Internet Engineering Task Force (IETF)
Request for Comments: 9552
Obsoletes: 7752, 9029
Category: Standards Track
ISSN: 2070-1721

K. Talaulikar, Ed.
Cisco Systems

December 2023

Distribution of Link-State and Traffic Engineering Information Using BGP

Abstract

In many environments, a component external to a network is called upon to perform computations based on the network topology and the 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 BGP Network Layer Reachability Information (NLRI) encoding format. The mechanism applies to physical and virtual (e.g., tunnel) 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 some small changes and clarifications to the previous specification. This document also obsoletes RFC 9029 by incorporating the updates that it made to RFC 7752.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at https://www.rfc-editor.org/info/rfc9552.

Table of Contents

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) 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 crosses the visibility of more than one TED or that requires 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 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 BGP Network Layer Reachability Information (NLRI) encoding format. The mechanism applies to physical and virtual (e.g., tunnel) 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 IGP metric, link 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 - Link State (BGP-LS) 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 Figure 1 below.

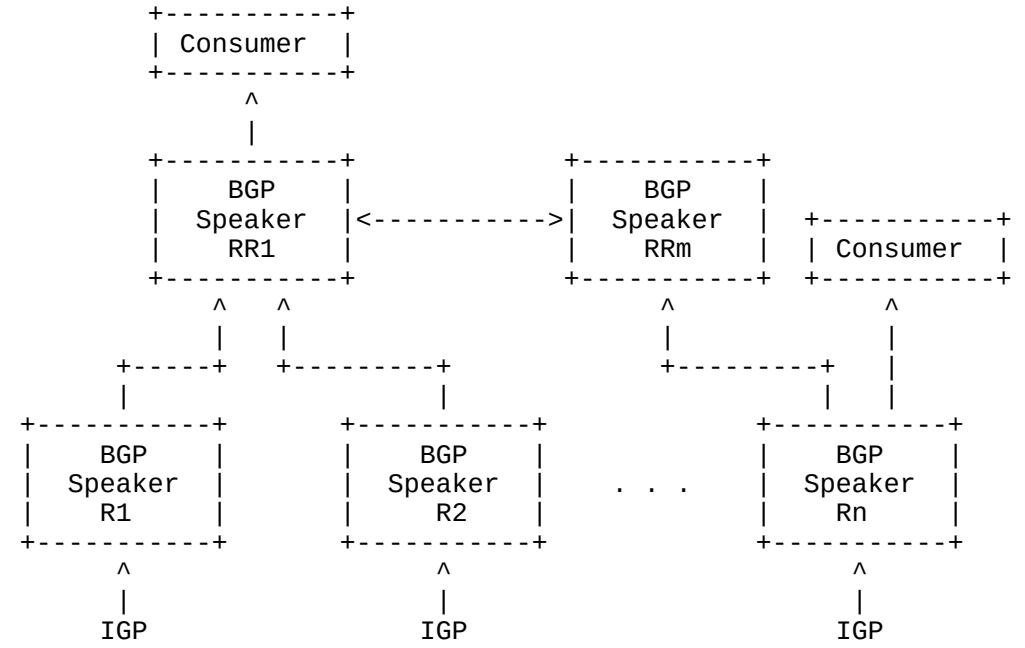


Figure 1: Collection of Link-State and TE Information

A BGP Speaker may apply a 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 in 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 focuses on the specifications related to the origination of IGP-derived information and their propagation via BGP-LS. It also describes the advertisement into BGP-LS of information, either configured or derived, that is local to a node. In general, the procedures in this document form part of the base BGP-LS protocol specification and apply to information from other sources that are introduced into BGP-LS.

This document obsoletes [RFC7752] by completely replacing that document. It makes some 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 an 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 selects the exit ABR and other ABRs or AS Border Routers (ASBRs) as loose-hops [RFC3209] and using the IGP-computed shortest path topology for the remainder of the path. This may lead to suboptimal paths, makes alternate/back-up path computation hard, and

might result in no TE path being found when one 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 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 suboptimal 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.

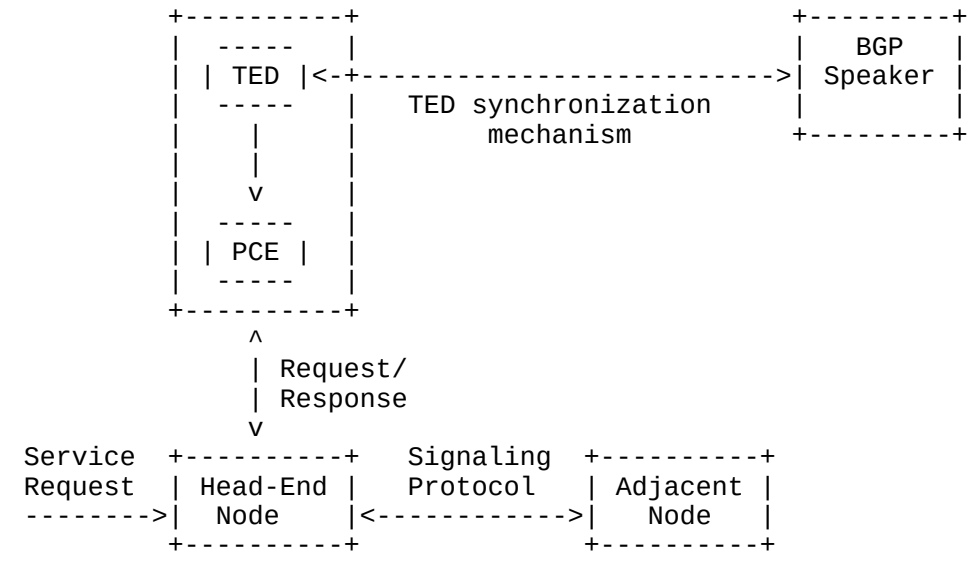


Figure 2: External PCE Node Using a TED Synchronization Mechanism

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 the 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 prefixes 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.

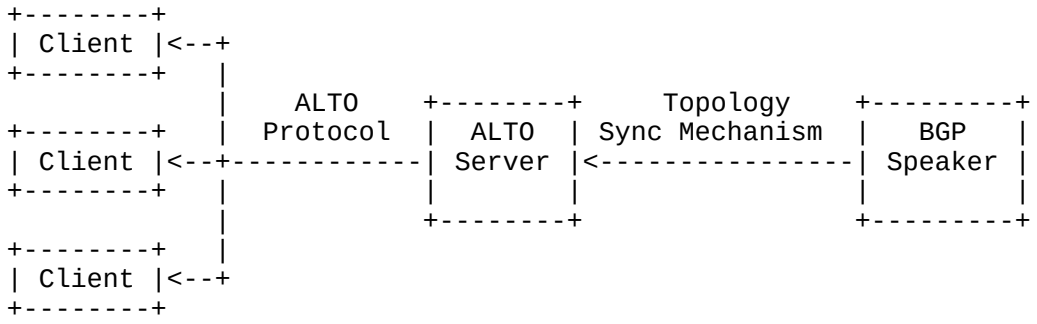


Figure 3: ALTO Server Using Network Topology Information

3. BGP Speaker Roles for BGP-LS

In 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 that are then used throughout the rest of this document.

BGP-LS Producer: The term BGP-LS Producer refers to a BGP Speaker that is originating link-state information into BGP. 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 flooding domain, then they would ordinarily originate the same link-state information into BGP-LS. R1 may also originate information from sources other than IGP, e.g., its local node information.

BGP-LS Consumer: The term BGP-LS Consumer refers to a consumer application/process and not a BGP Speaker. 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. This document only covers the BGP implementation. The consumer application and the design of the interface between BGP and the consumer application may be implementation specific and are outside the scope of this document. The communication of information MUST be unidirectional (i.e., from a BGP Speaker to the BGP-LS Consumer application), and a BGP-LS Consumer MUST NOT be able to send information to a BGP Speaker for origination into BGP-LS.

BGP-LS Propagator: The term BGP-LS Propagator refers to a BGP Speaker that is performing BGP protocol processing on the link-state information. BGP Speaker RRm propagates the BGP-LS information between BGP Speaker Rn and BGP Speaker RR1. The BGP implementation on RRm is propagating BGP-LS information. It performs handling of BGP-LS UPDATE messages and performs the BGP Decision Process as part of deciding what information is to be propagated. Similarly, 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 Decision Process.

The above roles are not mutually exclusive. The same BGP Speaker may be the BGP-LS Producer for some link-state information and BGP-LS Propagator for some other link-state information while also providing this information to a BGP-LS Consumer.

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. Advertising IGP Information into BGP-LS

The origination and propagation of IGP link-state information via BGP needs to provide a consistent and accurate view of the topology of the IGP domain. BGP-LS provides an abstraction of the IGP specifics, and BGP-LS Consumers may be varied types of applications.

The link-state information advertised in BGP-LS from the IGPs is derived from the IGP LSDB built using the OSPF Link-State Advertisements (LSAs) or the IS-IS Link-State Packets (LSPs). However, it does not serve as a verbatim reflection of the originating router's LSDB. It does not include the LSA/LSP sequence number information since a single link-state object may be put together with information that is coming from multiple LSAs/LSPs. Also, not all of the information carried in LSAs/LSPs may be required or suitable for advertisement via BGP-LS (e.g., ASBR reachability in OSPF, OSPF virtual links, link-local-scoped information, etc.). The LSAs/LSPs that are purged or aged out are not included in the BGP-LS advertisement even though they may be present in the LSDB (e.g., for the IGP flooding purposes). The information from the LSAs/LSPs that is invalid or malformed or that which needs to be ignored per the respective IGP protocol specifications are also not included in the BGP-LS advertisement.

The details of the interface between IGPs and BGP for the advertisement of link-state information are outside the scope of this document. In some cases, the information derived from IGP processing (e.g., combination of link-state object from across multiple LSAs/ LSPs, leveraging reachability and two-way connectivity checks, etc.) is required for the advertisement of link-state information into BGP-LS.

5. Carrying Link-State Information in BGP

The link-state information is carried in BGP UPDATE messages as: (1) BGP NLRI information carried within MP_REACH_NLRI and MP_UNREACH_NLRI attributes that describes link, node, or prefix objects and (2) a BGP path attribute (BGP-LS Attribute) that carries properties of the link, node, or prefix objects such as the link and prefix metric, 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 for a BGP-LS Producer to originate link-state information into BGP-LS.

5.1. TLV Format

Information in the 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.

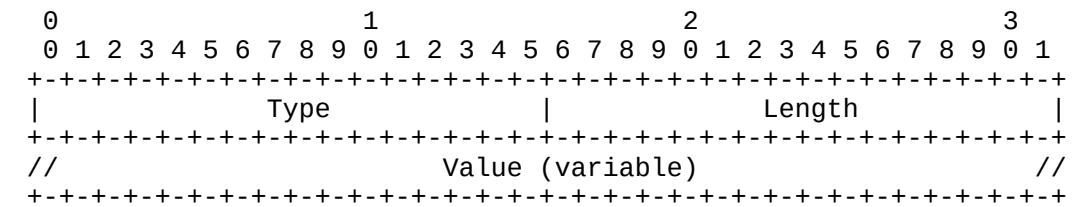


Figure 4: TLV Format

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 unknown or unexpected TLVs MUST NOT result in the NLRI or the BGP-LS Attribute being considered malformed. An example of an unexpected TLV is when a TLV is received along with an update for a link-state object other than the one that the TLV is specified as associated with.

