

Inter-Domain Routing
Internet-Draft
Intended status: Standards Track
Expires: January 16, 2013

H. Gredler
Juniper Networks, Inc.
J. Medved
S. Previdi
Cisco Systems, Inc.
A. Farrel
Juniper Networks, Inc.
July 15, 2012

North-Bound Distribution of Link-State and TE Information using BGP
draft-gredler-idr-ls-distribution-02

Abstract

In a number of environments, a component external to a network is called upon to perform computations based on the network topology and current state of the connections within the network, including traffic engineering 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]

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 16, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	4
2. Motivation and Applicability	5
2.1. MPLS-TE with PCE	5
2.2. ALTO Server Network API	7
3. Carrying Link State Information in BGP	8
3.1. TLV Format	8
3.2. The Link State NLRI	9
3.2.1. Node Descriptors	11
3.2.2. Link Descriptors	14
3.3. The LINK_STATE Attribute	15
3.3.1. Link Attribute TLVs	15
3.3.2. Node Attribute TLVs	19
3.4. Inter-AS Links	22
4. Link to Path Aggregation	22
4.1. Example: No Link Aggregation	23
4.2. Example: ASBR to ASBR Path Aggregation	23
4.3. Example: Multi-AS Path Aggregation	24
5. IANA Considerations	24
6. Manageability Considerations	25
6.1. Operational Considerations	25
6.1.1. Operations	25
6.1.2. Installation and Initial Setup	25
6.1.3. Migration Path	25
6.1.4. Requirements on Other Protocols and Functional Components	25
6.1.5. Impact on Network Operation	26
6.1.6. Verifying Correct Operation	26
6.2. Management Considerations	26
6.2.1. Management Information	26
6.2.2. Fault Management	26
6.2.3. Configuration Management	26
6.2.4. Accounting Management	27
6.2.5. Performance Management	27
6.2.6. Security Management	27
7. Security Considerations	27
8. Acknowledgements	27
9. References	28
9.1. Normative References	28
9.2. Informative References	29
Authors' Addresses	30

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.

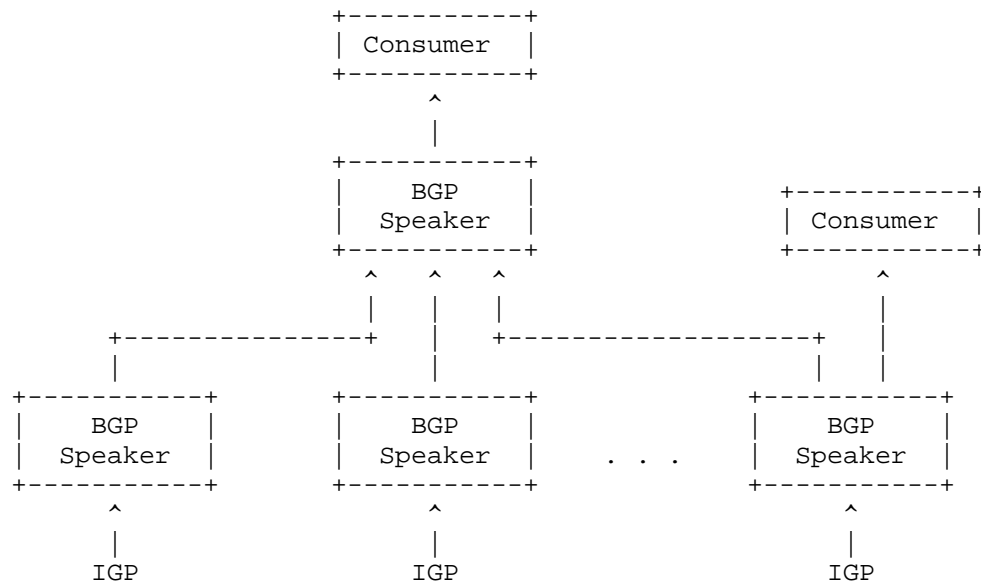


Figure 1: TE Link State info collection

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

2. Motivation and Applicability

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

2.1. MPLS-TE with PCE

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

- 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.

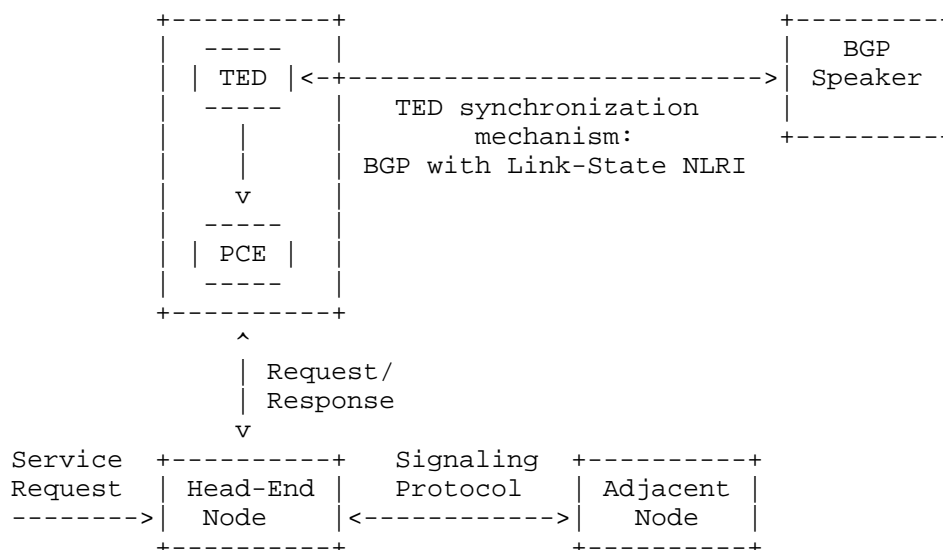


Figure 2: External PCE node using a TED synchronization mechanism

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.

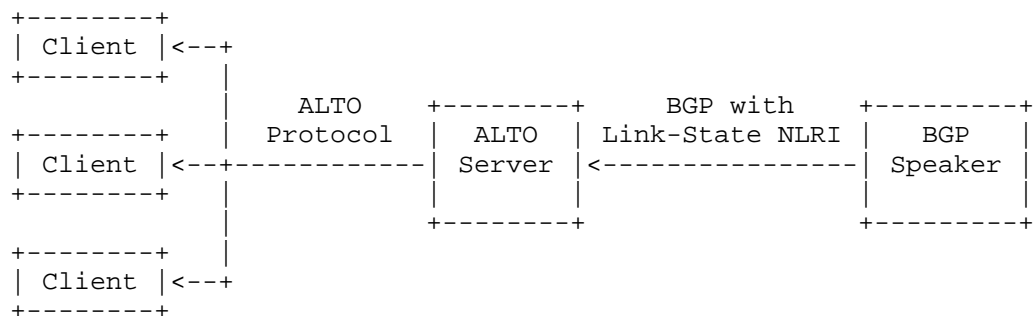


Figure 3: ALTO Server using network topology information

3. Carrying Link State Information in BGP

Two parts: a new BGP NLRI that describes links and nodes comprising IGP link state information, and a new BGP path attribute that carries link and node properties and attributes, such as the link metric or node properties.

3.1. TLV Format

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

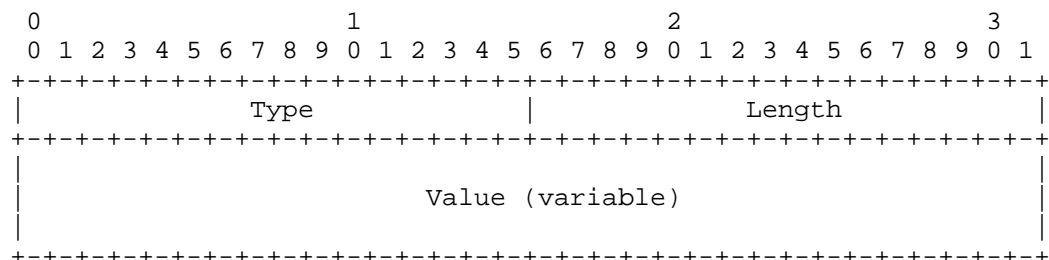


Figure 4: TLV format

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

3.2. The Link State 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.

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

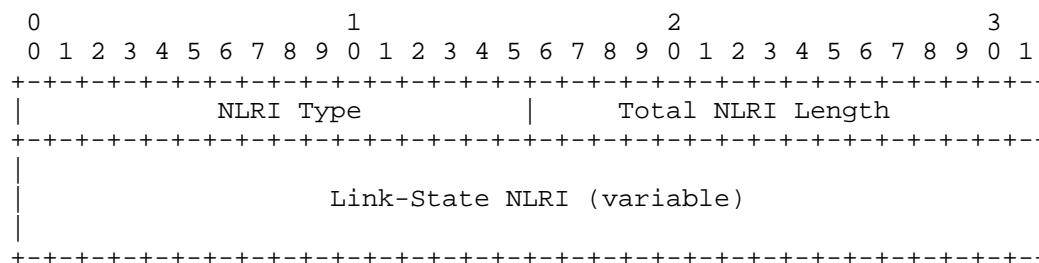


Figure 5: Link State SAFI 1 NLRI Format

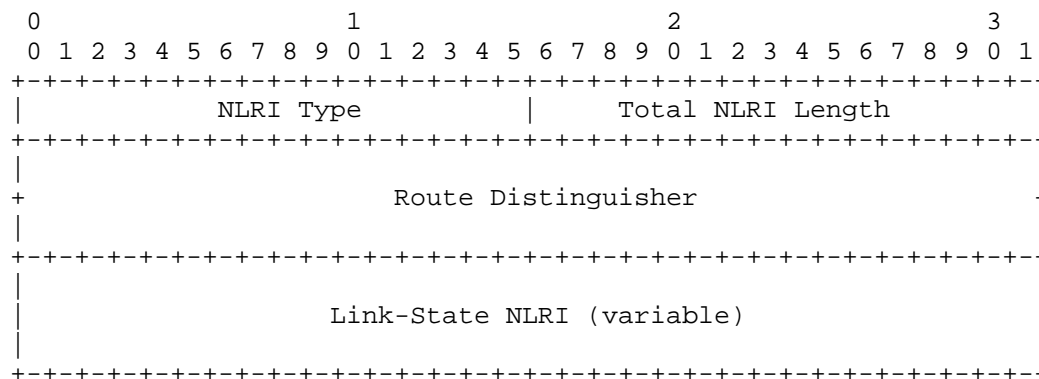


Figure 6: Link State SAFI 128 NLRI Format

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.

