

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: June 23, 2019

K. Patel
Arrcus, Inc.
A. Lindem
Cisco Systems
S. Zandi
Linkedin
W. Henderickx
Nokia
December 20, 2018

Shortest Path Routing Extensions for BGP Protocol
draft-ietf-lsvr-bgp-spf-04.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 June 23, 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.

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	6
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	7
4.1.	Node NLRI Usage and Modifications	7
4.2.	Link NLRI Usage	8
4.2.1.	BGP-LS Link NLRI Attribute Prefix-Length TLVs	9
4.2.2.	BGP-LS Link NLRI Attribute BGP SPF Status TLV	9
4.2.3.	BGP-LS Prefix NLRI Attribute SPF Status TLV	10
4.3.	Prefix NLRI Usage	10
4.4.	BGP-LS Attribute Sequence-Number TLV	10
5.	Decision Process with SPF Algorithm	11
5.1.	Phase-1 BGP NLRI Selection	12
5.2.	Dual Stack Support	13
5.3.	SPF Calculation based on BGP-LS NLRI	13
5.4.	NEXT_HOP Manipulation	16
5.5.	IPv4/IPv6 Unicast Address Family Interaction	16
5.6.	NLRI Advertisement and Convergence	17
5.6.1.	Link/Prefix Failure Convergence	17

5.6.2. Node Failure Convergence	17
5.7. Error Handling	18
6. IANA Considerations	18
7. Security Considerations	18
8. Management Considerations	18
8.1. Configuration	18
8.2. Operational Data	18
9. Acknowledgements	19
10. Contributors	19
11. References	19
11.1. Normative References	19
11.2. Information References	20
Authors' Addresses	22

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 additional BGP paths or other extensions. In short, the advantages of an IGP such as OSPF [RFC2328] are available 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 IGP 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. 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.

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 up and considered operational. This is much like the previous peering model only peering is on a single loopback address and the switch fabric links can be unnumbered. However, there will be the same unnumber of sessions as with the previous peering model unless there are parallel links between switches in the fabric.

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.

This peering model, known as sparse peering, allows for many fewer BGP sessions and, consequently, instances of the same NLRI received from multiple peers. It is discussed in greater detail in [I-D.ietf-lsvr-applicability].

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) [RFC8402]. 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 [RFC8402].

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.

Note that usage of Strict Shortest Path First (SPF) algorithm is defined in the IGP algorithm registry but usage is restricted to [I-D.ietf-idr-bgppls-segment-routing-epe]. Hence, its usage for BGP-LS SPF is out of scope.

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.

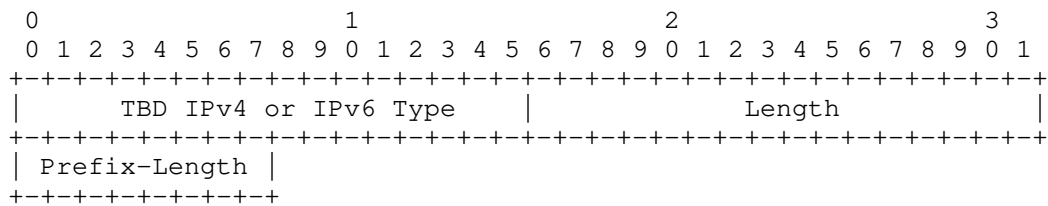
Link NLRI is advertised with local and remote node descriptors as described above and unique link identifiers dependent on the addressing. For IPv4 links, the links local IPv4 (TLV 259) and remote IPv4 (TLV 260) addresses will be used. For IPv6 links, the local IPv6 (TLV 261) and remote IPv6 (TLV 262)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

4.2.1. BGP-LS Link NLRI Attribute Prefix-Length TLVs

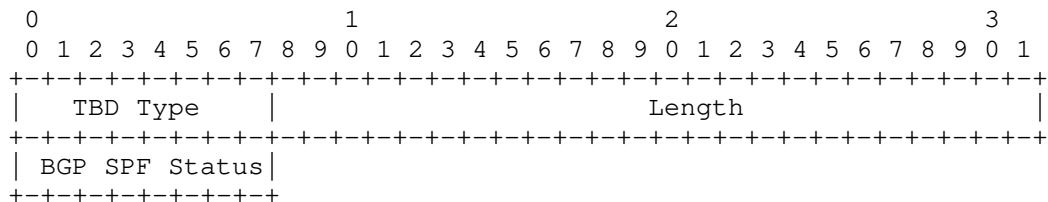
Two BGP-LS Attribute TLVs to BGP-LS Link NLRI are defined to advertise the prefix length associated with the IPv4 and IPv6 link prefixes. The prefix length is used for the optional installation of prefixes corresponding to Link NLRI as defined in Section 5.3.



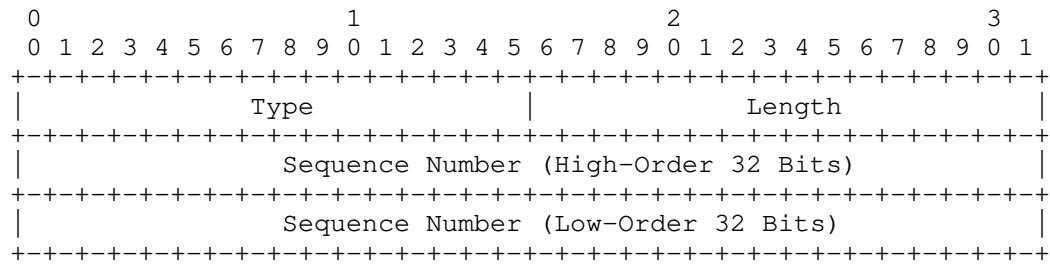
Prefix-length - A one-octet length restricted to 1-32 for IPv4 Link NLIR endpoint prefixes and 1-128 for IPv6 Link NLRI endpoint prefixes.

4.2.2. BGP-LS Link NLRI Attribute BGP SPF Status TLV

A BGP-LS Attribute TLV to BGP-LS 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 5.6.1. If the BGP SPF Status TLV is not included with the Link NLRI, the link is considered up and available.



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



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 MinASOriginationIntervalTimer [RFC4271] are not applicable to the BGP-LS-SPF SAFI. Rather, SPF calculations SHOULD be triggered and dampened consistent with the SPF backoff algorithm specified in [RFC8405].

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.

When a BGP speaker completely loses its sequence number state, i.e., due to a cold start, or in the unlikely possibility that that sequence number wraps, the BGP routing domain will still converge. This is due to the fact that BGP speakers adjacent to the router will always accept self-originated NLRI from the associated speaker as more recent (rule # 1). When BGP speaker reestablishes a connection with its peers, any existing session will be taken down and stale NLRI will be replaced by the new NLRI and stale NLRI will be discarded independent of whether or not BGP graceful restart is deployed, [RFC4724]. The adjacent BGP speaker will update their NLRI advertisements in turn until the BGP routing domain has converged.

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 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. SPF Calculation based on BGP-LS NLRI

This section details the BGP-LS SPF local routing information base (RIB) calculation. The router will use BGP-LS Node, Link, and Prefix NLRI to populate the local RIB using the following algorithm. This calculation yields the set of intra-area routes associated with the BGP-LS domain. A router calculates the shortest-path tree using itself as the root. Variations and optimizations of the algorithm are valid as long as it yields the same set of routes. The algorithm below supports Equal Cost Multi-Path (ECMP) routes. Weighted Unequal Cost Multi-Path 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.

- o Local Route Information Base (RIB) - This is abstract contains reachability information (i.e., next hops) for all prefixes (both IPv4 and IPv6) as well as the Node NLRI reachability. Implementations may choose to implement this as separate RIBs for each address family and/or Node NLRI.
- o Link State NLRI Database (LSNDB) - Database of BGP-LS NLRI that facilitates access to all Node, Link, and Prefix NLRI as well as all the Link and Prefix NLRI corresponding to a given Node NLRI. Other optimization, such as, resolving bi-directional connectivity associations between Link NLRI are possible but of scope of this document.
- o Candidate List - This is a list of candidate Node NLRI with the lowest cost Node NLRI at the front of the list. It is typically implemented as a heap but other concrete data structures have also been used.

The algorithm is comprised of the steps below:

1. The current local RIB is invalidated. The local RIB is built again from scratch. The existing routing entries are preserved for comparison to determine changes that need to be installed in the global RIB.
2. The computing router's Node NLRI is installed in the local RIB with a cost of 0 and as as the sole entry in the candidate list.
3. The Node NLRI with the lowest cost is removed from the candidate list for processing. 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 cost for each prefix is the metric advertised in the Prefix NLRI added to the cost to reach the Current Node.
 - * If the BGP-LS Prefix attribute includes an BGP-SPF Status TLV indicating the prefix is unreachable, the BGP-LS Prefix NLRI is considered unreachable and the next BGP-LS Prefix NLRI is examined.
 - * If the prefix is in the local RIB and the cost is greater than the Current route's metric, the Prefix NLRI does not contribute to the route and is ignored.

- * If the prefix is in the local RIB and the cost is less than the current route's metric, the Prefix is installed with the Current Node's next-hops replacing the local RIB route's next-hops and the metric being updated.
 - * If the prefix is in the local RIB and the cost is same as the current route's metric, the Prefix is installed with the Current Node's next-hops being merged with local RIB route's next-hops.
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 metric in the Link NLRI added to the cost to reach the Current Node.
- * Optionally, the prefix(es) associated with the Current Link are installed into the local RIB using the same rules as were used for Prefix NLRI in the previous steps.
 - * The Current Link's endpoint Node NLRI is accessed (i.e., the Node NLRI with the same Node identifiers as the Link endpoint). If it exists, it will be referred to as the Endpoint Node NLRI and the algorithm will proceed as follows:
 - + If the BGP-LS Link NLRI includes an BGP-SPF Status TLV indicating the link is down, the BGP-LS Link NLRI is considered down and the next BGP-LS Link NLRI is examined.
 - + All the Link NLRI corresponding the Endpoint Node NLRI will be searched for a back-link NLRI pointing to the current node. Both the Node identifiers and the Link endpoint identifiers in the Endpoint Node's Link NLRI must match for a match. If there is no corresponding Link NLRI corresponding to the Endpoint Node NLRI, the Endpoint Node NLIR fails the bi-directional connectivity test and is not processed further.
 - + If the Endpoint Node NLRI is not on the candidate list, it is inserted based on the link cost and BGP Identifier (the latter being used as a tie-breaker).
 - + If the Endpoint Node NLRI is already on the candidate list with a lower cost, it need not be inserted again.

- + If the Endpoint Node NLRI is already on the candidate list with a higher cost, it must be removed and reinserted with a lower cost.
 - * Return to step 3 to process the next lowest cost Node NLRI on the candidate list.
6. The local RIB is examined and changes (adds, deletes, modifications) are installed into the global RIB.

5.4. 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].

5.5. 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 routes, 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.

5.6. NLRI Advertisement and Convergence

5.6.1. 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 normal BGP advertisement, the link would continue to be used until the last copy of the BGP-LS Link NLRI is withdrawn. In order to avoid this delay, the originator of the Link NLRI will advertise a more recent version of the BGP-LS Link NLRI including the BGP-SPF Status TLV Section 4.2.2 indicating the link is down with respect to BGP-SPF. After some configurable period of time, e.g., 2-3 seconds, the BGP-LS Link NLRI can be withdrawn with no consequence. If the link becomes available in that period, the originator of the BGP-LS LINK NLRI will simply advertise a more recent version of the BGP-LS Link NLRI without the BGP-SPF status TLV in the BGP-LS Link Attributes.

Similarly, when a prefix becomes unreachable, a more recent version of the BGP-LS Prefix NLRI will be advertised with the BGP-SPF status TLV Section 4.2.3 indicating the prefix is unreachable in the BGP-LS Prefix Attributes and the prefix will be considered unreachable with respect to BGP SPF. After some configurable period of time, e.g., 2-3 seconds, the BGP-LS Prefix NLRI can be withdrawn with no consequence. If the prefix becomes reachable in that period, the originator of the BGP-LS Prefix NLRI will simply advertise a more recent version of the BGP-LS Prefix NLRI without the BGP-SPF status TLV in the BGP-LS Prefix Attributes.

