              Revised Error Handling for BGP Updates from External Neighbors
                     draft-chen-ebgp-error-handling-00.txt

Status of this Memo

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

    Internet-Drafts are working documents of the Internet Engineering
    Task Force (IETF), its areas, and its working groups. Note that
    other groups may also distribute working documents as Internet-
    Drafts.

    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."

    The list of current Internet-Drafts can be accessed at
    http://www.ietf.org/1id-abstracts.html

    The list of Internet-Draft Shadow Directories can be accessed at
    http://www.ietf.org/shadow.html

    This Internet-Draft will expire on March 9, 2011.

Abstract

   In this document we partially revise the error handling of an UPDATE
   message from an external BGP neighbor. The essence of the revision
   is to avoid resetting an external BGP session by using the "treat-
   as-withdraw" approach when the whole NLRI field of a malformed UPDATE
   message can be parsed.

1. Introduction

   The base BGP specification [RFC4271] requires that a BGP session be
   reset when an UPDATE message containing a malformed attribute is
   received. This behavior is undesirable in the case of optional
   transitive attributes as has been discussed and revised in [OPT-
   TRANS].

   However, there are other situations where the behavior is also
   undesirable, but are outside the scope of [OPT-TRANS]. For example,
   there have been a few occurrences in the field where the AS-PATH
   attribute is malformed for a small number of routes. Resetting the
   BGP session would impact all the other valid routes in these cases.

   Our goal is to minimize the scope of the network that is affected by
   a malformed UPDATE message, and also to limit the impact to only the
   routes involved. The constrain is that the protocol correctness must
   not be violated.

   In this document we partially revise the error handling of an UPDATE
   message from an external BGP neighbor. The essence of the revision
   is to avoid resetting an external BGP session by using the "treat-
   as-withdraw" approach specified in [OPT-TRANS] when the whole NLRI
   field of a malformed UPDATE message can be parsed.

1.1. Specification of Requirements

   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 [RFC2119].

2. Revision to Base Specification

   The revised error handling specified in this section is applicable
   only for processing an UPDATE message from an external BGP neighbor.

   The error handling of the following case described in Section 6.3 of
   [RFC4271] remains unchanged:

       If the Withdrawn Routes Length or Total Attribute Length
       is too large (i.e., if Withdrawn Routes Length + Total Attribute
       Length + 23 exceeds the message Length), then the Error Subcode
       MUST be set to Malformed Attribute List.

   The error handling of all other cases described in Section 6.3 of
   [RFC4271] that specify a sesion reset is conditionally revised as
   follows.

   If a path attribute in an UPDATE message from an external BGP
   neighbor is determined to be malformed, the message containing that
   attribute SHOULD be treated as though all contained routes had been
   withdrawn ("treat-as-withdraw") when the whole NLRI field in the
   message can be parsed.

   One exception is that the "attribute discard" approach [OPT-TRANS]
   SHOULD be used to handle a malformed optional transitive attribute
   for which the "attribute discard" approach is specified.

   A BGP speaker MUST provide debugging facilities to permit issues
   caused by malformed UPDATE messages to be diagnosed. At a minimum,
   such facilities SHOULD include logging an error when such an
   attribute is detected. The malformed UPDATE message SHOULD be
   analyzed, and the root cause SHOULD be investigated.

3. Parsing of NLRI Fields

   As described in [OPT-TRANS], we observe that in order to use the
   "treat-as-withdraw" approach for a malformed UPDATE, the NLRI field
   and/or MP_REACH and MP_UNREACH [RFC4760] attributes need to be
   successfully parsed. If this were not possible, the UPDATE would
   necessarily be malformed in some other way beyond the scope of this
   document and therefore, the procedures of [RFC4271] would continue to
   apply.

   To facilitate the determination of the NLRI field in an UPDATE with
   malformed attributes, we strongly RECOMMEND that the MP_REACH or
   MP_UNREACH attribute (if present) be encoded as the very first path
   attribute in an UPDATE.

Traditionally the NLRIs for the IPv4 unicast address family are carried immediately following all the attributes in an UPDATE [RFC4271]. When such an UPDATE is received, we observe that the NLRI field can be determined using the "Message Length" and the "Total Attribute Length" (when they are consistent) carried in the message instead of relying on the length of individual attributes in the message.

Furthermore, it is observed that the NLRIs for the IPv4 unicast address family can also be carried in the MP_REACH attribute of an UPDATE when the IPV4 unicast address family capability is shared (i.e., both advertised and received) over a BGP session. For the same sake of better debugging and fault handling, we also RECOMMEND that the MP_REACH attribute be used and be placed as the very first path attribute in an UPDATE in this case.

4. Discussion

As discussed in [OPT-TRANS], the approach of "treat-as-withdraw" is not always safe to use. In the case of internal BGP sessions, the resolution of recursive nexthops can result in forwarding loops and blakholes when the BGP speakers inside a network have inconsistent routing information.

Depending on the network topology, the routing table, routes involved, and whether "tunnels" are used inside a network, the approach of "treat-as-withdraw" may work for internal BGP sessions only in some specific cases. Thus it may be deployed for internal BGP sessions only as a temporary measure to stop continuous session flaps due to malformed UPDATE messages. Such deployment must be carefully evaluated on a case-by-case basis.

5. IANA Considerations

This document makes no request of IANA.

6. Security Considerations

    TBD

7. Acknowledgments

    We would like to thank Robert Raszuk, Naiming Shen and Tony Li for
    their review and discussions.

8. References

8.1. Normative References

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

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

8.2. Informative References

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

    [OPT-TRANS] Scudder, J. and E. Chen, "Error Handling for Optional
                 Transitive BGP Attributes", Work in Progress, March 2010.

9. Authors' Addresses

    Enke Chen
    Cisco Systems, Inc.
    170 W. Tasman Dr.
    San Jose, CA 95134

    EMail: enkechen@cisco.com

    Pradosh Mohapatra
    Cisco Systems, Inc.
    170 W. Tasman Dr.
    San Jose, CA 95134

EMail: pmohapat@cisco.com

Keyur Patel
Cisco Systems, Inc.
170 W. Tasman Dr.
San Jose, CA 95134

EMail: keyupate@cisco.com

                 Advertising Link-State Information in BGP
                        draft-gredler-bgp-te-01

Abstract

   This document defines a new Border Gateway Protocol Network Layer
   Reachability Information (BGP NLRI) encoding format that can be used
   to distribute a network topologies' link and node information. Links
   can be either physical links connecting physical nodes, or virtual
   paths between physical or abstract nodes. The network topology
   information is carried via the BGP, thereby reusing protocol
   algorithms, operational experience, and administrative processes,
   such as inter-provider peering agreements.

   The BGP protocol carrying Link State information would provide a
   well-defined, uniform, policy-controlled interface from the network
   to outside servers that need to learn the network topology in real-
   time, for example an ALTO Server or a Path Computation Server.
   Having Traffic Engineering (TE) information from remote areas and/or
   Autonomous Systems would allow path computation for inter-area and/or
   inter-AS source-routed unicast and multicast tunnels.

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 January 12, 2012.

Copyright Notice

Table of Contents

1. Introduction

   Today, the contents of a link-state database usually has the scope of
   an IGP area. There are several use cases that could benefit from
   knowing the topology in a remote area or Autonomous System, but today
   no mechanism exists to distribute this information beyond an IGP
   area. This draft proposes to use BGP as the distribution mechanism
   for exchanging link-state data between routers in different IGP areas
   and/or Autonomous Systems. The mechanism can also be used to
   exchange topology and TE data between the network and external
   network-aware applications, such as the Alto Servers.

   The Border Gateway Protocol (BGP [RFC4271]) has grown beyond its
   original intention of disseminating IPv4 Inter-domain routing paths.
   A modern BGP implementation can be viewed as a ubiquitous database
   replication mechanism, which allows replication of many different
   state information types across arbitrary distribution graphs. Its
   built-in loop protection mechanism (AS path, Cluster List attributes)
   enables building of stable and redundant distribution topologies. In
   addition to IP routing, applications that use BGP for state
   distribution are L2VPN, VPLS, MAC-VPN, Route-target information, and
   Flowspec for firewalling. Using BGP as a dissemination protocol for
   topology data is a logical consequence.

   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 are: local/remote IP addresses, local/
   remote interface indices, 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 one of the link-state databases and distribute it to
   peer BGP Speakers using the encoding specified in this draft.

   A BGP Speaker may distribute the real physical topology from the Link
   State database or the Traffic Engineering database, or 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.

   Consumers of the network topology and TE data are peer routers in
   other areas either in the router's own AS or in remote ASes, or
   entities outside the network that may need network and/or TE data to
   optimize their behavior.

2. Scope

   The scope of Link State NLRI are the static attributes / metrics of a
   path between two routers. The path can be a physical link or
   multiple links aggregated into a path. Dynamic data, such as
   reservable bandwidth or delay metrics, is out of scope of this draft.

3. Transcoding Link State Information into a BGP NLRI

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

   All link and node information shall be encoded using a TBD AFI / SAFI
   1 or SAFI 128 header into those attributes. SAFI 1 shall be used for
   Internet routing (Public) and SAFI 128 shall be used for VPN routing
   (Private) applications.

   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 (multiprotocol BGP), with an
   AFI of TBD and an SAFI of 1 or 128.

3.1. NLRI format

```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |          NLRI Type            |      Total NLRI Length         |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                                                               |
   |                   Link-State NLRI (variable)                  |
   |                                                               |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

                  Figure 1: Link State SAFI 1 NLRI Format

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           NLRI Type           |        Total NLRI Length      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
+                      Route Distinguisher                      +
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
|                  Link-State NLRI (variable)                   |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
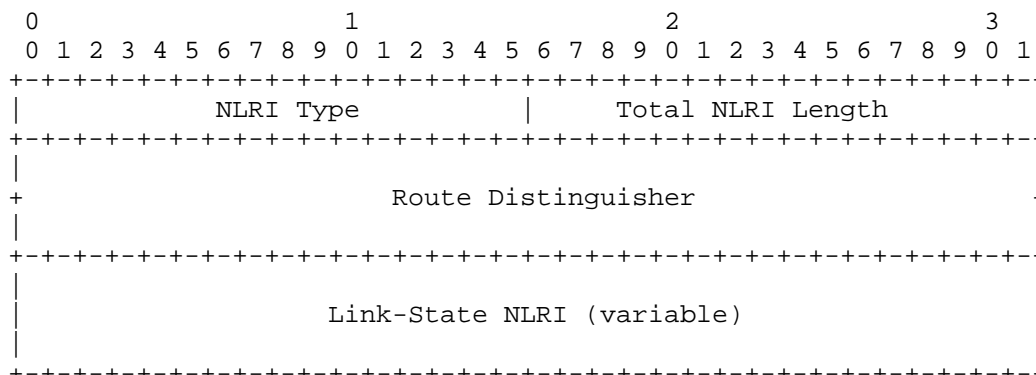
                  Figure 2: Link State SAFI 128 NLRI Format

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

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

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

   Type = 2: Node NLRI, contains node attributes

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

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Local Node Descriptors (variable)                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Remote Node Descriptors (variable)               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                 Link Descriptors (variable)                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  Link Attributes (variable)                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

                     Figure 3: The Link NLRI format

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

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               Local Node Descriptors (variable)               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  Node Attributes (variable)                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
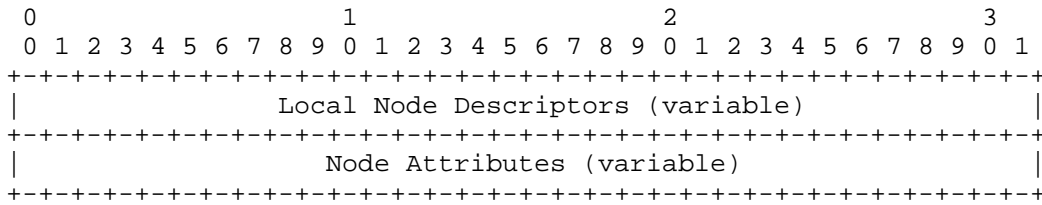
Figure 4: The Node NLRI format

3.2. TLV Format

   The Node Descriptors, Link Descriptors, Link Attribute, and Node
   Attribute fields are described using a set of Type/Length/Value
   triplets. The format of each TLV is shown in Figure 5.

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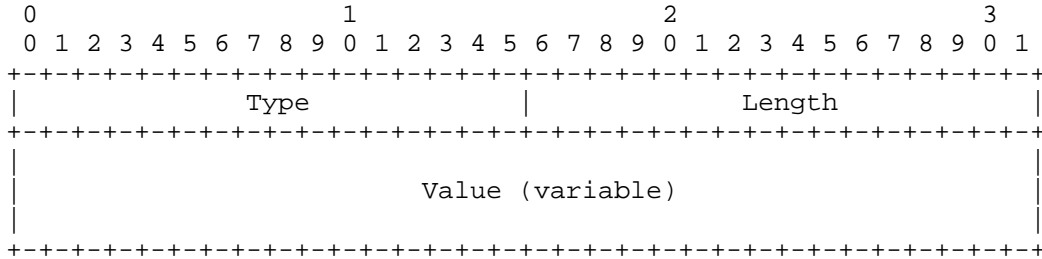
Figure 5: TLV format

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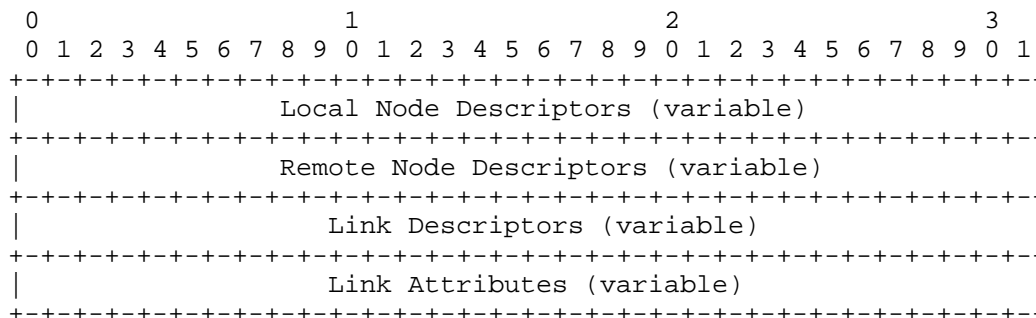   (thus a TLV with no value portion would have a length of zero). The
   TLV is not padded to four-octet alignment; Unrecognized types are
   ignored.

3.3. Node Descriptors

   Each link gets anchored by at least a pair of router-IDs. Since
   there are many Router-IDs formats (32 Bit IPv4 router-ID, 56 Bit ISO
   Node-ID and 128 Bit IPv6 router-ID) a link may be anchored by more
   than one Router-ID pair. The set of Local and Remote Node
   Descriptors describe which Protocols Router-IDs will be following to
   "anchor" the link described by the "Link attribute TLVs". There must
   be at least one "like" router-ID pair of a Local Node Descriptors and
   a Remote Node Descriptors per-protocol. If a peer sends an illegal
   combination in this respect, then this is handled as an NLRI error,
   described in [RFC4760].

   It is desirable that the Router-ID assignments inside the Node anchor
   are globally unique. However there may be router-ID spaces (e.g.

ISO) where not even a global registry exists, or worse, Router-IDs
have been allocated following private-IP RFC 1918 [RFC1918]
allocation. In order to disambiguate the Router-IDs the local and
remote Autonomous System number TLVs of the anchor nodes may be
included in the NLRI. The Local and Remote Autonomous System TLVs
are 4 octets wide as described in [RFC4893]. 2-octet AS Numbers shall
be expanded to 4-octet AS Numbers by zeroing the two MSB octets.

3.3.1. Local Node Descriptors

The Local Node Descriptors TLV (Type 256) contains Node Descriptors
for the node anchoring the local end of the link. The length of this
TLV is variable. The value contains one or more Node Descriptor Sub-
TLVs defined in Section 3.3.3.

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

Figure 6: Local Node Descriptors TLV format

3.3.2. Remote Node Descriptors

The Remote Node Descriptors TLV (Type 257) contains Node Descriptors
for the node anchoring the remote end of the link. The length of
this TLV is variable. The value contains one or more Node Descriptor
Sub-TLVs defined in Section 3.3.3.

```
     0                   1                   2                   3
     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |              Type             |             Length            |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                                                               |
    |                Node Descriptor Sub-TLVs (variable)            |
    |                                                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
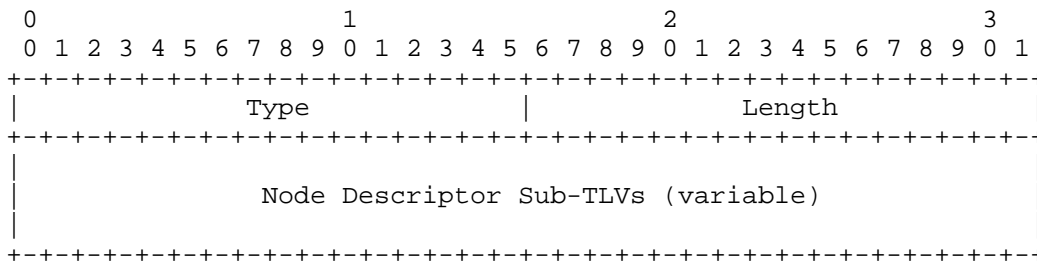
Figure 7: Remote Node Descriptors TLV format

3.3.3. Node Descriptor Sub-TLVs

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

```
          +------+------------------+--------+
          | Type | Description      | Length |
          +------+------------------+--------+
          |  258 | Autonomous System |     4 |
          |  259 | IPv4 Router-ID   |      4 |
          |  260 | IPv6 Router-ID   |     16 |
          |  261 | ISO Node-ID      |      7 |
          +------+------------------+--------+
```

                  Table 1: Node Descriptor Sub-TLVs

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

   Autonomous System: opaque value (32 Bit AS ID)

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

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

   ISO Node ID: ISO node-ID (6 octets ISO system-ID plus PSN octet)

3.3.4. Router-ID Anchoring Example: ISO Pseudonode

   IS-IS Pseudonodes are a good example for the variable Router-ID
   anchoring. Consider Figure 8. This represents a Broadcast LAN
   between a pair of routers. The "real" (=non pseudonode) routers have
   both an IPv4 Router-ID and IS-IS Node-ID. The pseudonode does not
   have an IPv4 Router-ID. Two unidirectional links (Node1, Pseudonode
   1) and (Pseudonode 1, Node 2) are being generated.

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

   The NLRI for (Pseudonode1, Node2) encodes a local ISO node-ID, remote
   IPv4 router-ID and remote ISO node-id.

```
      +----------------+   +----------------+   +----------------+
      |     Node1      |   |  Pseudonode 1  |   |     Node2      |
      |1921.6800.1001.00|-->|1921.6800.1001.02|-->|1921.6800.1002.00|
      |   192.168.1.1  |   |                |   |   192.168.1.2  |
      +----------------+   +----------------+   +----------------+
```
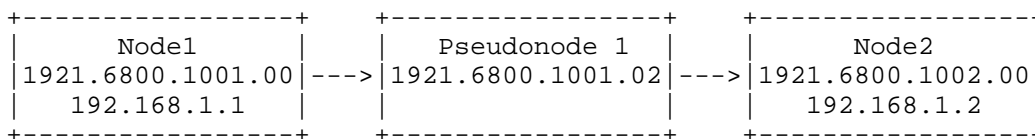
                    Figure 8: IS-IS Pseudonodes

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

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

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

3.4. Link Descriptors

   The 'Link Descriptor' field is a set of Type/Length/Value (TLV)
   triplets. The format of each TLV is shown in Figure 5. The 'Link
   descriptor' TLVs uniquely identify a link between a pair of anchor
   Routers.

   The encoding of 'Link Descriptor' TLVs, i.e. the Codepoints in
   'Type', and the 'Length' and 'Value' fields are the same as defined
   in [RFC5305], [RFC5307], and [RFC6119] for sub-TLVs in the Extended
   IS reachability TLV. The Codepoints are in the IANA Protocol
   Registry for IS-IS, sub-TLV Codepoints for TLV 22, [IANA-ISIS].
   Although the encodings for 'Link Descriptor' TLVs were originally
   defined for IS-IS, the TLVs can carry data sourced either by IS-IS or
   OSPF.