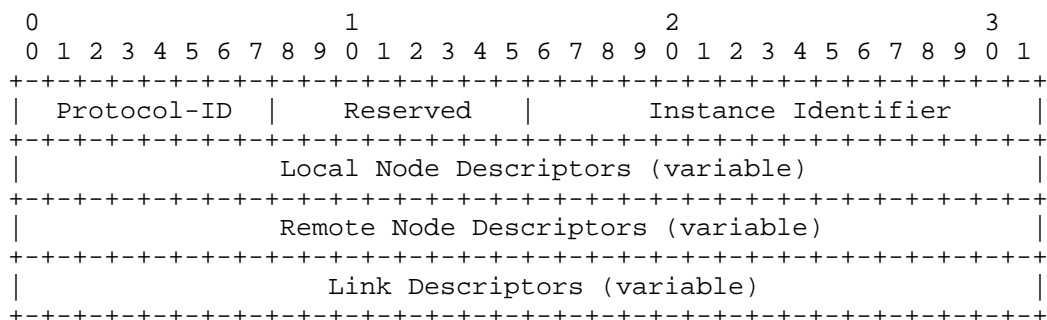


Figure 7: The Link NLRI format

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

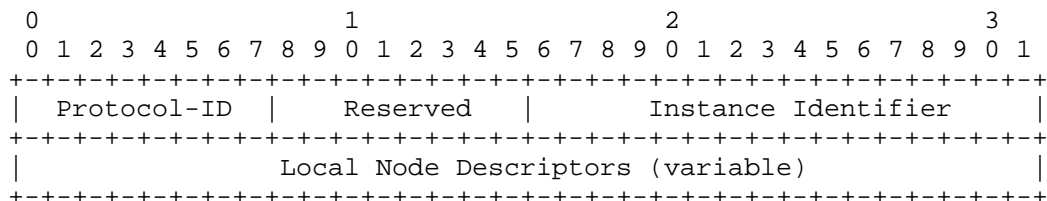


Figure 8: The Node NLRI format

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

Type = 4: Direct, The NLRI information has been sourced from local interface state

Type = 5: Static, The NLRI information has been sourced by static configuration

Both OSPF and IS-IS may run multiple routing protocol instances over the same link. See [I-D.ietf-isis-mi] and [RFC6549]. The 'Instance Identifier' field identifies the protocol instance.

Each Node Descriptor and Link Descriptor consists of one or more TLVs described in the following sections. The sender of an UPDATE message MUST order the TLVs within a Node Descriptor or a Link Descriptor in ascending order of TLV type."

3.2.1. 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. If the anchor node's AS is a member of an AS Confederation ([RFC5065]), then the Autonomous System number TLVs contains the confederations' AS Confederation Identifier and the Member-AS TLV is 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.

Type	Description	Length
258	Autonomous System	4
259	Member-AS	4
260	IPv4 Router-ID	5
261	IPv6 Router-ID	17
262	ISO Node-ID	7

Table 1: Node Descriptor Sub-TLVs

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

Autonomous System: opaque value (32 Bit AS ID)

Member-AS: opaque value (32 Bit AS ID); only included if the node is in an AS confederation.

IPv4 Router ID: opaque value (can be an IPv4 address or an 32 Bit router ID) followed by a LAN-ID octet in case LAN "Pseudonode" information gets advertised. The PSN octet must be zero for non-LAN "Pseudonodes".

IPv6 Router ID: opaque value (can be an IPv6 address or 128 Bit router ID) followed by a LAN-ID octet in case LAN "Pseudonode" information gets advertised. The PSN octet must be zero for non-LAN "Pseudonodes".

ISO Node ID: ISO node-ID (6 octets ISO system-ID) followed by a PSN octet in case LAN "Pseudonode" information gets advertised. The PSN octet must be zero for non-LAN "Pseudonodes".

3.2.1.4. 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.

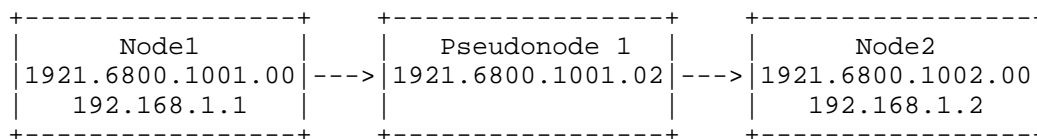


Figure 11: IS-IS Pseudonodes

3.2.1.5. 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.

3.2.2. Link Descriptors

The 'Link Descriptor' field is a set of Type/Length/Value (TLV) triplets. The format of each TLV is shown in Section 3.1. The 'Link descriptor' TLVs uniquely identify a link between a pair of anchor Routers. A link described by the Link descriptor TLVs actually is a "half-link", a unidirectional representation of a logical link. In order to fully describe a single logical link two originating routers need to advertise a half-link each, i.e. two link NLRI's will be advertised.

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

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

Type	Description	IS-IS TLV/Sub-TLV	Value defined in:
263	Link Local/Remote Identifiers	22/4	[RFC5307]/1.1
264	IPv4 interface address	22/6	[RFC5305]/3.2
265	IPv4 neighbor address	22/8	[RFC5305]/3.3
266	IPv6 interface address	22/12	[RFC6119]/4.2
267	IPv6 neighbor address	22/13	[RFC6119]/4.3
268	Multi Topology ID	---	Section 3.2.2.1

Table 2: Link Descriptor TLVs

3.2.2.1. Multi Topology ID TLV

The Multi Topology ID TLV (Type 268) 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.

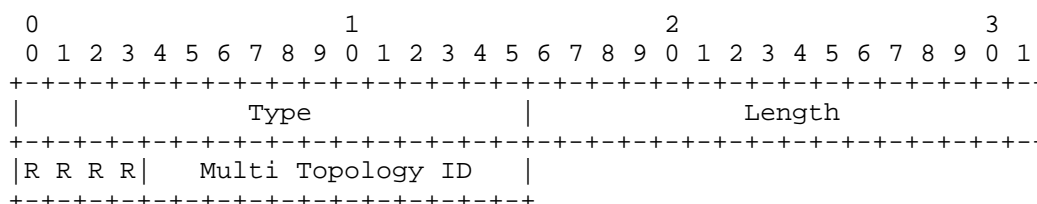


Figure 12: Multi Topology ID TLV format

3.3. The LINK_STATE Attribute

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

3.3.1. Link Attribute TLVs

Each 'Link Attribute' is a Type/Length/Value (TLV) triplet formatted as defined in Section 3.1. The format and semantics of the 'value' fields in some 'Link Attribute' TLVs correspond to the format and

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

The following 'Link Attribute' TLVs are are valid in the LINK_STATE attribute:

Type	Description	IS-IS TLV/Sub-TLV	Defined in:
269	Administrative group (color)	22/3	[RFC5305]/3.1
270	Maximum link bandwidth	22/9	[RFC5305]/3.3
271	Max. reservable link bandwidth	22/10	[RFC5305]/3.5
272	Unreserved bandwidth	22/11	[RFC5305]/3.6
273	Link Protection Type	22/20	[RFC5307]/1.2
274	MPLS Protocol Mask	---	Section 3.3.1.1
275	Metric	---	Section 3.3.1.2
276	Shared Risk Link Group	---	Section 3.3.1.3
277	OSPF specific link attribute	---	Section 3.3.1.4
278	IS-IS Specific Link Attribute	---	Section 3.3.1.5
279	Area ID	---	Section 3.3.1.6

Table 3: Link Attribute TLVs

3.3.1.1. MPLS Protocol Mask TLV

The MPLS Protocol TLV (Type 274) 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.

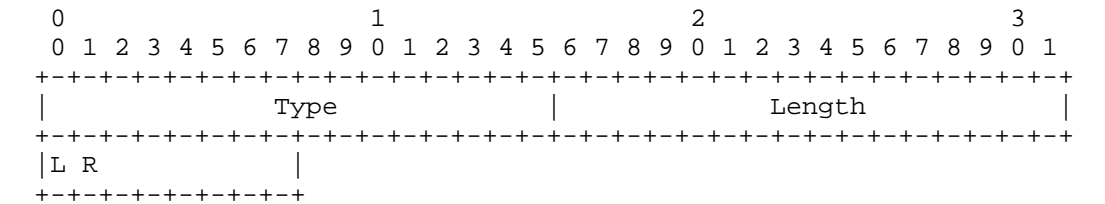


Figure 13: MPLS Protocol TLV

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 Mask TLV Codes

3.3.1.2. Metric TLV

The IGP Metric TLV (Type 275) carries the metric for this link. The length of this TLV is 3. If the length of the 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.

