Inter-Domain Routing
Internet-Draft
Intended status: Standards Track
Expires: April 25, 2013

H. Gredler
Juniper Networks, Inc.
J. Medved
S. Previdi
Cisco Systems, Inc.
A. Farrel
Juniper Networks, Inc.
S. Ray
Cisco Systems, Inc.
October 22, 2012

# North-Bound Distribution of Link-State and TE Information using BGP
draft-ietf-idr-ls-distribution-01

## Abstract

In a number of environments, a component external to a network is called upon to perform computations based on the network topology and current state of the connections within the network, including traffic engineering information. This is information typically distributed by IGP routing protocols within the network

This document describes a mechanism by which links state and traffic engineering information can be collected from networks and shared with external components using the BGP routing protocol. This is achieved using a new BGP Network Layer Reachability Information (NLRI) encoding format. The mechanism is applicable to physical and virtual links. The mechanism described is subject to policy control.

Applications of this technique include Application Layer Traffic Optimization (ALTO) servers, and Path Computation Elements (PCEs).

## 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

Copyright Notice

Table of Contents

1. Introduction . . . . . . . . . . . . . . . . . . . . . . . . 4
2. Motivation and Applicability . . . . . . . . . . . . . . . . 5
  2.1. MPLS-TE with PCE . . . . . . . . . . . . . . . . . . . . . 5
  2.2. ALTO Server Network API . . . . . . . . . . . . . . . . . 7
3. Carrying Link State Information in BGP . . . . . . . . . . . 8
  3.1. TLV Format . . . . . . . . . . . . . . . . . . . . . . . . 8
  3.2. The Link State NLRI . . . . . . . . . . . . . . . . . . . 9
    3.2.1. Node Descriptors . . . . . . . . . . . . . . . . . . . 11
    3.2.2. Link Descriptors . . . . . . . . . . . . . . . . . . . 15
    3.2.3. The Prefix NLRI . . . . . . . . . . . . . . . . . . . 16
  3.3. The LINK_STATE Attribute . . . . . . . . . . . . . . . . . 16
    3.3.1. Link Attribute TLVs . . . . . . . . . . . . . . . . . 16
    3.3.2. Node Attribute TLVs . . . . . . . . . . . . . . . . . 20
    3.3.3. Prefix Attributes TLVs . . . . . . . . . . . . . . . . 23
  3.4. BGP Next Hop Information . . . . . . . . . . . . . . . . . 27
  3.5. Inter-AS Links . . . . . . . . . . . . . . . . . . . . . . 27
4. Link to Path Aggregation . . . . . . . . . . . . . . . . . . 27
  4.1. Example: No Link Aggregation . . . . . . . . . . . . . . . 27
  4.2. Example: ASBR to ASBR Path Aggregation . . . . . . . . . . 28
  4.3. Example: Multi-AS Path Aggregation . . . . . . . . . . . . 28
5. IANA Considerations . . . . . . . . . . . . . . . . . . . . 29
6. Manageability Considerations . . . . . . . . . . . . . . . . 29
  6.1. Operational Considerations . . . . . . . . . . . . . . . . 29
    6.1.1. Operations . . . . . . . . . . . . . . . . . . . . . . 29
    6.1.2. Installation and Initial Setup . . . . . . . . . . . . 30
    6.1.3. Migration Path . . . . . . . . . . . . . . . . . . . . 30
    6.1.4. Requirements on Other Protocols and Functional
           Components . . . . . . . . . . . . . . . . . . . . . . 30
    6.1.5. Impact on Network Operation . . . . . . . . . . . . . 30
    6.1.6. Verifying Correct Operation . . . . . . . . . . . . . 30
  6.2. Management Considerations . . . . . . . . . . . . . . . . 31
    6.2.1. Management Information . . . . . . . . . . . . . . . . 31
    6.2.2. Fault Management . . . . . . . . . . . . . . . . . . . 31
    6.2.3. Configuration Management . . . . . . . . . . . . . . . 31
    6.2.4. Accounting Management . . . . . . . . . . . . . . . . 31
    6.2.5. Performance Management . . . . . . . . . . . . . . . . 31
    6.2.6. Security Management . . . . . . . . . . . . . . . . . 32
7. Security Considerations . . . . . . . . . . . . . . . . . . 32
8. Acknowledgements . . . . . . . . . . . . . . . . . . . . . . 32
9. References . . . . . . . . . . . . . . . . . . . . . . . . . 32
  9.1. Normative References . . . . . . . . . . . . . . . . . . . 32
  9.2. Informative References . . . . . . . . . . . . . . . . . . 34
Authors' Addresses . . . . . . . . . . . . . . . . . . . . . . . 34

## 1. Introduction

The contents of a Link State Database (LSDB) or a Traffic Engineering Database (TED) has the scope of an IGP area. Some applications, such as end-to-end Traffic Engineering (TE), would benefit from visibility outside one area or Autonomous System (AS) in order to make better decisions.

The IETF has defined the Path Computation Element (PCE) [RFC4655] as a mechanism for achieving the computation of end-to-end TE paths that cross the visibility of more than one TED or which require CPU-intensive or coordinated computations. The IETF has also defined the ALTO Server [RFC5693] as an entity that generates an abstracted network topology and provides it to network-aware applications.

Both a PCE and an ALTO Server need to gather information about the topologies and capabilities of the network in order to be able to fulfill their function

This document describes a mechanism by which Link State and TE information can be collected from networks and shared with external components using the BGP routing protocol [RFC4271]. This is achieved using a new BGP Network Layer Reachability Information (NLRI) encoding format. The mechanism is applicable to physical and virtual links. The mechanism described is subject to policy control.

A router maintains one or more databases for storing link-state information about nodes and links in any given area. Link attributes stored in these databases include: local/remote IP addresses, local/remote interface identifiers, link metric and TE metric, link bandwidth, reservable bandwidth, per CoS class reservation state, preemption and Shared Risk Link Groups (SRLG). The router's BGP process can retrieve topology from these LSDBs and distribute it to a consumer, either directly or via a peer BGP Speaker (typically a dedicated Route Reflector), using the encoding specified in this document.

