6. IPv6 Announce / Withdraw . . . . .	14
5.3.7. MPLS Label List . . . . .	14
5.3.8. MPLS IPv4 Announce / Withdraw . . . . .	15
5.3.9. MPLS IPv6 Announce / Withdraw . . . . .	15
6. Layer 2.5 and 3 Liveness . . . . .	16
7. The North/South Protocol . . . . .	16
7.1. Use BGP-LS as Much as Possible . . . . .	17
7.2. Extensions to BGP-LS . . . . .	17
8. Security Considerations . . . . .	17
9. IANA Considerations . . . . .	18
10. IEEE Considerations . . . . .	18
11. Acknowledgments . . . . .	18
12. References . . . . .	19
12.1. Normative References . . . . .	19

12.2. Informative References . . . . .	20
Authors' Addresses . . . . .	20

## 1. Introduction

The Massive Data Center (MDC) environment presents unusual problems of scale, e.g.  $O(10,000)$  switches, while its homogeneity presents opportunities for simple approaches. Approaches such as Jupiter Rising [JUPITER] use a central controller to deal with scaling, while BGP-SPF [I-D.ietf-lsvr-bgp-spf] provides massive scale out without centralization using a tried and tested scalable distributed control plane, offering a scalable routing solution in Clos and similar environments. But it needs link state and addressing data from the network to build the routing topology. LLDP has scaling issues, e.g. in extending a PDU beyond 1,500 bytes.

Link State Over Ethernet (LSOE) provides brutally simple mechanisms for devices to

- o Discover each other's MACs,
- o Run MAC keep-alives for liveness assurance,
- o Discover each other's unique IDs (ASN, RouterID, ...),
- o Negotiate mutually supported AFI/SAFIs,
- o Discover and maintain link IP/MPLS addresses,
- o Enable layer three link liveness such as BFD, and finally
- o Push these data up to BGP-SPF which computes the topology and builds routing and forwarding tables.

This protocol is more widely applicable than BGP-SPF, and has been designed to support a wide range of routing and similar protocols which need link discovery and characterisation.

## 2. Terminology

Even though it concentrates on the Ethernet layer, this document relies heavily on routing terminology. The following are some possibly confusing terms:

AFI/SAFI: Address Family Indicator and Subsequent Address Family Indicator. I.e. classes of addresses such as IPv4, IPv6, ...

ASN: Autonomous System Number [RFC4271], a BGP identifier for an originator of routing, particularly BGP, announcements, see [RFC4271].

RouterID: [RFC4271].

BGP-SPF A hybrid protocol using BGP transport but Dijkstra SPF decision process. See [I-D.ietf-lsvr-bgp-spf].

Clos: A hierarchic switch topology commonly used in data centers.

Frame The payload of an Ethernet packet.

MAC: Medium Access Control, essentially an Ethernet address, six octets.

MDC: Massive Data Center, O(1,000) TORs or more.

PDU: Protocol Data Unit, essentially an application layer message.

SPF: Shortest Path First, an algorithm for finding the shortest paths between nodes in a graph.

TOR: Top Of Rack switch, aggregates the servers in a rack and connects to the Clos spine.

ZTP: Zero Touch Provisioning gives devices initial addresses, credentials, etc. on boot/restart.

### 3. Background

LSOE assumes a Clos-like topology, though the acyclic constraint is not necessary.

While LSOE is designed for the MDC, there are no inherent reasons it could not run on a WAN; though it is not clear that this would be useful. The authentication and authorisation needed to run safely on the WAN are not (yet) included in this protocol.

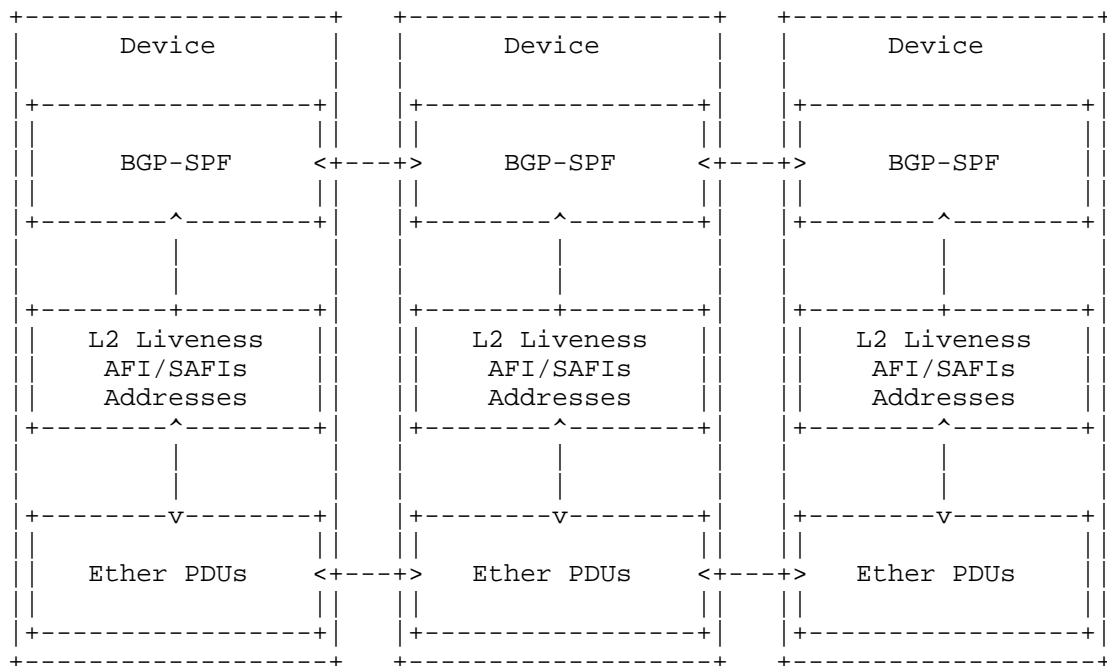
LLDP is not suitable because one can not extend a PDU beyond 1500 bytes without hitting an IPR barrier. It is also complex.

UDP is unsuitable as it would require prior knowledge of IP level addressing, one of the key purposes of this discovery protocol.

LSOE assumes a new IEEE assigned EtherType (TBD).

### 4. Top Level Overview

- o MAC Link State is exchanged over Ethernet
- o AFI/SAFI data are exchanged and IP-Level Liveness Checks done
- o BGP-SPF uses the data to discover and build the topology database



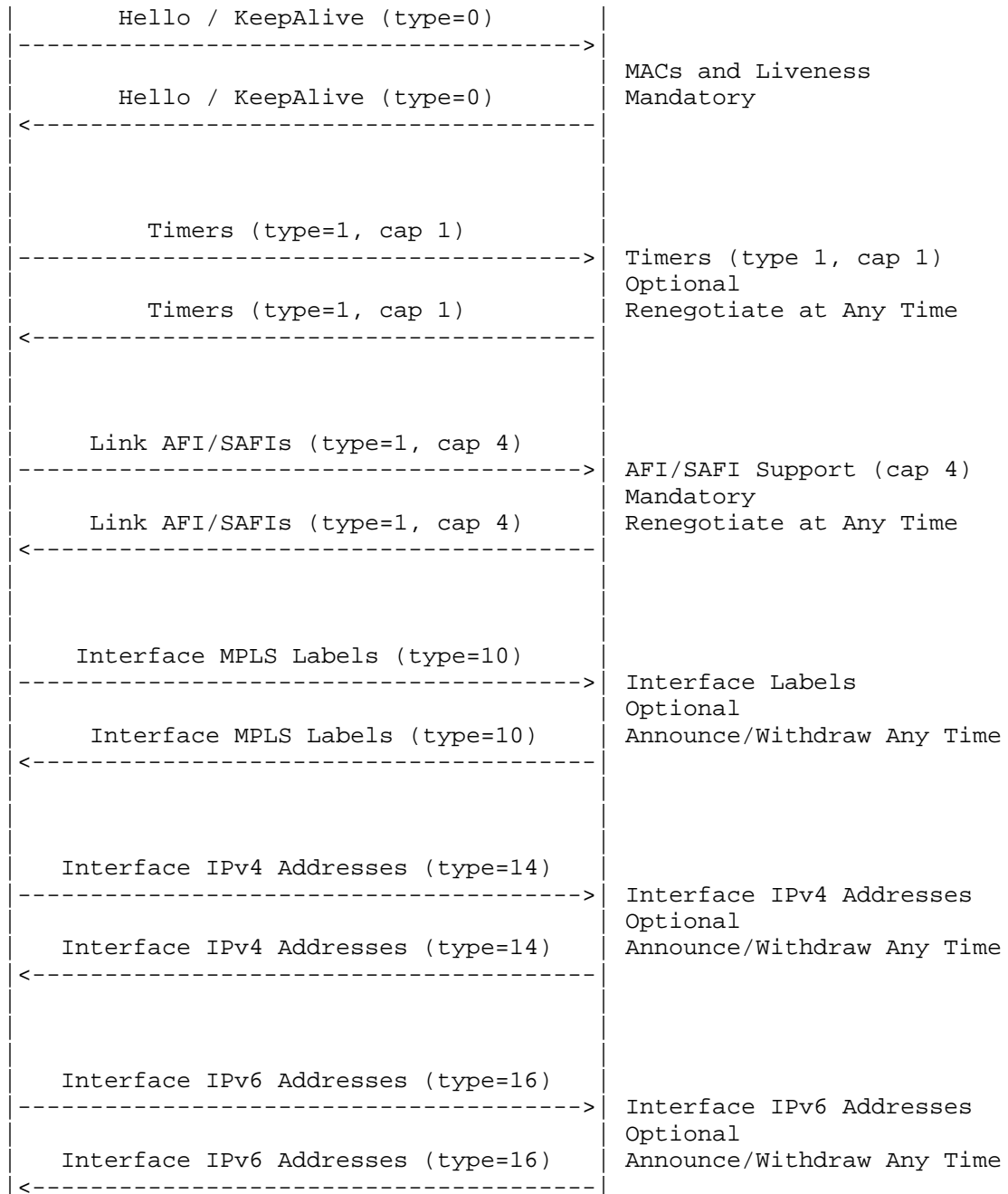
There are two sets of protocols:

- o Ethernet to Ethernet protocols are used to exchange layer 2 data, i.e. MACs, and layer 2.5 and 3 data, i.e. ASNs, AFI/SAFIs, and interface addresses.
- o A Link Layer to BGP protocol pushes these data up the stack to BGP-SPF, converting to the BGP-LS BGP-like data format.
- o And, of course, the BGP layer crosses all the devices, though it is not part of these LSOE protocols.

## 5. Ethernet to Ethernet Protocols

The basic Ethernet Framed protocols

## 5.1. Inter-Link Ether Protocol Overview



## 5.2. PDUs and Frames

This is all about inter-device Link State.

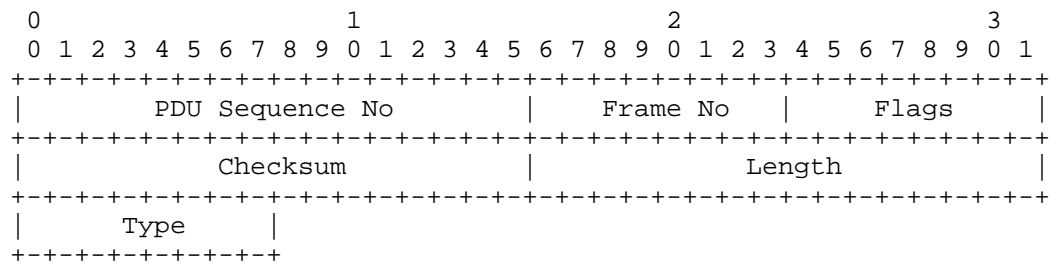
A PDU is one or more Ethernet Frames.

A Frame has a PDU Sequence Number and a Frame Number to allow assembly of out order frames.

Because BGP-SPF and Data Plane payloads are assumed to be IP over the same Ethernet, one needs to keep an eye on congestion.

### 5.2.1. Frame TLV

The basic Ethernet PDU is a typical TLV (Type Length Value) PDU, except it's really LTV for the sake of alignment :)



The fields of the basic Ethernet PDU are as follows:

PDU Sequence No: Semi-unique identifier of a TLV PDU (e.g. the low order 16 bits of UNIX time)

Frame No: 0..255 Frame Sequence Number Within a multi-frame PDU

Flags: A bit field

- 0 - Sender has been restarted
- 1 - One of a multi-Frame sequence
- 2 - last of a multi-Frame sequence
- 3-7 - Reserved

Checksum: One's complement over Frame, detect bit flips

Length: Total Bytes in PDU including all frames and fields

Type: An integer

- 0 - Hello / KeepAlive

- 1 - Capability
- 2-9 - Reserved
- 10 - AFI/SAFI ACK
- 11 - IPv4 Announce / Withdraw
- 12 - IPv6 Announce / Withdraw
- 13 - MPLS IPv4 Announce / Withdraw
- 14 - MPLS IPv6 Announce / Withdraw
- 15-255 Reserved

#### 5.2.1.1. The Checksum

There is a reason conservative folk use a checksum in UDP. And when the operators stretch to jumbo frames ...

One's complement is a bit silly, though trivial to implement and might be sufficient.

Sum up either 16-bit shorts in a 32-bit int, or 32-bit ints in a 64-bit long, then take the high-order section, shift it right, rotate, add it in, repeat until zero. -- smb off the top of his head

```
/* The F table from Skipjack, and it would work for the S-Box.
```

```
There are other S-Box sources as well. -- Russ Housley */
```

```
const BYTE sbox[256] = {
0xa3,0xd7,0x09,0x83,0xf8,0x48,0xf6,0xf4,0xb3,0x21,0x15,0x78,
0x99,0xb1,0xaf,0xf9,0xe7,0x2d,0x4d,0x8a,0xce,0x4c,0xca,0x2e,
0x52,0x95,0xd9,0x1e,0x4e,0x38,0x44,0x28,0x0a,0xdf,0x02,0xa0,
0x17,0xf1,0x60,0x68,0x12,0xb7,0x7a,0xc3,0xe9,0xfa,0x3d,0x53,
0x96,0x84,0x6b,0xba,0xf2,0x63,0x9a,0x19,0x7c,0xae,0xe5,0xf5,
0xf7,0x16,0x6a,0xa2,0x39,0xb6,0x7b,0x0f,0xc1,0x93,0x81,0x1b,
0xee,0xb4,0x1a,0xea,0xd0,0x91,0x2f,0xb8,0x55,0xb9,0xda,0x85,
0x3f,0x41,0xbf,0xe0,0x5a,0x58,0x80,0x5f,0x66,0x0b,0xd8,0x90,
0x35,0xd5,0xc0,0xa7,0x33,0x06,0x65,0x69,0x45,0x00,0x94,0x56,
0x6d,0x98,0x9b,0x76,0x97,0xfc,0xb2,0xc2,0xb0,0xfe,0xdb,0x20,
0xe1,0xeb,0xd6,0xe4,0xdd,0x47,0x4a,0x1d,0x42,0xed,0x9e,0x6e,
0x49,0x3c,0xcd,0x43,0x27,0xd2,0x07,0xd4,0xde,0xc7,0x67,0x18,
0x89,0xcb,0x30,0x1f,0x8d,0xc6,0x8f,0xaa,0xc8,0x74,0xdc,0xc9,
0x5d,0x5c,0x31,0xa4,0x70,0x88,0x61,0x2c,0x9f,0x0d,0x2b,0x87,
0x50,0x82,0x54,0x64,0x26,0x7d,0x03,0x40,0x34,0x4b,0x1c,0x73,
0xd1,0xc4,0xfd,0x3b,0xcc,0xfb,0x7f,0xab,0xe6,0x3e,0x5b,0xa5,
0xad,0x04,0x23,0x9c,0x14,0x51,0x22,0xf0,0x29,0x79,0x71,0x7e,
0xff,0x8c,0x0e,0xe2,0x0c,0xef,0xbc,0x72,0x75,0x6f,0x37,0xa1,
0xec,0xd3,0x8e,0x62,0x8b,0x86,0x10,0xe8,0x08,0x77,0x11,0xbe,
0x92,0x4f,0x24,0xc5,0x32,0x36,0x9d,0xcf,0xf3,0xa6,0xbb,0xac,
0x5e,0x6c,0xa9,0x13,0x57,0x25,0xb5,0xe3,0xbd,0xa8,0x3a,0x01,
0x05,0x59,0x2a,0x46
};
```

```
/* example C code, constant time even, thanks Rob Austein */
```

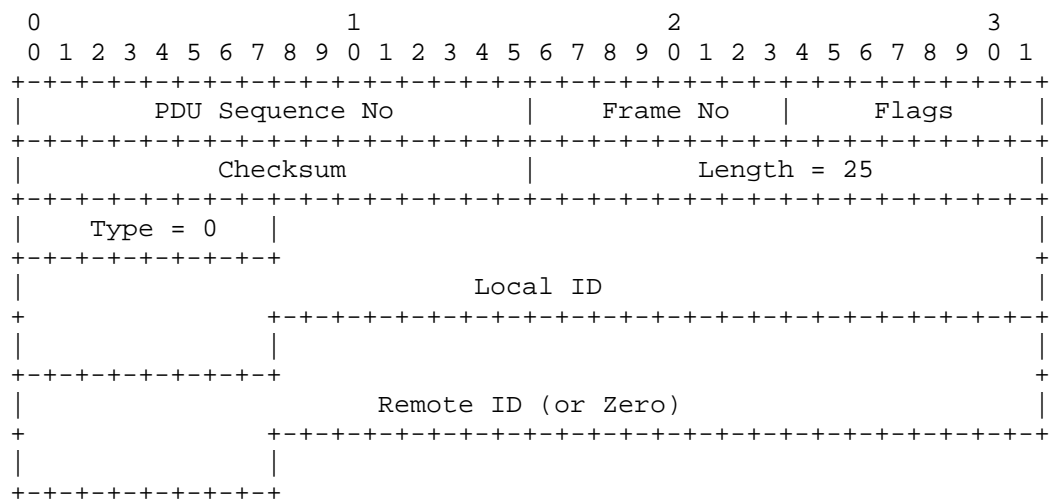
```
uint16_t sbox_checksum(const *b, const size_t n)
{
    uint32_t sum[2] = {0, 0};
    for (int i = 0; i < n; i++)
        sum[i & 1] += sbox[b[i]];
    uint32_t result = (sum[0] << 8) + sum[1];
    result = (result >> 16) + (result & 0xFFFF);
    result = (result >> 16) + (result & 0xFFFF);
    return (uint16_t) result;
}
```

### 5.2.2. Link Hello / KeepAlive

The Hello and KeepAlive PDUs are one and the same.

