Inter-Domain Routing
Internet-Draft
Intended status: Standards Track
Expires: April 18, 2016

H. Gredler, Ed.
Private Contributor
J. Medved
S. Previdi
Cisco Systems, Inc.
A. Farrel
Juniper Networks, Inc.
S. Ray

October 16, 2015

North-Bound Distribution of Link-State and TE Information using BGP
draft-ietf-idr-ls-distribution-13

Abstract

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

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

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

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

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

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

This Internet-Draft will expire on April 18, 2016.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction 3
2. Motivation and Applicability 5
2.1. MPLS-TE with PCE 5
2.2. ALTO Server Network API 6
3. Carrying Link State Information in BGP 7
3.1. TLV Format 7
3.2. The Link-State NLRI 8
3.2.1. Node Descriptors 12
3.2.2. Link Descriptors 16
3.2.3. Prefix Descriptors 17
3.3. The BGP-LS Attribute 19
3.3.1. Node Attribute TLVs 19
3.3.2. Link Attribute TLVs 23
3.3.3. Prefix Attribute TLVs 28
3.4. BGP Next Hop Information 31
3.5. Inter-AS Links 32
3.6. Router-ID Anchoring Example: ISO Pseudonode 32
3.7. Router-ID Anchoring Example: OSPF Pseudonode 33
3.8. Router-ID Anchoring Example: OSPFv2 to IS-IS Migration . 34
4. Link to Path Aggregation 34
4.1. Example: No Link Aggregation 35
4.2. Example: ASBR to ASBR Path Aggregation 35
4.3. Example: Multi-AS Path Aggregation 36
5. IANA Considerations 36

5.1. Guidance for Designated Experts 37
6. Manageability Considerations 37
6.1. Operational Considerations 37
6.1.1. Operations 37
6.1.2. Installation and Initial Setup 38
6.1.3. Migration Path 38
6.1.4. Requirements on Other Protocols and Functional Components 38
6.1.5. Impact on Network Operation 38
6.1.6. Verifying Correct Operation 38
6.2. Management Considerations 39
6.2.1. Management Information 39
6.2.2. Fault Management 39
6.2.3. Configuration Management 39
6.2.4. Accounting Management 40
6.2.5. Performance Management 40
6.2.6. Security Management 40
7. TLV/Sub-TLV Code Points Summary 40
8. Security Considerations 42
9. Contributors 43
10. Acknowledgements 43
11. References 43
11.1. Normative References 43
11.2. Informative References 45
Authors' Addresses 47

1. Introduction

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

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

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

This document describes a mechanism by which Link State and TE information can be collected from networks and shared with external components using the BGP routing protocol [RFC4271]. This is

achieved using a new BGP Network Layer Reachability Information (NLRI) encoding format. The mechanism is applicable to physical and virtual links. The mechanism described is subject to policy control.

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

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

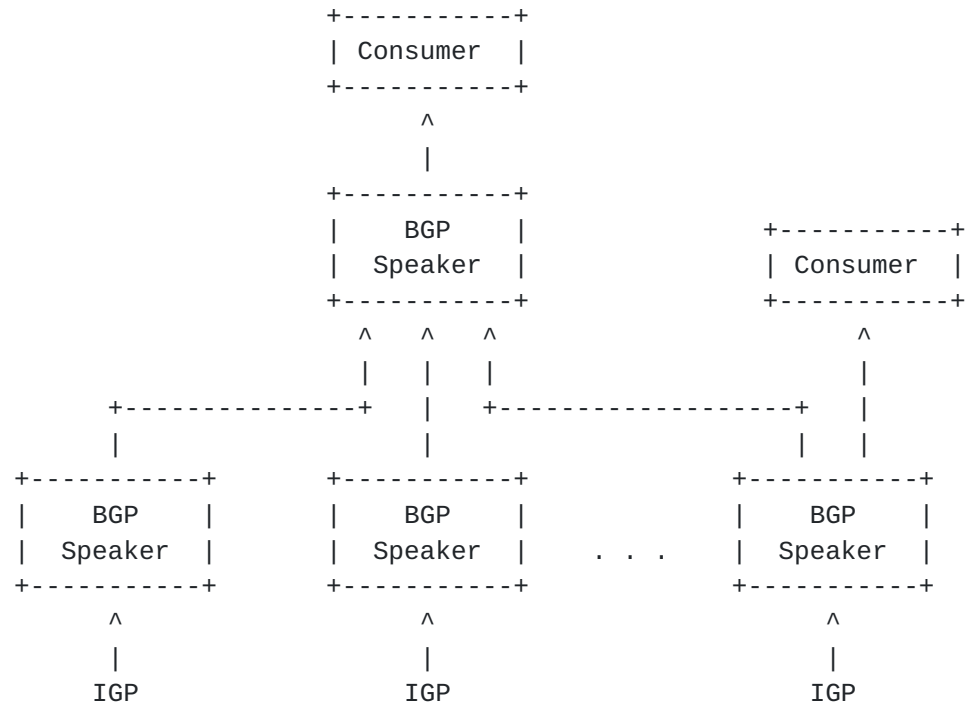


Figure 1: TE Link State info collection

A BGP Speaker may apply configurable policy to the information that it distributes. Thus, it may distribute the real physical topology from the LSDB or the TED. Alternatively, it may create an abstracted topology, where virtual, aggregated nodes are connected by virtual paths. Aggregated nodes can be created, for example, out of multiple routers in a POP. Abstracted topology can also be a mix of physical and virtual nodes and physical and virtual links. Furthermore, the

BGP Speaker can apply policy to determine when information is updated to the consumer so that there is reduction of information flow from the network to the consumers. Mechanisms through which topologies can be aggregated or virtualized are outside the scope of this document

2. Motivation and Applicability

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

2.1. MPLS-TE with PCE

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

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

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

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

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

high degree of churn, or when the PCE is responsible for multiple areas.

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

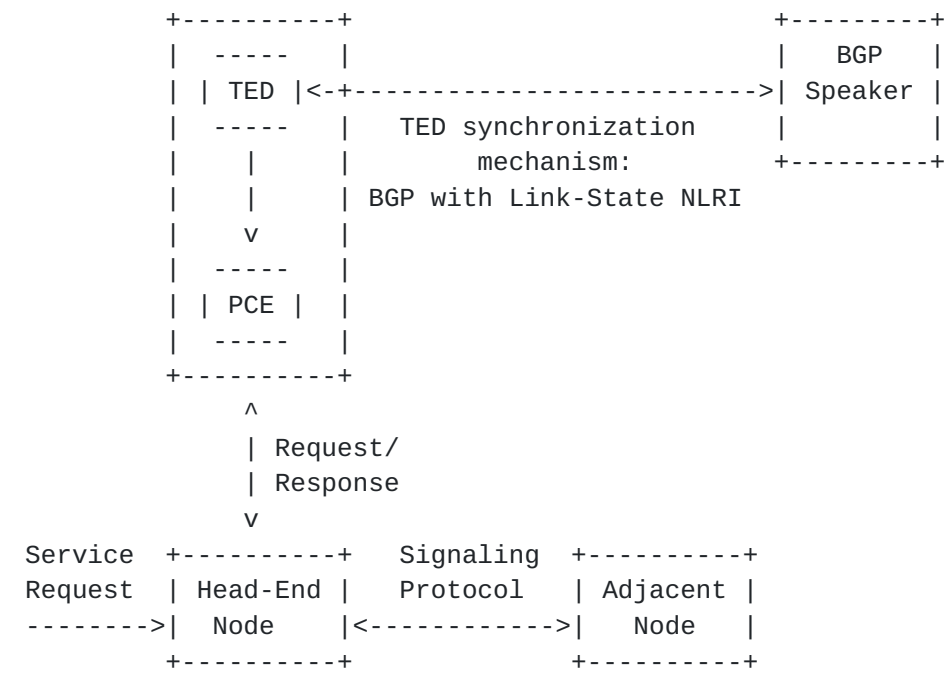


Figure 2: External PCE node using a TED synchronization mechanism

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

2.2. ALTO Server Network API

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

Cost Maps. Prefix data is carried and originated in BGP, TE data is originated and carried in an IGP. The mechanism defined in this document provides a single interface through which an ALTO Server can retrieve all the necessary prefix and network topology data from the underlying network. Note an ALTO Server can use other mechanisms to get network data, for example, peering with multiple IGP and BGP Speakers.