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

```
+------+-----------------------------+-----------------------+
| Type | Description                 | Defined in:           |
+------+-----------------------------+-----------------------+
|    4 | Link Local/Remote Identifiers | [RFC5307], Section 1.1 |
|    6 | IPv4 interface address      | [RFC5305], Section 3.2 |
|    8 | IPv4 neighbor address       | [RFC5305], Section 3.3 |
|   12 | IPv6 interface address      | [RFC6119], Section 4.2 |
|   13 | IPv6 neighbor address       | [RFC6119], Section 4.3 |
+------+-----------------------------+-----------------------+
```

                     Table 2: Link Descriptor TLVs

3.5. Link Attributes

   The 'Link Attributes' field is a set of Type/Length/Value (TLV)
   triplets. The format of each TLV is shown in Figure 5.

   For Codepoints < 255, the encoding of 'Link Attributes' TLVs, i.e.
   the Codepoints in 'Type', and the 'Length' and 'Value' fields are the
   same as defined in [RFC5305], [RFC5307], and [RFC6119] for sub-TLVs
   in the Extended IS reachability TLV. The Codepoints are in the IANA
   Protocol Registry for IS-IS, sub-TLV Codepoints for TLV 22,
   [IANA-ISIS]. Although the encodings for 'Link Attributes' TLVs were
   originally defined for IS-IS, the TLVs can carry data sourced either
   by IS-IS or OSPF.

   For Codepoints > 255, the encoding of 'Link Attributes' TLVs is
   described in subsequent sections.

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

```
+-------+-------------------------------+-----------------------+
|  Type | Description                   | Defined in:           |
+-------+-------------------------------+-----------------------+
|     3 | Administrative group (color)  | [RFC5305], Section 3.1 |
|     9 | Maximum link bandwidth        | [RFC5305], Section 3.3 |
|    10 | Max. reservable link bandwidth | [RFC5305], Section 3.5 |
|    11 | Unreserved bandwidth          | [RFC5305], Section 3.6 |
|    20 | Link Protection Type          | [RFC5307], Section 1.2 |
| 64509 | MPLS Protocol                 | Section 3.5.1          |
| 64510 | TE Default Metric             | Section 3.5.2          |
| 64511 | IGP Link Metric               | Section 3.5.3          |
| 64512 | Shared Risk Link Group        | Section 3.5.4          |
| 64513 | OSPF specific link attribute  | Section 3.5.5          |
| 64514 | IS-IS specific link attribute | Section 3.5.6          |
+-------+-------------------------------+-----------------------+
```

                     Table 3: Link Attribute TLVs

3.5.1. MPLS Protocol TLV

   The MPLS Protocol TLV (Type 64511) carries a bit mask describing
   which MPLS signaling protocols are enabled. The length of this TLV
   is 1. The value is a bit array of 8 flags, where each bit represents
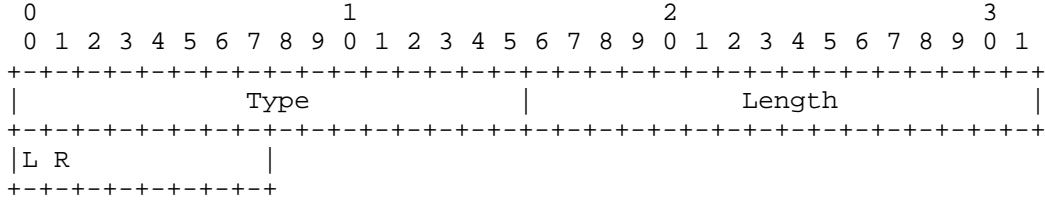   an MPLS Protocol capability.

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

                       Figure 9: MPLS Protocol TLV

   The following bits are defined:

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

                     Table 4: MPLS Protocol TLV Codes

3.5.2. TE Default Metric TLV

   The TE Default Metric TLV (Type 64512) carries the TE Default metric
   for this link. This TLV corresponds to the IS-IS TE Default metric
   sub-TLV (Type 18), defined in RFC5305, Section 3.7 [RFC5305], and the
   OSPF TE Metric sub-TLV (Type 5), defined in RFC3630, Section 2.5.5
   [RFC3630]. If the value in the TE Default metric TLV is derived from
   IS-IS TE Default Metric, then the upper 8 bits of this TLV are set to
   0.

```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |              Type             |             Length            |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                        TE Default Metric                      |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
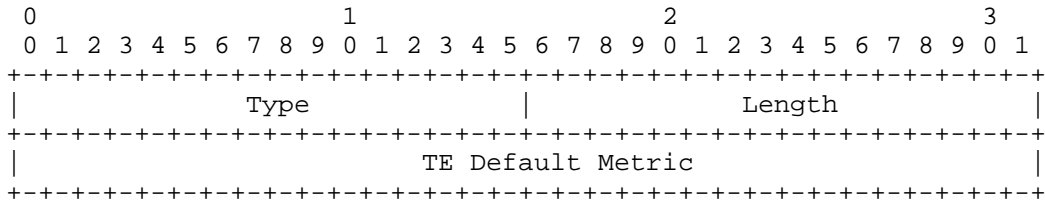
                  Figure 10: TE Default metric TLV format

3.5.3. IGP Link Metric TLV

   The IGP Metric TLV (Type 64513) carries the IGP metric for this link.
   This attribute is only present if the IGP link metric is different
   from the TE Default Metric (Type 18). The length of this TLV is 3.
   If the length of the IGP link metric from which the IGP Metric value
   is derived is less than 3 (e.g. for OSPF link metrics or non-wide
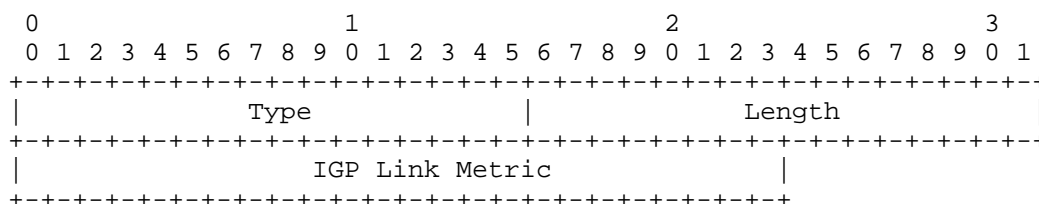   IS-IS metric), then the upper bits of the TLV are set to 0.

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

                   Figure 11: IGP Link Metric TLV format

3.5.4. Shared Risk Link Group TLV

   The Shared Risk Link Group (SRLG) TLV (Type 64514) 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 12. The length of this TLV is 4 *
   (number of SRLG values).

```
     0                   1                   2                   3
     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |              Type             |             Length            |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                  Shared Risk Link Group Value                 |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                          ...........                          |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                  Shared Risk Link Group Value                 |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
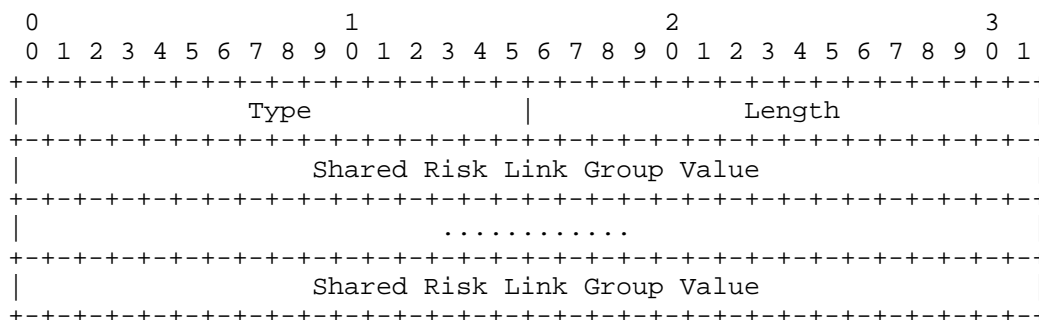
                Figure 12: Shared Risk Link Group TLV format

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

3.5.5. OSPF specific link attribute TLV

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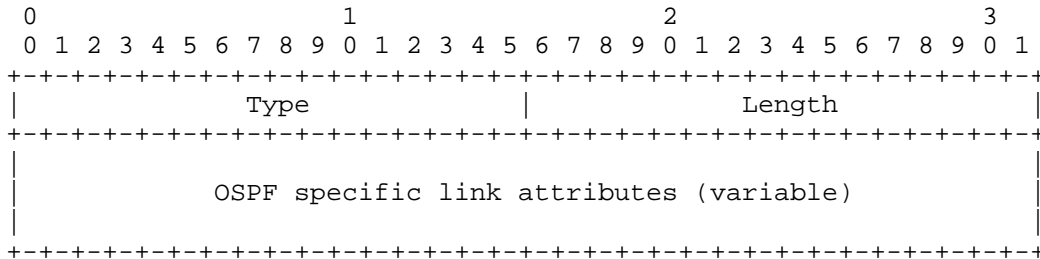   the BGP link-state NLRI.

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

                Figure 13: OSPF specific link attribute format

3.5.6. IS-IS specific link attribute TLV

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

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Type             |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
|          IS-IS specific link attributes (variable)            |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
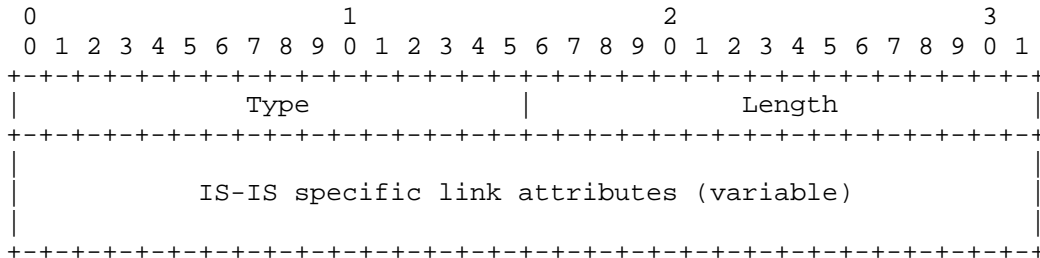
                Figure 14: IS-IS specific link attribute format

3.6. Node Attributes

   The following node attribute TLVs are valid in the Node NLRI:

```
            +-------+------------------------------+----------+
            | Type  | Description                  |  Length  |
            +-------+------------------------------+----------+
            | 65515 | Node Flag Bits               |        1 |
            | 65516 | OSPF Specific Node Properties | variable |
            | 65517 | IS-IS Specific Node Properties | variable |
            +-------+------------------------------+----------+
```

                      Table 5: Node Attribute TLVs

3.6.1. Node Flag Bits TLV

   The Node Flag Bits TLV (Type 1) carries a bit mask describing node
   attributes. The value is a bit array of 8 flags, where each bit
   represents an MPLS Protocol capability.

```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |              Type             |             Length            |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |    Flags      |
   +-+-+-+-+-+-+-+-+
```
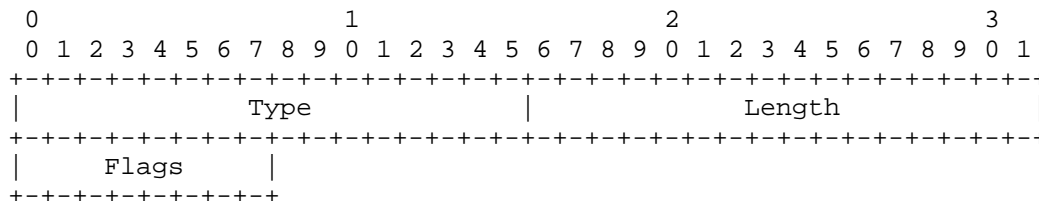
                  Figure 15: Node Flag Bits TLV format

   The bits are defined as follows:

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

                  Table 6: Node Flag Bits Definitions

3.6.2. OSPF Specific Node Properties TLV

   The OSPF Specific Node Properties TLV is an envelope that
   transparently carries optional node properties TLVs advertised by an
   OSPF router. The value field contains one or more optional OSPF node

property TLVs, such as the OSPF Router Informational Capabilities TLV
defined in [RFC4970], or the OSPF TE Node Capability Descriptor TLV
described in [RFC5073]. An originating router shall use this TLV for
encoding information specific to the OSPF protocol or new OSPF
extensions for which there is no protocol neutral representation in
the BGP link-state NLRI.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Type              |            Length             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
|            OSPF specific node properties (variable)           |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
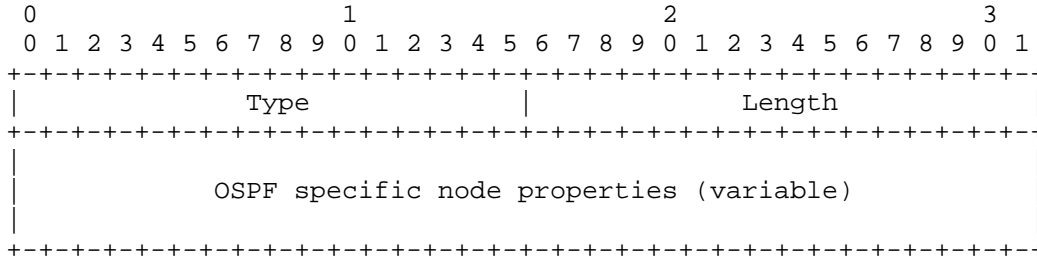
Figure 16: OSPF specific Node property format

3.6.3. IS-IS Specific Node Properties TLV

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

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Type              |            Length             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
|            IS-IS specific node properties (variable)          |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
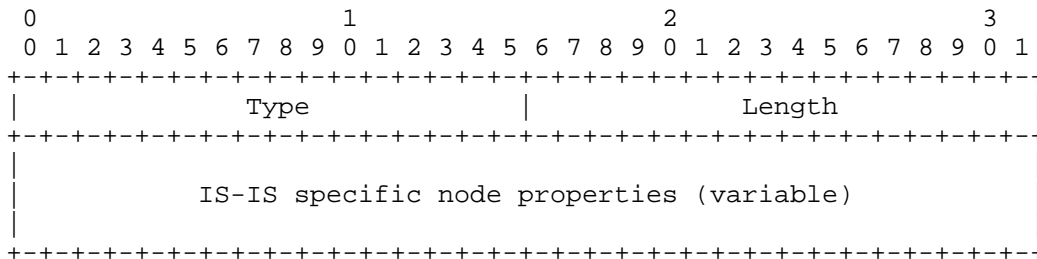
Figure 17: IS-IS specific Node property format

3.7. IGP Area Information

IGP Area information can be carried in BGP communities. An
implementation should support configuration that maps IGP areas to
BGP communities.

3.8. Inter-AS Links

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

4. Link to Path Aggregation

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

4.1. Example: No Link Aggregation

   Consider Figure 18. Both AS1 and AS2 operators want to protect their
   inter-AS {R1,R3}, {R2, R4} links using RSVP-FRR LSPs. If R1 wants to
   compute its link-protection LSP to R3 it needs to "see" an alternate
   path to R3. Therefore the AS2 operator exposes its topology. All
   BGP TE enabled routers in AS1 "see" the full topology of AS and
   therefore can compute a backup path. Note that the decision if the
   direct link between {R3, R4} or the {R4, R5, R3) path is used is made
   by the computing router.

```
        AS1   :   AS2
              :
         R1-------R3
          |   :   | \
          |   :   |  R5
          |   :   | /
         R2-------R4
              :
              :
```
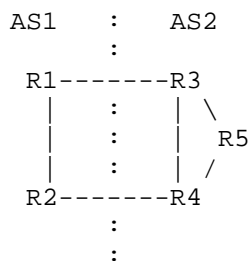
                     Figure 18: no-link-aggregation

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 19. The only link which gets advertised by AS2 is an
   "aggregate" link between R3 and R4. This is enough to tell AS1 that
   there is a backup path. However the actual links being used are
   hidden from the topology.

```
        AS1   :   AS2
              :
         R1-------R3
          |   :    |
          |   :    |
          |   :    |
         R2-------R4
              :
              :
```
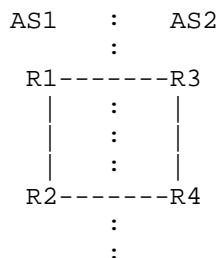
                    Figure 19: asbr-link-aggregation

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 20. Rather
   than exposing all specific R3 to R6 links, AS3 is modeled as a single
   node which connects to the border routers of the aggregated domain.

```
        AS1   :   AS2   :   AS3
              :         :
         R1-------R3-----
          |   :         : \
          |   :         :   vR0
          |   :         : /
         R2-------R4-----
              :         :
              :         :
```
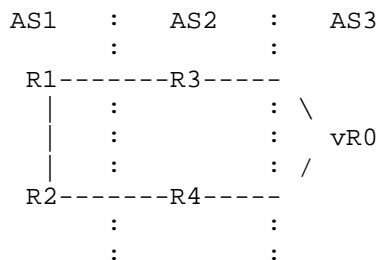
                    Figure 20: multi-as-aggregation

5. Originating the TED NLRI

   A BGP Speaker must be configured to originate TED NLRIs. Usually
   export of the TED database into BGP is enabled on ASBRs and ABRs.

   The BGP Speaker shall throttle the rate of TED NLRI updates. An
   implementation shall provide a configuration attribute for the

interval between updates. The minimum interval between updates is 30
seconds.

6. Receiving the TED NLRI

   This section describes the processing of TED NLRIs at the receiving
   BGP Speaker.

   TE attributes for a link received from an IGP have higher priority
   than TED NLRIs received via BGP. Multiple BGP Speakers may advertise
   the same TED NLRI; the receiving BGP Speaker can individually choose
   the source BGP Speaker for each NLRI.

   The AS_PATH attribute is used both for loop detection and for NLRI
   selection: the TED NLRI with shorter AS_PATH length is preferred.
   The Community and Extended Community path attributes are stored in
   the RIB and may be used in operator-defined policies. Communities
   can also be used to encode the IGP Area information. All other path
   attributes are ignored.

7. Use Cases

7.1. MPLS TE

   If a router wants to compute a MPLS TE path across IGP areas TED
   lacks visibility of the complete topology. This is an issue for
   large scale networks that need to segment their core networks into
   distinct areas because inter-area TE cannot get deployed there.
   Current solutions for inter area TE only compute the path for the
   first area. The router only has full topological visibility for the
   first area along the path, but not for subsequent areas. The best
   practice is to use a technique called "loose-hop-expansion" which
   uses the IGP computed shortest path topology for the remainder of the
   path. Therefore no non-SPF based path setup is possible across
   areas. This has disadvantages for path protection and path
   engineering applications, as shown in Figure 21.
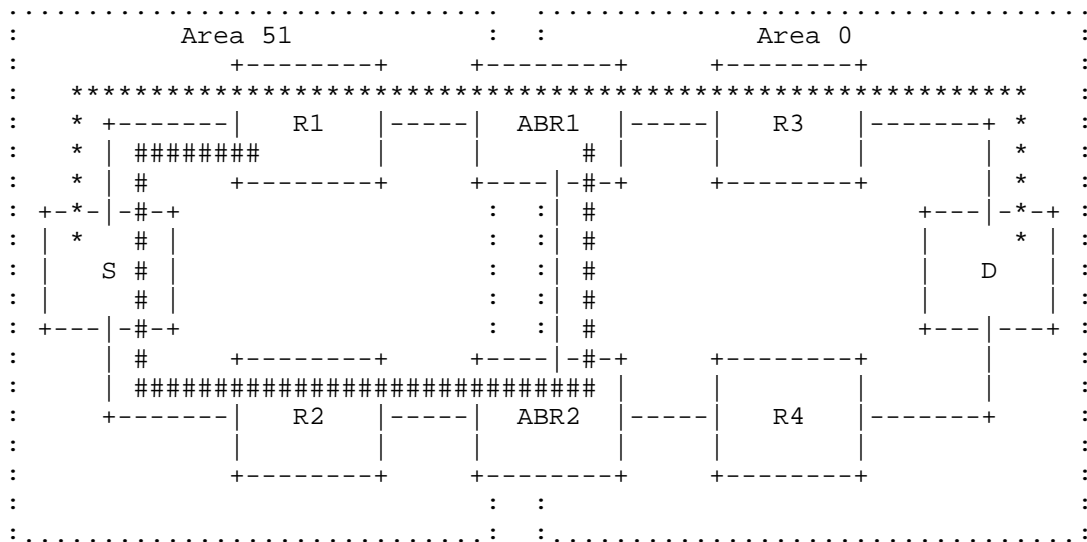
```
     ................................   .............................
     :           Area 51          :   :             Area 0         :
     :        +--------+      +--------+      +--------+           :
     :   *****************************************************************   :
     :   * +-------|   R1   |-----|  ABR1  |-----|   R3   |-------+ *   :
     :   * | ########     |      |      # |     |        |       | *   :
     :   * | #      +--------+      +----|-#-+     +--------+       | *   :
     : +-*-|-#-+                   :  :| #                  +---|-*-+ :
     : | *    # |                  :  :| #                  |    * | :
     : |   S  # |                  :  :| #                  |   D  | :
     : |      # |                  :  :| #                  |      | :
     : +---|-#-+                   :  :| #                  +---|---+ :
     :     | #     +--------+      +----|-#-+     +--------+       |   :
     :     | #############################  |     |        |       |   :
     :   +-------|   R2   |-----|  ABR2  |-----|   R4   |-------+   :
     :     |        |      |        |     |        |           :
     :     +--------+      +--------+     +--------+           :
     :                         :  :                           :
     :.........................:   :...........................:

                       ......
        ****   Primary LSP  :    : Area Boundary
        ####   Bypass LSP   :....:
```

                    Figure 21: MPLS TE Bypass LSP problem

   Router S sets up an RSVP LSP from S to D. Although it has only
   visibility into Area 51, the LSP setup ultimately succeeds, as
   shortest path first routing from ABR1 onwards routes the RSVP message
   towards destination D. What does not work is to setup a Link
   Protection bypass LSP protection for the R1 to ABR1 link as shown in
   the figure. The problem is that the TE database at Router R1 does
   not have path visibility of the link between ABR1 and ABR2, such that
   it can compute the Link Bypass LSP.

