

InterDomain Routing Group (IDR)  
Internet-Draft  
Updates: 4271 (if approved)  
Intended status: Standards Track  
Expires: August 29, 2013

S. Hares  
Huawei Technologies (USA)  
J. Scudder  
Juniper Networks  
February 25, 2013

Update Attribute Flag Low Bits Clarification  
draft-hares-idr-update-attrib-low-bits-fix-01

Abstract

This draft provides an update to RFC 4721 to clarify the use of the lower-order four bits of the Attribute flag in the Update message.

Status of this Memo

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

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

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

This Internet-Draft will expire on August 29, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

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

## Table of Contents

1. Introduction . . . . .	3
2. Change to RFC 4271 Section 4.3 . . . . .	3
3. Known BGP Implementation Habits . . . . .	4
4. IANA Considerations . . . . .	4
5. Security Considerations . . . . .	4
6. Normative References . . . . .	4
Authors' Addresses . . . . .	5

## 1. Introduction

[RFC4271] specifies in section 4.3 that that the low order four bits of the the Attribute Flags octet are unused, and MUST be zero when sent. There is a disagreement on what when sent means. This draft specifies the meaning.

The issue has been that one school of thought considers that "when sent" means when originated. Another holds that "when sent" means when originated or propagated. This draft takes the second approach of "when sent" being when originated or propagated.

The real issue is that reserved flags are only useful if there is some hope of someday using them for something. If implementations reset these flags on propagation, then a future revision to the BGP specification which introduces a new flag will not be able to propagate the new attribute flag end to end, since it would be very likely that some well-meaning intermediate router would zero on it. The effort to roll out implementations that transited the new flag would almost certainly be prohibitive.

## 2. Change to RFC 4271 Section 4.3

```

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+---+---+---+---+---+---+---+---+
| Attr. Flags |Attr. Type Code|
+---+---+---+---+---+---+---+---+

```

### Original Text:

The lower-order four bits of the Attribute Flags octet are unused. They MUST be zero when sent and MUST be ignored when received

### Corrected Text:

The lower-order four bits of the Attribute Flags octet are unused. They MUST be zero when originated or sent. When received, any MUST be accepted and ignored.

### 3. Known BGP Implementation Habits

The following are BGP implementation habits regarding the unused flag bits

- o always ignore bits received, and always send zero (originated or propagated);
- o always ignore bits received, always send zero bits (originated), and propagate what was received;
- o if non-zero bits are received, drop the peering session;
- o by special condition (policy) handle set bits or set bits, and propagate;and
- o always sets bits under special conditions, and propagates bits.

The reset of BGP sessions based on non-zero bits has been documented at:

<http://mailman.nanog.org/pipermail/nanog/2012-November/053754.html>

Compliance with this draft, as well as [RFC4271], means that routers should not reset BGP sessions if if non-zero lower bits are received.

### 4. IANA Considerations

This document includes no request to IANA.

### 5. Security Considerations

This document has no new security cases.

It clarifies some BGP UPDATE packet flag values and thus may aid in improving BGP security. In particular, it makes it even clearer that routers must not reset a session upon receiving unexpected flag values. Behaving otherwise exposes a router to a denial-of-service attack since a distant party might be able to inject such flag values.

### 6. Normative References

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

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

Authors' Addresses

Susan Hares  
Huawei Technologies (USA)  
2330 Central Expressway  
Santa Clara, CA 95050  
USA

Email: Susan.Hares@huawei.com

John Scudder  
Juniper Networks  
1194 N. Mathilda Ave  
Sunnyvale, CA 94089  
USA

Email: jgs@juniper.net



Inter-Domain Routing  
Internet-Draft  
Intended status: Standards Track  
Expires: August 28, 2013

H. Gredler  
Juniper Networks, Inc.  
J. Medved  
S. Previdi  
Cisco Systems, Inc.  
A. Farrel  
Juniper Networks, Inc.  
S. Ray  
Cisco Systems, Inc.  
February 24, 2013

North-Bound Distribution of Link-State and TE Information using BGP  
draft-ietf-idr-ls-distribution-02

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

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

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

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

This Internet-Draft will expire on August 28, 2013.

#### Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

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



## Table of Contents

1. Introduction . . . . .	4
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. Identifier TLV . . . . .	12
3.2.2. Node Descriptors . . . . .	14
3.2.3. Link Descriptors . . . . .	22
3.2.4. Prefix Descriptors . . . . .	23
3.3. The LINK_STATE Attribute . . . . .	23
3.3.1. Link Attribute TLVs . . . . .	24
3.3.2. Node Attribute TLVs . . . . .	27
3.3.3. Prefix Attributes TLVs . . . . .	29
3.4. BGP Next Hop Information . . . . .	33
3.5. Inter-AS Links . . . . .	33
4. Link to Path Aggregation . . . . .	33
4.1. Example: No Link Aggregation . . . . .	34
4.2. Example: ASBR to ASBR Path Aggregation . . . . .	34
4.3. Example: Multi-AS Path Aggregation . . . . .	35
5. IANA Considerations . . . . .	35
6. Manageability Considerations . . . . .	35
6.1. Operational Considerations . . . . .	35
6.1.1. Operations . . . . .	36
6.1.2. Installation and Initial Setup . . . . .	36
6.1.3. Migration Path . . . . .	36
6.1.4. Requirements on Other Protocols and Functional Components . . . . .	36
6.1.5. Impact on Network Operation . . . . .	36
6.1.6. Verifying Correct Operation . . . . .	37
6.2. Management Considerations . . . . .	37
6.2.1. Management Information . . . . .	37
6.2.2. Fault Management . . . . .	37
6.2.3. Configuration Management . . . . .	37
6.2.4. Accounting Management . . . . .	37
6.2.5. Performance Management . . . . .	37
6.2.6. Security Management . . . . .	38
7. TLV/SubTLV Code Points Summary . . . . .	38
8. Security Considerations . . . . .	40
9. Contributors . . . . .	40
10. Acknowledgements . . . . .	40
11. References . . . . .	40
11.1. Normative References . . . . .	40
11.2. Informative References . . . . .	42
Authors' Addresses . . . . .	43

## 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.

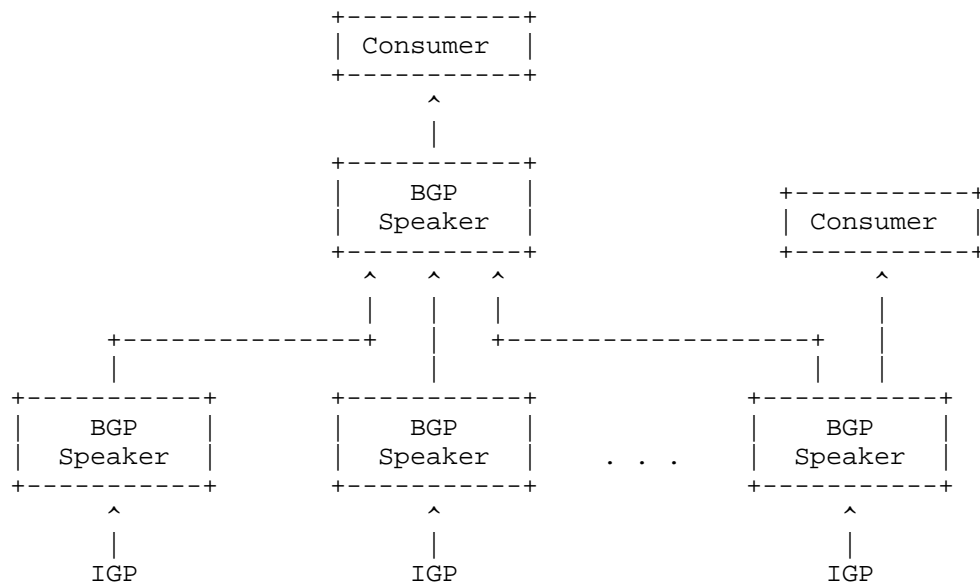


Figure 1: TE Link State info collection

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 from 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 use 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).

