

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 25 April 2022

H. Chen
Futurewei
G. Mishra
Verizon Inc.
A. Wang
China Telecom
Y. Liu
China Mobile
H. Wang
Huawei
Y. Fan
Casa Systems
22 October 2021

BGP-SPF Flooding Reduction
draft-chen-lsvr-flood-reduction-00

Abstract

This document describes extensions to Border Gateway Protocol (BGP) for flooding the link states on a topology that is a subgraph of the complete topology of a BGP-SPF domain, so that the amount of flooding traffic in the domain is greatly reduced. This would reduce convergence time with a more stable and optimized routing environment.

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on 25 April 2022.

Copyright Notice

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

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

Table of Contents

1. Introduction	3
2. Terminologies	3
3. Overview of BGP-SPF Link State Flooding	3
3.1. Flooding in RR Model	4
3.2. Flooding in Node Connections Model	5
3.3. Flooding in Directly-Connected Nodes Model	6
4. Revised Flooding Procedures	6
4.1. Revised Flooding Procedure for RR Model	6
4.2. Revised Flooding Procedure for Node Connections Model	8
5. BGP Extensions for Flooding Reduction	10
5.1. Extensions for RR Model	10
5.2. Extensions for Node Connections Model	12
5.2.1. New TLVs	12
5.2.2. Flooding Topology Distribution in Centralized Mode	16
5.2.3. An Algorithm for Distributed Mode	17
6. Security Considerations	18
7. Acknowledgements	19
8. IANA Considerations	19
9. References	19
9.1. Normative References	19
9.2. Informative References	19
Authors' Addresses	20

1. Introduction

For some networks such as dense Data Center (DC) networks with BGP-SPF, the existing Link State (LS) flooding mechanism defined in [I-D.ietf-lsvr-bgp-spf] for a BGP-SPF domain may not be efficient and may have some issues. The extra LS flooding consumes network bandwidth. Processing the extra LS flooding, including receiving, buffering and decoding the extra LSs, wastes memory space and processor time. This may cause scalability issues and affect the network convergence negatively.

This document describes extensions to Border Gateway Protocol (BGP) for flooding the link states on a topology that is a subgraph of the complete topology of a BGP-SPF domain, so that the amount of flooding traffic in the domain is greatly reduced.

2. Terminologies

The following terminologies are used in this document.

BGP: Border Gateway Protocol

LS: Link State

SPF: Shortest Path First

RR: Route Reflector

3. Overview of BGP-SPF Link State Flooding

[I-D.ietf-lsvr-bgp-spf] defines three BGP peering models:

- * BGP Peering in Route-Reflector or Controller Topology (RR or Sparse model for short).
- * BGP Single-Hop Peering on Network Node Connections (Node Connections model for short), and
- * BGP Peering Between Directly-Connected Nodes (Directly-Connected Nodes model for short).

This section briefs the BGP-SPF Link State Flooding in each of these models.

3.1. Flooding in RR Model

In RR model, BGP-SPF speakers/nodes peer solely with one or more Route Reflectors (RRs) or controllers. A BGP-SPF speaker sends/advertises its BGP-LS-SPF Link NLRI in a BGP update message to the RRs or controllers that the speaker peers with when it discovers that its corresponding link is up. After receiving the Link NLRI, each of the RRs or controllers sends the NLRI in a BGP update message to the other BGP-SPF speakers that peer with the RRs or controllers.

For example, Figure 1 shows a BGP-SPF domain, which contains two RRs RR1 and RR2, and three network nodes A, B and C. RR1 peers with all three nodes A, B and C in the network. RR2 also peers with all three nodes A, B and C in the network. There is a link between A and B, a link between A and C, and a link between B and C.

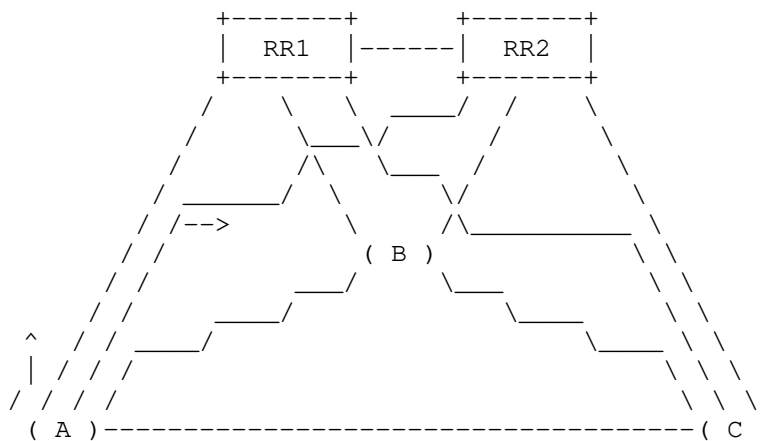


Figure 1: BGP-SPF Domain with two RRs

Each of the nodes A, B and C in the network sends/advertises its link NLRI in BGP update messages to both RR1 and RR2. After receiving a link NLRI in a BGP update message from a node (e.g., node A), each of RR1 and RR2 sends the NLRI in a BGP update message to the other nodes (e.g., nodes B and C). Each of the other nodes receives two copies of the same NLRI, one from RR1 and the other from RR2. One copy is enough, the other redundant copy should be reduced.

3.2. Flooding in Node Connections Model

In Node Connections model, EBGp single-hop sessions are established over direct point-to-point links interconnecting the nodes in the BGP-SPF routing domain. Once the session has been established and the BGP-LS-SPF AFI/SAFI capability has been exchanged 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 to all the nodes in the domain through all the BGP sessions over the links. If the session goes down, the corresponding Link NLRI will be withdrawn. The withdrawal is done through advertising a BGP update containing the NLRI in MP_UNREACH_NLRI to all the nodes in the domain using all BGP sessions over the links.

For example, Figure 2 shows a BGP-SPF domain, which contains four nodes A, B, C and D. These four nodes are connected by six links. There are two parallel links between A and B, a link between A and C, a link between A and D, a link between B and C and a link between C and D.