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

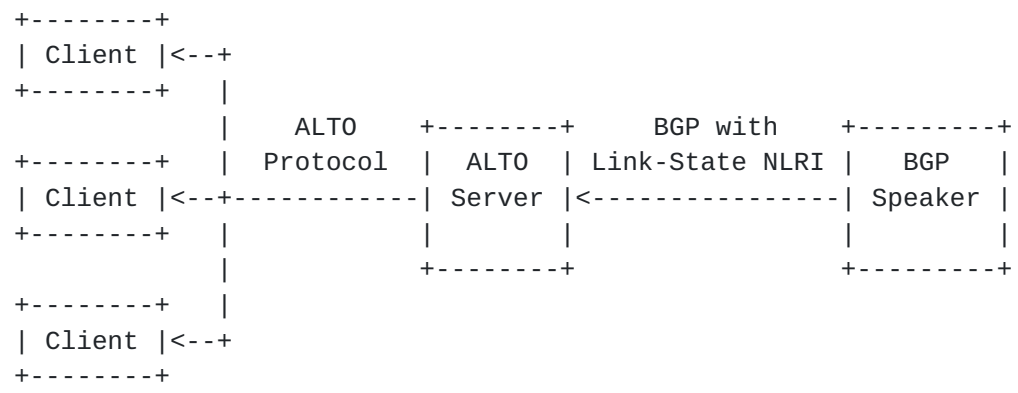


Figure 3: ALTO Server using network topology information

3. Carrying Link State Information in BGP

This specification contains two parts: definition of a new BGP NLRI that describes links, nodes and prefixes comprising IGP link state information, and definition of a new BGP path attribute (BGP-LS attribute) that carries link, node and prefix properties and attributes, such as the link and prefix metric or auxiliary Router-IDs of nodes, etc.

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

3.1. TLV Format

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

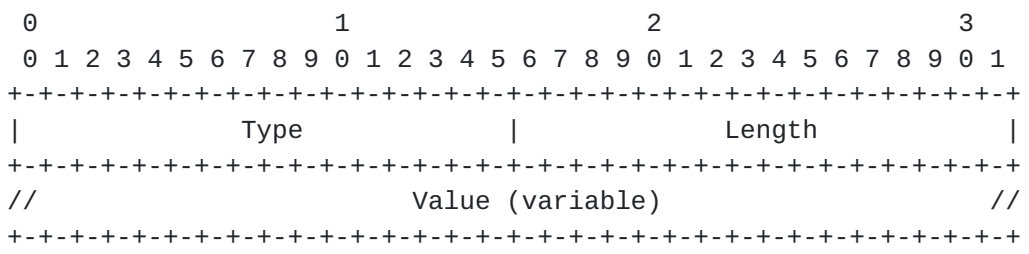


Figure 4: TLV format

The Length field defines the length of the value portion in octets (thus a TLV with no value portion would have a length of zero). The TLV is not padded to four-octet alignment. Unrecognized types MUST be preserved and propagated. In order to compare NLRIs with unknown TLVs all TLVs MUST be ordered in ascending order by TLV Type. If there are more TLVs of the same type, then the TLVs MUST be ordered in ascending order of the TLV value within the TLVs with the same type by treating the entire value field an opaque hexadecimal string and comparing leftmost octets first regardless of the length of the string. . All TLVs that are not specified as mandatory are considered optional.

3.2. The Link-State NLRI

The MP_REACH_NLRI and MP_UNREACH_NLRI attributes are BGP's containers for carrying opaque information. Each Link-State NLRI describes 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 TBD.

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

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

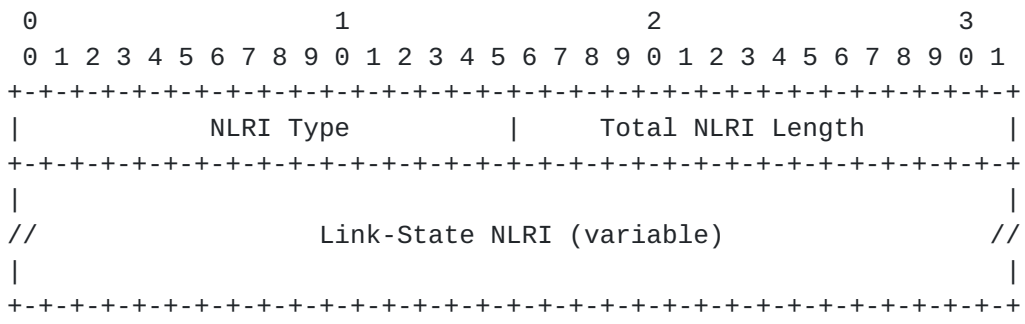


Figure 5: Link-State AFI 16388 / SAFI 71 NLRI Format

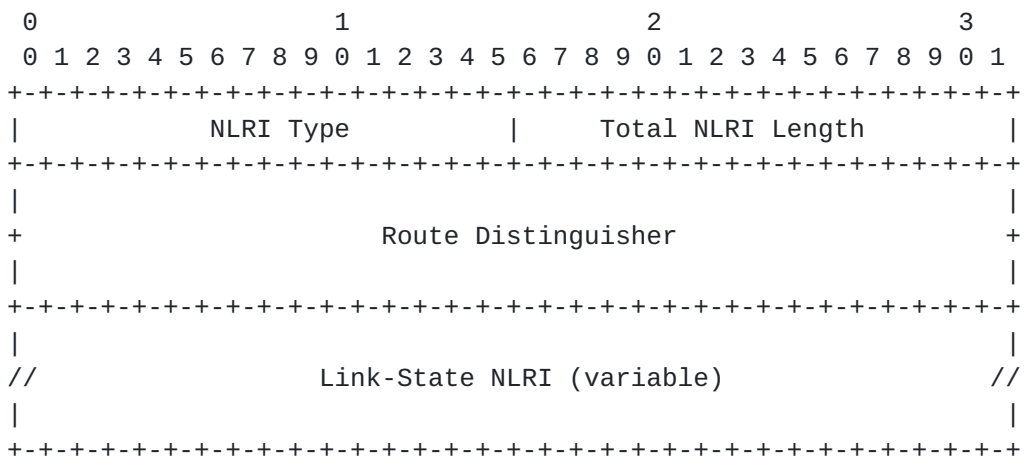


Figure 6: Link-State VPN AFI 16388 / SAFI TBD NLRI Format

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

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.