5.6.2. Node Failure Convergence

With BGP without graceful restart [RFC4724], all the NLRI advertised by 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 NLRI may be withdrawn while these versions are in-flight on longer paths. This will result the older version of the NLRI being used until the new versions arrive and, potentially, unnecessary route flaps. Therefore, BGP-LS SPF NLRI SHOULD always be retained before being implicitly withdrawn for a brief configurable interval, e.g., 2-3 seconds. 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 5.6.1 and the BGP-SPF calculation will failure the bi-directional connectivity check.

5.7. 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 four attribute TLVs for BGP LS NLRI. We request IANA to assign TLVs for the SPF capability, Sequence Number, IPv4 Link Prefix-Length, and IPv6 Link Prefix-Length 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 [RFC4271], [RFC4724], and [RFC7752].

8. Management Considerations

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

8.1. Configuration

In addition to configuration of the BGP-LS SPF address family, implementations SHOULD support the configuratio of the INITIAL_SPF_DELAY, SHORT_SPF_DELAY, LONG_SPF_DELAY, TIME_TO_LEARN, and HOLDDOWN_INTERVAL as documented in [RFC8405].

8.2. 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 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 of each type and the total number of SPF triggering events. Additionally, to troubleshoot SPF scheduling and backoff [RFC8405], the current SPF backoff state, remaining time-to-learn, remaining holddown, last trigger event time, last SPF time, and next SPF time should be available.

9. Acknowledgements

The authors would like to thank Sue Hares, Jorge Rabadan, Boris Hassanov, Dan Frost, and Fred Baker for their review and comments.

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

10. 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

11. References

11.1. 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-15 (work in progress), March 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>>.
- [RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <<https://www.rfc-editor.org/info/rfc8402>>.
- [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 IGP", RFC 8405, DOI 10.17487/RFC8405, June 2018, <<https://www.rfc-editor.org/info/rfc8405>>.

11.2. Information 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", draft-ietf-lsvr-applicability-00 (work in progress), July 2018.

- [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

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

Email: acee@cisco.com

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

Email: szandi@linkedin.com

Wim Henderickx
Nokia
Antwerp
Belgium

Email: 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].

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 (1180)										Length - (1 Octet)																													
SPF Algorithm																																							

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].

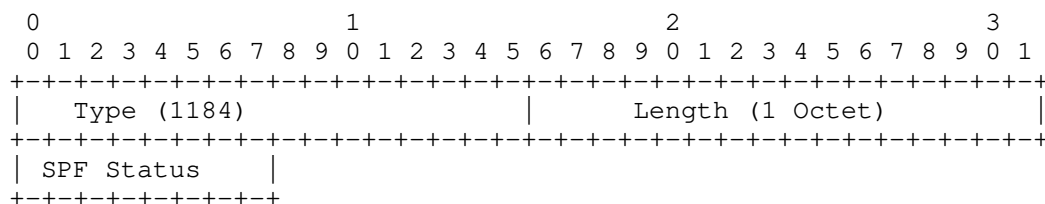
For a link to be used in Shortest Path Tree (SPT) for a given address family, i.e., IPv4 or IPv6, both routers connecting the link MUST have an address in the same subnet for that address family. However, an IPv4 or IPv6 prefix associated with the link MAY be installed without the corresponding address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

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.

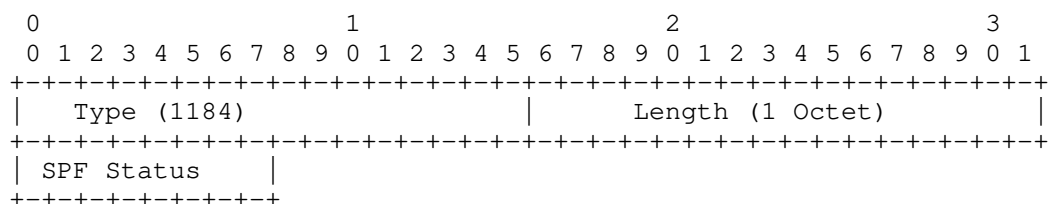
If a BGP SPF speaker received the Link NLRI but the SPF Status TLV is not received, then any previously received information is considered as implicitly withdrawn and the update is propagated to other BGP SPF speakers. A BGP SPF speaker receiving a BGP Update containing an SPF Status TLV in the BGP-LS 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

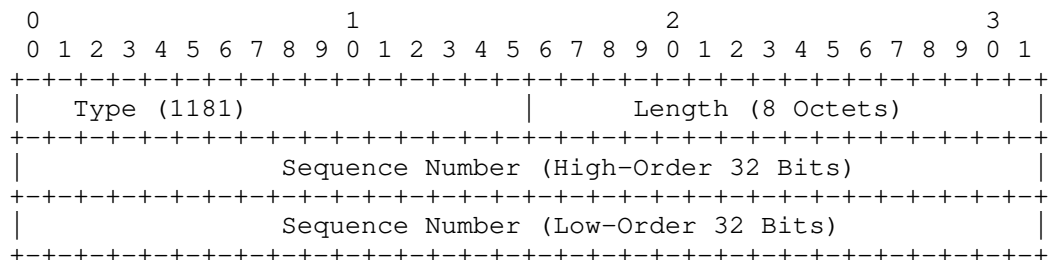
The BGP-LS-SPF Node Attribute SPF Status TLV, Link Attribute SPF Status TLV, and Prefix Attribute SPF Status TLV use the same TLV Type (1184). This implies that a BGP Update cannot contain multiple NLRI

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

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

Email: acee@cisco.com

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

Email: szandi@linkedin.com

Wim Henderickx
Nokia
Antwerp
Belgium

Email: wim.henderickx@nokia.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 22, 2019

R. Bush
Arrcus & IIJ
R. Austein
K. Patel
Arrcus
February 18, 2019

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

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 API (for LSVR), obviating the need for centralized topology distribution 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", "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.

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 August 22, 2019.

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
2. Terminology	4
3. Background	4
4. Top Level Overview	5
5. Ethernet to Ethernet Protocols	6
5.1. Inter-Link Ether Protocol Overview	7
6. Transport Layer	8
7. The Checksum	9
8. TLV PDUs	11
9. HELLO	11
10. OPEN	12
11. ACK	14
11.1. Retransmission	15
12. The Encapsulations	15
12.1. The Encapsulation PDU Skeleton	15
12.2. Prim/Loop Flags	17
12.3. IPv4 Encapsulation	17
12.4. IPv6 Encapsulation	17
12.5. MPLS Label List	18
12.6. MPLS IPv4 Encapsulation	18
12.7. MPLS IPv6 Encapsulation	19
13. KEEPALIVE - Layer 2 Liveness	19
14. VENDOR - Vendor Extensions	20
15. Layers 2.5 and 3 Liveness	20
16. The North/South Protocol	21
16.1. Use BGP-LS as Much as Possible	21
16.2. Extensions to BGP-LS	21
17. Discussion	22
17.1. HELLO Discussion	22
17.2. HELLO versus KEEPALIVE	22
18. VLANs/SVIs/Sub-interfaces	22

19. Implementation Considerations	23
20. Security Considerations	23
21. IANA Considerations	24
22. IEEE Considerations	25
23. Acknowledgments	25
24. References	25
24.1. Normative References	25
24.2. Informative References	27
Authors' Addresses	27

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.

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 session 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:

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. See [IEEE.802_2001].
MDC:	Massive Data Center, commonly thousands of TORs.
MTU:	Maximum Transmission Unit, the size in octets of the largest packet that can be sent on a medium, see [RFC1122] 1.3.3.
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].
Session:	An established, via OPEN PDUs, session between two LSoE capable devices,
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.

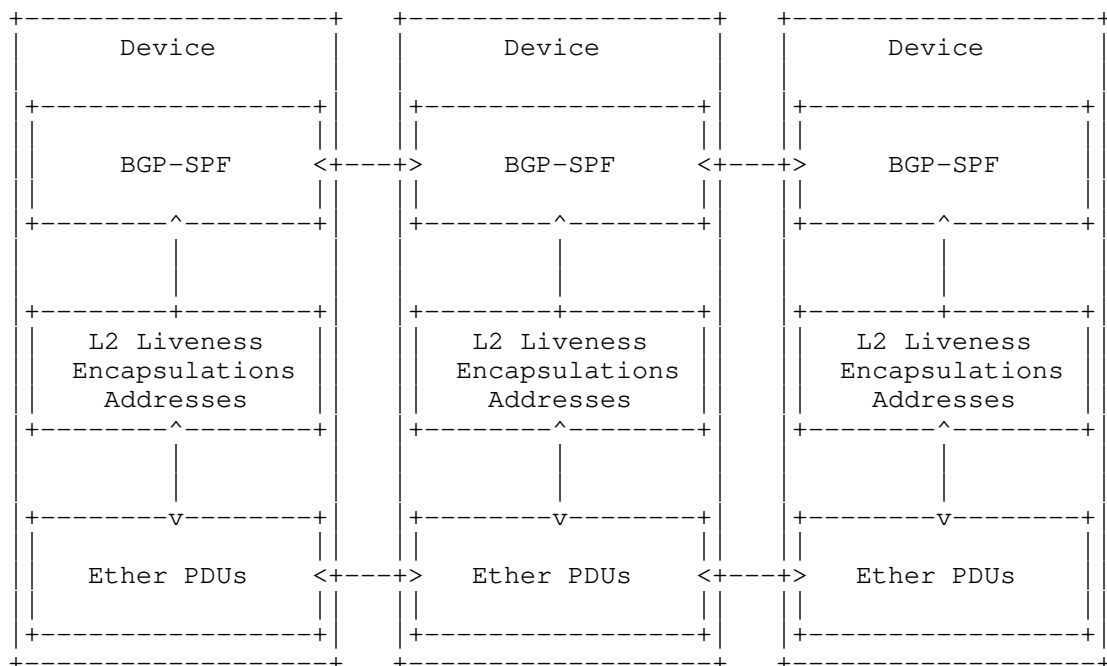
While LSoE is designed for the MDC, there are no inherent reasons it could not run on a WAN; though, as it is simply a discovery protocol, it is not clear that this would be useful. The authentication and authorisation needed to run safely on the WAN are not provided in detail in this version of the protocol, although future versions/extensions could expand on them.

LSoE assumes a new IEEE assigned EtherType (TBD).

As encapsulations may have an inordinate number of addresses, and security will further add to the length of PDUs, LLDP's limitation to 1,500 octets is judged to be too limiting.

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 17.1.