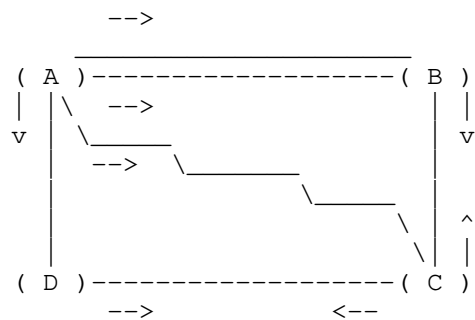


Figure 2: BGP-SPF Domain with parallel links

Suppose that the BGP sessions over all the links except for the session over the link between A and D have been established and the BGP-LS-SPF AFI/SAFI capability has been exchanged for the corresponding sessions. When the BGP session over the link between A and D is established and the BGP-LS-SPF AFI/SAFI capability is exchanged for the corresponding session, node A considers that the link from A to D is up and sends the BGP-LS-SPF Link NLRI for the link through its four BGP sessions (i.e., the session between A and B over the first parallel link between A and B, the session between A and B over the second parallel link between A and B, the session between A and C over the link between A and C, and the session between A and D over the link between A and D) to nodes B, C and D. After receiving the NLRI from node A, each of the nodes B, C and D

sends the NLRI to the other nodes that have BGP sessions with the node. Node B sends the NLRI to node C. Node C sends the NLRI to nodes B and D. Node D sends the NLRI to node C.

Similarly, when the BGP session over the link between A and D is established and the BGP-LS-SPF AFI/SAFI capability is exchanged for the corresponding session, node D considers that the link from D to A is up and sends the BGP-LS-SPF Link NLRI for the link through its two BGP sessions (i.e., the session between D and C over the link between D and C, and the session between D and A over the link between D and A) to nodes C and A. After receiving the NLRI from node D, each of the nodes A and C sends the NLRI to the other nodes that have BGP sessions with the node. Node C sends the NLRI to nodes A and B. Node A sends the NLRI to nodes B and C through two parallel BGP sessions to B and the BGP session to C.

3.3. Flooding in Directly-Connected Nodes Model

In Directly-Connected Nodes model, BGP-SPF speakers peer with all directly-connected nodes but the sessions may be between loopback addresses. 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. Since there are BGP sessions between every directly-connected nodes in the BGP-SPF routing domain, there is only a reduction in BGP sessions when there are parallel links between nodes comparing to node connections model.

4. Revised Flooding Procedures

4.1. Revised Flooding Procedure for RR Model

In RR model, the revised flooding procedure is as follows:

- * A BGP-SPF speaker/node sends its BGP-LS-SPF Link NLRI to some such as one of the RRs or controllers that the speaker peers with when it discovers that its corresponding link is up.
- * After receiving the Link NLRI, the RR or controller sends the NLRI to the other BGP-SPF speakers that peer with the RR or controller.

For example, for the BGP-SPF domain in Figure 1, using the revised flooding procedure, speaker/Node A sends its Link NLRI for link A to B to one RR RR1 when A discovers that link A to B is up. Node A does not send the NLRI to RR2. After receiving the Link NLRI for link A to B from speaker/node A, RR1 sends the NLRI to the other nodes B and C. Each of the other nodes receives only one copy of the same NLRI,

which is from RR1. There is no redundant copy of the same NLRI. Comparing to the normal flooding in RR model as illustrated in Figure 1, the revised flooding procedure reduces the amount of link states flooding by half.

In an option, for a number of RRs or controllers that peer with all the nodes/speakers in a network, the nodes are evenly divided into the number of groups. A first group of nodes send their link NLRIs to a first RR; a second group of nodes send their link NLRIs to a second RR; and so on. After receiving a NLRI from a node, a RR sends the NLRI to the other nodes in the network. This option may be used if each node peers with every RR or controller; otherwise, it should not be used.

In one implementation, the nodes (supposing there are m nodes in total) are divided into N groups through ordering the nodes by their IDs in ascending order and grouping the nodes. Each of the N groups has m/N nodes. The first m/N nodes in the ordered nodes are in the first group; the m/N nodes following the first group are in the second group; the m/N nodes following the second group are in the third group; and so on. The nodes following the second last group are in the N -th group (i.e., the last group).

For example, for the BGP-SPF domain in Figure 1, there are two RRs and three nodes, the nodes in the network are evenly divided into two groups. The first group contains one ($3/2 = 1$) node: node A. The second group contains the rest nodes: nodes B and C.

Node A in the first group sends its link NLRIs to RR1. After receiving a Link NLRI from node A, RR1 sends the NLRI to the other nodes B and C in the network. Nodes B and C in the second group send their link NLRIs to RR2. After receiving a Link NLRI from node B, RR2 sends the NLRI to the other nodes A and C in the network. After receiving a Link NLRI from node C, RR2 sends the NLRI to the other nodes A and B in the network.

Each of the other nodes receives only one copy of the same NLRI, which is from RR1 or RR2. There is no redundant copy of the same NLRI.

In this option, every group of nodes has about the same number of nodes as each of the other groups, the workload is balanced among the RRs (i.e., each of RRs has almost the same workload as any other RR).

In another option, for a number of RRs or controllers that peer with all the nodes/speakers in a network, the nodes in the network sends their link NLRIs to the same one or more of the RRs.

For example, for the BGP-SPF domain in Figure 1, nodes A, B and C in the network send their link NLRI to the same RR1. After receiving the Link NLRI from a node, RR1 sends the NLRI to the other nodes in the network. For example, after receiving the Link NLRI from node A, RR1 sends the NLRI to the other nodes B and C in the network. After receiving the Link NLRI from node B, RR1 sends the NLRI to the other nodes A and C in the network. After receiving the Link NLRI from node C, RR1 sends the NLRI to the other nodes A and B in the network.