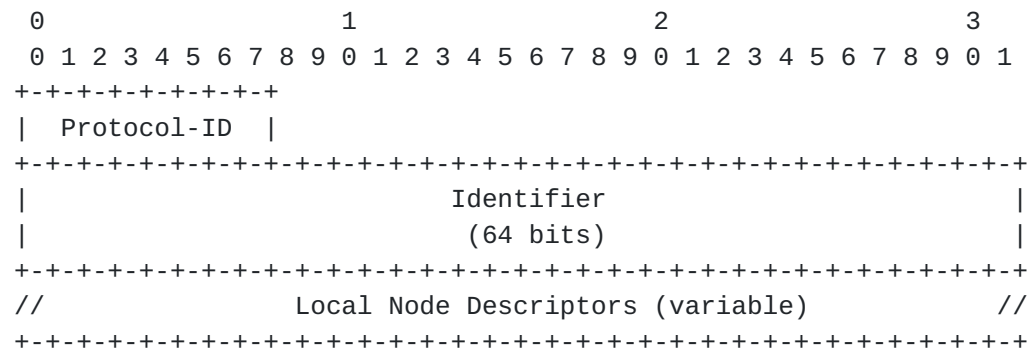


Figure 7: The Node NLRI format

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

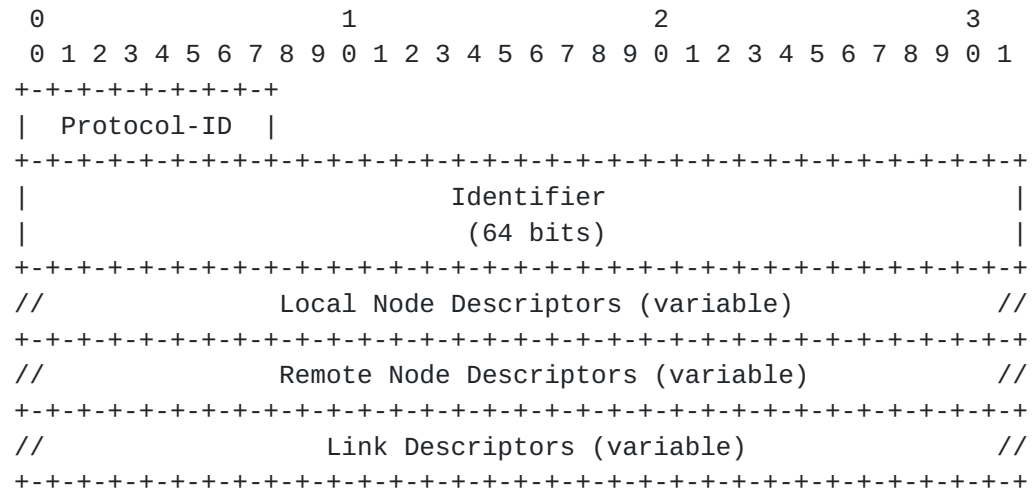


Figure 8: The Link NLRI format

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

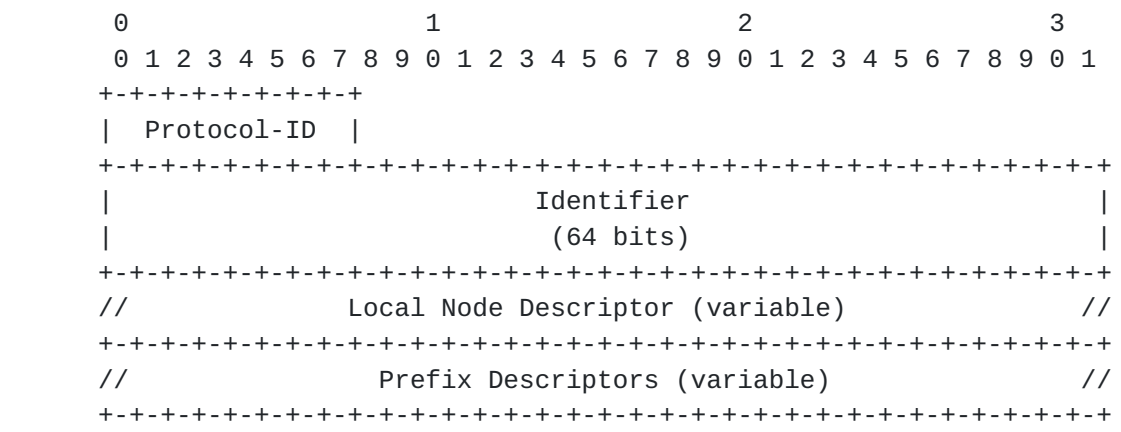


Figure 9: The IPv4/IPv6 Topology Prefix NLRI format

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 got direct access to interface information and wants to advertise a local link then the protocol-ID 'Direct' SHOULD be used. For modeling virtual links, like described in Section 4 the protocol-ID 'Static configuration' SHOULD be used.

Both OSPF and IS-IS MAY run multiple routing protocol instances over the same link. See [RFC6822] and [RFC6549]. These instances define independent "routing universes". The 64-Bit 'Identifier' field is used to identify the "routing universe" where the NLRI belongs. The NLRIs representing Link-state objects (nodes, links or prefixes) from the same routing universe MUST have the same 'Identifier' value. NLRIs with different 'Identifier' values MUST be considered to be from different routing universes. Table 3 lists the 'Identifier' values that are defined as well-known in this draft.

Identifier	Routing Universe
0	Default Layer 3 Routing topology
1-31	Reserved

Table 3: Well-known Instance Identifiers

If a given Protocol does not support multiple routing universes then it SHOULD set the 'Identifier' field according to Table 3. However an implementation MAY make the 'Identifier' configurable, for a given protocol.

Each Node Descriptor and Link Descriptor consists of one or more TLVs described in the following sections.

3.2.1. Node Descriptors

Each link is anchored by a pair of Router-IDs that are used by the underlying IGP, namely, 48 Bit ISO System-ID for IS-IS and 32 bit Router-ID for OSPFv2 and OSPFv3. An IGP may use one or more additional auxiliary Router-IDs, mainly for traffic engineering purposes. For example, IS-IS may have one or more IPv4 and IPv6 TE Router-IDs [RFC5305], [RFC6119]. These auxiliary Router-IDs MUST be included in the link attribute described in Section 3.3.2.

It is desirable that the Router-ID assignments inside the Node Descriptor are globally unique. However there may be Router-ID spaces (e.g. ISO) where no global registry exists, or worse, Router-IDs have been allocated following private-IP RFC 1918 [RFC1918] allocation. BGP-LS uses the Autonomous System (AS) Number and BGP-LS Identifier (see Section 3.2.1.4) to disambiguate the Router-IDs, as described in Section 3.2.1.1.

3.2.1.1. Globally Unique Node/Link/Prefix Identifiers

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

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

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

We define an "IGP domain" to be the set of nodes (hence, by extension links and prefixes), within which, each node has a unique IGP representation by using the combination of Area-ID, Router-ID, Protocol, Topology-ID, and Instance ID. The problem is that BGP may receive node/link/prefix information from multiple independent "IGP domains" and we need to distinguish between them. Moreover, we can't assume there is always one and only one IGP domain per AS. During IGP transitions it may happen that two redundant IGPs are in place.

In Section 3.2.1.4 a set of sub-TLVs is described, which allows specification of a flexible key for any given Node/Link information such that global uniqueness of the NLRI is ensured.

3.2.1.2. 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 length of this TLV is variable. The value contains one or more Node Descriptor Sub-TLVs defined in Section 3.2.1.4.

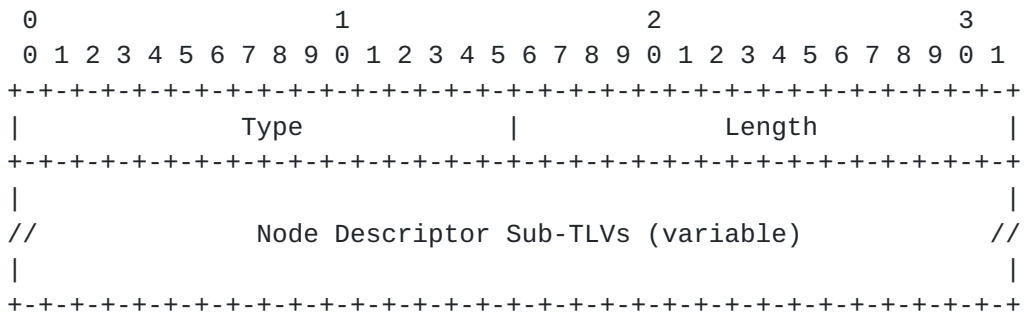


Figure 10: Local Node Descriptors TLV format

3.2.1.3. Remote Node Descriptors

The Remote Node Descriptors contains Node Descriptors for the node anchoring the remote end of the link. This is a mandatory TLV for link NLRIs. The length of this TLV is variable. The value contains one or more Node Descriptor Sub-TLVs defined in Section 3.2.1.4.

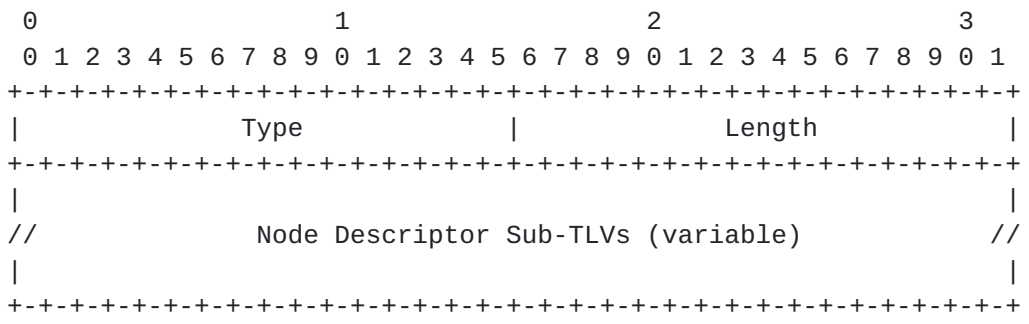


Figure 11: Remote Node Descriptors TLV format

3.2.1.4. Node Descriptor Sub-TLVs

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

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

Table 4: Node Descriptor Sub-TLVs

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

Autonomous System: opaque value (32 Bit AS Number)

BGP-LS Identifier: opaque value (32 Bit ID). In conjunction with ASN, uniquely identifies the BGP-LS domain. The combination of ASN and BGP-LS ID MUST be globally unique. All BGP-LS speakers within an IGP flooding-set (set of IGP nodes within which an LSP/LSA is flooded) MUST use the same ASN, BGP-LS ID tuple. If an IGP domain consists of multiple flooding-sets, then all BGP-LS speakers within the IGP domain SHOULD use the same ASN, BGP-LS ID tuple. The ASN, BGP Router-ID tuple (which is globally unique [RFC6286]) of one of the BGP-LS speakers within the flooding-set (or IGP domain) may be used for all BGP-LS speakers in that flooding-set (or IGP domain).

Area ID: It is used to identify the 32 Bit area to which the NLRI belongs. Area Identifier allows the different NLRIs of the same router to be discriminated.

IGP Router ID: opaque value. This is a mandatory TLV. For an IS-IS non-Pseudonode, this contains 6 octet ISO node-ID (ISO system-ID). For an IS-IS Pseudonode corresponding to a LAN, this contains 6 octet ISO node-ID of the "Designated Intermediate System" (DIS) followed by one 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 protocol identifier enables the decoder to determine the type of the node.

There can be at most one instance of each sub-TLV type present in any Node Descriptor. The sub-TLVs within a Node descriptor MUST be arranged in ascending order by sub-TLV type. This needs to be done in order to compare NLRIs, even when an implementation encounters an unknown sub-TLV. Using stable sorting an implementation can do binary comparison of NLRIs and hence allow incremental deployment of new key sub-TLVs.

3.2.1.5. Multi-Topology ID

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

Semantics of the IS-IS MT-ID are defined in RFC5120, Section 7.2 [RFC5120]. Semantics of the OSPF MT-ID are defined in RFC4915, Section 3.7 [RFC4915]. If the value in the MT-ID TLV is derived from OSPF, then the upper 9 bits MUST be set to 0. Bits R are reserved, SHOULD be set to 0 when originated and ignored on receipt.