The collection of Link State and TE link state information and its distribution to consumers is shown in the following figure.

```
                           +-----------+
                           | Consumer  |
                           +-----------+
                                 ^
                                 |
                           +-----------+
                           |    BGP    |               +-----------+
                           |  Speaker  |               | Consumer  |
                           +-----------+               +-----------+
                             ^   ^   ^                       ^
                             |   |   |                       |
             +---------------+   |   +-------------------+   |
             |                   |                       |   |
       +-----------+       +-----------+             +-----------+
       |    BGP    |       |    BGP    |             |    BGP    |
       |  Speaker  |       |  Speaker  |    . . .    |  Speaker  |
       +-----------+       +-----------+             +-----------+
             ^                   ^                         ^
             |                   |                         |
            IGP                 IGP                       IGP
```

Figure 1: TE Link State info collection

A BGP Speaker may apply configurable policy to the information that it distributes. Thus, it may distribute the real physical topology from the LSDB or the TED. Alternatively, it may create an abstracted topology, where virtual, aggregated nodes are connected by virtual paths. Aggregated nodes can be created, for example, out of multiple routers in a POP. Abstracted topology can also be a mix of physical and virtual nodes and physical and virtual links. Furthermore, the BGP Speaker can apply policy to determine when information is updated to the consumer so that there is reduction of information flow form the network to the consumers. Mechanisms through which topologies can be aggregated or virtualized are outside the scope of this document

## 2. Motivation and Applicability

This section describes uses cases from which the requirements can be derived.

### 2.1. MPLS-TE with PCE

As described in [RFC4655] a PCE can be used to compute MPLS-TE paths within a "domain" (such as an IGP area) or across multiple domains (such as a multi-area AS, or multiple ASes).

- Within a single area, the PCE offers enhanced computational power that may not be available on individual routers, sophisticated policy control and algorithms, and coordination of computation across the whole area.

- If a router wants to compute a MPLS-TE path across IGP areas its own TED lacks visibility of the complete topology. That means that the router cannot determine the end-to-end path, and cannot even select the right exit router (Area Border Router - ABR) for an optimal path. This is an issue for large-scale networks that need to segment their core networks into distinct areas, but which still want to take advantage of MPLS-TE.

Previous solutions used per-domain path computation [RFC5152]. The source router could only compute the path for the first area because the router only has full topological visibility for the first area along the path, but not for subsequent areas. Per-domain path computation uses a technique called "loose-hop-expansion" [RFC3209], and selects the exit ABR and other ABRs or AS Border Routers (ASBRs) using the IGP computed shortest path topology for the remainder of the path. This may lead to sub-optimal paths, makes alternate/ back-up path computation hard, and might result in no TE path being found when one really does exist.

The PCE presents a computation server that may have visibility into more than one IGP area or AS, or may cooperate with other PCEs to perform distributed path computation. The PCE obviously needs access to the TED for the area(s) it serves, but [RFC4655] does not describe how this is achieved. Many implementations make the PCE a passive participant in the IGP so that it can learn the latest state of the network, but this may be sub-optimal when the network is subject to a high degree of churn, or when the PCE is responsible for multiple areas.

The following figure shows how a PCE can get its TED information using the mechanism described in this document.

```
                   +----------+                           +---------+
                   |  -----   |                           |   BGP   |
                   | | TED |<-+-------------------------->| Speaker |
                   |  -----   |   TED synchronization     |         |
                   |    |     |        mechanism:         +---------+
                   |    |     | BGP with Link-State NLRI
                   |    v     |
                   |  -----   |
                   | | PCE |  |
                   |  -----   |
                   +----------+
                        ^
                        | Request/
                        | Response
                        v
          Service  +----------+   Signaling  +----------+
          Request  | Head-End |   Protocol   | Adjacent |
          -------->|  Node    |<------------>|   Node   |
                   +----------+              +----------+
```

Figure 2: External PCE node using a TED synchronization mechanism

The mechanism in this document allows the necessary TED information to be collected from the IGP within the network, filtered according to configurable policy, and distributed to the PCE as necessary.

### 2.2. ALTO Server Network API

An ALTO Server [RFC5693] is an entity that generates an abstracted network topology and provides it to network-aware applications over a web service based API.  Example applications are p2p clients or trackers, or CDNs.  The abstracted network topology comes in the form of two maps: a Network Map that specifies allocation of prefixes to PIDs, and a Cost Map that specifies the cost between PIDs listed in the Network Map. For more details, see [I-D.ietf-alto-protocol].

ALTO abstract network topologies can be auto-generated from the physical topology of the underlying network.  The generation would typically be based on policies and rules set by the operator.  Both prefix and TE data are required: prefix data is required to generate ALTO Network Maps, TE (topology) data is required to generate ALTO Cost Maps.  Prefix data is carried and originated in BGP, TE data is originated and carried in an IGP.  The mechanism defined in this document provides a single interface through which an ALTO Server can retrieve all the necessary prefix and network topology data from the underlying network.  Note an ALTO Server can use other mechanisms to get network data, for example, peering with multiple IGP and BGP Speakers.

The following figure shows how an ALTO Server can get network topology information from the underlying network using the mechanism described in this document.

```
     +--------+
     | Client |<--+
     +--------+   |
                  |    ALTO    +--------+     BGP with    +---------+
     +--------+   |  Protocol  |  ALTO  | Link-State NLRI |   BGP   |
     | Client |<--+------------| Server |<----------------| Speaker |
     +--------+   |            |        |                 |         |
                  |            +--------+                 +---------+
     +--------+   |
     | Client |<--+
     +--------+
```

Figure 3: ALTO Server using network topology information

## 3. Carrying Link State Information in BGP

Two parts: a new BGP NLRI that describes links and nodes comprising IGP link state information, and a new BGP path attribute that carries link and node properties and attributes, such as the link metric or node properties.

