BGP-SPF Flooding Reduction

draft-chen-lsvr-flood-reduction-00

Huaimo Chen
Gyan Mishra
Aijun Wang
Yisong Liu
Haibo Wang
Yanhe Fan

IETF 112

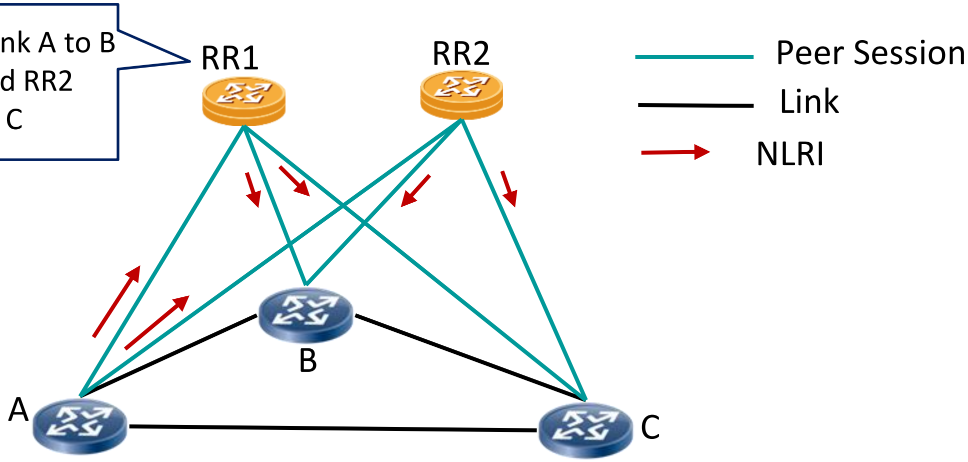
Overview on BGP-SPF Flooding Reduction

- BGP-SPF Flooding Overview
- Revised Flooding Procedures
- Protocol Extensions

BGP-SPF Flooding in RR Model

- BGP SPF speaker sends its Link NLRI to RRs
- After receiving it, RRs sends the NLRI to the other BGP SPF speakers.

For example,
Node A sends its link state to both RR1 and RR2.
RR1 and RR2 sends it to nodes B and C.



After receiving Link NLRI for link A to B from speaker/node A, RR1 and RR2 send the NLRI to nodes B and C

Speaker/Node A sends Link NLRI for link A to B to RR1 and RR2 when A discovers that link A to B is up

Figure 1. BGP-SPF Domain with two RRs

Nodes B and C receives two copies of the same NLRI, one from RR1 and the other from RR2.
Redundant copy should be reduced.

BGP-SPF Flooding in Node Connections Model

Once BGP-SPF speaker knows its link up or down, it triggers

advertising link state to all the nodes through all the BGP sessions.

For example, when node A considers that the link from A to D is up,
A sends the link state for the link up over four sessions (to B, to B, to C, to D)
B sends it to C, D sends it to C, C sends it to B and D

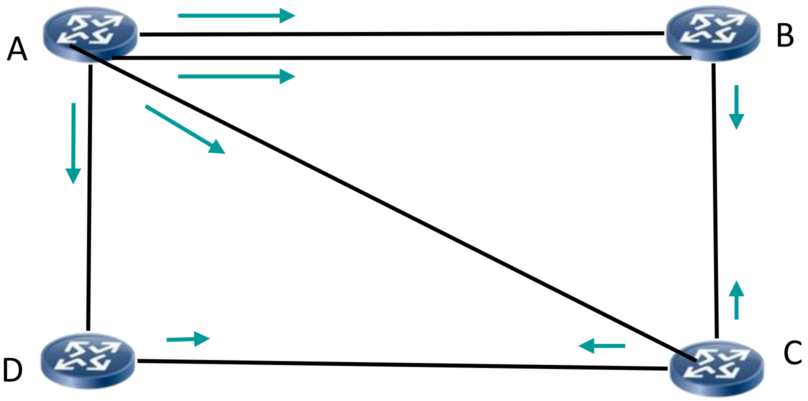


Figure 1a. BGP-SPF Domain with parallel links

Similarly, node D considers that the link from D to A is up and there are link state advertisements for link D→A in network.

Flooding in Directly-Connected Nodes Model

Similar to the one for Node Connections Model. But there is a single BGP session even if there are multiple direct connections between BGP SPF speakers.

