Inter-Domain Routing
Internet-Draft
Intended status: Standards Track
Expires: September 4, 2011

H. Gredler
J. Medved
Juniper Networks, Inc.

March 3, 2011

Advertising Traffic Engineering Information in BGP
draft-gredler-bgp-te-00

Abstract

This document defines a new Border Gateway Protocol Network Layer Reachability Information (BGP NLRI) encoding format that can be used to distribute Traffic Engineering (TE) link information. Links can be either physical links connecting physical nodes, or virtual paths between physical or abstract nodes. The TE 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 Traffic Engineering (TE) 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 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 will expire on September 4, 2011.

Copyright Notice

Table of Contents

1. Introduction

Today, the contents of the traffic engineering database usually has the scope of an IGP area. There are several use cases that could benefit from knowing the topology or Traffic Engineering (TE) data 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 traffic engineering 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 Traffic Engineering data is a logical consequence.

A router maintains a database for storing Traffic Engineering related data and link information. The Traffic Engineering Database (TED) is populated by a link-state IGP routing protocol that supports TE extensions: IS-IS or OSPF. The TED can be seen as a protocol-neutral representation of links in the area. Link attributes stored in the TED 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 the TE data from the TED database 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 TED, 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 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 TED 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 dynamic bandwidth or delay metrics, is out of scope of this draft.

3. Transcoding TE Link Information Into a BGP NLRI

The MP_REACH and MP_UNREACH attributes are BGP's containers for carrying opaque information. Each TED NLRI describes a single link 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 anchoring Router-IDs are carried in the Node Anchor TLVs.

All TE link 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 TE 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.