- o 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.
- o 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.

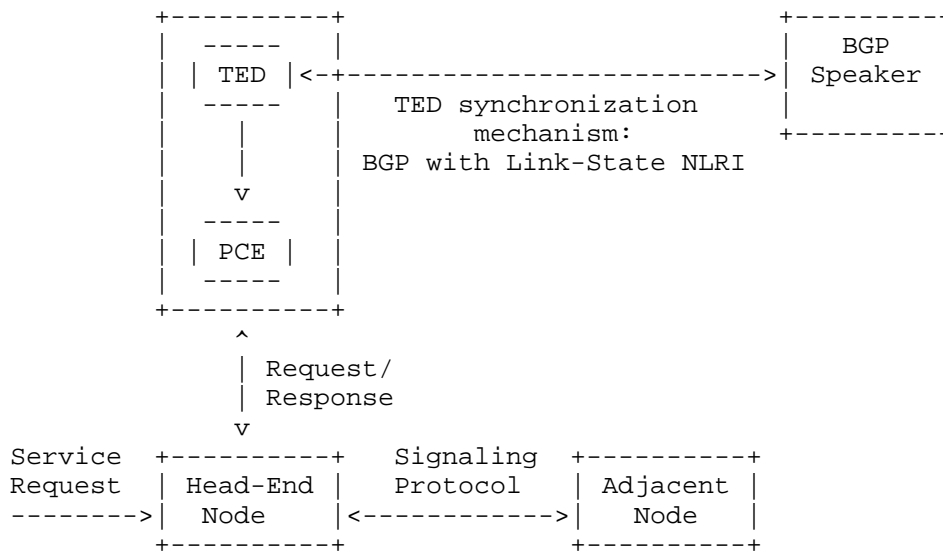


Figure 2: External PCE node using a TED synchronization mechanism

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 Partition Identifiers (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.

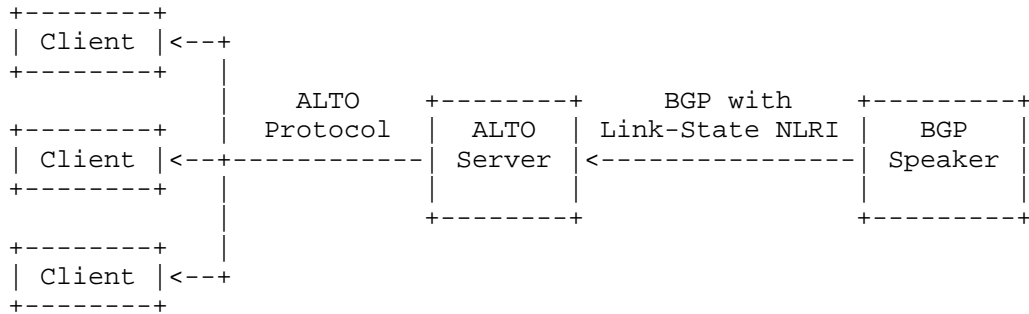


Figure 3: ALTO Server using network topology information

### 3. Carrying Link State Information in BGP

This specification contains two parts: definition of a new BGP NLRI that describes links, nodes and prefixes comprising IGP link state information, and definition of a new BGP path attribute that carries link, node and prefix properties and attributes, such as the link and prefix 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.

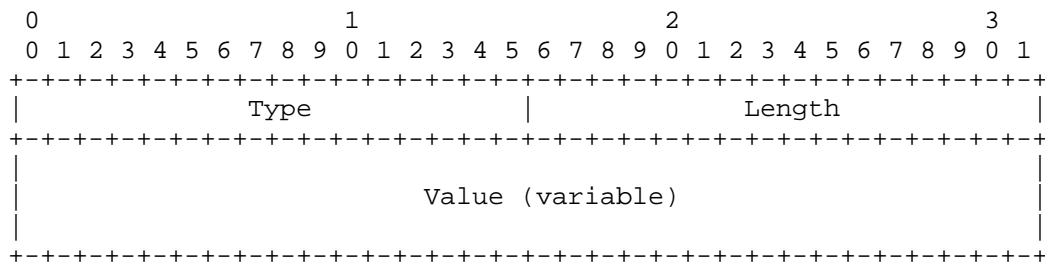


Figure 4: TLV format

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 node, a link or a prefix.

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.

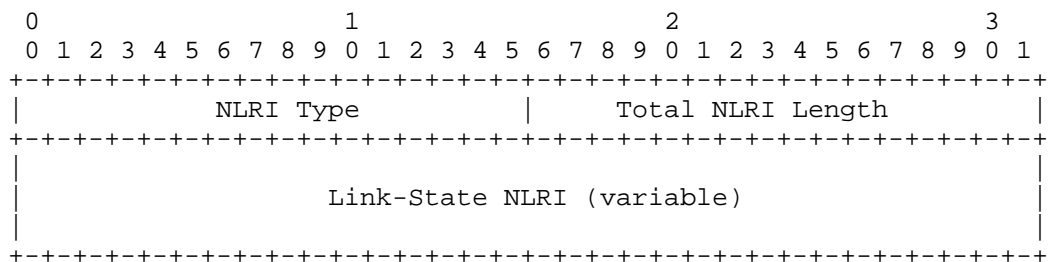


Figure 5: Link State SAFI (TBD) NLRI Format

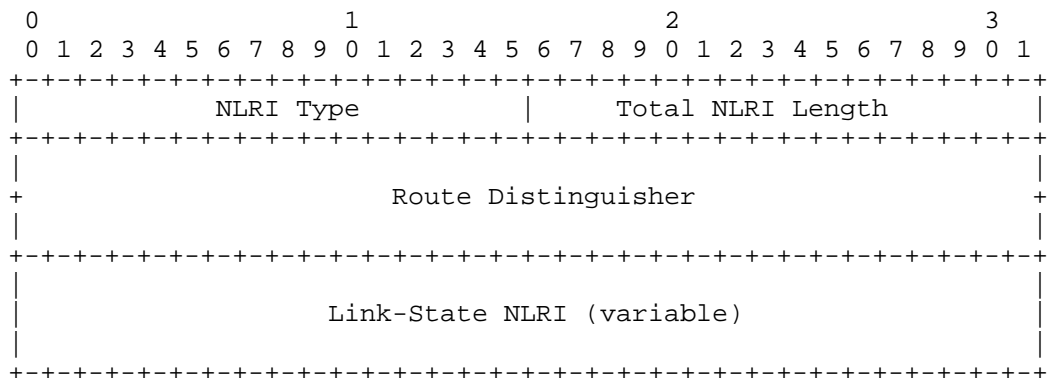


Figure 6: Link State SAFI 128 NLRI Format

The 'Total NLRI Length' field contains the cumulative length of rest of the NLRI not including the NLRI Type field or itself. 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.