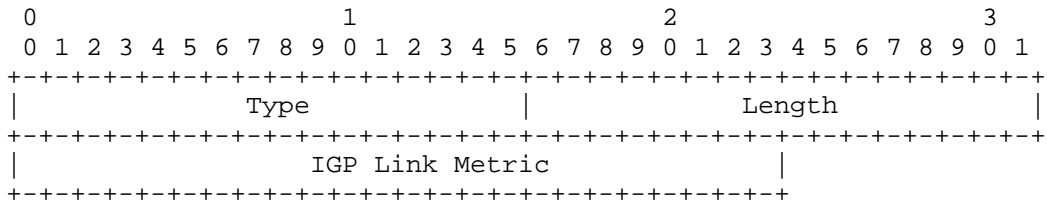


Figure 14: Metric TLV format

3.3.1.3. Shared Risk Link Group TLV

The Shared Risk Link Group (SRLG) TLV (Type 276) 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 15. The length of this TLV is 4 * (number of SRLG values).

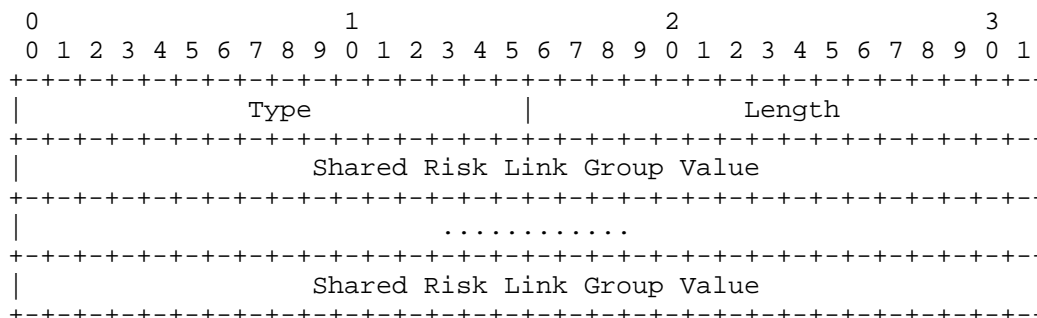


Figure 15: Shared Risk Link Group TLV format

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.

3.3.1.4. OSPF Specific Link Attribute TLV

The OSPF specific link attribute TLV (Type 277) 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.

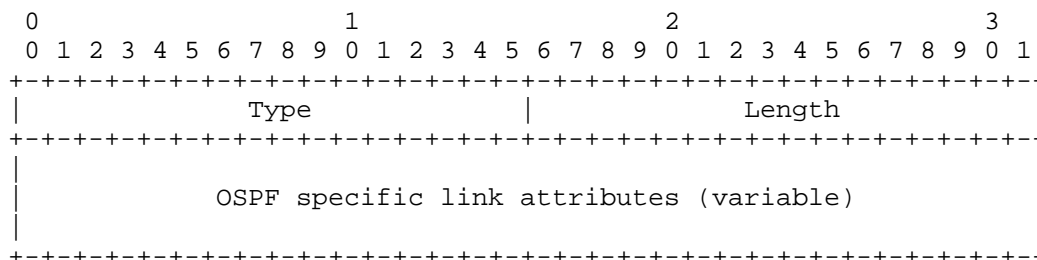


Figure 16: OSPF specific link attribute format

3.3.1.5. IS-IS specific link attribute TLV

The IS-IS specific link attribute TLV (Type 278) 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.

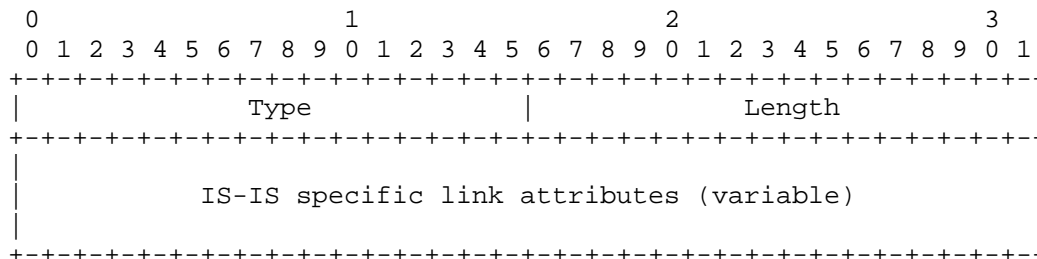


Figure 17: IS-IS specific link attribute format

3.3.1.6. Link Area TLV

The Area TLV (Type 279) 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.

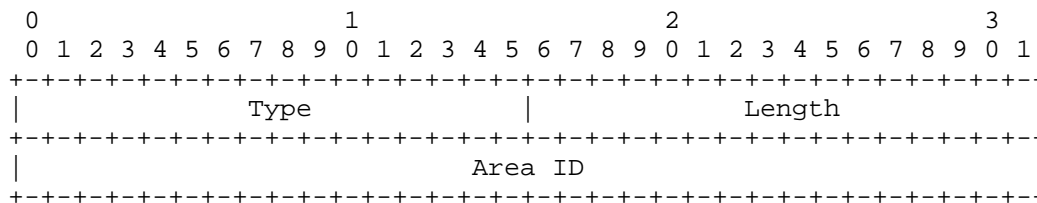


Figure 18: Link Area TLV format

3.3.2. Node Attribute TLVs

The following node attribute TLVs are defined:

Type	Description	Length
280	Multi Topology	2
281	Node Flag Bits	1
282	OSPF Specific Node Properties	variable
283	IS-IS Specific Node Properties	variable
284	Node Area ID	variable

Table 5: Node Attribute TLVs

3.3.2.1. Multi Topology Node TLV

The Multi Topology TLV (Type 280) carries the Multi Topology ID and topology specific flags for this node. The format and semantics of the 'value' field in 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.

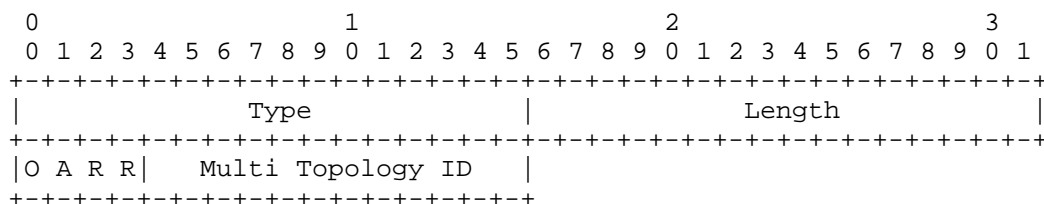


Figure 19: Multi Topology Node TLV format

3.3.2.2. Node Flag Bits TLV

The Node Flag Bits TLV (Type 281) carries a bit mask describing node attributes. The value is a bit array of 8 flags, where each bit represents an MPLS Protocol capability.

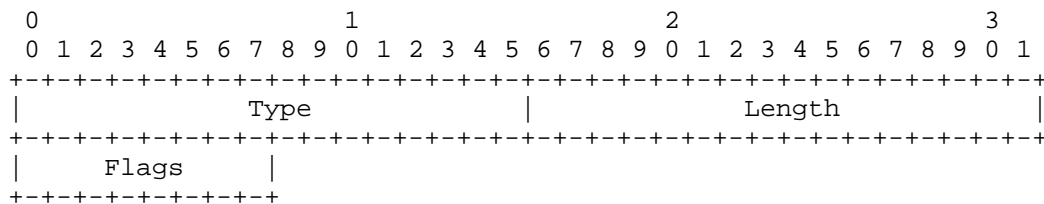


Figure 20: Node Flag Bits TLV format

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

3.3.2.3. OSPF Specific Node Properties TLV

The OSPF Specific Node Properties TLV (Type 282) 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.

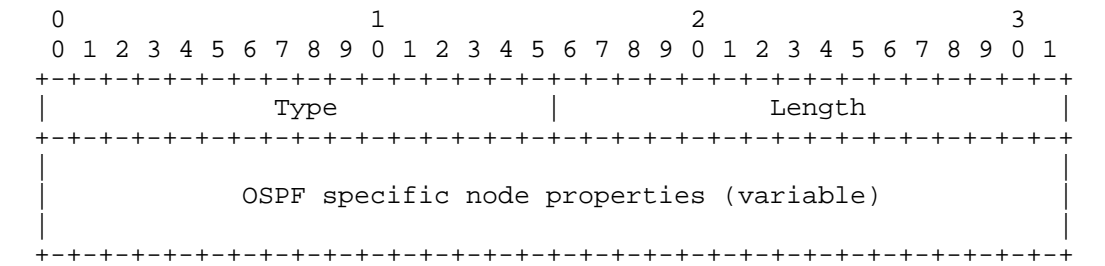


Figure 21: OSPF specific Node property format

3.3.2.4. IS-IS Specific Node Properties TLV

The IS-IS Router Specific Node Properties TLV (Type 283) 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.