### 3.1. TLV Format

Information in the new link state NLRIs and attributes is encoded in Type/Length/Value triplets. The TLV format is shown in Figure 4.

```
      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |              Type             |             Length            |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                                                               |
     |                       Value (variable)                        |
     |                                                               |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4: TLV format

The Length field defines the length of the value portion in octets (thus a TLV with no value portion would have a length of zero). The TLV is not padded to four-octet alignment; Unrecognized types are ignored.

### 3.2.  The Link State NLRI

The MP_REACH and MP_UNREACH attributes are BGP's containers for carrying opaque information.  Each Link State NLRI describes either a single node or link.

All link, node and prefix information SHALL be encoded using a TBD AFI / TBD SAFI header into those attributes.

In order for two BGP speakers to exchange Link-State NLRI, they MUST use BGP Capabilities Advertisement to ensure that they both are capable of properly processing such NLRI.  This is done as specified in [RFC4760], by using capability code 1 (multi-protocol BGP), with an AFI/SAFI TBD.

The format of the Link State NLRI is shown in the following figure.

```
      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
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |            NLRI Type          |     Total NLRI Length         |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                                                               |
     |                   Link-State NLRI (variable)                  |
     |                                                               |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 5: Link State SAFI 1 NLRI Format

```
      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
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |            NLRI Type          |     Total NLRI Length         |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                                                               |
     +                       Route Distinguisher                     +
     |                                                               |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                                                               |
     |                   Link-State NLRI (variable)                  |
     |                                                               |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 6: Link State SAFI 128 NLRI Format

The 'Total NLRI Length' field contains the cumulative length of all the TLVs in the NLRI.  For VPN applications it also includes the

length of the Route Distinguisher.

The 'NLRI Type' field can contain one of the following values:

   Type = 1: Link NLRI, contains link descriptors and link attributes

   Type = 2: Node NLRI, contains node attributes

   Type = 3: IPv4 Topology Prefix NLRI

   Type = 4: IPv6 Topology Prefix NLRI

The Link NLRI (NLRI Type = 1) is shown in the following figure.

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

Figure 7: The Link NLRI format

The Node NLRI (NLRI Type = 2) is shown in the following figure.

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

Figure 8: The Node NLRI format

The IPv4 and IPv6 Prefix NLRIs (NLRI Type = 3 and Type = 4) use the same format as shown in the following figure.

```
      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  Protocol-ID  |    Reserved    |       Instance Identifier     |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                    Node Descriptor                             |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                    Prefix NLRI (variable)                      |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 9: The IPv4/IPv6 Topology Prefix NLRI format

The 'Protocol-ID' field can contain one of the following values:

Protocol-ID = 0: Unknown, The source of NLRI information could not be determined

Protocol-ID: IS-IS Level 1, The NLRI information has been sourced by IS-IS Level 1

Protocol-ID: IS-IS Level 2, The NLRI information has been sourced by IS-IS Level 2

Protocol-ID = 3: OSPF, The NLRI information has been sourced by OSPF

Protocol-ID = 4: Direct, The NLRI information has been sourced from local interface state

Protocol-ID = 5: Static, The NLRI information has been sourced by static configuration

Both OSPF and IS-IS may run multiple routing protocol instances over the same link.  See [I-D.ietf-isis-mi] and [RFC6549].  The 'Instance Identifier' field identifies the protocol instance.

Each Node Descriptor and Link Descriptor consists of one or more TLVs described in the following sections.  The sender of an UPDATE message MUST order the TLVs within a Node Descriptor or a Link Descriptor in ascending order of TLV type."

### 3.2.1.  Node Descriptors

Each link gets anchored by at least a pair of router-IDs.  Since there are many Router-IDs formats (32 Bit IPv4 router-ID, 56 Bit ISO Node-ID and 128 Bit IPv6 router-ID) a link may be anchored by more than one Router-ID pair.  The set of Local and Remote Node Descriptors describe which Protocols Router-IDs will be following to

"anchor" the link described by the "Link attribute TLVs". There must be at least one "like" router-ID pair of a Local Node Descriptors and a Remote Node Descriptors per-protocol. If a peer sends an illegal combination in this respect, then this is handled as an NLRI error, described in [RFC4760].

It is desirable that the Router-ID assignments inside the Node anchor are globally unique. However there may be router-ID spaces (e.g. ISO) where not even a global registry exists, or worse, Router-IDs have been allocated following private-IP RFC 1918 [RFC1918] allocation. In order to disambiguate the Router-IDs the local and remote Autonomous System number TLVs of the anchor nodes MUST be included in the NLRI. If the anchor node's AS is a member of an AS Confederation ([RFC5065]), then the Autonomous System number TLV contains the confederations' AS Confederation Identifier and the Member-AS TLV is included in the NLRI. The Local and Remote Autonomous System TLVs are 4 octets wide as described in [RFC4893]. 2-octet AS Numbers SHALL be expanded to 4-octet AS Numbers by zeroing the two MSB octets.

### 3.2.1.1. Local Node Descriptors

The Local Node Descriptors TLV (Type 256) contains Node Descriptors for the node anchoring the local end of the link. The length of this TLV is variable. The value contains one or more Node Descriptor Sub-TLVs defined in Section 3.2.1.3.

```
      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |              Type             |             Length            |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                                                               |
     |               Node Descriptor Sub-TLVs (variable)             |
     |                                                               |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 10: Local Node Descriptors TLV format

### 3.2.1.2. Remote Node Descriptors

The Remote Node Descriptors TLV (Type 257) contains Node Descriptors for the node anchoring the remote end of the link. The length of this TLV is variable. The value contains one or more Node Descriptor Sub-TLVs defined in Section 3.2.1.3.