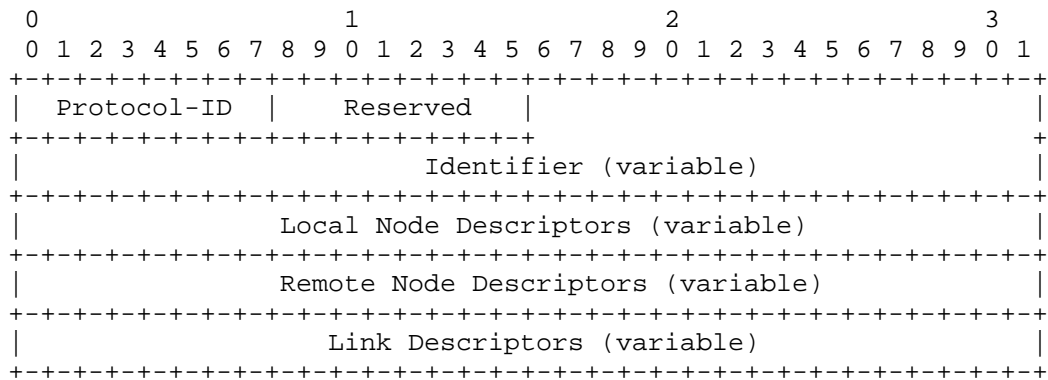


Figure 7: The Link NLRI format

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

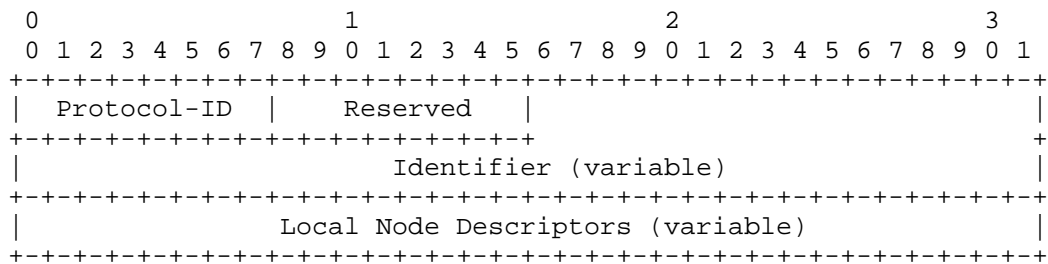


Figure 8: The Node NLRI format

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



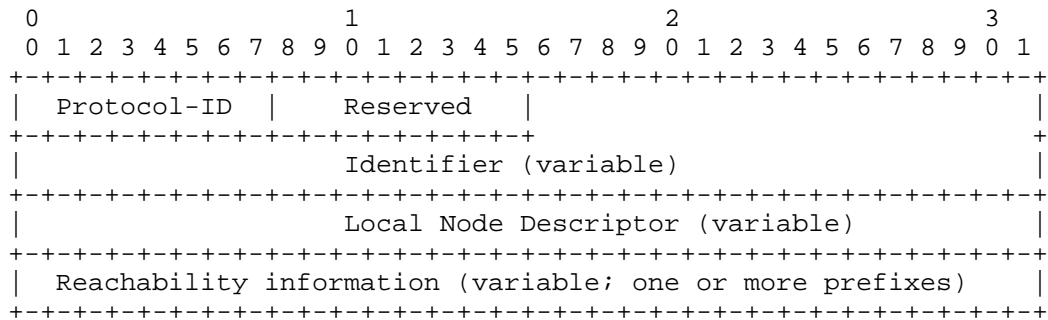


Figure 9: The IPv4/IPv6 Topology Prefix NLRI format

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

Protocol-ID = 2: 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

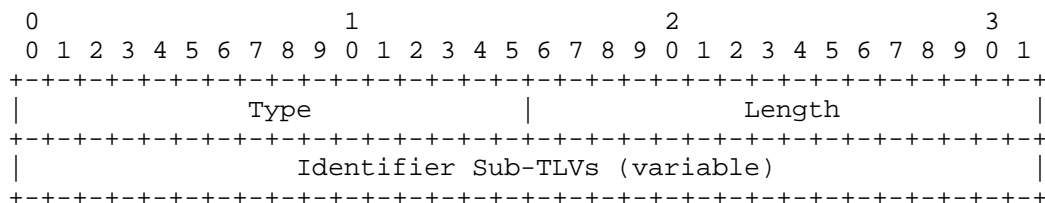
Both OSPF and IS-IS may run multiple routing protocol instances over the same link. See [RFC6822] and [RFC6549].

Identifier TLV is a mandatory TLV containing identifiers of the NLRI and used to associate the NLRI to an instance, a domain, an area or a prefix.

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. Identifier TLV

Identifier TLV (Type 256) is a mandatory TLV that appear in Node, Link and Prefix NLRIs. Identifier TLV carries all identifiers associated with the NLRI in a SubTLV format. Possible Sub TLVs are Instance Identifier, Domain Identifier, Area Identifier, OSPF Route Type and Multi-Topology ID.



Where:

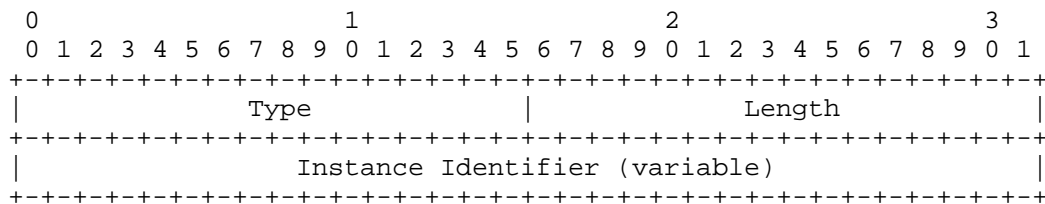
```
Type: 256
Length: variable
Identifier Sub-TLVs: Identifiers
```

Figure 10: Identifier TLV Format

An Identifier may be used to distinguish a Node, a Link or a Prefix with different types of identifiers. Therefore different SubTLVs are defined here below in order to address the different requirements.

#### 3.2.1.1. Instance Identifier SubTLV

Instance Identifier is a mandatory SubTLV that MUST be present in all NLRI's. It is used to identify the topology instance the content of the NLRI and attributes refers to.



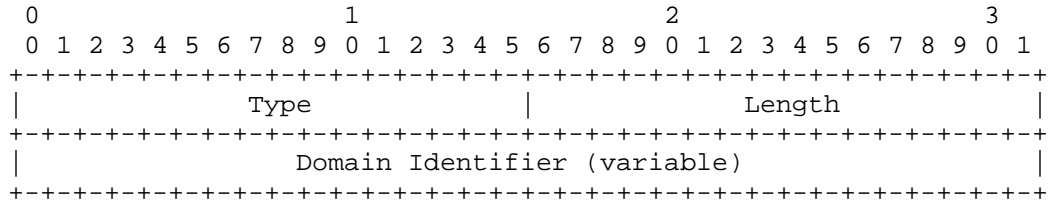
Where:

```
Type: 1
Length: variable
```

Figure 11: Instance Identifier Sub-TLV Format

### 3.2.1.2. Domain Identifier SubTLV

Domain Identifier is an optional SubTLV that MAY be present in all NLRIs. It is used to identify the domain (or sub-domain) to which the NLRI belongs.