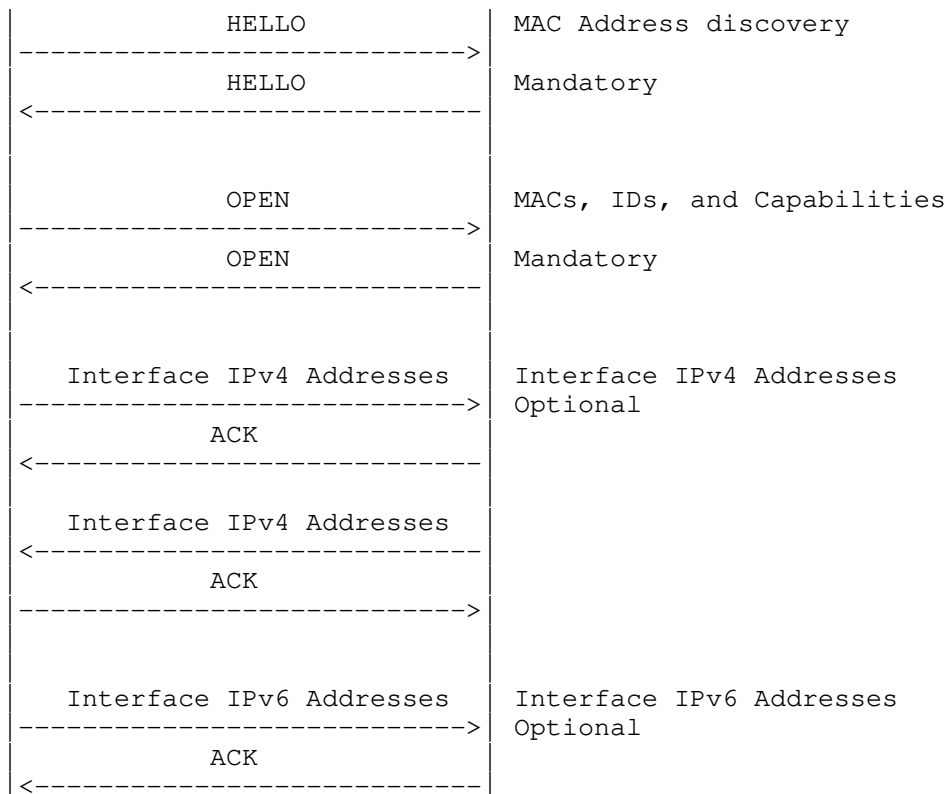
Once a new device is recognized, both devices attempt to negotiate and establish peering by sending unicast OPEN PDUs (Section 10). In an established peering, Encapsulations (Section 12) may be announced and modified. When two devices on a link have compatible Encapsulations and addresses, i.e. the same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

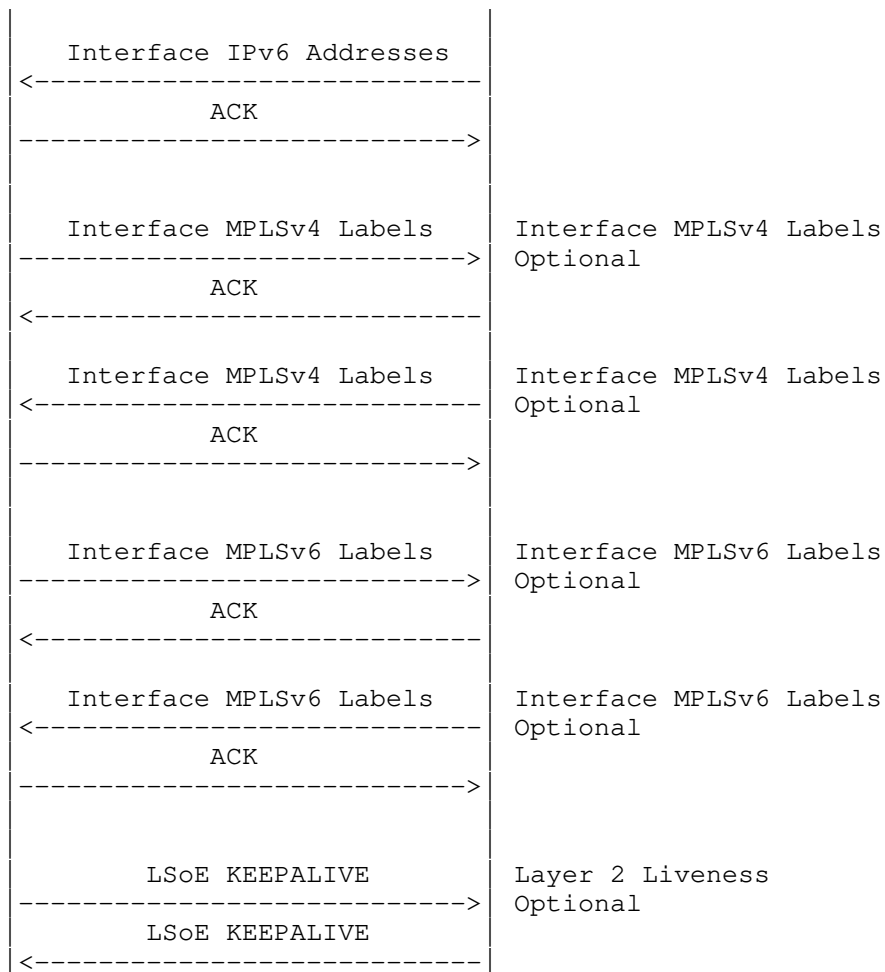
5.1. Inter-Link Ether Protocol Overview

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:

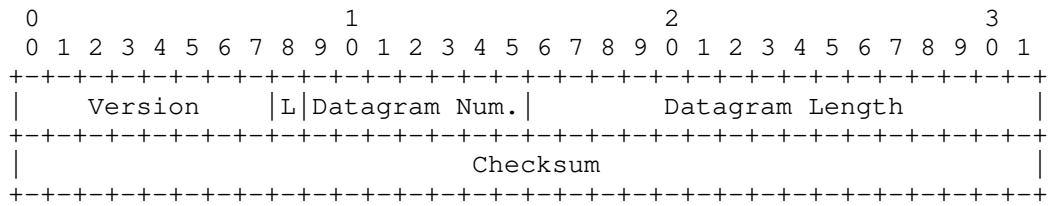




6. Transport Layer

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 errors.

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.

Datagram Number: 0..127, a monotonically increasing value, modulo 128, see [RFC1982].

Datagram 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.

The following code describes the suggested algorithm.

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.

```
<CODE BEGINS>
#include <stddef.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
};

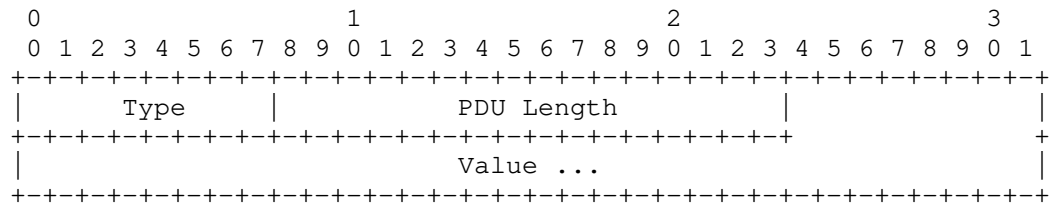
/* non-normative example C code, constant time even */

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;
}

<CODE ENDS>
```

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
- 8-254 Reserved
- 255 - VENDOR

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 17.1 for why multicast is used. The multicast MACs to be used MUST be one of the following, See Clause 9.2.2 of [IEEE802-2014]:

- 01-80-C2-00-00-0E: Nearest Bridge = Propagation constrained to a single physical link; stopped by all types of bridges (including MPRs (media converters)).
- 01-80-C2-00-00-03: Nearest non-TPMR Bridge = Propagation constrained by all bridges other than TPMRs; intended for use within provider bridged networks.

The Nonce enables detection of a duplicate OPEN PDU. It SHOULD be either a random number or time of day. It is needed to prevent session closure due to a repeated OPEN caused by a race or a dropped or delayed ACK.

My ID can be an ASN with high order bits zero, a classic RouterID with high order bits zero, a catenation of the two, a 80-bit ISO System-ID, or any other identifier unique to a single device in the topology. While a link is uniquely identified by a MAC pair, the same ID pair MAY occur on multiple links between the same two devices. IDs are big-endian.

AttrCount is the number of attributes in the Attribute List. Attributes are single octets whose semantics are user-defined.

A node may have zero or more user-defined attributes, e.g. spine, leaf, backbone, route reflector, arabica, ...

Attribute syntax and semantics are local to an operator or datacenter; hence there is no global registry. Nodes exchange their attributes only in the OPEN PDU.

Auth Length is a 16-bit field denoting the length in octets of the Authentication Data, not including the Auth Length itself. If there are no Authentication Data, the Auth Length MUST BE zero.

The Authentication Data are specific to the operational environment. A failure to authenticate is a failure to start the LSoE session, an ERROR PDU is sent (Error Code 2), and HELLOs MUST be restarted.

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 session semantics alive. The timing and acceptable drop of KEEPALIVE PDUs is discussed in Section 13.

If a sender of OPEN does not receive an ACK of the OPEN PDU Type, then they MUST resend the same OPEN PDU, with the same Nonce.

Resending an unacknowledged OPEN PDU, like other ACKed PDUs, SHOULD use exponential back-off, see [RFC1122].

If a properly authenticated OPEN arrives with a new Nonce from a device with which the receiving device believes it already has an LSoE session (OPENs have already been exchanged), the receiver MUST assume that the sending device has been reset. All discovered encapsulation date SHOULD be withdrawn via the BGP-LS API and the recipient MUST respond with a new OPEN. Then encapsulations SHOULD

NOT be kept because. while the new OPEN is likely to be followed by new encapsulation PDUs of the same data, the old session might have an encapsulation type not in the new session.

11. ACK

The ACK PDU acknowledges receipt of a PDU and reports any error condition which might have been raised.

```

      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  |      PDU Length = 8      |      PDU Type      |
+-----+-----+-----+-----+-----+-----+-----+-----+
| EType |      Error Code      |      Error Hint      |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

The ACK acknowledges receipt of an OPEN, Encapsulation, VENDOR PDU, etc.

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

If there was an error processing the received PDU, then the EType is non-zero. If the EType is zero, Error Code and Error Hint MUST also be zero.

A non-zero EType is the receiver's way of telling the PDU's sender that the receiver had problems processing the PDU. The Error Code and Error Hint will tell the sender more detail about the error.

The decimal value of EType gives a strong hint how the receiver sending the ACK believes things should proceed:

- 0 - No Error, Error Code and Error Hint MUST be zero
- 1 - Warning, something not too serious happened, continue
- 2 - Session should not be continued, try to restart
- 3 - Restart is hopeless, call the operator
- 4-15 - Reserved

Someone stuck in the 1990s might think of the error codes as 0x1zzz, 0x2zzz, etc. They might be right. Or not.

The Error Code indicates the type of error.

The Error Hint is any additional data the sender of the error PDU thinks will help the recipient or the debugger with the particular error.

11.1. Retransmission

If a PDU sender expects an ACK, e.g. for an OPEN, an Encapsulation, a VENDOR PDU, etc., and does not receive the ACK for a configurable time (default one second), the sender resends the PDU using exponential back-off, see [RFC1122].. 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 this 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 session is considered established, and the devices SHOULD announce their interface encapsulations, 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.

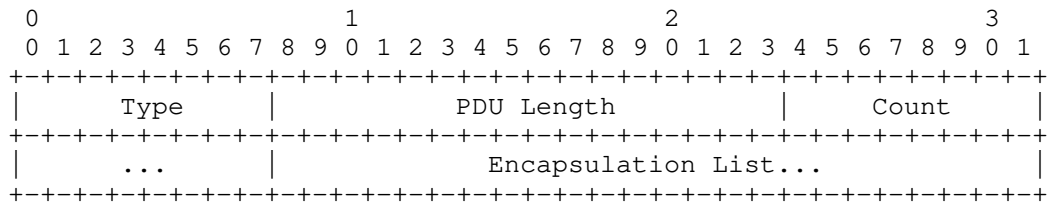
The sender of an Encapsulation PDU MUST NOT assume that the peer is capable of the same Encapsulation Type. An ACK (Section 11) merely acknowledges receipt. Only if both peers have sent the same Encapsulation Type is it safe to assume that they are compatible for that type.

A receiver of an encapsulation might recognize an addressing conflict, such as both ends of the link trying to use the same address. In this case, the receiver SHOULD respond with an ERROR (Error Code 1) instead of an ACK. As there may be other usable addresses or encapsulations, this error might log and continue, letting an upper layer topology builder deal with what works.

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, e.g. on the same IPvX subnet.

12.1. The Encapsulation PDU Skeleton

The header for all encapsulation PDUs is as follows:



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 a configurable interval (default one second), they SHOULD retransmit. After a user configurable number of failures, the LSoE session 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 and only one address SHOULD be configured to be marked as primary, see Section 12.2.

Loopback addresses are generally not seen directly on an external interface. One or more loopback addresses MAY be exposed by configuration on one or more LSoE speaking external interfaces, e.g. for iBGP peering. They SHOULD be marked as such, see Section 12.2.

If there is exactly one non-loopback address for an encapsulation type on an interface, it SHOULD be marked as primary.

If a sender has multiple links on the same interface, separate data, ACKs, etc. must be kept for each peer.

Over time, multiple Encapsulation PDUs may be sent for an interface as configuration changes.

12.2. Prim/Loop Flags

0	1	2	3	...	7
Primary	Loopback	Reserved ...			

Each Encapsulation interface address MAY be marked as a primary address, and/or a loopback, in which case the respective bit is set to one.

Only one address MAY be marked as primary for an encapsulation type.

12.3. IPv4 Encapsulation

The IPv4 Encapsulation describes a device's ability to exchange IPv4 packets on one or more subnets. It does so by stating the interface's addresses and the corresponding prefix lengths.

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 = 4	PDU Length	Count	
...	PrimLoop Flags	IPv4 Address	
...	PrefixLen	more ...	

The 16-bit Count is the number of IPv4 Encapsulations.

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 addresses and the corresponding prefix lengths.

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 = 6										PDU Length										Count																			
...										PrimLoop Flags										MPLS Label List ...																			
...																				IPv4 Address																			
...																				PrefixLen										more ...									

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 addresses, the corresponding prefix lengths, and the corresponding labels.

[illegible]

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

13. KEEPALIVE - Layer 2 Liveness

LSOE devices MUST beacon occasional Layer 2 KEEPALIVE PDUs to ensure session continuity.

They SHOULD be beaconed at a configured frequency. One per second is the default. Layer 3 liveness, such as BFD, will likely be more aggressive.

If a KEEPALIVE is not received from a peer with which a receiver has an open session for a configurable time (default one minute), the session SHOULD BE presumed closed. The devices MAY keep configuration state until a new session is established and new Encapsulation PDUs are received.