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 first in ascending order based on the Length field followed by ascending order based on the Value field. Comparison of the Value fields is performed by treating the entire field as opaque binary data and ordered lexicographically (i.e., treating each byte of binary data as a symbol to compare, with the symbols ordered by their numerical value). NLRIs having TLVs that do not follow the above ordering rules MUST be considered as malformed by a BGP-LS Propagator. This insistence on canonical ordering ensures that multiple variant copies of the same NLRI from multiple BGP-LS Producers and the ambiguity arising therefrom is prevented.

For both the NLRI and BGP-LS Attribute parts, all TLVs are considered as optional except where explicitly specified as mandatory or required in specific conditions.

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

The origination of the same link-state information by multiple BGP-LS Producers may result in differences and inconsistencies due to the inclusion or exclusion of optional TLVs. Different optional TLVs in the NLRI results in multiple NLRIs being generated for the same link-state object. Different optional TLVs in the BGP-LS Attribute may result in the propagation of partial information. To address these inconsistencies, the BGP-LS Consumer will need to recognize and merge the duplicate information or deal with missing information. The deployment of BGP-LS Producers that consistently originate the same set of optional TLVs is recommended to mitigate such situations.

5.2. 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 describe either a node, a link, or 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.

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 (multiprotocol 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 BGP-LS but

does not support a particular Link-State NLRI type. To allow the 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 Section 5.4 (paragraph 2) of [RFC7606] for Link-State address family. An implementation MUST handle unknown 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.

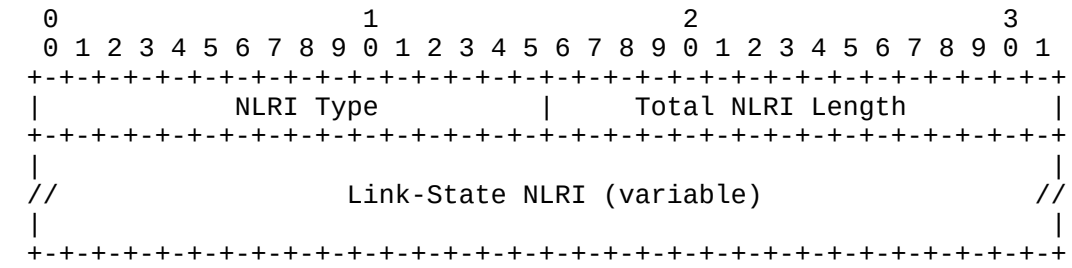


Figure 5: Link-State AFI 16388 / SAFI 71 NLRI Format

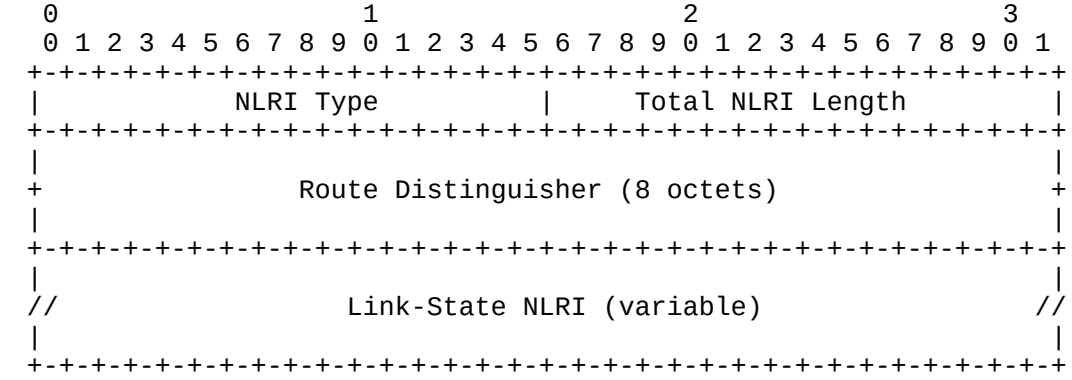


Figure 6: Link-State VPN AFI 16388 / SAFI 72 NLRI Format

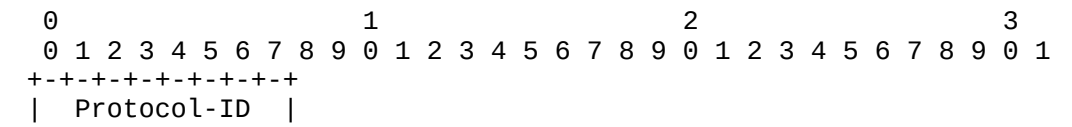
The Total NLRI Length field contains the cumulative length, in octets, of the rest of the NLRI, not including the NLRI Type field or itself. For VPN applications, it also includes the length of the Route Distinguisher.

Type	NLRI Type
1	Node NLRI
2	Link NLRI
3	IPv4 Topology Prefix NLRI
4	IPv6 Topology Prefix NLRI

Table 1: NLRI Types

Route Distinguishers are defined and discussed in [RFC4364].

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



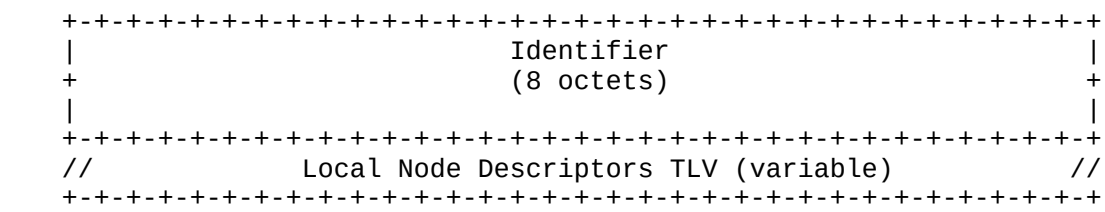


Figure 7: The Node NLRI Format

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

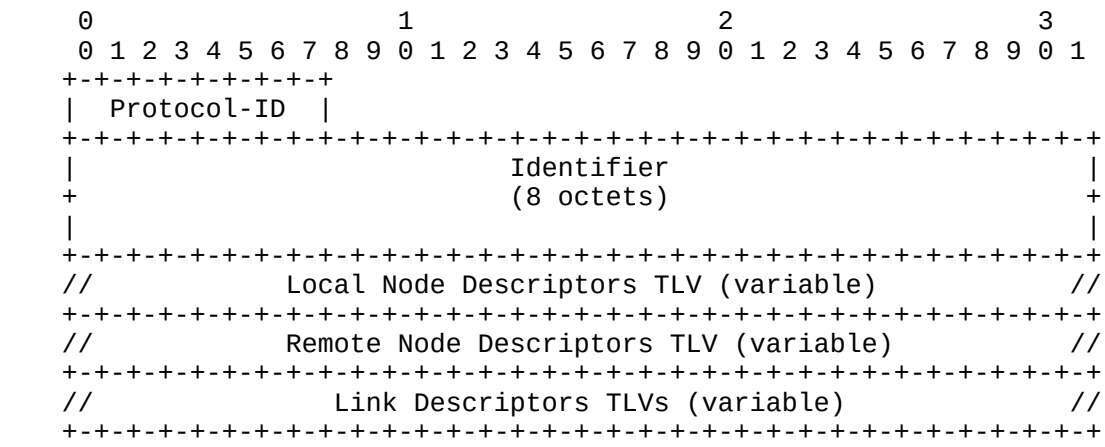


Figure 8: The Link NLRI Format

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

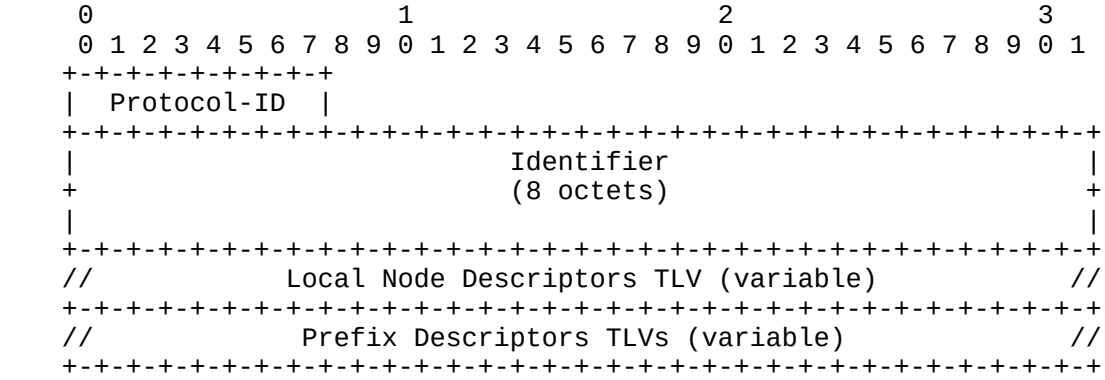


Figure 9: The IPv4/IPv6 Topology Prefix NLRI Format

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

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 6, the Protocol-ID 'Static configuration' SHOULD be used.

A router may run multiple protocol instances of OSPF or IS-IS whereby 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 Identifier field carries an 8-octet BGP-LS Instance Identifier (Instance-ID) number 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 should have the same BGP-LS Instance-ID. NLRIs with different BGP-LS Instance-IDs are considered to be from different IGP routing instances.

To support multiple IGP instances, an implementation needs to support the configuration of unique BGP-LS Instance-IDs at the routing protocol instance level. The BGP-LS Instance-ID 0 is RECOMMENDED to be used when there is only a single protocol instance in the network where BGP-LS is operational. The network operator MUST assign the same BGP-LS Instance-IDs on all BGP-LS Producers within a given IGP domain. Unique BGP-LS Instance-IDs MUST be assigned to routing protocol instances operating in different IGP domains. This can allow the BGP-LS Consumer to build an accurate segregated multi-domain topology based on the BGP-LS Instance-ID.

When the above-described semantics and recommendations are not followed, a BGP-LS Consumer may see more than one link-state object for the same node, link, or prefix (each with a different BGP-LS Instance-ID) when there are multiple BGP-LS Producers deployed. This may also result in the BGP-LS Consumers getting an inaccurate network-wide topology.

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 that are listed in Table 2. Documents extending BGP-LS specifications with new NLRI Types and/or protocols MUST specify the NLRI descriptors for them.

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 inconsistent link-state objects hanging around in the BGP-LS table.

5.2.1. 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]. When configured, these auxiliary TE Router-IDs (TLV 1028/1029) MUST be included in the node attribute described in Section 5.3.1 and MAY be included in the link attribute described in Section 5.3.2. The advertisement of the TE Router-IDs can 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 Descriptors 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 [RFC1918]. BGP-LS uses the Autonomous System Number to disambiguate the Router-IDs, as described in Section 5.2.1.1.