Revised Flooding Procedure in RR Model (1/2)

- ✓ BGP SPF speaker/node sends its Link NLRI to some such as one of RRs.
- ✓ After receiving it, the RR sends the NLRI to the other BGP SPF speakers.

For example,
Node A sends its link state one RR RR1 and does not send the NLRI to RR2.
After receiving the link state from A, RR1 sends it to the other nodes B and C.

After receiving Link NLRI for link A to B from speaker/node A, RR1 sends the NLRI to nodes B and C

Speaker/Node A sends Link NLRI for link A to B to RR1 when A discovers that link A to B is up. Node A does not send the NLRI to RR2.

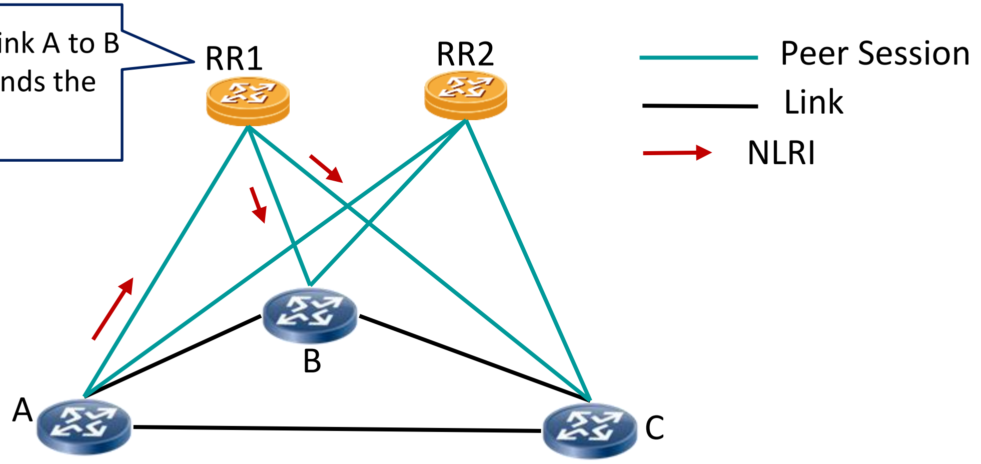


Figure 2. BGP-SPF Domain with two RRs

Nodes B and C receives only one copy of the same NLRI
Comparing to normal flooding in RR model, revised flooding reduced the amount of flooding by half.

Revised Flooding Procedure in RR Model (2/2)

- Nodes are evenly divided into the number of groups.
- Each group sends their link NLRIs to one RR. (i.e., A first group of nodes sends their link NLRIs to a first RR; a second group of nodes sends their link NLRIs to a second RR; and so on.)

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

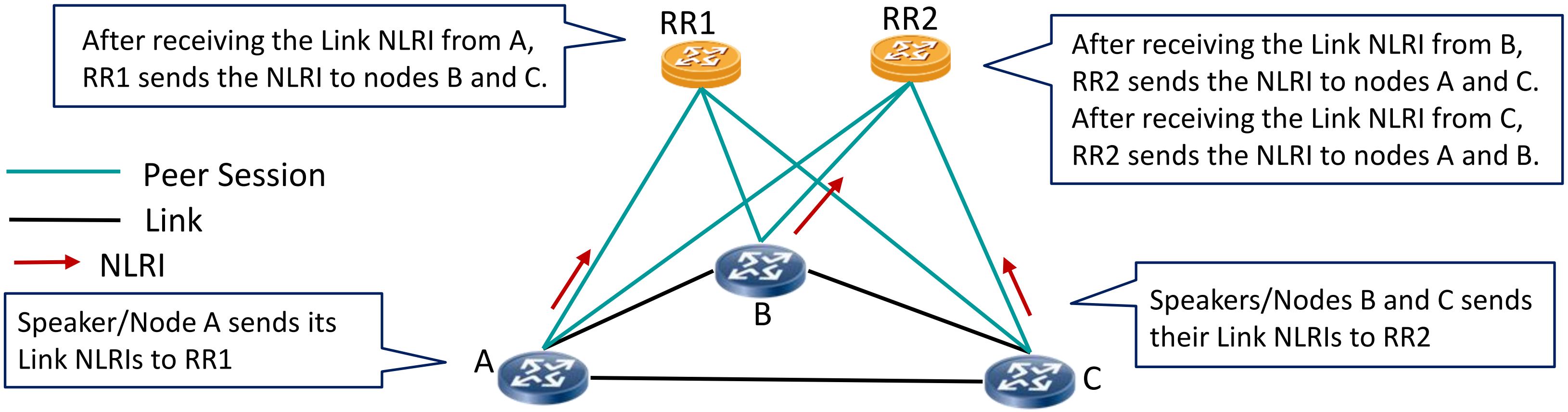
For example,

The nodes in the network are evenly divided into two groups. The first group contains A. The second contains B and C.