7.2. ALTO Server Network API

   An ALTO Server 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: the network map that specifies allocation of prefixes to PIDs,
   and the cost map that specifies the cost between the PIDs. For more
   details, see [I-D.ietf-alto-protocol].

   ALTO abstract network topologies can be auto-generated from the
   physical topology of the underlying network. The generation would
   typically be based on policies and rules set by the operator. Both

prefix and TE data are required: prefix data is required to generate
the network maps, TE (topology) data is required to generate the cost
maps. Prefix data is carried and originated in BGP, TE data is
originated and carried in an IGP. Without BGP TE NLRI the ALTO
Server would have to peer with both BGP Speakers and IGP in multiple
areas and/or ASes to obtain all the necessary network topology data.
The BGP TE NLRI allows for a single interface between the network and
the ALTO Server.

7.3. Path Computation Element (PCE) TED Synchronization Protocol

   RFC4655, Section 5.2, Figure 2 [RFC4655] describes a Path Computation
   Element (PCE) which synchronizes its traffic engineering database
   (TED) by use of a routing protocol. This memo describes the first
   standardized protocol for PCE to learn about inter-AS or inter-area
   TE information.

8. IANA Considerations

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

   This document requests creation of a new registry for node anchor,
   link descriptor and link attribute TLVs. The range of Codepoints in
   the registry is 0-65535. Values 0-255 will shadow Codepoints of the
   IANA Protocol Registry for IS-IS, sub-TLV Codepoints for TLV 22.
   Values 256-65535 will be used for Codepoints that are specific to the
   BGP TE NLRI. The registry will be initialized as shown in Table 2
   and Table 3. Allocations within the registry will require
   documentation of the proposed use of the allocated value and approval
   by the Designated Expert assigned by the IESG (see [RFC5226]).

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

9. Security Considerations

   This draft does not affect the BGP security model.

10. Acknowledgements

   We would like to thank Nischal Sheth from Juniper Networks for his
   input and contributions to this text. We would like to thank Alia
   Atlas, David Ward, John Scudder, Kaliraj Vairavakkalai, and Yakov
   Rekhter from Juniper Networks, Les Ginsberg and Mike Shand from Cisco

Systems, and Richard Woundy from Comcast for their comments.

11. References

11.1. Normative References

[IANA-ISIS]
          "IS-IS TLV Codepoint, Sub-TLVs for TLV 22", <http://
          www.iana.org/assignments/isis-tlv-codepoints/
          isis-tlv-codepoints.xml#isis-tlv-codepoints-3>.

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

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

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

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

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

[RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering
          (TE) Extensions to OSPF Version 2", RFC 3630,
          September 2003.

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

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

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

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

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

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

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

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

   [RFC6119] Harrison, J., Berger, J., and M. Bartlett, "IPv6 Traffic
              Engineering in IS-IS", RFC 6119, February 2011.

11.2. Informative References

   [I-D.ietf-alto-protocol]
              Alimi, R., Penno, R., and Y. Yang, "ALTO Protocol",
              draft-ietf-alto-protocol-08 (work in progress), May 2011.

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

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

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

Authors' Addresses

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

   Email: hannes@juniper.net

Jan Medved
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: jmedved@juniper.net

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

Email: sprevidi@cisco.com

Inter-Domain Routing   H. Gredler
Internet-Draft   J. Medved
Intended status: Standards Track   A. Farrel
Expires: March 24, 2012   Juniper Networks, Inc.
                                                            S. Previdi
                                                    Cisco Systems, Inc.
                                                    September 21, 2011

         North-Bound Distribution of Link-State and TE Information using BGP
                    draft-gredler-idr-ls-distribution-00

Abstract

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

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

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

Requirements Language

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

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 March 24, 2012.

Copyright Notice

Table of Contents

1. Introduction

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

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

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

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

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

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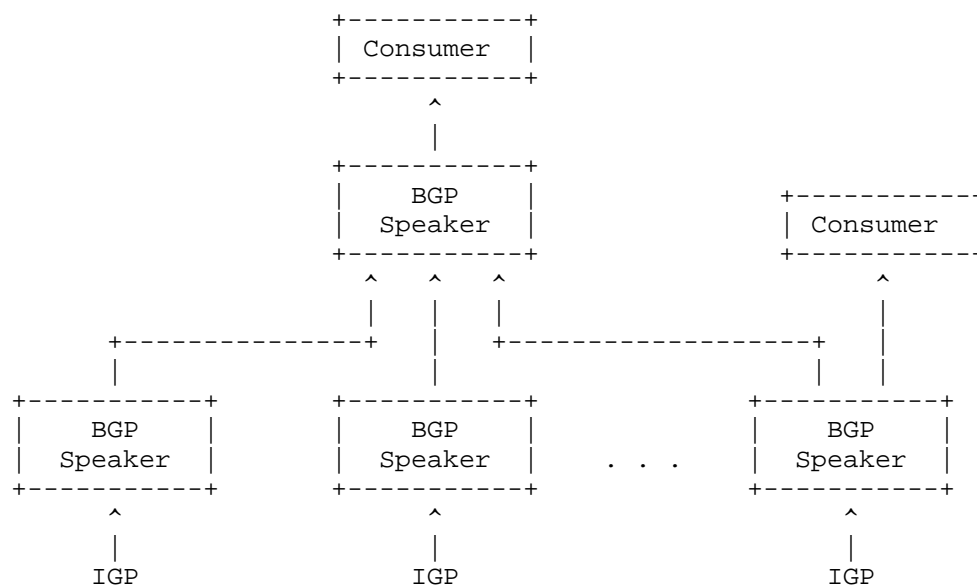   distribution to consumers is shown in the following figure.

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

              Figure 1: TE Link State info collection

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

2. Motivation and Applicability

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

2.1. MPLS-TE with PCE

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

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

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

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

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

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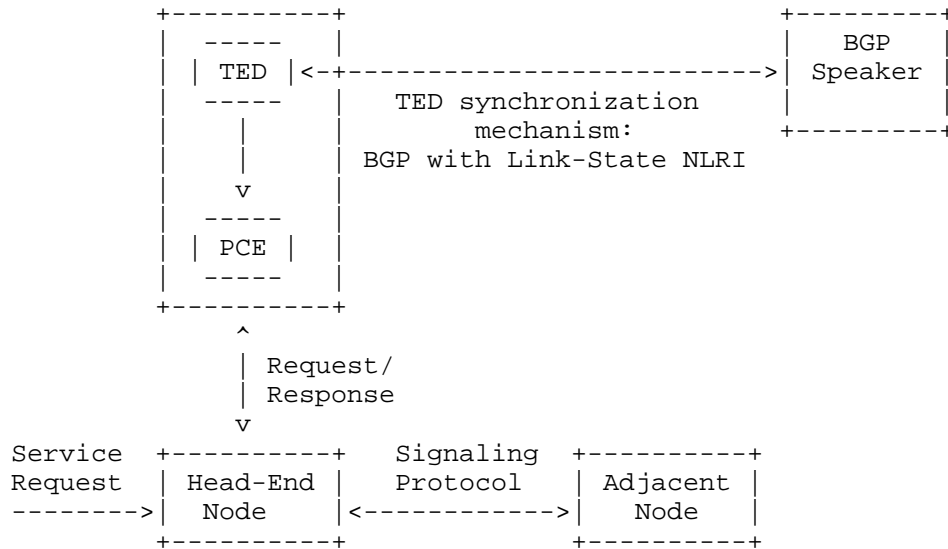   using the mechanism described in this document.

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

        Figure 2: External PCE node using a TED synchronization mechanism

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

2.2. ALTO Server Network API

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

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

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

```
  +--------+
  | Client |<--+
  +--------+   |
               |        ALTO      +--------+      BGP with    +---------+
  +--------+   |     Protocol  |  ALTO  | Link-State NLRI |  BGP    |
  | Client |<--+-----------|  Server |<----------------|  Speaker |
  +--------+   |           |        |                 |         |
               |           +--------+                 +---------+
  +--------+   |
  | Client |<--+
  +--------+
```
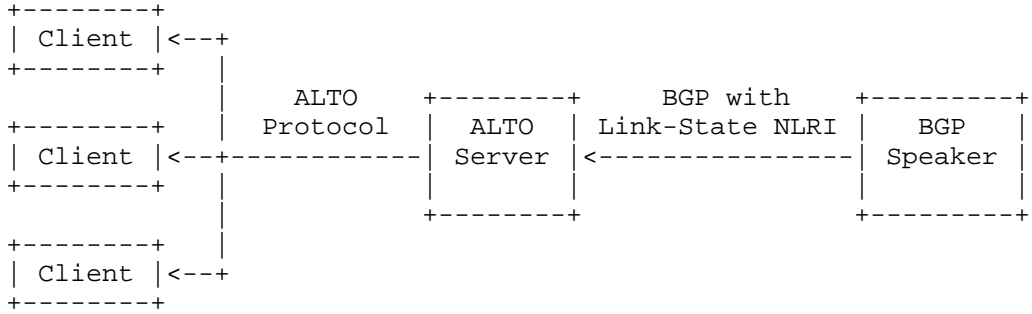
            Figure 3: ALTO Server using network topology information

3. Transcoding Link State Information into a BGP NLRI

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

   All link and node information SHALL be encoded using a TBD AFI / SAFI
   1 or SAFI 128 header into those attributes. SAFI 1 SHALL be used for
   Internet routing (Public) and SAFI 128 SHALL be used for VPN routing
   (Private) applications.

   In order for two BGP speakers to exchange Link-State NLRI, they MUST
   use BGP Capabilities Advertisement to ensure that they both are
   capable of properly processing such NLRI. This is done as specified
   in [RFC4760], by using capability code 1 (multi-protocol BGP), with
   an AFI of TBD and an SAFI of 1 or 128.

3.1. NLRI Format

```
     0                   1                   2                   3
     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |          NLRI Type         |       Total NLRI Length          |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                                                               |
    |                  Link-State NLRI (variable)                   |
    |                                                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4: Link State SAFI 1 NLRI Format


```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           NLRI Type           |        Total NLRI Length      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
+                      Route Distinguisher                      +
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
|                    Link-State NLRI (variable)                 |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
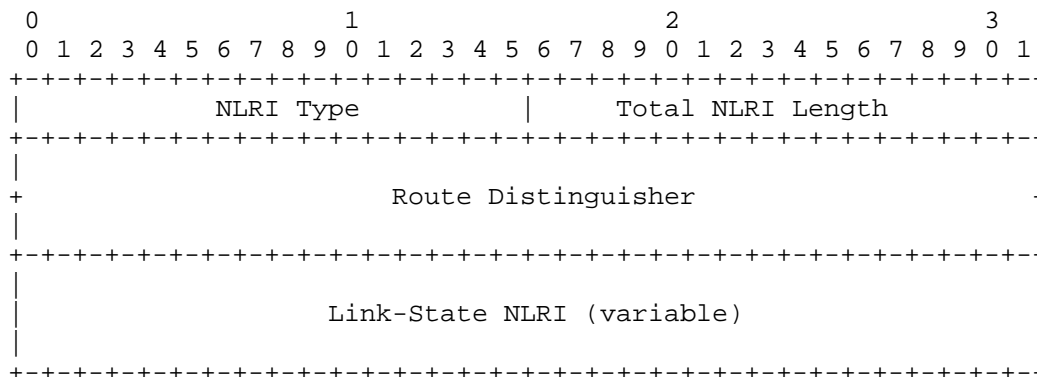
Figure 5: Link State SAFI 128 NLRI Format

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

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

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

   Type = 2: Node NLRI, contains node attributes

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

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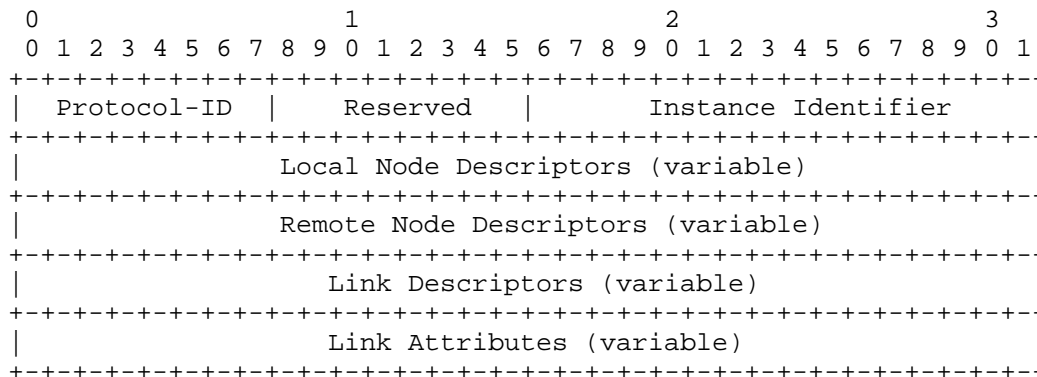
Figure 6: The Link NLRI format

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

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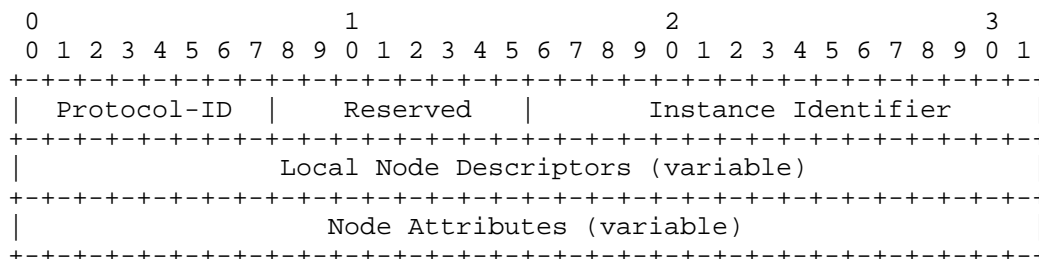
Figure 7: The Node NLRI format

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

   Type = 0: Unknown, The source of NLRI information could not be
   determined

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

   Type = 2: IS-IS Level 2, The NLRI information has been sourced by
   IS-IS Level 2

   Type = 3: OSPF, The NLRI information has been sourced by OSPF

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

3.2. TLV Format

The Node Descriptors, Link Descriptors, Link Attribute, and Node
Attribute fields are described using a set of Type/Length/Value
triplets. The format of each TLV is shown in Figure 8.

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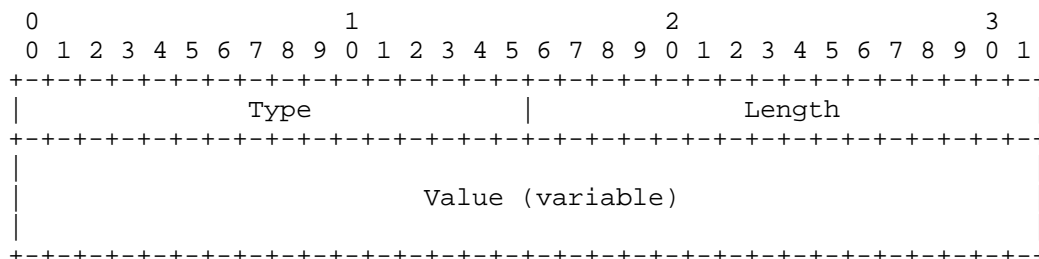
Figure 8: TLV format

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

3.3. Node Descriptors

Each link gets anchored by at least a pair of router-IDs. Since there are many Router-IDs formats (32 Bit IPv4 router-ID, 56 Bit ISO Node-ID and 128 Bit IPv6 router-ID) a link may be anchored by more than one Router-ID pair. The set of Local and Remote Node Descriptors describe which Protocols Router-IDs will be following to "anchor" the link described by the "Link attribute TLVs". There must be at least one "like" router-ID pair of a Local Node Descriptors and a Remote Node Descriptors per-protocol. If a peer sends an illegal combination in this respect, then this is handled as an NLRI error, described in [RFC4760].

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

3.3.1. Local Node Descriptors

The Local Node Descriptors TLV (Type 256) contains Node Descriptors for the node anchoring the local end of the link. The length of this TLV is variable. The value contains one or more Node Descriptor Sub-TLVs defined in Section 3.3.3.

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

Figure 9: Local Node Descriptors TLV format

3.3.2. Remote Node Descriptors

   The Remote Node Descriptors TLV (Type 257) contains Node Descriptors
   for the node anchoring the remote end of the link. The length of
   this TLV is variable. The value contains one or more Node Descriptor
   Sub-TLVs defined in Section 3.3.3.

```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |              Type             |             Length            |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                                                               |
   |             Node Descriptor Sub-TLVs (variable)               |
   |                                                               |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
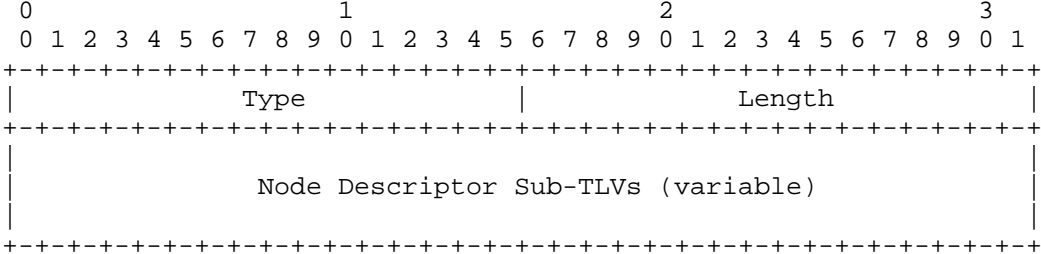
              Figure 10: Remote Node Descriptors TLV format

3.3.3. Node Descriptor Sub-TLVs

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

```
          +------+-------------------+--------+
          | Type | Description       | Length |
          +------+-------------------+--------+
          |  258 | Autonomous System |      4 |
          |  259 | IPv4 Router-ID    |      4 |
          |  260 | IPv6 Router-ID    |     16 |
          |  261 | ISO Node-ID       |      7 |
          +------+-------------------+--------+
```

                 Table 1: Node Descriptor Sub-TLVs

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

   Autonomous System: opaque value (32 Bit AS ID)

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

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

ISO Node ID: ISO node-ID (6 octets ISO system-ID plus PSN octet)

3.3.4. Router-ID Anchoring Example: ISO Pseudonode

IS-IS Pseudonodes are a good example for the variable Router-ID
anchoring. Consider Figure 11. This represents a Broadcast LAN
between a pair of routers. The "real" (=non pseudonode) routers have
both an IPv4 Router-ID and IS-IS Node-ID. The pseudonode does not
have an IPv4 Router-ID. Two unidirectional links (Node1, Pseudonode
1) and (Pseudonode 1, Node 2) are being generated.

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

The NLRI for (Pseudonode1, Node2) encodes a local ISO node-ID, remote
IPv4 router-ID and remote ISO node-id.