5.2.1.1. 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 OSPF Area-ID, Router-ID, Protocol-ID, Multi-Topology Identifier (MT-ID), and BGP-LS 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.

Furthermore, in deployments where BGP-LS is used to advertise topology from multiple ASes, the Autonomous System Number (ASN) is used to distinguish topology information reported from different ASes.

The BGP-LS Instance-ID carried in the Identifier field, as described earlier along with a set of sub-TLVs described in Section 5.2.1.4, allows specification of a flexible key for any given node/link information such that the global uniqueness of the NLRI is ensured. Since the BGP-LS Instance-ID is operator assigned, its allocation scheme can ensure that each IGP domain is uniquely identified even across a multi-AS network.

5.2.1.2. 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 5.2.1.4.

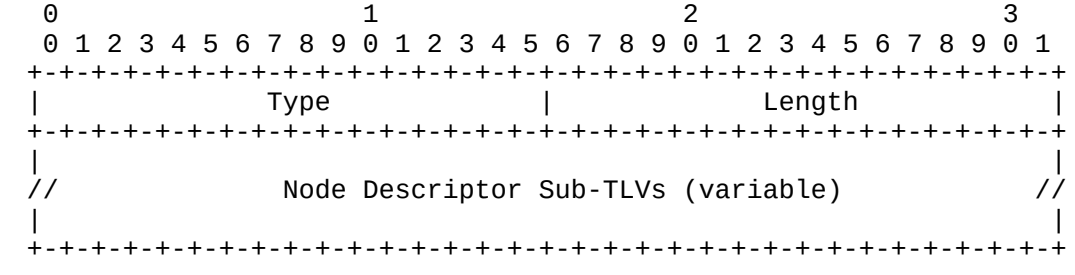


Figure 10: Local Node Descriptors TLV Format

5.2.1.3. 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 5.2.1.4.

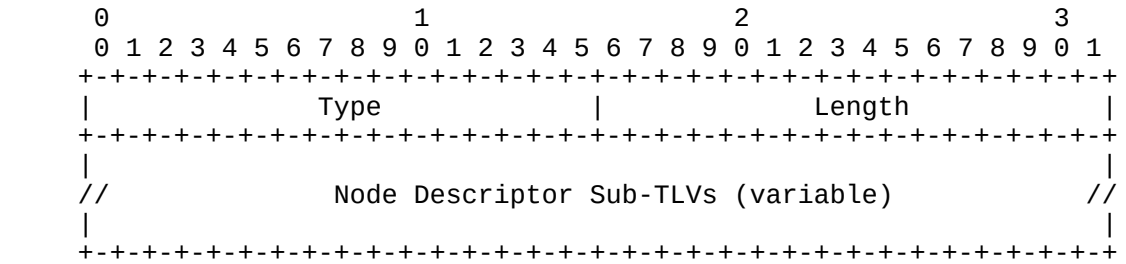


Figure 11: Remote Node Descriptors TLV Format

5.2.1.4. 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 different value, e.g., to avoid collisions when using private AS Numbers.

BGP-LS Identifier: Opaque value (32-bit ID). This is an optional TLV that has been deprecated by this document (refer to Appendix A for more details). It MAY be advertised for compatibility with [RFC7752] implementations. See the final paragraph of this section for further considerations and a recommended default value.

OSPF 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 OSPF Area Identifier allows different NLRIs of the same router to be differentiated on a per-area basis. It is not used for NLRIs when carrying information that is derived from AS-scope LSAs as that information 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 configuration'. 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. When the node is running an IGP protocol, an implementation MAY choose to use the IGP Router-ID for 'Direct' or 'Static configuration'.

At most, there MUST be 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 to compare NLRIs, even when an implementation encounters an unknown sub-TLV. Using stable sorting, an implementation can do a binary comparison of NLRIs and hence allow incremental deployment of new key sub-TLVs.

The BGP-LS Identifier was introduced by [RFC7752], and its use is being deprecated by this document. Implementations SHOULD support the advertisement of this sub-TLV for backward compatibility in deployments where there are BGP-LS Producer implementations that conform to [RFC7752] to ensure consistency of NLRI encoding for link-state objects. The default value of 0 is RECOMMENDED to be used when a BGP-LS Producer includes this sub-TLV when originating information into BGP-LS. Implementations SHOULD provide an option to configure this value for backward compatibility reasons. As a reminder, the use of the BGP-LS Instance-ID that is carried in the Identifier field is the way of segregation of link-state objects of different IGP domains in BGP-LS.

5.2.2. Link Descriptors

The Link Descriptor field is a set of Type/Length/Value (TLV) triplets. The format of each TLV is shown in Section 5.1. 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. To fully describe a single logical link, two anchor routers advertise a half-link each, i.e., two Link NLRIs are advertised for a given point-to-point link.

A link between two nodes is not considered as complete (or available) unless it is described by the two Link NLRIs corresponding to the half-link representation from the pair of anchor nodes. This check is similar to the 'two-way connectivity check' that is performed by link-state IGPs.

An implementation MAY suppress the advertisement of a Link NLRI, corresponding to a half-link, from a link-state IGP unless the IGP 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 can be leveraged before passing information for any half-link that is reported from these IGPs into BGP-LS. This ensures that only those link-state IGP adjacencies that are established get reported via Link NLRIs. Such a 'two-way connectivity check' could

also be 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, which are 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
258	Link Local/Remote Identifiers	22/4	[RFC5307], Section 1.1
259	IPv4 interface address	22/6	[RFC5305], Section 3.2
260	IPv4 neighbor address	22/8	[RFC5305], Section 3.3
261	IPv6 interface address	22/12	[RFC6119], Section 4.2
262	IPv6 neighbor address	22/13	[RFC6119], Section 4.3
263	Multi-Topology Identifier	---	Section 5.2.2.1

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 interface/neighbor 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. IPv4/IPv6 link-local addresses MUST NOT be carried in the IPv4/IPv6 interface/neighbor address TLVs (259/260/261/262) as descriptors of a link since 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 and in the case of links having only IPv6 link-local addressing on them.

The Multi-Topology Identifier TLV MUST be included as a 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 identifiers may not always be signaled in the IGPs unless their advertisement is enabled specifically. In such cases, it is valid to advertise a BGP-LS Link NLRI without any of these identifiers.

5.2.2.1. Multi-Topology Identifier

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

The semantics of the IS-IS MT-ID are defined in Sections 7.1 and 7.2 of [RFC5120]. The semantics of the OSPF MT-ID are defined in Section 3.7 of [RFC4915]. If the value in the MT-ID TLV is derived from OSPF, then the upper R bits of the MT-ID field MUST be set to 0 and only the values from 0 to 127 are valid for the MT-ID.

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

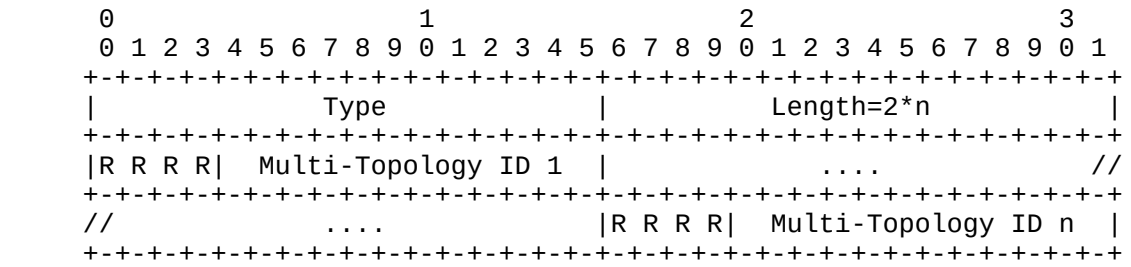


Figure 12: Multi-Topology Identifier TLV Format

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

The MT-ID TLV MAY be included as a Link Descriptor, as a Prefix Descriptor, or in the BGP-LS Attribute of a Node NLRI. When included as 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 or Prefix Descriptor, multiple NLRIs MUST be generated where each NLRI contains a single unique MT-ID. When used as a Link or Prefix Descriptor for IS-IS, the Bits R are reserved and MUST be set to 0 (as per Section 7.2 of [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 [RFC5120].

5.2.3. 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
263	Multi-Topology Identifier	variable	Section 5.2.2.1
264	OSPF Route Type	1	Section 5.2.3.1
265	IP Reachability Information	variable	Section 5.2.3.2

Table 5: Prefix Descriptor TLVs

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

5.2.3.1. OSPF Route Type

The OSPF Route Type TLV is an optional 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 OSPF Route Type TLV MUST be included in the advertisement when the type is either being signaled explicitly in the underlying LSA or can be determined via another LSA for the same prefix when it is not signaled explicitly (e.g., in the case of OSPFv2 Extended Prefix Opaque LSA [RFC7684]). The route type advertised in the OSPFv2 Extended Prefix TLV (Section 2.1 of [RFC7684]) does not make a distinction between Type 1 and 2 for AS external and Not-So-Stubby Area (NSSA) external routes. In this case, the route type to be used in the BGP-LS advertisement can be determined by checking the OSPFv2 External or NSSA External LSA for the prefix. A similar check for the base OSPFv2 LSAs can be done to determine the route type to be used when the route type value 0 is carried in the OSPFv2 Extended Prefix TLV.

The format of the OSPF Route Type TLV is shown in the following figure.

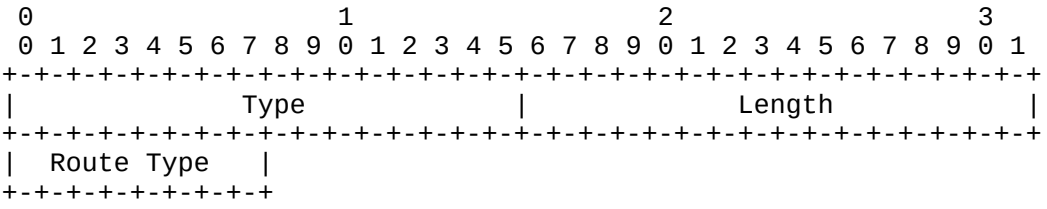


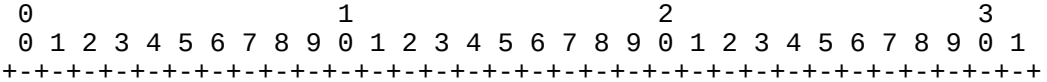
Figure 13: OSPF Route Type TLV Format

The Type and Length fields of the TLV are defined in Table 5. The Route Type field follows the route types 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)

5.2.3.2. 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. 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:



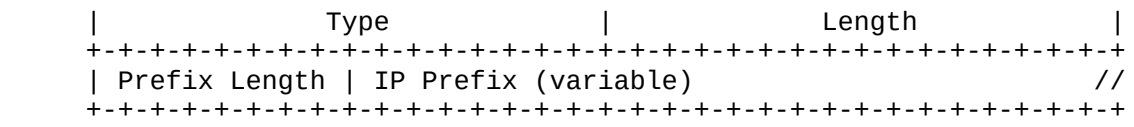


Figure 14: IP Reachability Information TLV Format

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 an IP address prefix followed by the minimum number of trailing bits needed to make the end of the field fall on an octet boundary. Any trailing bits MUST be set to 0. Thus, 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 up to 16, 3 octets for prefix length 17 up to 24, 4 octets for prefix length 25 up to 32, etc.

