Inter-Domain Routing
Internet-Draft
Obsoletes: 7752 (if approved)
Intended status: Standards Track
Expires: May 4, 2020

K. Talaulikar, Ed.
Cisco Systems

November 1, 2019

Distribution of Link-State and Traffic Engineering Information Using BGP

draft-ietf-idr-rfc7752bis-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 (TE) information. This is information typically distributed by IGP routing protocols within the network.

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. This is achieved using a new BGP Network Layer Reachability Information (NLRI) encoding format. The mechanism is applicable to physical and virtual IGP 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).

This document obsoletes RFC 7752 by completely replacing that document. It makes a number of small changes and clarifications to the previous specification.

Status of This Memo

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

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

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

This Internet-Draft will expire on May 4, 2020.

Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction 3
1.1. Requirements Language 5
2. Motivation and Applicability 5
2.1. MPLS-TE with PCE 5
2.2. ALTO Server Network API 7
3. BGP Speaker Roles for BGP-LS 7
4. Carrying Link-State Information in BGP 9
4.1. TLV Format 9
4.2. The Link-State NLRI 10
4.2.1. Node Descriptors 14
4.2.2. Link Descriptors 18
4.2.3. Prefix Descriptors 21
4.3. The BGP-LS Attribute 23
4.3.1. Node Attribute TLVs 23
4.3.2. Link Attribute TLVs 27
4.3.3. Prefix Attribute TLVs 32
4.4. Private Use 35
4.5. BGP Next-Hop Information 36
4.6. Inter-AS Links 36
4.7. Handling of Unreachable IGP Nodes 37
4.8. Router-ID Anchoring Example: ISO Pseudonode 38
4.9. Router-ID Anchoring Example: OSPF Pseudonode 39
4.10. Router-ID Anchoring Example: OSPFv2 to IS-IS Migration . 40
5. Link to Path Aggregation 41
5.1. Example: No Link Aggregation 41
5.2. Example: ASBR to ASBR Path Aggregation 41
5.3. Example: Multi-AS Path Aggregation 42
6. IANA Considerations 42
6.1. Guidance for Designated Experts 43
7. Manageability Considerations 44
7.1. Operational Considerations 44

7.1.1. Operations 44
7.1.2. Installation and Initial Setup 44
7.1.3. Migration Path 44
7.1.4. Requirements on Other Protocols and Functional Components 45
7.1.5. Impact on Network Operation 45
7.1.6. Verifying Correct Operation 45
7.2. Management Considerations 45
7.2.1. Management Information 45
7.2.2. Fault Management 45
7.2.3. Configuration Management 48
7.2.4. Accounting Management 48
7.2.5. Performance Management 48
7.2.6. Security Management 49
8. TLV/Sub-TLV Code Points Summary 49
9. Security Considerations 50
10. Contributors 51
11. Acknowledgements 52
12. References 52
12.1. Normative References 52
12.2. Informative References 55
Appendix A. Changes from RFC 7752 57
Author's Address 59

1. Introduction

The contents of a Link-State Database (LSDB) or of an IGP's Traffic Engineering Database (TED) describe only the links and nodes within 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 that 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 Class-of-Service (CoS) class reservation state, preemption, and Shared Risk Link Groups (SRLGs). 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.

An illustration of the collection of link-state and TE information and its distribution to consumers is shown in the Figure 1 below.

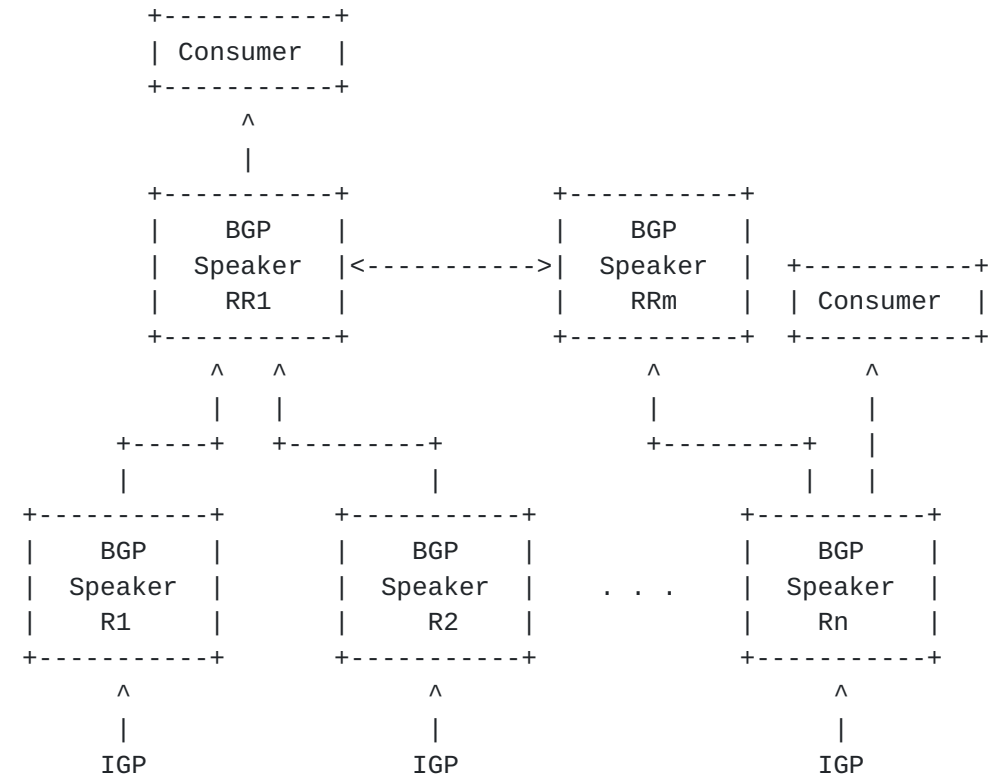


Figure 1: Collection of Link-State and TE Information

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 Point of Presence (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 a 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.

This document obsoletes [RFC7752] by completely replacing that document. It makes a number of small changes and clarifications to the previous specification as documented in Appendix A.

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. 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).

- 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, then 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 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.

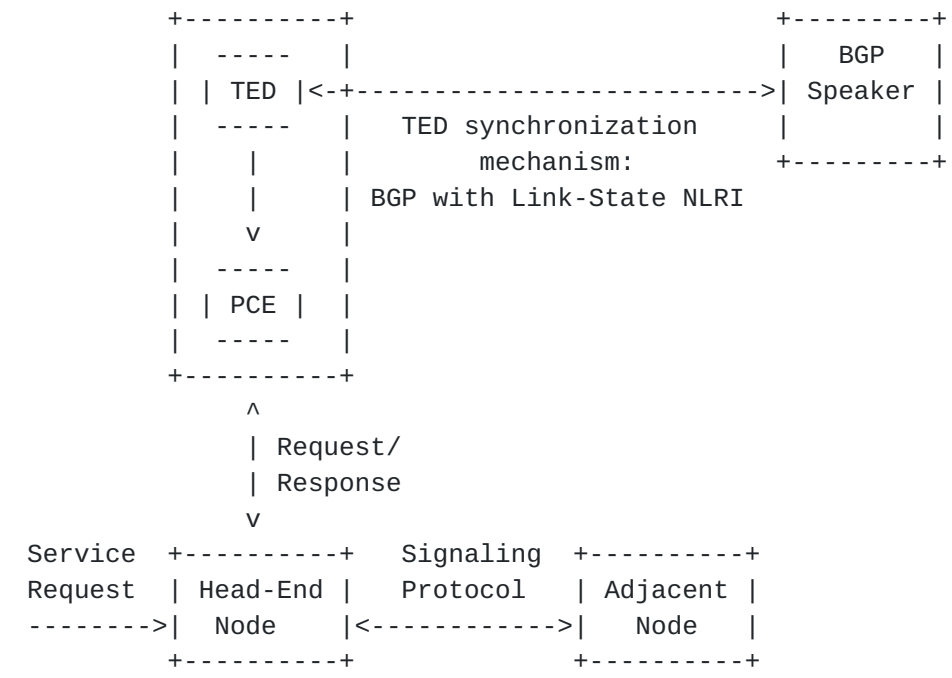


Figure 2: External PCE Node Using a TED Synchronization Mechanism

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 peer-to-peer (P2P) clients or trackers, or Content Distribution Networks (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 [RFC7285].

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, and TE (topology) data is required to generate ALTO Cost Maps. Prefix data is carried and originated in BGP, and 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 that 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.

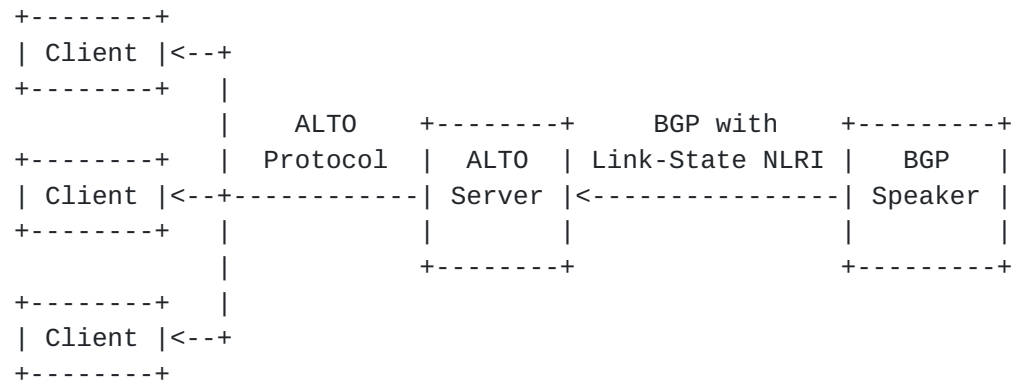


Figure 3: ALTO Server Using Network Topology Information

3. BGP Speaker Roles for BGP-LS

In the illustration shown in Figure 1, the BGP Speakers can be seen playing different roles in the distribution of information using BGP-LS. This section introduces terms that explain the different roles

of the BGP Speakers which are then used through the rest of this document.

- BGP-LS Producer: The BGP Speakers R1, R2, ... Rn, originate link-state information from their underlying link-state IGP protocols into BGP-LS. If R1 and R2 are in the same IGP area, then likely they are originating the same link-state information into BGP-LS. R1 may also source information from sources other than IGP, e.g. its local node information. The term BGP-LS Producer refers to the BGP Speaker that is originating link-state information into BGP.

- BGP-LS Consumer: The BGP Speakers RR1 and Rn are handing off the BGP-LS information that they have collected to a consumer application. The BGP protocol implementation and the consumer application may be on the same or different nodes. The term BGP-LS Consumer refers to the consumer application/process and not the BGP Speaker. This document only covers the BGP implementation. The consumer application and the design of interface between BGP and consumer application may be implementation specific and outside the scope of this document.

- BGP-LS Propagator: The BGP Speaker RRm propagates the BGP-LS information between the BGP Speaker Rn and the BGP Speaker RR1. The BGP implementation on RRm is doing the propagation of BGP-LS updates and performing BGP best path calculations. Similarly, the BGP Speaker RR1 is receiving BGP-LS information from R1, R2 and RRm and propagating the information to the BGP-LS Consumer after performing BGP best path calculations. The term BGP-LS Propagator refers to the BGP Speaker that is performing BGP protocol processing on the link-state information.

The above roles are not mutually exclusive. The same BGP Speaker may be the producer for some link-state information and propagator for some other link-state information while also providing this information to a consumer application. Nothing precludes a BGP implementation performing some of the validation and processing on behalf of the BGP-LS Consumer as long as it does not impact the semantics of its role as BGP-LS Propagator as described in this document.

The rest of this document refers to the role when describing procedures that are specific to that role. When the role is not specified, then the said procedure applies to all BGP Speakers.

4. 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 (BGP-LS Attribute) that carries link, node, and prefix properties and attributes, such as the link and prefix metric or auxiliary Router-IDs of nodes, etc.

It is desirable to keep the dependencies on the protocol source of this attribute to a minimum and represent any content in an IGP-neutral way, such that applications that want to learn about a link-state topology do not need to know about any OSPF or IS-IS protocol specifics.

This section mainly describes the procedures at a BGP-LS Producer that originate link-state information into BGP-LS.

4.1. TLV Format

Information in the new Link-State NLRIs and the BGP-LS Attribute is encoded in Type/Length/Value triplets. The TLV format is shown in Figure 4 and applies to both the NLRI and the BGP-LS Attribute encodings.

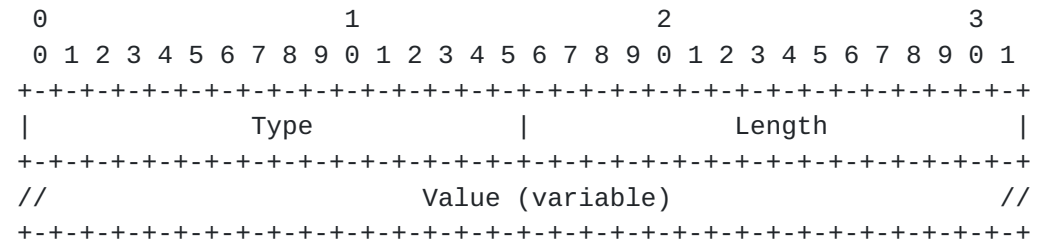


Figure 4: TLV Format

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 4-octet alignment. Unknown and unsupported types MUST be preserved and propagated within both the NLRI and the BGP-LS Attribute. The presence of unrecognized or unexpected TLVs MUST NOT result in the NLRI or the BGP-LS Attribute being considered as malformed.

In order to compare NLRIs with unknown TLVs, all TLVs within the NLRI MUST be ordered in ascending order by TLV Type. If there are multiple TLVs of the same type within a single NLRI, then the TLVs sharing the same type MUST be in ascending order based on the value field. Comparison of the value fields is performed by treating the

entire field as an opaque hexadecimal string. Standard string comparison rules apply. NLRIs having TLVs which do not follow the above ordering rules MUST be considered as malformed by a BGP-LS Propagator. This ensures that multiple copies of the same NLRI from multiple BGP-LS Producers and the ambiguity arising there from is prevented.