```
+-----------------+     +-----------------+     +-----------------+
|      Node1      |     |  Pseudonode 1   |     |      Node2      |
|1921.6800.1001.00|--->|1921.6800.1001.02|--->|1921.6800.1002.00|
|    192.168.1.1   |     |                 |     |    192.168.1.2   |
+-----------------+     +-----------------+     +-----------------+
```
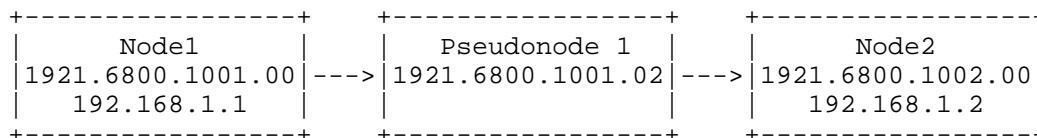
Figure 11: IS-IS Pseudonodes

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

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

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

3.4. Link Descriptors

The 'Link Descriptor' field is a set of Type/Length/Value (TLV)
triplets. The format of each TLV is shown in Figure 8. The 'Link
descriptor' TLVs uniquely identify a link between a pair of anchor
Routers.

The encoding of 'Link Descriptor' TLVs, i.e. the Codepoints in
'Type', and the 'Length' and 'Value' fields are the same as defined

in [RFC5305], [RFC5307], and [RFC6119] for sub-TLVs in the Extended
IS reachability TLV. The Codepoints are in the IANA Protocol
Registry for IS-IS, sub-TLV Codepoints for TLV 22, [IANA-ISIS].
Although the encodings for 'Link Descriptor' TLVs were originally
defined for IS-IS, the TLVs can carry data sourced either by IS-IS or
OSPF.

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

```
   +------+-----------------------------+------------------------+
   | Type | Description                 | Defined in:            |
   +------+-----------------------------+------------------------+
   |    4 | Link Local/Remote Identifiers | [RFC5307], Section 1.1 |
   |    6 | IPv4 interface address      | [RFC5305], Section 3.2 |
   |    8 | IPv4 neighbor address       | [RFC5305], Section 3.3 |
   |   12 | IPv6 interface address      | [RFC6119], Section 4.2 |
   |   13 | IPv6 neighbor address       | [RFC6119], Section 4.3 |
   |  222 | Multi Topology ID           | Section 3.5            |
   +------+-----------------------------+------------------------+
```

Table 2: Link Descriptor TLVs

3.5. Multi Topology ID TLV

The Multi Topology ID TLV (Type 222) carries the Multi Topology ID
for this link. The semantics of the Multi Topology ID are defined in
RFC5120, Section 7.2 [RFC5120], and the OSPF Multi Topology ID),
defined in RFC4915, Section 3.7 [RFC4915]. If the value in the Multi
Topology ID TLV is derived from OSPF, then the upper 9 bits of the
Multi Topology ID are set to 0.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Type             |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|R R R R|    Multi Topology ID  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
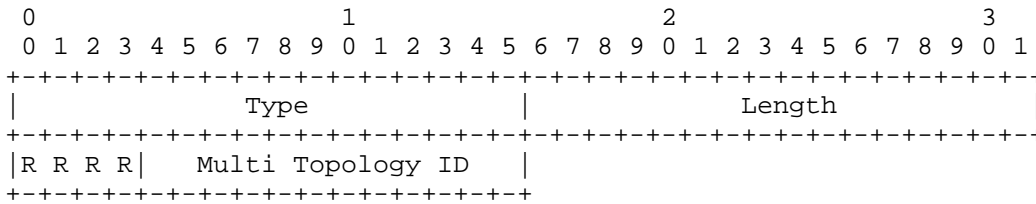
Figure 12: Multi Topology ID TLV format

3.6. Link Attributes

The 'Link Attributes' field is a set of Type/Length/Value (TLV)
triplets. The format of each TLV is shown in Figure 8.

For Codepoints < 255, the encoding of 'Link Attributes' TLVs, i.e.
the Codepoints in 'Type', and the 'Length' and 'Value' fields are the

same as defined in [RFC5305], [RFC5307], and [RFC6119] for sub-TLVs in the Extended IS reachability TLV. The Codepoints are in the IANA Protocol Registry for IS-IS, sub-TLV Codepoints for TLV 22, [IANA-ISIS]. Although the encodings for 'Link Attributes' TLVs were originally defined for IS-IS, the TLVs can carry data sourced either by IS-IS or OSPF.

For Codepoints > 255, the encoding of 'Link Attributes' TLVs is described in subsequent sections.

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

```
+-------+------------------------------+-----------------------+
| Type  | Description                  | Defined in:           |
+-------+------------------------------+-----------------------+
|    3  | Administrative group (color) | [RFC5305], Section 3.1 |
|    9  | Maximum link bandwidth       | [RFC5305], Section 3.3 |
|   10  | Max. reservable link bandwidth | [RFC5305], Section 3.5 |
|   11  | Unreserved bandwidth         | [RFC5305], Section 3.6 |
|   20  | Link Protection Type         | [RFC5307], Section 1.2 |
| 64509 | MPLS Protocol                | Section 3.6.1         |
| 64510 | TE Default Metric            | Section 3.6.2         |
| 64511 | IGP Link Metric              | Section 3.6.3         |
| 64512 | Shared Risk Link Group       | Section 3.6.4         |
| 64513 | OSPF specific link attribute | Section 3.6.5         |
| 64514 | IS-IS specific link attribute | Section 3.6.6        |
| 64515 | Area ID                      | Section 3.6.7         |
+-------+------------------------------+-----------------------+
```

Table 3: Link Attribute TLVs

3.6.1. MPLS Protocol TLV

The MPLS Protocol TLV (Type 64511) carries a bit mask describing which MPLS signaling protocols are enabled. The length of this TLV is 1. The value is a bit array of 8 flags, where each bit represents an MPLS Protocol capability.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Type             |             Length          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|L R           |
+-+-+-+-+-+-+-+-+
```
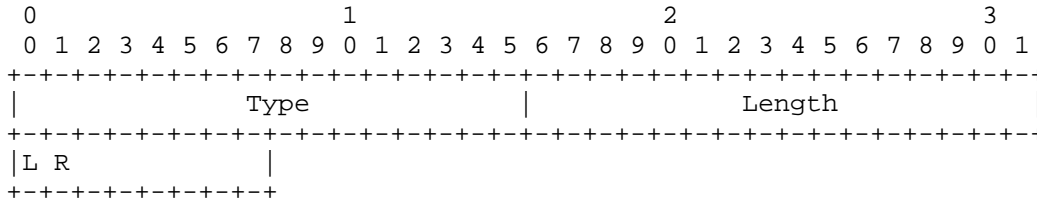
Figure 13: MPLS Protocol TLV

The following bits are defined:

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

                 Table 4: MPLS Protocol TLV Codes

3.6.2. TE Default Metric TLV

   The TE Default Metric TLV (Type 64512) carries the TE Default metric
   for this link. This TLV corresponds to the IS-IS TE Default metric
   sub-TLV (Type 18), defined in RFC5305, Section 3.7 [RFC5305], and the
   OSPF TE Metric sub-TLV (Type 5), defined in RFC3630, Section 2.5.5
   [RFC3630]. If the value in the TE Default metric TLV is derived from
   IS-IS TE Default Metric, then the upper 8 bits of this TLV are set to
   0.

```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |              Type             |             Length            |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                       TE Default Metric                       |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
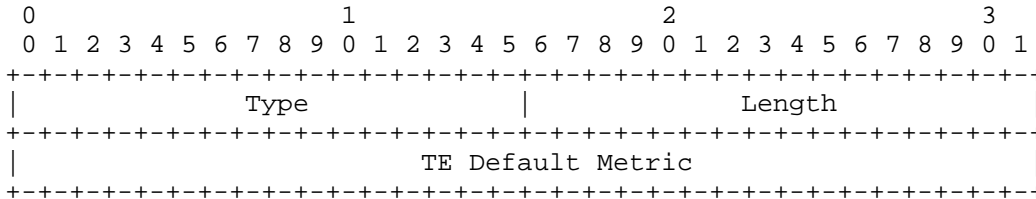
              Figure 14: TE Default metric TLV format

3.6.3. IGP Link Metric TLV

   The IGP Metric TLV (Type 64513) carries the IGP metric for this link.
   This attribute is only present if the IGP link metric is different
   from the TE Default Metric (Type 18). The length of this TLV is 3.
   If the length of the IGP link metric from which the IGP Metric value
   is derived is less than 3 (e.g. for OSPF link metrics or non-wide
   IS-IS metric), then the upper bits of the TLV are set to 0.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Type              |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     IGP Link Metric           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
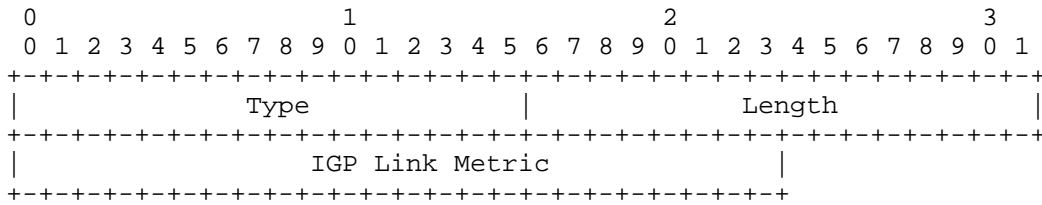
               Figure 15: IGP Link Metric TLV format

3.6.4. Shared Risk Link Group TLV

   The Shared Risk Link Group (SRLG) TLV (Type 64514) 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 16. The length of this TLV is 4 *
   (number of SRLG values).

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Type              |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                 Shared Risk Link Group Value                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       ...........                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                 Shared Risk Link Group Value                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
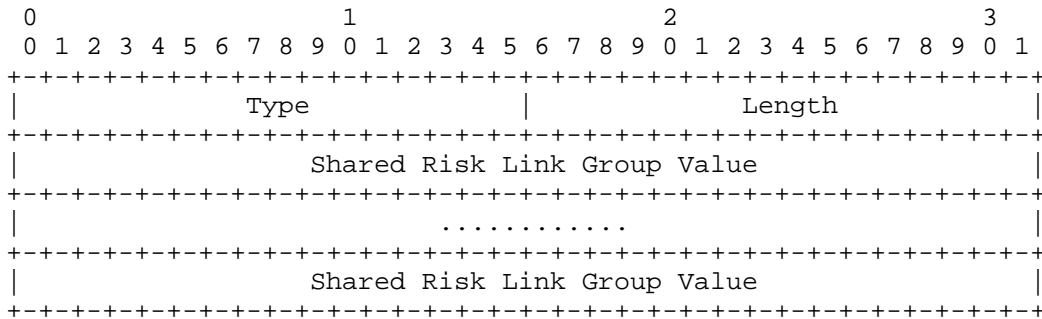
            Figure 16: Shared Risk Link Group TLV format

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

3.6.5. OSPF specific link attribute TLV

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

the BGP link-state NLRI.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Type             |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
|           OSPF specific link attributes (variable)           |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
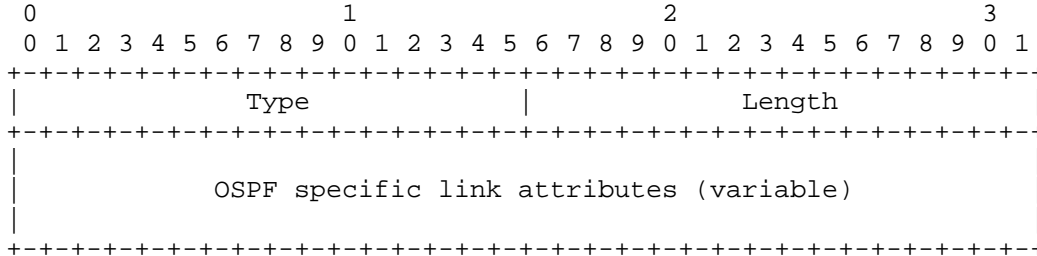
Figure 17: OSPF specific link attribute format

3.6.6. IS-IS specific link attribute TLV

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

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Type             |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
|          IS-IS specific link attributes (variable)           |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
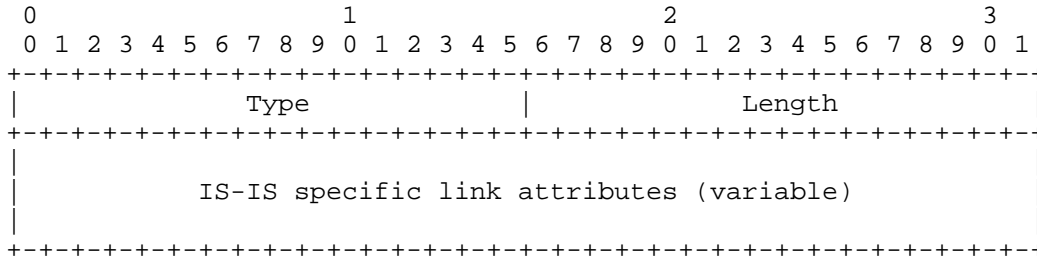
Figure 18: IS-IS specific link attribute format

3.6.7. Link Area TLV

The Area TLV (Type 64515) carries the Area ID which is assigned on
this link. If a link is present in more than one Area then several
occurrences of this TLV may be generated. Since only the OSPF
protocol carries the notion of link specific areas, the Area ID has a
fixed length of 4 octets.

```
0                   1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Type             |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Area ID                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
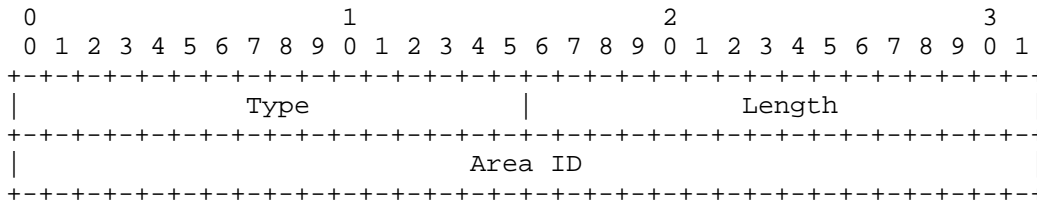
Figure 19: Link Area TLV format

3.7. Node Attributes

   The following node attribute TLVs are valid in the Node NLRI:

```
+-------+------------------------------+----------+
| Type  | Description                  | Length   |
+-------+------------------------------+----------+
|   229 | Multi Topology               |        2 |
| 65515 | Node Flag Bits               |        1 |
| 65516 | OSPF Specific Node Properties | variable |
| 65517 | IS-IS Specific Node Properties | variable |
| 65518 | Node Area ID                 | variable |
+-------+------------------------------+----------+
```

Table 5: Node Attribute TLVs

3.7.1. Multi Topology Node TLV

   The Multi Topology TLV (Type 229) carries the Multi Topology ID and
   topology specific flags for this node. The format of the Multi
   Topology TLV is defined in RFC5120, Section 7.1 [RFC5120]. If the
   value in the Multi Topology TLV is derived from OSPF, then the upper
   9 bits of the Multi Topology ID and the 'O' and 'A' bits are set to
   0.

```
0                   1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Type             |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|O A R R|   Multi Topology ID   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
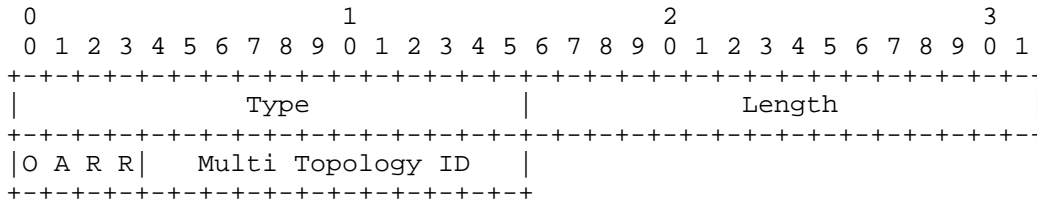
Figure 20: Multi Topology Node TLV format

3.7.2. Node Flag Bits TLV

   The Node Flag Bits TLV (Type 1) carries a bit mask describing node
   attributes. The value is a bit array of 8 flags, where each bit
   represents an MPLS Protocol capability.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Type             |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Flags     |
+-+-+-+-+-+-+-+-+
```
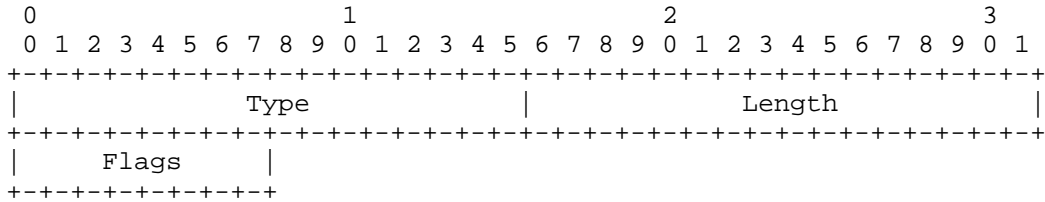
                  Figure 21: Node Flag Bits TLV format

   The bits are defined as follows:

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

                  Table 6: Node Flag Bits Definitions

3.7.3. OSPF Specific Node Properties TLV

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

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Type             |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
|            OSPF specific node properties (variable)           |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
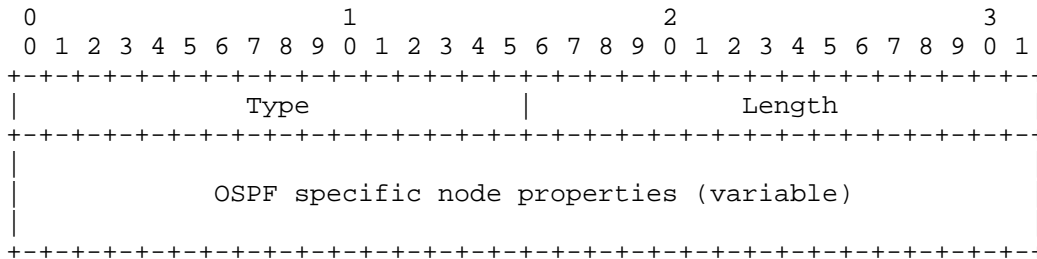
               Figure 22: OSPF specific Node property format

3.7.4. IS-IS Specific Node Properties TLV

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

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Type             |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
|           IS-IS specific node properties (variable)           |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
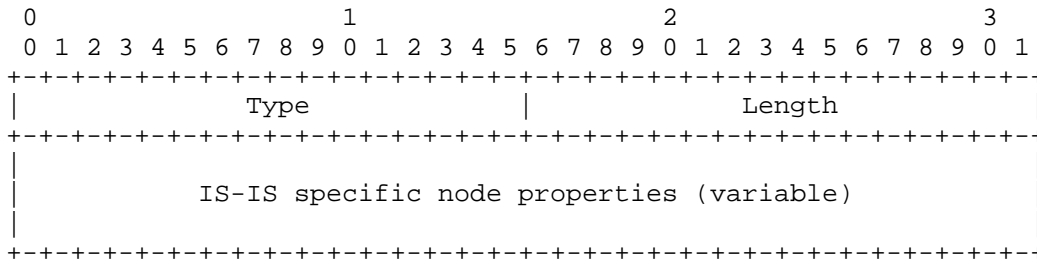
               Figure 23: IS-IS specific Node property format