5.3. The BGP-LS Attribute

The BGP-LS Attribute (assigned value 29 by IANA) 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, as described in the following section. This attribute SHOULD only be included with Link-State NLRIs. The use of this attribute for other address families is outside the scope of this document.

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 size of the BGP-LS Attribute may potentially grow large, 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 implementations support the extended message size for BGP [RFC8654] to accommodate a larger size of information within the BGP-LS Attribute. BGP-LS Producers MUST ensure that the TLVs included in the BGP-LS Attribute does not result in a BGP UPDATE message for a single Link-State NLRI that crosses the maximum limit for a BGP message.

An implementation MAY adopt mechanisms to avoid this problem that may be based on the BGP-LS Consumer applications' requirement; these mechanisms are beyond the scope of this specification. However, if an implementation chooses to mitigate the problem by excluding some TLVs from the BGP-LS Attribute, this exclusion SHOULD be done consistently by all BGP-LS Producers within a given BGP-LS domain. In the event of inconsistent exclusion of TLVs from the BGP-LS Attribute, the result would be a differing set of attributes of the link-state object being propagated to BGP-LS Consumers based on the BGP Decision Process at BGP-LS Propagators.

When a BGP-LS Propagator finds that it is exceeding the maximum BGP message size due to the addition or update of some other BGP Attribute (e.g., AS_PATH), it MUST consider the BGP-LS Attribute to be malformed, apply the 'Attribute Discard' error-handling approach [RFC7606], and handle the propagation as described in Section 8.2.2. When a BGP-LS Propagator needs to perform 'Attribute Discard' for reducing the BGP UPDATE message size as specified in Section 4 of [RFC8654], it MUST first discard the BGP-LS Attribute to enable the detection and diagnosis of this error condition as discussed in Section 8.2.2. This brings the deployment consideration that the consistent propagation of BGP-LS information with a BGP UPDATE message size larger than 4096 octets can only happen along a set of

BGP Speakers that all support the contents of [RFC8654].

5.3.1. 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
263	Multi-Topology Identifier	variable	Section 5.2.2.1
1024	Node Flag Bits	1	Section 5.3.1.1
1025	Opaque Node Attribute	variable	Section 5.3.1.5
1026	Node Name	variable	Section 5.3.1.3
1027	IS-IS Area Identifier	variable	Section 5.3.1.2
1028	IPv4 Router-ID of Local Node	4	[RFC5305], Section 4.3
1029	IPv6 Router-ID of Local Node	16	[RFC6119], Section 4.1

Table 6: Node Attribute TLVs

5.3.1.1. Node Flag Bits TLV

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

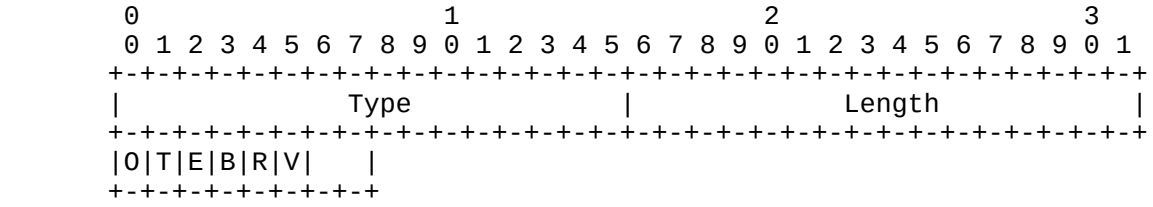


Figure 15: Node Flag Bits TLV Format

The bits are defined as follows:

Bit	Description	Reference
'O'	Overload Bit	[ISO10589]
'A'	Attached Bit	[ISO10589]
'E'	External Bit	[RFC2328]
'B'	ABR Bit	[RFC2328]
'R'	Router Bit	[RFC5340]
'V'	V6 Bit	[RFC5340]

Table 7: Node Flag Bits Definitions

The bits that are not defined MUST be set to 0 by the originator and MUST be ignored by the receiver.

5.3.1.2. IS-IS Area Identifier TLV

An IS-IS node can be part of only a single IS-IS area. However, a node can have multiple synonymous area addresses. 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.

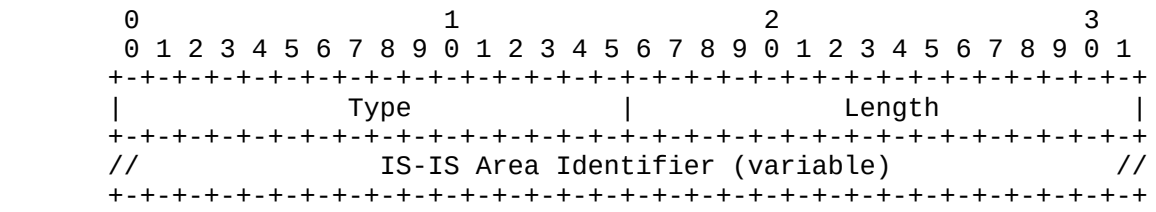


Figure 16: IS-IS Area Identifier TLV Format

5.3.1.3. Node Name TLV

The Node Name TLV is optional. The encoding semantics for the node name 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, a substring of the FQDN (e.g., a hostname), or any string that an operator wants to use for the router. The use of the FQDN or a substring 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, then the 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 the advertisement of the node name, which may be encoded in the Node Name TLV.

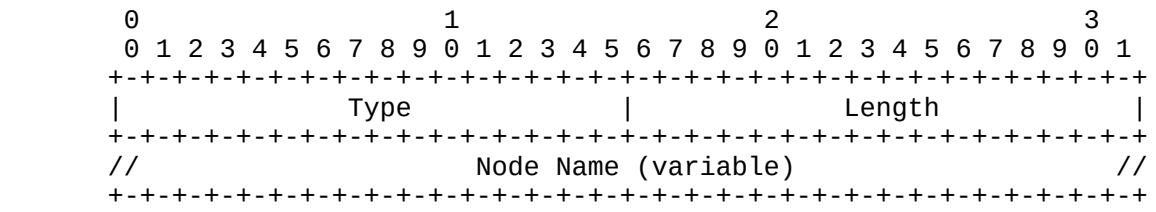


Figure 17: Node Name Format

5.3.1.4. 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 as a separate TLV.

5.3.1.5. 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 a new IGP link-state attribute and its protocol-neutral BGP-LS extension being defined. Once the protocol-neutral BGP-LS extensions are defined, the BGP-LS implementations may still need to advertise the information both within the Opaque Attribute TLV and the new TLV definition for incremental deployment and transition.

In the case of OSPF, this TLV MUST NOT be used to advertise TLVs other than those in the OSPF Router Information (RI) LSA [RFC7770].

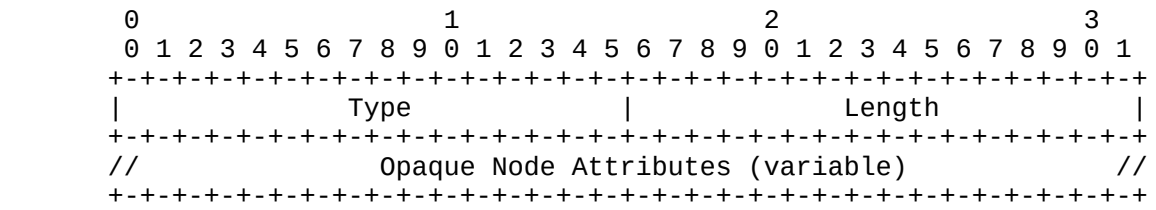


Figure 18: Opaque Node Attribute Format

The Type is as specified in Table 6. The length is variable.

5.3.2. 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 5.1. 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, which are 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
1028	IPv4 Router-ID of Local Node	134/---	[RFC5305], Section 4.3
1029	IPv6 Router-ID of Local Node	140/---	[RFC6119], Section 4.1
1030	IPv4 Router-ID of Remote Node	134/---	[RFC5305], Section 4.3
1031	IPv6 Router-ID of Remote Node	140/---	[RFC6119], Section 4.1
1088	Administrative group (color)	22/3	[RFC5305], Section 3.1
1089	Maximum link bandwidth	22/9	[RFC5305], Section 3.4

1090	Max. reservable link bandwidth	22/10	[RFC5305], Section 3.5
1091	Unreserved bandwidth	22/11	[RFC5305], Section 3.6
1092	TE Default Metric	22/18	Section 5.3.2.3
1093	Link Protection Type	22/20	[RFC5307], Section 1.2
1094	MPLS Protocol Mask	---	Section 5.3.2.2
1095	IGP Metric	---	Section 5.3.2.4
1096	Shared Risk Link Group	---	Section 5.3.2.5
1097	Opaque Link Attribute	---	Section 5.3.2.6
1098	Link Name	---	Section 5.3.2.7

Table 8: Link Attribute TLVs

5.3.2.1. 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.

5.3.2.2. MPLS Protocol Mask TLV

The MPLS Protocol Mask TLV carries a bitmask 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.

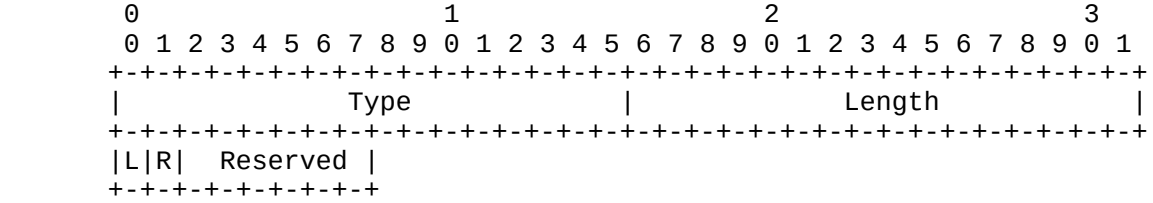


Figure 19: MPLS Protocol Mask TLV

The following bits are defined, and the reserved bits MUST be set to zero and SHOULD be ignored on receipt:

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

Table 9: MPLS Protocol Mask TLV Codes

The bits that are not defined MUST be set to 0 by the originator and MUST be ignored by the receiver.

5.3.2.3. 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 fewer than 32 bits, then the high-order bits of this field MUST be padded with zero.