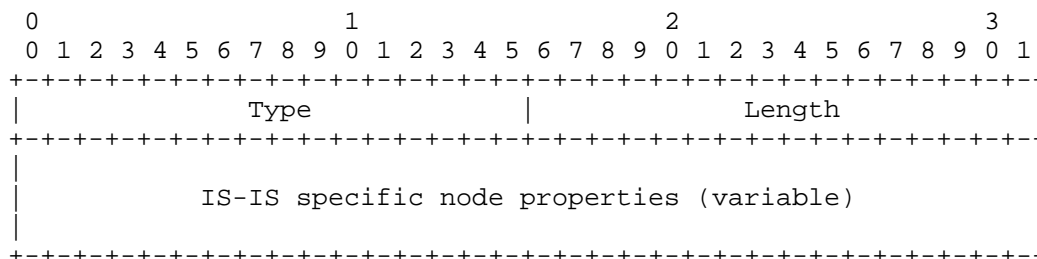


Figure 22: IS-IS specific Node property format

3.3.2.5. Area Node TLV

The Area TLV (Type 284) 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.

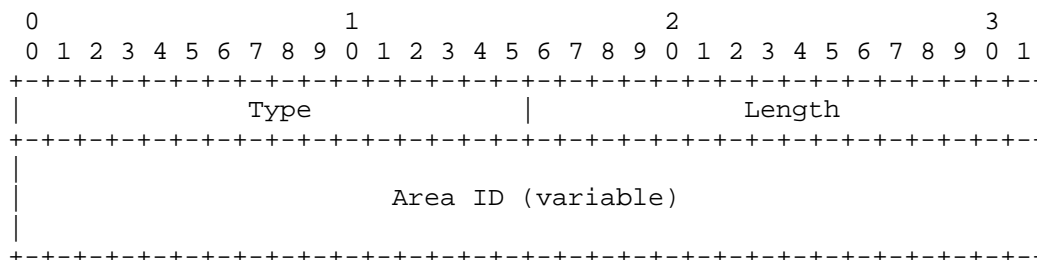


Figure 23: Area Node TLV format

3.4. 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 24. 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.

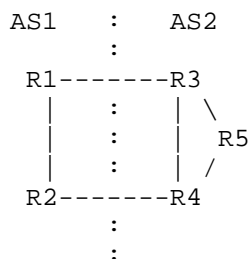


Figure 24: no-link-aggregation

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 25. 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.

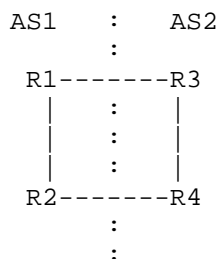


Figure 25: asbr-link-aggregation

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 26. 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.

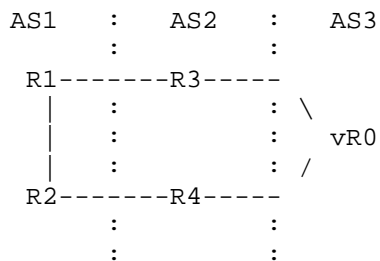


Figure 26: multi-as-aggregation

5. IANA Considerations

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

This document requests a code point from the BGP Path Attributes registry.

This document requests creation of a new registry for node anchor, link descriptor and link attribute TLVs. Values 0-255 are reserved. Values 256-65535 will be used for Codepoints. 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. It shall be noted that the NLRI information present in this document purely carries application level data that have no immediate corresponding forwarding state impact. As such, any churn in reachability information has different impact than regular BGP update which needs to change forwarding state for an entire router. Furthermore it is anticipated that distribution of this NLRI will be handled by dedicated route-reflectors providing a level of isolation and fault-containment between different NLRI types.

6.1.2. Installation and Initial Setup

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

- 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 200 updates per second.

6.1.3. Migration Path

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

6.1.4. Requirements on Other Protocols and Functional Components

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

6.1.5. Impact on Network Operation

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

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

6.1.6. Verifying Correct Operation

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

- 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, Derek Yeung, Murtuza Lightwala, John Scudder, Kaliraj Vairavakkalai, Les Ginsberg, Liem Nguyen, Manish Bhardwaj, Mike Shand, Peter Psenak, Rex Fernando, Richard Woundy, Robert Varga, Saikat Ray, Steven Luong, Tamas Mondal, Waqas Alam, and Yakov Rekhter for their comments.

9. References

9.1. Normative References

- [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.
- [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.
- [RFC5065] Traina, P., McPherson, D., and J. Scudder, "Autonomous System Confederations for BGP", RFC 5065, August 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, February 2008.

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-11 (work in progress), March 2012.
- [I-D.ietf-isis-mi]
Roy, A., Ward, D., Ginsberg, L., Shand, M., and S. Previdi, "IS-IS Multi-Instance", draft-ietf-isis-mi-06 (work in progress), February 2012.
- [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.

[RFC6549] Lindem, A., Roy, A., and S. Mirtorabi, "OSPFv2 Multi-
Instance Extensions", RFC 6549, March 2012.

Authors' Addresses

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

Email: hannes@juniper.net

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

Email: jmedved@cisco.com

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

Email: sprevidi@cisco.com

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

Email: afarrel@juniper.net

IDR
Internet-Draft
Intended status: Informational
Expires: January 15, 2013

P. Lapukhov
Microsoft Corp.
A. Premji
Arista Networks
July 14, 2012

Using BGP for routing in large-scale data centers
draft-lapukhov-bgp-routing-large-dc-01

Abstract

Some service providers build and operate data centers that support over 100,000 servers. In this document, such data-centers are referred to as "large-scale" data centers to differentiate them the from more common smaller infrastructures. The data centers of this scale have a unique set of network requirements, with emphasis on operational simplicity and network stability.

This document attempts to summarize the authors' experiences in designing and supporting large data centers, using BGP as the only control-plane protocol. The intent here is to describe a proven and stable routing design that could be leveraged by others in the industry.

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 15, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal

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

Table of Contents

1. Introduction	3
2. Traditional data center designs	3
2.1. Layer 2 Designs	3
2.2. Fully routed network designs	4
3. Document structure	5
4. Network design requirements	5
4.1. Traffic patterns	5
4.2. CAPEX minimization	6
4.3. OPEX minimization	6
4.4. Traffic Engineering	7
5. Requirement List	7
6. Network topology	7
6.1. Clos topology overview	8
6.2. Clos topology properties	8
6.3. Scaling Clos topology	9
7. Routing design	10
7.1. Choosing the routing protocol	10
7.2. BGP configuration for Clos topology	11
7.2.1. BGP Autonomous System numbering layout	11
7.2.2. Non-unique private BGP ASN's	12
7.2.3. Prefix advertisement	13
7.2.4. External connectivity	13
7.3. ECMP Considerations	14
7.3.1. Basic ECMP	14
7.3.2. BGP ECMP over multiple ASN	15
7.4. BGP convergence properties	16
7.4.1. Convergence timing	16
7.4.2. Failure impact scope	16
7.4.3. Third-party route injection	17
8. Security Considerations	17
9. IANA Considerations	17
10. Acknowledgements	17
11. Informative References	18
Authors' Addresses	19

1. Introduction

This document presents a practical routing design that can be used in large-scale data centers. Such data centers, also known as hyper-scale or warehouse scale data centers, have a unique attribute of supporting over a 100,000 end hosts. In order to support networks of such scale, operators are revisiting networking designs and platforms to address this need.. Contrary to the more traditional data center designs, the approach presented in this document does not have any dependency on building a large Layer-2 domain and instead relies on routing at every layer in the network. Implementing a pure Layer-3 design using BGP further ensures broad vendor support and almost guarantees interoperability between vendors given that BGP is one of the most widely deployed protocols on the Internet.

2. Traditional data center designs

This section provides an overview of two types of traditional data center designs - Layer-2 and fully routed Layer-3 topologies.

2.1. Layer 2 Designs

In the networking industry, a common design choice for data centers is to use a mix of Ethernet-based Layer 2 technologies. Network topologies typically look like a tree with redundant uplinks and three levels of hierarchy commonly named Core , Aggregation and Access layers (see Figure 1). To accommodate bandwidth demands, every next level has higher port density and bandwidth capacity, moving upwards in the topology. To keep terminology uniform, in this document, these topology layers will be referred to as "tiers", e.g. Tier 1, Tier 2 and Tier 3 instead of Core, Aggregation or Access layers.

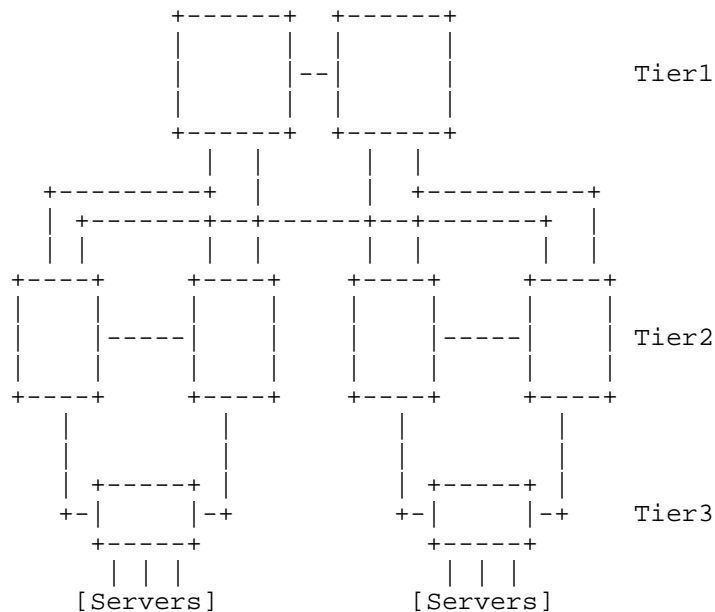


Figure 1: Typical Data Center network layout

IP routing is normally used only at the upper layers in the topology, e.g. Tier 1 or Tier 2. Some of the reasons for introducing such large (sometimes called stretched) layer-2 domains are:

- o Supporting legacy applications that may require direct Layer 2 adjacency or use non-IP protocols
- o Seamless mobility for virtual machines, to allow the preservation of IP addresses when a virtual machine moves across physical hosts
- o Simplified IP addressing - less IP subnets is required for the data-center
- o Application load-balancing may require direct layer-2 reachability to perform certain functions such as Level 2 Direct Server Return (DSR)

2.2. Fully routed network designs

Network designs that leverage IP routing down to the access layer (Tier 3) of the network have gained popularity as well. The main benefit of such designs is improved network stability and scalability, as a result of confining L2 broadcast domains. A common choice of routing protocol for data center designs would be an IGP, such as OSPF or ISIS. As data centers grow in scale, and server count exceeds tens of thousands, such fully routed designs become

more attractive.