3.7.5. Area Node TLV

   The Area TLV (Type 65518) carries the Area ID which is assigned to
   this node. If a node is present in more than one Area then several
   occurrences of this TLV may be generated. Since only the IS-IS
   protocol carries the notion of per-node areas, the Area ID has a
   variable length of 1 to 20 octets.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Type             |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
|                                                               |
|                     Area ID (variable)                        |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
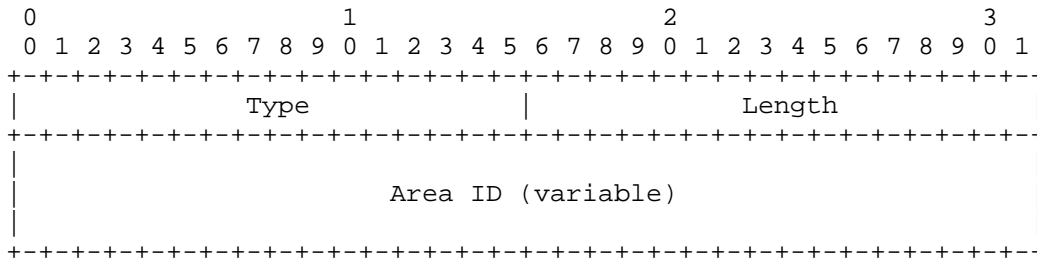
                    Figure 24: Area Node TLV format

3.8. Inter-AS Links

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

4. Link to Path Aggregation

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

4.1. Example: No Link Aggregation

   Consider Figure 25. Both AS1 and AS2 operators want to protect their
   inter-AS {R1,R3}, {R2, R4} links using RSVP-FRR LSPs. If R1 wants to
   compute its link-protection LSP to R3 it needs to "see" an alternate
   path to R3. Therefore the AS2 operator exposes its topology. All
   BGP TE enabled routers in AS1 "see" the full topology of AS and
   therefore can compute a backup path. Note that the decision if the
   direct link between {R3, R4} or the {R4, R5, R3) path is used is made
   by the computing router.

```
        AS1    :    AS2
               :
          R1-------R3
          |    :   | \
          |    :   |  R5
          |    :   | /
          R2-------R4
               :
               :
```
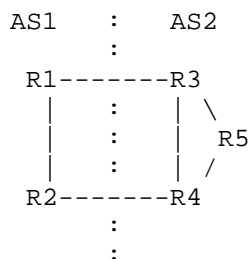
Figure 25: no-link-aggregation

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 26. The only link which gets advertised by AS2 is an
   "aggregate" link between R3 and R4. This is enough to tell AS1 that
   there is a backup path. However the actual links being used are
   hidden from the topology.

```
        AS1    :    AS2
               :
          R1-------R3
          |    :   |
          |    :   |
          |    :   |
          R2-------R4
               :
               :
```
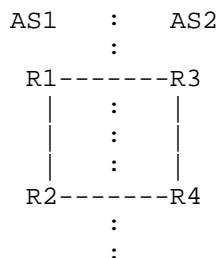
Figure 26: asbr-link-aggregation

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 27. Rather
   than exposing all specific R3 to R6 links, AS3 is modeled as a single
   node which connects to the border routers of the aggregated domain.

```
        AS1   :    AS2   :    AS3
              :          :
        R1-------R3-----
         |    :          : \
         |    :          :   vR0
         |    :          : /
        R2-------R4-----
              :          :
              :          :
```
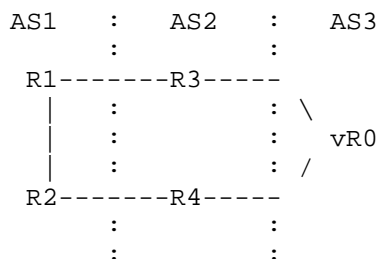
Figure 27: multi-as-aggregation

5. IANA Considerations

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

   This document requests creation of a new registry for node anchor,
   link descriptor and link attribute TLVs. The range of Codepoints in
   the registry is 0-65535. Values 0-255 will shadow Codepoints of the
   IANA Protocol Registry for IS-IS, sub-TLV Codepoints for TLV 22.
   Values 256-65535 will be used for Codepoints that are specific to the
   BGP TE NLRI. The registry will be initialized as shown in Table 2
   and Table 3. Allocations within the registry will require
   documentation of the proposed use of the allocated value and approval
   by the Designated Expert assigned by the IESG (see [RFC5226]).

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

6. Manageability Considerations

   This section is structured as recommended in [RFC5706].

6.1. Operational Considerations

6.1.1. Operations

   Existing BGP operation procedures apply. No new operation procedures
   are defined in this document.

6.1.2. Installation and Initial Setup

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

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

   o The maximum rate at which Link State NLRIs will be advertised/
      withdrawn from neighbors is set to ???.

6.1.3. Migration Path

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

6.1.4. Requirements on Other Protocols and Functional Components

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

6.1.5. Impact on Network Operation

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

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

6.1.6. Verifying Correct Operation

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

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

6.2. Management Considerations

6.2.1. Management Information

6.2.2. Fault Management

   TBD.

6.2.3. Configuration Management

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

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

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

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

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

6.2.4. Accounting Management

Not Applicable.

6.2.5. Performance Management

An implementation SHOULD provide the following statistics:

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

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

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

o Total number of locally originated Link-State NLRIs

6.2.6. Security Management

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

o Drop all updates from Consumer peers

7. Security Considerations

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

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

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

8. Acknowledgements

   We would like to thank Nischal Sheth for contributions to this
   document.

   We would like to thank Alia Atlas, David Ward, John Scudder, Kaliraj
   Vairavakkalai, Yakov Rekhter, Les Ginsberg, Mike Shand, and Richard
   Woundy for their comments.

9. References

9.1. Normative References

   [IANA-ISIS]
             "IS-IS TLV Codepoint, Sub-TLVs for TLV 22", <http://
             www.iana.org/assignments/isis-tlv-codepoints/
             isis-tlv-codepoints.xml#isis-tlv-codepoints-3>.

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

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

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

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

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

   [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering
             (TE) Extensions to OSPF Version 2", RFC 3630,
             September 2003.

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

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

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

                  January 2007.

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

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

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

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

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

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

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

   [RFC6119] Harrison, J., Berger, J., and M. Bartlett, "IPv6 Traffic
              Engineering in IS-IS", RFC 6119, February 2011.

9.2. Informative References

   [I-D.ietf-alto-protocol]
              Alimi, R., Penno, R., and Y. Yang, "ALTO Protocol",
              draft-ietf-alto-protocol-08 (work in progress), May 2011.

   [I-D.ietf-isis-mi]
              Previdi, S., Ginsberg, L., Shand, M., Roy, A., and D.
              Ward, "IS-IS Multi-Instance", draft-ietf-isis-mi-04 (work
              in progress), March 2011.

   [I-D.ietf-ospf-multi-instance]
              Lindem, A., Roy, A., and S. Mirtorabi, "OSPF Multi-
              Instance Extensions", draft-ietf-ospf-multi-instance-04
              (work in progress), April 2011.

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

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

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

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

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

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

Authors' Addresses

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

   Email: hannes@juniper.net

   Jan Medved
   Juniper Networks, Inc.
   1194 N. Mathilda Ave.
   Sunnyvale, CA 94089
   US

   Email: jmedved@juniper.net

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

Email: afarrel@juniper.net

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

Email: sprevidi@cisco.com

IDR Working Group   Rajiv Asati
Internet Draft   Cisco Systems
Updates: 4271 (if approved)
Intended status: Standards Track
Expires: December 5, 2019
                                                       June 5, 2019

                  BGP Bestpath Selection Criteria Enhancement
              draft-ietf-idr-bgp-bestpath-selection-criteria-12.txt

   Abstract

      BGP specification (RFC4271) prescribes 'BGP next-hop reachability'
      as one of the key 'Route Resolvability Condition' that must be
      satisfied before the BGP bestpath candidate selection. This
      condition, however, may not be sufficient (as explained in the
      Appendix section) and would desire further granularity.

      This document defines enhances the "Route Resolvability Condition"
      to facilitate the next-hop to be resolved in the chosen data plane.

Table of Contents

1. Introduction

   As per BGP specification [RFC4271], when a router receives a BGP
   path, BGP must qualify it as the valid candidate prior to the BGP
   bestpath selection using the 'Route Resolvability Condition'
   (section#9.1.2.1 of RFC4271]. After the path gets qualified as the
   bestpath candidate, it becomes eligible to be the bestpath, and may
   get advertised out to the neigbhor(s), if it became the bestpath.

   However, in BGP networks that utilize data plane protocol other than
   IP, such as MPLS [RFC3031] etc. to forward the received traffic
   towards the next-hop, the above qualification condition may not be
   sufficient. In fact, this may expose the BGP networks to experience
   traffic blackholing i.e. traffic loss, due to malfunctioning of the
   chosen data plane protocol to the next-hop. This is explained
   further in the Appendix section.

   This document defines further granularity to the "Route
   Resolvability Condition" by (a) resolving the BGP next-hop

reachability in the forwarding database of a particular data plane
protocol, and (b) optionally including the BGP next-hop "path
availability" check.

The goal is to enable BGP to select the bestpaths based on whether
or not the corresponding nexthop can be resolved in the valid data
plane.

2. Specification 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 [RFC2119].

3. Route Resolvability Condition - Modification

This document proposes two amendments to 'Route Resolvability
Condition', which is defined in RFC4271, in consideration for a
particular data plane protocol:

1) The next-hop reachability (check) SHOULD be resolved in a
   forwarding database of a particular data plane protocol.

   For example, if a BGP IPv4/v6 or VPNv4/v6 path wants to use
   MPLS data plane to the next-hop, as determined by the policy,
   then the BGP 'next-hop reachability' should be resolved using
   the MPLS forwarding database. In another example, if BGP path
   wants to use the IP data plane to the next-hop, as determined
   by the policy, then BGP 'next-hop reachability' should be
   resolved using the IP forwarding database. The latter example
   relates to MPLS-in-IP encapsulation techniques such as
   [RFC4817], [RFC4023] etc.

The selection of particular data plane is a matter of a policy, and
is outside the scope of this document. It is envisioned that the
policy would exist for either per-neighbor or per-SAFI or both. A
dynamic signaling such as BGP encapsulation SAFI (or tunnel encap
attribute) [RFC5512] may be used to convey the data plane protocol
chosen by the policy.

This check is about confirming the availability of the valid
forwarding entry for the next-hop in the forwarding database of the
chosen data plane protocol.

   2) The 'path availability' check for the BGP next-hop MAY be
      performed. This criterion checks for the functional data plane
      path to the next-hop in a particular data plane protocol.

   The path availability check may be performed by any of the OAM data-
   plane liveness mechanisms associated with the data plane that is
   used to reach the Next Hop. The data plane protocol for this
   criterion MUST be the same as the one selected by the previous
   criterion (#1).

   The mechanism(s) to perform the "path availability" check and the
   selection of particular data plane are a matter of a policy and
   outside the scope of this document.

        For example, if a BGP VPNv4 path wants to use the MPLS as the
        data plane protocol to the next-hop, then MPLS path
        availability to the next-hop should be evaluated i.e. liveness
        of MPLS LSP to the next-hop should be validated.

   This check is about confirming the availability of functioning path
   to the next-hop. Note that it is not necessary to trigger the data-
   plane liveness mechanism for a given next-hop as a consequence of
   this check, though it may be an option. Another option is to do it a
   priori. The selection of a particular option is deemed deployment
   specific and outside the scope of this document.

4. Conclusions

   Both amendments discussed in section 2 provide further clarity and
   granularity to help the BGP speaker to either continue to advertise
   a BGP path's reachability or withdraw the BGP path's reachability,
   based on the consideration for the path's next-hop reachability
   and/or availability in a particular data plane.

   It is not expected that the proposed amendments would negatively
   impact BGP convergence, barring any implementation specifics.

   The intention of this document is to help operators to build BGP
   networks that can avoid self-blackholing.

5. Security Considerations

   While this draft doesn't impose any additional security constraints,
   it can help with mitigating one particular type of routing attack in
   which a BGP speaker could receive routes with an arbitrary next-hop.
   If the next-hop is not reachable, then those routes/paths would not
   get selected.

6. IANA Considerations

   None.

7. Acknowledgments

   Yakov Rekhter provided critical suggestions and feedback to improve
   this document. Thanks to John Scudder and Chandrashekhar Appanna for
   contributing to the discussions that formed the basis of this
   document. Thanks to Ilya Varlashkin and Michael Benjamin, who made
   the case to revive this document and provided useful feedback. Also
   thanks to Robert Raszuk and Keyur Patel for constructive feedback.

   This document was prepared using 2-Word-v2.0.template.dot.

8. Appendix

8.1. Problem Applicability

   In IP networks using BGP, a router would continue to attract traffic
   by advertising the BGP prefix reachability to neighbor(s) as long as
   the router had a route to the next-hop in its routing table, but
   independent of whether the router has a functional forwarding path
   to the next-hop. This may cause the forwarded traffic to be dropped
   inside the IP network.

   In MPLS or MPLS VPN networks [RFC4364], the same problem is observed
   if the functional MPLS LSP to the next-hop is not available (due to
   the forwarding path error on any node along the path to the next-
   hop).

   The following MPLS/VPN topology clarifies the problem -

```
        <-eBGP/IGP-> <-------MP-BGP------> <-eBGP/IGP->

   CE1˜˜˜˜˜˜˜˜˜PE1˜˜˜MPLS Network˜˜˜PE2˜˜˜˜˜˜˜˜CE2˜˜
                                                    ^
            ======PE1-PE2 LSP==>                    ^
                                                    ^
                                               a.b.c.d
```
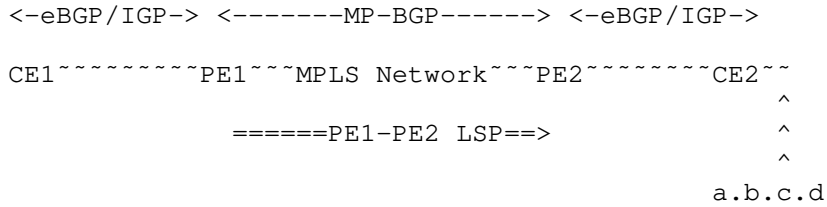
                   Figure 1 MPLS VPN Network

   In the network illustrated in Figure 1, the PE1 to PE2 LSP may be
   non-functional due to any reason such as corrupted MPLS Forwarding
   Table entry, or the missing MPLS Forwarding table entry, or LDP
   binding defect, or down LDP session between the P routers (with
   independent label distribution control) etc. In such a situation, it
   is clear that the CE1->CE2 traffic inserted into the MPLS network by
   PE1 will get dropped inside the MPLS network.

   It is undesirable to have PE1 continue to convey to the CE1 router
   that PE1 (and the MPLS network) is still the next-hop for the remote
   VPN reachability, without being sure of the corresponding LSP
   health.

8.1.1. Multi-Homed VPN Site

   If the remote VPN site is dual-homed to both PE2 and PE3, then PE1
   may learn two VPNv4 paths to the prefix a.b.c.d. via PE2 and PE3
   routers, as shown below in Figure 2. PE1 may select the bestpath for
   the prefix a.b.c.d via PE2 (say, for which the PE1->PE2 LSP is mal-
   functioning) and advertise that bestpath to CE1 in the context of
   figure 2.


```
                  <------MP-BGP------>

      CE1˜˜˜˜˜˜˜˜˜PE1˜˜˜MPLS Network˜˜˜PE2˜˜˜˜˜˜˜˜CE2˜˜
                        \                    /    ^
                         \˜˜˜˜˜˜˜˜˜PE3˜˜˜˜˜˜/     ^
                                                  ^
                                             a.b.c.d
```
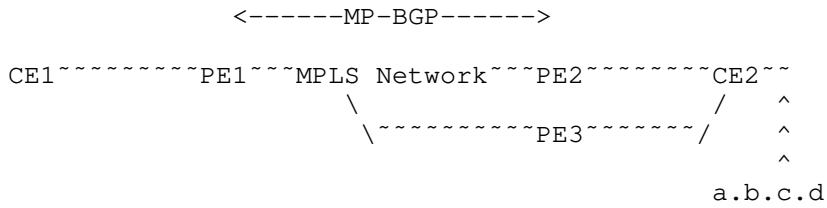

           Figure 2 MPLS VPN Network - CE2 Dual-Homing

This causes CE1 to likely send the traffic destined to prefix
a.b.c.d to the PE1 router, which forwards the traffic over the
malfunctioning LSP to PE2. It is clear that this MPLS encapsulated
VPN traffic ends up getting dropped or blackholed somewhere inside
the MPLS network.

It is desirable to force PE1 to select an alternate bestpath via
that next-hop (such as PE3), whose LSP is correctly functioning.

8.1.2. Single-Homed VPN Site with Site-to-Site Backup Connectivity

The local VPN site may have a backup/dial-up link available at the
CE router, but the backup link will not even be activated as long as
the CE's routing table continues to point to the PE router as the
next-hop (over the MPLS/VPN network).


```
                        <------MP-BGP------>

       CE1˜˜˜˜˜˜˜˜˜PE1˜˜˜MPLS Network˜˜˜PE2˜˜˜˜˜˜˜˜˜CE2˜˜
         \                                          /    ^
          \˜˜˜˜˜˜˜˜˜˜˜˜˜˜backup path˜˜˜˜˜˜˜˜˜˜˜˜˜˜/     ^
                                                         ^
                                                 a.b.c.d
```
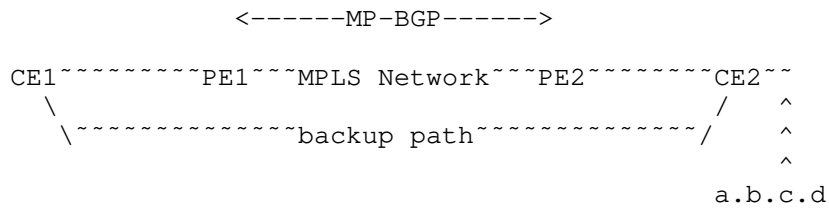
Figure 3 MPLS VPN Network - CE1-CE2 Backup connection

Unless PE2 withdraws the route via the routing protocol used on the
PE-CE link, CE1 will not be able to activate the backup link
(barring any tracking functionality) to the remote VPN site.

In summary, if PE1 could appropriately qualify the BGP VPNv4
bestpath, then the VPN traffic outage could likely be avoided. Even
if the VPN site was not multi-homed, it is desirable to force PE1 to
withdraw the path from CE1 to improve the CE-to-CE convergence. This
document proposes a mechanism to achieve the optimal BGP behavior at
PE.

8.1.3. 6PE or 6VPE

   This problem is very much applicable to the MPLS network that is
   providing either 6PE [RFC4978] or 6VPE [RFC4659] service to
   transport IPv6 packets over the MPLS network.

9. References

9.1. Normative References

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

   [RFC4364] Rosen E. and Rekhter Y., "BGP/MPLS IP Virtual Private
             Networks (VPNs)", RFC4364, February 2006.

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

9.2. Informative References

   [RFC3031] Rosen, et al., "Multiprotocol Label Switching
             Architecture", RFC3031, Jan 2001.

   [RFC5512] Rosen, E., Mohapatra, P., "BGP Encapsulation SAFI and BGP
             Tunnel Encapsulation Attribute", RFC5512, April 2009.

   [RFC4023] Rosen, et al., "Encapsulating MPLS in IP or Generic
             Routing Encapsulation", RFC4023, March 2005.

   [RFC4817] Townsley, et al., "Encapsulation of MPLS over Layer 2
             Tunneling Protocol Version 3", RFC4817, Nov 2006.

   [RFC4978] De Clercq, et al., "Connecting IPv6 Islands over IPv4 MPLS
             Using IPv6 Provider Edge Routers", RFC4978, Feb 2007.

   [RFC4659] De Clercq, et al., "BGP-MPLS IP VPN Extension for IPv6
             VPN", RFC4659, Sep 2006.

Author's Addresses

   Rajiv Asati
   Cisco Systems
   7025 Kit Creek Road
   RTP, NC 27560 USA
   Email: rajiva@cisco.com

                 Revised Error Handling for BGP UPDATE Messages
                   draft-ietf-idr-optional-transitive-04.txt

Abstract

   According to the base BGP specification, a BGP speaker that receives
   an UPDATE message containing a malformed attribute is required to
   reset the session over which the offending attribute was received.
   This behavior is undesirable as a session reset would impact not only
   routes with the offending attribute, but also other valid routes
   exchanged over the session. This document partially revises the
   error handling for UPDATE messages, and provides guidelines for the
   authors of documents defining new optional attributes. Finally, it
   revises the error handling procedures for several existing
   attributes.