Each device learns the other's MAC from its HELLO whining. I.e., all devices on a wire/interface know each others MACs and learn each other's IDs.

An ID can be an ASN with high order bits zero, a classic RouterID with high order bits zero, a catenation of the two, a 48-bit ISO System-ID, or any other identifier unique to a single device in the BGP-SPF routing space.



Once two devices know each other's MACs, Ethernet keep-alives may be started to ensure layer two liveness. The timing and acceptable drop of the keep-alives may be set with the Timer Negotiation capability exchange.

When the local sends a first Hello without knowing the remote device's ID, the Remote ID SHOULD be zero. The Local ID MUST never be zero.

### 5.2.3. Capability Exchange

Peers on the Ethernet exchange capabilities, such as timers, AFI/SAFIs supported, etc. There is a simple capability exchange.

By convention, the device with the lowest MAC sends first.



### 5.3.1. AFI/SAFI Capability Exchange

First they negotiate what AFI/SAFIs are supported on the link.

As before, the lowest MAC initiates the negotiation.

```

      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
+-----+-----+-----+-----+-----+-----+-----+-----+
|          PDU Sequence No          |      Frame No      |      Flags      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|          Checksum          |      Length = 13      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Type = 1      |      RADflag      |      Capability = 4      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|      AFI/SAFIs      |
+-----+-----+-----+-----+-----+-----+-----+

```

The AFI/SAFIs currently defined are as follows:

- 10 - IPv4
- 11 - IPv6
- 12 - MPLS IPv4
- 13 - MPLS IPv6
- ... - other tunnels (e.g. GRE)

### 5.3.2. The AFI/SAFI PDU Skeleton

Now both sides can exchange their actual interfaces addresses for all the negotiated AFI/SAFIs.

```

      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
+-----+-----+-----+-----+-----+-----+-----+-----+
|          PDU Sequence No          |      Frame No      |      Flags      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|          Checksum          |      Length          |
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Type = 42      |      Sequence Number      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|          |      AFI/SAFI Count          |      sub-PDUs...      |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

The AFI/SAFI Exchange is over an unreliable transport so there are Sequence Numbers and ACKs.

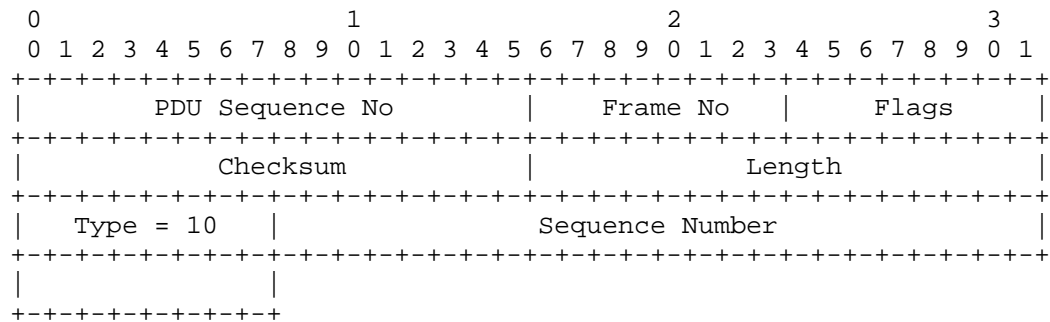
The Sequence Number is a point-to-point link announcement counter, incremented for each exchange in each direction on the link.

The Receiver will ACK it with a Type=10, see following PDU.

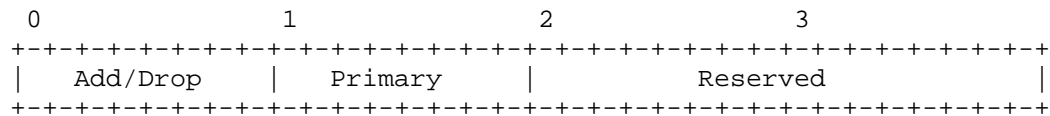
If the Sender does not receive an ACK in one second, they retransmit. Other delay timers may be negotiated using the Timing Capability.

If a sender has multiple links on the same interface, separate counters must be kept for each.

#### 5.3.3. AFI/SAFI ACK



#### 5.3.4. Add/Drop/Prim



Each AFI/SAFI interface address may be announced (Add/Drop == 1), or withdrawn (Add/Drop == 0).

An interface may have multiple AFI/SAFIs.

For each AFI/SAFI on an interface there might be multiple addresses.

One address per AFI/SAFI SHOULD be marked as primary (Primary == 1).

#### 5.3.5. IPv4 Announce / Withdraw



```

      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
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| Label Count |                               Label                               | Exp |S|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|                               Label                               | Exp |S| more ... |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

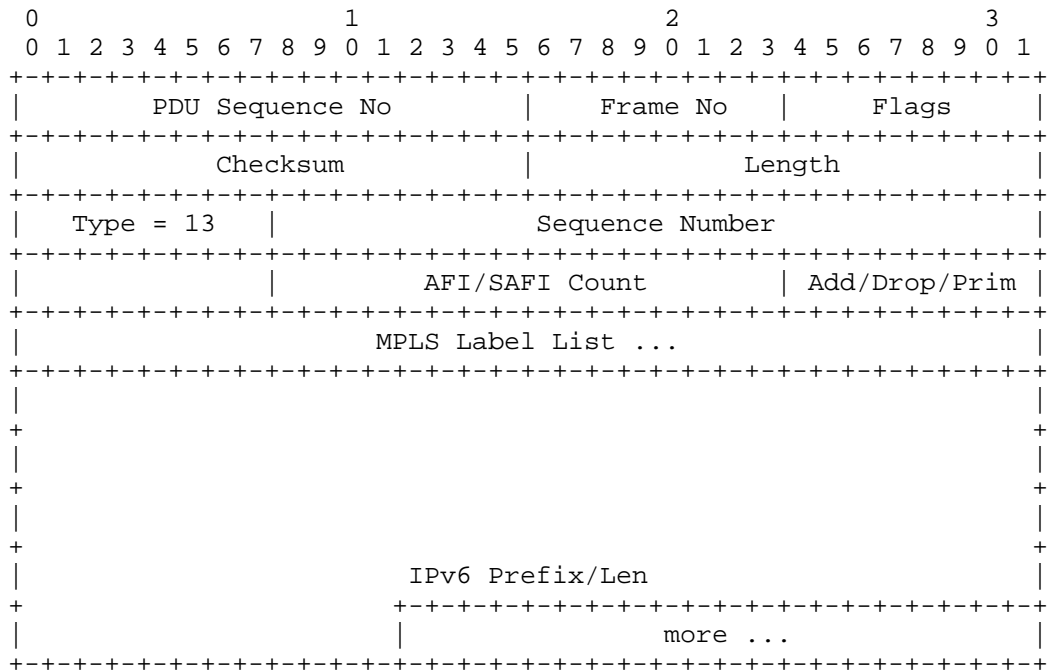
#### 5.3.8. MPLS IPv4 Announce / Withdraw

```

      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
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| PDU Sequence No | Frame No | Flags |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| Checksum | Length |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| Type = 13 | Sequence Number |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| AFI/SAFI Count | Add/Drop/Prim |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| MPLS Label List ... |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| IPv4 Prefix/Len |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| MPLS Label List ... |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| IPv4 Prefix/Len |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| more ... |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+

```

#### 5.3.9. MPLS IPv6 Announce / Withdraw



## 6. Layer 2.5 and 3 Liveness

Ether liveness is continuously tested by Hello Keep-Alives, see Section 5.2.2. Now IP/Label liveness may be tested. As layer 2.5 or layer 3 connectivity could still break, liveness above layer 2 should be frequently tested.

Assume one or more AFI/SAFI addresses will be used to ping, BFD, or whatever the operator configures.

## 7. The North/South Protocol

Thus far, we have a one-hop point-to-point link discovery protocol.

We know what unique node identifiers (ASNs, RouterIDs, ...) and AFI/SAFIs are on each Link Interface.

At the Ethernet layer we do not want to do topology discovery and Dijkstra a la IS-IS.

So the node identifiers, link AFI/SAFIs, and state changes are pushed North to BGP-SPF which discovers and maintains the topology, runs Dijkstra, and builds the routing database.

For example, if a neighbor's MAC changes, the device seeing the change pushes that change Northbound.