Although BGP is the de-facto standard protocol for routing on the Internet, having wide support from both the vendor and service provider communities, it is not generally deployed in data centers for a number of reasons:

- o BGP is perceived as a "WAN only protocol only" and not often considered for enterprise or data center applications.
- o BGP is believed to have a "much slower" routing convergence than traditional IGPs.
- o BGP deployment within an Autonomous System (iBGP mesh) is assumed to have a dependency on the presence of an IGP, which assists with recursive next-hop resolution.
- o BGP is perceived to require significant configuration overhead and does not support any form of neighbor auto-discovery.

In this document we demonstrate a practical approach for using BGP as the single routing protocol for data center networks.

3. Document structure

The remaining of this document is organized as following. First the design requirements for large scale data centers are presented. Next, the document gives an overview of Clos network topology and its properties. After that, the reasons for selecting BGP as the single routing protocols are presented. Finally, the document discusses the design in more details and covers specific BGP policy features.

4. Network design requirements

This section describes and summarizes network design requirement for a large-scale data center.

4.1. Traffic patterns

The primary requirement when building an interconnection network for large number of servers is to accommodate application bandwidth and latency requirements. Until recently it was quite common to see traffic flows mostly entering and leaving the data center (also known as north-south traffic) There were no intense, highly meshed flows or traffic patterns between the machines within the same tier. As a result, traditional "tree" topologies were sufficient to accommodate such flows, even with high oversubscription ratios in network equipment. If more bandwidth was required, it was added by "scaling up" the network elements, by upgrading line-cards or switch fabrics.

In contrast, large-scale data centers often host applications that generate significant amount of server to server traffic, also known as "east-west" traffic. Examples of such applications could be compute clusters such as Hadoop or live virtual machine migrations. Scaling up traditional tree topologies to match these bandwidth demands becomes either too expensive or impossible due to physical limitations.

4.2. CAPEX minimization

The cost of the network infrastructure alone (CAPEX) constitutes about 10-15% of total data center expenditure [GREENBERG2009]. However, The absolute cost is significant, and there is a need to constantly drive down the cost of networking elements themselves. This can be accomplished in two ways:

- o Unifying all network elements, preferably using the same hardware type or even the same device. This allows for bulk purchases with discounted pricing.
- o Driving costs down by introducing multiple network equipment vendors.

In order to allow for vendor diversity, it is important to minimize the software feature requirements for the network elements. Furthermore, this strategy provides the maximum flexibility of vendor equipment choices while enforcing interoperability using open standards

4.3. OPEX minimization

Operating large scale infrastructure could be expensive, provide that larger amount of elements will statistically fail more often. Having a simpler design and operating using a limited software feature-set ensures that failures will mostly result from hardware malfunction and not software issues.

An important aspect of OPEX minimization is reducing size of failure domains in the network. Ethernet networks are known to be susceptible to broadcast or unicast storms. The use of a fully routed design significantly reduces the size of the data-plane failure domains (e.g. limits to Tier-3 switches only). However, such designs also introduce the problem of distributed control-plane failures. This calls for simpler control-plane protocols that are expected to have less chances of network meltdown.

4.4. Traffic Engineering

In any data center, application load-balancing is a critical function performed by network devices. Traditionally, load-balancers are deployed as dedicated devices in the traffic forwarding path. The problem arises in scaling load-balancers under growing traffic demand. A preferable solution would be able to scale load-balancing layer horizontally, by adding more of the uniform nodes and distributing incoming traffic across these nodes

In situation like this, an ideal choice would to use network infrastructure itself to distribute traffic across a group of load-balancers. A combination of features such as Anycast prefix advertisement [RFC4786] along with Equal Cost Multipath (ECMP) functionality could be used to accomplish this. To allow for more granular load-distribution, it is beneficial for the network to support the ability to perform controlled per-hop traffic engineering. For example, it is beneficial to directly control the ECMP next-hop set for anycast prefixes at every level of network hierarchy.

5. Requirement List

This section summarizes the list of requirements, based on the discussion so far:

- o REQ1: Select a network topology where capacity could be scaled "horizontally" by adding more links and network switches of the same type, without requiring an upgrade to the network elements themselves.
- o REQ2: Define a narrow set of software features/protocols supported by a multitude of networking equipment vendors.
- o REQ3: Among the network protocols, choose the one that has a simpler implementation in terms of minimal programming code complexity.
- o REQ4: The network routing protocol should allow for explicit control of the routing prefix next-hop set on per-hop basis.

6. Network topology

This section outlines the most common choice for horizontally scalable topology in large scale data centers.

6.1. Clos topology overview

A common choice for a horizontally scalable topology is a folded Clos topology, sometimes called "fat-tree" (see, for example, [INTERCON] and [ALFARES2008]). This topology features odd number of stages (dimensions) and is commonly made of the same uniform elements, e.g. switches with the same port count. Therefore, the choice of Clos topology satisfies both REQ1 and REQ2. See Figure 2 below for an example of folded 3-stage Clos topology:

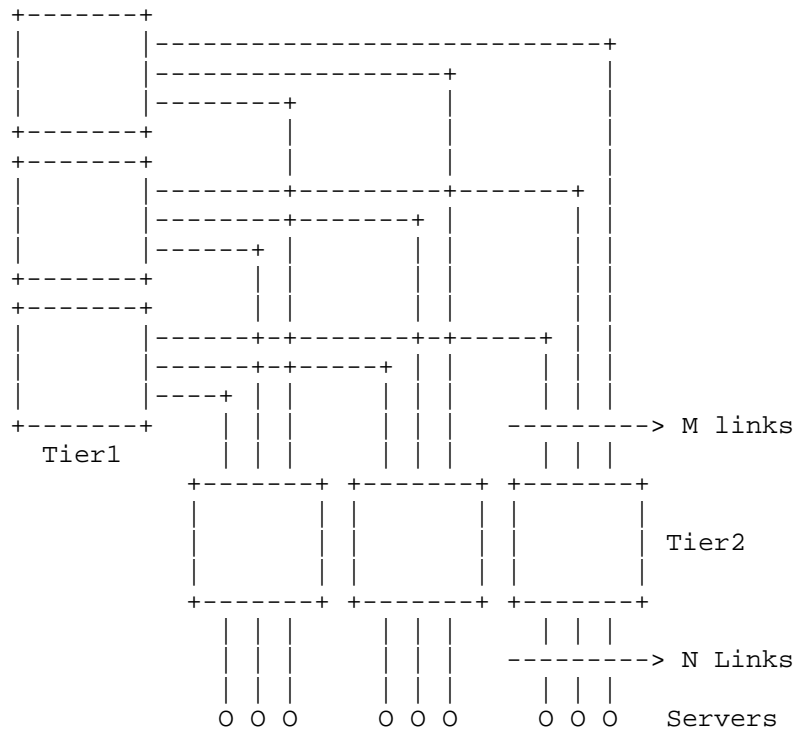


Figure 2: 3-Stage Folded Clos topology

In the networking industry, a topology like this is sometimes referred to as "Leaf and Spine" network, where "Spine" is the name given to the middle stage of the Clos topology (Tier 1) and "Leaf" is the name of input/output stage (Tier 2). However, for consistency, we will refer to these layers as "Tier n".