4.2. Revised Flooding Procedure for Node Connections Model

In Node Connections model, the revised flooding procedure is as follows:

- * A BGP-SPF speaker/node has a flooding topology of the BGP-SPF domain. In an option, the flooding topology is computed in a distributed mode, where every BGP-SPF speaker computes a flooding topology for the domain using a same algorithm. In another option, the flooding topology is computed in a centralized mode, where one BGP-SPF speaker elected as a leader computes a flooding topology for the domain and advertises the flooding topology to every BGP-SPF speaker in the domain.
- * A BGP-SPF speaker/node sends its link NLRI in a BGP update message for its link up or down to its peers that are directly connected on the flooding topology, and sends its link NLRI in a BGP update message for its link down to all its peers. When receiving the NLRI in a new BGP update message for a link up or down from a peer, the speaker sends the NLRI in a BGP update message to its other peers that are directly connected on the flooding topology.
- * When a BGP-SPF session is down, the BGP-SPF speaker/node that was connected to the session will not withdraw the link NLRI received from the session right away. It keeps the NLRI for some time.

Given a real network topology (RT), a flooding topology (FT) of the RT is a sub network topology of the RT and connects all the nodes in the RT.

For example, Figure 3 shows a flooding topology of the real topology in Figure 2.



Figure 3: A Flooding Topology

The flooding topology in Figure 3 is a sub network topology of the RT in Figure 2 and connects all the nodes (i.e., nodes A, B, C and D) in the RT in Figure 2.

Figure 4 shows a reduced flooding flow of a link NLRI in a BGP update message for a link up or down in the BGP-SPF domain, which is the same as the one in Figure 2.

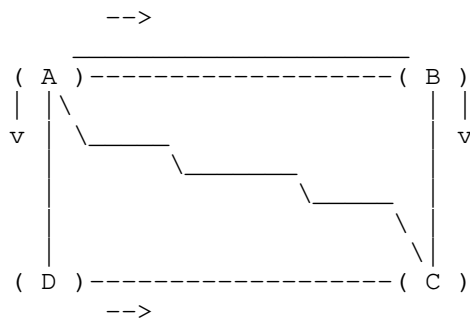


Figure 4: A Reduced Link State Flooding Flow

Speaker/Node A sends the NLRI in a BGP update message for its link to its peers B and D. Nodes B and D are peers of node A and are directly connected to A on the flooding topology (FT). Node A does not send the NLRI to its peer C since C is not directly connected to A on the FT.

After receiving the NLRI in the message from A, node B sends the NLRI in a BGP update message to B's other peer C (which is directly connected to B on the FT). After receiving the NLRI in a BGP update message from A, node D sends the NLRI in a BGP update message to D's other peer C (which is directly connected to D on the FT).

The number of NLRIs in messages flooded in Figure 4 is much less than that in Figure 2. The performance of network is improved using the revised flooding procedure.

5. BGP Extensions for Flooding Reduction

This section specifies BGP extensions for flooding reduction in two models: RR model and Node Connections model. The extensions for Directly-Connected Node model are included in the extensions for Node Connections model.

5.1. Extensions for RR Model

A single RR for a BGP-SPF domain is elected as a leader RR of the domain. The leader RR is the RR with the highest priority to become a leader in the domain. If there are more than one RRs having the same highest priority, the RR with the highest Node ID and the highest priority is the leader RR in the domain. In a deployment, only every RR advertises its priority for becoming a leader using a Leader Priority TLV defined below.

Two new TLVs are defined for flooding reduction in RR model.

- * Leader Priority TLV: A node uses it to advertise its priority for becoming a leader.
- * Node Flood TLV: A RR or controller uses it to tell every node the flooding behavior the node needs to follow.

The format of Leader Priority TLV is illustrated in Figure 5.

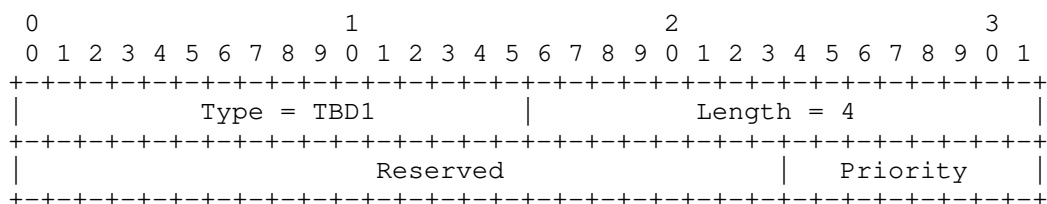


Figure 5: Leader Priority TLV

Type: It is to be assigned by IANA.

Length: 4.

Reserved: MUST be set to zero in transmission and should be ignored on reception.

Priority: A unsigned integer from 0 to 255 in one octet indicating priority to become a leader.

The format of Node Flood TLV is illustrated in Figure 6.

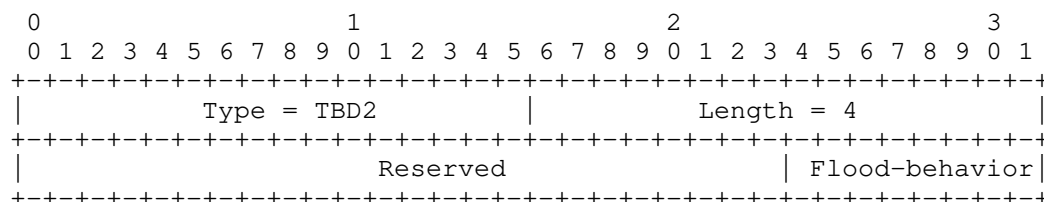


Figure 6: Node Flood TLV

Type: It is to be assigned by IANA.

Length: 4.

Reserved: MUST be set to zero in transmission and should be ignored on reception.

Flood-behavior: The following flooding behavior are defined.

- 0 - Reserved.
- 1 - send link states to the RR with the minimum ID
- 2 - send link states to the RR with the maximum ID
- 3 - balanced groups
- 4 - send link states to 2 RRs with smaller IDs
- 5 - send link states to 2 RRs with larger IDs
- 6 - balanced groups with redundancy of 2
- 7-127 - Standardized flooding behaviors for RR Model
- 128-254 - Private flooding behaviors for RR Model.