All TLVs within the NLRI that are not specified as mandatory are considered optional. All TLVs within the BGP-LS Attribute are considered optional unless specified otherwise.

The TLVs within the BGP-LS Attribute MAY be ordered in ascending order by TLV type. BGP-LS Attribute with unordered TLVs MUST NOT be considered malformed.

4.2. The Link-State NLRI

The MP_REACH_NLRI and MP_UNREACH_NLRI attributes are BGP's containers for carrying opaque information. This specification defines three Link-State NLRI types that describes either a node, a link, and a prefix.

All non-VPN link, node, and prefix information SHALL be encoded using AFI 16388 / SAFI 71. VPN link, node, and prefix information SHALL be encoded using AFI 16388 / SAFI 72.

In order for two BGP speakers to exchange Link-State NLRI, they MUST use BGP Capabilities Advertisement to ensure that they are both capable of properly processing such NLRI. This is done as specified in [RFC4760], by using capability code 1 (multi-protocol BGP), with AFI 16388 / SAFI 71 for BGP-LS, and AFI 16388 / SAFI 72 for BGP-LS-VPN.

New Link-State NLRI Types may be introduced in the future. Since supported NLRI type values within the address family are not expressed in the Multiprotocol BGP (MP-BGP) capability [RFC4760], it is possible that a BGP speaker has advertised support for Link-State but does not support a particular Link-State NLRI type. In order to allow introduction of new Link-State NLRI types seamlessly in the future, without the need for upgrading all BGP speakers in the propagation path (e.g. a route reflector), this document deviates from the default handling behavior specified by [RFC7606] for Link-State address-family. An implementation MUST handle unrecognized Link-State NLRI types as opaque objects and MUST preserve and propagate them.

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

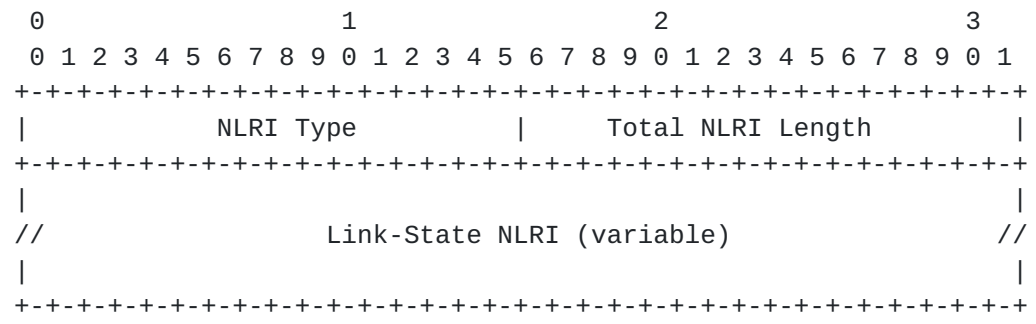


Figure 5: Link-State AFI 16388 / SAFI 71 NLRI Format

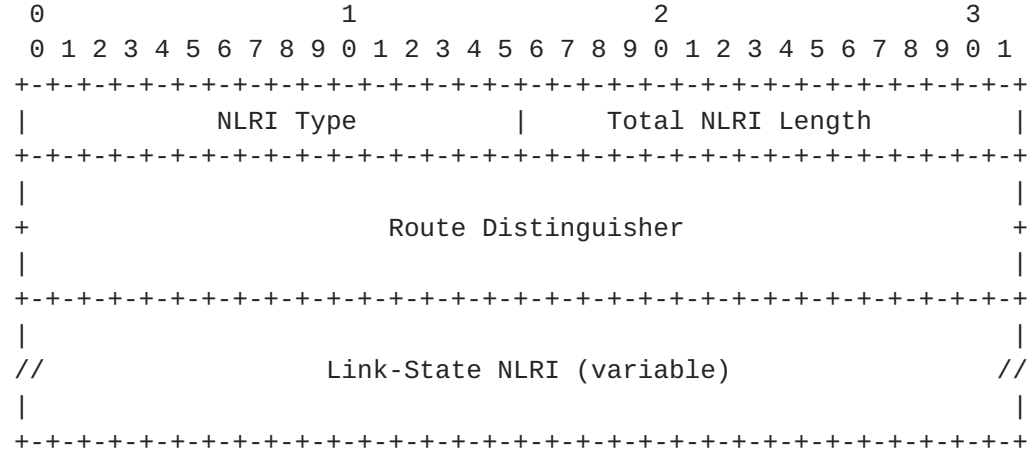


Figure 6: Link-State VPN AFI 16388 / SAFI 72 NLRI Format

The Total NLRI Length field contains the cumulative length, in octets, of the rest of the NLRI, not including the NLRI Type field or itself. For VPN applications, it also includes the length of the Route Distinguisher.

Type	NLRI Type
1	Node NLRI
2	Link NLRI
3	IPv4 Topology Prefix NLRI
4	IPv6 Topology Prefix NLRI
65000-65535	Private Use

Table 1: NLRI Types

Route Distinguishers are defined and discussed in [RFC4364].

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

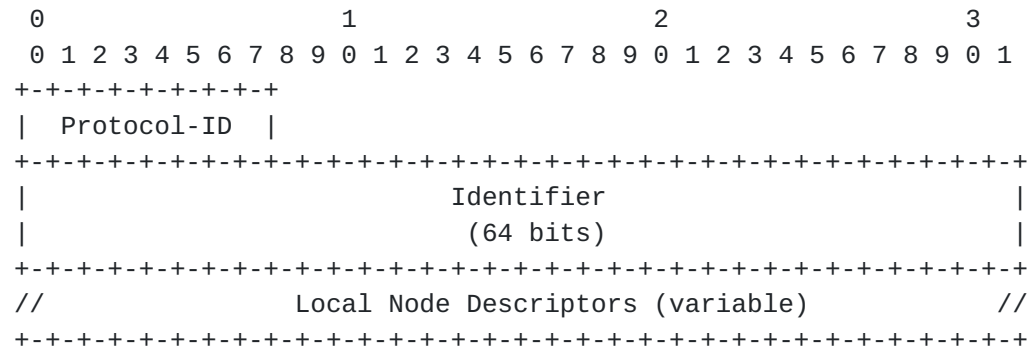


Figure 7: The Node NLRI Format

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

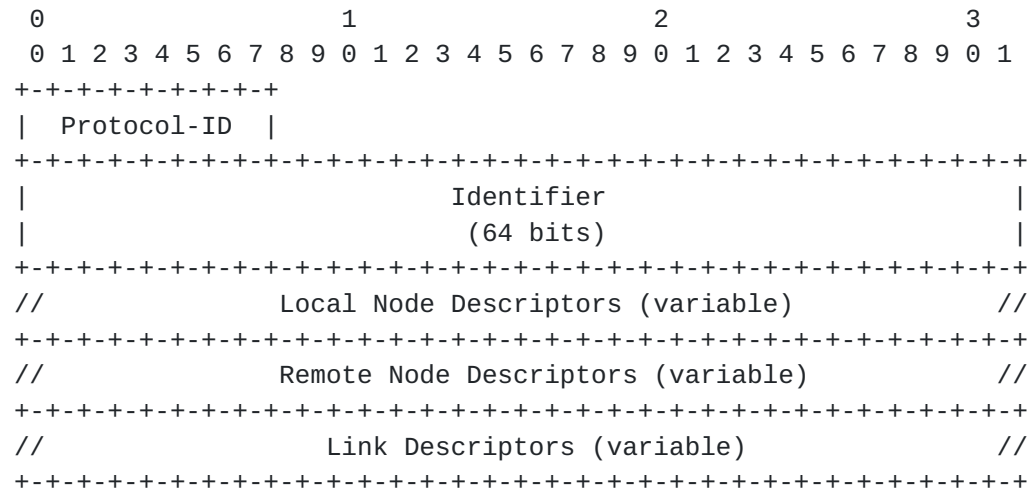


Figure 8: The Link NLRI Format

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

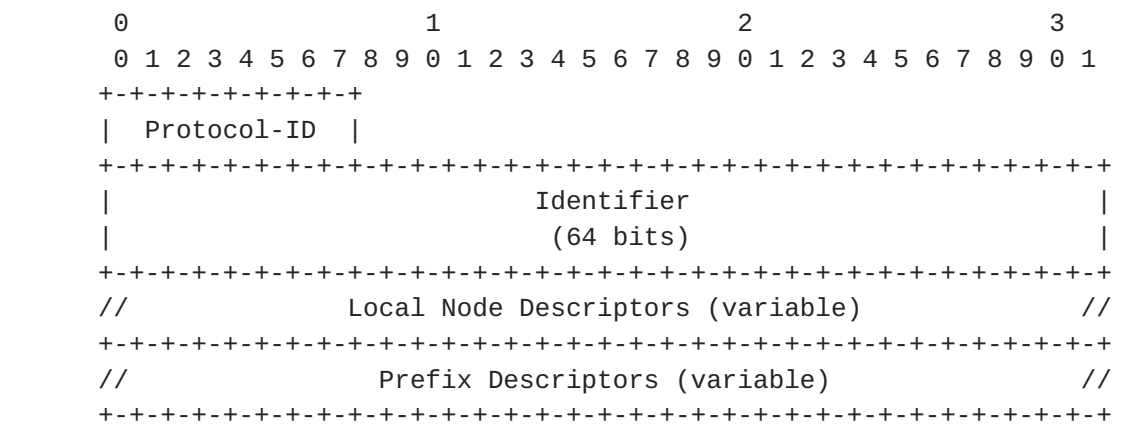


Figure 9: The IPv4/IPv6 Topology Prefix NLRI Format

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

Protocol-ID	NLRI information source protocol
1	IS-IS Level 1
2	IS-IS Level 2
3	OSPFv2
4	Direct
5	Static configuration
6	OSPFv3
200-255	Private Use

Table 2: Protocol Identifiers

The 'Direct' and 'Static configuration' protocol types SHOULD be used when BGP-LS is sourcing local information. For all information derived from other protocols, the corresponding Protocol-ID MUST be used. If BGP-LS has direct access to interface information and wants to advertise a local link, then the Protocol-ID 'Direct' SHOULD be used. For modeling virtual links, such as described in Section 5, the Protocol-ID 'Static configuration' SHOULD be used.

A router MAY run multiple protocol instances of OSPF or ISIS where by it becomes a border router between multiple IGP domains. Both OSPF and IS-IS MAY also run multiple routing protocol instances over the same link. See [RFC8202] and [RFC6549]. These instances define independent IGP routing domains. The 64-bit Identifier field carries a BGP-LS Instance Identifier (Instance-ID) that is used to identify the IGP routing domain where the NLRI belongs. The NLRIs representing link-state objects (nodes, links, or prefixes) from the same IGP routing instance MUST have the same Identifier field value.

NLRIs with different Identifier field values MUST be considered to be from different IGP routing instances. The Identifier field value 0 is RECOMMENDED to be used when there is only a single protocol instance in the network where BGP-LS is operational.

An implementation which supports multiple IGP instances MUST support the configuration of unique BGP-LS Instance-IDs at the routing protocol instance level. The network operator MUST assign consistent BGP-LS Instance-ID values on all BGP-LS Producers within a given IGP domain. Unique BGP-LS Instance-ID values MUST be assigned to routing protocol instances operating in different IGP domains. This allows the BGP-LS Consumer to build an accurate segregated multi-domain topology based on the Identifier field even when the topology is advertised via BGP-LS by multiple BGP-LS Producers in the network.

When the above described semantics and recommendations are not followed, a BGP-LS Consumer may see duplicate link-state objects for the same node, link or prefix when there are multiple BGP-LS Producers deployed. This may also result in the BGP-LS Consumers getting an inaccurate network-wide topology.

When adding, removing or modifying a TLV/sub-TLV from a Link-State NLRI, the BGP-LS Producer MUST withdraw the old NLRI by including it in the MP_UNREACH_NLRI. Not doing so can result in duplicate and in-consistent link-state objects hanging around in the BGP-LS table.

Each Node Descriptor, Link Descriptor and Prefix Descriptor consists of one or more TLVs, as described in the following sections. These Descriptor TLVs are applicable for the Node, Link and Prefix NLRI Types for the protocols listed in Table 2. Documents extending BGP-LS specifications with new NLRI Types and/or protocols MUST specify the NLRI Descriptors for them.

4.2.1. Node Descriptors

Each link is anchored by a pair of Router-IDs that are used by the underlying IGP, namely, a 48-bit ISO System-ID for IS-IS and a 32-bit Router-ID for OSPFv2 and OSPFv3. An IGP may use one or more additional auxiliary Router-IDs, mainly for Traffic Engineering purposes. For example, IS-IS may have one or more IPv4 and IPv6 TE Router-IDs [RFC5305] [RFC6119]. These auxiliary Router-IDs MUST be included in the node attribute described in Section 4.3.1 and MAY be included in link attribute described in Section 4.3.2. The advertisement of the TE Router-IDs help a BGP-LS Consumer to correlate multiple link-state objects (e.g. in different IGP instances or areas/levels) to the same node in the network.

It is desirable that the Router-ID assignments inside the Node Descriptor are globally unique. However, there may be Router-ID spaces (e.g., ISO) where no global registry exists, or worse, Router-IDs have been allocated following the private-IP allocation described in RFC 1918 [RFC1918]. BGP-LS uses the Autonomous System (AS) Number to disambiguate the Router-IDs, as described in Section 4.2.1.1.