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

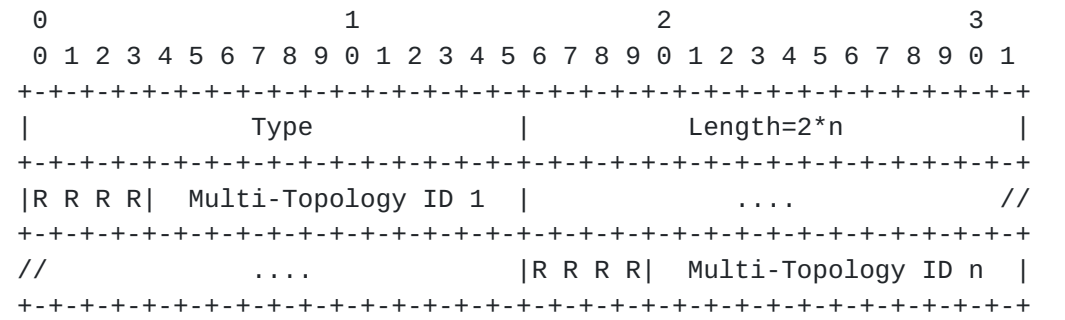


Figure 12: Multi-Topology ID TLV format

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

The MT-ID TLV MAY be present in a Link Descriptor, a Prefix Descriptor, or in the BGP-LS attribute of a node NLRI. In a Link or Prefix Descriptor, only a single MT-ID TLV containing the MT-ID of the topology where the link or the prefix is reachable is allowed. In case one wants to advertise multiple topologies for a given Link Descriptor or Prefix Descriptor, multiple NLRIs need to be generated where each NLRI contains an unique MT-ID. 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.

3.2.2. Link Descriptors

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

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

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

TLV Code Point	Description	IS-IS TLV /Sub-TLV	Value defined in:
258	Link Local/Remote Identifiers	22/4	[RFC5307]/1.1
259	IPv4 interface address	22/6	[RFC5305]/3.2
260	IPv4 neighbor address	22/8	[RFC5305]/3.3
261	IPv6 interface address	22/12	[RFC6119]/4.2
262	IPv6 neighbor address	22/13	[RFC6119]/4.3
263	Multi-Topology Identifier	---	Section 3.2.1.5

Table 5: Link Descriptor TLVs

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

If interface and neighbor addresses, either IPv4 or IPv6, are present, then the IP address TLVs are included in the link descriptor, but not the link local/remote Identifier TLV. The link local/remote identifiers MAY be included in the link attribute.

If interface and neighbor addresses are not present and the link local/remote identifiers are present, then the link local/remote Identifier TLV is included in the link descriptor.

The Multi-Topology Identifier TLV is included in link descriptor if that information is present.

3.2.3. 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 valid as Prefix Descriptors in the IPv4/IPv6 Prefix NLRI:

TLV Code Point	Description	Length	Value defined in:
263	Multi-Topology Identifier	variable	Section 3.2.1.5
264	OSPF Route Type	1	Section 3.2.3.1
265	IP Reachability Information	variable	Section 3.2.3.2

Table 6: Prefix Descriptor TLVs

3.2.3.1. OSPF Route Type

OSPF Route Type is an optional TLV that MAY be present in Prefix NLRIs. It is used to identify the OSPF route-type of the prefix. It is used when an OSPF prefix is advertised in the OSPF domain with multiple route-types. The Route Type TLV allows the discrimination of these advertisements. The format of the OSPF Route Type TLV is shown in the following figure.

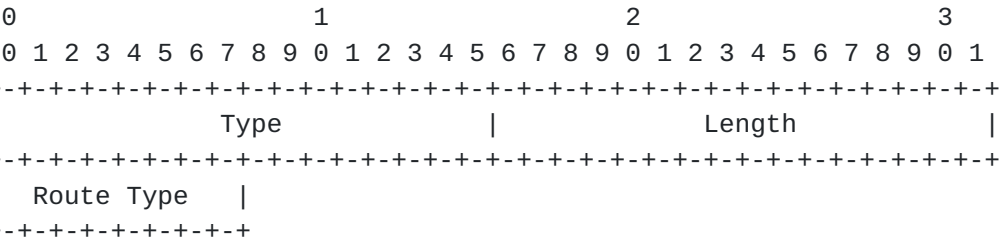


Figure 13: OSPF Route Type TLV Format

where the Type and Length fields of the TLV are defined in Table 6. The OSPF Route Type field values are defined in the OSPF protocol, and can be one of the following:

Intra-Area (0x1)

Inter-Area (0x2)

External 1 (0x3)

External 2 (0x4)

NSSA 1 (0x5)

NSSA 2 (0x6)

3.2.3.2. IP Reachability Information

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

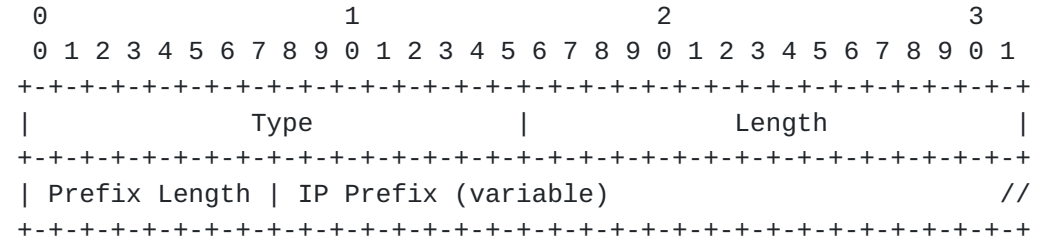


Figure 14: IP Reachability Information TLV Format

The Type and Length fields of the TLV are defined in Table 6. The following two fields determine the address-family reachability information. The 'Prefix Length' field contains the length of the prefix in bits. The 'IP Prefix' field contains the most significant octets of the prefix; i.e., 1 octet for prefix length 1 up to 8, 2 octets for prefix length 9 to 16, 3 octets for prefix length 17 up to 24 and 4 octets for prefix length 25 up to 32, etc.

3.3. The BGP-LS Attribute

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

3.3.1. Node Attribute TLVs

Node attribute TLVs are the TLVs that may be encoded in the BGP-LS attribute with a node NLRI. The following node attribute TLVs are defined:

TLV Code Point	Description	Length	Value defined in:
263	Multi-Topology Identifier	variable	Section 3.2.1.5
1024	Node Flag Bits	1	Section 3.3.1.1
1025	Opaque Node Properties	variable	Section 3.3.1.5
1026	Node Name	variable	Section 3.3.1.3
1027	IS-IS Area Identifier	variable	Section 3.3.1.2
1028	IPv4 Router-ID of Local Node	4	[RFC5305]/4.3
1029	IPv6 Router-ID of Local Node	16	[RFC6119]/4.1

Table 7: Node Attribute TLVs

3.3.1.1. Node Flag Bits TLV

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

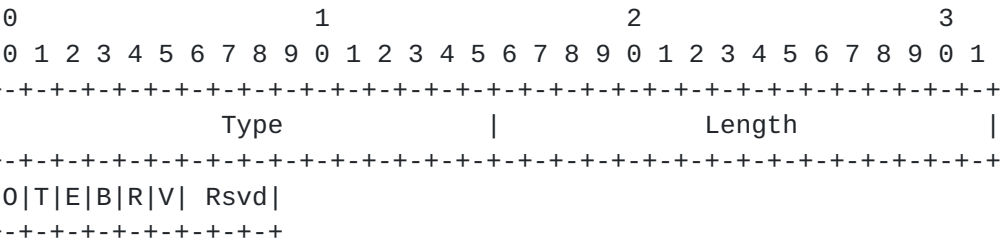


Figure 15: Node Flag Bits TLV format

The bits are defined as follows:

Bit	Description	Reference
'O'	Overload Bit	[RFC1195]
'T'	Attached Bit	[RFC1195]
'E'	External Bit	[RFC2328]
'B'	ABR Bit	[RFC2328]
'R'	Router Bit	[RFC5340]
'V'	V6 Bit	[RFC5340]
Reserved (Rsvd)	Reserved for future use	

Table 8: Node Flag Bits Definitions

3.3.1.2. IS-IS Area Identifier TLV

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

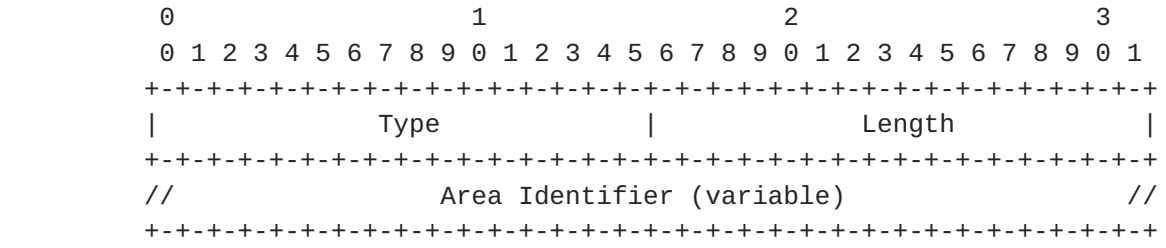


Figure 16: IS-IS Area Identifier TLV Format

3.3.1.3. Node Name TLV

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

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

Although [RFC5301] is an IS-IS specific extension, usage of the Node Name TLV is possible for all protocols. How a router derives and injects node names for e.g. OSPF nodes, is outside of the scope of this document.