In a deployment, the flooding behavior for every node is configured on a RR or controller such as the leader RR and the RR advertises the behavior to the other RRs and every node in the network though using a Node Flood TLV.

For example, if we want every node in the network to send its link states to only one RR, we configure this behavior on a RR and the RR advertises the behavior to every node using a Node Flood TLV with Flood-behavior set to one, which tells every node to send its link states to the RR with the minimum ID. If we want every node in the network to send its link states to two RRs for redundancy, we configure this behavior on a RR and the RR advertises the behavior to

every node using a Node Flood TLV with Flood-behavior set to 4, which tells every node to send its link states to the two RRs with smaller IDs (i.e., the RR with the minimum ID and the RR with the second minimum ID).

If we want to balance the traffic among RRs or controllers through dividing the nodes into groups and letting each group send their link states to a RR, we configure this behavior on a RR and the RR advertises the behavior to every node using a Node Flood TLV with Flood-behavior set to 3, which tells every node to divide the nodes in the network into a number of groups. A node in a group sends its link states to the RR corresponding to the group.

5.2. Extensions for Node Connections Model

There are two modes for the flooding topology computation: centralized mode and distributed mode. In a centralized mode, one BGP-SPF node is elected as a leader. The leader computes a flooding topology for the BGP-SPF domain and advertises the flooding topology to every BGP-SPF node in the domain. In a distributed mode, every BGP-SPF node computes a flooding topology for the BGP-SPF domain using a same algorithm. There is not any flooding topology distribution.

This section defines the new TLVs for the two modes, describes the flooding topology distribution in centralized mode and an algorithm that can be used by every node to compute its flooding topology in distributed mode.

5.2.1. New TLVs

Five new TLVs are defined for flooding reduction in Node Connections model.

- * Node Algorithm TLV: A leader uses this TLV to tell every node the algorithm to be used to compute a flooding topology.
- * Algorithms Support TLV: A node uses this TLV to indicate the algorithms that it supports for distributed mode.
- * Node IDs TLV: A leader uses this TLV to indicate the mapping from nodes to their indices for centralized mode.
- * Paths TLV: A leader uses this TLV to advertise a part of flooding topology for centralized mode.

- * Connection Used for Flooding TLV: A node uses this TLV to indicate that a connection/link is a part of the flooding topology and used for flooding.

5.2.1.1. Node Algorithm TLV

The format of Node Algorithm TLV is illustrated in Figure 7.

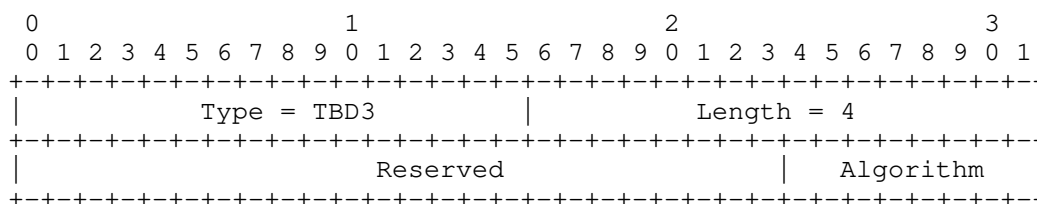


Figure 7: Node Algorithm TLV

Type: It is to be assigned by IANA.

Length: 4.

Reserved: MUST be set to zero in transmission and should be ignored on reception.

Algorithm:

- 0 - The leader computes a flooding topology using its own algorithm and advertises the flooding topology to every node.
- 1-127 - Every node computes its flooding topology using this standardized distributed algorithm.
- 128-254 - Private distributed algorithms.

A node such as the leader node can use this TLV to tell every node in the domain to use the flooding topology from the leader for flooding the link states through advertising the TLV with the Algorithm field set to zero, or to tell every node to compute its own flooding topology using the algorithm given by the Algorithm field in the TLV containing an identifier of an algorithm when the Algorithm field is not zero.

5.2.1.2. Algorithms Support TLV

The format of Algorithms Support TLV is illustrated in Figure 8.

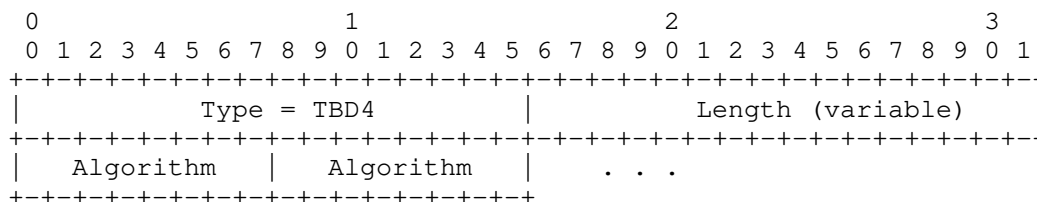


Figure 8: Algorithms Support TLV

Type: It is to be assigned by IANA.

Length: The number of Algorithms in the TLV.

Algorithm: A numeric identifier in the range 0-255 indicating the algorithm that can be used to compute the flooding topology.

5.2.1.3. Node IDs TLV

The format of Node IDs TLV is illustrated in Figure 9.

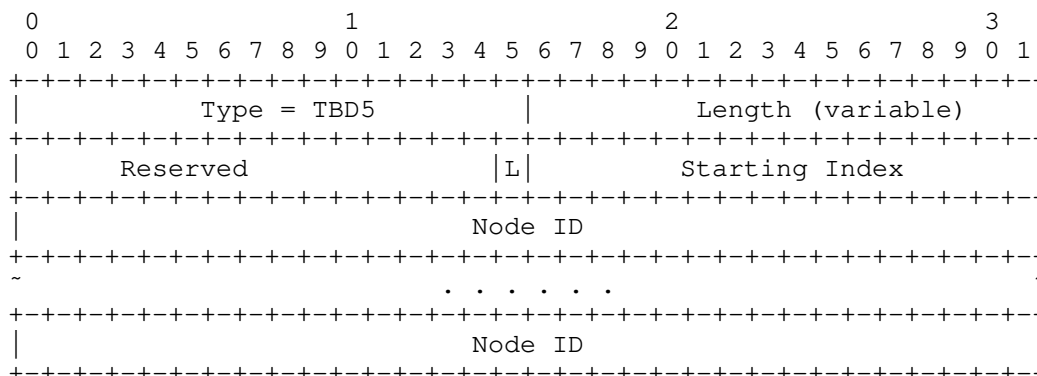


Figure 9: Node IDs TLV

Type: It is to be assigned by IANA.