4.2.1.1. Globally Unique Node/Link/Prefix Identifiers

One problem that needs to be addressed is the ability to identify an IGP node globally (by "globally", 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 (hence, by extension links and prefixes) within which each node has a unique IGP representation by using the combination of Area-ID, Router-ID, Protocol-ID, Multi-Topology ID, and Instance-ID. The problem is that BGP may receive node/link/prefix information 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 AS. During IGP transitions, it may happen that two redundant IGPs are in place.

The mapping of the Instance-ID to the Identifier field as described earlier along with a set of sub-TLVs described in Section 4.2.1.4, allows specification of a flexible key for any given node/link information such that global uniqueness of the NLRI is ensured.

4.2.1.2. Local Node Descriptors

The Local Node Descriptors TLV contains Node Descriptors for the node anchoring the local end of the link. This is a mandatory TLV in all three types of NLRIs (node, link, and prefix). The Type is 256. The length of this TLV is variable. The value contains one or more Node Descriptor Sub-TLVs defined in Section 4.2.1.4.

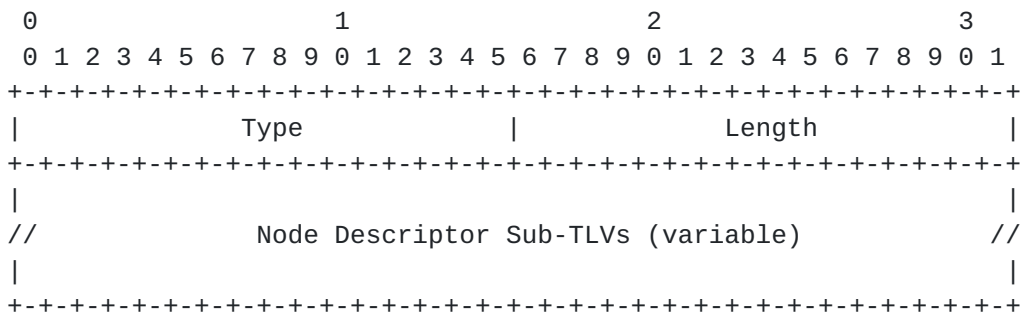


Figure 10: Local Node Descriptors TLV Format

4.2.1.3. Remote Node Descriptors

The Remote Node Descriptors TLV contains Node Descriptors for the node anchoring the remote end of the link. This is a mandatory TLV for Link NLRIs. The type is 257. The length of this TLV is variable. The value contains one or more Node Descriptor Sub-TLVs defined in Section 4.2.1.4.

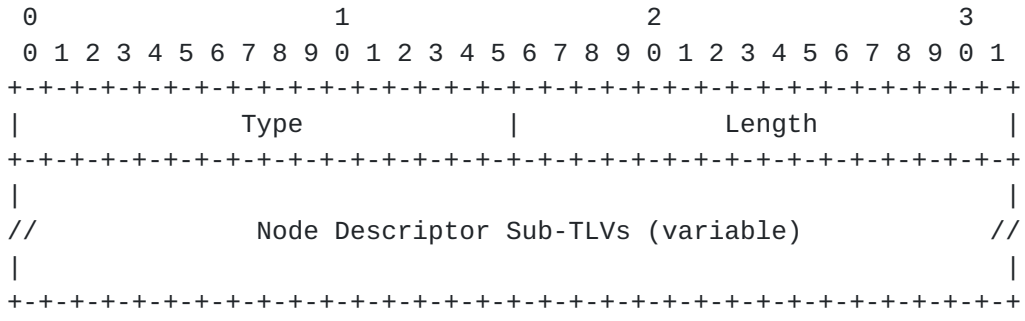


Figure 11: Remote Node Descriptors TLV Format

4.2.1.4. Node Descriptor Sub-TLVs

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

Sub-TLV Code Point	Description	Length
512	Autonomous System	4
513	BGP-LS Identifier (deprecated)	4
514	OSPF Area-ID	4
515	IGP Router-ID	Variable

Table 3: Node Descriptor Sub-TLVs

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

Autonomous System: Opaque value (32-bit AS Number). This is an optional TLV. The value SHOULD be set to the AS Number associated with the BGP process originating the link-state information. An implementation MAY provide a configuration option on the BGP-LS Producer to use a value different.

BGP-LS Identifier: Opaque value (32-bit ID). This is an optional TLV. In conjunction with Autonomous System Number (ASN), uniquely identifies the BGP-LS domain. The combination of ASN and BGP-LS ID MUST be globally unique. All BGP-LS speakers within an IGP flooding-set (set of IGP nodes within which an LSP/LSA is flooded) MUST use the same ASN, BGP-LS ID tuple. If an IGP domain consists of multiple flooding-sets, then all BGP-LS speakers within the IGP domain SHOULD use the same ASN, BGP-LS ID tuple.

Area-ID: Used to identify the 32-bit area to which the information advertised in the NLRI belongs. This is a mandatory TLV when originating information from OSPF that is derived from area-scope LSAs. The Area Identifier allows different NLRIs of the same router to be discriminated on a per area basis. It is not used for NLRIs when carrying information that is derived from AS-scope LSAs as it is not associated with a specific area.

IGP Router-ID: Opaque value. This is a mandatory TLV when originating information from IS-IS, OSPF, direct or static. For an IS-IS non-pseudonode, this contains a 6-octet ISO Node-ID (ISO system-ID). For an IS-IS pseudonode corresponding to a LAN, this contains the 6-octet ISO Node-ID of the Designated Intermediate System (DIS) followed by a 1-octet, nonzero PSN identifier (7 octets in total). For an OSPFv2 or OSPFv3 non-pseudonode, this contains the 4-octet Router-ID. For an OSPFv2 pseudonode representing a LAN, this contains the 4-octet Router-ID of the Designated Router (DR) followed by the 4-octet IPv4 address of the DR's interface to the LAN (8 octets in total). Similarly, for an OSPFv3 pseudonode, this contains the 4-octet Router-ID of the DR followed by the 4-octet interface identifier of the DR's interface to the LAN (8 octets in total). The TLV size in combination with the protocol identifier enables the decoder to determine the type of the node. For Direct or Static configuration, the value SHOULD be taken from an IPv4 or IPv6 address (e.g. loopback interface) configured on the node.

There can be at most one instance of each sub-TLV type present in any Node Descriptor. The sub-TLVs within a Node Descriptor MUST be arranged in ascending order by sub-TLV type. This needs to be done in order to compare NLRIs, even when an implementation

encounters an unknown sub-TLV. Using stable sorting, an implementation can do binary comparison of NLRIs and hence allow incremental deployment of new key sub-TLVs.

The BGP-LS Identifier was introduced by [RFC7752] and it's use is being deprecated by this document. Implementations MUST continue to support this sub-TLV for backward compatibility. The default value of 0 is RECOMMENDED to be use when a BGP-LS Producer includes this sub-TLV when originating information into BGP-LS. Implementations MAY provide an option to configure this value for backward compatibility reasons. The use of the Instance-ID in the Identifier field is the RECOMMENDED way of segregation of different IGP domains in BGP-LS.

4.2.2. Link Descriptors

The Link Descriptor field is a set of Type/Length/Value (TLV) triplets. The format of each TLV is shown in Section 4.1. The Link Descriptor TLVs uniquely identify a link among multiple parallel links 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 advertise a half-link each, i.e., two Link NLRIs are advertised for a given point-to-point link.

A BGP-LS Consumer should not consider a link between two nodes as being available unless it has received the two Link NLRIs corresponding to the half-link representation of that link from both the nodes. This check is similar to the 'two way connectivity check' that is performed by link-state IGPs and is also required to be done by BGP-LS Consumers of link-state topology.

A BGP-LS Producer MAY supress the advertisement of a Link NLRI, corresponding to a half link, from a link-state IGP unless it has verified that the link is being reported in the IS-IS LSP or OSPF Router LSA by both the nodes connected by that link. This 'two way connectivity check' is performed by link-state IGPs during their computation and may be leveraged before passing information for any half-link that is reported from these IGPs in to BGP-LS. This ensures that only those Link State IGP adjacencies which are established get reported via Link NLRIs. Such a 'two way connectivity check' may be also required in certain cases (e.g. with OSPF) to obtain the proper link identifiers of the remote node.

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 by either IS-IS or OSPF.

The following TLVs are defined as Link Descriptors in the Link NLRI:

TLV Code Point	Description	IS-IS TLV /Sub-TLV	Reference (RFC/Section)
258	Link Local/Remote Identifiers	22/4	[RFC5307] / 1.1
259	IPv4 interface address	22/6	[RFC5305] / 3.2
260	IPv4 neighbor address	22/8	[RFC5305] / 3.3
261	IPv6 interface address	22/12	[RFC6119] / 4.2
262	IPv6 neighbor address	22/13	[RFC6119] / 4.3
263	Multi-Topology Identifier	---	Section 4.2.2.1

Table 4: Link Descriptor TLVs

The information about a link present in the LSA/LSP originated by the local node of the link determines the set of TLVs in the Link Descriptor of the link.

If interface and neighbor addresses, either IPv4 or IPv6, are present, then the IP address TLVs MUST be included and the Link Local/Remote Identifiers TLV MUST NOT be included in the Link Descriptor. The Link Local/Remote Identifiers TLV MAY be included in the link attribute when available. IPv6 link-local addresses MUST NOT be carried in the IPv6 address TLVs as descriptors of a link as they are not considered unique.

If interface and neighbor addresses are not present and the link local/remote identifiers are present, then the Link Local/Remote Identifiers TLV MUST be included in the Link Descriptor. The Link Local/Remote Identifiers MUST be included in the Link Descriptor also in the case of links having only IPv6 link-local addressing on them.

The Multi-Topology Identifier TLV MUST be included in Link Descriptor if the underlying IGP link object is associated with a non-default topology.

The TLVs/sub-TLVs corresponding to the interface addresses and/or the local/remote identfiers may not always be signaled in the IGPs unless their advertisement is enabled specifically. In such cases, a BGP-LS Producer may not be able to generate valid Link NLRIs for such link advertisements from the IGPs.

4.2.2.1. Multi-Topology ID

The Multi-Topology ID (MT-ID) TLV carries one or more IS-IS or OSPF Multi-Topology IDs for a link, node, or prefix.

Semantics of the IS-IS MT-ID are defined in Section 7.1 and 7.2 of RFC 5120 [RFC5120]. Semantics of the OSPF MT-ID are defined in Section 3.7 of RFC 4915 [RFC4915]. If the value in the MT-ID TLV is derived from OSPF, then the upper 5 bits of the MT-ID field MUST be set to 0.

The format of the MT-ID TLV is shown in the following figure.

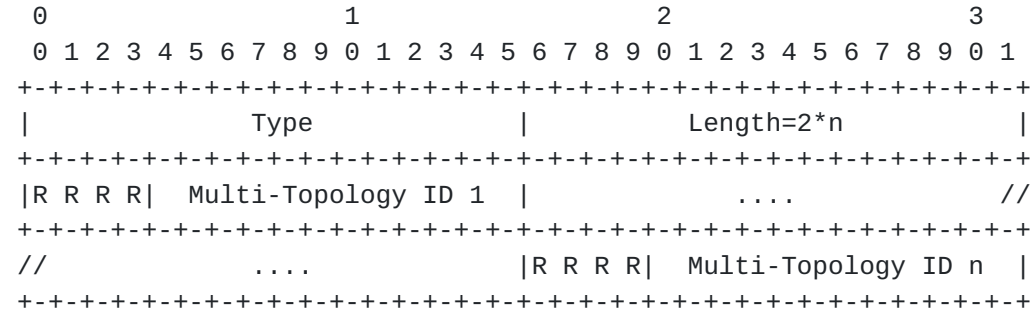


Figure 12: Multi-Topology ID TLV Format

where Type is 263, Length is 2*n, and n is the number of MT-IDs carried in the TLV.

The MT-ID TLV MAY be present in a Link Descriptor, a Prefix Descriptor, or the BGP-LS attribute of a Node NLRI. In a Link or Prefix Descriptor, only a single MT-ID TLV containing the MT-ID of the topology where the link or the prefix is reachable is allowed. In case one wants to advertise multiple topologies for a given Link Descriptor or Prefix Descriptor, multiple NLRIs MUST be generated where each NLRI contains a single unique MT-ID. When used in the Link or Prefix Descriptor TLV for IS-IS, the Bits R are reserved and MUST be set to 0 (as per Section 7.2 of RFC 5120 [RFC5120]) when originated and ignored on receipt.

In the BGP-LS attribute of a Node NLRI, one MT-ID TLV containing the array of MT-IDs of all topologies where the node is reachable is

allowed. When used in the Node Attribute TLV for IS-IS, the Bits R are set as per Section 7.1 of RFC 5120 [RFC5120].

4.2.3. Prefix Descriptors

The Prefix Descriptor field is a set of Type/Length/Value (TLV) triplets. Prefix Descriptor TLVs uniquely identify an IPv4 or IPv6 prefix originated by a node. The following TLVs are defined as Prefix Descriptors in the IPv4/IPv6 Prefix NLRI:

TLV Code Point	Description	Length	Reference (RFC/Section)
263	Multi-Topology Identifier	variable	Section 4.2.2.1
264	OSPF Route Type	1	Section 4.2.3.1
265	IP Reachability Information	variable	Section 4.2.3.2

Table 5: Prefix Descriptor TLVs

The Multi-Topology Identifier TLV MUST be included in Prefix Descriptor if the underlying IGP prefix object is associated with a non-default topology.

4.2.3.1. OSPF Route Type

The OSPF Route Type TLV is a mandatory TLV corresponding to Prefix NLRIs originated from OSPF. It is used to identify the OSPF route type of the prefix. An OSPF prefix MAY be advertised in the OSPF domain with multiple route types. The Route Type TLV allows the discrimination of these advertisements. The format of the OSPF Route Type TLV is shown in the following figure.