```

      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. VENDOR - Vendor 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 = 255   |           Length           |           ...           |
+-----+-----+-----+-----+-----+-----+-----+-----+
|           Enterprise Number           |           Ent Type           |
+-----+-----+-----+-----+-----+-----+-----+-----+
|           Enterprise Data ...           |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Vendors or enterprises may define TLVs beyond the scope of LSoE standards. This is done using a Private Enterprise Number [IANA-PEN] followed by free form data.

Ent Type allows a VENDOR PDU to be sub-typed in the event that the vendor/enterprise needs multiple PDU types.

As with Encapsulation PDUs, a receiver of a VENDOR PDU MUST respond with an ACK or an ERROR PDU. Similarly, a VENDOR PDU MUST only be sent over an open session.

15. 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.

16. 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.

16.1. 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.

16.2. 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 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.

17. Discussion

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

17.1. 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.

17.2. HELLO versus KEEPALIVE

Both HELLO and KEEPALIVE are periodic. KEEPALIVE might be eliminated in favor of keeping only HELLOs. But currently KEEPALIVE is unicast, and thus less noisy on the network, especially if HELLO is configured to transit layer-2-only switches.

This warrants discussion.

18. VLANs/SVIs/Sub-interfaces

One can think of the protocol as an instance (i.e. state machine) which runs on each link of a device.

As the upper routing layer must view VLAN topologies as separate graphs, LSoE treats VLAN ports as separate links.

LSoE PDUs learned over VLAN-ports may be interpreted by upper layer-3 routing protocols as being learned on the corresponding layer-3 SVI interface for the VLAN.

As Sub-Interfaces each have their own MAC, they act as separate interfaces, forming their own links.

19. Implementation Considerations

An implementation SHOULD provide the ability to configure an interface as LSoE speaking or not.

An implementation SHOULD provide the ability to configure whether HELLOs on an LSoE enabled interface send Nearest Bridge or Nearest non-TPMR Bridge multicast frames from that interface; see Section 9.

An implementation SHOULD provide the ability to distribute one or more loopback addresses or interfaces into LSoE on an external LSoE speaking interface.

An implementation SHOULD provide the ability to configure one of the addresses of an encapsulation as primary on an LSoE speaking interface. If there is only one address for a particular encapsulation, the implementation MAY mark it as primary by default.

20. Security Considerations

The protocol as it 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 sessions, etc.

If OPENs are not being authenticated, an attacker could forge an OPEN for an existing session and cause the session to be reset.

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

21. 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
8-254	Reserved
255	VENDOR

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

This document requests the IANA create a registry for LSoE Error Codes, a 16 bit integer. The name of the registry should be LSoE-Error-Codes. The policy for adding to the registry is RFC Required per [RFC5226], either standards track or experimental. The initial entries should be the following:

Error Code	Error Name
-----	-----
0	Reserved
1	Link Addressing Conflict
2	Authorisation Failure in OPEN

22. IEEE Considerations

This document requires a new EtherType.

23. Acknowledgments

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

24. References

24.1. 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-04 (work in progress), December
2018.
- [IANA-PEN]
"IANA Private Enterprise Numbers",
<[https://www.iana.org/assignments/enterprise-numbers/
enterprise-numbers](https://www.iana.org/assignments/enterprise-numbers/enterprise-numbers)>.
- [IEEE.802_2001]
IEEE, "IEEE Standard for Local and Metropolitan Area
Networks: Overview and Architecture", IEEE 802-2001,
DOI 10.1109/ieeestd.2002.93395, July 2002,
<<http://ieeexplore.ieee.org/servlet/opac?punumber=7732>>.

- [IEEE802-2014] Institute of Electrical and Electronics Engineers, "Local and Metropolitan Area Networks: Overview and Architecture", IEEE Std 802-2014, 2014.
- [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>>.

24.2. 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.
- [RFC1122] Braden, R., Ed., "Requirements for Internet Hosts - Communication Layers", STD 3, RFC 1122, DOI 10.17487/RFC1122, October 1989, <<http://www.rfc-editor.org/info/rfc1122>>.

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

INTERNET-DRAFT

N. Malhotra, Ed.
K. Patel
Arrcus

Intended Status: Proposed Standard

J. Rabadan
Nokia

Expires: Sept 12, 2019

Mar 11, 2019

LSoE-based PE-CE Control Plane for EVPN
draft-malhotra-bess-evpn-lsoe-00

Abstract

In an EVPN network, EVPN PEs provide VPN bridging and routing service to connected CE devices based on BGP EVPN control plane. At present, there is no PE-CE control plane defined for an EVPN PE to learn CE MAC, IP, and any other routes from a CE that may be distributed in EVPN control plane to enable unicast flows between CE devices. As a result, EVPN PEs rely on data plane based gleaning of source MACs for CE MAC learning, ARP/ND snooping for CE IPv4/IPv6 learning, and in some cases, local configuration for learning prefix routes behind a CE. A PE-CE control plane alternative to this traditional learning approach, where applicable, offers certain distinct advantages that in turn result in simplified EVPN operation.

This document defines a PE-CE control plane as an optional alternative to traditional non-control-plane based PE-CE learning in an EVPN network. It defines PE-CE control plane procedures and TLVs based on LSoE as the base protocol, enumerates advantages that may be achieved by using this PE-CE control plane, and discusses in detail EVPN use cases that are simplified as a result.

Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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".

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2017 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	4
1.1	Terminology	5
2.	PE <-> CE Control Plane Overview	7
3.	TLVs	9
3.1	Overlay IPv4 Encapsulation PDU	9
3.2	Overlay IPv6 Encapsulation PDU	10
3.3	Overlay IPv4 Prefix Encapsulation PDU	12
3.4	Overlay IPv6 Prefix Encapsulation PDU	13
4.	CE MAC/IP Learning on a PE AC	14
4.1	PE <-> CE LSoE Session Establishment	14
4.2	CE MAC/IP Learning	14
5.	PE Any-cast GW MAC/IP Learning on CE	15
6.	Remote CE MAC/IP Learning on CE	15
7.	PE <-> CE Control Plane with EVPN All-active Multi-Homing	16
7.1	All-active Multi-Homing Mode	16
7.2	Source MAC	17
7.3	CE MAC/IP Learning with EVPN All-active Multi-Homing	17
7.4	LAG Member Link Failure	18
7.4.1	Session Re-establishment	18
7.4.2	TLV Retention	18
7.4	LAG Failure	18
7.5	Example PE <-> CE Control Plane Flow with All-active	

Multi-Homing	19
8. Software Neighbor Tables	21
9. MAC/IP Learning Conflict Resolution	21
10. PE-CE Overlay Prefix Learning	22
11. Asymmetric EVPN-IRB	22
12. Centralized Gateway EVPN-IRB	22
13. Use Cases	22
13.1 Simplified EVPN Operations	22
13.1.1 EVPN All-active Multi-Homing	23
13.1.2 Convergence on CE Host Moves	24
13.1.2.1 Silent Hosts	24
13.1.2.2 Probing	25
13.1.3 ARP Gleaning Latency	26
13.2 Applicability to non-EVPN Use Cases	26
14. Summary	26
15. References	28
15.1 Normative References	28
15.2 Informative References	28
16. Acknowledgements	29
Contributors	29
Authors' Addresses	29

1 Introduction

In an EVPN network, CE devices typically connect to an EVPN PE via layer-2 interfaces that terminate in a BD on the PE. Multi-homed LAG interfaces together with EVPN all-active multi-homing procedures are used to achieve PE-CE link and PE node redundancy for fault-tolerance and load-balancing. PEs provide overlay bridging and, optionally, first-hop routing service for these CE devices based on an EVPN control plane that is used to distribute CE MAC, IP, and prefix reachability across PEs.

At present, there is no PE-CE control plane defined for an EVPN PE to learn connected CE host MACs and IPs. As a result, EVPN PEs rely on:

- o data plane based gleaning of source MAC for MAC learning,
- o ARP snooping for IPv4 + MAC learning, and
- o ND snooping for IPv6 + MAC learning.

A PE-CE control plane alternative to this traditional learning approach, where applicable, can offer some distinct advantages across various boot-up, mobility, and convergence scenarios:

- o PE-CE learning is decoupled from non-deterministic hashing of data, ARP, and ND packets from CEs over all-active multi-homed LAG interfaces.
- o PE-CE learning is decoupled from non-deterministic periodicity of data traffic from CEs or, in an extreme scenario, from CE device being silent for an extended period.
- o PE-CE learning is decoupled from non-deterministic CE behavior with respect to unsolicited ARPs and NAs following boot-up and moves.
- o PE-CE learning is decoupled from latencies associated with data packet triggered ARP and ND gleaning.

This in-turn results in simplification of certain EVPN operations such as aliasing, MAC and IP syncing across multi-homing PEs, and probing on MAC/IP moves. In addition, it helps achieve a deterministic convergence behavior across various boot-up, mobility, and failure scenarios.

A PE may also use local policy configuration for learning prefixes behind a CE that does not run a dynamic routing protocol. A PE-CE control plane can provide an operationally simpler alternative to local configuration for such use cases, where CE and PE devices are not under the same configuration management entity.

This document defines a new PE-CE control plane as an alternative to traditional data-plane and ARP/ND snooping based PE-CE host learning

and to local configuration-based PE-CE prefix learning. It defines PE-CE control plane procedures and TLVs based on [LSOE] as the base protocol, enumerates advantages that may be achieved by using this PE-CE control plane, and discusses in detail EVPN operations that are simplified as a result. Use of PE-CE control plane defined in this document is intended to be optional and backwards compatible with CEs that use traditional PE-CE learning within the same BD.