#### 7.1. Use BGP-LS as Much as Possible

BGP-LS [RFC7752] defines BGP-like PDUs describing link state (links, nodes, link prefixes, and many other things), and a new BGP path attribute providing Northbound transport, all of which can be ingested by BGP-SPF; see Section 4 of [I-D.ietf-lsvr-bgp-spf].

For IPv4 links, TLVs 259 and 260 are used. And for IPv6 links, TLVs 261 and 262. If there are multiple addresses on a link, multiple TLV pairs are pushed North, having the same ID pairs.

#### 7.2. Extensions to BGP-LS

The Northbound protocol needs a few minor extensions to BGP-LS. Luckily, others have needed the same extensions.

Similarly to BGP-SPF, the BGP protocol is used in the Protocol-ID field specified in table 1 of [I-D.ietf-idr-bgpls-segment-routing-epe]. The local and remote node descriptors for all NLRI are the ID's described in Section 5.2.2.

Label Sub-TLVs from [I-D.ietf-idr-bgp-ls-segment-routing-ext] Section 2.1.1, are used to associate one or more MPLS Labels with a link.

### 8. Security Considerations

The protocol as is MUST NOT be used outside a datacenter or similarly closed environment due to lack of authentication and authorisation. These will be worked on in a later effort, likely using credentials configured using ZTP.

Many MDC operators have a strange belief that physical walls and firewalls provide sufficient security. This is not credible. All MDC protocols need to be examined for exposure and attack surface.

On the wire Ethernet is assumed to be secure, though it could be tapped and data modified by an in-house on the wire attacker.

Malicious nodes/devices could mis-announce addressing, form malicious associations, etc.

## 9. IANA Considerations

This document requests the IANA create a registry for LSOE PDU Type, which may range from 0 to 255. The name of the registry should be LSOA-PDU-Type. The policy for adding to the registry is RFC Required per [RFC5226], either standards track or experimental. The initial entries should be the following:

PDU Code	PDU Name
-----	-----
0	Hello / KeepAlive
1	Capability
2-9	Reserved
10	AFI/SAFI ACK
11	IPv4 Announce / Withdraw
12	IPv6 Announce / Withdraw
13	MPLS IPv4 Announce / Withdraw
14	MPLS IPv6 Announce / Withdraw
15-255	Reserved

This document requests the IANA create a registry for LSOE AFI/SAFI Type, which may range from 0 to 255. The name of the registry should be LSOA-AFI/SAFI-Type. The policy for adding to the registry is RFC Required per [RFC5226], either standards track or experimental. The initial entries should be the following:

AFI/SAFI Type Code	AFI/SAFI Type Name
-----	-----
0-9	Reserved
10	IPv4
11	IPv6
12	MPLS IPv4
13	MPLS IPv6
14-255	Reserved

## 10. IEEE Considerations

This document needs a new EtherType.

## 11. Acknowledgments

The authors thank Cristel Pelsser for multiple reviews, Joe Clarke for a useful review, Martijn Schmidt for his contribution, Rob Austein for reviews and checksum code, Russ Housley for checksum discussion and sBox, and Steve Bellovin for checksum advice.

## 12. References

### 12.1. 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-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-lsvr-bgp-spf]  
Patel, K., Lindem, A., Zandi, S., and W. Henderickx,  
"Shortest Path Routing Extensions for BGP Protocol",  
draft-ietf-lsvr-bgp-spf-01 (work in progress), May 2018.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate  
Requirement Levels", BCP 14, RFC 2119,  
DOI 10.17487/RFC2119, March 1997,  
<<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y.,  
Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack  
Encoding", RFC 3032, DOI 10.17487/RFC3032, January 2001,  
<<http://www.rfc-editor.org/info/rfc3032>>.
- [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,  
<<http://www.rfc-editor.org/info/rfc4271>>.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an  
IANA Considerations Section in RFCs", RFC 5226,  
DOI 10.17487/RFC5226, May 2008,  
<<http://www.rfc-editor.org/info/rfc5226>>.
- [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,  
<<http://www.rfc-editor.org/info/rfc7752>>.

## 12.2. Informative References

[JUPITER] Singh, A., Germano, P., Kanagala, A., Liu, H., Provost, J., Simmons, J., Tanda, E., Wanderer, J., HAP.1zle, U., Stuart, S., Vahdat, A., Ong, J., Agarwal, A., Anderson, G., Armistead, A., Bannon, R., Boving, S., Desai, G., and B. Felderman, "Jupiter rising", Communications of the ACM Vol. 59, pp. 88-97, DOI 10.1145/2975159, August 2016.

## Authors' Addresses

Randy Bush  
Arrcus & IIJ  
5147 Crystal Springs  
Bainbridge Island, WA 98110  
United States of America

Email: randy@psg.com

Keyur Patel  
Arrcus  
2077 Gateway Place, Suite #250  
San Jose, CA 95119  
United States of America

Email: keyur@arrcus.com

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: May 10, 2019

R. Bush  
Arrcus & IIJ  
R. Austein  
K. Patel  
Arrcus  
November 6, 2018

Link State Over Ethernet  
draft-ymbk-lsvr-lsoe-03

Abstract

Used in Massive Data Centers (MDCs), BGP-SPF and similar protocols need link neighbor discovery, link encapsulation data, and Layer 2 liveness. The Link State Over Ethernet protocol provides link discovery, exchanges supported encapsulations (IPv4, IPv6, ...), discovers encapsulation addresses (Layer 3 / MPLS identifiers) over raw Ethernet, and provides layer 2 liveness checking. The interface data are pushed directly to a BGP-LS API, obviating the need for centralized controller architectures. This protocol is intended to be more widely applicable to other upper layer routing protocols which need link discovery and characterisation.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described in [RFC2119] only when they appear in all upper case. They may also appear in lower or mixed case as English words, without normative meaning. See [RFC8174].

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 May 10, 2019.

## Copyright Notice

Copyright (c) 2018 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
2. Terminology . . . . .	4
3. Background . . . . .	4
4. Top Level Overview . . . . .	5
5. Ethernet to Ethernet Protocols . . . . .	6
5.1. Inter-Link Ether Protocol Overview . . . . .	6
6. Transport Layer . . . . .	8
7. The Checksum . . . . .	8
8. TLV PDUs . . . . .	10
9. HELLO . . . . .	10
10. OPEN . . . . .	11
11. ACK . . . . .	13
11.1. Retransmission . . . . .	13
12. The Encapsulations . . . . .	13
12.1. The Encapsulation PDU Skeleton . . . . .	14
12.2. Prim/Loop Flags . . . . .	15
12.3. IPv4 Encapsulation . . . . .	15
12.4. IPv6 Encapsulation . . . . .	16
12.5. MPLS Label List . . . . .	16
12.6. MPLS IPv4 Encapsulation . . . . .	16
12.7. MPLS IPv6 Encapsulation . . . . .	17
13. KEEPALIVE - Layer 2 Liveness . . . . .	18
14. Layers 2.5 and 3 Liveness . . . . .	19
15. The North/South Protocol . . . . .	19
15.1. Use BGP-LS as Much as Possible . . . . .	19
15.2. Extensions to BGP-LS . . . . .	20
16. Discussion . . . . .	20
16.1. HELLO Discussion . . . . .	20
16.2. HELLO versus KEEPALIVE . . . . .	20
17. Open Issues . . . . .	21
18. Security Considerations . . . . .	21

19. IANA Considerations . . . . .	21
20. IEEE Considerations . . . . .	22
21. Acknowledgments . . . . .	22
22. References . . . . .	22
22.1. Normative References . . . . .	22
22.2. Informative References . . . . .	23
Authors' Addresses . . . . .	24

## 1. Introduction

The Massive Data Center (MDC) environment presents unusual problems of scale, e.g.  $O(10,000)$  devices, while its homogeneity presents opportunities for simple approaches. Approaches such as Jupiter Rising [JUPITER] use a central controller to deal with scaling, while BGP-SPF [I-D.ietf-lsvr-bgp-spf] provides massive scale-out without centralization using a tried and tested scalable distributed control plane, offering a scalable routing solution in Clos and similar environments. But BGP-SPF and similar higher level device-spanning protocols need link state and addressing data from the network to build the routing topology. LLDP has scaling issues, e.g. in extending a message beyond 1,500 bytes.

Link State Over Ethernet (LSOE) provides brutally simple mechanisms for devices to

- o Discover each other's Layer 2 (MAC) Addresses,
- o Run Layer 2 keep-alive messages for liveness continuity,
- o Discover each other's unique IDs (ASN, RouterID, ...),
- o Discover mutually supported encapsulations, e.g. IP/MPLS,
- o Discover Layer 3 and/or MPLS addressing of interfaces of the link encapsulations,
- o Enable layer 3 link liveness such as BFD, and finally
- o Present these data, using a very restricted profile of a BGP-LS [RFC7752] API, to BGP-SPF which computes the topology and builds routing and forwarding tables.

This protocol may be more widely applicable to a range of routing and similar protocols which need link discovery and characterisation.

## 2. Terminology

Even though it concentrates on the Ethernet layer, this document relies heavily on routing terminology. The following are some possibly confusing terms:

Association: An established, vis OPEN PDUs, session between two LSOE capable devices,

ASN: Autonomous System Number [RFC4271], a BGP identifier for an originator of Layer 3 routes, particularly BGP announcements.