6.2. Clos topology properties

The following are some key properties of the Clos topology:

- o Topology is fully non-blocking (or more accurately - non-interfering) if $M \geq N$ and oversubscribed by a factor of N/M otherwise. Here M and N is the uplink and downlink port count respectively, for Tier 2 switch, as shown on Figure 2
- o Implementing Clos topology requires a routing protocol supporting ECMP with the fan-out of M or more
- o Every Tier 1 device has exactly one path to every end host (server) in this topology
- o Traffic flowing from server to server is naturally load-balanced over all available paths using simple ECMP behavior

6.3. Scaling Clos topology

A Clos topology could be scaled either by increasing network switch port count or adding more stages, e.g. moving to a 5-stage Clos, as illustrated on Figure 3 below:

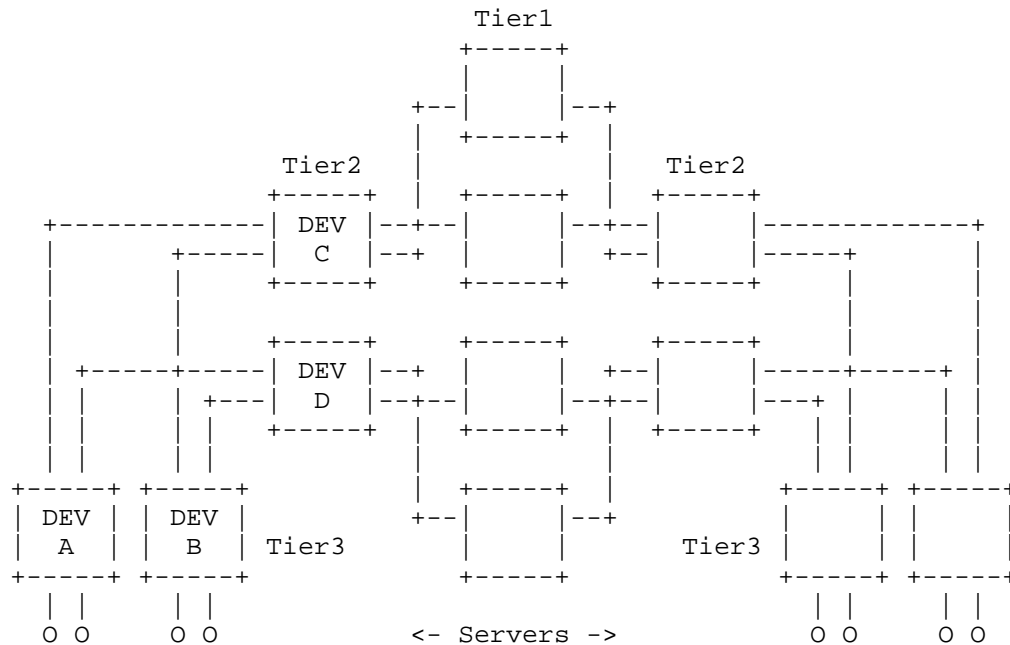


Figure 3: 5-Stage Clos topology

The topology on Figure 3 is built from switches with port count of 4 and provides full bisection bandwidth to all connected servers. We will refer to the collection of directly connected Tier 2 and Tier 3 switches as a "cluster" in this document. For example, devices A, B, C, and D on Figure 3 form a cluster.

In practice, the Tier 3 level of the network (typically top of rack switches, or ToRs) is where oversubscription is introduced to allow for packaging of more servers in data center. The main reason to limit oversubscription at a single layer of the network is to simplify application development that would otherwise need to account for two bandwidth pools: within the same access switch (e.g. rack) and outside of the local switch. Since oversubscription itself does not have any effect on routing, we will not be discussing it further in this document.

7. Routing design

This section discusses the motivation for choosing BGP as the routing protocol and BGP configuration for routing in Clos topology.

7.1. Choosing the routing protocol

The set of requirements discussed earlier call for a single routing protocol (REQ2) to reduce complexity and interdependencies. While it is common to rely on an IGP in this situation, the document proposes the use of BGP only. The advantages of using BGP are discussed below.

- o BGP inherently has less complexity within its protocol design - internal data structures and state-machines are simpler when compared to a link-state IGP. For example, instead of implementing adjacency formation, adjacency maintenance and/or flow-control, BGP simply relies on TCP as the underlying transport. This fulfills REQ1 and REQ2.
- o BGP information flooding overhead is less when compared to link-state IGPs. Indeed, since every BGP router normally re-calculates and propagates best-paths only, a network failure is masked as soon as the BGP speaker finds an alternate path. In contrary, the event propagation scope of a link-state IGP is single flooding domain, regardless of the failure type. Furthermore, all well-known link-state IGPs feature periodic refresh updates, while BGP does not expire routing state.
- o BGP supports third-party (recursively resolved) next-hops. This allows for ECMP or forwarding based on customer-defined forwarding paths. This satisfied REQ4 stated above. Some IGPs, such as OSPF, support similar functionality using special concepts such as "Forwarding Address", but do not satisfy other requirement, such as protocol simplicity.
- o Vanilla BGP configuration, without routing policies, is easier to troubleshoot for network reachability issues. For example, it is straightforward to dump contents of LocRIB and compare it to the router's RIB and FIB. Furthermore, every BGP neighbor has

corresponding AdjRIBIn and AdjRIBOut structures with incoming/outgoing NRLI information that could be easily correlated on both sides of the BGP peering session. Thus BGP fully satisfies REQ3.

7.2. BGP configuration for Clos topology

Topologies that have more than 5 stages are very uncommon due to the large numbers of interconnects required by such a design.

7.2.1. BGP Autonomous System numbering layout

The diagram below illustrates suggests BGP Autonomous System Number (BGP ASN) allocation scheme. The following is a list of guidelines that can be used:

- o All BGP peering sessions are external BGP (eBGP) established over direct point-to-point links interconnecting the network nodes.
- o 16-bit (two octet) BGP ASNs are used, since these are widely supported and have better vendor interoperability (e.g. no need to support BGP capability negotiation).
- o Private BGP ASNs from the range 64512-64534 are used so as to avoid ASN conflicts. The private ASN stripping feature can be leveraged as a result (see below).
- o A single BGP ASN is allocated to the Clos middle stage ("Tier 1"), e.g. ASN 64534 as shown in Figure 4
- o Unique BGP ASN is allocated per group of "Tier 2" switches. All Tier 2 switches in the same group share the BGP ASN.
- o Unique BGP ASN is allocated to every Tier 3 switch (e.g. ToR) in this topology.

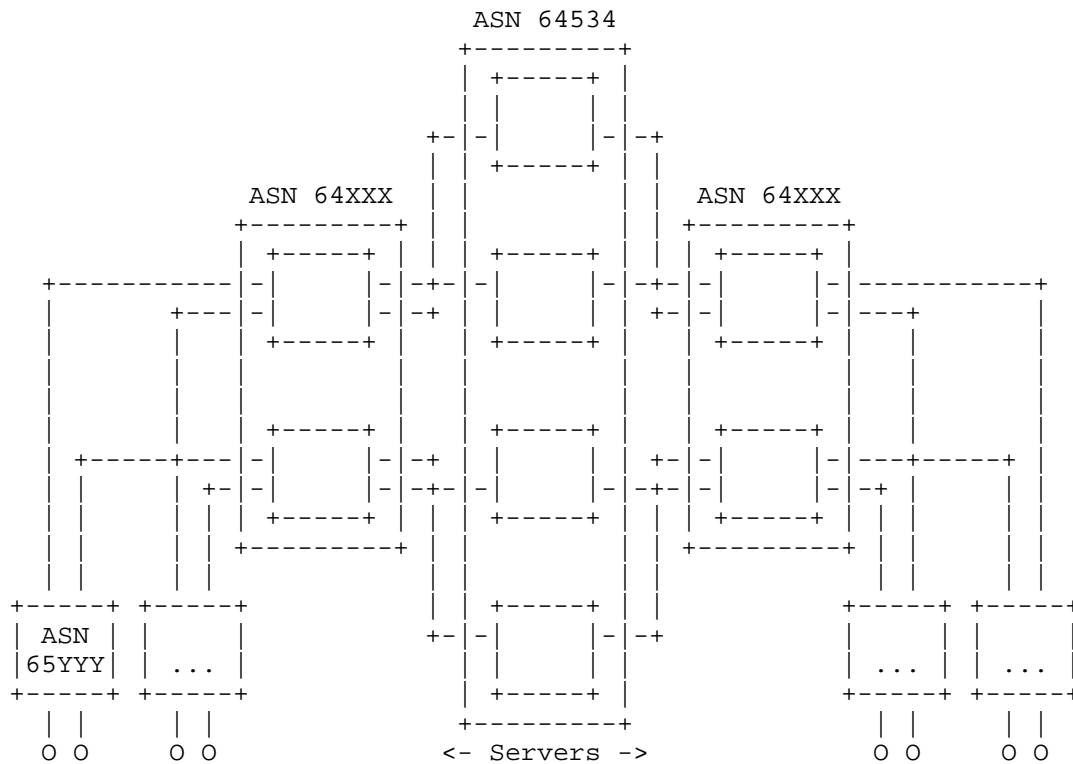


Figure 4: BGP ASN layout for 5-stage Clos

7.2.2. Non-unique private BGP ASN's

The use of private BGP ASNs limits to the usable range of 1022 unique numbers. Since it is very likely that the number of network switches could exceed this number, a workaround is required. One approach would be to re-use the private ASN's assigned to the Tier 3 switches across different clusters. For example, private BGP ASN's 65001, 65002 ... 65032 could be used within every individual cluster to be assigned to Tier 3 switches.

To avoid route suppression due to AS PATH loop prevention, upstream eBGP sessions on Tier 3 switches must be configured with the "AllowAS In" feature that allows accepting a device's own ASN in received route advertisements. Introducing this feature does not create the opportunity for routing loops under misconfiguration since the AS PATH is always incremented when routes are propagated from tier to tier.

Another solution to this problem would be to using four-octet (32-bit) BGP ASNs. However, there are no reserved private ASN range in the four-octet numbering scheme although efforts are underway to support this, see [I-D.mitchell-idr-as-private-reservation]. This will also require vendors to implement specific policy features, such as four-octet private AS removal from AS-PATH attribute.