1.1 Terminology

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 BCP14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

The following terms are used in this document:

- o LSoE: Link State over Ethernet Protocol defined in [LSOE]
- o EVPN-IRB: A BGP-EVPN distributed control plane based integrated routing and bridging fabric overlay discussed in [EVPN-IRB]
- o Underlay: IP or MPLS fabric core network that provides IP or MPLS routed reachability between EVPN PEs.
- o Overlay: VPN or service layer network consisting of EVPN PEs OR VPN provider-edge (PE) switch-router devices that runs on top of an underlay routed core.
- o EVPN PE: A PE switch-router in a data-center fabric that runs overlay BGP-EVPN control plane and connects to overlay CE host devices. An EVPN PE may also be the first-hop layer-3 gateway for CE/host devices. This document refers to EVPN PE as a logical function in a data-center fabric. This EVPN PE function may be physically hosted on a top-of-rack switching device (ToR) OR at layer(s) above the ToR in the Clos fabric. An EVPN PE is typically also an IP or MPLS tunnel end-point for overlay VPN flows.
- o CE: A tenant host device that has layer 2 connectivity to an EVPN PE switch-router, either directly OR via intermediate switching device(s).
- o Symmetric EVPN-IRB: An overlay fabric first-hop routing architecture as defined in [EVPN-IRB], wherein, overlay host-to-host routed inter-subnet flows are routed at both ingress and egress EVPN PEs.
- o Asymmetric EVPN-IRB: An overlay fabric first-hop routing architecture as defined in [EVPN-IRB], wherein, overlay host-to-host routed inter-subnet flows are routed and bridged at ingress PE and bridged at egress PEs.
- o Centralized EVPN-IRB: An overlay fabric first-hop routing architecture, wherein, overlay host-to-host routed inter-subnet

flows are routed at a centralized gateway, typically at the one of the spine layers, and where EVPN PEs are pure bridging devices.

- o ARP: Address Resolution Protocol [RFC 826].
- o ND: IPv6 Neighbor Discovery Protocol [RFC 4861].
- o Ethernet-Segment: physical Ethernet or LAG port that connects an access device to an EVPN PE, as defined in [RFC 7432].
- o ESI: Ethernet Segment Identifier as defined in [RFC 7432].
- o LAG: Layer-2 link-aggregation, also known as layer-2 bundle port-channel, or bond interface.
- o EVPN all-active multi-homing: PE-CE all-active multi-homing achieved via a multi-homed layer-2 LAG interface on a CE with member links to multiple PEs and related EVPN procedures on the PEs.
- o EVPN Aliasing: multi-homing procedure as defined in [RFC 7432].
- o BD: Broadcast Domain.
- o Bridge Table: An instantiation of a broadcast domain on a MAC-VRF.
- o AC: A PE Attachment Circuit. This may be an access (untagged) or trunk (tagged) layer-2 interface that is a member of a local VLAN or a BD.

2. PE <-> CE Control Plane Overview

The Link State over Ethernet (LSoE) protocol is defined in [LSoE] as a protocol over Ethernet links to auto-discover connected neighbor's layer 2, layer 3 attributes, and encapsulations for the purpose of bringing up upper layer routing protocols. This document leverages LSoE as a PE-CE protocol in an EVPN network fabric on access links between an EVPN PE and CE. Specifically,

- o PE-CE control plane based on LSoE protocol is proposed for CE MAC learning as an alternative to data-plane based source MAC learning.
- o PE-CE control plane based on LSoE protocol is proposed for CE MAC-IP adjacency learning as an alternative to MAC-IP learning based on ARP/ND snooping.
- o PE-CE control plane based on LSoE is proposed for learning of IP Prefixes and associated overlay indexes, as an alternative to local configuration on the PE for use case defined in section 4.1 of [EVPN-PREFIX-ADV].

Note that any specification related to base LSoE protocol itself is considered out of scope for this document and will continue to be covered in the base protocol spec. This document will instead focus on procedures and TLV extensions needed to achieve the above learning on PE-CE links in an EVPN network. Any text that relates to the base protocol included in this document is simply background information in the context of use cases covered in this document. The reader should refer to the base LSoE protocol document for the exact LSoE protocol specification.

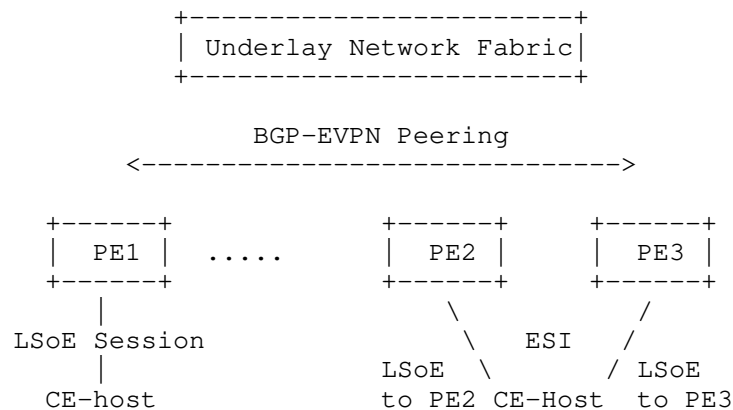


Figure 1

An LSoE session is established on layer-2 access interfaces between the EVPN PE and each connected CE host device. A session end-point is identified by a peer device MAC address on a layer-2 interface. LSoE HELLO messages are used for end-point discovery and OPEN messages are exchanged between two end-points to establish an LSoE peering. Once LSoE peering is established, encapsulation TLVs are exchanged for learning.

In the context of an EVPN network, CE Attachment Circuits (AC logical interfaces) typically terminate in a BD on the PE, with multi-homed LAG interfaces used for EVPN all-active multi-homing. CE hosts may be directly connected to EVPN PEs via access ports, or may be connected on trunk-ports via another switch. In a common EVPN-IRB design, EVPN PEs also function as distributed first-hop gateways for hosts in a BD. While symmetric and asymmetric IRB designs are possible as discussed in [EVPN IRB], procedures described in subsequent sections assume symmetric IRB with distributed any-cast gateways on EVPN PEs. Any deviations from these procedures for asymmetric IRB design or a centralized IRB design will be covered in future updates to this document.

The next few sections will focus on additional LSoE TLVs and procedures needed for PE-CE learning on EVPN PE ACs without and with all-active multi-homing.

3. TLVs

This section defines new TLVs that are used by PE-CE control plane defined in this document.

3.1 Overlay IPv4 Encapsulation PDU

A new encapsulation PDU type is defined for the purpose of carrying overlay IPv4 and MAC bindings. Alternatively, it may also be used to carry an overlay MAC with a NULL IPv4 address in a non-IRB use case.

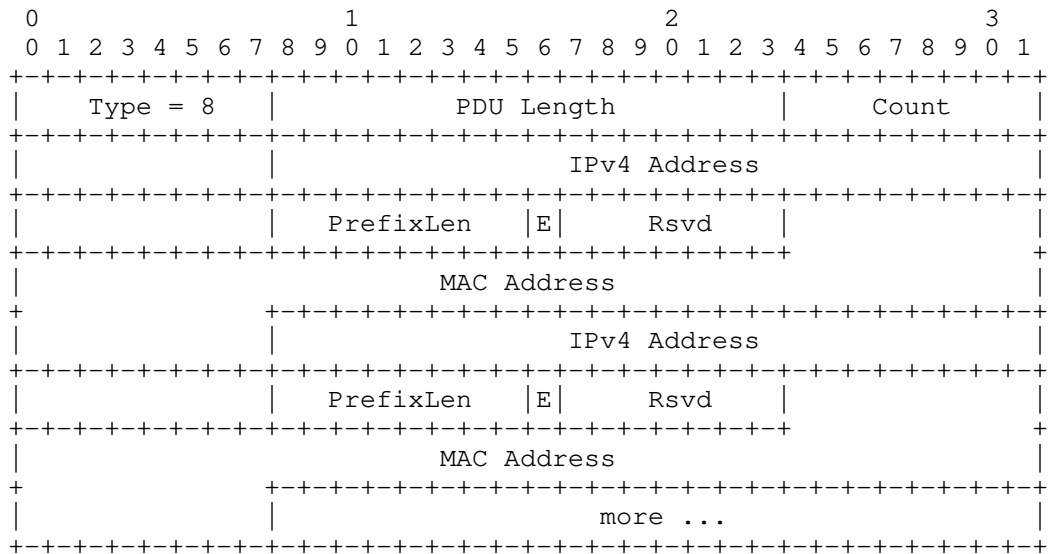


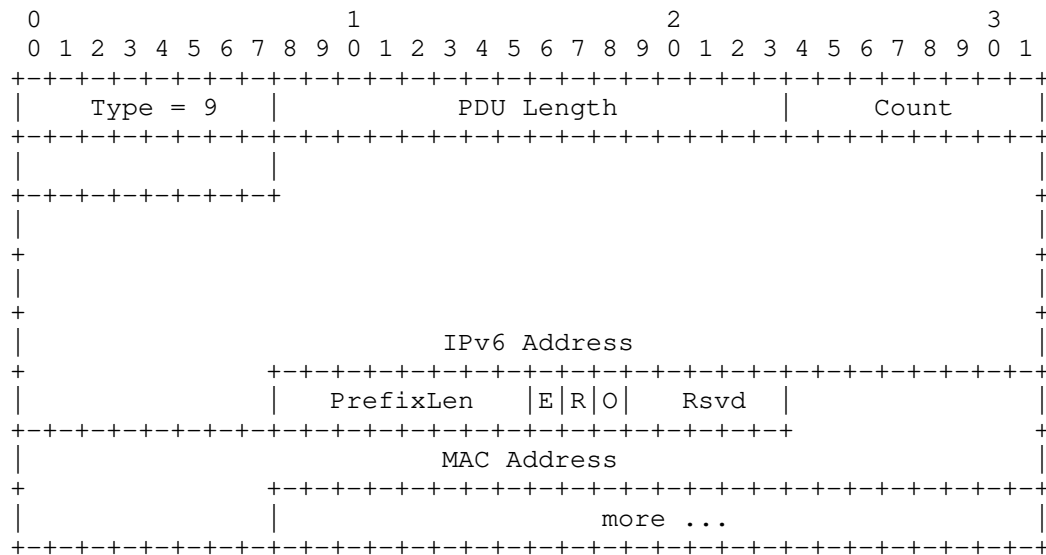
Figure 2

- o A new LSoE PDU type (8) is requested for this PDU.
- o The IPv4 Address is that of an overlay.
- o MAC address carries the MAC binding for the particular IPv4 address if one is set in the PDU. If an IPv4 address is not set, it simply signals an overlay MAC address.
- o EVPN flag 'E' indicates if this encapsulation is being sent on behalf of a remote host learnt via EVPN. Use of this flag is covered in a later section.

This PDU is used to carry PE's any-cast gateway IPv4 address and MAC bindings to a CE host device. Optionally, it may also be used to relay a remote CE's IPv4 address and MAC bindings to a local CE host within a subnet, as well as to send local CE IPv4 address and MAC binding to the PE. Procedures related to use of this PDU are

The encapsulation list in this PDU MUST follow full replace semantics as in the LSoE protocol specification.

A new encapsulation PDU type is defined for the purpose of carrying overlay IPv6 and MAC bindings:



- o A new LSoE PDU type (9) is requested for this PDU.
- o The IPv6 Address is that of an overlay.
- o MAC address carries the MAC binding for IPv6 address in the PDU.
- o An EVPN flag 'E' indicates if this encapsulation is being sent on behalf of a remote host learnt via EVPN. Usage of this flag is covered in a later section.
- o A Router flag 'R' is used to carry "Router Flag" or "R-bit" as defined in [RFC4861]. Usage of this flag for the purpose of installing ND cache entries based on learning via this TLV is as defined in [RFC4861]

- o An Override flag 'O' is used to carry "Override Flag" or "O-bit" as defined in [RFC4861]. Usage of this flag for the purpose of installing ND cache entries based on learning via this TLV is as defined in [RFC4861]

This PDU is used to carry PE's any-cast gateway IPv6 address and MAC bindings to a CE host device. Optionally, it may also be used to relay a remote CE's IPv6 address and MAC bindings to a local CE within a subnet, as well as to send local CE IPv6 address and MAC bindings to the PE. Procedures related to usage of this PDU are discussed in subsequent sections.