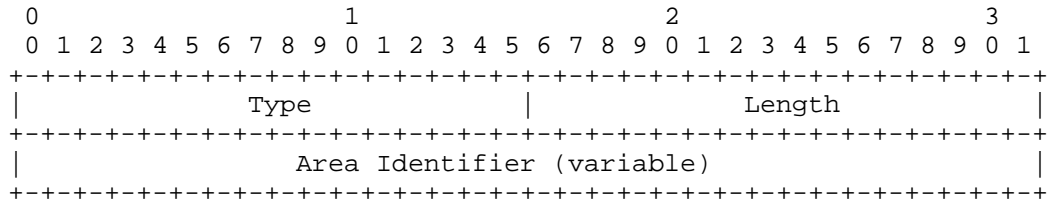
Where:

Type: 2  
Length: variable

Figure 12: Domain Identifier Sub-TLV Format

### 3.2.1.3. Area Identifier SubTLV

Area Identifier is an optional SubTLV that MAY be present in all NLRIs. It is used to identify the area to which the NLRI belongs. Example: an OSPF ABR router advertises itself multiple time (one for each area it participates into). Area Identifier allows the different NLRIs of the same router to be discriminated.



Where:

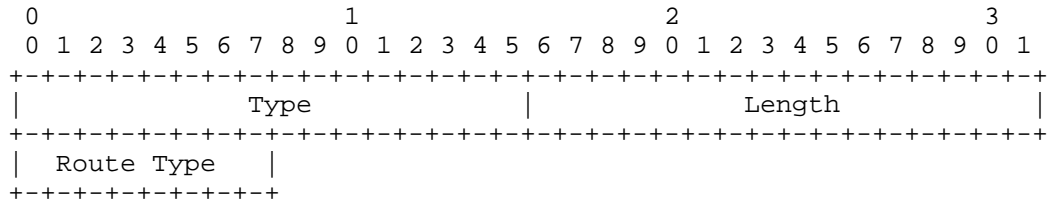
Type: 3  
Length:variable

Figure 13: Area Identifier Sub-TLV Format

### 3.2.1.4. OSPF Route Type SubTLV

Route Type is an optional SubTLV that MAY be present in the Prefix NLRIs. It is used to identify the OSPF route-type of the prefix. It is used when an OSPF prefix is advertised in the OSPF domain with multiple different route-types. The Route Type Identifier allows to

discriminate these advertisements.



Where:

Type: 4  
Length: 1

Figure 14: OSPF Route Type Sub-TLV Format

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.

#### 3.2.1.5. Multi Topology ID SubTLV

The Multi Topology ID SubTLV (type: 5) carries the Multi Topology ID for the link, node or prefix. 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.

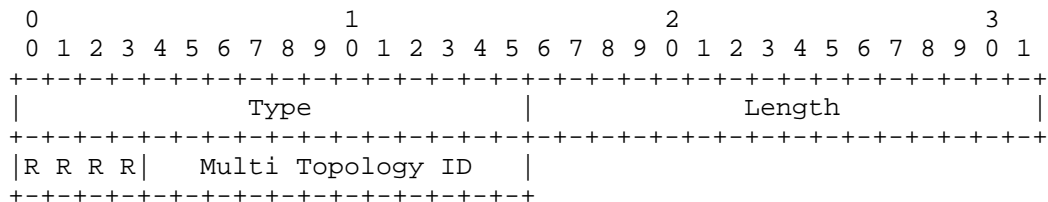


Figure 15: Multi Topology ID SubTLV format

The Multi Topology Identifier SubTLV is present in any NLRI Type.

#### 3.2.2. 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. We use AS Number (or Confederation ID) and BGP Identifier in order to disambiguate the Router-IDs, as described in Section 3.2.2.4.

### 3.2.2.1. Local Node Descriptors

The Local Node Descriptors TLV (Type 257) 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 3.2.2.3.

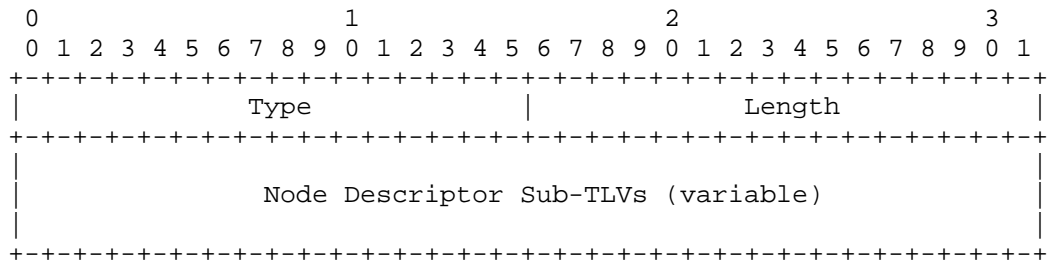


Figure 16: Local Node Descriptors TLV format

### 3.2.2.2. Remote Node Descriptors

The Remote Node Descriptors TLV (Type 258) 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 3.2.2.3.

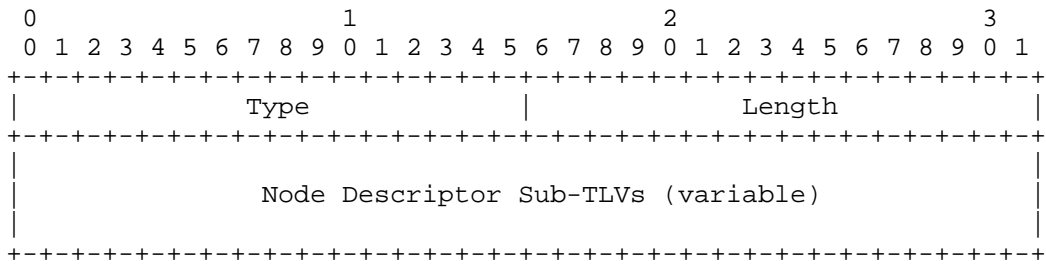


Figure 17: Remote Node Descriptors TLV format

### 3.2.2.3. Node Descriptor Sub-TLVs

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

TLV/SubTLV	Description	Length
259	Autonomous System	4
260	BGP Identifier	4
261	ISO Node-ID	7
262	IPv4 Router-ID	variable
263	IPv6 Router-ID	16

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 Number)

BGP-Identifier: opaque value (32 Bit AS ID); uniquely identifying the BGP-LS speaker within an AS.

IPv4 Router ID: opaque value (can be an IPv4 address or an 32 Bit router ID). When encoding an OSPF Designated Router ID, the length is 8 (first 4 bytes is the Router-ID originating the Type-2 LSA and next 4 bytes are taken from the Type-2 LSA ID). In other cases, the length is 4.

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 259 and 260 specify administrative context in which TLVs 261-263 are to be evaluated. The first TLV from range 261-263 is to be interpreted as the primary node identifier 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.

#### 3.2.2.4. Globally Unique BGP-LS Identifiers

One problem that needs to be addressed is the ability to identify an IGP node globally (by "global", we mean within the BGP-LS database collected by all BGP-LS speakers that talk to each other). This can be expressed through the following two requirements:

(A) The same node must not be represented by two keys (otherwise one node will look like two nodes).