Length: $4 * (\text{number of Node IDs} + 1)$.

Reserved: MUST be set to zero in transmission and should be ignored on reception.

L: This bit is set to one if the index of the last node ID in this TLV is equal to the last index in the full list of node IDs for the BFP-SPF domain.

Starting Index: The index of the first node ID in this TLV is Starting Index; the index of the second node ID in this TLV is Starting Index + 1; the index of the third node ID in this TLV is Starting Index + 2; and so on.

Node ID: The BGP identifier of a node in the BGP-SPF domain.

5.2.1.4. Paths TLV

The format of Paths TLV is illustrated in Figure 10. A leader uses this TLV to advertise a part of flooding topology for centralized mode. A path may be described as a sequence of indices: (Index 1, Index 2, Index 3, ...), denoting a connection between the node with index 1 and the node with index 2, a connection between the node with index 2 and the node with index 3, and so on. A single link/connection is a simple case of a path that only connects two nodes. A single link path may be encoded in a paths TLV of 8 bytes with two indices.

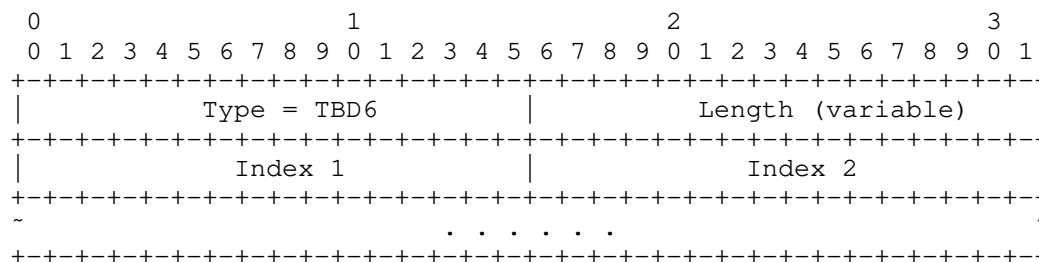


Figure 10: Paths TLV

Type: It is to be assigned by IANA.

Length: $2 * (\text{number of indices in the path})$ when the TLV contains the indices for one path.

Index 1: The index of the first node in the path.

Index 2: The index of the second (next) node in the path.

Multiple such as N paths may be encoded in one paths TLV. Each of the multiple paths is represented as a sequence of indices of the nodes on the path, and two paths (i.e., two sequences of indices for the two paths) are separated by a special index value such as 0xFFFF. In this case, there are (N - 1) special indices as separators to separate N paths, and the Length field has a value of $2 * (\text{number of indices in N paths} + N - 1)$.

When there are a number such as N of single link paths, using one paths TLV to represent them is more efficient than using N paths TLVs to represent them (i.e., each paths TLV represents a single link path). Using one TLV consumes $4 + 2 * (2*N + N - 1) = 6*N + 2$ bytes. Using N TLVs occupies $N * (4 + 4) = 8*N$ bytes. The space used by the former is about three quarters of the space used by the latter for a big N such as 30.

5.2.1.5. Connection Used for Flooding TLV

The format of Connection Used for Flooding TLV is illustrated in Figure 11.

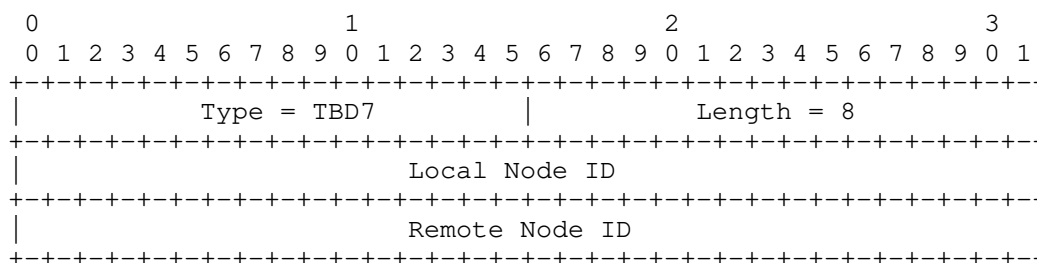


Figure 11: Connection Used for Flooding TLV

Type: It is to be assigned by IANA.

Length: 8.

Local Node ID: The BGP ID of the local node of the session over the connection on the flooding topology which is used for flooding link states.

Remote Node ID: The BGP ID of the remote node of the session over the connection on the flooding topology which is used for flooding link states.

5.2.2. Flooding Topology Distribution in Centralized Mode

In centralized mode, the leader computes a flooding topology for the domain whenever there is a change in the real network topology of the domain and advertises the flooding topology to every node in the domain.

After the current leader has failed, a new leader is elected. The new leader computes a flooding topology for the domain and advertises the flooding topology to every node in the domain.

For a brand new flooding topology of the domain computed, the leader advertises the whole flooding topology to every node in the domain. The leader advertises the mappings between all the node IDs and their indices to every node in the domain using a number of node IDs TLVs first. These node IDs TLVs contain the IDs of all the nodes in the domain and indicates the index corresponding to each of the node IDs and are advertised under MP_REACH_NLRI in BGP update messages. And then the leader advertises the connections/links on the flooding topology to every node in the domain using a number of paths TLVs. These paths TLVs contain all the connections/links on the flooding topology and are advertised under MP_REACH_NLRI in BGP update messages.

After advertising a flooding topology to every node in the domain, which is called the current flooding topology, for a new flooding topology computed for the updated real network topology of the domain, the leader advertises only the changes in the new flooding topology comparing to the current flooding topology to every node in the domain. The leader advertises the changes in the mappings between all the node IDs and their indices to every node in the domain using node IDs TLVs first, and then advertises the changes in the flooding topology to every node in the domain using paths TLVs.