```
      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |              Type             |             Length            |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                                                               |
     |               Node Descriptor Sub-TLVs (variable)             |
     |                                                               |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 11: Remote Node Descriptors TLV format

### 3.2.1.3. Node Descriptor Sub-TLVs

The Node Descriptor Sub-TLV type codepoints and lengths are listed in the following table:

| Type | Description | Length |
|---|---|---|
| 258 | Autonomous System | 4 |
| 259 | Member-AS | 4 |
| 260 | ISO Node-ID | 7 |
| 261 | IPv4 Router-ID | 5 |
| 262 | IPv4 Router-ID | 17 |

Table 1: Node Descriptor Sub-TLVs

The TLV values in Node Descriptor Sub-TLVs are defined as follows:

Autonomous System:  opaque value (32 Bit AS ID)

Member-AS:  opaque value (32 Bit AS ID); only included if the node is in an AS confederation.

IPv4 Router ID:  opaque value (can be an IPv4 address or an 32 Bit router ID).

IPv6 Router ID:  opaque value (can be an IPv6 address or 128 Bit router ID).

ISO Node ID:  ISO node-ID (6 octets ISO system-ID) followed by a PSN octet in case LAN "Pseudonode" information gets advertised.  The PSN octet must be zero for non-LAN "Pseudonodes".

There can be at most one instance of each TLV type present in any Node Descriptor. The TLV ordering within a Node descriptor MUST be kept in order of increasing numeric value of type. TLVs 258 and 259 specify administrative context in which TLVs 260-262 are to be evaluated. The first TLV from range 260-262 is to be interpreted as the primary node identifier, e.g. it acts as the unique key by which the node can be referenced within its administrative contexts. Any further TLVs are to be treated as secondary identifiers, which may be used for cross-reference, but are to be treated as if they are object attributes.

### 3.2.1.4. Router-ID Anchoring Example: ISO Pseudonode

IS-IS Pseudonodes are a good example for the variable Router-ID anchoring. Consider Figure 12. This represents a Broadcast LAN between a pair of routers. The "real" (=non pseudonode) routers have both an IPv4 Router-ID and IS-IS Node-ID. The pseudonode does not have an IPv4 Router-ID. Two unidirectional links (Node1, Pseudonode 1) and (Pseudonode 1, Node 2) are being generated.

The NRLI for (Node1, Pseudonode1) encodes local IPv4 router-ID, local ISO node-ID and remote ISO node-id)

The NLRI for (Pseudonode1, Node2) encodes a local ISO node-ID and remote ISO node-id.

```
  +-----------------+    +-----------------+    +-----------------+
  |      Node1      |    |  Pseudonode 1   |    |      Node2      |
  |1920.0000.2001.00|--->|1921.6800.1001.02|--->|1920.0000.2002.00|
  |     192.0.2.1   |    |                 |    |     192.0.2.2   |
  +-----------------+    +-----------------+    +-----------------+
```

Figure 12: IS-IS Pseudonodes

### 3.2.1.5. Router-ID Anchoring Example: OSPFv2 to IS-IS Migration

Migrating gracefully from one IGP to another requires congruent operation of both routing protocols during the migration period. The target protocol (IS-IS) supports more router-ID spaces than the source (OSPFv2) protocol. When advertising a point-to-point link between an OSPFv2-only router and an OSPFv2 and IS-IS enabled router the following link information may be generated. Note that the IS-IS router also supports the IPv6 traffic engineering extensions RFC 6119 [RFC6119] for IS-IS.

The NRLI encodes local IPv4 router-id, remote IPv4 router-id, remote ISO node-id and remote IPv6 node-id.

### 3.2.2. Link Descriptors

The 'Link Descriptor' field is a set of Type/Length/Value (TLV) triplets. The format of each TLV is shown in Section 3.1. The 'Link descriptor' TLVs uniquely identify a link between a pair of anchor Routers. A link described by the Link descriptor TLVs actually is a "half-link", a unidirectional representation of a logical link. In order to fully describe a single logical link two originating routers need to advertise a half-link each, i.e. two link NLRIs will be advertised.

The format and semantics of the 'value' fields in most 'Link Descriptor' TLVs correspond to the format and semantics of value fields in IS-IS Extended IS Reachability sub-TLVs, defined in [RFC5305], [RFC5307] and [RFC6119]. Although the encodings for 'Link Descriptor' TLVs were originally defined for IS-IS, the TLVs can carry data sourced either by IS-IS or OSPF.

The following link descriptor TLVs are valid in the Link NLRI:

| Type | Description | IS-IS TLV/Sub-TLV | Value defined in: |
|---|---|---|---|
| 263 | Link Local/Remote Identifiers | 22/4 | [RFC5307]/1.1 |
| 264 | IPv4 interface address | 22/6 | [RFC5305]/3.2 |
| 265 | IPv4 neighbor address | 22/8 | [RFC5305]/3.3 |
| 266 | IPv6 interface address | 22/12 | [RFC6119]/4.2 |
| 267 | IPv6 neighbor address | 22/13 | [RFC6119]/4.3 |
| 268 | Multi Topology ID | --- | Section 3.2.2.1 |

Table 2: Link Descriptor TLVs

#### 3.2.2.1. Multi Topology ID TLV

The Multi Topology ID TLV (Type 268) carries the Multi Topology ID for this link. The semantics of the Multi Topology ID are defined in RFC5120, Section 7.2 [RFC5120], and the OSPF Multi Topology ID), defined in RFC4915, Section 3.7 [RFC4915]. If the value in the Multi Topology ID TLV is derived from OSPF, then the upper 9 bits of the Multi Topology ID are set to 0.

```
      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |              Type             |             Length            |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |R R R R|   Multi Topology ID   |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 13: Multi Topology ID TLV format

### 3.2.3. The Prefix NLRI

The Prefix NLRI is a variable length field that contains an IP address prefix (IPv4 or IPv6) originally advertised in the IGP topology. The distinction between IPv4 and IPv6 prefixes is given by the NLRI Type filed in the Link State NLRI. Reachability information is encoded as one or more 2-tuples of the form <length, prefix>, whose fields are described below:

```
                     +---------------------------+
                     |   Length (1 octet)        |
                     +---------------------------+
                     |   Prefix (variable)       |
                     +---------------------------+
```

Figure 14: Prefix NLRI format

### 3.3. The LINK_STATE Attribute

This is an optional, transitive BGP attribute that is used to carry link, node and prefix parameters and attributes. It is defined as a set of Type/Length/Value (TLV) triplets, described in the following section. This attribute SHOULD only be included with Link State NLRIs. This attribute MUST be ignored for all other NLRIs.

### 3.3.1. Link Attribute TLVs

Each 'Link Attribute' is a Type/Length/Value (TLV) triplet formatted as defined in Section 3.1. The format and semantics of the 'value' fields in some 'Link Attribute' TLVs correspond to the format and semantics of value fields in IS-IS Extended IS Reachability sub-TLVs, defined in [RFC5305] and [RFC5307]. Other 'Link Attribute' TLVs are defined in this document. Although the encodings for 'Link Attribute' TLVs were originally defined for IS-IS, the TLVs can carry data sourced either by IS-IS or OSPF.

The following 'Link Attribute' TLVs are are valid in the LINK_STATE attribute:

| Type | Description | IS-IS TLV/Sub-TLV | Defined in: |
|---|---|---|---|
| 269 | Administrative group (color) | 22/3 | [RFC5305]/3.1 |
| 270 | Maximum link bandwidth | 22/9 | [RFC5305]/3.3 |
| 271 | Max. reservable link bandwidth | 22/10 | [RFC5305]/3.5 |
| 272 | Unreserved bandwidth | 22/11 | [RFC5305]/3.6 |
| 273 | Link Protection Type | 22/20 | [RFC5307]/1.2 |
| 274 | MPLS Protocol Mask | --- | Section 3.3.1.1 |
| 275 | Metric | --- | Section 3.3.1.2 |
| 276 | Shared Risk Link Group | --- | Section 3.3.1.3 |
| 277 | OSPF specific link attribute | --- | Section 3.3.1.4 |
| 278 | IS-IS Specific Link Attribute | --- | Section 3.3.1.5 |
| 279 | Area ID | --- | Section 3.3.1.6 |

Table 3: Link Attribute TLVs

#### 3.3.1.1. MPLS Protocol Mask TLV

The MPLS Protocol TLV (Type 274) carries a bit mask describing which MPLS signaling protocols are enabled. The length of this TLV is 1. The value is a bit array of 8 flags, where each bit represents an MPLS Protocol capability.

```
  0                   1                   2                   3
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |              Type             |             Length            |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |L R            |
 +-+-+-+-+-+-+-+-+
```

Figure 15: MPLS Protocol TLV

The following bits are defined:

| Bit | Description | Reference |
|---|---|---|
| 0 | Label Distribution Protocol (LDP) | [RFC5036] |
| 1 | Extension to RSVP for LSP Tunnels (RSVP-TE) | [RFC3209] |
| 2-7 | Reserved for future use | |

Table 4: MPLS Protocol Mask TLV Codes

### 3.3.1.2.  Metric TLV

The IGP Metric TLV (Type 275) carries the metric for this link.  The length of this TLV is 3.  If the length of the metric from which the IGP Metric value is derived is less than 3 (e.g. for OSPF link metrics or non-wide IS-IS metric), then the upper bits of the TLV are set to 0.

```
      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |              Type             |             Length            |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                  IGP Link Metric              |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 16: Metric TLV format

### 3.3.1.3.  Shared Risk Link Group TLV

The Shared Risk Link Group (SRLG) TLV (Type 276) carries the Shared Risk Link Group information (see Section 2.3, "Shared Risk Link Group Information", of [RFC4202]).  It contains a data structure consisting of a (variable) list of SRLG values, where each element in the list has 4 octets, as shown in Figure 17.  The length of this TLV is 4 * (number of SRLG values).

```
      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |             Type              |             Length            |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                  Shared Risk Link Group Value                 |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                          ............                         |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                  Shared Risk Link Group Value                 |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 17: Shared Risk Link Group TLV format

Note that there is no SRLG TLV in OSPF-TE. In IS-IS the SRLG information is carried in two different TLVs: the IPv4 (SRLG) TLV (Type 138) defined in [RFC5307], and the IPv6 SRLG TLV (Type 139) defined in [RFC6119]. Since the Link State NLRI uses variable Router-ID anchoring, both IPv4 and IPv6 SRLG information can be carried in a single TLV.

### 3.3.1.4. OSPF Specific Link Attribute TLV

The OSPF specific link attribute TLV (Type 277) is an envelope that transparently carries optional link properties TLVs advertised by an OSPF router. The value field contains one or more optional OSPF link attribute TLVs. An originating router shall use this TLV for encoding information specific to the OSPF protocol or new OSPF extensions for which there is no protocol neutral representation in the BGP link-state NLRI.

```
      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |             Type              |             Length            |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                                                               |
     |            OSPF specific link attributes (variable)           |
     |                                                               |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 18: OSPF specific link attribute format

### 3.3.1.5. IS-IS specific link attribute TLV

The IS-IS specific link attribute TLV (Type 278) is an envelope that transparently carries optional link properties TLVs advertised by an IS-IS router. The value field contains one or more optional IS-IS

link attribute TLVs. An originating router shall use this TLV for encoding information specific to the IS-IS protocol or new IS-IS extensions for which there is no protocol neutral representation in the BGP link-state NLRI.

```
     0                   1                   2                   3
     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |              Type             |             Length            |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                                                               |
    |           IS-IS specific link attributes (variable)           |
    |                                                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 19: IS-IS specific link attribute format

### 3.3.1.6. Link Area TLV

The Area TLV (Type 279) carries the Area ID which is assigned on this link. If a link is present in more than one Area then several occurrences of this TLV may be generated. Since only the OSPF protocol carries the notion of link specific areas, the Area ID has a fixed length of 4 octets.