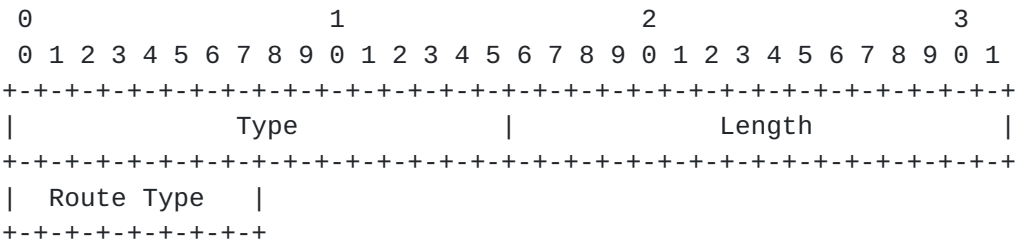


Figure 13: OSPF Route Type TLV Format

where the Type and Length fields of the TLV are defined in Table 5. The OSPF Route Type field values are defined in the OSPF protocol and can be one of the following:

- Intra-Area (0x1)
- Inter-Area (0x2)
- External 1 (0x3)
- External 2 (0x4)
- NSSA 1 (0x5)
- NSSA 2 (0x6)

4.2.3.2. IP Reachability Information

The IP Reachability Information TLV is a mandatory TLV for IPv4 & IPv6 Prefix NLRI types. The TLV contains one IP address prefix (IPv4 or IPv6) originally advertised in the IGP topology. Its purpose is to glue a particular BGP service NLRI by virtue of its BGP next hop to a given node in the LSDB. A router SHOULD advertise an IP Prefix NLRI for each of its BGP next hops. The format of the IP Reachability Information TLV is shown in the following figure:

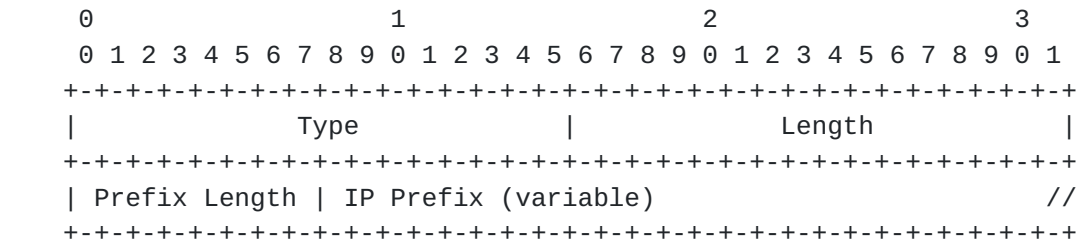


Figure 14: IP Reachability Information TLV Format

The Type and Length fields of the TLV are defined in Table 5. The following two fields determine the reachability information of the address family. The Prefix Length field contains the length of the prefix in bits. The IP Prefix field contains the most significant octets of the prefix, 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, 4 octets for prefix length 25 up to 32, etc.

4.3. The BGP-LS Attribute

The BGP-LS Attribute 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 address families.

The Node Attribute TLVs, Link Attribute TLVs and Prefix Attribute TLVs are sets of TLVs that may be encoded in the BGP-LS Attribute associated with a Node NLRI, Link NLRI and Prefix NLRI respectively.

The BGP-LS Attribute may potentially grow large in size depending on the amount of link-state information associated with a single Link-State NLRI. The BGP specification [RFC4271] mandates a maximum BGP message size of 4096 octets. It is RECOMMENDED that an implementation support [RFC8654] in order to accommodate larger size of information within the BGP-LS Attribute. BGP-LS Producers MUST ensure that they limit the TLVs included in the BGP-LS Attribute to ensure that a BGP update message for a single Link-State NLRI does not cross the maximum limit for a BGP message. The determination of the types of TLVs to be included MAY be made by the BGP-LS Producer based on the BGP-LS Consumer applications requirement and is outside the scope of this document. When a BGP-LS Propagator finds that it is exceeding the maximum BGP message size due to addition or update of some other BGP Attribute (e.g. AS_PATH), it MUST consider the BGP-LS Attribute to be malformed and handle the propagation as described in Section 7.2.2.

4.3.1. Node Attribute TLVs

The following Node Attribute TLVs are defined for the BGP-LS Attribute associated with a Node NLRI:

TLV Code Point	Description	Length	Reference (RFC/Section)
263	Multi-Topology Identifier	variable	Section 4.2.2.1
1024	Node Flag Bits	1	Section 4.3.1.1
1025	Opaque Node Attribute	variable	Section 4.3.1.5
1026	Node Name	variable	Section 4.3.1.3
1027	IS-IS Area Identifier	variable	Section 4.3.1.2
1028	IPv4 Router-ID of Local Node	4	[RFC5305] / 4.3
1029	IPv6 Router-ID of Local Node	16	[RFC6119] / 4.1

Table 6: Node Attribute TLVs

4.3.1.1. Node Flag Bits TLV

The Node Flag Bits TLV carries a bit mask describing node attributes. The value is a 1 octet length bit array of flags, where each bit represents a node operational state or attribute.

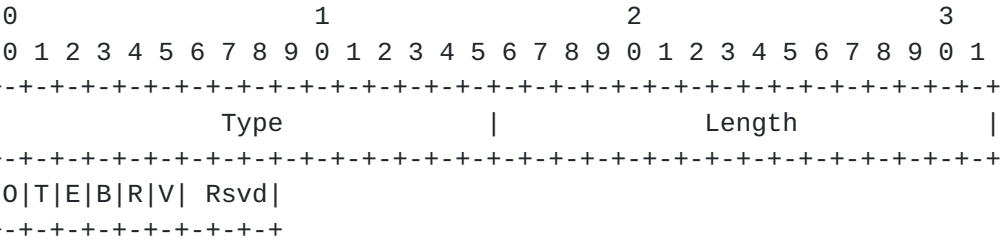


Figure 15: Node Flag Bits TLV Format

The bits are defined as follows:

Bit	Description	Reference
'O'	Overload Bit	[ISO10589]
'T'	Attached Bit	[ISO10589]
'E'	External Bit	[RFC2328]
'B'	ABR Bit	[RFC2328]
'R'	Router Bit	[RFC5340]
'V'	V6 Bit	[RFC5340]
Reserved (Rsvd)	Reserved for future use	

Table 7: Node Flag Bits Definitions

4.3.1.2. IS-IS Area Identifier TLV

An IS-IS node can be part of one or more IS-IS areas. Each of these area addresses is carried in the IS-IS Area Identifier TLV. If multiple area addresses are present, multiple TLVs are used to encode them. The IS-IS Area Identifier TLV may be present in the BGP-LS attribute only when advertised in the Link-State Node NLRI.

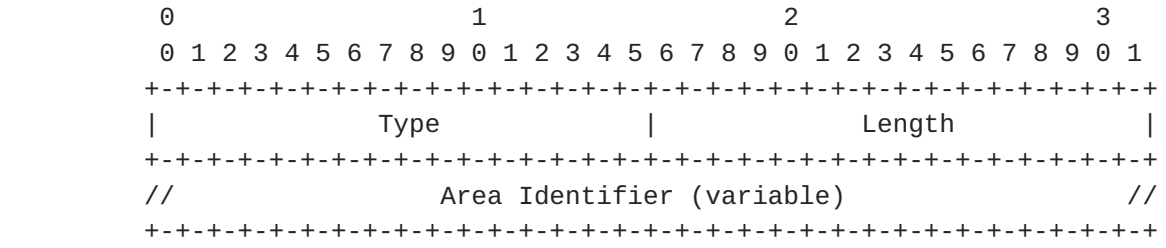


Figure 16: IS-IS Area Identifier TLV Format

4.3.1.3. Node Name TLV

The Node Name TLV is optional. Its structure and encoding has been borrowed from [RFC5301]. The Value field identifies the symbolic name of the router node. This symbolic name can be the Fully Qualified Domain Name (FQDN) for the router, it can be a subset of the FQDN (e.g., a hostname), or it can be any string operators want to use for the router. The use of FQDN or a subset of it is strongly RECOMMENDED. The maximum length of the Node Name TLV is 255 octets.

The Value field is encoded in 7-bit ASCII. If a user interface for configuring or displaying this field permits Unicode characters, that user interface is responsible for applying the ToASCII and/or ToUnicode algorithm as described in [RFC5890] to achieve the correct format for transmission or display.

[RFC5301] describes an IS-IS-specific extension and [RFC5642] describes an OSPF extension for advertisement of Node Name which MAY encoded in the Node Name TLV.

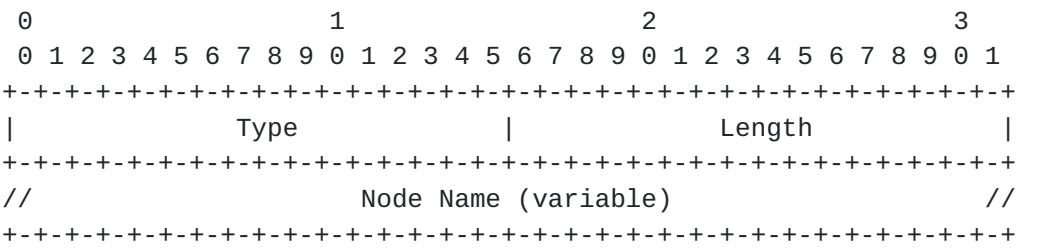


Figure 17: Node Name Format

4.3.1.4. Local IPv4/IPv6 Router-ID TLVs

The local IPv4/IPv6 Router-ID TLVs are used to describe auxiliary Router-IDs that the IGP might be using, e.g., for TE and migration purposes such as correlating a Node-ID between different protocols. If there is more than one auxiliary Router-ID of a given type, then each one is encoded in its own TLV.

4.3.1.5. Opaque Node Attribute TLV

The Opaque Node Attribute TLV is an envelope that transparently carries optional Node Attribute TLVs advertised by a router. An originating router shall use this TLV for encoding information specific to the protocol advertised in the NLRI header Protocol-ID field or new protocol extensions to the protocol as advertised in the NLRI header Protocol-ID field for which there is no protocol-neutral representation in the BGP Link-State NLRI. The primary use of the Opaque Node Attribute TLV is to bridge the document lag between, e.g., a new IGP link-state attribute being defined and the protocol-neutral BGP-LS extensions being published. A router, for example, could use this extension in order to advertise the native protocol's Node Attribute TLVs, such as the OSPF Router Informational Capabilities TLV defined in [RFC7770] or the IGP TE Node Capability Descriptor TLV described in [RFC5073].

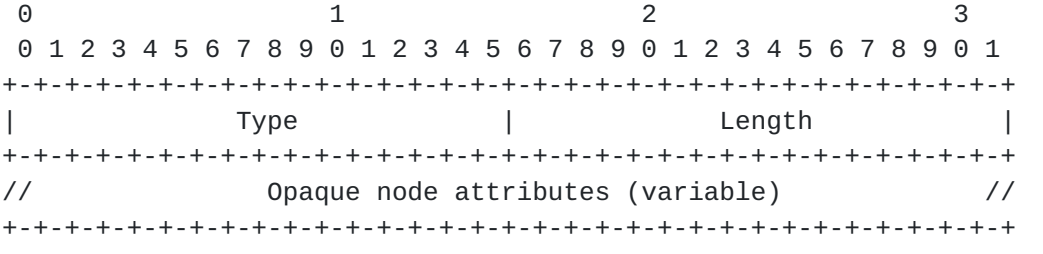


Figure 18: Opaque Node Attribute Format

4.3.2. Link Attribute TLVs

Link Attribute TLVs are TLVs that may be encoded in the BGP-LS attribute with a Link NLRI. Each 'Link Attribute' is a Type/Length/Value (TLV) triplet formatted as defined in Section 4.1. The format and semantics of the Value fields in some Link Attribute TLVs correspond to the format and semantics of the 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 by either IS-IS or OSPF.

The following Link Attribute TLVs are defined for the BGP-LS Attribute associated with a Link NLRI:

TLV Code Point	Description	IS-IS TLV /Sub-TLV	Reference (RFC/Section)
1028	IPv4 Router-ID of Local Node	134/---	[RFC5305] / 4.3
1029	IPv6 Router-ID of Local Node	140/---	[RFC6119] / 4.1
1030	IPv4 Router-ID of Remote Node	134/---	[RFC5305] / 4.3
1031	IPv6 Router-ID of Remote Node	140/---	[RFC6119] / 4.1
1088	Administrative group (color)	22/3	[RFC5305] / 3.1
1089	Maximum link bandwidth	22/9	[RFC5305] / 3.4
1090	Max. reservable link bandwidth	22/10	[RFC5305] / 3.5
1091	Unreserved bandwidth	22/11	[RFC5305] / 3.6
1092	TE Default Metric	22/18	Section 4.3.2.3
1093	Link Protection Type	22/20	[RFC5307] / 1.2
1094	MPLS Protocol Mask	---	Section 4.3.2.2
1095	IGP Metric	---	Section 4.3.2.4
1096	Shared Risk Link Group	---	Section 4.3.2.5
1097	Opaque Link Attribute	---	Section 4.3.2.6
1098	Link Name	---	Section 4.3.2.7

Table 8: Link Attribute TLVs

4.3.2.1. IPv4/IPv6 Router-ID TLVs

The local/remote IPv4/IPv6 Router-ID TLVs are used to describe auxiliary Router-IDs that the IGP might be using, e.g., for TE purposes. All auxiliary Router-IDs of both the local and the remote node MUST be included in the link attribute of each Link NLRI. If there is more than one auxiliary Router-ID of a given type, then multiple TLVs are used to encode them.

4.3.2.2. MPLS Protocol Mask TLV

The MPLS Protocol Mask TLV 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.

Generation of the MPLS Protocol Mask TLV is only valid for and SHOULD only be used with originators that have local link insight, for example, the Protocol-IDs 'Static configuration' or 'Direct' as per Table 2. The MPLS Protocol Mask TLV MUST NOT be included in NLRIs with the other Protocol-IDs listed in Table 2.