For the new nodes added into the domain, the leader advertises the mappings between the IDs of the new nodes and their indices using a node IDs TLV under MP_REACH_NLRI in a BGP update message to add the mappings. For the dead nodes removed from the domain, the leader advertises the mappings between the IDs of the dead nodes and their indices using a node IDs TLV under MP_UNREACH_NLRI in a BGP update message to withdraw the mappings.

For the new connections/links added into the current flooding topology, the leader advertises the new connections/links using a paths TLV under MP_REACH_NLRI in a BGP update message to add the new connections/inks to the current flooding topology. For the old connections/links removed from the current flooding topology, the leader advertises the old connections/links using a paths TLV under MP_UNREACH_NLRI in a BGP update message to withdraw the old connections/links from the current flooding topology.

5.2.3. An Algorithm for Distributed Mode

This section specifies an algorithm that can be used by every node to compute its flooding topology.

The algorithm for computing a flooding topology of a BGP-SPF domain (real topology) is described as follows.

- * Select a node R0 with the smallest node ID and without the status indicating that the node does not support transit;
- * Build a tree using R0 as root of the tree (details below);
- * And then connect a leaf to the tree to have a flooding topology (details follow).

The algorithm starts from

- * a variable MaxD with an initial value 3,
 - * an initial flooding topology $FT = \{(R0, D=0, PHs=\{\})\}$ with node R0 as root, where R0's Degree $D = 0$, Previous Hops $PHs = \{\}$;
 - * an initial candidate queue $Cq = \{(R1, D=0, PHs=\{R0\}), (R2, D=0, PHs=\{R0\}), \dots, (Rm, D=0, PHs=\{R0\})\}$, where each of nodes R1 to Rm is connected to R0, its Degree $D = 0$ and Previous Hops $PHs = \{R0\}$, R1 to Rm are in increasing order by their IDs.
1. Find and remove the first element with node A from Cq that is not on FT and one PH's D in $PHs < MaxD$, and add the element with A into FT; Set A's D to one, increase A's PH's D by one. If no element in Cq satisfies the conditions, algorithm is restarted with ++MaxD, the initial FT and Cq.
 2. If all the nodes are on the FT, then goto step 4;
 3. Suppose that node Xi ($i = 1, 2, \dots, n$) is connected to node A and not on FT, and X1, X2, ..., Xn are in increasing order by their IDs (i.e., X1's ID < X2's ID < ... < Xn's ID). If they are not ordered, then make them in the order. If Xi is not in Cq, then add it into the end of Cq with $D = 0$ and $PHs = \{A\}$; otherwise (i.e., Xi is in Cq), add A into the end of Xi's PHs; Goto step 1.
 4. For each node B on FT whose D is one (from minimum to maximum node ID), find a link L attached to B such that L's remote node R can transit traffic and has minimum D and ID (if there is no node R which can transit traffic, then find a link L to node R whose D and ID are minimum), add link L between B and R into FT and increase B's D and R's D by one. Return FT.
6. Security Considerations

TBD

7. Acknowledgements

The authors of this document would like to thank Donald E. Eastlake for the comments.

8. IANA Considerations

TBD

9. References

9.1. Normative References

- [I-D.ietf-lsvr-bgp-spf]
Patel, K., Lindem, A., Zandi, S., and W. Henderickx, "BGP Link-State Shortest Path First (SPF) Routing", Work in Progress, Internet-Draft, draft-ietf-lsvr-bgp-spf-15, 1 July 2021, <<https://www.ietf.org/archive/id/draft-ietf-lsvr-bgp-spf-15.txt>>.
- [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>>.
- [RFC4721] Perkins, C., Calhoun, P., and J. Bharatia, "Mobile IPv4 Challenge/Response Extensions (Revised)", RFC 4721, DOI 10.17487/RFC4721, January 2007, <<https://www.rfc-editor.org/info/rfc4721>>.
- [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>>.

9.2. Informative References

- [I-D.ietf-lsr-dynamic-flooding]
Li, T., Psenak, P., Ginsberg, L., Chen, H., Przygienda, T., Cooper, D., Jalil, L., Dontula, S., and G. S. Mishra, "Dynamic Flooding on Dense Graphs", Work in Progress, Internet-Draft, draft-ietf-lsr-dynamic-flooding-09, 9 June 2021, <<https://www.ietf.org/archive/id/draft-ietf-lsr-dynamic-flooding-09.txt>>.
- [I-D.ietf-lsr-flooding-topo-min-degree]
Chen, H., Toy, M., Yang, Y., Wang, A., Liu, X., Fan, Y., and L. Liu, "Flooding Topology Minimum Degree Algorithm",

Work in Progress, Internet-Draft, draft-ietf-lsr-flooding-
topo-min-degree-02, 1 June 2021,
<[https://www.ietf.org/archive/id/draft-ietf-lsr-flooding-
topo-min-degree-02.txt](https://www.ietf.org/archive/id/draft-ietf-lsr-flooding-topo-min-degree-02.txt)>.

Authors' Addresses

Huaimo Chen
Futurewei
Boston, MA,
United States of America

Email: huaimo.chen@futurewei.com

Gyan S. Mishra
Verizon Inc.
13101 Columbia Pike
Silver Spring, MD 20904
United States of America

Phone: 301 502-1347
Email: gyan.s.mishra@verizon.com

Aijun Wang
China Telecom
Beiqijia Town, Changping District
Beijing
102209
China

Email: wangaj3@chinatelecom.cn

Yisong Liu
China Mobile

Email: liuyisong@chinamobile.com

Haibo Wang
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing
100095
China

Email: rainsword.wang@huawei.com

Yanhe Fan
Casa Systems
United States of America

Email: yfan@casa-systems.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 2, 2022

K. Patel
Arrcus, Inc.
A. Lindem
Cisco Systems
S. Zandi
LinkedIn
W. Henderickx
Nokia
July 1, 2021

BGP Link-State Shortest Path First (SPF) Routing
draft-ietf-lsvr-bgp-spf-15

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 January 2, 2022.