```
     0                   1                   2                   3
     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |              Type             |             Length            |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                            Area ID                            |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 20: Link Area TLV format

### 3.3.2. Node Attribute TLVs

The following node attribute TLVs are defined:

| Type | Description | Length |
|---|---|---|
| 280 | Multi Topology | 2 |
| 281 | Node Flag Bits | 1 |
| 282 | OSPF Specific Node Properties | variable |
| 283 | IS-IS Specific Node Properties | variable |
| 284 | Node Area ID | variable |

Table 5: Node Attribute TLVs

### 3.3.2.1. Multi Topology Node TLV

The Multi Topology TLV (Type 280) carries the Multi Topology ID and topology specific flags for this node. The format and semantics of the 'value' field in the Multi Topology TLV is defined in RFC5120, Section 7.1 [RFC5120]. If the value in the Multi Topology TLV is derived from OSPF, then the upper 9 bits of the Multi Topology ID and the 'O' and 'A' bits are set to 0.

```
     0                   1                   2                   3
     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |              Type             |             Length            |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |O A R R|   Multi Topology ID   |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 21: Multi Topology Node TLV format

### 3.3.2.2. Node Flag Bits TLV

The Node Flag Bits TLV (Type 281) carries a bit mask describing node attributes. The value is a bit array of 8 flags, where each bit represents a node capability.

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

Figure 22: Node Flag Bits TLV format

The bits are defined as follows:

| Bit | Description | Reference |
|---|---|---|
| 0 | Overload Bit | [RFC1195] |
| 1 | Attached Bit | [RFC1195] |
| 2 | External Bit | [RFC2328] |
| 3 | ABR Bit | [RFC2328] |

Table 6: Node Flag Bits Definitions

#### 3.3.2.3. OSPF Specific Node Properties TLV

The OSPF Specific Node Properties TLV (Type 282) is an envelope that transparently carries optional node properties TLVs advertised by an OSPF router.  The value field contains one or more optional OSPF node property TLVs, such as the OSPF Router Informational Capabilities TLV defined in [RFC4970], or the OSPF TE Node Capability Descriptor TLV described in [RFC5073].  An originating router shall use this TLV for encoding information specific to the OSPF protocol or new OSPF extensions for which there is no protocol neutral representation in the BGP link-state NLRI.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Type             |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
|            OSPF specific node properties (variable)          |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 23: OSPF specific Node property format

#### 3.3.2.4. IS-IS Specific Node Properties TLV

The IS-IS Router Specific Node Properties TLV (Type 283) is an envelope that transparently carries optional node specific TLVs advertised by an IS-IS router.  The value field contains one or more optional IS-IS node property TLVs, such as the IS-IS TE Node Capability Descriptor TLV described in [RFC5073].  An originating router shall use this TLV for encoding information specific to the IS-IS protocol or new IS-IS extensions for which there is no protocol neutral representation in the BGP link-state NLRI.

```
      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |              Type             |             Length            |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                                                               |
     |           IS-IS specific node properties (variable)           |
     |                                                               |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 24: IS-IS specific Node property format

#### 3.3.2.5.  Area Node TLV

The Area TLV (Type 284) carries the Area ID which is assigned to this node.  If a node is present in more than one Area then several occurrences of this TLV may be generated.  Since only the IS-IS protocol carries the notion of per-node areas, the Area ID has a variable length of 1 to 20 octets.

```
      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |              Type             |             Length            |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                                                               |
     |                       Area ID (variable)                      |
     |                                                               |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 25: Area Node TLV format

### 3.3.3.  Prefix Attributes TLVs

Prefixes are learned from the IGP topology (ISIS or OSPF) with a set of IGP attributes (such as metric, route tags, route type, etc.) that MUST be reflected into the LINK_STATE attribute.  This section describes the different attributes related to the IPv4/IPv6 prefixes. Prefix Attributes TLVs SHOULD be used when advertising NLRI types 3 and 4 only.  The following attributes TLVs are defined:

| Type | Description | Length | Reference |
|---|---|---|---|
| 285 | IGP Flags | 4 | |
| 286 | Route Tag | 4 | [RFC5130] |
| 287 | Extended Tag | 8 | [RFC5130] |
| 288 | Metric | 4 | [RFC5305] |
| 289 | OSPF Forwarding Address | 4 | [RFC2328] |

Table 7: Prefix Attribute TLVs

### 3.3.3.1. IGP Flags TLV

IGP Flags TLV contains ISIS and OSPF flags and bits originally assigned to the prefix.  The IGP Flags TLV is encoded as follows:

```
  0                   1                   2                   3
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |              Type             |             Length            |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |                            IGP Flags                          |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 26: IGP Flag TLV format

where:

Type is 285

Length is 4

The following bits are defined according to the table here below:

| Bit | Description | Reference |
|---|---|---|
| 0 | ISIS Up/Down Bit | [RFC5305] |
| 1-3 | OSPF Route Type | [RFC2328] |
| 4-15 | RESERVED | |

Table 8: IGP Flag Bits Definitions

OSPF Route Type can be either: Intra-Area (0x1), Inter-Area (0x2), External 1 (0x3), External 2 (0x4), NSSA (0x5) and is encoded in a 3 bits number.  For prefixes learned from IS-IS, this field MUST to be

set to 0x0 on transmission.

### 3.3.3.2. Route Tag

Route Tag TLV carries the original IGP TAG (ISIS or OSPF) of the prefix and is encoded as follows:

```
      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |              Type             |             Length            |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                           Route Tag                           |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 27: IGP Route TAG TLV format

where:

Type is 286

Length is 4

Route Tag contains the original tags as learned in the IGP topology.

### 3.3.3.3. Extended Route Tag

Extended Route Tag TLV carries the ISIS Extended Route TAG of the prefix and is encoded as follows:

```
      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |              Type             |             Length            |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                      Extended Route Tag                       |
     |                                                               |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 28: Extended IGP Route TAG TLV format

where:

Type is 287

Length is 8

Extended Route Tag contains the original ISIS Extended Tag as learned

in the IGP topology.