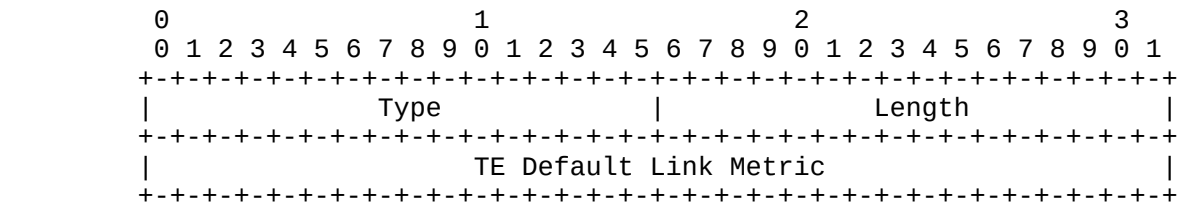


Figure 20: TE Default Metric TLV Format

5.3.2.4. 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 are 6 bits in size but are encoded in a 1-octet field; therefore, the two most significant bits of the field MUST be set to 0 by the originator and MUST be ignored by the receiver. OSPF link metrics have a length of 2 octets. IS-IS wide metrics have a length of 3 octets.

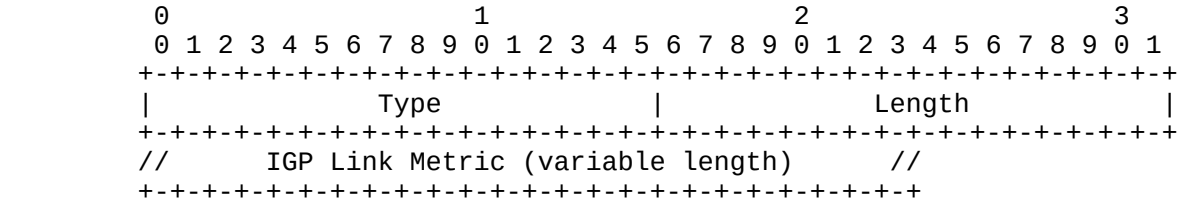
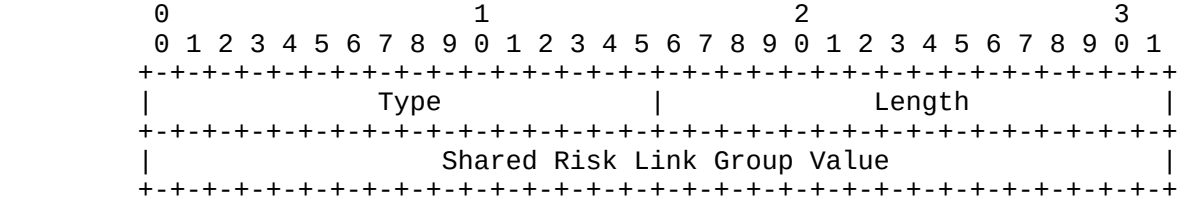


Figure 21: IGP Metric TLV Format

5.3.2.5. 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).



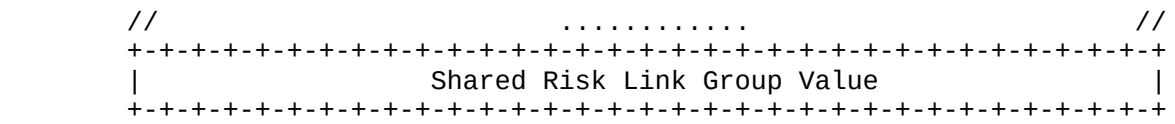


Figure 22: Shared Risk Link Group TLV Format

The SRLG TLV for OSPF-TE is defined in [RFC4203]. In IS-IS, the SRLG information is carried in two different TLVs: the GMPLS-SRLG TLV (for IPv4) (Type 138) defined in [RFC5307] and the IPv6 SRLG TLV (Type 139) defined in [RFC6119]. Both IPv4 and IPv6 SRLG information is carried in a single TLV.

5.3.2.6. 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 a new IGP link-state attribute and its 'protocol-neutral' BGP-LS extension being defined. Once the protocol-neutral BGP-LS extensions are defined, the BGP-LS implementations may still need to advertise the information both within the Opaque Attribute TLV and the new TLV definition for incremental deployment and transition.

In the case of OSPFv2, this TLV MUST NOT be used to advertise information carried using TLVs other than those in the OSPFv2 Extended Link Opaque LSA [RFC7684]. In the case of OSPFv3, this TLV MUST NOT be used to advertise TLVs other than those in the OSPFv3 E-Router-LSA or E-Link-LSA [RFC8362].

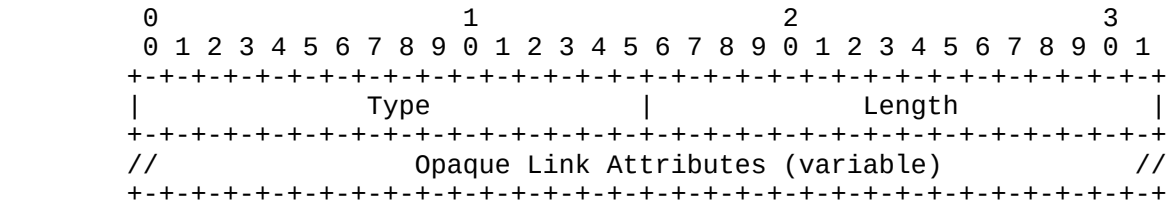


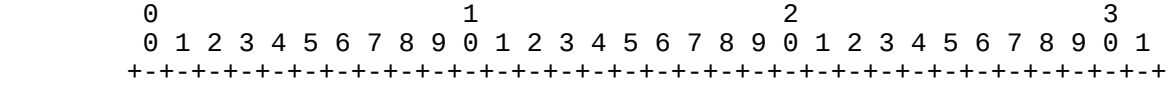
Figure 23: Opaque Link Attribute TLV Format

5.3.2.7. 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, a substring of the FQDN, or any string that an operator wants to use for the link. The use of the FQDN or a substring 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, then the 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.



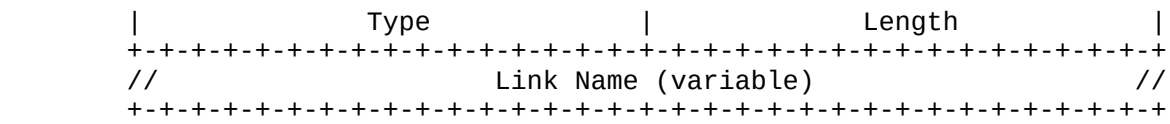


Figure 24: Link Name TLV Format

5.3.3. 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 5.3.3.1
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 5.3.3.6

Table 10: Prefix Attribute TLVs

5.3.3.1. IGP Flags TLV

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

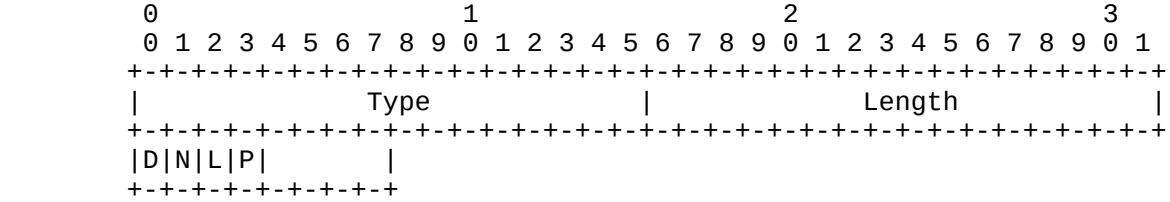


Figure 25: IGP Flag TLV Format

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]

Table 11: IGP Flag Bits Definitions

The bits that are not defined MUST be set to 0 by the originator and MUST be ignored by the receiver.

5.3.3.2. 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:

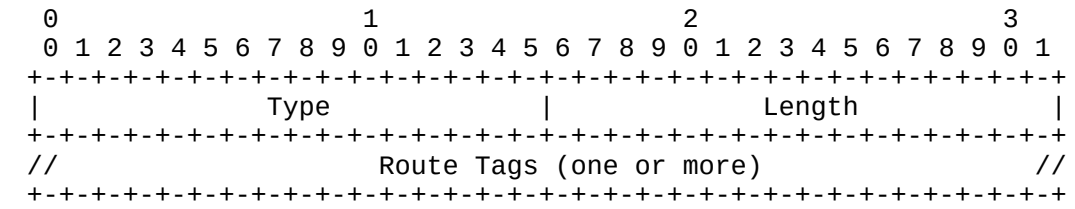


Figure 26: IGP Route Tag TLV Format

The length is a multiple of 4.

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

5.3.3.3. IGP Extended Route Tag TLV

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

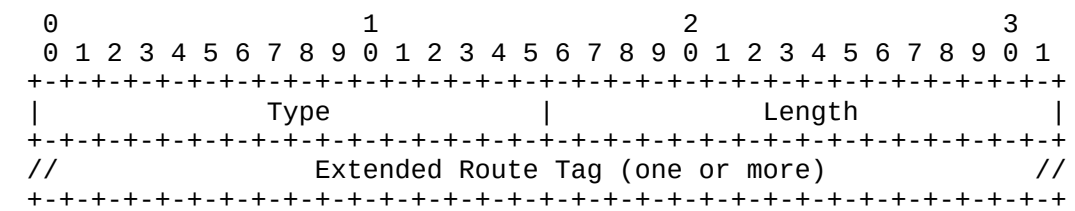


Figure 27: IGP Extended Route Tag TLV Format

The length is a multiple of 8.

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

5.3.3.4. 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.

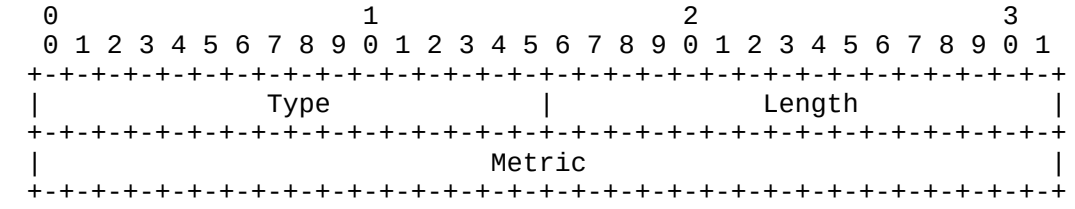


Figure 28: Prefix Metric TLV Format

The length is 4.

5.3.3.5. OSPF Forwarding Address TLV

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

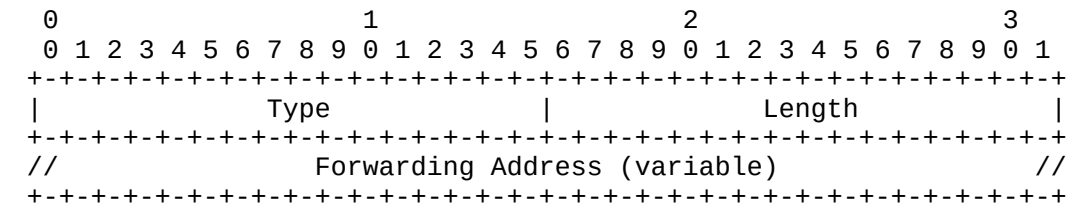


Figure 29: OSPF Forwarding Address TLV Format

The length is 4 for an IPv4 forwarding address and 16 for an IPv6 forwarding address.

5.3.3.6. 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 it shall use new protocol extensions for 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 a new IGP link-state attribute and its protocol-neutral BGP-LS extension being defined. Once the protocol-neutral BGP-LS extensions are defined, the BGP-LS implementations may still need to advertise the information both within the Opaque Attribute TLV and the new TLV definition for incremental deployment and transition.