(B) Two different nodes must not be represented by the same key (otherwise, two nodes will look like one node).

We define an "IGP domain" to be the set of nodes (and links), within which, each node has a unique IGP representation by using the combination of area-id, IGP router-id, Level, instance ID, etc. The problem is that BGP brings nodes from multiple independent "IGP domains" and we need to distinguish between them. Moreover, we can't assume there is always one and only one IGP domain per Autonomous System (or Autonomous System confederation member). Following cases illustrate scenario's where IGP domain and ASs boundaries do not match.

(i) Stub ASs or non-contiguous AS: One can have an AS that has disjoint parts, each running an independent IGP domain.

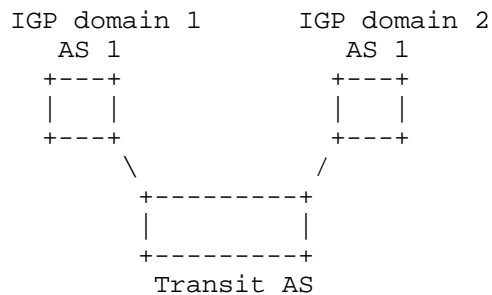


Figure 18: Stub-ASs or non-contiguous AS

Using ASN to globally identify IGP node may break requirement (B).

(ii) It is possible to run the same IGP domain across multiple AS.

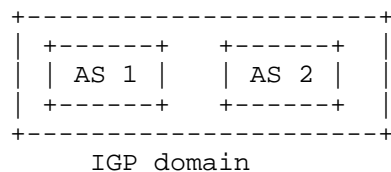


Figure 19: IGP Domain

Using ASN to globally identify IGP node will break requirement (A).

(iii) It is possible to run IGP across member-ASs in a confederation.

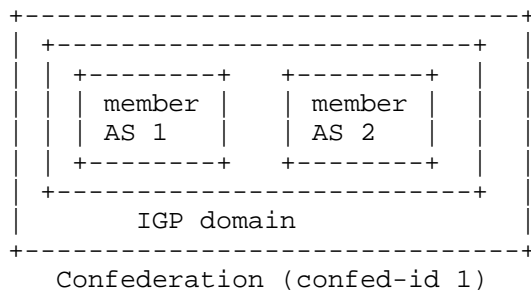


Figure 20: Confederation

Using a Confederation/MemberAS identifier to globally identify IGP node will break requirement (A).

(iv) It is possible to run more than one IGP domain within an AS by setting up "transit BGP speakers".



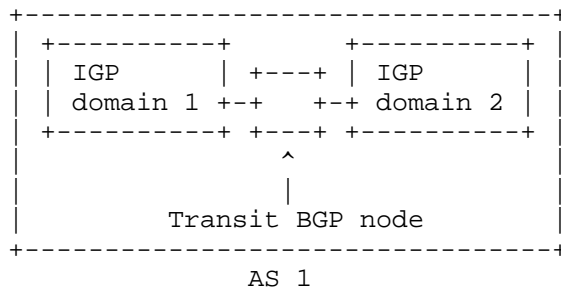


Figure 21: Transit BGP Node

Using ASN to globally identify IGP node may break requirement (A).

In summary, there is no strict relation between BGP AS division and IGP domains. Therefore, the following mechanism is proposed to address the requirements. We assume that a BGP-LS speaker is collocated with one and only one IGP node. The BGP-LS speaker originates BGP-LS NLRIs that correspond to the objects in the LSDB of that IGP node.

We embed a "string" (identifier) in the node descriptor to globally identify the node. The question is how we construct such a string, and what should be the scope of such a string so that the construction of the string can be simple. Let the set of IGP nodes within which LSA/LSP flooding is limited to be the "flooding set". Consider a given "flooding set". We have the following three possibilities:

Case a) There is no BGP LS speaker running on any node in the flooding set.

Case b) There is one BGP LS speaker running on one node in the flooding set.

Case c) There is more than one BGP LS speakers running on the nodes in the flooding set.

For Case a), the nodes in that flooding set do not appear in BGP LS database. So we can ignore that case for this discussion. To satisfy requirement (B), the string we use in different IGP domains must be different. One possible approach is as follows:

Approach 1) The user configures a unique "string" on all BGP LS speakers within one IGP domain.

Now we make an observation that simplifies the task: it is sufficient

to have a unique "string" per flooding set.

When we have a unique string per flooding set, then two nodes in different IGP domains, which by definition belong to different flooding sets, would have different "strings". So requirement B) is satisfied. On the other hand, a given node appears only in the LSDB of the nodes in the same flooding set. So a given node will always have only one "string" and we satisfy requirement A). Given this, we have:

Approach 2) Each BGP LS speaker uses the <Autonomous System Number, BGP Identifier> as the string.

The combination of <Autonomous System, BGP Identifier> is globally unique, as per [RFC6286].

For Case b), which is the simplest BGP-LS deployment scenario, this approach requires no additional configuration from the user.

For Case c), however, if each BGP-LS speaker in the given flooding set attaches its own <Autonomous System, BGP Identifier>, then we will violate requirement A). So that case, the user needs to choose one of the BGP-LS speakers in the flooding set as the "chosen speaker" and configure the rest of the BGP-LS speakers in that flooding set to use the <Autonomous System, BGP Identifier> combination of the "chosen speaker".

When an IGP node belongs to two or more flooding sets, it views itself as a collocation of one node per flooding set and accordingly encodes the NLRIs. Consider the following example:

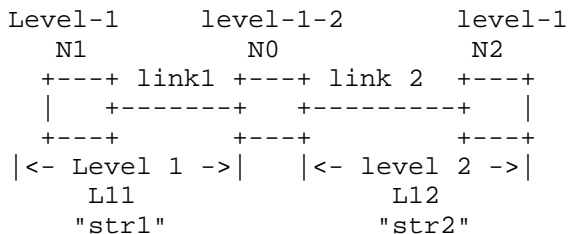


Figure 22: IGP Node in multiple flooding sets

The node N0 is a level 1-2 node. Link1 belongs to level 1 area L11, which has string "str1". Link2 belongs to level 1 area L12 which has string "str2". N0 has both link1 and link2 in its LSDB. If BGP LS speaker is running on N0, then N0 views itself as a collocation of two nodes: N0(L11) and N0(L12) and originate <str1, N1, N0> and <str2, N0, N2>.

To sum up, the mechanism works as follows:

1. We use <Autonomous System, BGP Identifier> as the disambiguating string.
2. By default, a BGP-LS speaker uses its own ASN, BGP identifier (router-id) for these fields for the NLRIs it originates.
3. Operator has the ability to configure other <ASN, BGP ID> per flooding set the IGP node underneath belongs to. In that case, the node descriptor(s) for a given NLRI uses the string corresponding to the flooding set where the node belongs.

The operator needs to provide the configuration if there are multiple BGP-LS speakers running in the same flooding set.

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

IS-IS Pseudonodes are a good example for the variable Router-ID anchoring. Consider Figure 23. 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 NLRI 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.