Node A in the first group sends its link NLRIs to RR1.

Nodes B and C in the second group send their link NLRIs to RR2.

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



After receiving the Link NLRI from A, RR1 sends the NLRI to nodes B and C.

After receiving the Link NLRI from B, RR2 sends the NLRI to nodes A and C. After receiving the Link NLRI from C, RR2 sends the NLRI to nodes A and B.

Speaker/Node A sends its Link NLRIs to RR1

Speakers/Nodes B and C sends their Link NLRIs to RR2

Figure 3. BGP-SPF Domain with two RRs

Revised Flooding Procedure in Node Connections Model (1/2)

Similar to the one in ietf-lsr-dynamic-flooding:

- ✓ Each node has a flooding topology (FT).
  - In an option, FT is computed in a distributed mode, where every BGP SPF speaker computes a FT for the network using a same algorithm.
  - In another, FT is computed in a centralized mode, where one BGP SPF speaker elected as a leader computes a FT and advertises FT to every BGP SPF speaker. For a new FT computed, only changes are advertised.
- ✓ Each node sends link NLRI to its peers on FT (i.e., are connected by the links on the FT).

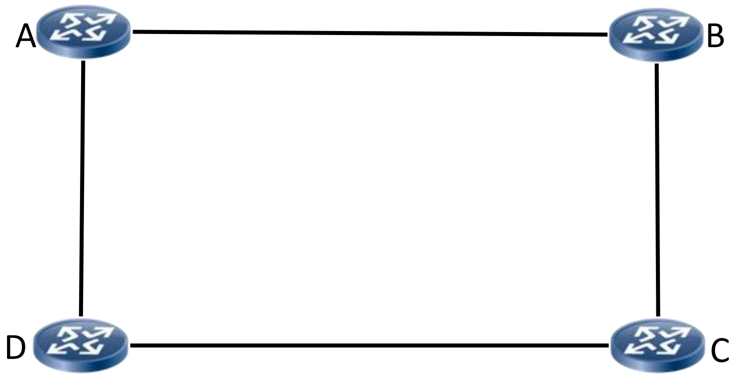


Figure 2a. A Flooding Topology

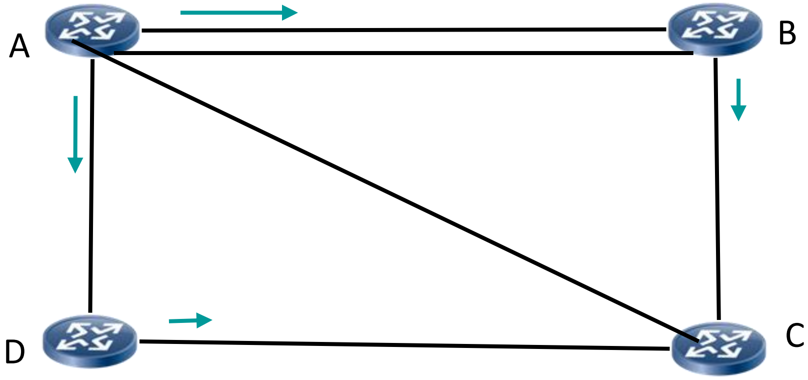


Figure 3a. Advertise NLRI Using Flooding Topology

For example, Figure 3a shows a flooding flow of a link NLRI.

A sends the NLRI to its peers B and D. B and D are peers of node A and on the FT. A does not send the NLRI to its peer C since C is not connected to A on the FT.

After receiving it from A, B sends it to C; D sends it to C.