In the case of OSPFv2, this TLV MUST NOT be used to advertise information carried using TLVs other than those in the OSPFv2 Extended Prefix Opaque LSA [RFC7684]. In the case of OSPFv3, this TLV MUST NOT be used to advertise TLVs other than those in the OSPFv3 E-Inter-Area-Prefix-LSA, E-Intra-Area-Prefix-LSA, E-AS-External-LSA, and E-NSSA-LSA [RFC8362].

The format of the TLV is as follows:

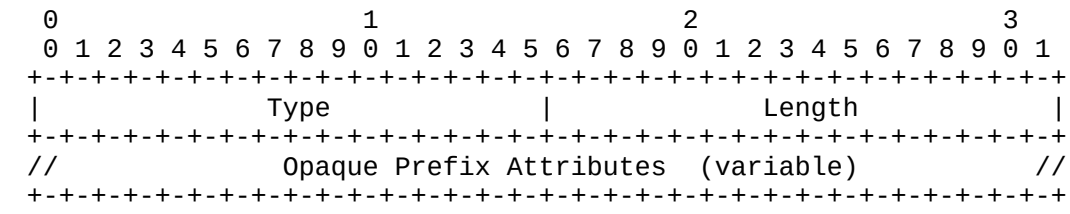


Figure 30: Opaque Prefix Attribute TLV Format

The Type is as specified in Table 10. The length is variable.

5.4. Private Use

TLVs for Vendor Private Use are supported using the code point range reserved as indicated in Section 7. For such TLV use in the NLRI or BGP-LS Attribute, the format described in Section 5.1 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 5.2 to carry the Enterprise Code [ENTNUM]. This enables the use of vendor-specific extensions without conflicts.

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

5.5. 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 an IPv6 link-local address. The IPv6 link-local 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 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 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.

5.6. 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] [RFC9346]. In other cases, an implementation SHOULD provide a means to inject inter-AS links into BGP-LS. The exact mechanism used to advertise the inter-AS links is outside the scope of this document.

5.7. OSPF Virtual Links and Sham Links

In an OSPF [RFC2328] [RFC5340] network, OSPF virtual links serve to connect physically separate components of the backbone to establish/maintain continuity of the backbone area. While OSPF virtual links are modeled as point-to-point, unnumbered links in the OSPF topology, their characteristics and purpose are different from other types of links in the OSPF topology. They are advertised using a distinct "virtual link" type in OSPF LSAs. The mechanism for the advertisement of OSPF virtual links via BGP-LS is outside the scope of this document.

In an OSPF network, sham links [RFC4577] [RFC6565] are used to provide intra-area connectivity between VPN Routing and Forwarding (VRF) instances on Provider Edge (PE) routers over the VPN provider’s network. These links are advertised in OSPF as point-to-point, unnumbered links and represent connectivity over a service provider network using encapsulation mechanisms like MPLS. As such, the mechanism for the advertisement of OSPF sham links follows the same procedures as other point-to-point, unnumbered links as described previously in this document.

5.8. OSPFv2 Type 4 Summary-LSA & OSPFv3 Inter-Area-Router-LSA

OSPFv2 [RFC2328] defines the type 4 summary-LSA and OSPFv3 [RFC5340]

defines the inter-area-router-LSA for an Area Border Router (ABR) to advertise reachability to an AS Border Router (ASBR) that is external to the area yet internal to the AS. The nature of information advertised by OSPF using this type of LSA does not map to either a node, a link, or a prefix as discussed in this document. Therefore, the mechanism for the advertisement of the information carried by these LSAs is outside the scope of this document.

5.9. Handling of Unreachable IGP Nodes

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 are in area 1 as indicated by the a0 and a1 references respectively against the links.

A BGP-LS Consumer talks to BGP route reflector RR0, which is a BGP-LS Propagator that is aggregating the BGP-LS feed from BGP-LS Producers R2 and R3. Here, R2 and R3 provide a redundant topology feed via BGP-LS to RR0. Normally, RR0 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 Decision Process.

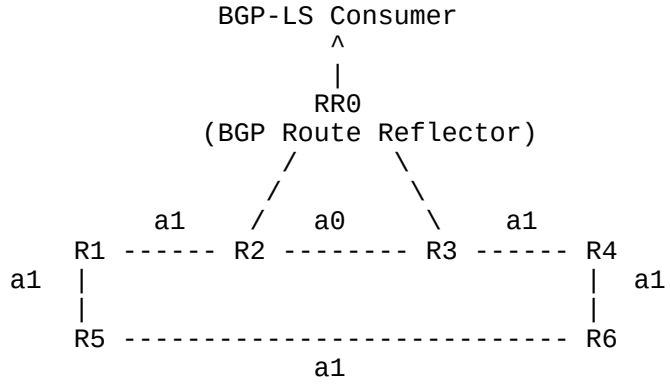


Figure 31: Incorrect Reporting Due to BGP Path Selection

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

Now, R5 will remove the link R5-R6 from its Router LSA, and this updated LSA is available at R2. R2 also has a stale copy of R6's Router LSA that still has the link R6-R5 in it. Based on this view in its LSDB, R2 will advertise only the half-link R6-R5 that it derives from R6's stale Router LSA.

At the same time, R6 has removed the link R6-R5 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 R5-R6 in it. Based on its LSDB, R3 will advertise only the half-link R5-R6 that it derives 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 RR0. When viewed together, it would not detect or realize that area 1 is partitioned. Also, if R2 continues to report Node and Prefix NLRIs corresponding to the stale copy of R4's and R6's Router LSAs, then RR0 could prefer them over the valid Node and Prefix NLRIs for R4 and R6 that it is receiving from R3 depending on RR0's BGP Decision Process. This would result in the BGP-LS Consumer getting stale and inaccurate topology information. This problem 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 SHOULD withdraw all link-state objects advertised by it in BGP when the node that originated its corresponding LSPs/LSAs is determined to have become unreachable in the IGP. An implementation MAY continue to advertise link-state objects corresponding to unreachable nodes in a deployment use case where the BGP-LS Consumer is interested in receiving a topology feed corresponding to a complete IGP LSDB view. In such deployments, it is expected that the problem described above is mitigated by the BGP-LS Consumer via appropriate handling of such a topology feed in addition to the use of either a direct BGP peering with the BGP-LS Producer nodes or mechanisms such as those described in [RFC7911] when using RRs. Details of these mechanisms are outside the scope of this document.

If the BGP-LS Producer does withdraw link-state objects associated with an IGP node based on the failure of reachability check for that node, then it MUST re-advertise those link-state objects after that node becomes reachable again in the IGP domain.

5.10. Router-ID Anchoring Example: ISO Pseudonode

The 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 an 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.

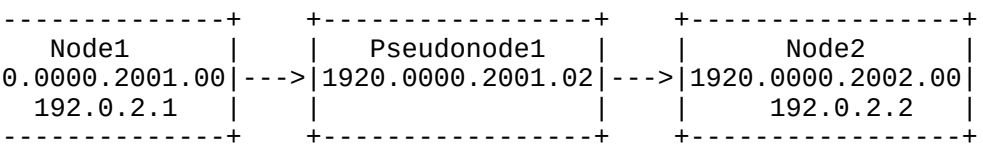


Figure 32: IS-IS Pseudonodes

5.11. Router-ID Anchoring Example: OSPF Pseudonode

The 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 198.51.100.1 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: 192.0.2.1

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

* Remote Node Descriptor

  TLV #515: IGP Router-ID: 192.0.2.1:198.51.100.1

  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: 192.0.2.1:198.51.100.1

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

* Remote Node Descriptor

  TLV #515: IGP Router-ID: 192.0.2.2

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

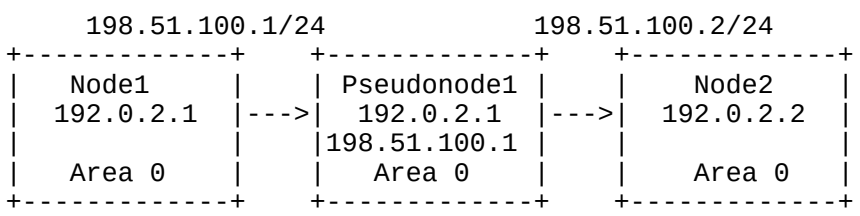


Figure 33: OSPF Pseudonodes

The LAN subnet 198.51.100.0/24 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.

5.12. 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 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, which initially were OSPFv2-only routers and then had 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 protocols.

6. Link to Path Aggregation

Distribution of all links available on the global Internet is certainly possible; however, it is 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 of 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.

6.1. 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 - Fast Reroute (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.

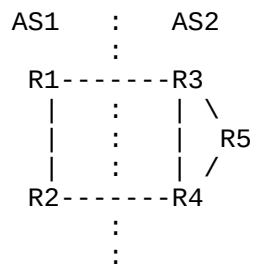


Figure 34: No Link Aggregation

6.2. 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.

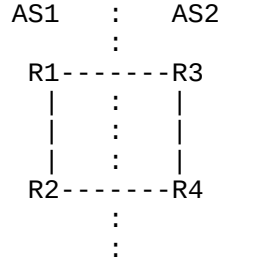


Figure 35: ASBR Link Aggregation

6.3. 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.

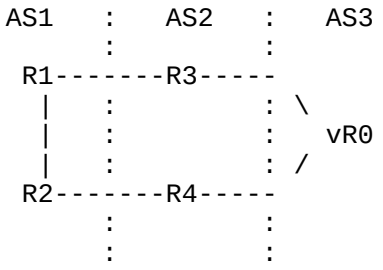


Figure 36: Multi-AS Aggregation

7. IANA Considerations

As this document obsoletes [RFC7752] and [RFC9029], IANA has updated all registration information that references those documents to instead reference this document.

IANA has assigned address family number 16388 (BGP-LS) in the "Address Family Numbers" registry.

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

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

IANA has created a "Border Gateway Protocol - Link-State (BGP-LS) Parameters" registry group at <https://www.iana.org/assignments/bgp-ls-parameters>.

This section also incorporates all the changes to the allocation procedures for the BGP-LS IANA registry group as well as the guidelines for designated experts introduced by [RFC9029].

7.1. BGP-LS Registries

All of the registries listed in the following subsections are specific to BGP-LS and are accessible under this registry.

7.1.1. BGP-LS NLRI Types Registry

The "BGP-LS NLRI Types" registry has been set up for assignment for the two-octet-sized code points for BGP-LS NLRI types and populated with the values shown below:

Type	NLRI Type	Reference
0	Reserved	RFC 9552
1	Node NLRI	RFC 9552
2	Link NLRI	RFC 9552
3	IPv4 Topology Prefix NLRI	RFC 9552
4	IPv6 Topology Prefix NLRI	RFC 9552

| 65000-65535 | Private Use | RFC 9552 |

Table 12: BGP-LS NLRI Types

A range is reserved for Private Use [RFC8126]. All other allocations within the registry are to be made using the "Expert Review" policy [RFC8126], which requires documentation of the proposed use of the allocated value and approval by the designated expert assigned by the IESG.