Status of this Memo

Copyright Notice

1. Introduction

According to the base BGP specification [RFC4271], a BGP speaker that
receives an UPDATE message containing a malformed attribute is
required to reset the session over which the offending attribute was
received. This behavior is undesirable as a session reset would
impact not only routes with the offending attribute, but also other
valid routes exchanged over the session. In the case of optional
transitive attributes, the behavior is especially troublesome and may
present a potential security vulnerability. The reason is that such
attributes may have been propagated without being checked by
intermediate routers that do not recognize the attributes -- in
effect the attribute may have been tunneled, and when they do reach a
router that recognizes and checks them, the session that is reset may
not be associated with the router that is at fault.

The goal for revising the error handling for UPDATE messages is to
minimize the impact on routing by a malformed UPDATE message, while
maintaining protocol correctness to the extent possible. This can be
achieved largely by maintaining the established session and keeping
the valid routes exchanged, but removing the routes carried in the
malformed UPDATE from the routing system.

This document partially revises the error handling for UPDATE
messages, and provides guidelines for the authors of documents
defining new optional attributes. Finally, it revises the error
handling procedures for several existing attributes. Specifically,
the error handling procedures of [RFC4271], [RFC1997], and [RFC4360]
are revised.

1.1. Specification of Requirements

   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].

2. Revision to Base Specification

   The first paragraph of Section 6.3 of [RFC4271] is revised as
   follows:

   Old Text:

      All errors detected while processing the UPDATE message MUST be
      indicated by sending the NOTIFICATION message with the Error Code
      UPDATE Message Error. The error subcode elaborates on the specific
      nature of the error.

   New text:

      An error detected while processing the UPDATE message for which a
      session reset is specified MUST be indicated by sending the
      NOTIFICATION message with the Error Code UPDATE Message Error.
      The error subcode elaborates on the specific nature of the error.

   The error handling of the following case described in Section 6.3 of
   [RFC4271] remains unchanged:

      If the Withdrawn Routes Length or Total Attribute Length
      is too large (i.e., if Withdrawn Routes Length + Total Attribute
      Length + 23 exceeds the message Length), then the Error Subcode
      MUST be set to Malformed Attribute List.

   The error handling of the following case described in Section 6.3 of
   [RFC4271] is revised

      If any recognized attribute has Attribute Flags that conflict with
      the Attribute Type Code, then the Error Subcode MUST be set to
      Attribute Flags Error. The Data field MUST contain the erroneous
      attribute (type, length, and value).

   as follows:

      If any attribute has Attribute Flags that conflict with the
      Attribute Type Code, then the error SHOULD be logged, and the
      Attribute Flags MUST be reset to the correct value. The UPDATE
      message MUST continue to be processed.

The error handling of all other cases described in Section 6.3 of
[RFC4271] that specify a session reset is revised as follows.

When a path attribute in an UPDATE message is determined to be
malformed, the UPDATE message containing that attribute MUST be
treated as though all contained routes had been withdrawn just as if
they had been listed in the WITHDRAWN ROUTES field (or in the
MP_UNREACH_NLRI attribute [RFC4760bis] if appropriate) of the UPDATE
message, thus causing them to be removed from the Adj-RIB-In
according to the procedures of [RFC4271]. In the case of an
attribute which has no effect on route selection or installation, the
malformed attribute MAY instead be discarded and the UPDATE message
continue to be processed. For the sake of brevity, the former
approach is termed "treat-as-withdraw", and the latter as "attribute
discard".

The approach of "treat-as-withdraw" MUST be used for the error
handling of the cases described in Section 6.3 of [RFC4271] that
specify a session reset and involve any of the following attributes:
ORIGIN, AS_PATH, NEXT_HOP, MULTI_EXIT_DISC, and LOCAL_PREF.

The approach of "attribute discard" MUST be used for the error
handling of the cases described in Section 6.3 of [RFC4271] that
specify a session reset and involve any of the following attributes:
ATOMIC_AGGREGATE and AGGREGATOR.

When multiple malformed attributes exist in an UPDATE message, if the
same approach (either "treat-as-withdraw" or "attribute discard") is
specified for the handling of these malformed attributes, then the
specified approach MUST be used. Otherwise "treat-as-withdraw" MUST
be used.

A document which specifies a new attribute MUST provide specifics
regarding what constitutes an error for that attribute and how that
error is to be handled.

Finally, we observe that in order to use the approach of "treat-as-
withdraw", the entire NLRI field and/or MP_REACH and MP_UNREACH
[RFC4760bis] attributes need to be successfully parsed. If this is
not possible, the procedures of [RFC4271] continue to apply.
Alternatively the error handling procedures specified in [RFC4760bis]
for disabling a particular AFI/SAFI MAY be followed.

3. Parsing of NLRI Fields

   To facilitate the determination of the NLRI field in an UPDATE with a
   malformed attribute, the MP_REACH or MP_UNREACH attribute (if
   present) SHOULD be encoded as the very first path attribute in an
   UPDATE as recommended by [RFC4760bis]. An implementation, however,
   MUST still be prepared to receive these fields in any position.

   If the encoding of [RFC4271] is used, the NLRI field for the IPv4
   unicast address family is carried immediately following all the
   attributes in an UPDATE. When such an UPDATE is received, we observe
   that the NLRI field can be determined using the "Message Length",
   "Withdrawn Route Length" and "Total Attribute Length" (when they are
   consistent) carried in the message instead of relying on the length
   of individual attributes in the message.

4. Operational Considerations

   Although the "treat-as-withdraw" error-handling behavior defined in
   Section 2 makes every effort to preserve BGP's correctness, we note
   that if an UPDATE received on an IBGP session is subjected to this
   treatment, inconsistent routing within the affected Autonomous System
   may result. The consequences of inconsistent routing can include
   long-lived forwarding loops and black holes. While lamentable, this
   issue is expected to be rare in practice, and more importantly is
   seen as less problematic than the session-reset behavior it replaces.

   When a malformed attribute is indeed detected over an IBGP session,
   we recommend that routes with the malformed attribute be identified
   and traced back to the ingress router in the network where the routes
   were sourced or received externally, and then a filter be applied on
   the ingress router to prevent the routes from being sourced or
   received. This will help maintain routing consistency in the
   network.

   Even if inconsistent routing does not arise, the "treat-as-withdraw"
   behavior can cause either complete unreachability or sub-optimal
   routing for the destinations whose routes are carried in the affected
   UPDATE message.

   Note that "treat-as-withdraw" is different from discarding an UPDATE
   message. The latter violates the basic BGP principle of incremental
   update, and could cause invalid routes to be kept. (See also
   Appendix A.)

   For any malformed attribute which is handled by the "attribute
   discard" instead of the "treat-as-withdraw" approach, it is critical

to consider the potential impact of doing so. In particular, if the
attribute in question has or may have an effect on route selection or
installation, the presumption is that discarding it is unsafe, unless
careful analysis proves otherwise. The analysis should take into
account the tradeoff between preserving connectivity and potential
side effects.

Because of these potential issues, a BGP speaker MUST provide
debugging facilities to permit issues caused by a malformed attribute
to be diagnosed. At a minimum, such facilities MUST include logging
an error listing the NLRI involved, and containing the entire
malformed UPDATE message when such an attribute is detected. The
malformed UPDATE message SHOULD be analyzed, and the root cause
SHOULD be investigated.

5. Error Handling Procedures for Existing Optional Attributes

5.1. AGGREGATOR

The error handling of [RFC4271] is revised as follows:

The AGGREGATOR attribute SHALL be considered malformed if any of the
following applies:

   o Its length is not 6 (when the "4-octet AS number capability" is
      not advertised to, or not received from the peer [RFC4893]).

   o Its length is not 8 (when the "4-octet AS number capability" is
      both advertised to, and received from the peer).

An UPDATE message with a malformed AGGREGATOR attribute SHALL be
handled using the approach of "attribute discard".

5.2. Community

The error handling of [RFC1997] is revised as follows:

The Community attribute SHALL be considered malformed if its length
is not a nonzero multiple of 4.

An UPDATE message with a malformed Community attribute SHALL be
handled using the approach of "treat-as-withdraw".

5.3. Extended Community

   The error handling of [RFC4360] is revised as follows:

   The Extended Community attribute SHALL be considered malformed if its
   length is not a nonzero multiple of 8.

   An UPDATE message with a malformed Extended Community attribute SHALL
   be handled using the approach of "treat-as-withdraw".

   Note that a BGP speaker MUST NOT treat an unrecognized Extended
   Community Type or Sub-Type as an error.

6. IANA Considerations

   This document makes no request of IANA.

7. Security Considerations

   This specification addresses the vulnerability of a BGP speaker to a
   potential attack whereby a distant attacker can generate a malformed
   optional transitive attribute that is not recognized by intervening
   routers (which thus propagate the attribute unchecked) but that
   causes session resets when it reaches routers that do recognize the
   given attribute type.

   In other respects, this specification does not change BGP's security
   characteristics.

8. Acknowledgments

   The authors wish to thank Ron Bonica, Mach Chen, Andy Davidson, Dong
   Jie, Rex Fernando, Joel Halpern, Akira Kato, Miya Kohno, Tony Li,
   Alton Lo, Shin Miyakawa, Tamas Mondal, Jonathan Oddy, Robert Raszuk,
   Yakov Rekhter, Rob Shakir, Naiming Shen, Shyam Sethuram, Ananth
   Suryanarayana, and Kaliraj Vairavakkalai for their observations and
   discussion of this topic, and review of this document.

9. Normative References

    [RFC1997] Chandrasekeran, R., Traina, P., and T. Li, "BGP
               Communities Attribute", RFC 1997, August 1996.

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

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

    [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended
               Communities Attribute", RFC 4360, February 2006.

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

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

    [RFC4760bis]
               Bates, T., Chandra, R., Katz, D., and Y. Rekhter,
               "Multiprotocol Extensions for BGP-4",
               draft-ietf-idr-rfc4760bis-03.txt, work in progress,
               August 2011.

Appendix A. Why not discard UPDATE messages?

   A commonly asked question is "why not simply discard the UPDATE
   message instead of treating it like a withdraw? Isn't that safer and
   easier?" The answer is that it might be easier, but it would
   compromise BGP's correctness so is unsafe. Consider the following
   example of what might happen if UPDATE messages carrying bad
   attributes were simply discarded:

                        AS1 ---- AS2
                          \   /
                           \   /
                            \ /
                            AS3
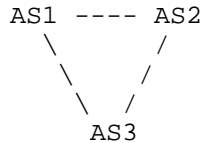
   o AS1 prefers to reach AS3 directly, and advertises its route to
      AS2.

o AS2 prefers to reach AS3 directly, and advertises its route to
   AS1.

o Connections AS3-AS1 and AS3-AS2 fail simultaneously.

o AS1 switches to prefer AS2's route, and sends an update message
   which includes a withdraw of its previous announcement. The
   withdraw is bundled with some advertisements. It includes a bad
   attribute. As a result, AS2 ignores the message.

o AS2 switches to prefer AS1's route, and sends an update message
   which includes a withdraw of its previous announcement. The
   withdraw is bundled with some advertisements. It includes a bad
   attribute. As a result, AS1 ignores the message.

The end result is that AS1 forwards traffic for AS3 towards AS2, and
AS2 forwards traffic for AS3 towards AS1. This is a permanent (until
corrected) forwarding loop.

Although the example above discusses route withdraws, we observe that
in BGP the announcement of a route also withdraws the route
previously advertised. The implicit withdraw can be converted into a
real withdraw in a number of ways; for example, the previously-
announced route might have been accepted by policy, but the new
announcement might be rejected by policy. For this reason, the same
concerns apply even if explicit withdraws are removed from
consideration.

10. Authors' Addresses

   John G. Scudder
   Juniper Networks

   Email: jgs@juniper.net

   Enke Chen
   Cisco Systems, Inc.

   EMail: enkechen@cisco.com

   Pradosh Mohapatra
   Cisco Systems, Inc.

   EMail: pmohapat@cisco.com

Keyur Patel
Cisco Systems, Inc.

EMail: keyupate@cisco.com

                 Automatic Route Target Filtering for legacy PEs
                        draft-l3vpn-legacy-rtc-00

Abstract

   This document describes a simple procedure that allows "legacy" BGP
   speakers to exchange route target membership information in BGP
   without using mechanisms specified in RFC 4684. The intention of the
   proposed technique is to help in partial deployment scenarios and is
   not meant to replace RFC 4684.

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 September 3, 2011.

Table of Contents

1. Introduction

   [RFC4684], "Constrained Route Distribution for Border Gateway
   Protocol/ MultiProtocol Label Switching (BGP/MPLS) Internet Protocol
   (IP) Virtual Private Networks (VPNs)" provides a powerful and general
   means for BGP speakers to exchange and propagate Route Target
   reachability information and constrain VPN route distribution to
   achieve high scale. However, it requires that all the BGP speakers
   in the network are upgraded to support this functionality. For
   example, in a network with route reflectors (RR), if one PE client in
   the cluster doesn't support constrained distribution, the cluster
   degenerates into storing and processing all the VPN routes. The
   route reflectors need to request and store all the network routes
   since they do not receive route target membership information from
   the legacy PEs. The RR will also generate all those routes to the
   legacy PEs and the legacy PEs will end up filtering the routes and
   store the subset of VPN routes that are of interest.

   This document specifies a mechanism for such legacy PE devices using
   existing configuration and toolset to provide similar benefits as
   [RFC4684]. At the same time, it is backward-compatible with the
   procedures defined in [RFC4684]. It also allows graceful upgrade of
   the legacy router to be [RFC4684] capable.

1.1. 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].

2. Basic Idea

   The basic idea is to make use of VPN unicast route exchange from the
   legacy PEs to a new BGP speaker (e.g. an RR) to signal RT membership.
   The legacy PEs announce a set of "special" routes with mapped RTs to
   the RR along with a standard community (defined in this document).
   The presence of the community triggers the RR to extract the RTs and
   build RT membership information.

3. Detailed Operation

3.1. Legacy PE Behavior

   The following simple steps are performed on the legacy PE device:

o Collect the "import route targets" of all the configured customer
   VRFs. Let's call this set 'IRTS'.

o Create a special "route-filter VRF" with a route distinguisher(RD)
   that's configured with the same value across the network for all
   legacy PE devices. Note: the equivalence of the RD value is for
   optimization - the operator may choose to use different values.

o Originate one or more routes in this VRF and attach a subset of
   'IRTS' as "translated route-target extended communities" with each
   route so as to evenly distribute the RTs (and to make sure they
   can fit into one BGP UPDATE message). Collectively, the union of
   the "translated route-target extended communities" of all these
   routes is equal to the set 'IRTS'. The translated RTs are
   attached as export route-targets for the routes originated in the
   route-filter VRF.