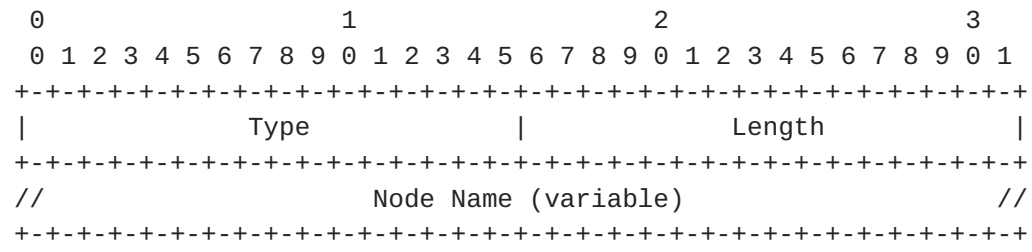


Figure 17: Node Name format

3.3.1.4. Local IPv4/IPv6 Router-ID

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 like correlating a Node-ID between different protocols. If there is more than one auxiliary Router-ID of a given type, then each one is encoded in its own TLV.

3.3.1.5. Opaque Node Attribute TLV

The Opaque Node Attribute TLV is an envelope that transparently carries optional node attribute TLVs advertised by a router. An originating router shall use this TLV for encoding information specific to the protocol advertised in the NLRI header Protocol-ID field or new protocol extensions to the protocol as advertised in the NLRI header Protocol-ID field for which there is no protocol neutral representation in the BGP link-state NLRI. The primary use of the Opaque Node Attribute TLV is to bridge the document lag between e.g. a new IGP Link-state attribute being defined and the 'protocol-neutral' BGP-LS extensions being published. A router for example could use this extension in order to advertise the native protocols node attribute TLVs, such as the OSPF Router Informational Capabilities TLV defined in [RFC4970], or the IGP TE Node Capability Descriptor TLV described in [RFC5073].

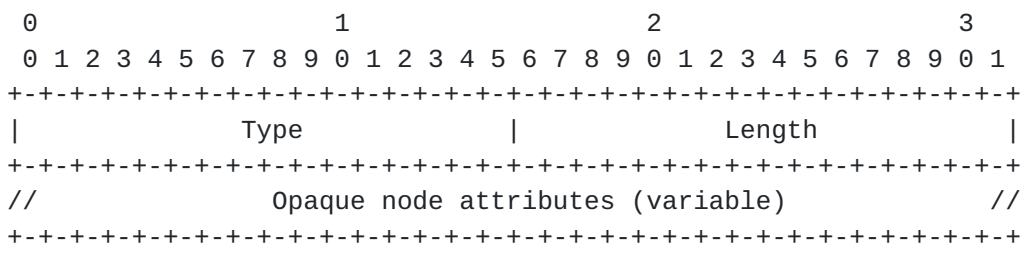


Figure 18: Opaque Node attribute format

3.3.2. 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 3.1. The format and semantics of the 'value' fields in some 'Link Attribute' TLVs correspond to the format and semantics of value fields in IS-IS Extended IS Reachability sub-TLVs, defined in [RFC5305] and [RFC5307]. Other 'Link Attribute' TLVs are defined in this document. Although the encodings for 'Link Attribute' TLVs were originally defined for IS-IS, the TLVs can carry data sourced either by IS-IS or OSPF.

The following 'Link Attribute' TLVs are valid in the BGP-LS attribute with a link NLRI:

TLV Code Point	Description	IS-IS TLV /Sub-TLV	Defined in:
1028	IPv4 Router-ID of Local Node	134/---	[RFC5305]/4.3
1029	IPv6 Router-ID of Local Node	140/---	[RFC6119]/4.1
1030	IPv4 Router-ID of Remote Node	134/---	[RFC5305]/4.3
1031	IPv6 Router-ID of Remote Node	140/---	[RFC6119]/4.1
1088	Administrative group (color)	22/3	[RFC5305]/3.1
1089	Maximum link bandwidth	22/9	[RFC5305]/3.3
1090	Max. reservable link bandwidth	22/10	[RFC5305]/3.5
1091	Unreserved bandwidth	22/11	[RFC5305]/3.6
1092	TE Default Metric	22/18	Section 3.3.2.3/
1093	Link Protection Type	22/20	[RFC5307]/1.2
1094	MPLS Protocol Mask	---	Section 3.3.2.2
1095	IGP Metric	---	Section 3.3.2.4
1096	Shared Risk Link Group	---	Section 3.3.2.5
1097	Opaque link attribute	---	Section 3.3.2.6
1098	Link Name attribute	---	Section 3.3.2.7

Table 9: Link Attribute TLVs

3.3.2.1. IPv4/IPv6 Router-ID

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 are more than one auxiliary Router-ID of a given type, then multiple TLVs are used to encode them.

3.3.2.2. MPLS Protocol Mask TLV

The MPLS Protocol Mask TLV carries a bit mask describing which MPLS signaling protocols are enabled. The length of this TLV is 1. The

value is a bit array of 8 flags, where each bit represents an MPLS Protocol capability.

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

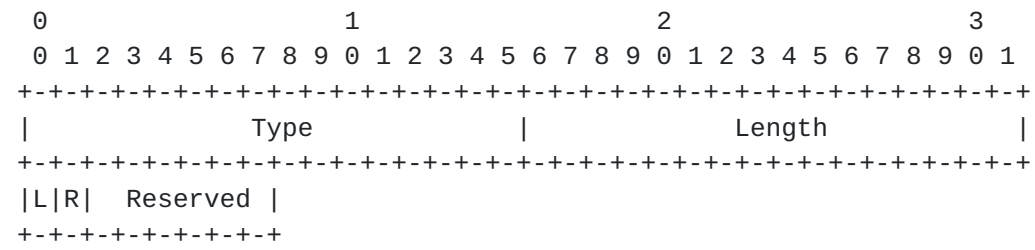


Figure 19: MPLS Protocol Mask TLV

The following bits are defined:

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

Table 10: MPLS Protocol Mask TLV Codes

3.3.2.3. TE Default Metric TLV

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

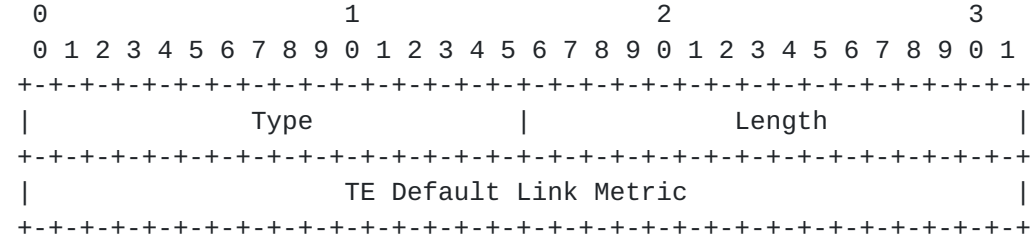


Figure 20: TE Default Metric TLV format

3.3.2.4. IGP Metric TLV

The IGP Metric TLV carries the metric for this link. The length of this TLV is variable, depending on the metric width of the underlying protocol. IS-IS small metrics have a length of 1 octet (the two most significant bits are ignored). OSPF link metrics have a length of two octets. IS-IS wide-metrics have a length of three octets.

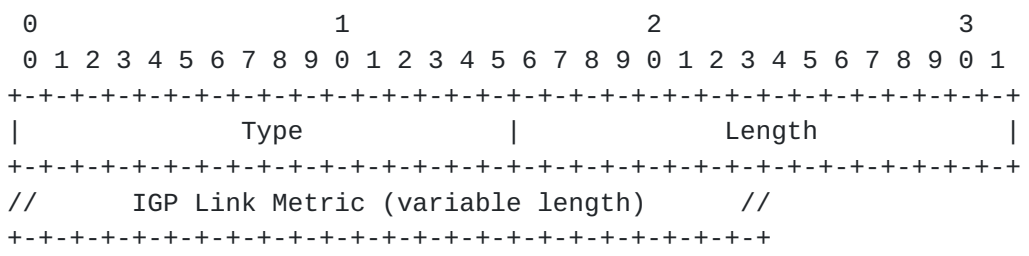


Figure 21: Metric TLV format

3.3.2.5. Shared Risk Link Group TLV

The Shared Risk Link Group (SRLG) TLV carries the Shared Risk Link Group information (see Section 2.3, "Shared Risk Link Group Information", of [RFC4202]). It contains a data structure consisting of a (variable) list of SRLG values, where each element in the list has 4 octets, as shown in Figure 22. The length of this TLV is 4 * (number of SRLG values).