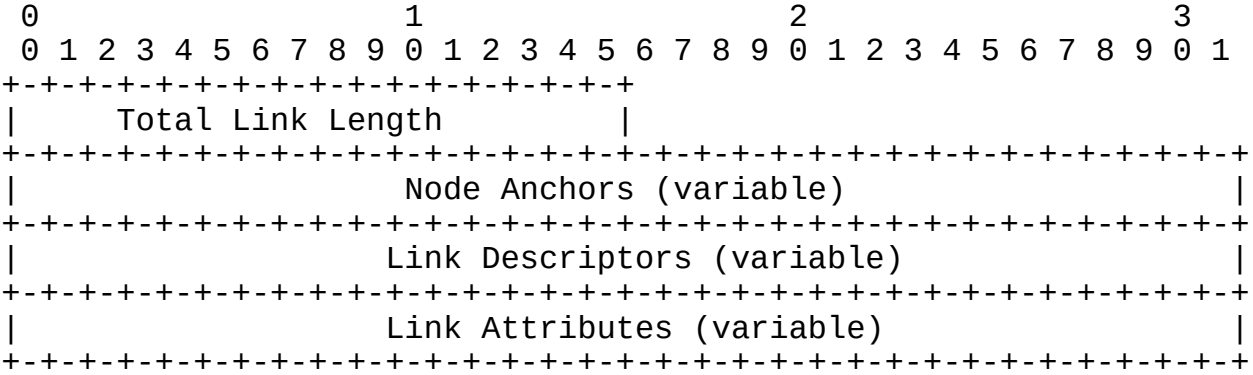

Figure 1: TED SAFI 1 NLRI Format

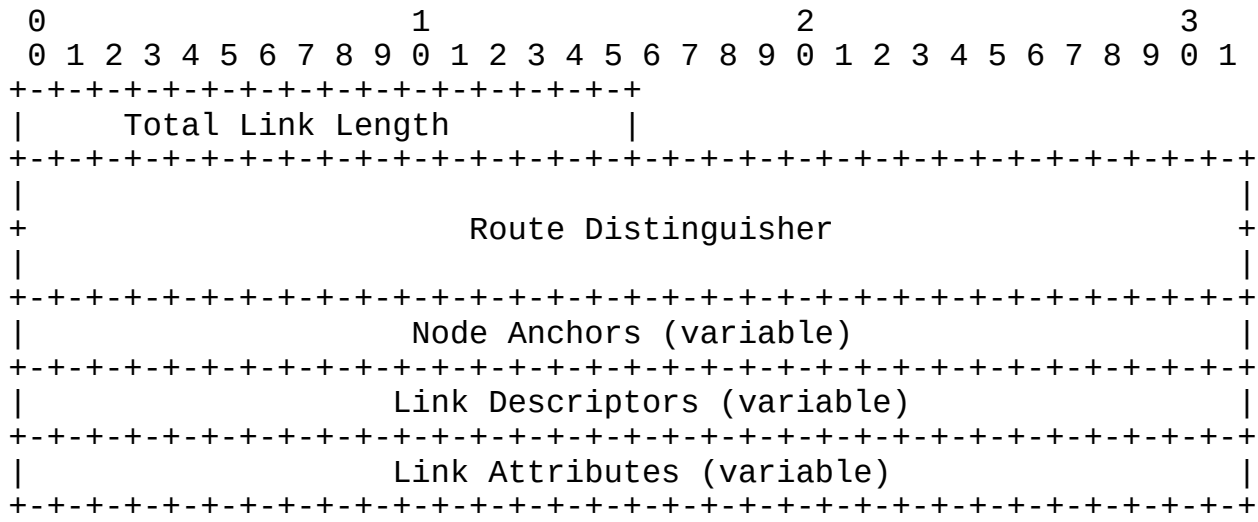

Figure 2: TED SAFI 128 NLRI Format

The 'Total Link Length" field contains the cumulative length of all the TLVs, describing the Node Anchors, Link descriptors and Link Attributes. For VPN applications it also includes the length of the Route Distinguisher.

3.1. TLV Format

The Node anchor, Link descriptor and Link attribute fields are described using a set of Type/Length/Value triplets. The format of each TLV is shown in Figure 3

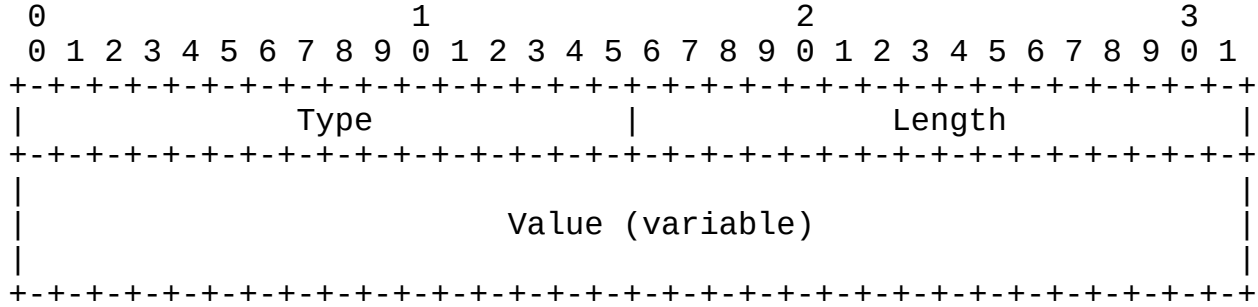

Figure 3: TLV format

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

3.2. Node anchors

The set of Node Anchor TLVs describes 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 per-protocol. If a peer sends an illegal combination in this respect, then this is handled as an NLRI error, described in [RFC4760].

Type	Description	Length
256	Local Autonomous System	4
257	Local IPv4 Router-ID	4
258	Local IPv6 Router-ID	16
259	Local ISO Node-ID	7
260	Remote Autonomous System	4
261	Remote IPv4 Router-ID	4
262	Remote IPv6 Router-ID	16
263	Remote ISO Node-ID	7

Table 1: Node Anchor TLVs

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

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

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

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

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

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

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.

3.2.1. Router-ID Anchoring Example: ISO Pseudonode

IS-IS Pseudonodes are a good example for the variable Router-ID anchoring. Consider Figure 4. 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.

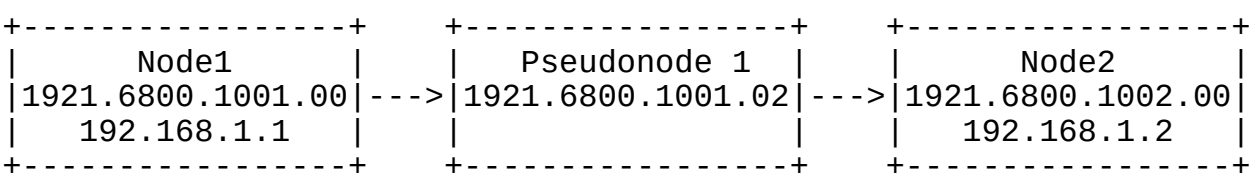

Figure 4: IS-IS Pseudonodes

3.2.2. 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) does support 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 does encode local IPv4 router-id, remote IPv4 router-id, remote ISO node-id and remote IPv6 node-id.

3.3. Link Descriptors

The ’Link Descriptor’ field is a set of Type/Length/Value (TLV) triplets. The format of each TLV is shown in Figure 3. 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 TED 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

3.4. Link Attributes

The 'Link Attributes' field is a set of Type/Length/Value (TLV) triplets. The format of each TLV is shown in Figure 3.

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 TED 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
64512	TE Default Metric	Section 3.4.1
64513	IGP Link Metric	Section 3.4.2
64514	Shared Risk Link Group	Section 3.4.3

Table 3: Link Attribute TLVs

3.4.1. 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.