7.2.3. Prefix advertisement

A Clos topology has a large number of point-to-point links and associated prefixes. Advertising all of these routes into BGP may create FIB overload conditions. There are two possible solutions that can help prevent FIB overload:

- o Do not advertise any of the point-to-point links into BGP. Since eBGP peering changes the next-hop address anyways at every node, distant networks will automatically be reachable via the advertising eBGP peer
- o Advertising point-to-point links, but summarizing them on every advertising device. This requires proper address allocation, for example allocating a consecutive block of IP addresses per Tier 1 and Tier 2 device to be used for point-to-point interface addressing.

Server facing subnets on Tier 3 switches are announced into BGP without using summarization on Tier 2 and Tier 1 switches. Summarizing subnets in the Clos topology will result in route black-holing under a single link failure (e.g. between Tier 2 and Tier 3 switch) and hence must be avoided. The use of peer links within the same tier to resolve the black-holing problem is undesirable due to $O(N^2)$ complexity of the peering mesh and waste of ports on the switches.

7.2.4. External connectivity

A dedicate cluster (or clusters) in the Clos topology could be used solely for the purpose of connecting to the Wide Area Network (WAN) edge devices, or WAN Routers. Tier 3 switches in such a cluster would be replaced with WAN Routers, but eBGP peering would be used again, though WAN routers are likely to belong to a public ASN.

The Tier 2 devices in such a dedicated cluster will be referred to as "Border Routers" in this document. These devices have to perform a few special functions:

- o Hide network topology information when advertising paths to WAN routers, i.e. remove private BGP ASNs from the AS-PATH attribute. This is typically done to avoid BGP ASN number collisions across

the data centers. A BGP policy feature called "Remove Private AS" is commonly used to accomplish this. This feature strips a contiguous sequence of private ASNs found in AS PATH attribute prior to advertising the path to a neighbor. This assumes that all BGP ASN's used for intra data center numbering are from the private ASN range.

- o Originate a default route to the data center devices. This is the only place where default route could be originated, as route summarization is highly undesirable for the "scale-out" topology. Alternatively, Border Routers may simply relay the default route learned from WAN routers.

7.3. ECMP Considerations

This section covers the Equal Cost Multipath (ECMP) functionality for Clos topology and discusses a few special requirements.

7.3.1. Basic ECMP

ECMP is the fundamental load-sharing mechanism used by a Clos topology. Effectively, every lower-tier switch will use all of its directly attached upper-tier devices to load-share traffic destined to the same prefix. Number of ECMP paths between two input/output switches in Clos topology equals to the number of the switches in the middle stage (Tier 1). For example, Figure 5 illustrates the topology where Tier 3 device A has four paths to reach servers X and Y, via Tier 2 devices B and C and then Tier 1 devices 1, 2, 3, and 4 respectively.

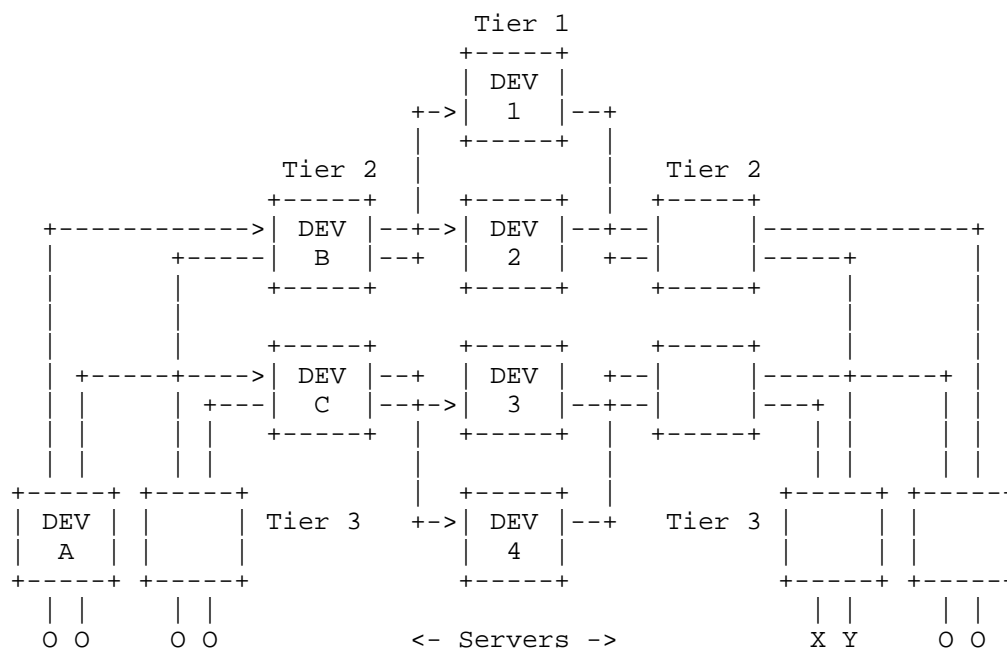


Figure 5: ECMP fan-out tree from A to X and Y

The ECMP requirement implies that the BGP implementation must support multi-path fan-out for up to the maximum number of devices directly attached at any point in the topology. Normally, this number does not exceed half of the ports found on a switch in the topology. For example, an ECMP max-path of 32 would be required when building a Clos network using 64-port devices.

Most implementations declare paths to be equal from ECMP perspective if they match up to and including step (e) in Section 9.1.2.2 of [RFC4271]. In the proposed network design there is no underlying IGP, so all IGP costs are automatically assumed to be zero (or otherwise the same value across all paths). Loop prevention is assumed to be handled by the BGP best-path selection process.

7.3.2. BGP ECMP over multiple ASN

For application load-balancing purposes we may want the same prefix to be advertised from multiple Tier-3 switches. From the perspective of other devices, such a prefix would have BGP paths with different AS PATH attribute values, though having the same AS PATH attribute lengths. Therefore, the BGP implementations must support load-sharing over above-mentioned paths. This feature is sometimes known

as "AS PATH multipath relax" and effectively allows for ECMP to be done across different neighboring ASNs.

7.4. BGP convergence properties

This section reviews routing convergence properties of BGP in the proposed design. A case is made that sub-second convergence is achievable provided that implementation supports fast BGP peering session shutdown upon failure of an associated link.

7.4.1. Convergence timing

BGP typically relies on an IGP to route around link/node failures inside an AS, and implements either a polling based or an event-driven mechanism to obtain updates on IGP state changes. The proposed routing design omits the use of an IGP, so the only mechanisms that could be used for fault detection are BGP keep-alives and link-failure triggers.

Relying solely on BGP keep-alive packets may result in high convergence delays, in the order of multiple seconds (normally, the minimum recommended BGP hold time value is 3 seconds). However, many BGP implementations can shut down local eBGP peering sessions in response to the "link down" event for the outgoing interface used for BGP peering. This feature is sometimes called as "fast fail-over". Since the majority of the links in modern data centers are point to point fiber connections, a physical interface failure is often detected in milliseconds and subsequently triggers a BGP re-convergence.

Furthermore, popular link technologies, such as 10Gbps Ethernet, may support a simple form of OAM for failure signaling such as [FAULTSIG10GE], which makes failure detection more robust. Alternatively, as opposed to relying on physical layer for fault signaling, some platforms may support Bidirectional Forwarding Detection ([RFC5880]) to allow for sub-second failure detection and fault signaling to the BGP process. This, however, presents additional requirements to vendor software and possibly hardware, and may contradict REQ1.

7.4.2. Failure impact scope

BGP is inherently a distance-vector protocol, and as such some of failures could be masked if the local node can immediately find a backup path. The worst case is that all devices in data center topology would have to either withdraw a prefix completely, or recalculate the ECMP paths in the FIB. Reducing the fault domain using summarization is not possible with the proposed design, since

using this technique may create route black-holing issues as mentioned previously. Thus, the control-plane failure impact scope is the network as a whole. It is worth pointing that such property is not a result of choosing BGP, but rather a result of using the "scale-out" Clos topology.

7.4.3. Third-party route injection

BGP allows for a third-party BGP speaker (not necessarily directly attached to the network devices) to inject routes anywhere in the network topology. This could be achieved by peering an external speaker using an eBGP multi-hop session with some or even all devices in the topology. Furthermore, BGP diverse path distribution [I-D.ietf-grow-diverse-bgp-path-dist] could be used to inject multiple next-hop for the same prefix to facilitate load-balancing. Using such a technique would make it possible to implement unequal-cost load-balancing across multiple clusters in the data-center, by associating the same prefix with next-hops mapped to different clusters.

For example, a third-party BGP speaker may peer with Tier 3 and Tier 1 switches, injecting the same prefix, but using a special set of BGP next-hops for Tier 1 devices. Those next-hops are assumed to resolve recursively via BGP, and could be, for example, IP addresses on Tier 3 switches. The resulting forwarding table programming could provide desired traffic proportion distribution among different clusters.

8. Security Considerations

The design does not introduce any additional security concerns. For control plane security, BGP peering sessions could be authenticated using TCP MD5 signature extension header [RFC2385]. Furthermore, BGP TTL security [I-D.gill-btsh] could be used to reduce the risk of session spoofing and TCP SYN flooding attacks against the control plane.

9. IANA Considerations

There are no considerations associated with IANA for this document.

10. Acknowledgements

This publication summarizes work of many people who participated in developing, testing and deploying the proposed design. Their names, in alphabetical order, are George Chen, Parantap Lahiri, Dave Maltz,

Edet Nkposong, Robert Toomey, and Lihua Yuan. Authors would also like to thank Jon Mitchell, Linda Dunbar and Susan Hares for reviewing and providing valuable feedback on the document.