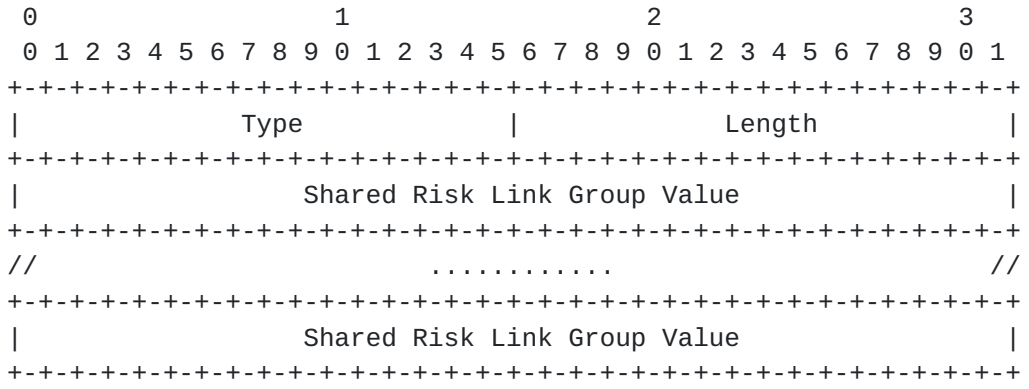


Figure 22: Shared Risk Link Group TLV format

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

3.3.2.6. Opaque Link Attribute TLV

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

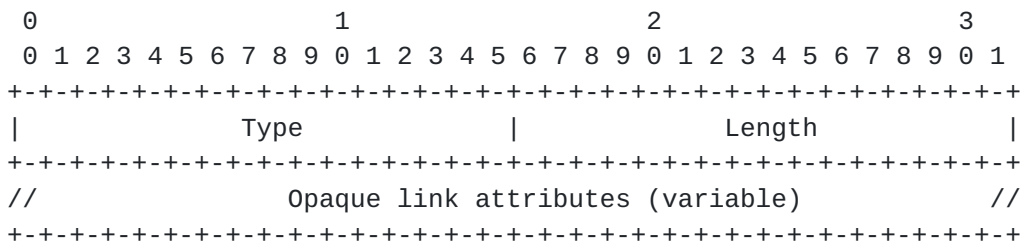


Figure 23: Opaque link attribute format

3.3.2.7. Link Name TLV

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

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

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

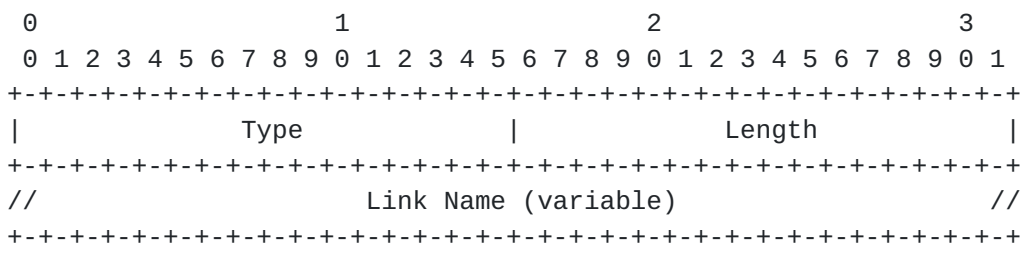


Figure 24: Link Name format

3.3.3. 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 MUST be reflected into the BGP-LS attribute with a link NLRI. This section describes the different attributes related to the IPv4/IPv6 prefixes. Prefix Attributes TLVs SHOULD be used when advertising NLRI types 3 and 4 only. The following attributes TLVs are defined:

TLV Code Point	Description	Length	Reference
1152	IGP Flags	1	Section 3.3.3.1
1153	Route Tag	4*n	Section 3.3.3.2
1154	Extended Tag	8*n	Section 3.3.3.3
1155	Prefix Metric	4	Section 3.3.3.4
1156	OSPF Forwarding Address	4	Section 3.3.3.5
1157	Opaque Prefix Attribute	variable	Section 3.3.3.6

Table 11: Prefix Attribute TLVs

3.3.3.1. IGP Flags TLV

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

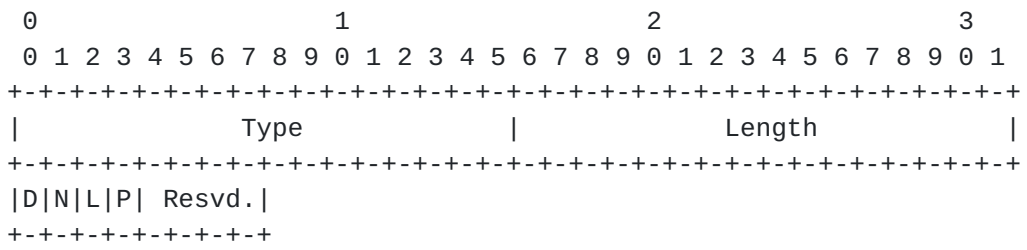


Figure 25: IGP Flag TLV format

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

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

Table 12: IGP Flag Bits Definitions

3.3.3.2. Route Tag

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

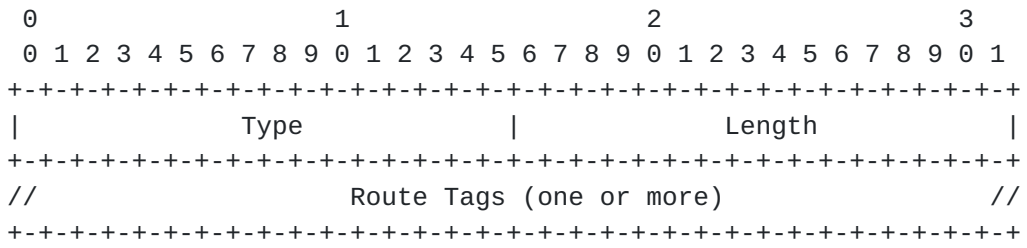


Figure 26: IGP Route TAG TLV format

Length is a multiple of 4.

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

3.3.3.3. Extended Route Tag

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

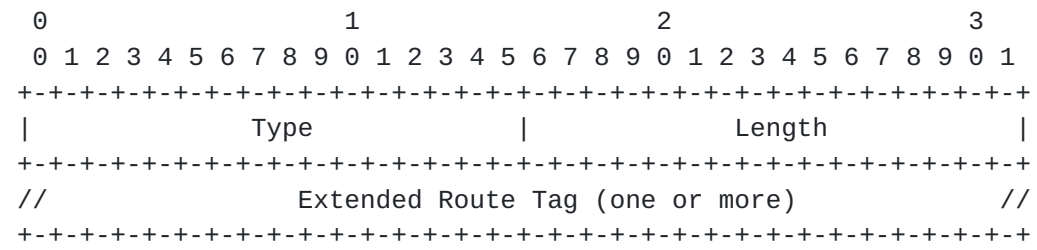


Figure 27: Extended IGP Route TAG TLV format

Length is a multiple of 8.

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

3.3.3.4. Prefix Metric TLV

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

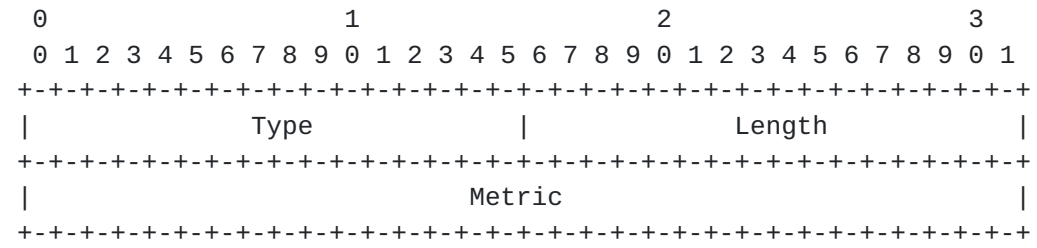


Figure 28: Prefix Metric TLV Format

Length is 4.

3.3.3.5. OSPF Forwarding Address TLV

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

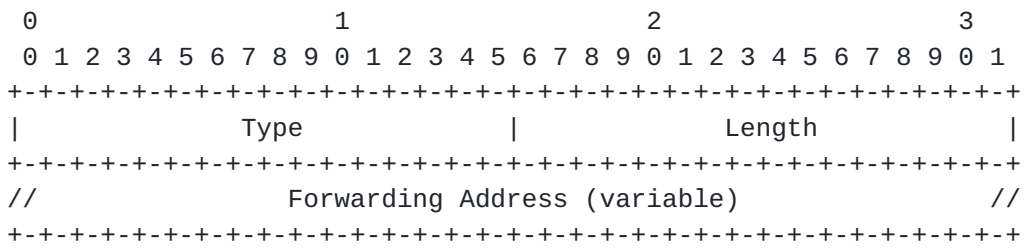


Figure 29: OSPF Forwarding Address TLV Format

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

3.3.3.6. Opaque Prefix Attribute TLV

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

The format of the TLV is as follows:

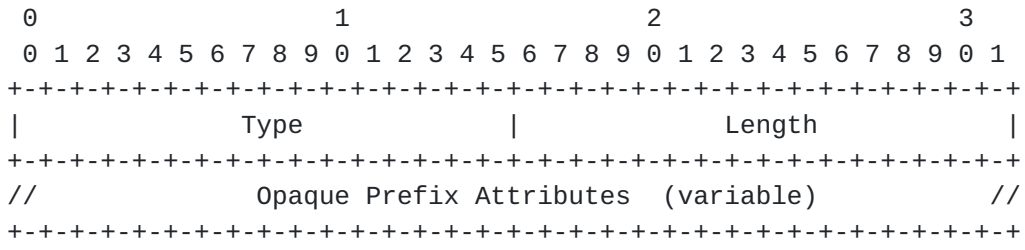


Figure 30: Opaque Prefix Attribute TLV Format

Type is as specified in Table 11 and Length is variable.