Copyright Notice

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

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

Table of Contents

1. Introduction	3
1.1. Terminology	4
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	8
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	14
5.2.3. IPv4/IPv6 Prefix NLRI Usage	15
5.2.3.1. BGP-LS-SPF Prefix NLRI Attribute SPF Status TLV	16
5.2.4. BGP-LS Attribute Sequence-Number TLV	16
5.3. NEXT_HOP Manipulation	17
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	20
6.3. SPF Calculation based on BGP-LS-SPF NLRI	20

6.4.	IPv4/IPv6 Unicast Address Family Interaction	25
6.5.	NLRI Advertisement	25
6.5.1.	Link/Prefix Failure Convergence	25
6.5.2.	Node Failure Convergence	26
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	30
10.	Management Considerations	31
10.1.	Configuration	31
10.1.1.	Link Metric Configuration	31
10.1.2.	backoff-config	31
10.2.	Operational Data	32
11.	Implementation Status	32
12.	Acknowledgements	33
13.	Contributors	33
14.	References	33
14.1.	Normative References	33
14.2.	Informational References	35
	Authors' Addresses	37

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 NLRIs. 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 liveness detection for the interconnecting links are independent of the BGP protocol. For example, liveness detection could be done using the BFD protocol [RFC5880]. Precisely how discovery and liveness 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 liveness 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. NLRIs having TLVs which do not follow the ordering rules MUST be considered as malformed and discarded with appropriate error logging.

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

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

5.1.2. BGP-LS Attribute

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

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

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

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

5.2. Extensions to BGP-LS

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

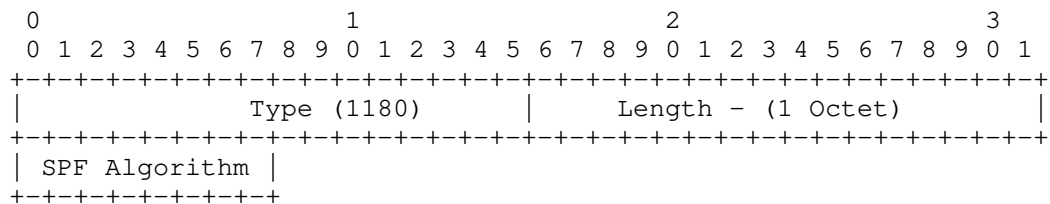
The protocol identifier specified in the Protocol-ID field [RFC7752] will represent the origin of the advertised NLRI. For Node NLRI and Link NLRI, this MUST be the direct protocol (4). Node or Link NLRI with a Protocol-ID other than direct will be considered malformed. For Prefix NLRI, the specified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

5.2.1. Node NLRI Usage

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

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

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



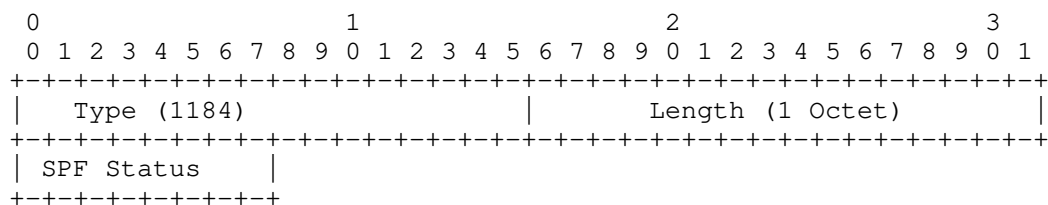
The SPF algorithm inherits the values from the IGP Algorithm Types registry [RFC8665]. Algorithm 0, (Shortest Path Algorithm (SPF))

based on link metric, is supported and described in Section 6.3. Support for other algorithm types is beyond the scope of this specification.

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

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

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



BGP Status Values:

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

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

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

5.2.2. Link NLRI Usage

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

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

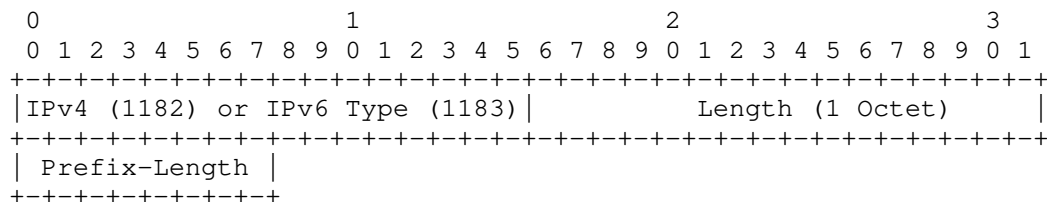
For a link to be used in Shortest Path Tree (SPT) for a given address family, i.e., IPv4 or IPv6, both routers connecting the link MUST have an address in the same subnet for that address family. However, an IPv4 or IPv6 prefix associated with the link MAY be installed without the corresponding address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

The link IGP metric attribute TLV (TLV 1095) MUST be advertised. If a BGP SPF speaker receives a Link NLRI without an IGP metric attribute TLV, then it SHOULD consider the received NLRI as a malformed and the receiving BGP SPF speaker MUST handle such malformed NLRI as 'Treat-as-withdraw' [RFC7606]. The BGP SPF metric length is 4 octets. Like OSPF [RFC2328], a cost is associated with the output side of each router interface. This cost is configurable by the system administrator. The lower the cost, the more likely the interface is to be used to forward data traffic. One possible default for metric would be to give each interface a cost of 1 making it effectively a hop count. Algorithms such as setting the metric inversely to the link speed as supported in the OSPF MIB [RFC4750] MAY be supported. However, this is beyond the scope of this document. Refer to Section 10.1.1 for operational guidance.