The encapsulation list contained in this PDU MUST follow full replace semantics as in the LSoE protocol specification.

3.3 Overlay IPv4 Prefix Encapsulation PDU

A new encapsulation PDU type is defined for the purpose of carrying overlay IPv4 prefix routes for prefixes behind a CE that does not run a dynamic routing protocol for use-case as defined in section 4.1 of [EVPN-PREFIX-ADV]:

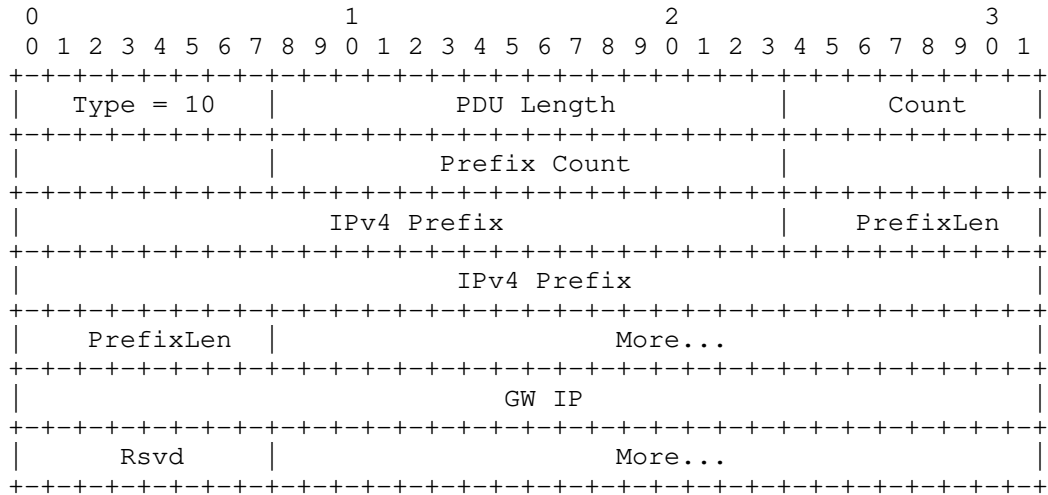


Figure 4

A CE device as defined in [EVPN-PREFIX-ADV], with prefixes behind it MAY use the above PDU to send these prefixes to an EVPN PE with itself as the GW. An EVPN PE MAY then advertise prefixes received via this PDU as RT-5, with TS as the GW, as defined in [EVPN-PREFIX-ADV].

- o A new LSoE PDU type (10) is requested for this PDU.
- o IPv4 Prefix is set to a prefix behind a CE.
- o PrefixLen is set to IPv4 prefix length for the advertised prefix.
- o GW-IP is set to the CE IPv4 address (advertised via Type 8 PDU).

Multiple prefixes may be set for a single GW IP. The encapsulation list contained in this PDU MUST follow full replace semantics as in the LSoE protocol specification.

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	2	3	4	5	6	7	8	9
Type = 10										PDU Length										Count																			
										Prefix Count																													
IPv6 Prefix																																							
																														PrefixLen									
IPv4 Prefix																																							
PrefixLen										more...																													
Rsvd										more...																													

A CE device as defined in [EVPN-PREFIX-ADV], with prefixes behind it

MAY use the above PDU to send these prefixes to an EVPN PE with itself as the GW. An EVPN PE MAY then advertise prefixes received via this PDU as RT-5, with TS as the GW, as defined in [EVPN-PREFIX-ADV].

- o A new LSoE PDU type (11) is requested for this PDU.
- o IPv6 Prefix is set to an IPv6 prefix behind a CE.
- o PrefixLen is set to IPv6 prefix length for the advertised prefix.
- o GW-IP is set to the CE IPv6 address (advertised via Type 9 PDU).

Multiple prefixes may be set for a single GW IP. The encapsulation list contained in this PDU MUST follow full replace semantics as in the LSoE protocol specification.

4. CE MAC/IP Learning on a PE AC

This section defines procedures for learning a connected CE MAC and IP on a PE local attachment circuit (AC).

4.1 PE <-> CE LSoE Session Establishment

On an EVPN PE,

- o A HELLO and/or OPEN PDU sent from a CE host source MAC is received on a tagged or untagged interface that is member of a local BD, referred here to as an AC.
- o OPEN messages are exchanged with the host on the AC.
- o LSoE session is established to the host source MAC and bound to a local AC.

4.2 CE MAC/IP Learning

Overlay IPv4 and IPv6 encapsulation PDU types 8/9 from a CE are used for the purpose of CE MAC/IP learning on a PE:

- o The EVPN flag 'E' MUST NOT be set in type 8/9 PDU from a CE.
- o A MAC entry for the MAC received in a type 8/9 PDU MUST be installed in the MAC-VRF table pointing to the AC to which the session is bound.
- o If an IPv4/IPv6 address is set in the PDU, an IPv4/IPv6 neighbor binding MUST be established for the IPv4/IPv6 address in the PDU to the MAC address in the PDU. In other words, a next-hop re-write for these IPv4/IPv6 neighbor entries MUST be installed using the MAC address in the PDU, and if required by forwarding logic, bound to the AC associated with the LSoE session.
- o Note that an IPv4/IPv6 address MAY NOT be set in a type 8/9 PDU received from a CE, in which case this PDU is only used for MAC learning. This MAY be the case in a non-IRB EVPN network, wherein, an EVPN PE is not a first-hop router for the attached CEs.

5. PE Any-cast GW MAC/IP Learning on CE

If LSoE based host learning is enabled on a PE with a distributed any-cast gateway on the EVPN PE,

- o EVPN PE MUST send type 8/9 Overlay Encapsulation PDUs on associated ACs with LSoE sessions toward CE hosts.
- o Type 8/9 PDUs from an EVPN PE MUST be encoded with the any-cast gateway IPv4/IPv6 address and any-cast gateway MAC address.
- o EVPN flag 'E' MUST NOT be set in this PDU.
- o A CE MAY process type 8/9 PDUs to establish GW IP to MAC bindings and learn gateway MAC to LAG AC bindings, similar to handling of type 8/9 PDUs on the PE described above.

Handling of type 8/9 PDUs for the purpose of gateway learning on the host is desirable but optional. A CE MAY continue to use ARP and ND for this purpose.

6. Remote CE MAC/IP Learning on CE

For CE to CE intra-subnet flows across the overlay, CE needs to learn and install a neighbor IP to MAC binding for remote CEs. This is handled today either by flooding ARP/ND requests across the overlay bridge and optionally implementing an ARP/ND suppression cache on the PE that is populated via MAC+IP EVPN route-type 2. ARP/ND request frames are trapped on the PE that does a local ARP/ND reply on behalf of the remote CE. If LSoE based learning is enabled in the fabric, LSoE may be used for this purpose to avoid overlay ARP/ND flooding, data frame triggered ARP learning, and to avoid maintaining an ARP suppression cache on the PE.

- o Remote MAC-IP routes learned via BGP EVPN route-type 2 that are imported to a local MAC-VRF MAY also be sent as type 8/9 PDUs on LSoE sessions to CEs over local ACs in that BD.
- o EVPN flag 'E' MUST be set in this encapsulation in the PDU.
- o A CE MAY install IPv4/IPv6 neighbor MAC bindings for remote CEs within a subnet based on 'E' flagged type 8/9 PDUs received from the PE.

Handling of type 8/9 PDUs for this purpose is optional but desirable to get full benefit of a fabric that is completely setup on boot-up, avoids overlay flooding, and is decoupled from latencies associated with data plane driven ARP and ND learning.

7. PE <-> CE Control Plane with EVPN All-active Multi-Homing

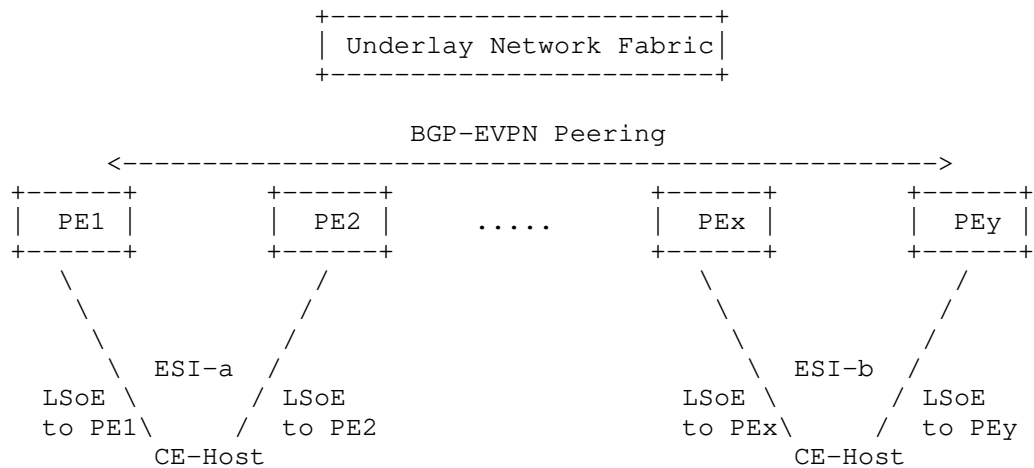


Figure 6

In an EVPN all-active multi-homing setup, a LAG interface on the CE includes member physical ports that connect to multiple PE devices. A subset of these member ports that terminate at a PE are configured as members of a local LAG interface at that PE. A LAG AC at the PE is a logical interface in a BD, identified by this LAG interface and optionally, an Ethernet Tag in case of trunk ports.

In order for LSoE based learning to work with EVPN all-active multi-homing, a separate LSoE peering MUST be established between the CE host and each PE device. For this reason, while an EVPN PE MAY form an LSoE peering to a CE host on its local LAG AC, the CE host MUST form an LSoE peering to a PE on a local LAG "member physical port".

A configurable All-active Multi-Homing mode is defined below in order to be able to bind an LSoE peering to a LAG member-port as opposed to a LAG interface.

7.1 All-active Multi-Homing Mode

When configured to run on a local LAG port in this mode,

- o LSoE HELLO messages MUST be replicated on ALL LAG member ports.
- o An LSoE OPEN message sent in response to a HELLO MUST be sent on the LAG member port on which the HELLO was received.
- o An LSoE session MUST be bound to the local LAG member port on

which the OPEN message was received.

- o LSoE encapsulation PDUs MUST be sent on the local LAG member port on which the session was bound.
- o LSoE Keep-Alives MUST be sent on the local LAG member port on which the session was bound.

Note that this may result in a PE receiving multiple HELLO PDUs from a CE MAC. This however is harmless, as per the [LSOE] specification. A PE simply drops redundant HELLOs from a MAC that it has already replied to with an OPEN, within a retry time window.

7.2 Source MAC

LSoE relies on the source MAC address in the Ethernet frame to establish a peering. When running LSoE on a LAG port (in all-active multi-homing mode or regular mode), LSoE frames MUST use the LAG interface MAC as the source MAC address in the Ethernet frame.

7.3 CE MAC/IP Learning with EVPN All-active Multi-Homing

In order to accomplish MAC/IP learning of CE host devices multi-homed to EVPN fabric PEs via EVPN All-active Multi-Homing:

- o A multi-homed CE device MUST be configured to run LSoE on a local LAG interfaces in All-active Multi-Homing mode defined above.
- o EVPN PE MAY run LSoE on local LAG interfaces to multi-homed CE devices in regular mode.
- o EVPN PEs that share the same Ethernet Segment MUST use unique source MACs (that of the local LAG) in HELLO/OPEN messages to establish separate LSoE sessions to a CE.

With the above rules in place,

- o An LSoE session on the CE is bound to a local LAG member-port.
- o An LSoE session on the PE is bound to a local LAG AC port.
- o A single LSoE session is established at the PE to a CE on the local LAG AC.
- o 'N' LSoE sessions are established at the CE, one to each PE on a local LAG member interface, where N = number of multi-homing PEs in an Ethernet Segment.