7.1.2. BGP-LS Protocol-IDs Registry

The "BGP-LS Protocol-IDs" registry has been set up for assignment for the one-octet-sized code points for BGP-LS Protocol-IDs and populated with the values shown below:

Protocol-ID	NLRI information source protocol	Reference
0	Reserved	RFC 9552
1	IS-IS Level 1	RFC 9552
2	IS-IS Level 2	RFC 9552
3	OSPFv2	RFC 9552
4	Direct	RFC 9552
5	Static configuration	RFC 9552
6	OSPFv3	RFC 9552
200-255	Private Use	RFC 9552

Table 13: BGP-LS Protocol-IDs

A range is reserved for Private Use [RFC8126]. All other allocations within the registry are to be made using the "Expert Review" policy [RFC8126], which requires documentation of the proposed use of the allocated value and approval by the designated expert assigned by the IESG.

7.1.3. BGP-LS Well-Known Instance-IDs Registry

The "BGP-LS Well-Known Instance-IDs" registry that was set up via [RFC7752] is no longer required. IANA has marked this registry obsolete and changed its registration procedure to "registry closed".

7.1.4. BGP-LS Node Flags Registry

The "BGP-LS Node Flags" registry has been created for the one-octet-sized flags field of the Node Flag Bits TLV (1024) and populated with the initial values shown below:

Bit	Description	Reference
0	Overload Bit (O-bit)	RFC 9552
1	Attached Bit (A-bit)	RFC 9552
2	External Bit (E-bit)	RFC 9552

3	ABR Bit (B-bit)	RFC 9552
4	Router Bit (R-bit)	RFC 9552
5	V6 Bit (V-bit)	RFC 9552
6-7	Unassigned	

Table 14: BGP-LS Node Flags

Allocations within the registry are to be made using the "Expert Review" policy [RFC8126], which requires documentation of the proposed use of the allocated value and approval by the designated expert assigned by the IESG.

7.1.5. BGP-LS MPLS Protocol Mask Registry

The "BGP-LS MPLS Protocol Mask" registry has been created for the one-octet-sized flags field of the MPLS Protocol Mask TLV (1094) and populated with the initial values shown below:

Bit	Description	Reference
0	Label Distribution Protocol (L-bit)	RFC 9552
1	Extension to RSVP for LSP Tunnels (R-bit)	RFC 9552
2-7	Unassigned	

Table 15: BGP-LS MPLS Protocol Mask

Allocations within the registry are to be made using the "Expert Review" policy [RFC8126], which requires documentation of the proposed use of the allocated value and approval by the designated expert assigned by the IESG.

7.1.6. BGP-LS IGP Prefix Flags Registry

The "BGP-LS IGP Prefix Flags" registry has been created for the one-octet-sized flags field of the IGP Flags TLV (1152) and populated with the initial values shown below:

Bit	Description	Reference
0	IS-IS Up/Down Bit (D-bit)	RFC 9552
1	OSPF "no unicast" Bit (N-bit)	RFC 9552
2	OSPF "local address" Bit (L-bit)	RFC 9552
3	OSPF "propagate NSSA" Bit (P-bit)	RFC 9552
4-7	Unassigned	

Table 16: BGP-LS IGP Prefix Flags

Allocations within the registry are to be made using the "Expert Review" policy [RFC8126], which requires documentation of the proposed use of the allocated value and approval by the designated expert assigned by the IESG.

7.1.7. BGP-LS TLVs Registry

The "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs" registry was created via [RFC7752]. Per this document, IANA has renamed that registry to "BGP-LS NLRI and Attribute TLVs" and removed the column for "IS-IS TLV/Sub-TLV". The registration procedures are as follows:

TLV Code Point	Registration Process
0-255	Reserved (not to be allocated)
256-64999	Expert Review
65000-65535	Private Use

Table 17: BGP-LS TLVs Registration Process

A range is reserved for Private Use [RFC8126]. All other allocations except for the reserved range within the registry are to be made using the "Expert Review" policy [RFC8126], which requires documentation of the proposed use of the allocated value and approval by the designated expert assigned by the IESG.

The registry was pre-populated with the values shown in Table 18, and the reference for each allocation has been changed to this document and the respective section where those TLVs are specified.

7.2. Guidance for Designated Experts

In all cases of review by the designated expert described here, the designated expert is expected to check the clarity of purpose and use of the requested code points. The following points apply to the registries discussed in this document:

1. Application for a code point allocation may be made to the designated experts at any time and MUST be accompanied by technical documentation explaining the use of the code point. Such documentation SHOULD be presented in the form of an Internet-Draft but MAY arrive in any form that can be reviewed and exchanged among reviewers.

2. The designated experts SHOULD only consider requests that arise from Internet-Drafts that have already been accepted as working group documents or that are planned for progression as AD-Sponsored documents in the absence of a suitably chartered working group.

3. In the case of working group documents, the designated experts MUST check with the working group chairs that there is a consensus within the working group to allocate at this time. In the case of AD-Sponsored documents, the designated experts MUST check with the AD for approval to allocate at this time.

4. If the document is not adopted by the IDR Working Group (or its successor), the designated expert MUST notify the IDR mailing list (or its successor) of the request and MUST provide access to the document. The designated expert MUST allow two weeks for any response. Any comments received MUST be considered by the designated expert as part of the subsequent step.

5. The designated experts MUST then review the assignment requests on their technical merit. The designated experts MAY raise issues related to the allocation request with the authors and on

the IDR (or successor) mailing list for further consideration before the assignments are made.

6. The designated expert MUST ensure that any request for a code point does not conflict with work that is active or already published within the IETF.

7. Once the designated experts have approved, IANA will update the registry by marking the allocated code points with a reference to the associated document.

8. In the event that the document is a working group document or is AD-Sponsored and fails to progress to publication as an RFC, the working group chairs or AD SHOULD contact IANA to coordinate about marking the code points as deprecated. A deprecated code point is not marked as allocated for use and is not available for allocation in a future document. The WG chairs may inform IANA that a deprecated code point can be completely deallocated (i.e., made available for new allocations) at any time after it has been deprecated if there is a shortage of unallocated code points in the registry.

8. Manageability Considerations

This section is structured as recommended in [RFC5706].

8.1. Operational Considerations

8.1.1. 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. Distribution of the BGP-LS NLRIs SHOULD be handled by dedicated route reflectors in most deployments providing a level of isolation and fault containment between different BGP address families. 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.

It is RECOMMENDED that operators deploying BGP-LS enable two or more BGP-LS Producers in each IGP flooding domain to achieve redundancy in the origination of link-state information into BGP-LS. It is also RECOMMENDED that operators ensure BGP peering designs that ensure redundancy in the BGP update propagation paths (e.g., using at least a pair of route reflectors) and ensure that BGP-LS Consumers are receiving the topology information from at least two BGP-LS Speakers.

In a multi-domain IGP network, the correct provisioning of the BGP-LS Instance-IDs on the BGP-LS Producers is required for consistent reporting of the multi-domain link-state topology. Refer to Section 5.2 for more details.

8.1.2. Installation and Initial Setup

Configuration parameters defined in Section 8.2.3 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.

8.1.3. 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 8.2.3), so the extension can be gradually rolled out in the network.

8.1.4. Requirements for Other Protocols and Functional Components

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

8.1.5. Impact on Network Operation

The frequency of Link-State NLRI updates could interfere with regular BGP prefix distribution. A network operator should use a dedicated route reflector infrastructure to distribute Link-State NLRIs as discussed in Section 8.1.1.

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.

8.1.6. 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.

8.2. Management Considerations

8.2.1. 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.

8.2.2. Fault Management

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

A Link-State NLRI MUST NOT be considered malformed or invalid based on the inclusion/exclusion of TLVs or contents of the TLV fields (i.e., semantic errors), as described in Sections 5.1 and 5.2.

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

* The sum of all TLV lengths found in the BGP MP_REACH_NLRI attribute corresponds to the BGP MP_REACH_NLRI length.

* The sum of all TLV lengths found in the BGP MP_UNREACH_NLRI attribute corresponds to the BGP MP_UNREACH_NLRI length.

* The sum of all TLV lengths found in a Link-State NLRI corresponds to the Total NLRI Length field of all its descriptors.

* The length of the TLVs and, when the TLV is recognized then, the length of its sub-TLVs in the NLRI are valid.

* The syntactic correctness of the NLRI fields has been verified as per [RFC7606].

* The rule regarding the ordering of TLVs has been followed as described in Section 5.1.

* For NLRIs carrying either a Local or Remote Node Descriptor TLV, there is not more than one instance of a sub-TLV present.

When the error that is determined allows for the router to skip the malformed NLRI(s) and continue the processing of the rest of the BGP UPDATE message (e.g., when the TLV ordering rule is violated), then it MUST handle such malformed NLRIs as 'NLRI discard' (i.e., processing similar to what is described in Section 5.4 of [RFC7606]). 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 a 'session reset' when the session is only being used for BGP-LS or if 'AFI/SAFI disable' action is not possible.

A BGP-LS Attribute MUST NOT be considered malformed or invalid based on the inclusion/exclusion of TLVs or contents of the TLV fields (i.e., semantic errors), as described in Sections 5.1 and 5.3.

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

* The sum of all TLV lengths found in the BGP-LS Attribute corresponds to the BGP-LS Attribute length.

* The syntactic correctness of the Attributes (including the BGP-LS Attribute) have been verified as per [RFC7606].

* The length of each TLV and, when the TLV is recognized then, the length of its sub-TLVs in the BGP-LS Attribute are valid.

When the error that is determined allows for the router to skip the malformed BGP-LS Attribute and continue the processing of the rest of the BGP 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, 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, unless denied by local policy, 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/withdrawal. This also makes it possible for a network operator to trace back to the BGP-LS Propagator that detected

the fault with the BGP-LS Attribute.

An implementation SHOULD log a message for any errors found during syntax validation for further analysis.

A BGP-LS Propagator, even when it has a coexisting BGP-LS Consumer on the same node, 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):

* presence of a mandatory TLV

* the length of a fixed-length TLV is correct or the length of a variable length TLV is valid or permissible

* the values of TLV fields are valid or permissible

* the inclusion and use of TLVs/sub-TLVs with specific Link-State NLRI types is valid

Each TLV may indicate the valid and permissible values and their semantics that can 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.

8.2.3. 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 MUST allow the operator to configure an 8-octet BGP-LS Instance-ID. Refer to Section 5.2 for guidance to the operator for the configuration of BGP-LS Instance-ID.

An implementation SHOULD allow the operator to configure Autonomous System Number (ASN) and BGP-LS identifiers (refer to Section 5.2.1.4).

An implementation SHOULD allow the operator to configure a 4096-byte size limit for a BGP-LS UPDATE message on a BGP-LS Producer or allow larger values when they know that all BGP-LS Speakers support the extended message size [RFC8654].

8.2.4. Accounting Management

Not Applicable.

8.2.5. 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 the system or session start time. An implementation MAY also enhance this information by recording peak per-second counts in each case.