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

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

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



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

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

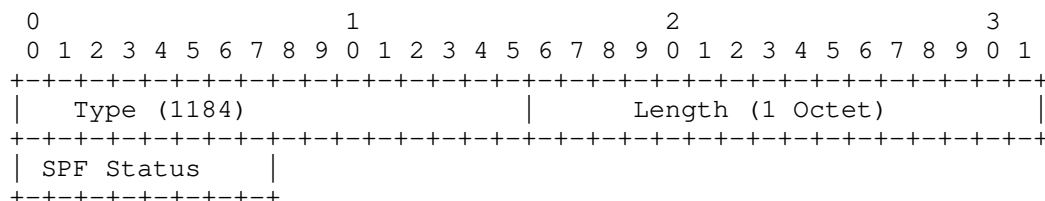
The maximum prefix-length for IPv4 Prefix-Length TLV is 32 bits. A prefix-length field indicating a larger value than 32 bits MUST be discarded as an error and the received TLV is not passed to other BGP peers as specified in [RFC7606]. The corresponding Link NLRI is considered as malformed and MUST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

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

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

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

calculation Section 6.3. A single TLV type will be shared by the Node, Link, and Prefix NLRI. The TLV type 1184 will be assigned by IANA.



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

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

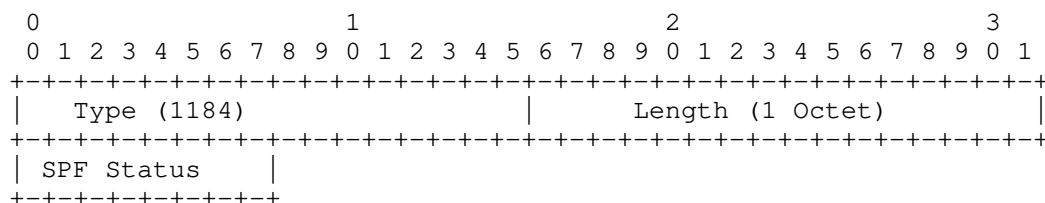
If a BGP SPF speaker received the Link NLRI but the SPF Status TLV is not received, then any previously received information is considered as implicitly withdrawn and the update is propagated to other BGP SPF speakers. A BGP SPF speaker receiving a BGP Update containing an SPF Status TLV in the BGP-LS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

5.2.3. IPv4/IPv6 Prefix NLRI Usage

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

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

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



BGP Status Values:

- 0 - Reserved
- 1 - Prefix Unreachable with respect to SPF
- 2-254 - Undefined
- 255 - Reserved

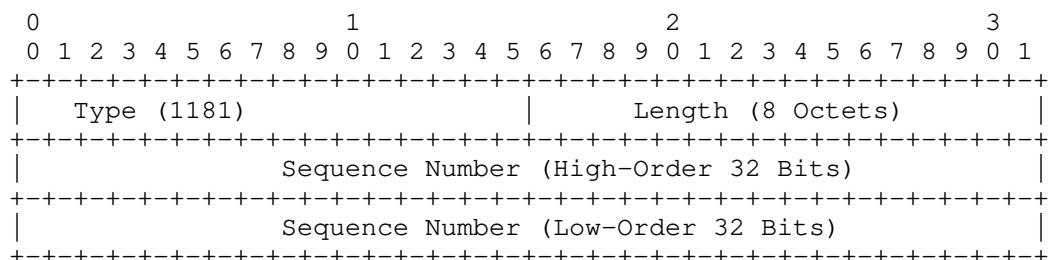
The BGP-LS-SPF Node Attribute SPF Status TLV, Link Attribute SPF Status TLV, and Prefix Attribute SPF Status TLV use the same TLV Type (1184). This implies that a BGP Update cannot contain multiple NLRI with differing status. If the BGP-LS-SPF Status TLV is advertised and the advertised value is not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

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

5.2.4. BGP-LS Attribute Sequence-Number TLV

A BGP-LS Attribute TLV of the BGP-LS-SPF NLRI types is defined to assure the most recent version of a given NLRI is used in the SPF computation. The Sequence-Number TLV is mandatory for BGP-LS-SPF NLRI. The TLV type 1181 has been assigned by IANA. The BGP-LS

Attribute TLV will contain an 8-octet sequence number. The usage of the Sequence Number TLV is described in Section 6.1.



Sequence Number

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

When incrementing the sequence number for each self-originated NLRI, the sequence number should be treated as an unsigned 64-bit value. If the lower-order 32-bit value wraps, the higher-order 32-bit value should be incremented and saved in non-volatile storage. If a BGP SPF speaker completely loses its sequence number state (e.g., the BGP SPF speaker hardware is replaced or experiences a cold-start), the BGP NLRI selection rules (see Section 6.1) will insure convergence, albeit not immediately.

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

5.3. NEXT_HOP Manipulation

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

IPv6 address whenever the length of the Next-Hop address is 16 octets.

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

6. Decision Process with SPF Algorithm

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

The SPF based Decision process replaces the BGP Decision process described in [RFC4271]. This process starts with selecting only those Node NLRI whose SPF capability TLV matches with the local BGP SPF speaker's SPF capability TLV value. Since Link-State NLRI always contains the local node descriptor Section 5.2, each NLRI is uniquely originated by a single BGP SPF speaker in the BGP SPF routing domain (the BGP node matching the NLRI's Node Descriptors). Instances of the same NLRI originated by multiple BGP SPF speakers would be indicative of a configuration error or a masquerading attack (Section 9). These selected Node NLRI and their Link/Prefix NLRI are used to build a directed graph during the SPF computation as described below. The best routes for BGP prefixes are installed in the RIB as a result of the SPF process.

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

The Phase 3 decision function of the Decision Process [RFC4271] is also simplified since under normal SPF operation, a BGP SPF speaker MUST advertise the changed NLRI 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 NLRIs.

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.

- o 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.
- o 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.
- o 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's 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.

- o 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].
- o 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".
- o Link State NLRI Database (LSNDB) - Database of BGP-LS-SPF NLRI that facilitates access to all Node, Link, and Prefix NLRI.
- o Candidate List (CAN-LIST) - This is a list of candidate Node NLRIs 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. The 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 match 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", draft-ietf-lsvr-applicability-05 (work in progress), March 2020.
- [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", draft-psarkar-lsvr-bgp-spf-impl-00 (work in progress), June 2020.
- [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
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