3.4. BGP Next Hop Information

BGP link-state information for both IPv4 and IPv6 networks can be carried over either an IPv4 BGP session, or an IPv6 BGP session. If 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 end-point

address of the BGP session. The next hop address MUST be encoded as described in [RFC4760]. The length field of the next hop address will specify the next hop address-family. If the next hop length is 4, then the next hop is an IPv4 address; if the next hop length is 16, then it is a global IPv6 address and if the next hop length is 32, then there is one global IPv6 address followed by a link-local IPv6 address. The link-local IPv6 address should be used as described in [RFC2545]. For VPN SAFI, as per custom, an 8 byte route-distinguisher set to all zero is prepended to the next hop.

The BGP Next Hop attribute is used by each BGP-LS speaker to validate the NLRI it receives. In case identical NLRIs are sourced by multiple originators the BGP next hop attribute is used to tie-break as per the standard BGP path decision process. This specification doesn't mandate any rule regarding the re-write of the BGP Next Hop attribute.

3.5. Inter-AS Links

The main source of TE information is the IGP, which is not active on inter-AS links. In some cases, the IGP may have information of inter-AS links ([RFC5392], [RFC5316]). In other cases, an implementation SHOULD provide a means to inject inter-AS links into BGP-LS. The exact mechanism used to provision the inter-AS links is outside the scope of this document

3.6. Router-ID Anchoring Example: ISO Pseudonode

Encoding of a broadcast LAN in IS-IS provides a good example of how Router-IDs are encoded. Consider Figure 31. This represents a Broadcast LAN between a pair of routers. The "real" (=non pseudonode) routers have both an IPv4 Router-ID and IS-IS Node-ID. The pseudonode does not have an IPv4 Router-ID. Node1 is the DIS for the LAN. Two unidirectional links (Node1, Pseudonode 1) 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 containing ISO-ID of Node1, 1920.0000.2001; the IGP Router-ID TLV of the remote node descriptor is 7 octets long containing 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 containing the ISO-ID of Pseudonode1, 1920.0000.2001.02; the IGP Router-ID TLV of the remote node descriptor is 6 octets long

containing 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.

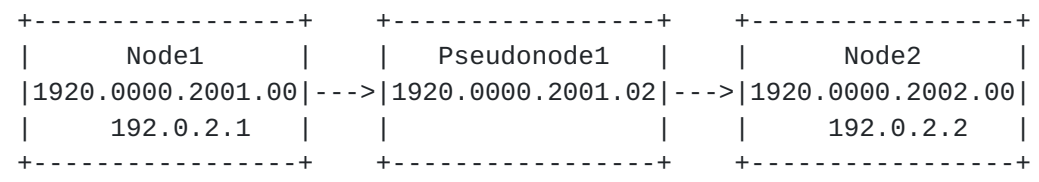


Figure 31: IS-IS Pseudonodes

3.7. Router-ID Anchoring Example: OSPF Pseudonode

Encoding of a broadcast LAN in OSPF provides a good example of how Router-IDs and local Interface IPs are encoded. Consider Figure 32. This represents a Broadcast LAN between a pair of routers. The "real" (=non pseudonode) routers have both an IPv4 Router-ID and an Area Identifier. The pseudonode does have an IPv4 Router-ID, an IPv4 interface Address (for disambiguation) and an OSPF Area. Node1 is the DR for the LAN, hence its local IP address 10.1.1.1 is used both as the Router-ID and Interface IP for the Pseudonode keys. Two unidirectional links (Node1, Pseudonode 1) and (Pseudonode1, Node2) are being generated.

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

o Local Node Descriptor

  TLV #515: IGP Router ID: 11.11.11.11

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

o Remote Node Descriptor

  TLV #515: IGP Router ID: 11.11.11.11:10.1.1.1

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

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

o Local Node Descriptor

  TLV #515: IGP Router ID: 11.11.11.11:10.1.1.1

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

o Remote Node Descriptor

TLV #515: IGP Router ID: 33.33.33.34

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

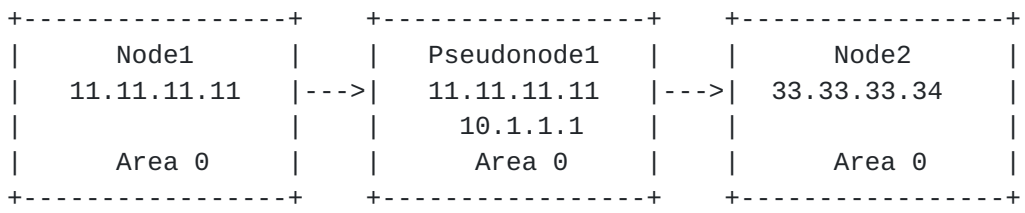


Figure 32: OSPF Pseudonodes

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

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

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

4. Link to Path Aggregation

Distribution of all links available in the global Internet is certainly possible, however not desirable from a scaling and privacy point of view. Therefore an implementation may support link to path aggregation. Rather than advertising all specific links of a domain, an ASBR may advertise an "aggregate link" between a non-adjacent pair of nodes. The "aggregate link" represents the aggregated set of link properties between a pair of non-adjacent nodes. The actual methods to compute the path properties (of bandwidth, metric) are outside the scope of this document. The decision whether to advertise all

specific links or aggregated links is an operator's policy choice. To highlight the varying levels of exposure, the following deployment examples are discussed.

4.1. Example: No Link Aggregation

Consider Figure 33. Both AS1 and AS2 operators want to protect their inter-AS {R1,R3}, {R2, R4} links using RSVP-FRR LSPs. If R1 wants to compute its link-protection LSP to R3 it needs to "see" an alternate path to R3. Therefore the AS2 operator exposes its topology. All BGP TE enabled routers in AS1 "see" the full topology of AS2 and therefore can compute a backup path. Note that the decision if the direct link between {R3, R4} or the {R4, R5, R3) path is used is made by the computing router.

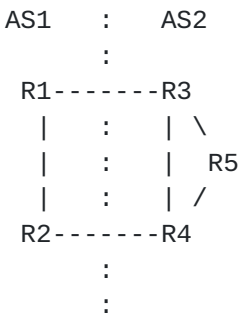


Figure 33: No link aggregation

4.2. Example: ASBR to ASBR Path Aggregation

The brief difference between the "no-link aggregation" example and this example is that no specific link gets exposed. Consider Figure 34. The only link which gets advertised by AS2 is an "aggregate" link between R3 and R4. This is enough to tell AS1 that there is a backup path. However the actual links being used are hidden from the topology.

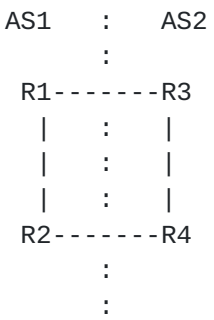


Figure 34: ASBR link aggregation

4.3. Example: Multi-AS Path Aggregation

Service providers in control of multiple ASes may even decide to not expose their internal inter-AS links. Consider Figure 35. AS3 is modeled as a single node which connects to the border routers of the aggregated domain.



Figure 35: Multi-AS aggregation

5. IANA Considerations

This document is the reference for Address Family Number 16388, 'BGP-LS'.

This document requests code point 71 from the registry of Subsequent Address Family Numbers named 'BGP-LS'.

This document requests a code point from the registry of Subsequent Address Family Numbers named 'BGP-LS-VPN'. The SAFI assignment does not need to be out of the range 1-63 and may come out of the "First Come First Served" range 128-240.

This document requests a code point from the BGP Path Attributes registry. As per early allocation procedure this is Path Attribute 29.

All the following Registries are BGP-LS specific and shall be accessible under the following URL: "http://www.iana.org/assignments/bgp-ls-parameters" Title "Border Gateway Protocol - Link State (BGP-LS) Parameters"

This document requests creation of a new registry for BGP-LS NLRI-Types. Value 0 is reserved. The maximum value is 65535. The registry will be initialized as shown in Table 1. Allocations within the registry will require documentation of the proposed use of the allocated value (=Specification required) and approval by the Designated Expert assigned by the IESG (see [RFC5226]).

This document requests creation of a new registry for BGP-LS Protocol-IDs. Value 0 is reserved. The maximum value is 255. The registry will be initialized as shown in Table 2. Allocations within the registry will require documentation of the proposed use of the allocated value (=Specification required) and approval by the Designated Expert assigned by the IESG (see [RFC5226]).

This document requests creation of a new registry for BGP-LS Well-known Instance-IDs. The registry will be initialized as shown in Table 3. Allocations within the registry will require documentation of the proposed use of the allocated value (=Specification required) and approval by the Designated Expert assigned by the IESG (see [RFC5226]).