BGP-LS: A mechanism by which link-state and TE information can be collected from networks and shared with external components using the BGP routing protocol. See [RFC7752].

BGP-SPF: A hybrid protocol using BGP transport but a Dijkstra SPF decision process. See [I-D.ietf-lsvr-bgp-spf].

Clos: A hierarchic subset of a crossbar switch topology commonly used in data centers.

Datagram: The LSOE content of a single Ethernet frame. A full LSOE PDU may be packaged in multiple Datagrams.

Encapsulation: Address Family Indicator and Subsequent Address Family Indicator (AFI/SAFI). I.e. classes of addresses such as IPv4, IPv6, MPLS, ...

Frame: An Ethernet Layer 2 packet.

MAC Address: Media Access Control Address, essentially an Ethernet address, six octets.

MDC: Massive Data Center, commonly thousands of TORs.

PDU: Protocol Data Unit, an LSOE application layer message. A PDU may need to be broken into multiple Datagrams to make it through MTU or other restrictions.

RouterID: An 32-bit identifier unique in the current routing domain, see [RFC4271] updated by [RFC6286].

SPF: Shortest Path First, an algorithm for finding the shortest paths between nodes in a graph; AKA Dijkstra's algorithm.

TOR: Top Of Rack switch, aggregates the servers in a rack and connects to aggregation layers of the Clos tree, AKA the Clos spine.

ZTP: Zero Touch Provisioning gives devices initial addresses, credentials, etc. on boot/restart.

## 3. Background

LSOE assumes a datacenter scale and topology, but can accommodate richer topologies which contain potential cycles.

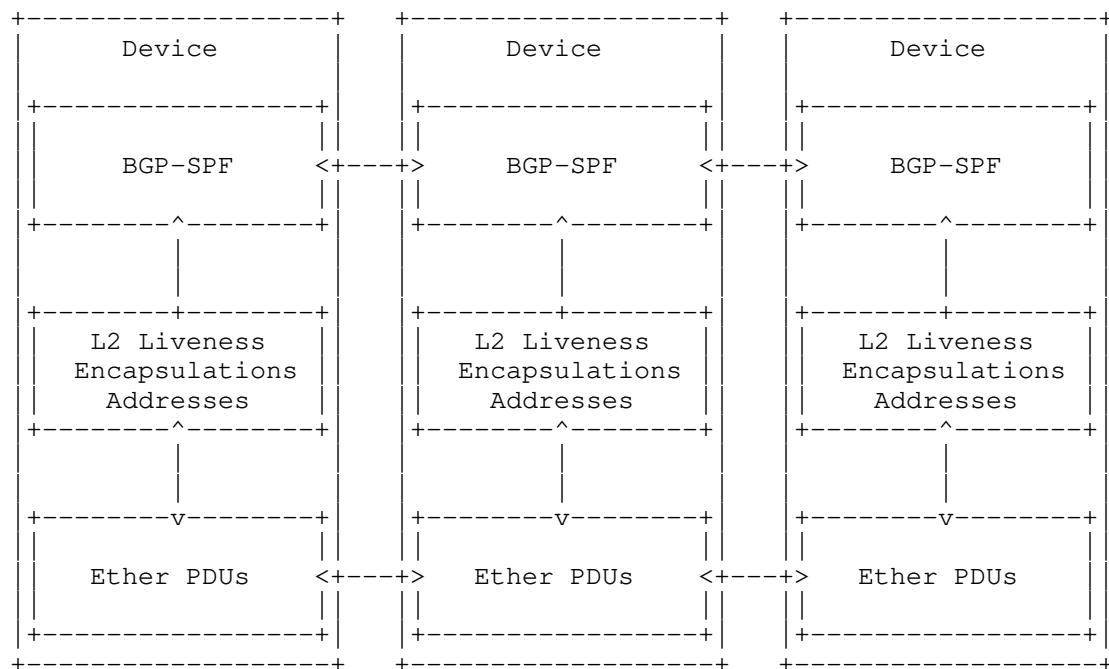
While LSOE is designed for the MDC, there are no inherent reasons it could not run on a WAN; though, as it is simply a discovery protocol, it is not clear that this would be useful. The authentication and

authorisation needed to run safely on the WAN are not provided in detail in this version of the protocol, although future versions/extensions could expend on them.

LSOE assumes a new IEEE assigned EtherType (TBD).

#### 4. Top Level Overview

- o Devices discover each other on Ethernet links
- o MAC addresses and Link State are exchanged over Ethernet
- o Layer 2 Liveness Checks are begun
- o Encapsulation data are exchanged and IP-Level Liveness Checks done
- o A BGP-like protocol is assumed to use these data to discover and build a topology database



There are two protocols, the Ethernet discovery and the interface to the upper level BGP-like protocol:

- o Layer 2 Ethernet protocols are used to exchange Layer 2 data, i.e. MAC addresses, and layer 2.5 and 3 identifiers (not payloads), i.e. ASNs, Encapsulations, and interface addresses.
- o A Link Layer to BGP API presents these data up the stack to a BGP protocol or an other device-spanning upper layer protocol, presenting them using the BGP-LS BGP-like data format.

The upper layer BGP family routing protocols cross all the devices, though they are not part of these LSOE protocols.

To simplify this document, Layer 2 Ethernet framing is not shown.

## 5. Ethernet to Ethernet Protocols

Two devices discover each other and their respective MAC addresses by sending multicast HELLO PDUs (Section 9). To allow discovery of new devices coming up on a multi-link topology, devices send periodic HELLOs forever, see Section 16.1.

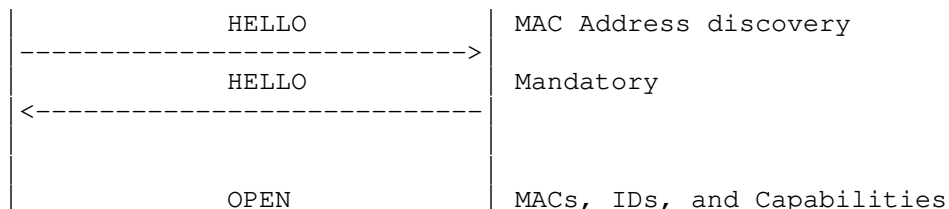
Once a new device is recognized, both devices attempt to negotiate and establish peering by sending unicast OPEN PDUs (Section 10). In an established peering, Encapsulations (Section 12) may be announced and modified. When two devices on a link have compatible Encapsulations and addresses, i.e. the same AFI/SAFI and the same subnet, the link is announced via the BGP-LS API.

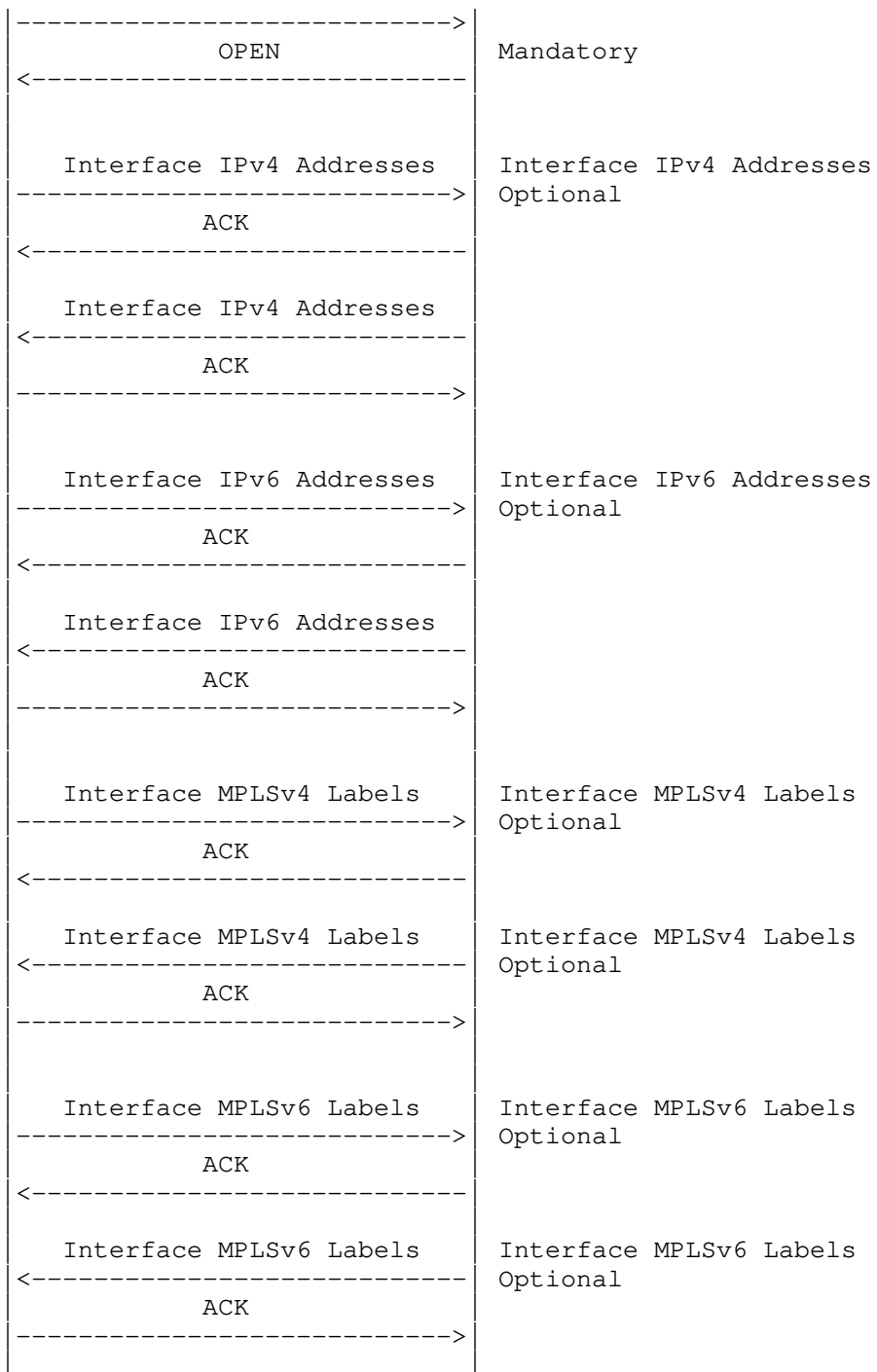
### 5.1. Inter-Link Ether Protocol Overview