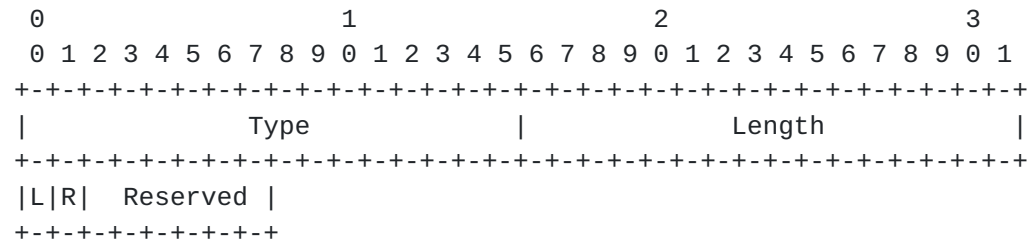


Figure 19: MPLS Protocol Mask TLV

The following bits are defined:

Bit	Description	Reference
'L'	Label Distribution Protocol (LDP)	[RFC5036]
'R'	Extension to RSVP for LSP Tunnels (RSVP-TE)	[RFC3209]
'Reserved'	Reserved for future use	

Table 9: MPLS Protocol Mask TLV Codes

4.3.2.3. TE Default Metric TLV

The TE Default Metric TLV carries the Traffic Engineering metric for this link. The length of this TLV is fixed at 4 octets. If a source protocol uses a metric width of less than 32 bits, then the high-order bits of this field MUST be padded with zero.

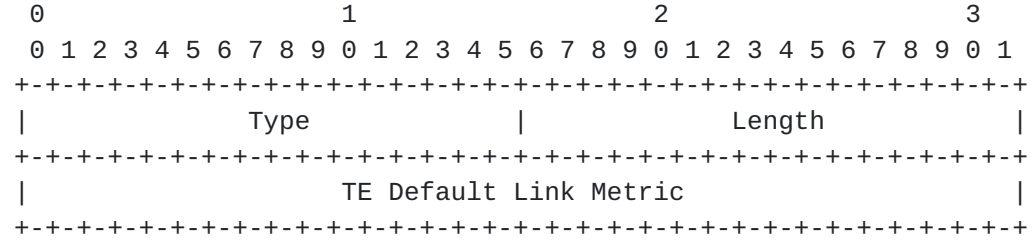


Figure 20: TE Default Metric TLV Format

4.3.2.4. IGP Metric TLV

The IGP Metric TLV carries the metric for this link. The length of this TLV is variable, depending on the metric width of the underlying protocol. IS-IS small metrics have a length of 1 octet. Since the ISIS small metrics are of 6 bit size, the two most significant bits MUST be set to 0 and MUST be ignored by receiver. OSPF link metrics have a length of 2 octets. IS-IS wide metrics have a length of 3 octets.

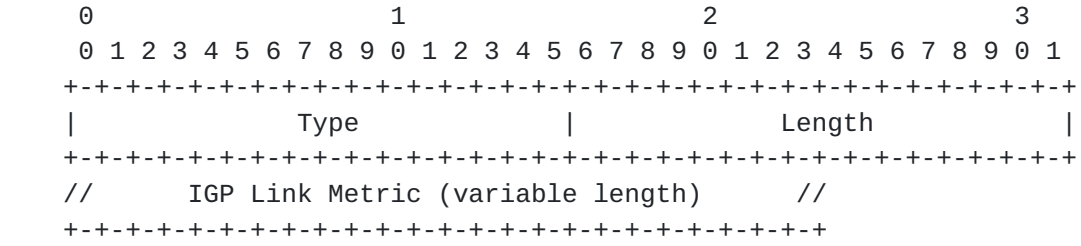


Figure 21: IGP Metric TLV Format

4.3.2.5. Shared Risk Link Group TLV

The Shared Risk Link Group (SRLG) TLV 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 22. The length of this TLV is 4 * (number of SRLG values).

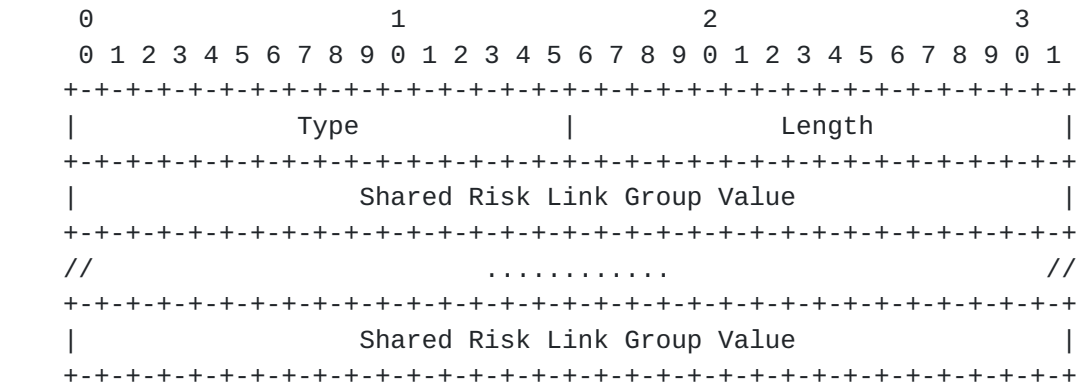


Figure 22: Shared Risk Link Group TLV Format

The SRLG TLV for OSPF-TE is defined in [RFC4203]. 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]. In Link-State NLRI, both IPv4 and IPv6 SRLG information are carried in a single TLV.

4.3.2.6. Opaque Link Attribute TLV

The Opaque Link Attribute TLV is an envelope that transparently carries optional Link Attribute TLVs advertised by a router. An originating router shall use this TLV for encoding information specific to the protocol advertised in the NLRI header Protocol-ID field or new protocol extensions to the protocol as advertised in the NLRI header Protocol-ID field for which there is no protocol-neutral representation in the BGP Link-State NLRI. The primary use of the Opaque Link Attribute TLV is to bridge the document lag between, e.g., a new IGP link-state attribute being defined and the 'protocol-neutral' BGP-LS extensions being published.

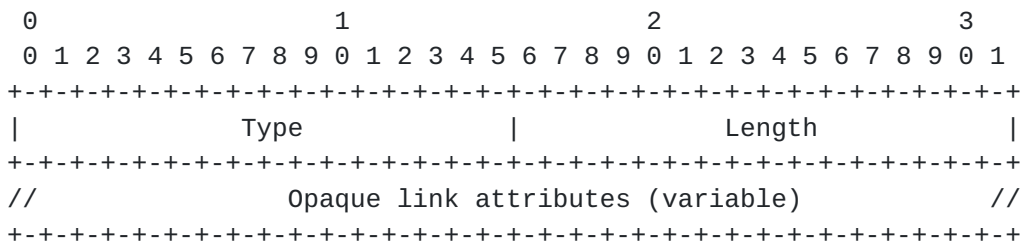


Figure 23: Opaque Link Attribute TLV Format

4.3.2.7. Link Name TLV

The Link Name TLV is optional. The Value field identifies the symbolic name of the router link. This symbolic name can be the FQDN for the link, it can be a subset of the FQDN, or it can be any string operators want to use for the link. The use of FQDN or a subset of it is strongly RECOMMENDED. The maximum length of the Link Name TLV is 255 octets.

The Value field is encoded in 7-bit ASCII. If a user interface for configuring or displaying this field permits Unicode characters, that user interface is responsible for applying the ToASCII and/or ToUnicode algorithm as described in [RFC5890] to achieve the correct format for transmission or display.

How a router derives and injects link names is outside of the scope of this document.

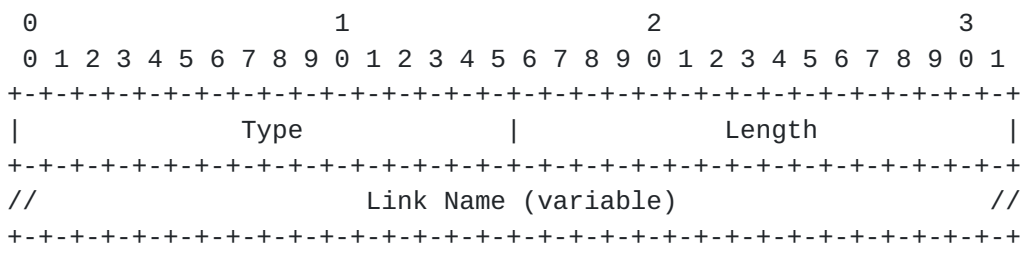


Figure 24: Link Name TLV Format

4.3.3. Prefix Attribute TLVs

Prefixes are learned from the IGP topology (IS-IS or OSPF) with a set of IGP attributes (such as metric, route tags, etc.) that are advertised in the BGP-LS Attribute with Prefix NLRI types 3 and 4.

The following Prefix Attribute TLVs are defined for the BGP-LS Attribute associated with a Prefix NLRI:

TLV Code Point	Description	Length	Reference
1152	IGP Flags	1	Section 4.3.3.1
1153	IGP Route Tag	4*n	[RFC5130]
1154	IGP Extended Route Tag	8*n	[RFC5130]
1155	Prefix Metric	4	[RFC5305]
1156	OSPF Forwarding Address	4	[RFC2328]
1157	Opaque Prefix Attribute	variable	Section 4.3.3.6

Table 10: Prefix Attribute TLVs

4.3.3.1. IGP Flags TLV

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

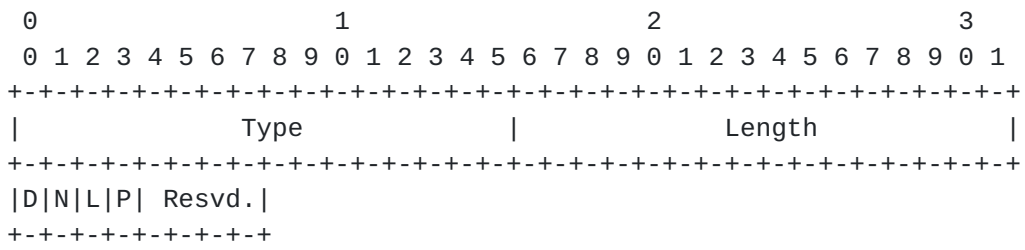


Figure 25: IGP Flag TLV Format

The Value field contains bits defined according to the table below:

Bit	Description	Reference
'D'	IS-IS Up/Down Bit	[RFC5305]
'N'	OSPF "no unicast" Bit	[RFC5340]
'L'	OSPF "local address" Bit	[RFC5340]
'P'	OSPF "propagate NSSA" Bit	[RFC5340]
Reserved	Reserved for future use.	

Table 11: IGP Flag Bits Definitions

4.3.3.2. IGP Route Tag TLV

The IGP Route Tag TLV carries original IGP Tags (IS-IS [RFC5130] or OSPF) of the prefix and is encoded as follows:

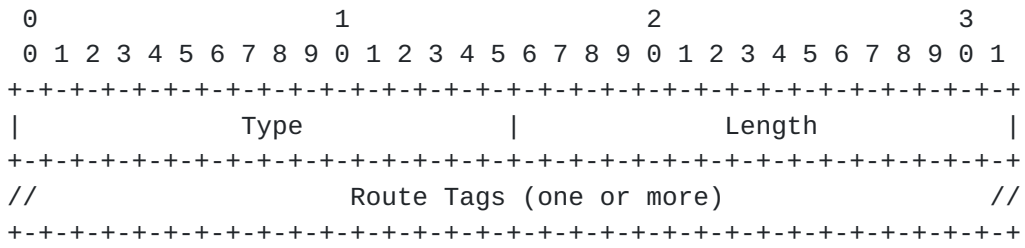


Figure 26: IGP Route Tag TLV Format

Length is a multiple of 4.

The Value field contains one or more Route Tags as learned in the IGP topology.

4.3.3.3. Extended IGP Route Tag TLV

The Extended IGP Route Tag TLV carries IS-IS Extended Route Tags of the prefix [RFC5130] and is encoded as follows:

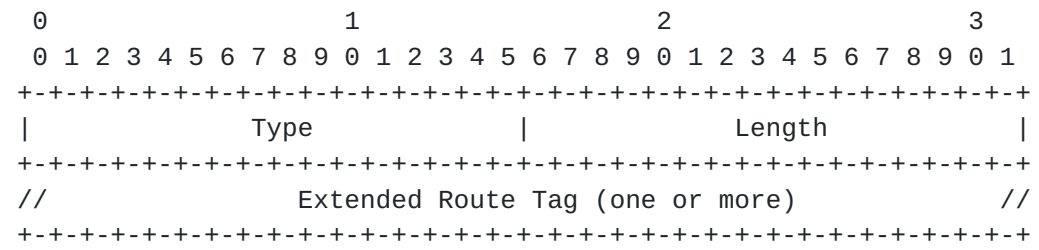


Figure 27: Extended IGP Route Tag TLV Format

Length is a multiple of 8.

The Extended Route Tag field contains one or more Extended Route Tags as learned in the IGP topology.

4.3.3.4. Prefix Metric TLV

The Prefix Metric TLV is an optional attribute and may only appear once. If present, it carries the metric of the prefix as known in the IGP topology as described in Section 4 of [RFC5305] (and therefore represents the reachability cost to the prefix). If not present, it means that the prefix is advertised without any reachability.

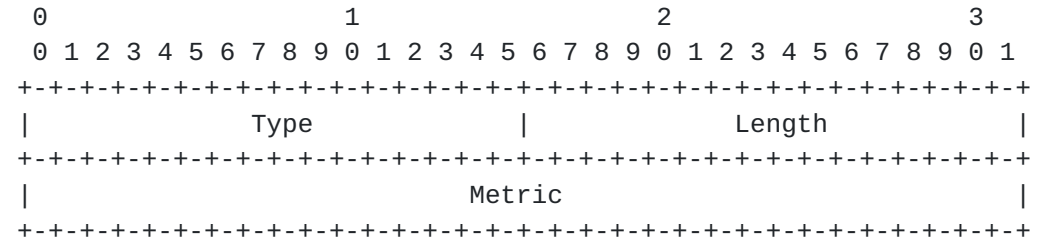


Figure 28: Prefix Metric TLV Format

Length is 4.