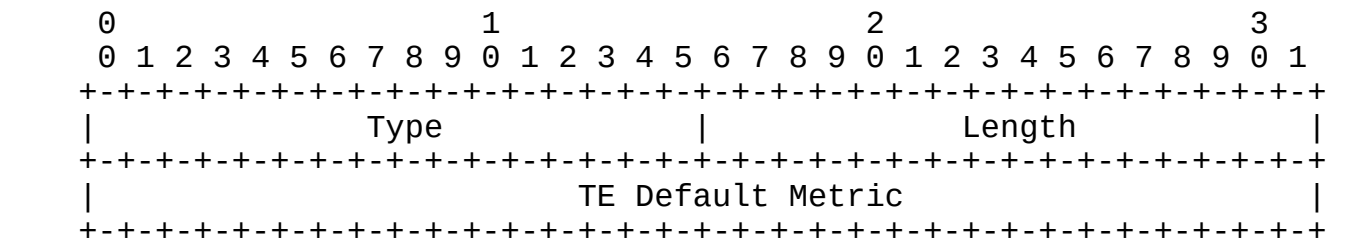

Figure 5: TE Default metric TLV format

3.4.2. 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.

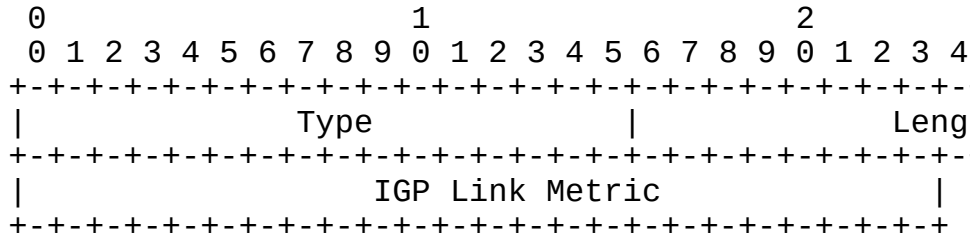

Figure 6: IGP Link Metric TLV format

3.4.3. 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 7. The length of this TLV is 4 * (number of SRLG values).

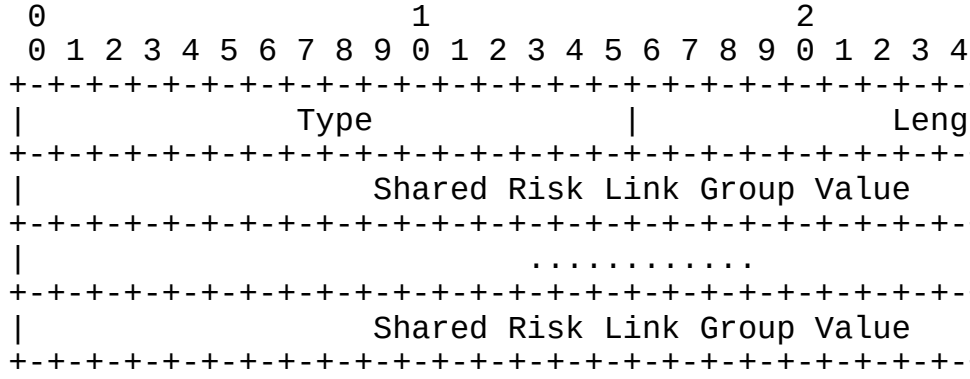

Figure 7: Shared Risk Link Group TLV format

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 BGP TED NLRI uses variable Router-ID anchoring, both IPv4 and IPv6 SRLG information can be carried in a single TLV.

3.5. IGP Area Information

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

3.6. 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 traffic-engineering database (TED) an implementation must support configuration of static TE 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 8. Both AS1 and AS2 operators want to protect their inter-AS {R1,R3}, {R2, R4} links using RSVP-FRR LSPs. If R1 wants to compute its link-protection LSP to R3 it needs to "see" an alternate path to R3. Therefore the AS2 operator exposes its topology. All BGP TE enabled routers in AS1 "see" the full topology of AS and therefore can compute a backup path. Note that the decision if the direct link between {R3, R4} or the {R4, R5, R3) path is used is made by the computing router.

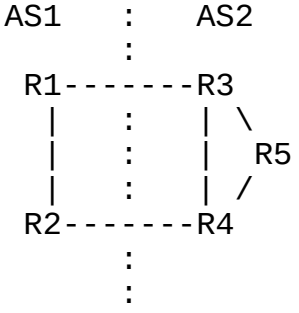

Figure 8: no-link-aggregation

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

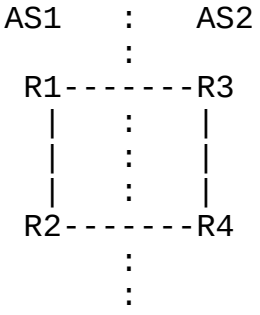

Figure 9: asbr-link-aggregation

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 10. 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.