o The translation of the IRTs is necessary in order to refrain from
   importing "route-filter" VRF routes into VPN VRFs that would
   import the same route-targets. The translation of the IRTS is
   done as follows. For a given IRT, the equivalent translated RT
   (TRT) is constructed by means of swapping the value of the high-
   order octet of the Type field for the IRT (as defined in
   [RFC4360]).

```
 0                   1               0                   1
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      0x00     |     0x02      |   |     0x01      |     0x02      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|2B AS                          |   |2B AS => IP(high)              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+<=>+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Local Admin(high)              |   |Local Admin(high) => IP(low)   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Local Admin(low)               |   |Local Admin(low) => Local Admin|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

 0                   1               0                   1
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      0x01     |     0x02      |   |     0x02      |     0x02      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|IP(high)                       |   |IP(high) => 4B AS(high)        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+<=>+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|IP(low)                        |   |IP(low) => 4B AS(low)          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Local Admin                    |   |Local Admin => Local Admin     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

 0                   1               0                   1
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      0x02     |     0x02      |   |     0x00      |     0x02      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|4B AS(high)                    |   |4B AS(high) => 2B AS           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+<=>+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|4B AS(low)                     |   |4B AS(low) => Local Admin(high)|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Local Admin                    |   |Local Admin => Local Admin(low)|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

As an example, if IRT R= 65500:12244(hex: 0x0002ffdc00002fd4),
equivalent route-filter TRT: 255.220.0.0:12244(hex:
0x0102ffdc00002fd4). One shortcoming of the translation mechanism
is a possible collision between IRTs and TRTs if the network has
been configured with RTs of multiple higher order octet types
(2-byte AS, IP address, and 4-byte AS). It is expected that such
a configuration is rare in practice.

o As an alternative to the translation of the IRTS, the subset of
   the 'IRTS' can be attached as-is (without swapping the type field
   as described earlier) as "export route-target extended
   communities" with each route so as to evenly distribute the RTs

(and to make sure they can fit into one BGP UPDATE message). In
this case, the IRT subsets can be attached in outbound policy to
avoid the route-filter VRFs from being imported into VPN VRFs.
Also in this case, the route-filter VRF routes must be tagged with
a different special community (from that associated with the
translated RTs) as described in Section 4 so that the receiving
BGP speaker can distinguish the two cases.

   o The routes are marked with NO_ADVERTISE and NO_EXPORT well-known
      communities as well as the appropriate new community that's
      defined in this document Section 4. Note that there is no
      specific provision made to disallow configuration of subsequent
      route policies that can potentially alter the set of communities
      attached to "route-filter" VRF routes. The protocol behavior in
      such a case is undefined and the use of those policy statements is
      discouraged.

3.2. RR behavior

   Upon receiving the "route-filter" routes, the BGP speaker does its
   usual processing to store them in its local RIB. It recognizes them
   as route-filter routes based on the association of the new standard
   community as defined in this document. If required (as indicated by
   the community value), it translates the attached route-target
   extended communities (TRT) to equivalent import route-targets (IRT).
   Finally it creates the route-target filter list for each legacy
   client by collecting the entire set of route targets. From this
   point onwards, the behavior is similar to that defined in [RFC4684].
   The RR does not propagate the routes further because of their
   association with NO_ADVERTISE community. Also the VPN EoR that is
   sent by the legacy PE should also be used as an indication that the
   legacy PE is done sending the route-filter information as per the
   procedures defined in [RFC4684] for implementing a EoR mechanism to
   signal the completion of initial RT membership exchange.

3.2.1. Generating Route Target Membership NLRIs for the legacy PE
        clients

   The RR MAY also translate the received extended communities from
   legacy clients into route target membership NLRIs as if it had
   received those NLRIs from the client itself. This is useful for
   further propagation of the NLRIs to rest of the network to create RT
   membership flooding graph. When the route_filter routes are received
   with same RD (from all legacy PE speakers), processing of the paths
   to generate equivalent NLRIs becomes fairly easy.

4. ROUTE_FILTER community

   This memo defines four BGP communities that are attached to BGP
   UPDATE messages at the legacy PE devices and processed by the route
   reflectors as defined above. They are as follows:

   +---------------------------+------------------------------------+
   | Community | Meaning |
   +---------------------------+------------------------------------+
ROUTE_FILTER_v4	RTs are attached as-is for VPNv4
	route filtering
...	...
ROUTE_FILTER_v6	RTs are attached as-is for VPNv6
	route filtering
...	...
ROUTE_FILTER_TRANSLATED_v4	Translated RTs are attached for
	VPNv4 route filtering
...	...
ROUTE_FILTER_TRANSLATED_v6	Translated RTs are attached for
	VPNv6 route filtering
   +---------------------------+------------------------------------+

   In the absence of (or lack of support of) AF specific communities
   (ROUTE_FILTER_v6, ROUTE_FILTER_TRANSLATED_v6), the ROUTE_FILTER_v4 or
   ROUTE_FILTER_TRANSLATED_v4 MAY be treated by an implementation as a
   default VPN route-filter community to build a combination VPN filter
   for all VPN AFs (VPNv4, VPNv6) present on the RR. This is in
   accordance with the procedures in [RFC4684] to build combination
   route-filters for VPN AFs and AF specific route-filters defined in
   [I-D.keyur-bgp-af-specific-rt-constrain]. If this is the case, then
   subsequent receipt of any "route-filter" routes with AF specific
   communities (ROUTE_FILTER_v6, ROUTE_FILTER_TRANSLATED_v6) will
   override the default filters sent with ROUTE_FILTER_v4 or
   ROUTE_FILTER_TRANSLATED_v4 for the VPNv6 AFI when support for the AF
   specific communities exists.

5. Deployment Considerations

   When both the legacy PE and the RR support extended community based
   Outbound Route Filtering as in
   [I-D.draft-chen-bgp-ext-community-orf-00] this may be used as a
   alternate solution for the legacy PE to signal RT membership
   information, in order to realize the same benefits as [RFC4684].
   Also extended community ORF can be used amongst the RRs in lieu of
   [RFC4684] to realize similar benefits.

6. Contributors

   Significant contributions were made by Luis M Tomotaki and James
   Uttaro which the authors would like to acknowledge.

7. Acknowledgements

8. IANA Considerations

   IANA shall assign new code points from BGP first-come first-serve
   communities for the four communities as listed in Section 4.

9. Security Considerations

   None.

10. Normative References

   [I-D.chen-bgp-ext-community-orf]
             Chen, E. and Y. Rekhter, "Extended Community Based
             Outbound Route Filter for BGP-4",
             draft-chen-bgp-ext-community-orf-00 (work in progress),
             June 2006.

   [I-D.keyur-bgp-af-specific-rt-constrain]
             Patel, K., Raszuk, R., Djernaes, M., Dong, J., and M.
             Chen, "AFI Specific Route Target Distribution",
             draft-keyur-bgp-af-specific-rt-constrain-00 (work in
             progress), October 2010.

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

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

   [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended
             Communities Attribute", RFC 4360, February 2006.

   [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private
             Networks (VPNs)", RFC 4364, February 2006.

   [RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk,
             R., Patel, K., and J. Guichard, "Constrained Route

              Distribution for Border Gateway Protocol/MultiProtocol
              Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual
              Private Networks (VPNs)", RFC 4684, November 2006.

Authors' Addresses

   Pradosh Mohapatra
   Cisco Systems
   170 W. Tasman Drive
   San Jose, CA 95134
   USA

   Email: pmohapat@cisco.com

   Arjun Sreekantiah
   Cisco Systems
   170 W. Tasman Drive
   San Jose, CA 95134
   USA

   Email: asreekan@cisco.com

   Keyur Patel
   Cisco Systems
   170 W. Tasman Drive
   San Jose, CA 95134
   USA

   Email: keyupate@cisco.com

   Alton Lo
   Cisco Systems
   170 W. Tasman Drive
   San Jose, CA 95134
   USA

   Email: altonlo@cisco.com

Network Working Group   J. Uttaro
Internet-Draft   AT&T
Intended status: Standards Track   A. Simpson
Expires: April 22, 2012   Alcatel-Lucent
                                                              R. Shakir
                                                                    C&W
                                                            C. Filsfils
                                                            P. Mohapatra
                                                          Cisco Systems
                                                             B. Decraene
                                                          France Telecom
                                                              J. Scudder
                                                              Y. Rekhter
                                                        Juniper Networks
                                                        October 20, 2011

                            BGP Persistence
                  draft-uttaro-idr-bgp-persistence-00

Abstract

   For certain AFI/SAFI combinations it is desirable that a BGP speaker
   be able to retain routing state learned over a session that has
   terminated. By maintaining routing state forwarding may be
   preserved. This technique works effectively as long as the AFI/SAFI
   is primarily used to realize services that do not depend on
   exchanging BGP routing state with peers or customers. There may be
   exceptions based upon the amount and frequency of route exchange that
   allow for this technique. Generally the BGP protocol tightly couples
   the viability of a session and the routing state that is learned over
   it. This is driven by the history of the protocol and it's
   application in the internet space as a vehicle to exchange routing
   state between administrative authorities. This document addresses
   new services whose requirements for persistence diverge from the
   Internet routing point of view.

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 22, 2012.

Copyright Notice

Table of Contents

1. Introduction

   In certain scenarios, a BGP speaker may maintain forwarding in spite
   of BGP session termination. Currently all routing state learned
   between two speakers is flushed upon either normal or abnormal
   session termination. There are techniques that are useful for
   maintaining routing when a session abnormally terminates i.e BGR
   Graceful RestartR (RFC 4724) or normal termination such as
   increasing timers but they do not change the fundamental problem.
   The technique of BGP persistence works effectively as long as the
   expectation is that there is a decoupling of session viability and
   the correct service delivery, and the delivery uses the routing state
   learned over that session. This document proposes a modification to
   BGP's behavior by enabling persistence of BGP learned routing state
   in spite of normal or abnormal session termination.

1.1. 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].

2. Communities

   This memo defines three new communities that are used to identify the
   capability of a path to persist and whether or not that path is live
   or stale.

2.1. PERSIST

   This memo defines a new transitive BGP community, PERSIST, with value
   TBD (to be assigned by IANA). Attaching of the PERSIST community
   SHOULD be controlled by configuration. Attaching the PERSIST
   community indicates that the peer should maintain forwarding in the
   case of a session failure. The functionality SHOULD default to being
   disabled.

2.2. DO_NOT_PERSIST

   This memo defines a new transitive BGP community, DO_NOT_PERSIST,
   with value TBD (to be assigned by IANA). Attaching of the
   DO_NOT_PERSIST community SHOULD be controlled by configuration. The
   functionality SHOULD default to being disabled.

2.3. STALE

   This memo defines a new transitive BGP community, STALE, with value
   TBD (to be assigned by IANA). Attaching of the STALE community is
   limited to a path that currently has the PERSIST community attached

3. Configuration (Persistence Timer, PERSIST and DO_NOT_PERSIST
    Community)

    Persistence must be configured on a per session basis. A speaker
    configures the ability to persist independently of it's peer. There
    is no negotiation between the peers. A timer must be configured
    indicating the time to persist stale state from a peer where the
    session is no longer viable. This timer is designated as the
    persist-timer. A speaker must also attach persistence community
    value indicating if a path to a route should persist.

3.1. Settings for Different Applications

    The setting of the persist-timer should be based upon the field of
    use. BGP is used in a many different applications that each bring a
    unique requirement for retaining state. The following is not meant
    as a comprehensive listing but to suggest timer settings for a subset
    of AFI/SAFIs.

    L2VPN This AFI/SAFI requires the exchange of routing state in order
       to establish PWs to realize a VPLS VPN, or a VPWS PW. This AFI/
       SAFI does not require exchange of routing state with a customer
       and there is no eBGP session established. The persist-timer
       should be set to a large value on the order of days to infinity.

    L3VPN This AFI/SAFI requires the exchange of routing state to create
       a private VPN. This AFI/SAFI requires exchange of state with
       customers via eBGP and is dynamic. The SP needs to consider the
       possibility that stale state may not reflect the latest route
       updates and therefore may be incorrect from the customer
       perspective. The persist-timer should be set to a large value on
       the order of hours to a few days. this is built upon the notion
       some incorrectness is preferable to a large outage.

4. Operation

   Assuming a session failure has occurred a BGP persistent router must
   retain local forwarding state for those paths that are Persistent/
   Stale and propagate paths to downstream speakers that indicate that a
   given path is now stale.

4.1. Attaching the STALE Community Value and Propagation of Paths

   The following rules must be followed.

   o Identify paths learned over a failed session that have the PERSIST
      capable community value attached.

   o For those paths attach the STALE community value and propagate to
      all peers.

   o For those paths learned over the failed session that do not have
      PERSIST capable community value or are marked with the
      DO_NOT_PERSIST community follow BGP rules and generate withdrawals
      to all peers for those paths.

4.2. Forwarding

   The following rules must be followed to ensure valid forwarding:

   o All forwarding state must be retained i.e labels for BGP labeled
      unicast.

   o Forwarding must ensure that the Next Hop to a "stale" route is
      viable.

   o Forwarding to a "stale" route is only used if there are no other
      paths available to that route. In other words an active path
      always wins regardless of path selection. "Stale" state is always
      considered to be less preferred when compared with an active path.

   o Forwarding should be retained through an advertisement. When the
      session is re-established forwarding should only change if the new
      state is either different or better in terms of path selection. A
      make before break strategy should be employed.

   o Stale state may be retained indefinitely or may be programmed to
      expire via configuration.

   o The Receiving Speaker MUST replace the stale routes by the routing
      updates received from the peer. Once the End-of-RIB marker for an
      address family is received from the peer, it MUST immediately

remove any paths from the peer that are still marked as stale for
that address family.

o There is no restriction on whether the session is internal or
   external.

4.3. Example Behaviour

   Upon session establishment a speaker S2 may receive paths from S1
   that are marked with PERSIST, DO_NOT_PERSIST or neither. Assume S2
   is also peered with a downstream speaker S3.. Implementations MUST
   follow the specifications outlined below for.

   Upon recognition of the failure to S1, S2 will identify paths that
   had been marked with PERSIST, DO_NOT_PERSIST or neither learned from
   S1. S2 MUST implement the following behavior:

```
if ( P1 is tagged with PERSIST ) {

Retain Forwarding
  Attach the STALE Community to all paths that were marked with PERSIST
  Advertise STALE paths to all peers including S3
}
else ( P1 is marked with DO_NOT_PERSIST || not marked )

Tear down the forwarding structure for P1
Follow normal BGP rules i.e Best path, withdrawal etc.

fi
```

5. Deployment Considerations

   BGP Persistence as described in this document is useful within a
   single autonomous system or across autonomous systems.

6. Applications

   This technique may be useful in a wide array of applications where
   routing state is either fairly static or, the state is localized
   within a routing context. Some applications that come immediately to
   mind are L2 and L3 VPN.

6.1. Persistence in L2VPN (VPLS/VPWS)

   VPLS/VPWS VPNs use BGP to exchange routing state between two PEs.
   This exchange allows for the creation of a PW within a VPN context
   between those PEs. By definition, L2VPN does not exchange any
   routing state with customers via BGP. BGP persistence is very useful
   here as the state is quite constant. The only time state is
   exchanged is when a PW endpoint is provisioned, deleted or when a
   speaker reboots.

   Referring to Figure 1, PE1 and PE2 have advertised BGP routing state
   in order to create PWs between PE1 and PE2. The RRs are only
   responsible to reflect this state between the PEs. The use of a
   unique RD makes every path unique from the RRs perspective.

   Assume that the both RR experience catastrophic failure.

   Case 1 - All BGP speakers are persistent capable.

   The PWs created between PE1 and PE2 persist. Forwarding
   uninterrupted.

   Case 2 - PE1 and the RRs are persistent capable, PE2 is not.

   In this case the path advertised from PE2 via the RRs is persistent
   at PE1, the PW from PE1 to PE2 is not torn down. PE2 will remove the
   path from PE1 and tear down the PW from PE2 to PE1. THe effect is
   that MAC state learned at PE2 is valid as the PW is still valid. MAC
   state learned at PE1 is removed as the PW is no longer valid.
   Eventually MAC destinations recursed to the PW at PE1 destined for
   PE2 over the valid PW will time out.

   Assume that the RRs are valid but the iBGP sessions are torn down..

   Case 3 - All BGP speakers are persistent capable.

   The PWs created between PE1 and PE2 persist. Forwarding
   uninterrupted.

```
            VPNA                    VPNA
             PW++++++++++++++++++PW


           CE1-------PE1--------RR1-------PE2------CE2
            |                |
            |                |
            ---------RR2---------

           <--iBGP---><---iBGP-->
```
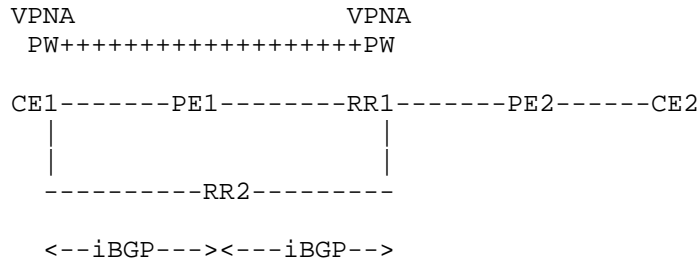
                          Figure 1

6.2. Persistence in L3VPN

```
                   --------RR1-------
                  / A             C \
         CE1 ----- PE1 --Forwarding Path-- PE2 ---- CE2
                  \ B             D /
                   ------- RR2 ------
```
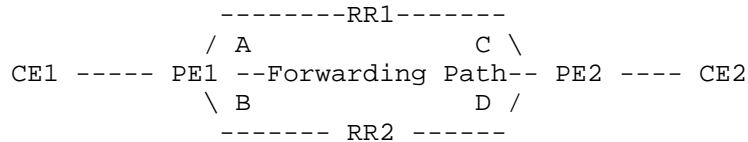
                          Figure 2

   In the case of a Layer 3 VPN topology, during the failure of a route
   reflector device at the current time, all routing information
   propagated via BGP is purged from the routing database. In this
   case, forwarding is interrupted within such a topology due to the
   lack of signalling information, rather than an outage to the
   forwarding path between the PE devices. With the addition of BGP
   persistence, a complete service outage can be avoided.

   The topology shown in Figure 2 is a simple L3VPN topology consisting
   of two customer edge (CE) devices, along with two provider edge (PE),
   and route reflector (RR) devices. In this case, where an RFC4364 VPN
   topology is utilised a BGP session exists between PE1 to both RR1 and
   RR2, and from PE2 to RR1 and RR2, in order to propagate the VPN
   topology.

   Case 1: No BGP speakers are persistence capable:

   o In this scenario, during a simultaneous failure of RR1 and RR2
      (which are extremely likely to share route reflector clients) both
      PE1 and PE2 remove all routing information from the VPN from their
      RIB, and hence a complete service outage is experienced.

   o Where either sessions A and B, or C and D fail simultaneously,
      routing information from either PE1 (in the case of A and B), or
      PE2 (in the case of C and D) are withdrawn, and a partial service
      topology exists.

o Both of the states described reflect a service outage where the
   forwarding path between the PE devices is not interrupted.

Case 2: All BGP speakers are persistence capable:

o PE1 continues to forward utilising the label information received
   from PE2 via the working forwarding path for the duration of the
   persistence timer (and vice versa).

o This condition occurs regardless of the session(s) that fail. In
   the worst case where sessions A, B, C and D fail simultaneously,
   the network continues to operate in the state in which it was at
   the time of the failure.

Case 3: PE1 and RR[12] are persistence capable - PE2 is not.

o During a failure of BGP session A or B, PE1 will continue to
   forward utilising the routing information received from the RRs
   for PE2 for the duration of the persistence timer. PE2 will
   continue to forward utilising the routing information received
   from the RRs, again for the duration of the persistence timer.

o In the case that either BGP session C or D fails, all routes will
   be withdrawn by RR[12] towards PE1 since these routes are not
   valid to be persisted by the RRs. The end result of this will be
   that the routes advertised by CE2 into the VPN will be withdrawn.

o Where the worst case failure occurs (i.e. sessions A, B, C and D
   fail) the routes advertised by CE1 into the VPN will be
   persistently advertised by the RR devices, whereas those
   advertised by CE2 will be withdrawn. Clearly in the example shown
   in the figure this results in a service outage, but where multiple
   PE devices exist within a topology, service is maintained for the
   subset of CEs attached to PE devices supporting the persistence
   capability.

Within the Layer 3 VPN deployment it should be noted that routing
information is less static than that of the many Layer 2 VPNs since
typically multiple routes exist within the topology rather than an
individual MAC address or egress interface per CE device on the PE
device. As such, the L3VPN operates with the routing databases in
the 'core' of the network reflecting those at the time of failure.
Should there be re-convergence for any path between the PE and CE
devices, this will result in invalid routing information, should the
egress PE device not hold alternate routing information for the
prefixes undergoing such re-convergence. It is expected that where
each PE maintains multiple paths to each egress prefix (where an
alternate path is available), it is expected that the egress PE will

forward packets towards an alternative egress PE for the prefix in
question where the topology is no longer valid.

The lack of convergence within a Layer 3 topology during the
persistent state SHOULD be considered since it may adversely affect
services, however, an assumption is made that a degraded service is
preferable to a complete service outage during a large-scale BGP
control plane failure.

7. Security Considerations

   The security implications of the persistence mechanism defined within
   in this document are akin to those incurred by the maintenance of
   stale routing information within a network. This is particularly
   relevant when considering the maintenance of routing information that
   is utilised for service segregation - such as MPLS label entries.

   For MPLS VPN services, the effectiveness of the traffic isolation
   between VPNs relies on the correctness of the MPLS labels between
   ingress and egress PEs. In particular, when an egress PE withdraws a
   label L1 allocated to a VPN1 route, this label MUST not be assigned
   to a VPN route of a different VPN until all ingress PEs stop using
   the old VPN1 route using L1.

   Such a corner case may happen today, if the propagation of VPN routes
   by BGP messages between PEs takes more time than the label re-
   allocation delay on a PE. Given that we can generally bound worst
   case BGP propagation time to a few minutes (e.g. 2-5), the security
   breach will not occur if PEs are designed to not reallocate a
   previous used and withdrawn label before a few minutes.

   The problem is made worse with BGP GR between PEs as VPN routes can
   be stalled for a longer period of time (e.g. 20 minutes).

   This is further aggravated by the BGP persistent extension proposed
   in this document as VPN routes can be stalled for a much longer
   period of time (e.g. 2 hours, 1 day).

   Therefore, to avoid VPN breach, before enabling BGP persistence, SPs
   needs to check how fast a given label can be reused by a PE, taking
   into account:

   o The load of the BGP route churn on a PE (in term of number of VPN
      label advertised and churn rate).

   o The label allocation policy on the PE (possibly depending upon the
      size of pool of the VPN labels (which can be restricted by
      hardware consideration or others MPLS usages), the label
      allocation scheme (e.g. per route or per VRF/CE), the re-
      allocation policy (e.g. least recently used label...)

   In addition to these considerations, the persistence mechanism
   described within this document is considered to be complex to exploit
   maliciously - in order to inject packets into a topology, there is a
   requirement to engineer a specific persistence state between two PE
   devices, whilst engineering label reallocation to occur in a manner
   that results in the two topologies overlapping. Such allocation is

particularly difficult to engineer (since it is typically an internal
mechanism of an LSR).

8. IANA Considerations

   IANA shall assigned community values from BGP well-known communities
   registry for the PERSIST, DO-NOT-PERSIST and STALE communities. No
   additional IANA action is required.

9. Acknowledgements

   We would like to acknowledge Roberto Fragassi (Alcatel-Lucent), John
   Medamana, (AT&T) Han Nguyen (AT&T), Jeffrey Haas (Juniper), Nabil
   Bitar (Verizon), Nicolai Leymann (DT) for their contributions to this
   document.

10. Normative References

   [RFC1997] Chandrasekeran, R., Traina, P., and T. Li, "BGP
              Communities Attribute", RFC 1997, August 1996.

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

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

   [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private
              Networks (VPNs)", RFC 4364, February 2006.

Authors' Addresses

    James Uttaro
    AT&T
    200 S. Laurel Avenue
    Middletown, NJ 07748
    USA

    Email: ju1738@att.com

    Adam Simpson
    Alcatel-Lucent
    600 March Road
    Ottawa, Ontario K2K 2E6
    Canada

    Email: adam.simpson@alcatel-lucent.com

    Rob Shakir
    Cable&Wireless Worldwide
    London
    UK

    Email: rjs@cw.net
    URI: http://www.cw.com/

    Clarence Filsfils
    Cisco Systems
    Brussels 1000
    BE

    Email: cf@cisco.com

    Pradosh Mohapatra
    Cisco Systems
    170 W. Tasman Drive
    San Jose, CA 95134
    USA

    Email: pmohapat@cisco.com

Bruno Decraene
France Telecom
38-40 Rue de General Leclerc
92794 Issy Moulineaux cedex 9
France

Email: bruno.decraene@orange.com

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

Email: jgs@juniper.net

Yakov Rekhter
Juniper Networks

Email: yakov@juniper.net

                 Carrying next-hop cost information in BGP
                     draft-varlashkin-bgp-nh-cost-02

Abstract

   This document describes new BGP SAFI to exchange cost information to
   next-hops for the purpose of calculating best path from a peer
   perspective rather than local BGP speaker own perspective.

Table of Contents

1. Motivation

    In certain situation route-reflector clients may not get optimum path
    to certain destinations. ADDPATH solves this problem by letting
    route-reflector to advertise multiple paths for given prefix. If
    number of advertised paths sufficiently big, route-reflector clients
    can choose same route as they would in case of full-mesh. This
    approach however places additional burden on the control plane.
    Solutions proposed by [BGP-ORR] use different approach - instead of
    calculating best path from local speaker own perspective the
    calculations are done using cost from the client to the next-hops.
    Although they eliminate need for transmitting redundant routing
    information between peers, there are scenarios where cost to the
    next-hop cannot be obtained accurately using this methods. For
    example, if next-hop information itself has been learned via BGP then
    simple SPF run on link-state database won't be sufficient to obtain
    cost information. To address such scenarios this document proposes a
    solution where cost information to the next-hops is carried within
    BGP itself using dedicated SAFI.

2. NEXT-HOP INFORMATION BASE

    To facilitate further description of the proposed solution we
    introduce new table for all known next hops and costs to it from
    various routers on the network.

    Next-Hop Information Base (NHIB) stores cost to reach next-hop from
    arbitrary router on the network. This information is essential for
    choosing best path from a peer perspective rather than BGP-speaker
    own perspective. In canonical form NHIB entry is triplet (router,
    next-hop, cost), however this specification does not impose any
    restriction on how BGP implementations store that information
    internally. The cost in NHIB is does not have to be an IGP cost, but
    all costs in NHIB MUST be comparable with each other.

    NHIB can be populated from various sources both static and dynamic.
    This document focuses on populating NHIB using BGP. However it is
    possible that protocols other than BGP could be also used to populate
    NHIB.

3. BGP BEST PATH SELECTION MODIFICATION

    This section applies regardless of method used to populate NHIB.

    When BGP speaker conforming to this specification selects routes to
    be advertised to a peer it SHOULD use cost information from NHIB

rather than its own IGP cost to the next-hop after step (d) of
9.1.2.2 in [RFC4271].

4.   USING BGP TO POPULATE NHIB

   This section describes extension to base BGP specification that
   allows BGP to be used for exchanging next-hop information between BGP
   speakers via new SAFI in order to populate NHIB. Although next-hops
   costs are exchanged via dedicated SAFI, this information is vital to
   best path selection process for other AFI/SAFI (e.g. IPv4 and IPv6
   unicast). It's therefore recommended that next-hop cost information
   is exchanged before other AFI/SAFI.

4.1.   NEXT-HOP SAFI

   This document introduces Next-Hop SAFI (NH SAFI) with value to be
   assigned by IANA and purpose of exchanging information about cost to
   next-hops.

4.2.   CAPABILITY ADVERTISEMENT

   A BGP speaker willing to exchange next-hop information MUST advertise
   this in the OPEN message using BGP Capability Code 1 (Multiprotocol
   Extensions, see [RFC4760]) setting AFI appropriately to indicate IPv4
   or IPv6 and SAFI to the value assigned by IANA for NH SAFI. Note
   that if BGP speaker whishes to exchange cost information for both
   IPv4 and IPv6, then it MUST advertise two capabilities: one NH SAFI
   for IPv4 and one NH SAFI for IPv6.

4.3.   INFORMATION ENCODING

   To request cost to a next-hop from peer or to inform peer about cost
   to a next-hop BGP attribute 14 is used as follow:

   1.   AFI is set to indicate IPv4 or IPv6 (whichever is appropriate)

   2.   SAFI is set to NH SAFI

   3.   Network Address of Next-Hop field is zeroed out

   4.   NLRI field is encoded as shown in the next figure

      +-------------+------------+
      | NEXT_HOP | cost |
      +-------------+------------+

   Where cost is 32-bit unsigned integer (value described below), and

NEXT_HOP is AFI-specific address of the next-hop cost to which is
being communicated or requested. Size of NEXT_HOP field is inferred
from total length of attribute 14.

To request cost to arbitrary next-hop from a peer, BGP speaker sets
cost field to zero.

To inform peer about cost to a next-hop BGP speaker sets cost to
actual cost value.

To inform peer that a next-hop is not reachable the cost is set to
all-ones (0xFFFFFFFF).

4.4. SESSION ESTABLISHMENT

BGP speakers willing to exchange next-hop information SHOULD NOT
establish more then one session for given AFI and NH SAFI, even using
different transport addresses. This can be ensured for example by
checking peer's Router Id.

4.5. INFORMATION EXCHANGE

Typically NH SAFI sessions will be established between route-
reflectors and its internal peers (both clients and non-clients). As
soon as the NH SAFI session is ESTABLISHED requests for next-hop cost
and information information about next-hop costs MAY be sent
independently. That is, route-reflector MAY send multiple requests
without waiting for response, and its peers MAY send cost information
before or after receiving such request. On the other hand, Router
Reflectors SHOULD request cost information from their internal peers
as soon as possible (due to reasons stated in section "BGP best path
selection modification"). BGP speaker does not need to track
outstanding requests to the peer.

When a BGP speaker receives request for cost information it MUST
reply with actual cost (not necessarily IGP cost, but whatever has
been chosen to be carried in NH SAFI) to given next-hop or with cost
set to all-ones indicating that next-hop is unreachable.

Note that BGP speaker MUST use longest match rather than exact match
for the next-hop.

When a BGP speaker detects change in cost to previously advertised
next-hop with delta equal or exceeding configured advertisement
threshold, it SHOULD inform peer by advertising new cost or
0xFFFFFFFF.

When a BGP speaker discovers new next-hop among candidate routes it

   SHOULD request cost information from the peer.

4.6. TERMINATION OF NH SAFI SESSION

   When BGP speaker terminates (for whatever reason) NH SAFI session
   with a peer, it SHOULD remove all cost information received from that
   peer unless instructed by configuration to do otherwise.

4.7. GRACEFUL RESTART AND ROUTE REFRESH

   NH SAFI sessions could use graceful restart and route refresh
   mechanisms in the same way as it's used for IPv4 and IPv6 unicast.

5. Security considerations

   No new security issues are introduced to the BGP protocol by this
   specification.

6. IANA Considerations

   IANA is requested to allocate value for Next-Hop Subsequent Address
   Family Identifier.

7. References

7.1. Normative References

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

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

7.2. Informative References

   [I-D.raszuk-bgp-optimal-route-reflection]
              Raszuk, R., Cassar, C., Aman, E., and B. Decraene, "BGP
              Optimal Route Reflection (BGP-ORR)",
              draft-raszuk-bgp-optimal-route-reflection-01 (work in
              progress), March 2011.

   [RFC2918] Chen, E., "Route Refresh Capability for BGP-4", RFC 2918,
              September 2000.

Appendix A.   USAGE SCENARIOS

A.1. Trivial case