Once an LSoE session is established as above, all other host learning procedures defined earlier for CE MAC/IP learning on a PE's AC port apply as is to a LAG AC in an EVPN all-active multi-homing setup.

7.4 LAG Member Link Failure

On a CE that is running in all-active multi-homing mode, an LSoE session to a PE is bound to a LAG member interface. If the link that the LSoE session is bound to fails, LSoE session will get torn down at the CE by virtue of the session interface going down. If the CE has additional active member link(s) to this PE, a new LSoE session must be established on one of the active member links via HELLO PDUs sent by the CE on its remaining active member links to the PE.

7.4.1 Session Re-establishment

LSoE session at the CE is torn down immediately following the session interface failure. While the LAG interface at the PE is still operationally UP, LSoE session at the PE is subject to Keep Alive PDUs received from the CE. Once the session expires at the PE because of missed Keep Alive PDUs from the CE, PE will respond to HELLO on one of the active member link with an OPEN to re-establish a new session. Note that the new session is still bound to the LAG AC at the PE and to a new member link at the CE.

7.4.2 TLV Retention

TLVs learnt from a CE over a failed session MUST be retained at the PE if the PE LAG AC is still operationally up following a member link failure because of active member link(s) in the LAG. TLV retention logic at the PE MAY be based on an age-out time, that is a local matter at the PE. TLV age-out time MUST be higher than the missed Keep Alive duration, after which the session is considered closed. Once a new LSoE session is established, PE MUST implement a mark and sweep logic to reconcile retained TLVs from the CE peer with the new set of TLVs received from this CE.

7.4 LAG Failure

When a LAG member link failure results in the LAG interface being operationally down, TLV age-out logic discussed above MUST NOT be in effect. LSoE session MAY be considered as DOWN immediately on the LAG being down at the PE. This is so that, in the event of a total connectivity loss between a PE and CE, CE learnt routes can be withdrawn immediately.

7.5 Example PE <-> CE Control Plane Flow with All-active Multi-Homing

An example LSoE over all-active multi-homing session flow is discussed below for clarity.

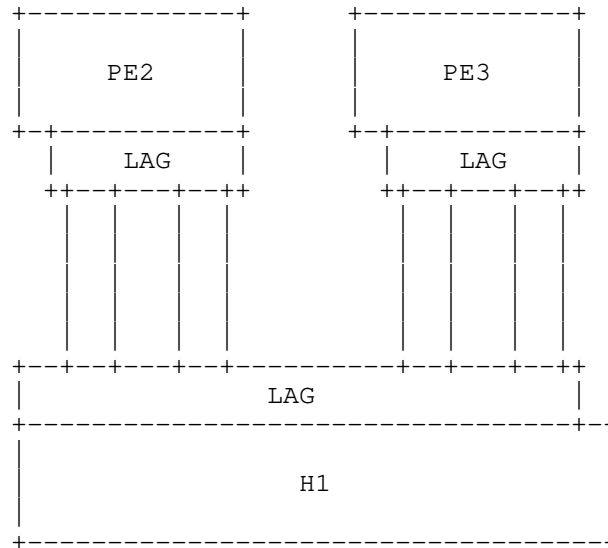


Figure 7

Example topology with CE H1 multi-homed to PE2 and PE3 via EVPN all-active multi-homing LAG with four member ports to each PE:

H1 member ports to PE2: i121, i122, i123, i124

PE2 member ports to H1: i211, i212, i213, i214

H1 member ports to PE3: i131, i132, i133, i134

PE3 member ports to H1: i311, i312, i313, i314

H1 LAG port to PE2/PE3: MLAG1

PE2 LAG port to H1: LAG2

PE3 LAG port to H1: LAG3

H1 LAG MAC: LMAC1

PE2 LAG MAC: LMAC2

PE3 LAG MAC: LMAC3

H1 running LSoE on MLAG1 in All-active Multi-Homing mode

PE2 running LSoE on LAG2 in regular mode

PE3 running LSoE on LAG3 in regular mode

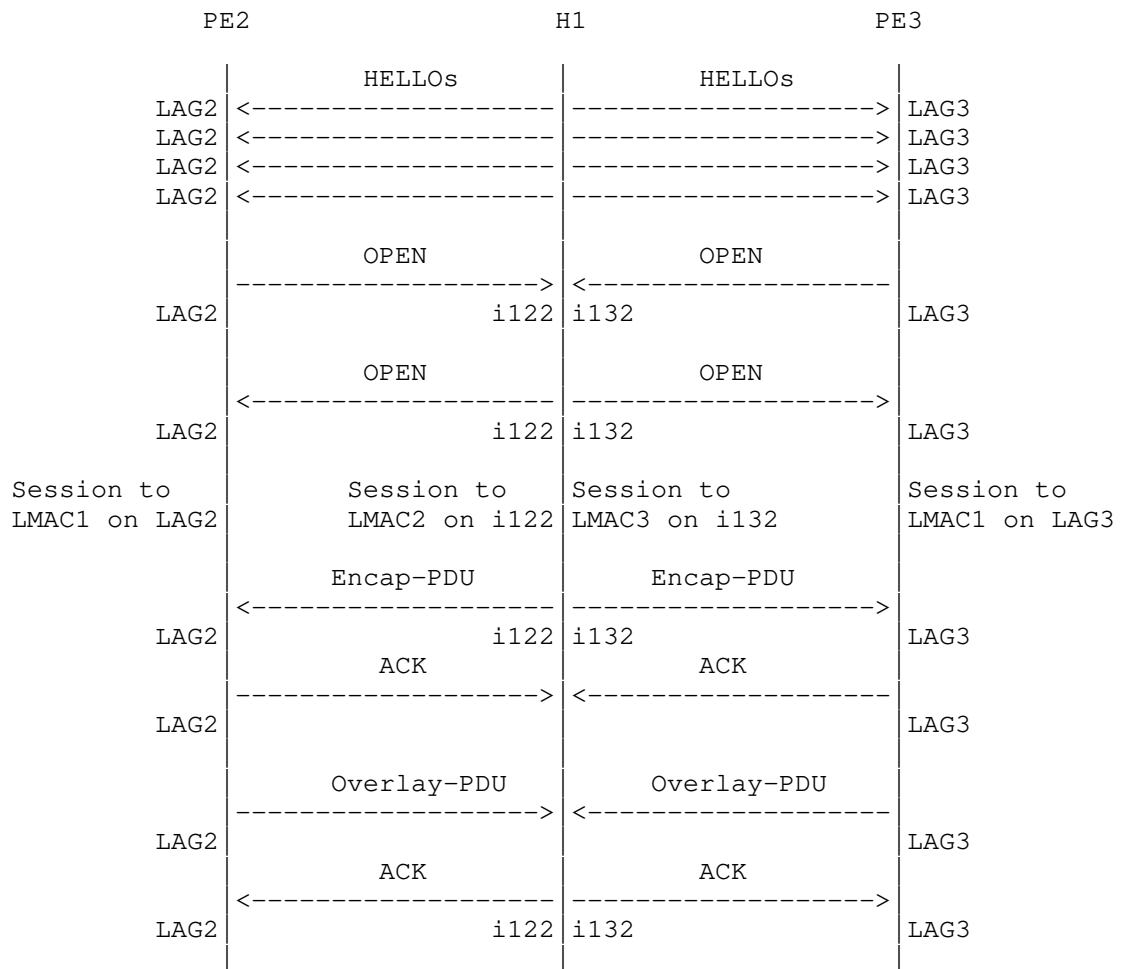


Figure 8

In an example flow shown above:

- o H1: originates HELLO(SMAC=LMAC2) on all MLAG member ports
- o PE2: Multiple HELLO(SMAC=LMAC2) copies received on port LAG2
- o PE3: Multiple HELLO(SMAC=LMAC2) copies received on port LAG3
- o PE2: A single OPEN(SMAC=LMAC2, DMAC=LMAC1) sent on port LAG2
- o PE3: A single OPEN(SMAC=LMAC3, DMAC=LMAC1) sent on port LAG3
- o PE2/PE3: duplicate HELLOs from same source LMAC2 are ignored
- o H1: OPEN(SMAC=LMAC2, DMAC=LMAC1) received on member port i122
- o H1: OPEN(SMAC=LMAC1, DMAC=LMAC2) sent on member port i122
- o H1: Session established to LMAC2 on MLAG1 member port i122

- o PE2: Session established to LMAC1 on LAG AC LAG2
- o H1: OPEN(SMAC=LMAC3, DMAC=LMAC1) received on member port i132
- o H1: OPEN(SMAC=LMAC1, DMAC=LMAC3) sent on member port i132
- o H1: Session established to LMAC3 on MLAG member port i132
- o PE3: Session established to LMAC1 on LAG AC LAG3
- o H1: IP encapsulation PDUs (type 4/5) sent to LMAC2 and LMAC3
- o PE2/PE3: H1 MAC and IP are learned
- o PE2/PE3: overlay IP encapsulation PDUs (type 8/9) sent to LMAC1
- o H1: Any-cast GW MAC and IP are learned
- o H1: Remote host MAC and IP are learned

8. Software Neighbor Tables

Some networking stack implementations rely on ARP and ND populated neighbor tables for software forwarding. In order to inter-work with such an implementation, an LSoE learned IPv4/IPv6 neighbor entry MAY also be installed in ARP and ND neighbor table as a static / permanent entry.

In addition,

- o Pre-installing LSoE learned neighbor entries may help reduce potential conflict with ARP or ND learned neighbor entries.
- o Pre-installing LSoE learned neighbor entries may help reduce reliance on data traffic triggered ARP requests / ND solicitations and associated learning latency.

With respect to installing IPv6 entries learnt via LSoE in IPv6 ND cache, Router flag (R-bit) and Override flag (O-bit) received in LSoE PDU should be handled as defined in [RFC4861].

9. MAC/IP Learning Conflict Resolution

If LSoE learned neighbor entries are not already installed as static entries in ARP/ND neighbor table, it is possible that a neighbor IPv4/IPv6 adjacency may be learned both via LSoE and ARP/ND. Even if LSoE learned entries were pre-installed in neighbor table, a race condition is still possible leading to a potential conflict between ARP/ND learned and LSoE learned neighbor IP adjacency. In such scenarios, LSoE learned entry should be preferred for the purpose of programming neighbor IP adjacencies in forwarding.

With respect to MAC-VRF entries, it is recommended that data plane learning be turned off when LSoE based learning is enabled. However, if it is not, data plane learned entries MUST be reconciled with LSoE learned entries in software and, in case of a conflict, LSoE learned entries preferred if LSoE based learning is enabled.

10. PE-CE Overlay Prefix Learning

[EVPN-PREFIX-ADV] section 4.1 defines a use case, wherein, a PE may advertise IP prefixes and subnets behind a CE. In this use case, CE device does not run a dynamic routing protocol. Instead, these prefixes are learnt on the PE via local policy or configuration. Prefixes are then advertised by PE as RT-5 with the CE as the GW.

PE-CE control plane defined in this document MAY be used to learn these prefixes from a CE as an alternative to local configuration on the PE. Once an LSoE session is established between a CE and a PE, as discussed earlier,

- o A CE MAY send type 10/11 PDUs with these IPv4/IPv6 prefixes over an LSoE session to a PE with the CE IP as the GW IP.
- o A PE MAY advertise prefixes learnt via type 10/11 PDUs as RT-5 with CE IP as the GW IP.

To summarize, A PE would advertise:

- o RT-2 for the CE MAC-IP learnt via type 8/9 PDU
- o RT-5 for Prefixes learnt via type 10/11 PDU with GW IP = CE IP

11. Asymmetric EVPN-IRB

Any deviations from the above procedures proposed in this document for asymmetric IRB design will be covered in subsequent updates to this document.

12. Centralized Gateway EVPN-IRB

Any deviations from the above procedures proposed in this document for centralized GW based IRB design will be covered in subsequent updates to this document.