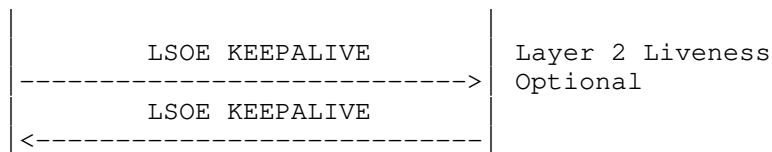
The HELLO, Section 9, is a priming message. It is an Ethernet multicast frame with a small LSOE PDU with the simple goal of discovering the Ethernet MAC address(es) of devices reachable via an interface.

The HELLO and OPEN, Section 10, PDUs, which are used to discover and exchange MAC address and IDs, are mandatory; other PDUs are optional; though at least one encapsulation MUST be agreed at some point.

The following is a ladder-style sketch of the Ethernet protocol exchanges:



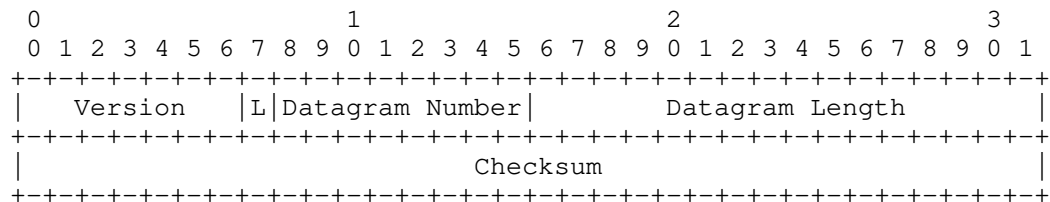




## 6. Transport Layer

LSOE PDU are carried by a simple transport layer which allows long PDUs to occupy multiple Ethernet frames. The LSOE data in each frame is referred to as a Datagram.

The LSOE Transport Layer encapsulates each Datagram using a common transport header.



The fields of the LSOE Transport Header are as follows:

Version: Version number of the protocol, currently 0. Values other than 0 are treated as failure.

Datagram Number: 0..255, a monotonically increasing value, modulo 256, see [RFC1982].

L: A bit that set to 1 if this Datagram is the last Datagram of the PDU. For a PDU which fits in only one Datagram, it is set to one.

PDU Length: Total number of octets in the Datagram including all payloads and fields.

Checksum: A 32 bit hash over the Datagram to detect bit flips, see Section 7.

## 7. The Checksum

There is a reason conservative folk use a checksum in UDP. And as many operators stretch to jumbo frames (over 1,500 octets) longer checksums are the conservative approach.

For the purpose of computing a checksum, the checksum field itself is assumed to be zero.

Sum up 32-bit unsigned ints in a 64-bit long, then take the high-order section, shift it right, rotate, add it in, repeat until zero.

```
#include <stdint.h>
#include <stdint.h>

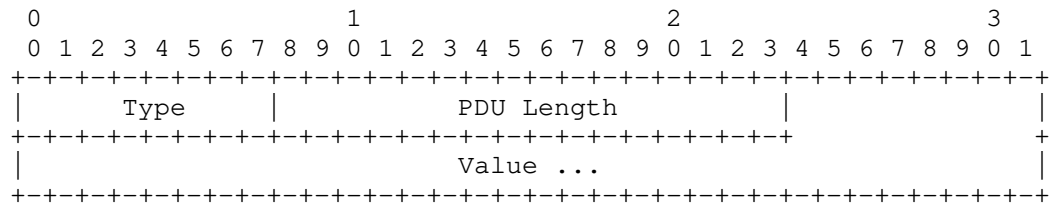
/* The F table from Skipjack, and it would work for the S-Box. */
static const uint8_t sbox[256] = {
0xa3,0xd7,0x09,0x83,0xf8,0x48,0xf6,0xf4,0xb3,0x21,0x15,0x78,
0x99,0xb1,0xaf,0xf9,0xe7,0x2d,0x4d,0x8a,0xce,0x4c,0xca,0x2e,
0x52,0x95,0xd9,0x1e,0x4e,0x38,0x44,0x28,0x0a,0xdf,0x02,0xa0,
0x17,0xf1,0x60,0x68,0x12,0xb7,0x7a,0xc3,0xe9,0xfa,0x3d,0x53,
0x96,0x84,0x6b,0xba,0xf2,0x63,0x9a,0x19,0x7c,0xae,0xe5,0xf5,
0xf7,0x16,0x6a,0xa2,0x39,0xb6,0x7b,0x0f,0xc1,0x93,0x81,0x1b,
0xee,0xb4,0x1a,0xea,0xd0,0x91,0x2f,0xb8,0x55,0xb9,0xda,0x85,
0x3f,0x41,0xbf,0xe0,0x5a,0x58,0x80,0x5f,0x66,0x0b,0xd8,0x90,
0x35,0xd5,0xc0,0xa7,0x33,0x06,0x65,0x69,0x45,0x00,0x94,0x56,
0x6d,0x98,0x9b,0x76,0x97,0xfc,0xb2,0xc2,0xb0,0xfe,0xdb,0x20,
0xe1,0xeb,0xd6,0xe4,0xdd,0x47,0x4a,0x1d,0x42,0xed,0x9e,0x6e,
0x49,0x3c,0xcd,0x43,0x27,0xd2,0x07,0xd4,0xde,0xc7,0x67,0x18,
0x89,0xcb,0x30,0x1f,0x8d,0xc6,0x8f,0xaa,0xc8,0x74,0xdc,0xc9,
0x5d,0x5c,0x31,0xa4,0x70,0x88,0x61,0x2c,0x9f,0x0d,0x2b,0x87,
0x50,0x82,0x54,0x64,0x26,0x7d,0x03,0x40,0x34,0x4b,0x1c,0x73,
0xd1,0xc4,0xfd,0x3b,0xcc,0xfb,0x7f,0xab,0xe6,0x3e,0x5b,0xa5,
0xad,0x04,0x23,0x9c,0x14,0x51,0x22,0xf0,0x29,0x79,0x71,0x7e,
0xff,0x8c,0x0e,0xe2,0x0c,0xef,0xbc,0x72,0x75,0x6f,0x37,0xa1,
0xec,0xd3,0x8e,0x62,0x8b,0x86,0x10,0xe8,0x08,0x77,0x11,0xbe,
0x92,0x4f,0x24,0xc5,0x32,0x36,0x9d,0xcf,0xf3,0xa6,0xbb,0xac,
0x5e,0x6c,0xa9,0x13,0x57,0x25,0xb5,0xe3,0xbd,0xa8,0x3a,0x01,
0x05,0x59,0x2a,0x46
};

/* non-normative example C code, constant time even */

uint32_t sbox_checksum_32(const uint8_t *b, const size_t n)
{
    uint32_t sum[4] = {0, 0, 0, 0};
    uint64_t result = 0;
    for (size_t i = 0; i < n; i++)
        sum[i & 3] += sbox[*b++];
    for (int i = 0; i < sizeof(sum)/sizeof(*sum); i++)
        result = (result << 8) + sum[i];
    result = (result >> 32) + (result & 0xFFFFFFFF);
    result = (result >> 32) + (result & 0xFFFFFFFF);
    return (uint32_t) result;
}
```

## 8. TLV PDUs

The basic LSOE application layer PDU is a typical TLV (Type Length Value) PDU. It may be broken into multiple Datagrams, see Section 6



The fields of the basic LSOE header are as follows:

Type: An integer differentiating PDU payload types

- 0 - HELLO
- 1 - OPEN
- 2 - KEEPALIVE
- 3 - ACK
- 4 - IPv4 Announcement
- 5 - IPv6 Announcement
- 6 - MPLS IPv4 Announcement
- 7 - MPLS IPv6 Announcement
- 8-255 Reserved

PDU Length: Total number of octets in the PDU including all payloads and fields

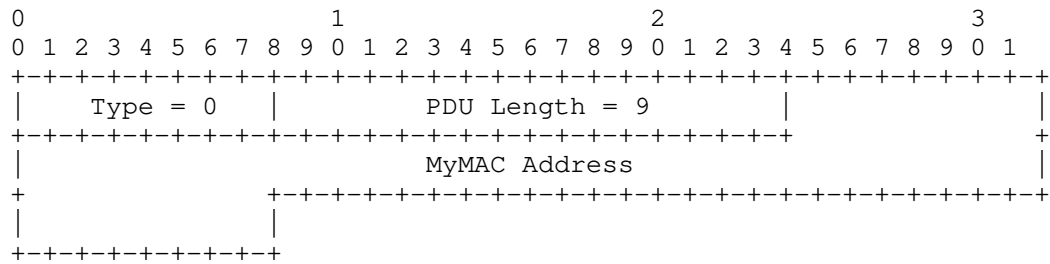
Value: Any application layer content of the LSOE PDU beyond the type.