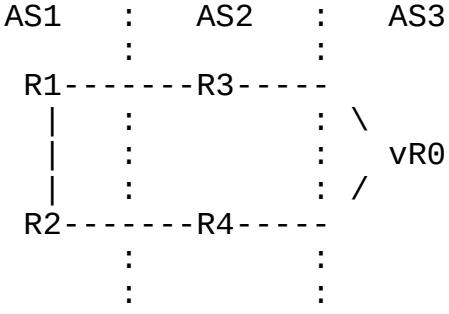

Figure 10: multi-as-aggregation

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 11.

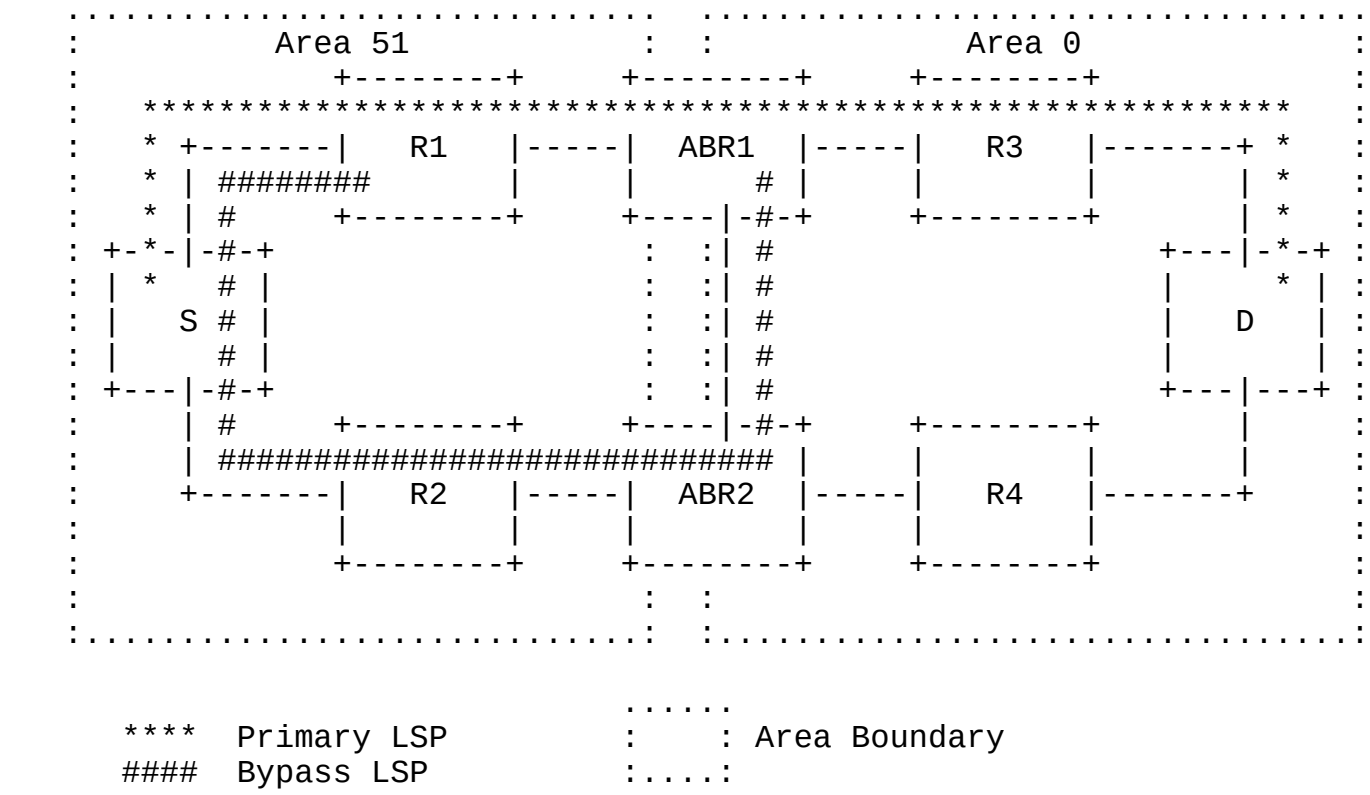

Figure 11: MPLS TE Bypass LSP problem

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 Alia Atlas, David Ward, John Scudder, Kaliraj Vairavakkalai, Nischal Sheth and Yakov Rekhter from Juniper Networks, Inc. and Richard Woundy from Comcast for their invaluable input and 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>.

[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.

[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.

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

[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-06 (work in progress), October 2010.

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