4.3.3.5. OSPF Forwarding Address TLV

The OSPF Forwarding Address TLV [RFC2328] [RFC5340] carries the OSPF forwarding address as known in the original OSPF advertisement. Forwarding address can be either IPv4 or IPv6.

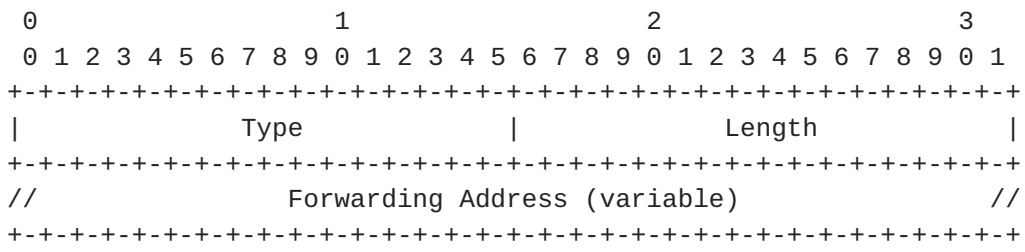


Figure 29: OSPF Forwarding Address TLV Format

Length is 4 for an IPv4 forwarding address, and 16 for an IPv6 forwarding address.

4.3.3.6. Opaque Prefix Attribute TLV

The Opaque Prefix Attribute TLV is an envelope that transparently carries optional Prefix Attribute TLVs advertised by a router. An originating router shall use this TLV for encoding information specific to the protocol advertised in the NLRI header Protocol-ID field or new protocol extensions to the protocol as advertised in the NLRI header Protocol-ID field for which there is no protocol-neutral representation in the BGP Link-State NLRI. The primary use of the Opaque Prefix Attribute TLV is to bridge the document lag between, e.g., a new IGP link-state attribute being defined and the protocol-neutral BGP-LS extensions being published.

The format of the TLV is as follows:

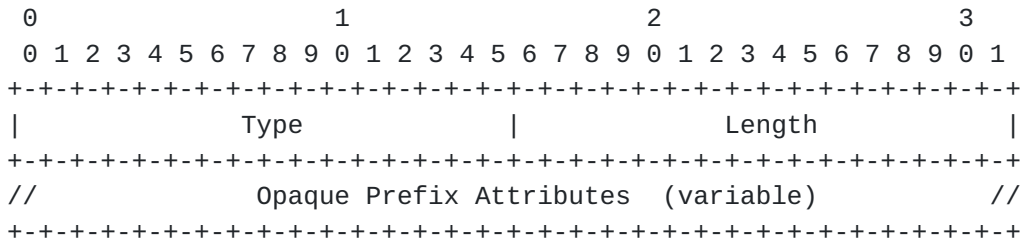


Figure 30: Opaque Prefix Attribute TLV Format

Type is as specified in Table 10. Length is variable.

4.4. Private Use

TLVs for Vendor Private use are supported using the code point range reserved as indicated in Section 6. For such TLV use in the NLRI or BGP-LS Attribute, the format as described in Section 4.1 is to be used and a 4 octet field MUST be included as the first field in the value to carry the Enterprise Code. For a private use NLRI Type, a 4 octet field MUST be included as the first field in the NLRI

immediately following the Total NLRI Length field of the Link-State NLRI format as described in Section 4.2 to carry the Enterprise Code. The Enterprise Codes are listed at <http://www.iana.org/assignments/enterprise-numbers>. This enables use vendor specific extensions without conflicts.

Multiple instances of private-use TLVs MAY appear in the BGP-LS Attribute.

4.5. 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 an IPv4 BGP session is used, then the next hop in the MP_REACH_NLRI SHOULD be an IPv4 address. Similarly, if an 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 endpoint 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]. For VPN Subsequent Address Family Identifier (SAFI), as per custom, an 8-byte Route Distinguisher set to all zero is prepended to the next hop.

The BGP Next Hop attribute is used by each BGP-LS speaker to validate the NLRI it receives. In case identical NLRIs are sourced by multiple BGP-LS Producers, the BGP Next Hop attribute is used to tiebreak as per the standard BGP path decision process. This specification doesn't mandate any rule regarding the rewrite of the BGP Next Hop attribute.

4.6. Inter-AS Links

The main source of TE information is the IGP, which is not active on inter-AS links. In some cases, the IGP may have information of inter-AS links [RFC5392] [RFC5316]. In other cases, an implementation SHOULD provide a means to inject inter-AS links into BGP-LS. The exact mechanism used to provision the inter-AS links is outside the scope of this document and are described in [I-D.ietf-idr-bgpls-inter-as-topology-ext].

4.7. Handling of Unreachable IGP Nodes

The origination and propagation of IGP link-state information via BGP needs to provide a consistent and true view of the topology of the IGP domain. BGP-LS provides an abstraction of the protocol specifics and BGP-LS Consumers may be varied types of applications. While the information propagated via BGP-LS from a link-state routing protocol is sourced from that protocol's LSDB, it does not serve as a true reflection of the originating router's LSDB since it does not include the LSA/LSP sequence number information. The sequence numbers are not included since a single NLRI update may be put together with information that is coming from multiple LSAs/LSPs.

Consider an OSPF network as shown in Figure 31, where R2 and R3 are the BGP-LS Producers and also the OSPF Area Border Routers (ABRs). The link between R2 and R3 is in area 0 while the other links shown are in area 1.

A BGP-LS Consumer talks to a BGP route-reflector (RR) R0 which is aggregating the BGP-LS feed from the BGP-LS Producers R2 and R3. Here R2 and R3 provide a redundant topology feed via BGP-LS to R0. Normally, R0 would receive two identical copies of all the Link-State NLRIs from both R2 and R3 and it would pick one of them (say R2) based on the standard BGP best path decision process.

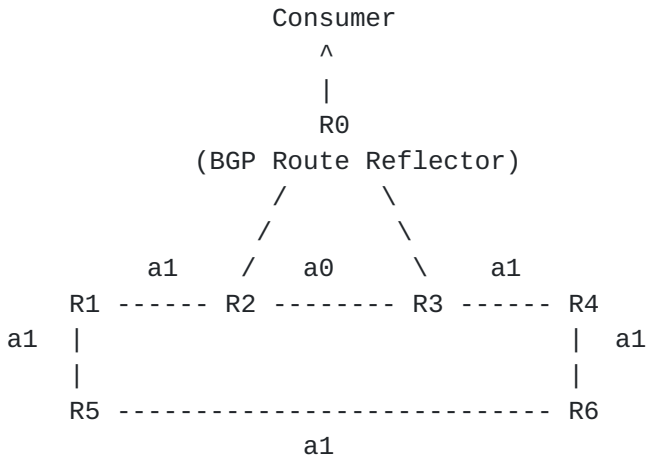


Figure 31: Incorrect Reporting due to BGP Path Selection

Consider a scenario where the link between R5 and R6 is lost (thereby partitioning the area 1) and its impact on the OSPF LSDB at R2 and R3.

Now, R5 will remove the link 5-6 from its Router LSA and this updated LSA is available at R2. R2 also has a stale copy of R6's Router LSA

which still has the link 6-5 in it. Based on this view in its LSDB, R2 will advertise only the half-link 6-5 that it derives from R6's stale Router LSA.

At the same time, R6 has removed the link 6-5 from its Router LSA and this updated LSA is available at R3. Similarly, R3 also has a stale copy of R5's Router LSA having the link 5-6 in it. Based on it's LSDB, R3 will advertise only the half-link 5-6 that it has derived from R5's stale Router LSA.

Now, the BGP-LS Consumer receives both the Link NLRIs corresponding to the half-links from R2 and R3 via R0. When viewed together, it would not detect or realize that the area 1 is actually partitioned. Also if R2 continues to report Link-State NLRIs corresponding to the stale copy of Router LSA of R4 and R6 nodes then R0 would prefer them over the valid Link-State NLRIs for R4 and R6 that it is receiving from R3 based on its BGP decision process. This would result in the BGP-LS Consumer getting stale and inaccurate topology information. This problems scenario is avoided if R2 were to not advertise the link-state information corresponding to R4 and R6 and if R3 were to not advertise similarly for R1 and R5.

A BGP-LS Producer MUST withdraw all link-state objects advertised by it in BGP when the node that originated its corresponding LSP/LSAs is determined to have become unreachable in the IGP and it MUST re-advertise those link-state objects only after that node becomes reachable again in the IGP domain.

4.8. Router-ID Anchoring Example: ISO Pseudonode

Encoding of a broadcast LAN in IS-IS provides a good example of how Router-IDs are encoded. Consider Figure 32. 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. Node1 is the DIS for the LAN. Two unidirectional links (Node1, Pseudonode1) and (Pseudonode1, Node2) are being generated.

The Link NLRI of (Node1, Pseudonode1) is encoded as follows. The IGP Router-ID TLV of the local Node Descriptor is 6 octets long and contains the ISO-ID of Node1, 1920.0000.2001. The IGP Router-ID TLV of the remote Node Descriptor is 7 octets long and contains the ISO-ID of Pseudonode1, 1920.0000.2001.02. The BGP-LS attribute of this link contains one local IPv4 Router-ID TLV (TLV type 1028) containing 192.0.2.1, the IPv4 Router-ID of Node1.

The Link NLRI of (Pseudonode1, Node2) is encoded as follows. The IGP Router-ID TLV of the local Node Descriptor is 7 octets long and

contains the ISO-ID of Pseudonode1, 1920.0000.2001.02. The IGP Router-ID TLV of the remote Node Descriptor is 6 octets long and contains the ISO-ID of Node2, 1920.0000.2002. The BGP-LS attribute of this link contains one remote IPv4 Router-ID TLV (TLV type 1030) containing 192.0.2.2, the IPv4 Router-ID of Node2.

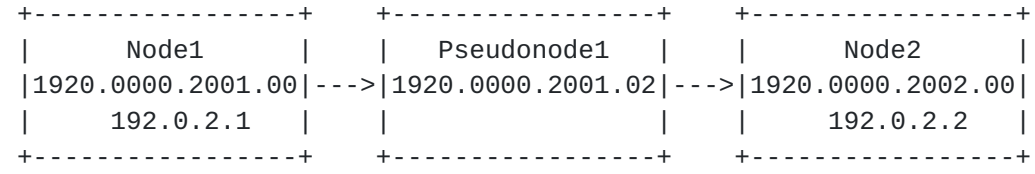


Figure 32: IS-IS Pseudonodes

4.9. Router-ID Anchoring Example: OSPF Pseudonode

Encoding of a broadcast LAN in OSPF provides a good example of how Router-IDs and local Interface IPs are encoded. Consider Figure 33. This represents a Broadcast LAN between a pair of routers. The "real" (non-pseudonode) routers have both an IPv4 Router-ID and an Area Identifier. The pseudonode does have an IPv4 Router-ID, an IPv4 Interface Address (for disambiguation), and an OSPF Area. Node1 is the DR for the LAN; hence, its local IP address 10.1.1.1 is used as both the Router-ID and Interface IP for the pseudonode keys. Two unidirectional links, (Node1, Pseudonode1) and (Pseudonode1, Node2), are being generated.

The Link NLRI of (Node1, Pseudonode1) is encoded as follows:

- Local Node Descriptor

  TLV #515: IGP Router-ID: 11.11.11.11

  TLV #514: OSPF Area-ID: ID:0.0.0.0

- Remote Node Descriptor

  TLV #515: IGP Router-ID: 11.11.11.11:10.1.1.1

  TLV #514: OSPF Area-ID: ID:0.0.0.0

The Link NLRI of (Pseudonode1, Node2) is encoded as follows:

- Local Node Descriptor

  TLV #515: IGP Router-ID: 11.11.11.11:10.1.1.1

  TLV #514: OSPF Area-ID: ID:0.0.0.0

o Remote Node Descriptor

TLV #515: IGP Router-ID: 33.33.33.34

TLV #514: OSPF Area-ID: ID:0.0.0.0

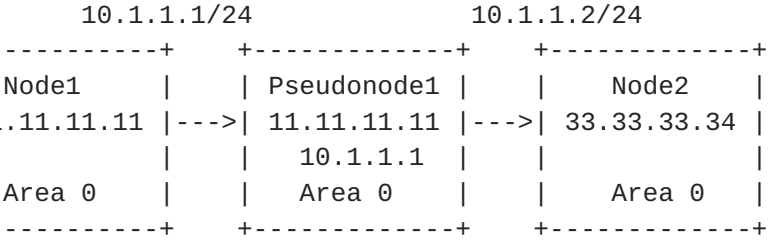


Figure 33: OSPF Pseudonodes

The LAN subnet 10.1.1.0/24 is not included in the Router LSA of Node1 or Node2. The Network LSA for this LAN advertised by the DR Node1 contains the subnet mask for the LAN along with the DR address. A Prefix NLRI corresponding to the LAN subnet is advertised with the Pseudonode1 used as the Local node using the DR address and the subnet mask from the Network LSA.

4.10. Router-ID Anchoring Example: OSPFv2 to IS-IS Migration

Graceful migration from one IGP to another requires coordinated operation of both protocols during the migration period. Such a coordination requires identifying a given physical link in both IGPs. The IPv4 Router-ID provides that "glue", which is present in the Node Descriptors of the OSPF Link NLRI and in the link attribute of the IS-IS Link NLRI.