## 9. HELLO

The HELLO PDU is unique in that it is a multicast Ethernet frame. It solicits response(s) from other device(s) on the link. See Section 16.1 for why multicast is used.

All other LSOE PDUs are unicast Ethernet frames, as the peer's MAC Address is known after the HELLO exchange.

When an interface is turned up on a device, it SHOULD issue a HELLO periodically. The interval is set by configuration.



If more than one device responds, one adjacency is formed for each unique (MAC address) response. LSOE treats the adjacencies as separate links.

When a HELLO is received from a MAC address where there is no established LSOE adjacency, the receiver SHOULD respond with an OPEN PDU. The two devices establish an LSOE adjacency by exchanging OPEN PDUs.

The PDU Length is the octet count of the entire PDU, including the Type, the Datagram Length field itself, and the MyMAC Address payload.

A particular MAC address SHOULD arrive on frames from only one interface.

#### 10. OPEN

Each device has learned the other's MAC address from the HELLO exchange, see Section 9. Therefore the OPEN and subsequent PDUs are unicast, as opposed to the HELLO's multicast, Ethernet frames.



Once two devices know each other's MAC addresses, and have ACKed each other's OPEN PDUs, Layer 2 KEEPALIVES (see Section 13) SHOULD be started to ensure Layer 2 liveness and keep the association semantics alive. The timing and acceptable drop of the KEEPALIVE PDUs SHOULD be configured.

If a properly authenticated OPEN arrives from a device with which the receiving device believes it already has an LSOE association (OPENs have already been exchanged), the receiver MUST assume that the sending device has been reset. All discovered data MUST BE withdrawn via the BGP-LS API and the recipient MUST respond with a new OPEN.

## 11. ACK

```

      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
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|   Type = 3   |           Length = 4           |   PDU Type   |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

The ACK acknowledges receipt of an OPEN or an Encapsulation PDU.

The PDU Type is the Type of the PDU being acknowledged, OPEN or one of the Encapsulations.

### 11.1. Retransmission

If a PDU sender expects an ACK, e.g. for an OPEN or an Encapsulation, and does not receive the ACK for a configurable time (default one second), the sender resends the PDU. This cycle MAY be repeated a configurable number of times (default three) before it is considered a failure. The session is considered closed in case of an ACK failure.

## 12. The Encapsulations

Once the devices know each other's MAC addresses, know each other's upper layer identities, have means to ensure link state, etc., the LSOE 'association' is considered established, and the devices SHOULD announce their interface encapsulation, addresses, (and labels).

The Encapsulation types the peers exchange may be IPv4 Announcement (Section 12.3), IPv6 Announcement (Section 12.4), MPLS IPv4 Announcement (Section 12.6), MPLS IPv6 Announcement (Section 12.7), and/or possibly others not defined here.

The sender of an Encapsulation PDU MUST NOT assume that the peer is capable of the same Encapsulation Type. An ACK (Section 11) merely

acknowledges receipt. Only if both peers have sent the same Encapsulation Type is it safe to assume that they are compatible for that type.

Further, to consider a link of a type to formally be established so that it may be pushed up to upper layer protocols, the addressing for the type must be compatible, i.e. on the same IPvX subnet.

### 12.1. The Encapsulation PDU Skeleton

The header for all encapsulation PDUs is as follows:

```

      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
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|      Type      |      PDU Length      |      Count      |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|      ...      |      Encapsulation List...      |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

The 16-bit Count is the number of Encapsulations in the Encapsulation list.

If the length of an Encapsulation PDU exceeds the Datagram size limit on media, the PDU is broken into multiple Datagrams. See Section 8.

The Receiver MUST acknowledge the Encapsulation PDU with a Type=3, ACK PDU (Section 11) with the Encapsulation Type being that of the encapsulation being announced, see Section 11.

If the Sender does not receive an ACK in one second, they SHOULD retransmit. After a user configurable number of failures, the LSOE association should be considered dead and the OPEN process SHOULD be restarted.

An Encapsulation PDU describes zero or more addresses of the encapsulation type.

An Encapsulation PDU of Type T replaces all previous encapsulations of Type T.

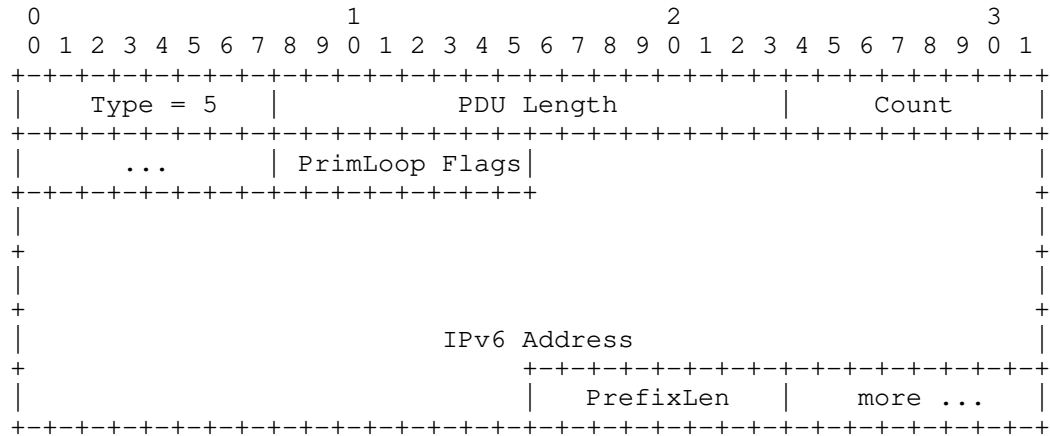
To remove all encapsulations of Type T, the sender uses a Count of zero.

If an interface has multiple addresses for an encapsulation type, one address SHOULD be marked as primary, see Section 12.2.



#### 12.4. IPv6 Encapsulation

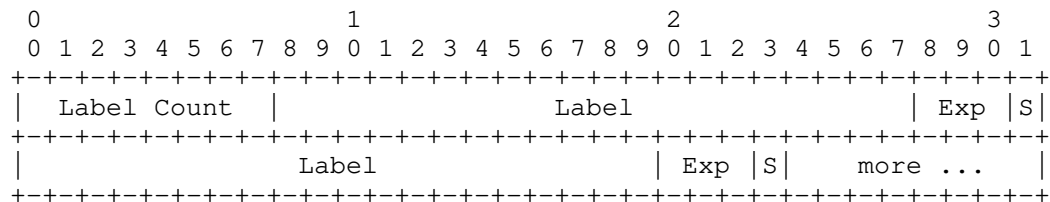
The IPv6 Encapsulation describes a device's ability to exchange IPv6 packets on one or more subnets. It does so by stating the interface's address and the prefix length.



The 16-bit Count is the number of IPv6 Encapsulations.

#### 12.5. MPLS Label List

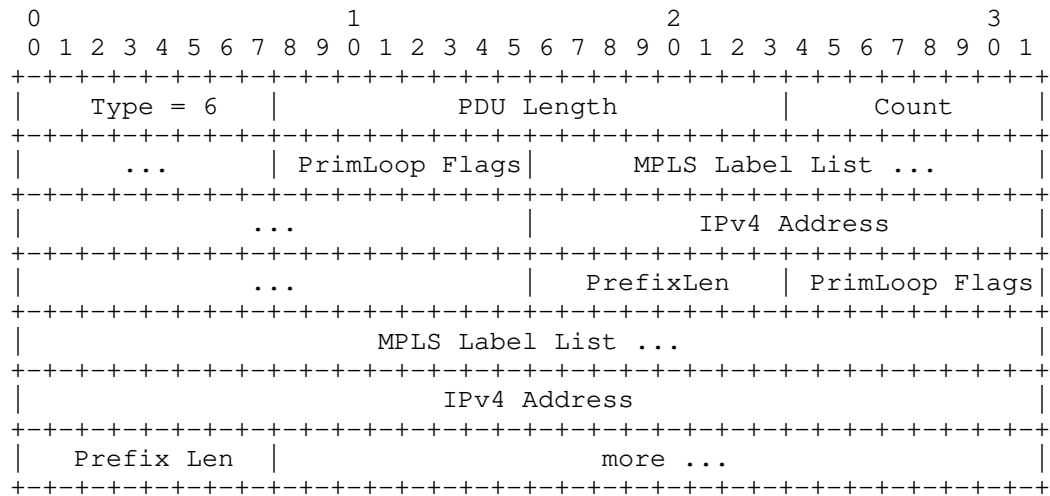
As an MPLS enabled interface may have a label stack, see [RFC3032], a variable length list of labels is needed.