```
    --+---NetA---+--
      |          |
     r1          r2
      |          |
      R1--RR-----R2
      | \        |
      |  +------R4
     R3
```
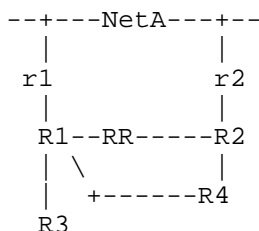
   In this scenario r1 and r3 along with NetA are part of AS1; and R1-R4
   along with RR are in AS2.

   If RR implements non-optimized route-reflection, then it will choose
   path to NetA via R1 and advertise it to both R3 and R4. Such choice
   is good from R3 perspective, but it results in suboptimal traffic
   flow from R4 to NetA.

   Using NH SAFI the route-reflector will learn that cost from R4 to R1
   is 8 whereas to R2 it's only 1. RR will announce NetA to R4 with
   next-hop set to R2, while its announce to R3 will still have R1 as
   next-hop. Both R3 and R4 now will send traffic to NetA via closest
   exit, achieving same behaviour as if full iBGP mesh would have been
   configured.

A.2. Non-IGP based cost

   When it's desirable to direct traffic over an exit other than the one
   with smallest IGP cost, NH SAFI can be used to convey cost which is
   not based on IGP. For example, network operator may arrange exit
   points in order of administrative preference and configure routers to
   send this instead of IGP cost. Route reflector then will then
   calculate best path based on administrative preference rather than
   IGP metrics.

   Network operators should excercise care to ensure that all routers up
   to and including exit point do not devert packets on to a different
   path, otherwise routing loops may occur. One way to achieve this is
   to have consistent administrative preference among all routers.
   Another option is to use a tunneling mechanism (e.g. MPLS-TE tunnel)
   between source and the exit point, provided that the router serving
   as exit point will send packets out of the network rather than
   diverting them to another exit point.

A.3. Multiple route-reflectors

   This example demonstrates that NH SAFI peerings are necessary only
   between routers that already exchange other AFI/SAFI.

```
                           |
    R1----R3---------R5----R7--+
         |           |         |
        RR1          |        NetA
         |          RR2        |
         |           |         |
    R2----R4---------R6----R8--+
                               |
```
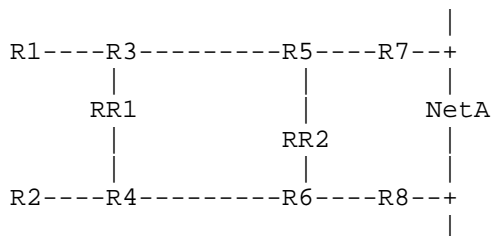
   In the above network the routers R1-R4 are clients of RR1, and R5-R8
   are clients of RR2. RR1 and RR2 also peer with each other and use
   ADDPATH.

   RR2 learns about NetA from R7 and R8. Since it sends not just best-
   path but all prefixes to RR1, there is no need for RR2 to learn cost
   information from R1 and R2 towards R7 and R8. On the other hand RR1
   does exchange NH SAFI information with R1 and R2 so that each of them
   can receive routes, which are best from their perspective.

   As addition to ADDPATH a mechanism could be devised that would allow
   RR2 to learn how many alternative routes does it need to send to RR1.
   For example, if NetA would also be connected to R9 (not shown) but
   all clients of RR1 prefer R7 as exit point and R9 as next-best, then
   there is no need for RR2 to send NetA routes with next-hop R8 to RR1.

   Discussion: authors would like to solicit discussion whether there is
   sufficient interest in such mechanism.

A.4. Inter-AS MPLS VPN

   Previous example could be transposed to Inter-AS MPLS VPN Option C
   scenario. In this case route reflectors RR1 and RR2 can be from
   different autonomous system. Essentially the behaviour of routers
   remains as already described.

A.5. Corner case

```
       --+---NetA--+--
         |         |
    RR---R1        R2
          \       /
          R3---R4
```
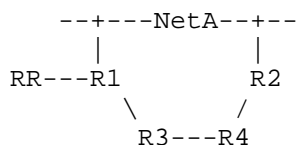
In the above network cost from R3 to R1 is 10, all other costs are 1.
If RR advertises NetA to R3 based on cost information received from
R3, but uses its own cost when advertising NetA to R4, there will be
a loop formed. This is the reason why section "BGP best path
selection modification" requires RR to have next-hop cost information
for every next-hop and every peer.

Note that the problem is the same as if RR would not use extensions
described in this document and R3 would peer directly with R1 and R2,
while R4 would peer only with RR.

Authors' Addresses

    Ilya Varlashkin
    Easynet Global Services

    Email: ilya.varlashkin@easynet.com

    Robert Raszuk
    NTT MCL Inc.
    101 S Ellsworth Avenue Suite 350
    San Mateo, CA 94401
    US

    Email: robert@raszuk.net

            Maximum Transmission Unit Extended Community for BGP-4
                  draft-zeng-idr-bgp-mtu-extension-01

Abstract

   Proper functioning of path Maximum Transmission Unit (MTU) discovery
   [RFC1191] requires that IP routers have knowledge of the MTU for each
   link to which they are connected. As MPLS progresses, [RFC3988]
   specifies extensions to LDP in support of LDP LSP MTU discovery. For
   the LSP created using Border Gateway Protocol (BGP) [RFC3107], it
   does not have the ability to signal the path MTU to the ingress Label
   Switching Router (LSR). In the absence of this functionality, the
   MTU for the BGP LSP must be statically configured by network
   operators or by equivalent off-line mechanisms.

   This document defines the MTU Extended Community for BGP in support
   of BGP LSP MTU discovery.

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 May 3, 2012.

Copyright Notice

Table of Contents

1. Introduction

   Proper functioning of [RFC1191] path Maximum Transmission Unit (MTU)
   discovery requires that IP routers have knowledge of the MTU for each
   link to which they are connected. As MPLS progresses, [RFC3988]
   specifies some extensions to LDP in support of LDP LSP MTU discovery.
   For the LSP created using Border Gateway Protocol (BGP) [RFC3107], it
   does not have the ability to signal the path MTU to the ingress Label
   Switching Router (LSR). Without knowledge of the path MTU of the
   whole BGP LSP, ingress BGP LSRs may transmit packets along that LSP
   which are either too big or too small, thus these packets may either
   be silently discarded by LSRs or be transmitted inefficiently. In
   the absence of MTU discovery functionality, the MTU for each BGP LSP
   must be statically configured by network operators or by equivalent
   off-line mechanisms.

   This document defines the MTU Extended Community for BGP in support
   of BGP LSP MTU discovery.

2. Problem Statement

   For some inter-AS services and also for network scalability, the LSPs
   need to be established using Labeled BGP [RFC3107]. Typical
   scenarios include inter-AS VPN Option C, Carrier's Carrier [RFC4364]
   and Seamless MPLS [I-D.ietf-mpls-seamless-mpls].

   Taking "Inter-AS IP VPN Option C" as an example. An ASBR must
   maintain labeled IPv4 /32 routes to the PE routers within its AS.
   And it uses EBGP to distribute these labeled /32 routes to other ASes
   using mechanism in [RFC3107]. ASBRs in transit ASes will also use
   BGP to pass along the labeled /32 routes. In the AS of ingress PEs
   (from data plane perspective), the labeled /32 routes can be
   distributed to the PE routers using IBGP. The /32 routes may also be
   redistributed into IGP of the Ingress AS (from data plane
   perspective). Intra-AS LSPs between the PE nodes and ASBRs can be
   established using LDP [RFC5036] or RSVP-TE [RFC3209].

   For intra-AS LSPs established using LDP or RSVP-TE, Path MTU of the
   LSP could be discovered using mechanisms defined in [RFC3988] and
   [RFC3209] respectively. But for the inter-AS LSP which is
   established using BGP, some mechanism is needed to discover the Path
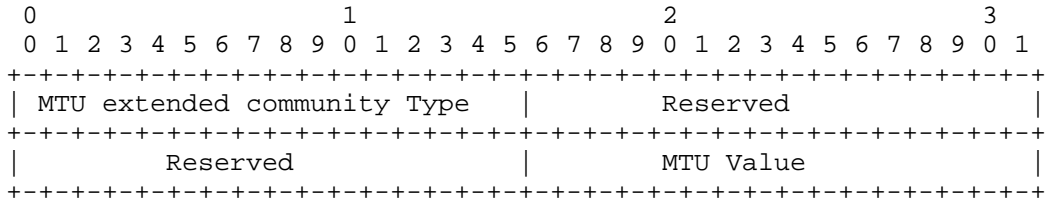   MTU.

3. BGP LSP MTU Discovery

3.1. Definitions

   BGP LSP Path MTU: The Path MTU of the LSP from a given BGP LSR to a
   specific prefix. It is carried as a Extended Community with the BGP
   labeled IPv4 (or IPv6) route. This size includes the IP header and
   data (or other payload) and the part of the label stack that is
   considered payload of this BGP LSP.

   BGP LSR Link MTU: If the two BGP LSRs are directly adjacent, the BGP
   LSR Link MTU is the interface MTU; If the two BGP LSRs are not
   directly adjacent, the BGP LSR Link MTU is the Path MTU of the
   underlying tunnel. If there are multiple links between the two BGP
   LSRs, the BGP LSR Link MTU is the minimum of those link MTUs.

3.2. MTU Extended Community

   BGP LSP Path MTU is carried in the MTU extended community for BGP-4.
   The MTU extended community is an optional transitive attribute.

```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   | MTU extended community Type   |            Reserved           |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |           Reserved            |           MTU Value           |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
   The MTU extended community type is to be assigned by IANA. The first
   four octets of the value field should be reserved, and the MTU value
   is carried in the following two octets of the value field.

3.3. Signaling

   The MTU is advertised hop-by-hop from BGP egress LSR to BGP ingress
   LSR along an BGP LSP. The steps are as follows:

   A. If BGP speaker A is the originator of the labeled BGP route, and
   there is a intra-AS LSP to the prefix, A SHOULD set its BGP LSP Path
   MTU to the path MTU value it has discovered to this prefix, and
   advertise the labeled BGP route with the MTU Extended Community to
   its BGP Peer (its upstream BGP LSR). If the prefix belongs to BGP
   speaker A, the BGP LSP Path MTU SHOULD be set to 65535.

   B. BGP speaker B receives the labeled BGP route with BGP LSP Path MTU
   from its BGP peer.

   a) B SHOULD compute the BGP LSR Link MTU to the Next Hop of the
   received message, then sets its BGP LSP Path MTU to the minimum of
   the received BGP LSP Path MTU and (the BGP LSR Link MTU - 4 octets).

b). If B distributes the route with the Next Hop attribute
unchanged, it MUST keep the MTU Extended Community unchanged when
advertising the message to its upstream BGP LSRs.

c). If B would change the Next Hop attribute to itself in the
subsequent advertisement, it SHOULD set the MTU Extended Community in
the message with its BGP LSP Path MTU obtained through step a).

3.4. Considerations on Route Flapping

Normally change of BGP path attributes would result in advertising a
BGP update for the route. In order to throttle the route updates
caused by changes of BGP path MTU , this section specifies rules of
route update when BGP LSP Path MTU changes:

1. If the BGP LSP Path MTU decreases, a new update SHOULD be
advertised immediately;

2. If the BGP LSP Path MTU increases, the BGP speaker MAY hold down
the update until there are changes of some other BGP attributes.

3.5. BGP LSP and LDP LSP Stitching

In scenarios where the labeled BGP routes are redistributed into IGP
on a border router and an LDP LSP is established and stitched to the
BGP LSP, the border router SHOULD use its BGP path MTU as the LDP LSP
MTU, and the path MTU discovery of the LDP LSP will be performed
according to [RFC3988].

4. Applicability Considerations

The BGP MTU Extended Community is applicable to labeled BGP defined
in [RFC3107]. The application of BGP MTU Discovery may also be used
for other inter-AS/inter-area routing scenarios. Such use cases are
for further study.

5. IANA Considerations

IANA is requested to assign a type and sub-type value for BGP MTU
extended community.

6. Security Considerations

This extension to BGP does not change the underlying security issues
in [RFC4271].

7. Contributors

   The following individuals contributed to this document:

   Haibo Wang rainsword.wang@huawei.com

   Haijun Xu xuhaijun@huawei.com

8. Acknowledgements

   The authors would like to thank Jeff Haas, Nagendra Kumar and David
   Freedman for their valuable discussions and suggestions.

9. References

9.1. Normative References

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

   [RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in
              BGP-4", RFC 3107, May 2001.

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

   [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended
              Communities Attribute", RFC 4360, February 2006.

9.2. Informative References

   [I-D.ietf-mpls-seamless-mpls]
              Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz,
              M., and D. Steinberg, "Seamless MPLS Architecture",
              draft-ietf-mpls-seamless-mpls-00 (work in progress),
              May 2011.

   [RFC1191] Mogul, J. and S. Deering, "Path MTU discovery", RFC 1191,
              November 1990.

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

   [RFC3988] Black, B. and K. Kompella, "Maximum Transmission Unit
              Signalling Extensions for the Label Distribution

              Protocol", RFC 3988, January 2005.

   [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private
              Networks (VPNs)", RFC 4364, February 2006.

   [RFC4659] De Clercq, J., Ooms, D., Carugi, M., and F. Le Faucheur,
              "BGP-MPLS IP Virtual Private Network (VPN) Extension for
              IPv6 VPN", RFC 4659, September 2006.

Authors' Addresses

   Qing Zeng
   Huawei Technologies
   Huawei Building, No.156 Beiqing Rd.
   Beijing 100095
   China

   Email: zengqing@huawei.com

   Jie Dong
   Huawei Technologies
   Huawei Building, No.156 Beiqing Rd.
   Beijing 100095
   China

   Email: jie.dong@huawei.com