Consider a point-to-point link between two routers, A and B, that initially were OSPFv2-only routers and then IS-IS is enabled on them. Node A has IPv4 Router-ID and ISO-ID; node B has IPv4 Router-ID, IPv6 Router-ID, and ISO-ID. Each protocol generates one Link NLRI for the link (A, B), both of which are carried by BGP-LS. The OSPFv2 Link NLRI for the link is encoded with the IPv4 Router-ID of nodes A and B in the local and remote Node Descriptors, respectively. The IS-IS Link NLRI for the link is encoded with the ISO-ID of nodes A and B in the local and remote Node Descriptors, respectively. In addition, the BGP-LS attribute of the IS-IS Link NLRI contains the TLV type 1028 containing the IPv4 Router-ID of node A, TLV type 1030 containing the IPv4 Router-ID of node B, and TLV type 1031 containing the IPv6 Router-ID of node B. In this case, by using IPv4 Router-ID, the link (A, B) can be identified in both the IS-IS and OSPF protocol.

5. Link to Path Aggregation

Distribution of all links available in the global Internet is certainly possible; however, it not desirable from a scaling and privacy point of view. Therefore, an implementation may support a 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, etc.) 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 are discussed.

5.1. Example: No Link Aggregation

Consider Figure 34. 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 AS2 and therefore can compute a backup path. Note that the computing router decides if the direct link between {R3, R4} or the {R4, R5, R3} path is used.

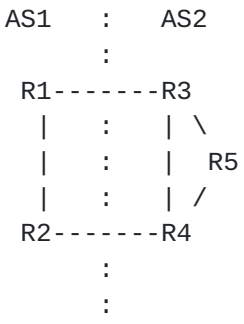


Figure 34: No Link Aggregation

5.2. 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 35. The only link that 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.

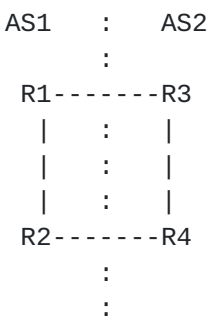


Figure 35: ASBR Link Aggregation

5.3. 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 36. AS3 is modeled as a single node that connects to the border routers of the aggregated domain.



Figure 36: Multi-AS Aggregation

6. IANA Considerations

IANA has assigned address family number 16388 (BGP-LS) in the "Address Family Numbers" registry with [RFC7752] as a reference.

IANA has assigned SAFI values 71 (BGP-LS) and 72 (BGP-LS-VPN) in the "SAFI Values" sub-registry under the "Subsequent Address Family Identifiers (SAFI) Parameters" registry.

IANA has assigned value 29 (BGP-LS Attribute) in the "BGP Path Attributes" sub-registry under the "Border Gateway Protocol (BGP) Parameters" registry.

IANA has created a new "Border Gateway Protocol - Link State (BGP-LS) Parameters" registry at <http://www.iana.org/assignments/bgp-ls-parameters>. All of the following registries are BGP-LS specific and are accessible under this registry:

- "BGP-LS NLRI-Types" registry

  Value 0 is reserved. The maximum value is 65535. The range 65000-65535 is for Private Use. The registry has been populated with the values shown in Table 1. Allocations within the registry under the "Expert Review" policy require documentation of the proposed use of the allocated value and approval by the Designated Expert assigned by the IESG (see [RFC8126]).

- "BGP-LS Protocol-IDs" registry

  Value 0 is reserved. The maximum value is 255. The range 200-255 is for Private Use. The registry has been populated with the values shown in Table 2. Allocations within the registry under the "Expert Review" policy require documentation of the proposed use of the allocated value and approval by the Designated Expert assigned by the IESG (see [RFC8126]).

- "BGP-LS Well-Known Instance-IDs" registry

  This registry was setup via [RFC7752] and is no longer required. It may be retained as deprecated.

- "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs" registry

  Values 0-255 are reserved. Values 256-65535 will be used for code points. The range 65000-65535 is for Private Use. The registry has been populated with the values shown in Table 12. Allocations within the registry under the "Expert Review" policy require documentation of the proposed use of the allocated value and approval by the Designated Expert assigned by the IESG (see [RFC8126]).

6.1. Guidance for Designated Experts

In all cases of review by the Designated Expert (DE) described here, the DE is expected to ascertain the existence of suitable documentation (a specification) as described in [RFC8126]. The DE is also expected to check the clarity of purpose and use of the requested code points. Additionally, the DE must verify that any request for one of these code points has been made available for review and comment within the IETF: the DE will post the request to the IDR Working Group mailing list (or a successor mailing list designated by the IESG). If the request comes from within the IETF, it should be documented in an Internet-Draft. Lastly, the DE must ensure that any other request for a code point does not conflict with work that is active or already published within the IETF.

The IANA also requests (per [RFC8126]) that all registries with "Expert review" allocation policies have a "Change Controller" assigned. For these four registries, the assigned "Change Controllers" are the chairs of the IDR working group or a successor as designated by the IESG.

7. Manageability Considerations

This section is structured as recommended in [RFC5706].

7.1. Operational Considerations

7.1.1. Operations

Existing BGP operational procedures apply. No new operation procedures are defined in this document. It is noted that the NLRI information present in this document carries purely application-level data that has no immediate impact on the corresponding forwarding state computed by BGP. As such, any churn in reachability information has a different impact than regular BGP updates, which need to change the forwarding state for an entire router. It is expected that the distribution of this NLRI SHOULD be handled by dedicated route reflectors in most deployments providing a level of isolation and fault containment between different NLRI types. In the event of dedicated route reflectors not being available, other alternate mechanisms like separation of BGP instances or separate BGP sessions (e.g. using different addresses for peering) for Link-State information distribution SHOULD be used.

7.1.2. Installation and Initial Setup

Configuration parameters defined in Section 7.2.3 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.

7.1.3. Migration Path

The proposed extension is only activated between BGP peers after capability negotiation. Moreover, the extensions can be turned on/ off on an individual peer basis (see Section 7.2.3), so the extension can be gradually rolled out in the network.

7.1.4. 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.

7.1.5. 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.

7.1.6. 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.

7.2. Management Considerations

7.2.1. Management Information

The IDR working group has documented and continues to document parts of the Management Information Base and YANG models for managing and monitoring BGP speakers and the sessions between them. It is currently believed that the BGP session running BGP-LS is not substantially different from any other BGP session and can be managed using the same data models.

7.2.2. Fault Management

This section describes the fault management actions, as described in [RFC7606] , that are to be performed for handling of BGP update messages for BGP-LS.

A Link-State NLRI MUST NOT be considered as malformed or invalid based on the inclusion/exclusion of TLVs or contents of the TLV fields (i.e. semantic errors), as described in Section 4.1 and Section 4.2.

A BGP-LS Speaker MUST perform the following syntactic validation of the Link-State NLRI to determine if it is malformed.

- Does the sum of all TLVs found in the BGP MP_REACH_NLRI attribute correspond to the BGP MP_REACH_NLRI length?

- Does the sum of all TLVs found in the BGP MP_UNREACH_NLRI attribute correspond to the BGP MP_UNREACH_NLRI length?

- Does the sum of all TLVs found in a Link-State NLRI correspond to the Total NLRI Length field of all its Descriptors?

- Is the length of the TLVs and, when the TLV is recognized then, its sub-TLVs in the NLRI valid?

- Has the syntactic correctness of the NLRI fields been verified as per [RFC7606]?

- Has the rule regarding ordering of TLVs been followed as described in Section 4.1?

When the error determined allows for the router to skip the malformed NLRI(s) and continue processing of the rest of the update message (e.g. when the TLV ordering rule is violated), then it MUST handle such malformed NLRIs as 'Treat-as-withdraw'. In other cases, where the error in the NLRI encoding results in the inability to process the BGP update message (e.g. length related encoding errors), then the router SHOULD handle such malformed NLRIs as 'AFI/SAFI disable' when other AFI/SAFI besides BGP-LS are being advertised over the same session. Alternately, the router MUST perform 'session reset' when the session is only being used for BGP-LS or when it 'AFI/SAFI disable' action is not possible.

A BGP-LS Attribute MUST NOT be considered as malformed or invalid based on the inclusion/exclusion of TLVs or contents of the TLV fields (i.e. semantic errors), as described in Section 4.1 and Section 4.3.

A BGP-LS Speaker MUST perform the following syntactic validation of the BGP-LS Attribute to determine if it is malformed.

- Does the sum of all TLVs found in the BGP-LS Attribute correspond to the BGP-LS Attribute length?

- Has the syntactic correctness of the Attributes (including BGP-LS Attribute) been verified as per [RFC7606]?

- Is the length of each TLV and, when the TLV is recognized then, its sub-TLVs in the BGP-LS Attribute valid?

When the error determined allows for the router to skip the malformed BGP-LS Attribute and continue processing of the rest of the update message (e.g. when the BGP-LS Attribute length and the total Path Attribute Length are correct but some TLV/sub-TLV length within the BGP-LS Attribute is invalid), then it MUST handle such malformed BGP-LS Attribute as 'Attribute Discard'. In other cases, where the error in the BGP-LS Attribute encoding results in the inability to process the BGP update message then the handling is the same as described above for the malformed NLRI.

Note that the 'Attribute Discard' action results in the loss of all TLVs in the BGP-LS Attribute and not the removal of a specific malformed TLV. The removal of specific malformed TLVs may give a wrong indication to a BGP-LS Consumer of that specific information being deleted or not available.

When a BGP Speaker receives an update message with Link-State NLRI(s) in the MP_REACH_NLRI but without the BGP-LS Attribute, it is most likely an indication that a BGP Speaker preceding it has performed the 'Attribute Discard' fault handling. An implementation SHOULD preserve and propagate the Link-State NLRIs in such an update message so that the BGP-LS Consumers can detect the loss of link-state information for that object and not assume its deletion/withdraw. This also makes it possible for a network operator to trace back to the BGP-LS Propagator which actually detected a fault with the BGP-LS Attribute.

An implementation SHOULD log an error for any errors found during syntax validation for further analysis.

A BGP-LS Propagator SHOULD NOT perform semantic validation of the Link-State NLRI or the BGP-LS Attribute to determine if it is malformed or invalid. Some types of semantic validation that are not to be performed by a BGP-LS Propagator are as follows (and this is not to be considered as an exhaustive list):

- is a mandatory TLV present or not?
- is the length of a fixed length TLV correct or the length of a variable length TLV a valid/permissible?
- are the values of TLV fields valid or permissible?
- are the inclusion and use of TLVs/sub-TLVs with specific Link-State NLRI types valid?

Each TLV MAY indicate the valid and permissible values and their semantics that can to be used only by a BGP-LS Consumer for its

semantic validation. However, the handling of any errors may be specific to the particular application and outside the scope of this document. A BGP-LS Consumer should ignore unrecognized and unexpected TLV types in both the NLRI and BGP-LS Attribute portions and not consider their presence as an error.

7.2.3. 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 a router's Routing Information Base (RIB).

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

An implementation SHOULD allow the operator to configure a 64-bit Instance-ID.

An implementation SHOULD allow the operator to configure ASN and BGP-LS identifiers (refer Section 4.2.1.4).

An implementation SHOULD allow the operator to configure the maximum size of the BGP-LS Attribute that may be used on a BGP-LS Producer.

7.2.4. Accounting Management

Not Applicable.

7.2.5. 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

These statistics should be recorded as absolute counts since system or session start time. An implementation MAY also enhance this information by recording peak per-second counts in each case.

7.2.6. Security Management

An operator SHOULD define an import policy to limit inbound updates as follows:

- Drop all updates from peers that are only serving BGP-LS Consumers.

An implementation MUST have the means to limit inbound updates.

8. TLV/Sub-TLV Code Points Summary

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

TLV Code Point	Description	IS-IS TLV/ Sub-TLV	Reference (RFC/Section)
256	Local Node Descriptors	---	Section 4.2.1.2
257	Remote Node Descriptors	---	Section 4.2.1.3
258	Link Local/Remote Identifiers	22/4	[RFC5307] / 1.1
259	IPv4 interface address	22/6	[RFC5305] / 3.2
260	IPv4 neighbor address	22/8	[RFC5305] / 3.3
261	IPv6 interface address	22/12	[RFC6119] / 4.2
262	IPv6 neighbor address	22/13	[RFC6119] / 4.3
263	Multi-Topology ID	---	Section 4.2.2.1
264	OSPF Route Type	---	Section 4.2.3
265	IP Reachability Information	---	Section 4.2.3
512	Autonomous System	---	Section 4.2.1.4
513	BGP-LS Identifier (deprecated)	---	Section 4.2.1.4
514	OSPF Area-ID	---	Section 4.2.1.4
515	IGP Router-ID	---	Section 4.2.1.4
1024	Node Flag Bits	---	Section 4.3.1.1
1025	Opaque Node	---	Section 4.3.1.5

	Attribute		
1026	Node Name	variable	Section 4.3.1.3
1027	IS-IS Area Identifier	variable	Section 4.3.1.2
1028	IPv4 Router-ID of Local Node	134/---	[RFC5305] / 4.3
1029	IPv6 Router-ID of Local Node	140/---	[RFC6119] / 4.1
1030	IPv4 Router-ID of Remote Node	134/---	[RFC5305] / 4.3
1031	IPv6 Router-ID of Remote Node	140/---	[RFC6119] / 4.1
1088	Administrative group (color)	22/3	[RFC5305] / 3.1
1089	Maximum link bandwidth	22/9	[RFC5305] / 3.4
1090	Max. reservable link bandwidth	22/10	[RFC5305] / 3.5
1091	Unreserved bandwidth	22/11	[RFC5305] / 3.6
1092	TE Default Metric	22/18	Section 4.3.2.3
1093	Link Protection Type	22/20	[RFC5307] / 1.2
1094	MPLS Protocol Mask	---	Section 4.3.2.2
1095	IGP Metric	---	Section 4.3.2.4
1096	Shared Risk Link Group	---	Section 4.3.2.5
1097	Opaque Link Attribute	---	Section 4.3.2.6
1098	Link Name	---	Section 4.3.2.7
1152	IGP Flags	---	Section 4.3.3.1
1153	IGP Route Tag	---	[RFC5130]
1154	IGP Extended Route Tag	---	[RFC5130]
1155	Prefix Metric	---	[RFC5305]
1156	OSPF Forwarding Address	---	[RFC2328]
1157	Opaque Prefix Attribute	---	Section 4.3.3.6

Table 12: Summary Table of TLV/Sub-TLV Code Points

9. Security Considerations

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

section of [RFC4271] for a discussion of BGP security. Also refer to [RFC4272] and [RFC6952] for analysis of security issues for BGP.