This document requests creation of a new registry for node anchor, link descriptor and link attribute TLVs. Values 0-255 are reserved. Values 256-65535 will be used for code points. The registry will be initialized as shown in Table 13. Allocations within the registry will require documentation of the proposed use of the allocated value (=Specification required) and approval by the Designated Expert assigned by the IESG (see [RFC5226]).

5.1. Guidance for Designated Experts

In all cases of review by Designated Expert (DE) described here, the DE is expected to ascertain the existence of suitable documentation (a specification) as described in [RFC5226], and to verify the permanent and publically ready availability of the document. The DE is also expected to check the clarity of purpose and use of the requested code points. Lastly, the DE must verify that any specification produced in the IETF that requests one of these code points has been made available for review by the IDR working group, and that any specification produced outside the IETF does not conflict with work that is active or already published within the IETF.

6. Manageability Considerations

This section is structured as recommended in [RFC5706].

6.1. Operational Considerations

6.1.1. 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 purely carries application level data that has no immediate corresponding forwarding state impact. As

such, any churn in reachability information has different impact than regular BGP updates which need to change forwarding state for an entire router. Furthermore it is anticipated that distribution of this NLRI will be handled by dedicated route-reflectors providing a level of isolation and fault-containment between different NLRI types.

6.1.2. Installation and Initial Setup

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

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

6.1.3. Migration Path

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

6.1.4. Requirements on Other Protocols and Functional Components

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

6.1.5. Impact on Network Operation

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

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

6.1.6. Verifying Correct Operation

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

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

6.2. Management Considerations

6.2.1. Management Information

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.

6.2.2. Fault Management

If an implementation of BGP-LS detects a malformed attribute, then it MUST use the 'Attribute Discard' action as per [RFC7606] Section 2.

An implementation of BGP-LS MUST perform the following syntactic checks for determining if a message is malformed.

- Does the sum of all TLVs found in the BGP LS attribute correspond to the BGP LS path attribute length?
- Does the sum of all TLVs found in the BGP MP_REACH_NLRI attribute correspond to the BGP MP_REACH_NLRI length?
- Does the sum of all TLVs found in the BGP MP_UNREACH_NLRI attribute correspond to the BGP MP_UNREACH_NLRI length?
- Does the sum of all TLVs found in a Node-, Link or Prefix Descriptor NLRI attribute correspond to the Node-, Link- or Prefix Descriptors 'Total NLRI Length' field?
- Does any fixed length TLV correspond to the TLV Length field in this document?

6.2.3. Configuration Management

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

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

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

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

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

An implementation SHOULD allow the operator to configure a pair of ASN and BGP-LS identifier (Section 3.2.1.4) per flooding set in which the node participates.

6.2.4. Accounting Management

Not Applicable.

6.2.5. Performance Management

An implementation SHOULD provide the following statistics:

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

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

6.2.6. Security Management

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

- Drop all updates from Consumer peers

An implementation MUST have means to limit inbound updates.

7. TLV/Sub-TLV Code Points Summary

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

TLV Code Point	Description	IS-IS TLV/ Sub-TLV	Value defined in:

256	Local Node Descriptors	---	Section 3.2.1.2
257	Remote Node Descriptors	---	Section 3.2.1.3
258	Link Local/Remote Identifiers	22/4	[RFC5307]/1.1
259	IPv4 interface address	22/6	[RFC5305]/3.2
260	IPv4 neighbor address	22/8	[RFC5305]/3.3
261	IPv6 interface address	22/12	[RFC6119]/4.2
262	IPv6 neighbor address	22/13	[RFC6119]/4.3
263	Multi-Topology ID	---	Section 3.2.1.5
264	OSPF Route Type	---	Section 3.2.3
265	IP Reachability Information	---	Section 3.2.3
512	Autonomous System	---	Section 3.2.1.4
513	BGP-LS Identifier	---	Section 3.2.1.4
514	OSPF Area ID	---	Section 3.2.1.4
515	IGP Router-ID	---	Section 3.2.1.4
1024	Node Flag Bits	---	Section 3.3.1.1
1025	Opaque Node Properties	---	Section 3.3.1.5
1026	Node Name	variable	Section 3.3.1.3
1027	IS-IS Area Identifier	variable	Section 3.3.1.2
1028	IPv4 Router-ID of Local Node	134/---	[RFC5305]/4.3
1029	IPv6 Router-ID of Local Node	140/---	[RFC6119]/4.1
1030	IPv4 Router-ID of Remote Node	134/---	[RFC5305]/4.3
1031	IPv6 Router-ID of Remote Node	140/---	[RFC6119]/4.1
1088	Administrative group (color)	22/3	[RFC5305]/3.1
1089	Maximum link bandwidth	22/9	[RFC5305]/3.3
1090	Max. reservable link bandwidth	22/10	[RFC5305]/3.5
1091	Unreserved bandwidth	22/11	[RFC5305]/3.6
1092	TE Default Metric	22/18	Section 3.3.2.3
1093	Link Protection Type	22/20	[RFC5307]/1.2

1094	MPLS Protocol Mask	---	Section 3.3.2.2
1095	IGP Metric	---	Section 3.3.2.4
1096	Shared Risk Link Group	---	Section 3.3.2.5
1097	Opaque link attribute	---	Section 3.3.2.6
1098	Link Name attribute	---	Section 3.3.2.7
1152	IGP Flags	---	Section 3.3.3.1
1153	Route Tag	---	[RFC5130]
1154	Extended Tag	---	[RFC5130]
1155	Prefix Metric	---	[RFC5305]
1156	OSPF Forwarding Address	---	[RFC2328]
1157	Opaque Prefix Attribute	---	Section 3.3.3.6

Table 13: Summary Table of TLV/Sub-TLV code points

8. Security Considerations

Procedures and protocol extensions defined in this document do not affect the BGP security model. See the 'Security Considerations' section of [RFC4271] for a discussion of BGP security. Also refer to [RFC4272] and [RFC6952] for analysis of security issues for BGP.

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

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

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

9. Contributors

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

10. Acknowledgements

We would like to thank Nischal Sheth, Alia Atlas, David Ward, Derek Yeung, Murtuza Lightwala, John Scudder, Kaliraj Vairavakkalai, Les Ginsberg, Liem Nguyen, Manish Bhardwaj, 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.

11. References

11.1. Normative References

[RFC1195] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, DOI 10.17487/RFC1195, December 1990, <http://www.rfc-editor.org/info/rfc1195>.

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

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <http://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, <http://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, <http://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, <http://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, <http://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, <http://www.rfc-editor.org/info/rfc4271>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <http://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, <http://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, <http://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, <http://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, <http://www.rfc-editor.org/info/rfc5130>.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, DOI 10.17487/RFC5226, May 2008, <http://www.rfc-editor.org/info/rfc5226>.

[RFC5301] McPherson, D. and N. Shen, "Dynamic Hostname Exchange Mechanism for IS-IS", RFC 5301, DOI 10.17487/RFC5301, October 2008, <http://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, <http://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, <http://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, <http://www.rfc-editor.org/info/rfc5340>.

[RFC5890] Klensin, J., "Internationalized Domain Names for Applications (IDNA): Definitions and Document Framework", RFC 5890, DOI 10.17487/RFC5890, August 2010, <http://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, <http://www.rfc-editor.org/info/rfc6119>.

[RFC6286] Chen, E. and J. Yuan, "Autonomous-System-Wide Unique BGP Identifier for BGP-4", RFC 6286, DOI 10.17487/RFC6286, June 2011, <http://www.rfc-editor.org/info/rfc6286>.

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

[RFC6822] Previdi, S., Ed., Ginsberg, L., Shand, M., Roy, A., and D. Ward, "IS-IS Multi-Instance", RFC 6822, DOI 10.17487/RFC6822, December 2012, <http://www.rfc-editor.org/info/rfc6822>.

[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, <http://www.rfc-editor.org/info/rfc7606>.

11.2. Informative References

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

[RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <http://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, <http://www.rfc-editor.org/info/rfc4364>.

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

[RFC4970] Lindem, A., Ed., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 4970, DOI 10.17487/RFC4970, July 2007, <http://www.rfc-editor.org/info/rfc4970>.

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

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

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

[RFC5392] Chen, M., Zhang, R., and X. Duan, "OSPF Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5392, DOI 10.17487/RFC5392, January 2009, <http://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, <http://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, <http://www.rfc-editor.org/info/rfc5706>.

[RFC6952] Jethanandani, M., Patel, K., and L. Zheng, "Analysis of BGP, LDP, PCEP, and MSDP Issues According to the Keying and Authentication for Routing Protocols (KARP) Design Guide", RFC 6952, DOI 10.17487/RFC6952, May 2013, <http://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, <http://www.rfc-editor.org/info/rfc7285>.

Authors' Addresses

Hannes Gredler (editor)
Private Contributor

Email: hannes@gredler.at

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

Email: jmedved@cisco.com

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

Email: sprevidi@cisco.com

Adrian Farrel
Juniper Networks, Inc.

Email: adrian@olddog.co.uk

Saikat Ray

Email: raysaikat@gmail.com