### 3.3.3.4. Prefix Metric TLV

Prefix Metric TLV carries the metric of the prefix as known in the IGP topology.  The attribute is mandatory and can only appear once.

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

Figure 29: Prefix Metric TLV Format

where:

Type is 288

Length is 4

### 3.3.3.5. OSPF Forwarding Address TLV

OSPF Forwarding Address TLV carries the OSPF forwarding address as known in the original OSPF advertisement.  Forwarding address can be either IPv4 or IPv6.

```
      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |              Type             |             Length            |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                 Forwarding Address (variable)                 |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 30: OSPF Forwarding Address TLV Format

where:

Type is 289

Length is 4 for an IPv4 forwarding address an 16 for an IPv6 forwarding address

### 3.4. BGP Next Hop Information

BGP link-state information for both IPv4 and IPv6 networks can be carried over either an IPv4 BGP session, or an IPv6 BGP session. If IPv4 BGP session is used, then the next hop in the MP_REACH_NLRI SHOULD be an IPv4 address. Similarly, if IPv6 BGP session is used, then the next hop in the MP_REACH_NLRI SHOULD be an IPv6 address. Usually the next hop will be set to the local end-point address of the BGP session. The next hop address MUST be encoded as described in [RFC4760]. The length field of the next hop address will specify the next hop address-family. If the next hop length is 4, then the next hop is an IPv4 address; if the next hop length is 16, then it is a global IPv6 address and if the next hop length is 32, then there is one global IPv6 address followed by a link-local IPv6 address. The link-local IPv6 address should be used as described in [RFC2545].

### 3.5. Inter-AS Links

The main source of TE information is the IGP, which is not active on inter-AS links. In order to inject a non-IGP enabled link into the BGP link-state RIB an implementation must support configuration of static links.

## 4. Link to Path Aggregation

Distribution of all links available in the global Internet is certainly possible, however not desirable from a scaling and privacy point of view. Therefore an implementation may support link to path aggregation. Rather than advertising all specific links of a domain, an ASBR may advertise an "aggregate link" between a non-adjacent pair of nodes. The "aggregate link" represents the aggregated set of link properties between a pair of non-adjacent nodes. The actual methods to compute the path properties (of bandwidth, metric) are outside the scope of this document. The decision whether to advertise all specific links or aggregated links is an operator's policy choice. To highlight the varying levels of exposure, the following deployment examples shall be discussed.

### 4.1. Example: No Link Aggregation

Consider Figure 31. Both AS1 and AS2 operators want to protect their inter-AS {R1,R3}, {R2, R4} links using RSVP-FRR LSPs. If R1 wants to compute its link-protection LSP to R3 it needs to "see" an alternate path to R3. Therefore the AS2 operator exposes its topology. All BGP TE enabled routers in AS1 "see" the full topology of AS and therefore can compute a backup path. Note that the decision if the direct link between {R3, R4} or the {R4, R5, R3) path is used is made

by the computing router.

```
          AS1   :   AS2
                :
           R1-------R3
            |   :   | \
            |   :   |  R5
            |   :   | /
           R2-------R4
                :
                :
```

Figure 31: no-link-aggregation

## 4.2.  Example: ASBR to ASBR Path Aggregation

The brief difference between the "no-link aggregation" example and this example is that no specific link gets exposed.  Consider Figure 32.  The only link which gets advertised by AS2 is an "aggregate" link between R3 and R4.  This is enough to tell AS1 that there is a backup path.  However the actual links being used are hidden from the topology.

```
          AS1   :   AS2
                :
           R1-------R3
            |   :   |
            |   :   |
            |   :   |
           R2-------R4
                :
                :
```

Figure 32: asbr-link-aggregation

## 4.3.  Example: Multi-AS Path Aggregation

Service providers in control of multiple ASes may even decide to not expose their internal inter-AS links.  Consider Figure 33.  Rather than exposing all specific R3 to R6 links, AS3 is modeled as a single node which connects to the border routers of the aggregated domain.