In the context of the BGP peerings associated with this document, a BGP speaker MUST NOT accept updates from a peer that is only providing information to a BGP-LS Consumer. That is, a participating BGP speaker should be aware of the nature of its relationships for link-state relationships and should protect itself from peers sending updates that either represent erroneous information feedback loops or are false input. Such protection can be achieved by manual configuration of consumer peers at the BGP speaker.

An operator SHOULD employ a mechanism to protect a BGP speaker against DDoS attacks from BGP-LS Consumers. The principal attack a consumer may apply is to attempt to start multiple sessions either sequentially or simultaneously. Protection can be applied by imposing rate limits.

Additionally, it may be considered that the export of link-state and TE information as described in this document constitutes a risk to confidentiality of mission-critical or commercially sensitive information about the network. BGP peerings are not automatic and require configuration; thus, it is the responsibility of the network operator to ensure that only trusted consumers are configured to receive such information.

10. Contributors

The following persons contributed significant text to RFC7752 and this document. They should be considered as co-authors.

Hannes Gredler
Rtbrick
Email: hannes@rtbrick.com

Jan Medved
Cisco Systems Inc.
USA
Email: jmedved@cisco.com

Stefano Previdi
Huawei Technologies
Italy
Email: stefano@previdi.net

Adrian Farrel
Old Dog Consulting
Email: adrian@olddog.co.uk

Saikat Ray
Individual
USA
Email: raysaikat@gmail.com

11. Acknowledgements

This document update to the BGP-LS specification [RFC7752] is a result of feedback and inputs from the discussions in the IDR working group. It also incorporates certain details and clarifications based on implementation and deployment experience with BGP-LS.

Cengiz Alaettinoglu and Parag Amritkar brought forward the need to clarify the advertisement of LAN subnet for OSPF.

We would like to thank Balaji Rajagopalan, Srihari Sangli, Shraddha Hegde, Andrew Stone, Jeff Tantsura, Acee Lindem, Jie Dong, Aijun Wang and Nandan Saha for their review and feedback on this document.

We would like to thank Robert Varga for the significant contribution he gave to RFC7752.

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, Matt Miller, Mike Shand, Peter Psenak, Rex Fernando, Richard Woundy, Steven Luong, Tamas Mondal, Waqas Alam, Vipin Kumar, Naiming Shen, Carlos Pignataro, Balaji Rajagopalan, Yakov Rekhter, Alvaro Retana, Barry Leiba, and Ben Campbell for their comments on RFC7752.

12. References

12.1. Normative References

[ISO10589] International Organization for Standardization, "Intermediate System to Intermediate System intra-domain routeing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode network service (ISO 8473)", ISO/IEC 10589, November 2002.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <https://www.rfc-editor.org/info/rfc2328>.

[RFC2545] Marques, P. and F. Dupont, "Use of BGP-4 Multiprotocol Extensions for IPv6 Inter-Domain Routing", RFC 2545, DOI 10.17487/RFC2545, March 1999, <https://www.rfc-editor.org/info/rfc2545>.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <https://www.rfc-editor.org/info/rfc3209>.

[RFC4202] Kompella, K., Ed. and Y. Rekhter, Ed., "Routing Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4202, DOI 10.17487/RFC4202, October 2005, <https://www.rfc-editor.org/info/rfc4202>.

[RFC4203] Kompella, K., Ed. and Y. Rekhter, Ed., "OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, DOI 10.17487/RFC4203, October 2005, <https://www.rfc-editor.org/info/rfc4203>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <https://www.rfc-editor.org/info/rfc4760>.

[RFC4915] Psenak, P., Mirtorabi, S., Roy, A., Nguyen, L., and P. Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF", RFC 4915, DOI 10.17487/RFC4915, June 2007, <https://www.rfc-editor.org/info/rfc4915>.

[RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "LDP Specification", RFC 5036, DOI 10.17487/RFC5036, October 2007, <https://www.rfc-editor.org/info/rfc5036>.

[RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, DOI 10.17487/RFC5120, February 2008, <https://www.rfc-editor.org/info/rfc5120>.

[RFC5130] Previdi, S., Shand, M., Ed., and C. Martin, "A Policy Control Mechanism in IS-IS Using Administrative Tags", RFC 5130, DOI 10.17487/RFC5130, February 2008, <https://www.rfc-editor.org/info/rfc5130>.

[RFC5301] McPherson, D. and N. Shen, "Dynamic Hostname Exchange Mechanism for IS-IS", RFC 5301, DOI 10.17487/RFC5301, October 2008, <https://www.rfc-editor.org/info/rfc5301>.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <https://www.rfc-editor.org/info/rfc5305>.

[RFC5307] Kompella, K., Ed. and Y. Rekhter, Ed., "IS-IS Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 5307, DOI 10.17487/RFC5307, October 2008, <https://www.rfc-editor.org/info/rfc5307>.

[RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, DOI 10.17487/RFC5340, July 2008, <https://www.rfc-editor.org/info/rfc5340>.

[RFC5642] Venkata, S., Harwani, S., Pignataro, C., and D. McPherson, "Dynamic Hostname Exchange Mechanism for OSPF", RFC 5642, DOI 10.17487/RFC5642, August 2009, <https://www.rfc-editor.org/info/rfc5642>.

[RFC5890] Klensin, J., "Internationalized Domain Names for Applications (IDNA): Definitions and Document Framework", RFC 5890, DOI 10.17487/RFC5890, August 2010, <https://www.rfc-editor.org/info/rfc5890>.

[RFC6119] Harrison, J., Berger, J., and M. Bartlett, "IPv6 Traffic Engineering in IS-IS", RFC 6119, DOI 10.17487/RFC6119, February 2011, <https://www.rfc-editor.org/info/rfc6119>.

[RFC6549] Lindem, A., Roy, A., and S. Mirtorabi, "OSPFv2 Multi-Instance Extensions", RFC 6549, DOI 10.17487/RFC6549, March 2012, <https://www.rfc-editor.org/info/rfc6549>.

[RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015, <https://www.rfc-editor.org/info/rfc7606>.

[RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <https://www.rfc-editor.org/info/rfc7752>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8202] Ginsberg, L., Previdi, S., and W. Henderickx, "IS-IS Multi-Instance", RFC 8202, DOI 10.17487/RFC8202, June 2017, <https://www.rfc-editor.org/info/rfc8202>.

[RFC8654] Bush, R., Patel, K., and D. Ward, "Extended Message Support for BGP", RFC 8654, DOI 10.17487/RFC8654, October 2019, <https://www.rfc-editor.org/info/rfc8654>.

12.2. Informative References

[I-D.ietf-idr-bgpls-inter-as-topology-ext] Wang, A., Chen, H., Talaulikar, K., Zhuang, S., and S. Ma, "BGP-LS Extension for Inter-AS Topology Retrieval", draft-ietf-idr-bgpls-inter-as-topology-ext-07 (work in progress), September 2019.

[RFC1918] Rekhter, Y., Moskowitz, B., Karrenberg, D., de Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, DOI 10.17487/RFC1918, February 1996, <https://www.rfc-editor.org/info/rfc1918>.

[RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <https://www.rfc-editor.org/info/rfc4272>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, DOI 10.17487/RFC4655, August 2006, <https://www.rfc-editor.org/info/rfc4655>.

[RFC5073] Vasseur, J., Ed. and J. Le Roux, Ed., "IGP Routing Protocol Extensions for Discovery of Traffic Engineering Node Capabilities", RFC 5073, DOI 10.17487/RFC5073, December 2007, <https://www.rfc-editor.org/info/rfc5073>.

[RFC5152] Vasseur, JP., Ed., Ayyangar, A., Ed., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, DOI 10.17487/RFC5152, February 2008, <https://www.rfc-editor.org/info/rfc5152>.

[RFC5316] Chen, M., Zhang, R., and X. Duan, "ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5316, DOI 10.17487/RFC5316, December 2008, <https://www.rfc-editor.org/info/rfc5316>.

[RFC5392] Chen, M., Zhang, R., and X. Duan, "OSPF Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5392, DOI 10.17487/RFC5392, January 2009, <https://www.rfc-editor.org/info/rfc5392>.

[RFC5693] Seedorf, J. and E. Burger, "Application-Layer Traffic Optimization (ALTO) Problem Statement", RFC 5693, DOI 10.17487/RFC5693, October 2009, <https://www.rfc-editor.org/info/rfc5693>.

[RFC5706] Harrington, D., "Guidelines for Considering Operations and Management of New Protocols and Protocol Extensions", RFC 5706, DOI 10.17487/RFC5706, November 2009, <https://www.rfc-editor.org/info/rfc5706>.

[RFC6952] Jethanandani, M., Patel, K., and L. Zheng, "Analysis of BGP, LDP, PCEP, and MSDP Issues According to the Keying and Authentication for Routing Protocols (KARP) Design Guide", RFC 6952, DOI 10.17487/RFC6952, May 2013, <https://www.rfc-editor.org/info/rfc6952>.

[RFC7285] Alimi, R., Ed., Penno, R., Ed., Yang, Y., Ed., Kiesel, S., Previdi, S., Roome, W., Shalunov, S., and R. Woundy, "Application-Layer Traffic Optimization (ALTO) Protocol", RFC 7285, DOI 10.17487/RFC7285, September 2014, <https://www.rfc-editor.org/info/rfc7285>.

[RFC7770] Lindem, A., Ed., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 7770, DOI 10.17487/RFC7770, February 2016, <https://www.rfc-editor.org/info/rfc7770>.

Appendix A. Changes from RFC 7752

This section lists the high-level changes from RFC 7752 and provides reference to the document sections wherein those have been introduced.

1. Update the Figure 1 in Section 1 and added Section 3 to illustrate the different roles of a BGP implementation in conveying link-state information.

2. In Section 4.1, clarification about the TLV handling aspects that are applicable to both the NLRI and BGP-LS Attribute parts and those that are applicable only for the NLRI portion. An implementation may have missed the part about handling of unrecognized TLV and so, based on [RFC7606] guidelines, might discard the unknown NLRI types. This aspect is now unambiguously clarified in Section 4.2. Also, the TLVs in the BGP-LS Attribute that are not ordered are not to be considered as malformed.

3. Clarification of mandatory and optional TLVs in both NLRI and BGP-LS Attribute portions all through the document.

4. Handling of the growth of the BGP-LS Attribute is covered in Section 4.3.

5. Clarified that the document describes the NLRI descriptor TLVs for the protocols and NLRI types specified in this document and future BGP-LS extensions must describe the same for other protocols and NLRI types that they introduce.

6. Clarification on the use of Identifier field in the Link-State NLRI in Section 4.2 is provided. It was defined ambiguously to refer to only mutli-instance IGP on a single link while it can also be used for multiple IGP protocol instances on a router. The IANA registry is accordingly being removed.

7. The BGP-LS Identifier TLV in the Node Descriptors has been deprecated. Its use was not well specified by [RFC7752] and there has been some amount of confusion between implementators on its usage for identification of IGP domains as against the use of the Identifier doing the same functionality as the Instance-ID when running multiple instances of IGP routing protocols.

8. Clarification that the Area-ID TLV is mandatory in the Node Descriptor for origination of information from OSPF except for

when sourcing information from AS-scope LSAs where this TLV is not applicable.

9. Moved MT-ID TLV from the Node Descriptor section to under the Link Descriptor section since it is not a Node Descriptor sub-TLV. Fixed the ambiguity in the encoding of OSPF MT-ID in this TLV. Updated the IS-IS specification reference section and describe the differences in the applicability of the R flags when MT-ID TLV is used as link descriptor TLV and Prefix Attribute TLV. MT-ID TLV use is now elevated to SHOULD when it is enabled in the underlying IGP.

10. Clarified that IPv6 Link-Local Addresses are not advertised in the Link Descriptor TLVs and the local/remote identifiers are to be used instead for links with IPv6 link-local addresses only.

11. Update the usage of OSPF Route Type TLV to mandate its use for OSPF prefixes in Section 4.2.3.1 since this is required for segregation of intra-area prefixes that are used to reach a node (e.g. a loopback) from other types of inter-area and external prefixes.

12. Clarification on the length of the Node Flag Bits TLV to be one octet.

13. Updated the Node Name TLV in Section 4.3.1.3 with the OSPF specification.

14. Clarification on the size of the IS-IS Narrow Metric advertisement via the IGP Metric TLV and the handling of the unused bits.

15. Clarified the advertisement of the prefix corresponding to the LAN segment in an OSPF network in Section 4.9.

16. Introduced Private Use TLV code point space and specified their encoding in Section 4.4.

17. Introduced Section 4.7 where issues related to consistency of reporting IGP link-state along with their solutions are covered.

18. Handling of large size of BGP-LS Attribute with growth in BGP-LS information is explained in Section 4.3 along with mitigation of errors arising out of it.

19. Added recommendation for isolation of BGP-LS sessions from other BGP route exchange to avoid errors and faults in BGP-LS affecting the normal BGP routing.

20. Updated the Fault Management section with detailed rules based on the role in the BGP-LS information propagation flow.

21. Change to the management of BGP-LS IANA registries from "Specification Required" to "Expert Review" along with updated guidelines for Designated Experts.

Author's Address

Ketan Talaulikar (editor)
Cisco Systems
India

Email: ketant@cisco.com