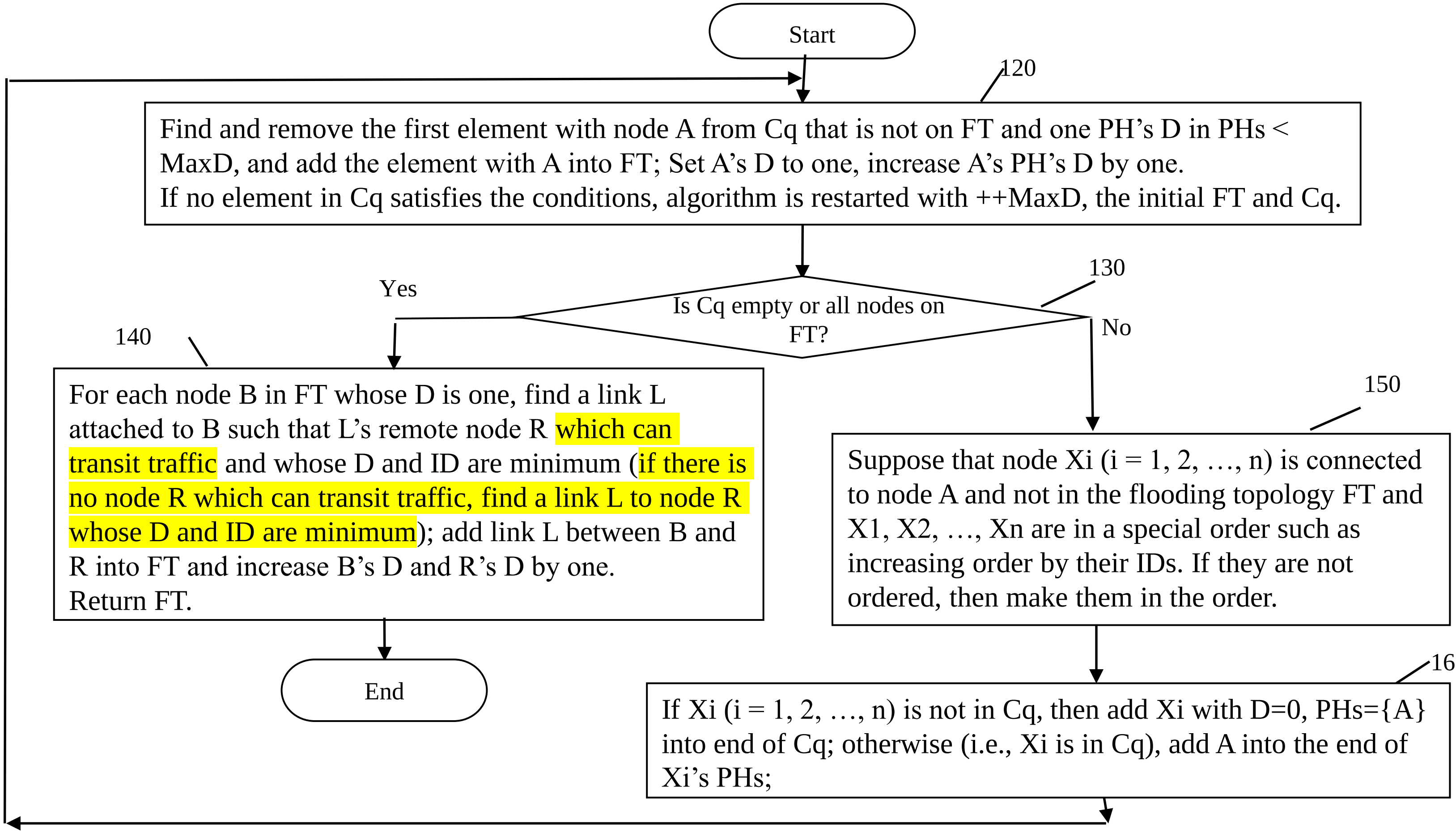
The number of NLRIs flooded in revised flooding is much less than that in normal flooding.

Revised Flooding Procedure in Node Connections Model (2/2)

An algorithm for computing a BGP-SPF flooding topology is similar to the one in ietf-lsr-flooding-topo-min-degree.

1. Select a node R according to a rule such as the node with the biggest/smallest node ID and without the status indicating that the node doesn't support transit
2. Build a tree using R as root of the tree (details below)
3. and then connect a leaf to the tree to have a flooding topology (details follow).