8.2.6. Security Management

An operator MUST 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.

9. 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	Reference Section
256	Local Node Descriptors	Section 5.2.1.2
257	Remote Node Descriptors	Section 5.2.1.3
258	Link Local/Remote Identifiers	Section 5.2.2
259	IPv4 interface address	Section 5.2.2
260	IPv4 neighbor address	Section 5.2.2
261	IPv6 interface address	Section 5.2.2
262	IPv6 neighbor address	Section 5.2.2
263	Multi-Topology Identifier	Section 5.2.2.1
264	OSPF Route Type	Section 5.2.3.1
265	IP Reachability Information	Section 5.2.3.2
512	Autonomous System	Section 5.2.1.4
513	BGP-LS Identifier (deprecated)	Section 5.2.1.4
514	OSPF Area-ID	Section 5.2.1.4
515	IGP Router-ID	Section 5.2.1.4
1024	Node Flag Bits	Section 5.3.1.1
1025	Opaque Node Attribute	Section 5.3.1.5

	1026	Node Name	Section 5.3.1.3
	1027	IS-IS Area Identifier	Section 5.3.1.2
	1028	IPv4 Router-ID of Local Node	Sections 5.3.1.4 and 5.3.2.1
	1029	IPv6 Router-ID of Local Node	Sections 5.3.1.4 and 5.3.2.1
	1030	IPv4 Router-ID of Remote Node	Section 5.3.2.1
	1031	IPv6 Router-ID of Remote Node	Section 5.3.2.1
	1088	Administrative group (color)	Section 5.3.2
	1089	Maximum link bandwidth	Section 5.3.2
	1090	Max. reservable link bandwidth	Section 5.3.2
	1091	Unreserved bandwidth	Section 5.3.2
	1092	TE Default Metric	Section 5.3.2.3
	1093	Link Protection Type	Section 5.3.2
	1094	MPLS Protocol Mask	Section 5.3.2.2
	1095	IGP Metric	Section 5.3.2.4
	1096	Shared Risk Link Group	Section 5.3.2.5
	1097	Opaque Link Attribute	Section 5.3.2.6
	1098	Link Name	Section 5.3.2.7
	1152	IGP Flags	Section 5.3.3.1
	1153	IGP Route Tag	Section 5.3.3.2
	1154	IGP Extended Route Tag	Section 5.3.3.3
	1155	Prefix Metric	Section 5.3.3.4
	1156	OSPF Forwarding Address	Section 5.3.3.5
	1157	Opaque Prefix Attribute	Section 5.3.3.6

Table 18: Summary Table of TLV/Sub-TLV Code Points

10. 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.

The operator should ensure that a BGP-LS Speaker does not accept UPDATE messages from a peer that only provides information to a BGP-LS Consumer by using the policy configuration options discussed in Sections 8.2.3 and 8.2.6. Generally, an operator is aware of the

BGP-LS Speaker's role and link-state peerings. Therefore, the operator can protect the BGP-LS Speaker from peers sending updates that may represent erroneous information, feedback loops, or false input.

An error or tampering of the link-state information that is originated into BGP-LS and propagated through the network for use by BGP-LS Consumers applications can result in the malfunction of those applications. Some examples of such risks are the origination of incorrect information that is not present or consistent with the IGP LSDB at the BGP-LS Producer, incorrect ordering of TLVs in the NLRI, or inconsistent origination from multiple BGP-LS Producers and updates to either the NLRI or BGP-LS Attribute during propagation (including discarding due to errors). These are not new risks from a BGP protocol perspective; however, in the case of BGP-LS, impact reflects on the consumer applications instead of BGP routing functionalities.

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 BGP Speakers are configured to receive such information. Similar security considerations also arise on the interface between BGP Speakers and BGP-LS Consumers, but their discussion is outside the scope of this document.

11. References

11.1. Normative References

[ENTNUM] IANA, "Private Enterprise Numbers (PENs)", <https://www.iana.org/assignments/enterprise-numbers/>.

[ISO10589] ISO, "Information technology - Telecommunications and information exchange between systems - 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 10589:2002, 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>.

[RFC4577] Rosen, E., Psenak, P., and P. Pillay-Esnault, "OSPF as the Provider/Customer Edge Protocol for BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4577, DOI 10.17487/RFC4577, June 2006, <https://www.rfc-editor.org/info/rfc4577>.

[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>.

[RFC6565] Pillay-Esnault, P., Moyer, P., Doyle, J., Ertekin, E., and M. Lundberg, "OSPFv3 as a Provider Edge to Customer Edge (PE-CE) Routing Protocol", RFC 6565, DOI 10.17487/RFC6565, June 2012, <https://www.rfc-editor.org/info/rfc6565>.

[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>.

[RFC7684] Psenak, P., Gredler, H., Shakir, R., Henderickx, W., Tantsura, J., and A. Lindem, "OSPFv2 Prefix/Link Attribute Advertisement", RFC 7684, DOI 10.17487/RFC7684, November 2015, <https://www.rfc-editor.org/info/rfc7684>.

[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>.

[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>.

[RFC8362] Lindem, A., Roy, A., Goethals, D., Reddy Vallem, V., and F. Baker, "OSPFv3 Link State Advertisement (LSA) Extensibility", RFC 8362, DOI 10.17487/RFC8362, April 2018, <https://www.rfc-editor.org/info/rfc8362>.

[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>.

11.2. Informative References

[RFC1918] Rekhter, Y., Moskowitz, B., Karrenberg, D., de Groot, G. J., 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.-P., 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>.

[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>.

[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>.

[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>.

[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>.

[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>.

[RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", RFC 7911, DOI 10.17487/RFC7911, July 2016, <https://www.rfc-editor.org/info/rfc7911>.

[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>.

[RFC9029] Farrel, A., "Updates to the Allocation Policy for the Border Gateway Protocol - Link State (BGP-LS) Parameters Registries", RFC 9029, DOI 10.17487/RFC9029, June 2021, <https://www.rfc-editor.org/info/rfc9029>.

[RFC9346] Chen, M., Ginsberg, L., Previdi, S., and D. Xiaodong, "IS-IS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 9346, DOI 10.17487/RFC9346, February 2023, <https://www.rfc-editor.org/info/rfc9346>.

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

2. Clarified aspects related to advertisement of link-state information from IGPs into BGP-LS in Section 4.

3. In Section 5.1, clarified aspects about TLV handling that apply to both the NLRI and BGP-LS Attribute parts as well as those that are applicable only for the NLRI portion. An implementation may have missed the part about the handling of an unknown TLV and so, based on [RFC7606] guidelines, might discard the unknown NLRI types. This aspect is now unambiguously clarified in Section 5.2. Also, the TLVs in the BGP-LS Attribute that are not ordered are not to be considered malformed.

4. Clarified aspects of mandatory and optional TLVs in both NLRI and BGP-LS Attribute portions all through the document.

5. In Section 5.3, the handling of a large-sized BGP-LS Attribute with growth in BGP-LS information is explained along with mitigation of errors arising out of it.

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

7. In Section 5.2, clarified the use of the Identifier field in the Link-State NLRI. It was defined ambiguously to refer to only multi-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.

8. 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 implementors on its usage for identification of IGP domains as against the use of the Identifier field carrying the BGP-LS Instance-ID when running multiple instances of IGP routing protocols. The original purpose of the BGP-LS Identifier was that, in conjunction with the ASN, it would uniquely identify the BGP-LS domain and that the combination of ASN and BGP-LS ID would be globally unique. However, the BGP-LS Instance-ID carried in the Identifier field in the fixed part of the NLRI also provides a similar functionality. Hence, the inclusion of the BGP-LS Identifier TLV is not necessary. If advertised, all BGP-LS Speakers within an IGP flooding-set (set of IGP nodes within which an LSP/LSA is flooded) had to use the same (ASN, BGP-LS ID) tuple, and if an IGP domain consists of multiple flooding-sets, then all BGP-LS Speakers within the IGP domain had to use the same (ASN, BGP-LS ID) tuple.

9. Clarified that the Area-ID TLV is mandatory in the Node Descriptor for the origination of information from OSPF except for when sourcing information from AS-scope LSAs where this TLV is not applicable. Also clarified the IS-IS area and area addresses.

10. Moved the 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 described the differences in the applicability of the R flags when the MT-ID TLV is used as the Link Descriptor TLV and Prefix

Attribute TLV. The MT-ID TLV use is now elevated to SHOULD when it is enabled in the underlying IGP.

11. 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.

12. Updated the usage of OSPF Route Type TLV to mandate its use for OSPF prefixes in Section 5.2.3.1 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.

13. Clarified the specific OSPFv2 and OSPFv3 protocol TLV space to be used in the Node, Link, and Prefix Opaque Attribute TLVs.

14. Clarified that the length of the Node Flag Bits and IGP Flags TLVs are to be one octet.

15. Updated the Node Name TLV in Section 5.3.1.3 with the OSPF specification.

16. Clarified the size of the IS-IS Narrow Metric advertisement via the IGP Metric TLV and the handling of the unused bits.

17. Clarified the advertisement of the prefix corresponding to the LAN segment in an OSPF network in Section 5.11.

18. Clarified the advertisement and support for OSPF-specific concepts like virtual links, sham links, and Type 4 LSAs in Sections 5.7 and 5.8.

19. Introduced the Private Use TLV code point space and specified their encoding in Section 5.4.

20. In Section 5.9, introduced where issues related to the consistency of reporting IGP link-state along with their solutions are covered.

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

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

23. Changed the management of BGP-LS IANA registries from "Specification Required" to "Expert Review" along with updated guidelines for designated experts, more specifically, the inclusion of changes introduced via [RFC9029] that are obsoleted by this document.

24. Added BGP-LS IANA registries with "Expert Review" policy for the flag fields of various TLVs that was missed out. Renamed the BGP-LS TLV registry and removed the "IS-IS TLV/Sub-TLV" column from it.

Acknowledgements

This document update to the BGP-LS specification [RFC7752] is a result of feedback and input 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 a LAN subnet for OSPF.

We would like to thank Balaji Rajagopalan, Srihari Sangli, Shraddha Hegde, Andrew Stone, Jeff Tantsura, Acee Lindem, Les Ginsberg, Jie Dong, Aijun Wang, Nandan Saha, Joel Halpern, and Gyan Mishra for their review and feedback on this document. Thanks to Tom Petch for his review and comments on the IANA Considerations section. We would also like to thank Jeffrey Haas for his detailed shepherd review and input for improving the document.

The detailed AD review by Alvaro Retana and his suggestions have helped improve this document significantly.

We would like to thank Robert Varga for his significant contribution 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].

Contributors

The following persons contributed significant text to [RFC7752] and this document. They should be considered coauthors.

Hannes Gredler
Rtbrick
Email: hannes@rtbrick.com

Jan Medved
Cisco Systems Inc.
United States of America
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
United States of America
Email: raysaikat@gmail.com

Author’s Address

Ketan Talaulikar (editor)
Cisco Systems
India
Email: ketant.ietf@gmail.com