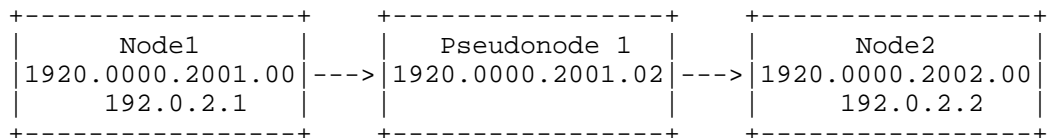


Figure 23: IS-IS Pseudonodes

#### 3.2.2.6. 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.

### 3.2.3. 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 NLRI's 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:

TLV/SubTLV	Description	IS-IS TLV/Sub-TLV	Value defined in:
264	Link Local/Remote Identifiers	22/4	[RFC5307]/1.1
265	IPv4 interface address	22/6	[RFC5305]/3.2
266	IPv4 neighbor address	22/8	[RFC5305]/3.3
267	IPv6 interface address	22/12	[RFC6119]/4.2
268	IPv6 neighbor address	22/13	[RFC6119]/4.3
256/5	Multi Topology ID	---	Section 3.2.1.5

Table 2: Link Descriptor TLVs

## 3.2.4. Prefix Descriptors

The 'Prefix descriptor' TLVs uniquely identify a Prefix (IPv4 or IPv6) originated by a Node.

The following Prefix descriptor TLVs are valid in the IPv4/IPv6 Prefix NLRI:

TLV/SubTLV	Description	IS-IS TLV/Sub-TLV	Value defined in:
256/5	Multi Topology ID	---	Section 3.2.1.5

Table 3: Prefix Descriptor TLVs

## 3.2.4.1. The Prefix NLRI

The Prefix NLRI is a variable length field that contains one or more IP address prefixes (IPv4 or IPv6) originally advertised in the IGP topology. The NLRI Type determines the address-family. 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 24: Prefix NLRI format

The 'Length' field contains the length of the prefix in bits. Only the most significant octets of the prefix are encoded. I.e. 1 octet for prefix length 1 up to 8, 2 octets for prefix length 9 to 16, 3 octets for prefix length 17 up to 24 and 4 octets for prefix length 25 up to 32, etc.

## 3.3. The LINK\_STATE Attribute

This is an optional, non-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 valid in the LINK\_STATE attribute:

TLV/SubTLV	Description	IS-IS TLV/Sub-TLV	Defined in:
256/3	Area Identifier	---	Section 3.2.1.3
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	TE Default Metric	22/18	[RFC5305]/3.7
274	Link Protection Type	22/20	[RFC5307]/1.2
275	MPLS Protocol Mask	---	Section 3.3.1.1
276	Metric	---	Section 3.3.1.2
277	Shared Risk Link Group	---	Section 3.3.1.3
278	OSPF specific link attribute	---	Section 3.3.1.4
279	IS-IS Specific Link Attribute	---	Section 3.3.1.5

Table 4: Link Attribute TLVs

## 3.3.1.1. MPLS Protocol Mask TLV

The MPLS Protocol TLV (Type 275) 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.

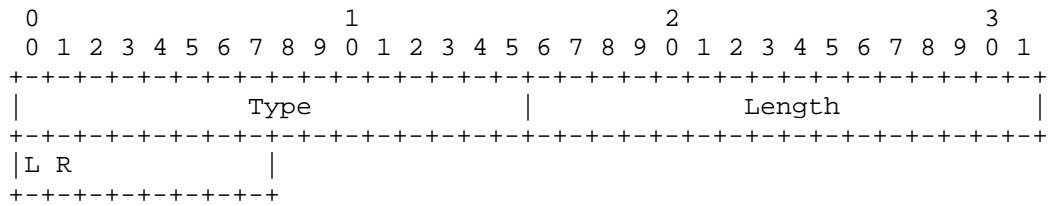


Figure 25: MPLS Protocol TLV

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 5: MPLS Protocol Mask TLV Codes

#### 3.3.1.2. Metric TLV

The IGP Metric TLV (Type 276) 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.

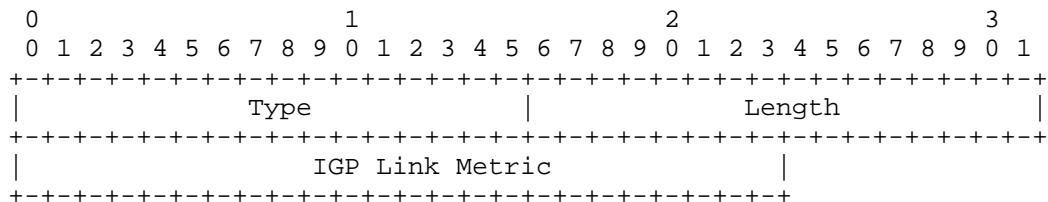


Figure 26: Metric TLV format

#### 3.3.1.3. Shared Risk Link Group TLV

The Shared Risk Link Group (SRLG) TLV (Type 277) 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 27. The length of this TLV is 4 \* (number of SRLG values).

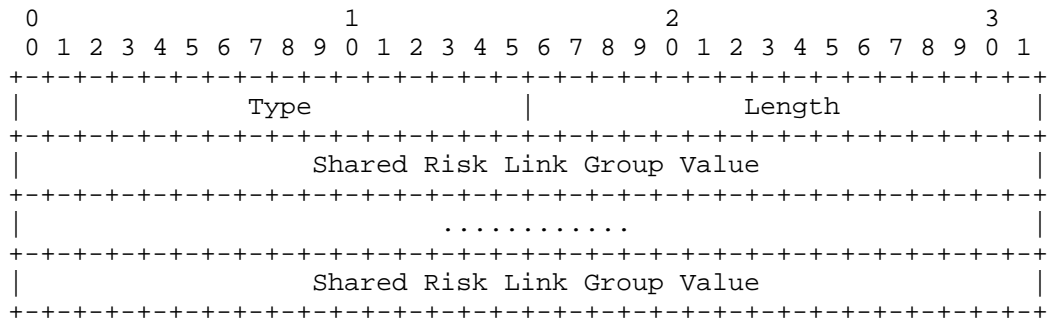


Figure 27: Shared Risk Link Group TLV format

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 278) 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.

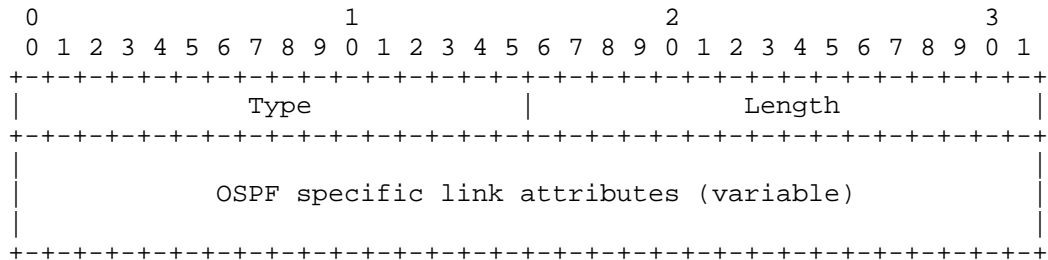


Figure 28: OSPF specific link attribute format

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

The IS-IS specific link attribute TLV (Type 279) 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.