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, PHs = { }; an initial candidate queue Cq = {(R1,D=0, PHs={R0}), (R2,D=0, PHs={R0}), …, (Rm,D=0, PHs={R0})}, where each of nodes R1 to Rm is connected to R0, its Degree D = 0 and PHs ={R0}, R1 to Rm are in a special order such as increasing order by their IDs.



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.

Is Cq empty or all nodes on FT?

For each node B in FT whose D is one, find a link L attached to B such that L's remote node R which can transit traffic and whose D and ID are minimum (if there is no node R which can transit traffic, 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.

Suppose that node Xi (i = 1, 2, …, n) is connected to node A and not in the flooding topology FT and X1, X2, …, Xn are in a special order such as increasing order by their IDs. If they are not ordered, then make them in the order.

If Xi (i = 1, 2, …, n) is not in Cq, then add Xi with D=0, PHs={A} into end of Cq; otherwise (i.e., Xi is in Cq), add A into the end of Xi's PHs;

Protocol Extensions for RR Model

Node Flood TLV: Leader RR uses this TLV to tell every node how to flood its link states.
(Leader RR is the RR with the highest priority and node ID)

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           Type = TBDa         |          Length = 4           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                   Reserved                    |Flood-behavior |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure E1 Node Flood TLV

```
Type: TBDa
Length: 4
Reserved: Must be set to zero when transmitting it and ignored when receiving it
Flood-behavior: 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.
```

Protocol Extensions for Node Connections Model (1/5)

Leader Preference TLV: A node uses this TLV to indicate its priority for becoming a leader.

```
  0                   1                   2                   3
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |           Type = TBD1         |           Length = 4          |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |          Reserved             |   Priority    |   Algorithm   |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure E2 Leader Preference TLV

```
Type: TBD1
Length: 4
Priority: 0-255, unsigned integer indicating priority to become a leader
Algorithm:    0 - Centralized computation by the Leader.
          1-127 - Standardized distributed algorithms.
        128-254 - Private distributed algorithms.
```

Leader use this TLV to tell every node to

- use the flooding topology from the leader through advertising the TLV with Algorithm = 0, or
- compute its own flooding topology using the algorithm given by the Algorithm =/= 0 in the TLV

Protocol Extensions for Node Connections Model (2/5)

Algorithm Support TLV: A node uses this TLV to indicate the algorithms that it supports for distributed mode.

```
  0                   1                   2                   3
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |           Type = TBD2         |       Length (variable)       |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |   Algorithm   |   Algorithm   |    . . .
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure E3 Algorithm Support TLV

Type: TBD2
Length: number of Algorithms (0 or more)
Algorithm: a numeric identifier in the range 0-255 indicating the algorithm that can be used to compute the flooding topology

Protocol Extensions for Node Connections Model (3/5)

Node IDs TLV: A leader uses this TLV to indicate the mapping from nodes to their indexes for centralized mode.

```
  0                   1                   2                   3
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |           Type = TBD3         |       Length (variable)       |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |      Reserved               |L|        Starting Index         |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |                             Node ID                           |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 ~                          . . . . . .                          ~
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |                             Node ID                           |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure E4   Node IDs TLV

Type: TBD3

Length: 4 * (number of Node IDs + 1)

L (Last): This bit is set 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.

Protocol Extensions for Node Connections Model (4/5)

Paths TLV: A leader uses this TLV to advertise a part of flooding topology for centralized mode.

Index 1, Index 2, Index 3, ... in TLV, denoting a connection from the node with index 1 to the node with index 2, a connection from the node with index 2 to the node with index 3, and so on.

```
  0                   1                   2                   3
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |          Type = TBD4          |       Length (variable)       |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |            Index 1            |            Index 2            |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 ~                           . . . . . .                         ~
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure E5 Paths TLV

Type: TBD4
Length: 2 * (number of indices in the 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 paths may be encoded in one TLV to improve efficiency
Two paths are separated by a special index value such as 0xFFFF.

Protocol Extensions for Node Connections Model (5/5)

Connection Used for Flooding (CUF) TLV: A node indicates that a connection/link is a part of the flooding topology and used for flooding by its (local) node ID and remote node ID of the session over the connection/link.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           Type = TBD5         |          Length = 8           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Local Node ID                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Remote Node ID                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure E6   Connection Used for Flooding TLV

```
Type: TBD5
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.
```

Next Step

Comments