A Label Count of zero is an implicit withdraw of all labels for that prefix on that interface.

#### 12.6. MPLS IPv4 Encapsulation

The MPLS IPv4 Encapsulation describes a device's ability to exchange labeled IPv4 packets on one or more subnets. It does so by stating the interface's address and the prefix length.



The 16-bit Count is the number of MPLSv6 Encapsulations.

#### 12.7. MPLS IPv6 Encapsulation

The MPLS IPv6 Encapsulation describes a device's ability to exchange labeled IPv6 packets on one or more subnets. It does so by stating the interface's address and the prefix length.



```

      0               1               2
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3
+---+---+---+---+---+---+---+---+---+---+---+---+---+
|   Type = 2   |           Length = 3           |
+---+---+---+---+---+---+---+---+---+---+---+---+

```

#### 14. Layers 2.5 and 3 Liveness

Ethernet liveness is continuously tested by KEEPALIVE PDUs, see Section 13. As layer 2.5 or layer 3 connectivity could still break, liveness above layer 2 SHOULD be frequently tested using BFD ([RFC5880]) or a similar technique.

This protocol assumes that one or more Encapsulation addresses will be used to ping, BFD, or whatever the operator configures.

#### 15. The North/South Protocol

Thus far, a one-hop point-to-point link discovery protocol has been defined.

The nodes know the unique node identifiers (ASNs, RouterIDs, ...) and Encapsulations on each link interface.

Full topology discovery is not appropriate at the Ethernet layer, so Dijkstra a la IS-IS etc. is assumed to be done by higher level protocols.

Therefore the node identifiers, link Encapsulations, and state changes are pushed North via a small subset of the BGP-LS API. The upper layer routing protocol(s), e.g. BGP-SPF, learn and maintain the topology, run Dijkstra, and build the routing database(s).

For example, if a neighbor's IPv4 Encapsulation address changes, the devices seeing the change push that change Northbound.

##### 15.1. Use BGP-LS as Much as Possible

BGP-LS [RFC7752] defines BGP-like Datagrams describing link state (links, nodes, link prefixes, and many other things), and a new BGP path attribute providing Northbound transport, all of which can be ingested by upper layer protocols such as BGP-SPF; see Section 4 of [I-D.ietf-lsvr-bgp-spf].

For IPv4 links, TLVs 259 and 260 are used. For IPv6 links, TLVs 261 and 262. If there are multiple addresses on a link, multiple TLV pairs are pushed North, having the same ID pairs.

## 15.2. Extensions to BGP-LS

The Northbound protocol needs a few minor extensions to BGP-LS. Luckily, others have needed the same extensions.

Similarly to BGP-SPF, the BGP protocol is used in the Protocol-ID field specified in table 1 of [I-D.ietf-idr-bgppls-segment-routing-epe]. The local and remote node descriptors for all NLRI are the ID's described in Section 10. This is equivalent to an adjacency SID or a node SID if the address is a loopback address.

Label Sub-TLVs from [I-D.ietf-idr-bgp-ls-segment-routing-ext] Section 2.1.1, are used to associate one or more MPLS Labels with a link.

## 16. Discussion

This section explores some trade-offs taken and some considerations.

### 16.1. HELLO Discussion

There is the question of whether to allow an intermediate switch to be transparent to discovery. We consider that an interface on a device is a Layer 2 or a Layer 3 interface. In theory it could be a Layer 3 interface with no encapsulation or Layer 3 addressing currently configured.

A device with multiple Layer 2 interfaces, traditionally called a switch, may be used to forward frames and therefore packets from multiple devices to one interface, I, on an LSOE speaking device. Interface I could discover a peer J across the switch. Later, a prospective peer K could come up across the switch. If I was not still sending and listening for HELLOs, the potential peering with K could not be discovered. Therefore, interfaces MUST continue to send HELLOs as long as they are turned up.

### 16.2. HELLO versus KEEPALIVE

Both HELLO and KEEPALIVE are periodic. KEEPALIVE might be eliminated in favor of keeping only HELLOs. But currently KEEPALIVE is unicast, has a checksum, is acknowledged, and thus more firmly verifies association existence.

This warrants discussion.

## 17. Open Issues

VLANs/SVIs/Subinterfaces

## 18. Security Considerations

The protocol as is MUST NOT be used outside a datacenter or similarly closed environment due to lack of formal definition of the authentication and authorisation mechanism. These will be worked on in a later effort, likely using credentials configured using ZTP or similar configuration automation.

Many MDC operators have a strange belief that physical walls and firewalls provide sufficient security. This is not credible. All MDC protocols need to be examined for exposure and attack surface.

It is generally unwise to assume that on the wire Ethernet is secure. Strange/unauthorized devices may plug into a port. Mis-wiring is very common in datacenter installations. A poisoned laptop might be plugged into a device's port.

Malicious nodes/devices could mis-announce addressing, form malicious associations, etc.

For these reasons, the OPEN PDU's authentication data exchange SHOULD be used. [ A mandatory to implement authentication is in development. ]

## 19. IANA Considerations

This document requests the IANA create a registry for LSOE PDU Type, which may range from 0 to 255. The name of the registry should be LSOE-PDU-Type. The policy for adding to the registry is RFC Required per [RFC5226], either standards track or experimental. The initial entries should be the following:

PDU Code	PDU Name
0	HELLO
1	OPEN
2	KEEPALIVE
3	ACK
4	IPv4 Announce / Withdraw
5	IPv6 Announce / Withdraw
6	MPLS IPv4 Announce / Withdraw
7	MPLS IPv6 Announce / Withdraw
8-255	Reserved

This document requests the IANA create a registry for LSOE PL Flag Bits, which may range from 0 to 7. The name of the registry should be LSOE-PL-Flag-Bits. The policy for adding to the registry is RFC Required per [RFC5226], either standards track or experimental. The initial entries should be the following:

Bit	Bit Name
----	-----
0	Primary
1	Loopback
2-7	Reserved

## 20. IEEE Considerations

This document requires a new EtherType.

## 21. Acknowledgments

The authors thank Cristel Pelsser for multiple reviews, Jeff Haas for review and comments, Joe Clarke for a useful review, John Scudder deeply serious review and comments, Larry Kreeger for a lot of layer 2 clue, Martijn Schmidt for his contribution, Russ Housley for checksum discussion and sBox, and Steve Bellovin for checksum advice.

## 22. References

### 22.1. 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-11 (work in progress), October 2018.

[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-17 (work in progress), October 2018.

[I-D.ietf-lsvr-bgp-spf]

Patel, K., Lindem, A., Zandi, S., and W. Henderickx, "Shortest Path Routing Extensions for BGP Protocol", draft-ietf-lsvr-bgp-spf-03 (work in progress), September 2018.

- [RFC1982] Elz, R. and R. Bush, "Serial Number Arithmetic", RFC 1982, DOI 10.17487/RFC1982, August 1996, <<http://www.rfc-editor.org/info/rfc1982>>.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, DOI 10.17487/RFC3032, January 2001, <<http://www.rfc-editor.org/info/rfc3032>>.
- [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, <<http://www.rfc-editor.org/info/rfc4271>>.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", RFC 5226, DOI 10.17487/RFC5226, May 2008, <<http://www.rfc-editor.org/info/rfc5226>>.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <<http://www.rfc-editor.org/info/rfc5880>>.
- [RFC6286] Chen, E. and J. Yuan, "Autonomous-System-Wide Unique BGP Identifier for BGP-4", RFC 6286, DOI 10.17487/RFC6286, June 2011, <<http://www.rfc-editor.org/info/rfc6286>>.
- [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, <<http://www.rfc-editor.org/info/rfc7752>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<http://www.rfc-editor.org/info/rfc8174>>.

## 22.2. Informative References

[JUPITER] Singh, A., Germano, P., Kanagala, A., Liu, H., Provost, J., Simmons, J., Tanda, E., Wanderer, J., HAP.1zle, U., Stuart, S., Vahdat, A., Ong, J., Agarwal, A., Anderson, G., Armistead, A., Bannon, R., Boving, S., Desai, G., and B. Felderman, "Jupiter rising", Communications of the ACM Vol. 59, pp. 88-97, DOI 10.1145/2975159, August 2016.

#### Authors' Addresses

Randy Bush  
Arrcus & IIJ  
5147 Crystal Springs  
Bainbridge Island, WA 98110  
United States of America

Email: randy@psg.com

Rob Austein  
Arrcus, Inc

Email: sra@hactrn.net

Keyur Patel  
Arrcus  
2077 Gateway Place, Suite #400  
San Jose, CA 95119  
United States of America

Email: keyur@arrcus.com