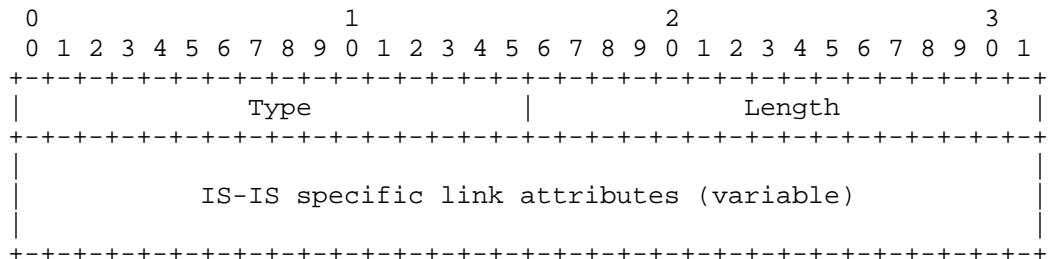


Figure 29: IS-IS specific link attribute format

#### 3.3.1.6. IS-IS Area Address attribute TLV

The area address is carried in the Area Identifier SubTLV of the Identifier TLV and consists of the Area Address which is assigned to the link. If more than one Area Addresses are present, only the lower address is encoded. Note that the Area Identifier SubTLV may appear in all NLRI types (Link, Node and Prefix) and is defined in Section 3.2.1.3.

#### 3.3.2. Node Attribute TLVs

The following node attribute TLVs are defined:

TLV/SubTLV	Description	Length
256/5	Multi Topology	2
280	Node Flag Bits	1
281	OSPF Specific Node Properties	variable
282	IS-IS Specific Node Properties	variable
256	IS-IS Area Address/Domain Identifier	variable

Table 6: Node Attribute TLVs

##### 3.3.2.1. Node Multi Topology ID

The Node Multi Topology ID is carried in the Multi Topology ID SubTLV (type 5) of Identifier ID TLV TLV (Type 256) and 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 Section 3.2.1.5. 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.

### 3.3.2.2. Node Flag Bits TLV

The Node Flag Bits TLV (Type 280) carries a bit mask describing node attributes. The value is a variable length bit array of flags, where each bit represents a node capability.

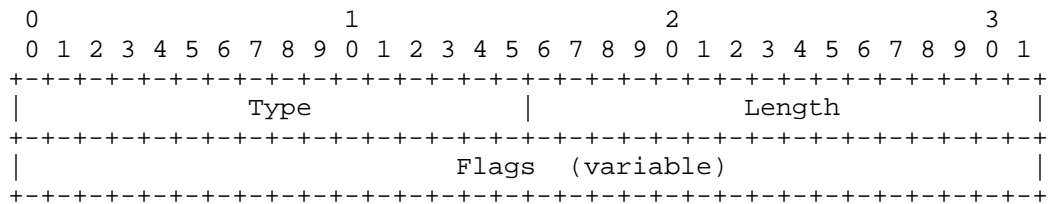


Figure 30: Node Flag Bits TLV format

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 7: Node Flag Bits Definitions

### 3.3.2.3. OSPF Specific Node Properties TLV

The OSPF Specific Node Properties TLV (Type 281) 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.

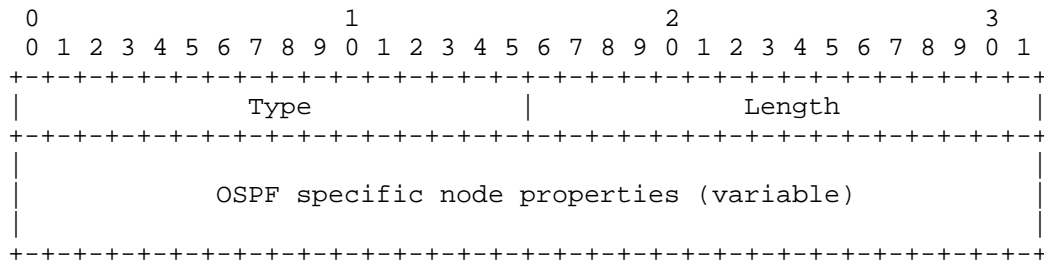


Figure 31: OSPF specific Node property format

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

The IS-IS Router Specific Node Properties TLV (Type 282) 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.

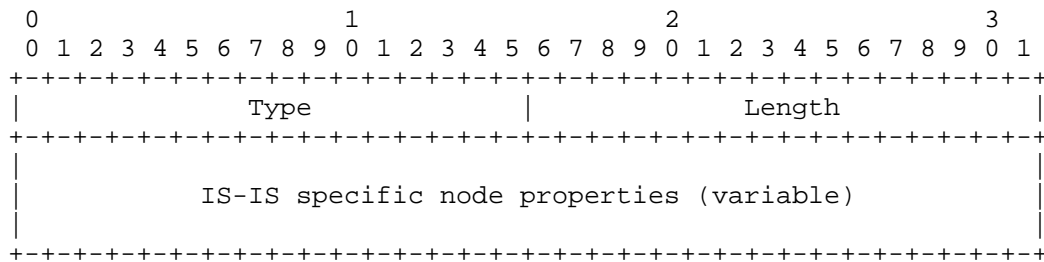


Figure 32: IS-IS specific Node property format

#### 3.3.2.5. ISIS Area Address TLV

The area address is carried in the Area Identifier SubTLV of the Identifier TLV and consists of the Area Address which is assigned to the node. If more than one Area Addresses are present, only the lower address is encoded. Note that the Area Identifier SubTLV may appear in all NLRI types (Link, Node and Prefix) and is defined in Section 3.2.1.3.

#### 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, 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:

TLV/SubTLV	Description	Length	Reference
283	IGP Flags	4	284
Route Tag	4*n	[RFC5130]	285
Extended Tag	8*n	[RFC5130]	286
Prefix Metric	4	[RFC5305]	287
OSPF Forwarding Address	4	[RFC2328]	

Table 8: 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:

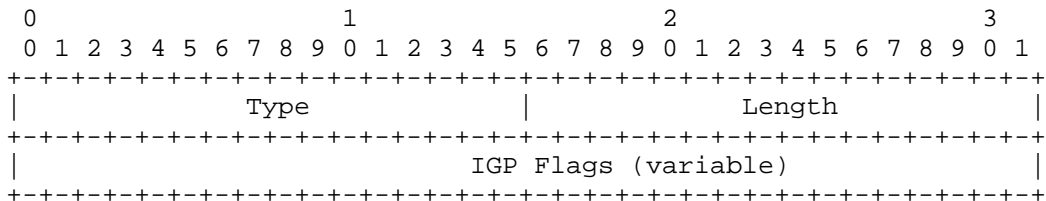


Figure 33: IGP Flag TLV format

where:

Type is 283

Length is variable

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 9: 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:

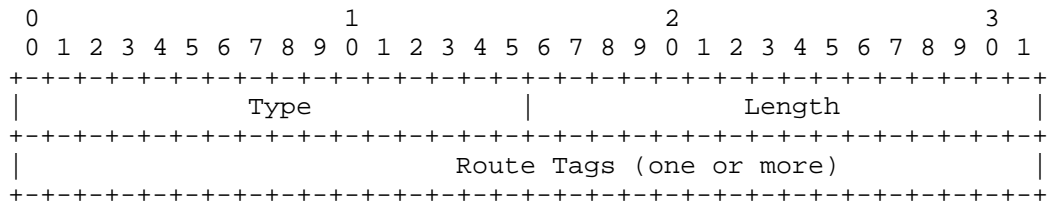


Figure 34: IGP Route TAG TLV format

where:

Type is 284

Length is a multiple of 4

One or more Route 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:

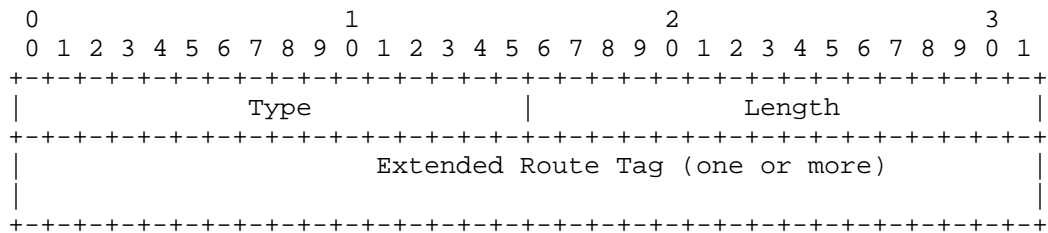


Figure 35: Extended IGP Route TAG TLV format

where:

Type is 285

Length is a multiple of 8

Extended Route Tag contains one or more Extended Route Tags 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.

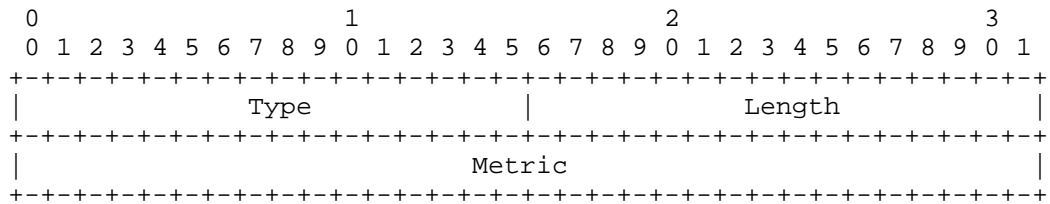


Figure 36: Prefix Metric TLV Format

where:

Type is 286

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.

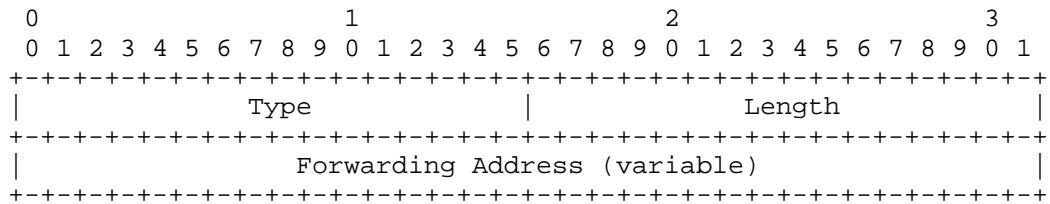


Figure 37: OSPF Forwarding Address TLV Format

where:

Type is 287

Length is 4 for an IPv4 forwarding address and 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].

The BGP Next Hop attribute is used by each BGP-LS speaker to validate the NLRI it receives. However, this specification doesn't mandate any rule regarding the re-write of the BGP Next Hop attribute.

### 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 38. 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.

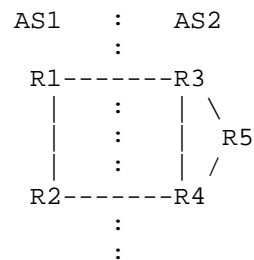


Figure 38: no-link-aggregation

## 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 39. 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.

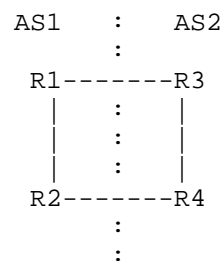


Figure 39: asbr-link-aggregation



#### 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 40. AS3 is modeled as a single node which connects to the border routers of the aggregated domain.

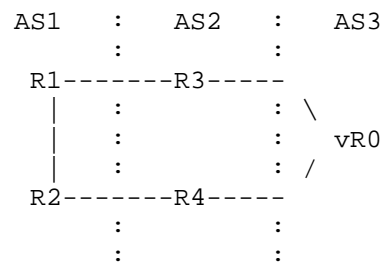


Figure 40: multi-as-aggregation

#### 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 4. 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 change 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:

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

- o 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 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.

An implementation SHOULD allow the operator to configure a pair of ASN and BGP identifier per flooding set the node participates in.

##### 6.2.4. Accounting Management

Not Applicable.

##### 6.2.5. Performance Management

An implementation SHOULD provide the following statistics:

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

- o Number of errored received Link-State NLRI updates, per neighbor
- o Total number of locally originated Link-State NLRIs

#### 6.2.6. Security Management

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

- o Drop all updates from Consumer peers

### 7. TLV/SubTLV Code Points Summary

This section contains the global table of all TLVs/SubTLVs defined in this document.

TLV/SubTLV	Description	IS-IS TLV/Sub-TLV	Value defined in:
256	Identifier	--	Section 3.2.1
257	Local Node Descriptors	--	Section 3.2.2.1
258	Remote Node Descriptors	--	Section 3.2.2.2
259	Autonomous System	--	Section 3.2.2.3
260	BGP Identifier	--	Section 3.2.2.3
261	ISO Node-ID	--	Section 3.2.2.3
262	IPv4 Router-ID	--	Section 3.2.2.3
263	IPv6 Router-ID	--	Section 3.2.2.3
264	Link Local/Remote Identifiers	22/4	[RFC5307]/1.1
265	IPv4 interface address	22/6	[RFC5305]/3.2
266	IPv4 neighbor address	22/8	[RFC5305]/3.3
267	IPv6 interface address	22/12	[RFC6119]/4.2
268	IPv6 neighbor address	22/13	[RFC6119]/4.3
256/5	Multi Topology ID	--	Section 3.2.1.5
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	TE Default Metric	22/18	[RFC5305]/3.7
274	Link Protection Type	22/20	[RFC5307]/1.2
275	MPLS Protocol Mask	--	Section 3.3.1.1
276	Metric	--	Section 3.3.1.2
277	Shared Risk Link Group	--	Section 3.3.1.3
278	OSPF specific link attribute	--	Section 3.3.1.4
279	IS-IS Specific Link Attribute	--	Section 3.3.1.5
280	Node Flag Bits	--	Section 3.3.2.2
281	OSPF Specific Node Properties	--	Section 3.3.2.3

282	IS-IS Specific Node Properties	--	Section 3.3.2.4
283	IGP Flags	--	Section 3.3.3.1
284	Route Tag	--	[RFC5130]
285	Extended Tag	--	[RFC5130]
286	Prefix Metric	--	[RFC5305]
287	OSPF Forwarding Address	--	[RFC2328]

Table 10: Summary Table of TLV/SubTLV Codepoints

## 8. 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.

## 9. Contributors

We would like to thank Robert Varga for the significant contribution he gave to this document.

## 10. Acknowledgements

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

## 11. References

### 11.1. 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.
- [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.
- [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.
- [RFC6822] Previdi, S., Ginsberg, L., Shand, M., Roy, A., and D. Ward, "IS-IS Multi-Instance", RFC 6822, December 2012.

## 11.2. 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.
- [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.
- [RFC6286] Chen, E. and J. Yuan, "Autonomous-System-Wide Unique BGP Identifier for BGP-4", RFC 6286, June 2011.
- [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