```
          AS1   :   AS2   :   AS3
                :         :
           R1-------R3-----
            |   :         : \
            |   :         :   vR0
            |   :         : /
           R2-------R4-----
                :         :
                :         :
```

Figure 33: multi-as-aggregation

## 5. IANA Considerations

This document requests a code point from the registry of Address Family Numbers.

This document requests a code point from the BGP Path Attributes registry.

This document requests creation of a new registry for node anchor, link descriptor and link attribute TLVs. Values 0-255 are reserved. Values 256-65535 will be used for Codepoints. The registry will be initialized as shown in Table 2 and Table 3. Allocations within the registry will require documentation of the proposed use of the allocated value and approval by the Designated Expert assigned by the IESG (see [RFC5226]).

Note to RFC Editor: this section may be removed on publication as an RFC.

## 6. Manageability Considerations

This section is structured as recommended in [RFC5706].

### 6.1. Operational Considerations

#### 6.1.1. Operations

Existing BGP operation procedures apply. No new operation procedures are defined in this document. It shall be noted that the NLRI information present in this document purely carries application level data that have no immediate corresponding forwarding state impact. As such, any churn in reachability information has different impact than regular BGP update which needs to chaange forwarding state for an entire router. Furthermore it is anticipated that distribution of

this NLRI will be handled by dedicated route-reflectors providing a level of isolation and fault-containment between different NLRI types.

#### 6.1.2. Installation and Initial Setup

Configuration parameters defined in Section 6.2.3 SHOULD be initialized to the following default values:

- The Link-State NLRI capability is turned off for all neighbors.

- The maximum rate at which Link State NLRIs will be advertised/ withdrawn from neighbors is set to 200 updates per second.

#### 6.1.3. Migration Path

The proposed extension is only activated between BGP peers after capability negotiation. Moreover, the extensions can be turned on/ off an individual peer basis (see Section 6.2.3), so the extension can be gradually rolled out in the network.

#### 6.1.4. Requirements on Other Protocols and Functional Components

The protocol extension defined in this document does not put new requirements on other protocols or functional components.

#### 6.1.5. Impact on Network Operation

Frequency of Link-State NLRI updates could interfere with regular BGP prefix distribution. A network operator MAY use a dedicated Route-Reflector infrastructure to distribute Link-State NLRIs.

Distribution of Link-State NLRIs SHOULD be limited to a single admin domain, which can consist of multiple areas within an AS or multiple ASes.

#### 6.1.6. Verifying Correct Operation

Existing BGP procedures apply. In addition, an implementation SHOULD allow an operator to:

- List neighbors with whom the Speaker is exchanging Link-State NLRIs

### 6.2. Management Considerations

#### 6.2.1. Management Information

#### 6.2.2. Fault Management

TBD.

#### 6.2.3. Configuration Management

An implementation SHOULD allow the operator to specify neighbors to which Link-State NLRIs will be advertised and from which Link-State NLRIs will be accepted.

An implementation SHOULD allow the operator to specify the maximum rate at which Link State NLRIs will be advertised/withdrawn from neighbors

An implementation SHOULD allow the operator to specify the maximum rate at which Link State NLRIs will be accepted from neighbors

An implementation SHOULD allow the operator to specify the maximum number of Link State NLRIs stored in router's RIB.

An implementation SHOULD allow the operator to create abstracted topologies that are advertised to neighbors; Create different abstractions for different neighbors.

#### 6.2.4. Accounting Management

Not Applicable.

#### 6.2.5. Performance Management

An implementation SHOULD provide the following statistics:

- Total number of Link-State NLRI updates sent/received

- Number of Link-State NLRI updates sent/received, per neighbor

- Number of errored received Link-State NLRI updates, per neighbor

- Total number of locally originated Link-State NLRIs

### 6.2.6. Security Management

An operator SHOULD define ACLs to limit inbound updates as follows:

- Drop all updates from Consumer peers

## 7. Security Considerations

Procedures and protocol extensions defined in this document do not affect the BGP security model.

A BGP Speaker SHOULD NOT accept updates from a Consumer peer.

An operator SHOULD employ a mechanism to protect a BGP Speaker against DDOS attacks from Consumers.

## 8. Acknowledgements

We would like to thank Nischal Sheth, Alia Atlas, Robert Varga, David Ward, Derek Yeung, Murtuza Lightwala, John Scudder, Kaliraj Vairavakkalai, Les Ginsberg, Liem Nguyen, Manish Bhardwaj, Mike Shand, Peter Psenak, Rex Fernando, Richard Woundy, Saikat Ray, Steven Luong, Tamas Mondal, Waqas Alam, and Yakov Rekhter for their comments.

## 9. References

### 9.1. Normative References

[RFC1195]  Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, December 1990.

[RFC1918]  Rekhter, Y., Moskowitz, R., Karrenberg, D., Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, February 1996.

[RFC2119]  Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC2328]  Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.

[RFC2545]  Marques, P. and F. Dupont, "Use of BGP-4 Multiprotocol Extensions for IPv6 Inter-Domain Routing", RFC 2545, March 1999.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

[RFC4202] Kompella, K. and Y. Rekhter, "Routing Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4202, October 2005.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

[RFC4893] Vohra, Q. and E. Chen, "BGP Support for Four-octet AS Number Space", RFC 4893, May 2007.

[RFC4915] Psenak, P., Mirtorabi, S., Roy, A., Nguyen, L., and P. Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF", RFC 4915, June 2007.

[RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

[RFC5065] Traina, P., McPherson, D., and J. Scudder, "Autonomous System Confederations for BGP", RFC 5065, August 2007.

[RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, February 2008.

[RFC5130] Previdi, S., Shand, M., and C. Martin, "A Policy Control Mechanism in IS-IS Using Administrative Tags", RFC 5130, February 2008.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.

[RFC5307] Kompella, K. and Y. Rekhter, "IS-IS Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 5307, October 2008.

[RFC6119] Harrison, J., Berger, J., and M. Bartlett, "IPv6 Traffic

Engineering in IS-IS", RFC 6119, February 2011.

### 9.2. Informative References

[I-D.ietf-alto-protocol]
Alimi, R., Penno, R., and Y. Yang, "ALTO Protocol", draft-ietf-alto-protocol-13 (work in progress), September 2012.

[I-D.ietf-isis-mi]
Previdi, S., Ginsberg, L., Shand, M., Roy, A., and D. Ward, "IS-IS Multi-Instance", draft-ietf-isis-mi-08 (work in progress), October 2012.

[RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.

[RFC4970] Lindem, A., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 4970, July 2007.

[RFC5073] Vasseur, J. and J. Le Roux, "IGP Routing Protocol Extensions for Discovery of Traffic Engineering Node Capabilities", RFC 5073, December 2007.

[RFC5152] Vasseur, JP., Ayyangar, A., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, February 2008.

[RFC5693] Seedorf, J. and E. Burger, "Application-Layer Traffic Optimization (ALTO) Problem Statement", RFC 5693, October 2009.

[RFC5706] Harrington, D., "Guidelines for Considering Operations and Management of New Protocols and Protocol Extensions", RFC 5706, November 2009.

[RFC6549] Lindem, A., Roy, A., and S. Mirtorabi, "OSPFv2 Multi-Instance Extensions", RFC 6549, March 2012.

## Authors' Addresses

Hannes Gredler
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA  94089
US

Email: hannes@juniper.net

Jan Medved
Cisco Systems, Inc.
170, West Tasman Drive
San Jose, CA  95134
US

Email: jmedved@cisco.com

Stefano Previdi
Cisco Systems, Inc.
Via Del Serafico, 200
Rome  00142
Italy

Email: sprevidi@cisco.com

Adrian Farrel
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA  94089
US

Email: afarrel@juniper.net

Saikat Ray
Cisco Systems, Inc.
170, West Tasman Drive
San Jose, CA  95134
US

Email: sairay@cisco.com