13. Use Cases

13.1 Simplified EVPN Operations

This section will discuss in detail, benefits and simplifications that may be achieved in the context of an EVPN network, if one chooses to implement PE-CE control plane defined in this document as opposed to using traditional data-plane and ARP/ND snooping based PE-CE learning.

13.1.1.1 EVPN All-active Multi-Homing

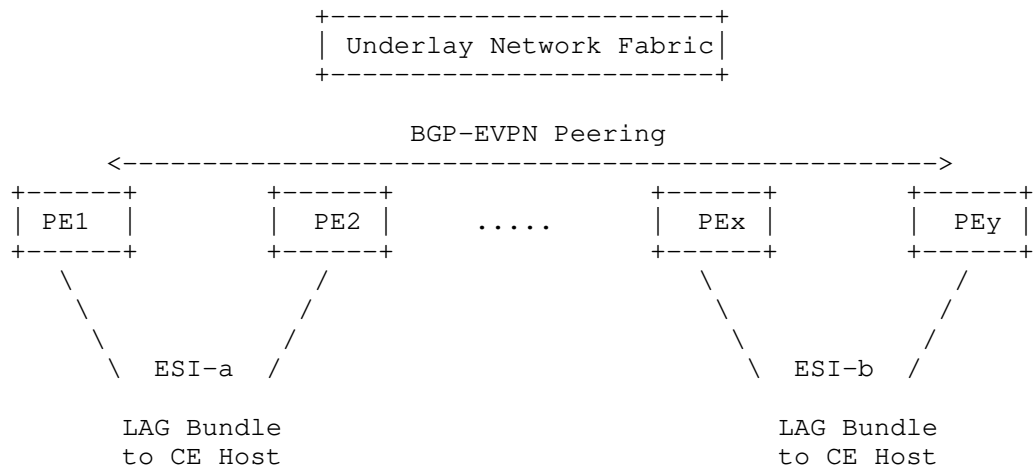


Figure 9

Data plane and ARP/ND snooping based MAC/IP learning on PE-CE all-active multi-homed LAG ports is subject to unpredictable hashing of ARP, ND, and data frames from host to PE. As an example, an ARP request for a connected host might originate at PE1 but the resulting ARP response from the host might be received at PE2. Redundant EVPN PEs in all-active multi-homing mode typically handle this unpredictability via combination of methods below:

- o PEs can handle unsolicited ARP and ND response frames.
- o PEs can implement additional mechanism to SYNC ARP, ND, and MAC tables across all PEs in a redundancy group for optimal forwarding to locally connected hosts.
- o PEs can implement EVPN aliasing procedures discussed in [RFC 7432] OR re-originate SYNCed MAC-IP adjacencies as local RT-2 to achieve MAC ECMP across the overlay.
- o PEs can also re-originate SYNCed MAC-IP adjacencies as local RT-2 to achieve IP ECMP across the overlay OR implement IP aliasing procedures discussed in [EVPN-IP-ALIASING].
- o PEs can also ensure EVPN sequence number SYNC for local MAC entries for EVPN mobility procedures to work correctly, as discussed in [EVPN-IRB-MOBILITY].

The PE-CE control plane learning alternative defined in this document fully decouples MAC and IP learning over MLAG ports from unpredictable hashing of data, AR, ND frames on all-active multi-

homed LAG member links. As a result, above procedures that essentially result from data-plane PE-CE learning on all-active multi-homed LAGs can be simplified via the PE-CE control plane alternative defined in this document.

13.1.2 Convergence on CE Host Moves

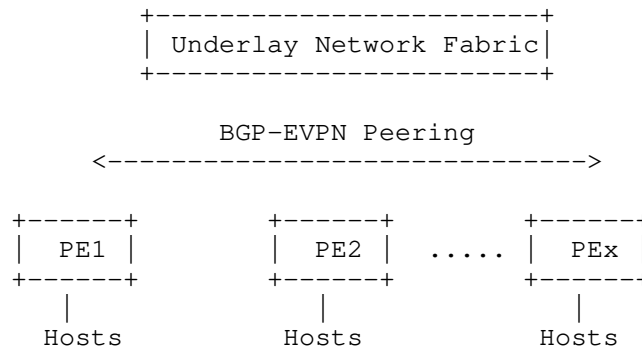


Figure 10

Host mobility across EVPN PE switches is a common occurrence in a data center fabric for flexibility in work load placement across a DC. Further, a host move must result in minimal, if any, disruption to traffic flows / services to / from the device.

Data plane and ARP/ND snooping based PE-CE learning may result in unpredictable convergence times, following host moves for the following cases:

- o A host may or may not send any data packet immediately following a move.
- o A host may or may not send an unsolicited ARP following a move.

While probing procedures, discussed in the next sub-sections are typically used to minimize convergence time, certain scenarios discussed below may still result in extended convergence times and flooding.

13.1.2.1 Silent Hosts

If a host is silent for an extended period following a move from PE1 to PE2, any bridged traffic flow destined to this host will continue to be black-holed by PE1 until the MAC ages out at PE1. Once the the MAC ages out at PE1, any bridged traffic flow destined to the host is

flooded across the overlay bridge. Flooding of unknown unicast traffic on the overlay is enabled for this purpose. In summary, PE-CE learning that is based on data-plane and AR/ND snooping may be subject to non-deterministic convergence time and flooding following host moves because of being heavily dependent on unpredictable CE behavior.

PE-CE control plane based learning defined in this document fully decouples convergence in such scenarios from non-deterministic data flows and unsolicited ARP/ND behavior on a CE.

13.1.2.2 Probing

ARP and ND probing procedures are typically used to achieve host re-learning and convergence following host moves across the overlay:

- o Following a host move from PE1 to PE2, the host's MAC is discovered at PE2 as a local MAC via a data frames received from the host. If PE2 has a prior REMOTE MAC-IP host route for this MAC from PE1, an ARP probe is typically triggered at PE2 to learn the MAC-IP as a local IP adjacency and triggers EVPN RT-2 advertisement for this MAC-IP across the overlay with new reachability via PE2.
- o Following a host move from PE1 to PE2, once PE1 receives a MAC or MAC-IP route from PE2 with a higher sequence number, an ARP probe is triggered at PE1 to clear the stale local MAC-IP neighbor adjacency OR re-learn the local MAC-IP in case the host has moved back or is duplicate.
- o Following a local MAC age-out, if there is a local IP adjacency with this MAC, an ARP probe is triggered for this IP to either re-learn the local MAC and maintain local l3 and l2 reachability to this host OR to clear the ARP entry in case the host is indeed no longer local. Note that clearing of stale ARP entries, following a move is required for traffic to converge in the event that the host was silent and not discovered at its new location. Once stale ARP entry for the host is cleared, routed traffic flow destined for the host can re-trigger ARP discovery for this host at the new location. ARP flooding on the overlay MUST also be done to enable ARP discovery via routed flows.
- o Alternatively, ARP probing timer may be tuned to be smaller than the MAC aging timer to avoid MAC age-out.

PE-CE control plane learning alternative defined in this document decouples host learning following moves from unpredictable host behavior with respect to sending data traffic and unsolicited ARPs,

and as a result from ARP probing and MAC aging timer settings. Host move handling is hence greatly simplified to a very predictable and deterministic behavior.

13.1.3 ARP Gleaning Latency

If a CE's ARP binding is not already learned on a PE via an unsolicited ARP sent by the CE following events such as boot-up, flaps, and moves, a data frame that needs to be routed to the CE triggers ARP or ND discovery process on the PE. On a typical hardware switching platform, an IP packet that does not resolve to a link layer re-write would be punted to host stack that delivers packets with incomplete link-layer resolution to ARP or ND for resolution. An ARP request / ND Solicitation is generated for the CE IP and an ARP response or NA results in installing a link-layer re-write for the CE IP. In an EVPN multi-homing environment, this procedure is further complicated as the response is only received by one of the PEs that may or may not be the one that generated the ARP or ND request. Learned neighbor binding is SYNCed to other PEs that share the multi-homed Ethernet Segment. Routed flows can now be forwarded to the host via all PEs. Latency associated with such data frame driven ARP discovery may result in significant initial convergence hit, following triggers that warrant re-gleaning of CE IP to MAC binding.

PE-CE control plane learning alternative defined in this document results in proactive host learning following these scenarios, potentially avoiding a convergence hit on initial data packets.

13.2 Applicability to non-EVPN Use Cases

While the LSoE based host learning procedure described in this document focuses on EVPN-IRB overlay fabric use case, it may also have benefits and applicability in non-EVPN use cases. Applicability of procedures described in this document to non-EVPN use cases is a topic for further study.

14. Summary

PE-CE control plane is proposed as an alternative to data plane and ARP/ND snooping based PE-CE host MAC/IP learning and for PE-CE prefix learning. With a PE-CE control plane, CE host MAC and IP are deterministically learned on host boot-up, on host configuration, across host moves, on convergence triggers such as link failures, flaps, and PE re-boots and on all-active multi-homing LAG links. A PE-CE control plane decouples CE MAC and IP learning from traffic flows sourced by a CE, from varying CE behavior with respect to sending unsolicited ARP/ND frames, and from hashing of CE sourced frames over all-active multi-homed LAG links. As a result, it helps

achieve a predictable and reliable convergence behavior across these triggers and helps simplify certain EVPN procedures that are otherwise needed with a data-plane and ARP/ND snooping based PE-CE learning. In addition, it may also be used for non-host learning use cases such as prefix learning.

15. References

15.1 Normative References

- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<http://www.rfc-editor.org/info/rfc7432>>.
- [LSOE] Bush, R., Austein R., Patel, K., "Link State Over Ethernet", Feb 2019, <<https://tools.ietf.org/html/draft-ietf-lsvr-lsoe-01>>.
- [EVPN-IRB] Sajassi, A., Salem, S., Thoria S., Drake J., Rabadan J., "Integrated Routing and Bridging in EVPN", July 2018, <<https://tools.ietf.org/html/draft-ietf-bess-evpn-inter-subnet-forwarding-05>>.
- [EVPN-PREFIX-ADV] Rabadan J., Henderickx W., Drake J., Lin W., Sajassi, A., "IP Prefix Advertisement in EVPN", May 2018, <<https://tools.ietf.org/html/draft-ietf-bess-evpn-prefix-advertisement-11>>.
- [EVPN-IRB-MOBILITY] Malhotra, N., Sajassi, A., Rabadan, J., Drake J., Lingala A., Patekar A., "Extended Mobility Procedures for EVPN-IRB", Jan 2019, <<https://tools.ietf.org/html/draft-malhotra-bess-evpn-irb-extended-mobility-04>>.
- [EVPN-IP-ALIASING] Sajassi, A., Badoni, G., "L3 Aliasing and Mass Withdrawal Support for EVPN", July 2017, <<https://tools.ietf.org/html/draft-sajassi-bess-evpn-ip-aliasing-00>>.
- [RFC2119] S. Bradner, "Key words for use in RFCs to Indicate Requirement Levels", March 1997, <<https://tools.ietf.org/html/rfc2119>>.
- [RFC8174] B. Leiba, "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", May 2017, <<https://tools.ietf.org/html/rfc8174>>.

15.2 Informative References

16. Acknowledgements

Authors would like to thank Randy Bush and Rob Austein for detailed review and feedback to ensure consistency with base LSOE protocol specification, as well as for helping build detailed LSOE flows included in this document.

Authors would like to thank Ali Sajassi and John Drake for detailed review and very valuable input on PE-CE protocol design for EVPN use cases as well as structuring this document for EVPN use cases.

Contributors

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

Email: randy@psg.com

Authors' Addresses

Neeraj Malhotra (Editor)
Arrcus
2077 Gateway Place, Suite #400
San Jose, CA 95119, USA

Email: neeraj.ietf@gmail.com

Keyur Patel
Arrcus
2077 Gateway Place, Suite #400
San Jose, CA 95119, USA

Email: keyur@arrcus.com

Jorge Rabadan
Nokia
777 E. Middlefield Road
Mountain View, CA 94043, USA

Email: jorge.rabadan@nokia.com