11. Informative References

- [RFC4786] Abley, J. and K. Lindqvist, "Operation of Anycast Services", BCP 126, RFC 4786, December 2006.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC2385] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC 2385, August 1998.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.
- [I-D.ietf-grow-diverse-bgp-path-dist]
Raszuk, R., Fernando, R., Patel, K., McPherson, D., and K. Kumaki, "Distribution of diverse BGP paths.", draft-ietf-grow-diverse-bgp-path-dist-07 (work in progress), May 2012.
- [I-D.mitchell-idr-as-private-reservation]
Mitchell, J., "Autonomous System (AS) Reservation for Private Use", draft-mitchell-idr-as-private-reservation-00 (work in progress), June 2012.
- [I-D.gill-btsh]
Gill, V., Heasley, J., and D. Meyer, "The BGP TTL Security Hack (BTSH)", draft-gill-btsh-02 (work in progress), May 2003.
- [GREENBERG2009]
Greenberg, A., Hamilton, J., and D. Maltz, "The Cost of a Cloud: Research Problems in Data Center Networks", January 2009.
- [FAULTSIG10GE]
Frazier, H. and S. Muller, "Remote Fault & Break Link Proposal for 10-Gigabit Ethernet", September 2000.
- [INTERCON]
Dally, W. and B. Towles, "Principles and Practices of Interconnection Networks", ISBN 978-0122007514, January 2004.

[ALFARES2008]

Al-Fares, M., Loukissas, A., and A. Vahdat, "A Scalable,
Commodity Data Center Network Architecture", August 2008.

Authors' Addresses

Petr Lapukhov
Microsoft Corp.
One Microsoft Way
Redmond, WA 98052
US

Phone: +1 425 7032723 X 32723
Email: petrlapu@microsoft.com
URI: <http://microsoft.com/>

Ariff Premji
Arista Networks
5470 Great America Parkway
Santa Clara, CA 95054
US

Phone: +1 408-547-5699
Email: ariff@aristanetworks.com
URI: <http://aristanetworks.com/>

Network Working Group
Internet-Draft
Updates: 1930 (if approved)
Intended status: Informational
Expires: December 22, 2012

J. Mitchell
Microsoft Corporation
June 20, 2012

Autonomous System (AS) Reservation for Private Use
draft-mitchell-idr-as-private-reservation-00

Abstract

This document describes the reservation of Autonomous System numbers (ASNs) that may be used within networks but should not be advertised to the Internet, known as private use ASNs. This document enlarges the total space available for private use ASNs by documenting the reservation of a second larger range and updates RFC 1930.

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 December 22, 2012.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

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

described in the Simplified BSD License.

1. Introduction

The original IANA reservation of ASNs for private use was a block of 1023 ASNs. This was documented by IETF in Section 10 of [RFC1930] that specified the private use ASN range as 64512 through 65535 (implying the inclusion of ASN 65535). Since that range was reserved and documented over a decade ago, BGP has seen much wider deployment in Service Provider, Enterprise and Content Provider networks. The use cases in these networks for private use ASNs include networks that are attached to the Internet, utilizing implementation specific features to remove them upon advertisement to Internet peers, and networks that are not attached to the Internet. The displacement of Frame Relay and ATM based VPNs by BGP/MPLS IP VPNs [RFC4364] has also increased the deployment of BGP to a larger number of sites, especially for networks with requirements for multi-homing or provider redundancy.

The limited size of the current range of private use ASNs has led to the usage of a number of implementation specific features that manipulate the AS_PATH or remove AS_PATH based loop prevention described in Section 9 of [RFC4271]. These workarounds have increased the operational complexity of the networks since the implementations of these functions vary and have been largely out of scope of existing BGP standards.

Since the introduction of BGP Support for Four-octet AS Number Space [RFC4893], the total size of the ASN space has increased dramatically, and a larger subset of the space should be available to network operators to deploy in private use cases. The existing range of private use ASNs is widely deployed and the ability to renumber this resource for reassignment in existing networks cannot be coordinated given these ASNs by definition are not registered. Therefore this document clarifies the existing ASN range, while introducing a second, larger, range that can also be utilized.

2. Private Use ASNs

To allow the continued growth of usage of the BGP protocol in networks that utilize private ASNs two ranges of ASNs are reserved by this document in Section 5. The first which was previously defined in [RFC1930] out of the original 16-bit Autonomous System range and a second, larger, reserved block available out of the higher part of the Four-Octet AS Number Space [RFC4893].

3. Operational Considerations

If private use ASNs are used and prefixes are originated from these private use ASNs which are destined to the Internet, private use ASNs must be removed from the AS_PATH before being advertised to the global Internet. Prior to making use of the second, numerically higher, range of these ASNs network operators should be confident any implementation specific features or filters that recognize private use ASNs have been updated to recognize both ranges correctly so that no unintended announcement of private use ASNs to the Internet occurs.

4. Acknowledgements

The author would like to acknowledge Christopher Morrow and Jason Schiller for their advice on how to pursue this change.

5. IANA Considerations

[Note to IANA, not for publication: The IANA may wish to consider updating the existing private use AS number reservation to explicitly include ASN 65535 that has been documented in RFC 1930 as well as implemented in a number of implementations that recognize private use ASNs as part of the existing range. Further, to maintain consistency from an operator standpoint, it is suggested that the end of the "32-bit number set" be reserved for Private Use, and a size of 967,295 is suggested corresponding to the range of 4294000000 to 4294967295 inclusive, primarily motivated by being visibly recognizable while not consuming a large portion of the total ASN space.]

[Note to WG, not for publication: a small poll of network implementors was taken and it was not felt optimizing the range to a bit boundary was important for control plane performance of any features that recognize private use ASNs - others implementors in WG may want to weigh in here. Also very interested in feedback on appropriate sizing of the future range given the many current and future uses of BGP in networks]

(If approved) IANA has reserved, for Private Use, a contiguous block of TBD (1023 or 1024) Autonomous System numbers from the "16-bit Autonomous System Numbers" registry, namely 64512 - TBD1 (65534 or 65535) inclusive.

(If approved) IANA has also reserved, for Private Use, a contiguous block of TBD Autonomous System numbers from the "32-bit Autonomous System Numbers" registry, namely TBD2 - TBD3 inclusive.

(If approved) These reservations have been documented in the IANA AS Number Registry [IANA.AS].

6. Security Considerations

This document does not introduce any additional security concerns in regards to private use ASNs.

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

7.2. Informative References

- [IANA.AS] IANA, "Autonomous System (AS) Numbers", June 2012, <<http://www.iana.org/assignments/as-numbers/>>.
- [RFC1930] Hawkinson, J. and T. Bates, "Guidelines for creation, selection, and registration of an Autonomous System (AS)", BCP 6, RFC 1930, March 1996.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

Author's Address

Jon Mitchell
Microsoft Corporation
12012 Sunset Hills Road
Reston, VA 20190
USA

Email: Jon.Mitchell@microsoft.com

Network Working Group
Internet Draft
Intended status: Standards Track
July 16, 2012
Expires: Jan 16, 2013

Jim Uttaro
AT&T
Matthieu Texier
Arbor Networks
David Smith
Pradosh Mohapatra
Cisco Systems
Wim Henderickx
Adam Simpson
Alcatel-Lucent
July 16, 2012

BGP Flow-Spec Extended Community for Traffic Redirect to IP Next Hop
draft-simpson-idr-flowspec-redirect-01.txt

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), 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/ietf/lid-abstracts.txt>

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

This Internet-Draft will expire on Jan 16, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must

include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Abstract

Flow-spec is an extension to BGP that allows for the dissemination of traffic flow specification rules. This has many possible applications but the primary one for many network operators is the distribution of traffic filtering actions for DDoS mitigation. The flow-spec standard [RFC 5575] defines a redirect-to-VRF action for policy-based forwarding but this mechanism can be difficult to use, particularly in networks without L3 VPNs.

This draft proposes a new redirect-to-IP flow-spec action that provides a simpler method of policy-based forwarding. This action is indicated by the presence of a new BGP extended community in the flow-spec route. Many routers already support a redirect-to-IP filter action and, in this case, the only new functionality implied by this draft is the ability to signal the action using flow-spec.

Table of Contents

1. Introduction.....	3
2. Terminology.....	3
3. Redirect to IP Extended Community.....	3
4. Security Considerations.....	5
5. IANA Considerations.....	5
6. References.....	5
6.1. Normative References.....	5
6.2. Informative References.....	5
7. Acknowledgments.....	6

1. Introduction

Flow-spec is an extension to BGP that allows for the dissemination of traffic flow specification rules. This has many possible applications but the primary one for many network operators is the distribution of traffic filtering actions for DDoS mitigation.

Every flow-spec route is effectively a rule, consisting of a matching part (encoded in the NLRI field) and an action part (encoded as a BGP extended community). The flow-spec standard [RFC 5575] defines widely-used filter actions such as discard and rate limit; it also defines a redirect-to-VRF action for policy-based forwarding. Using the redirect-to-VRF action for redirecting traffic towards an alternate destination is useful for DDoS mitigation but it can be complex and cumbersome, particularly in networks without L3 VPNs.

This draft proposes a new redirect-to-IP flow-spec action that provides a simpler method of policy-based forwarding. This action is indicated by the presence of a new BGP extended community in the flow-spec route. Many routers already support a redirect-to-IP filter action and, in this case, the only new functionality implied by this draft is the ability to signal the action using flow-spec.

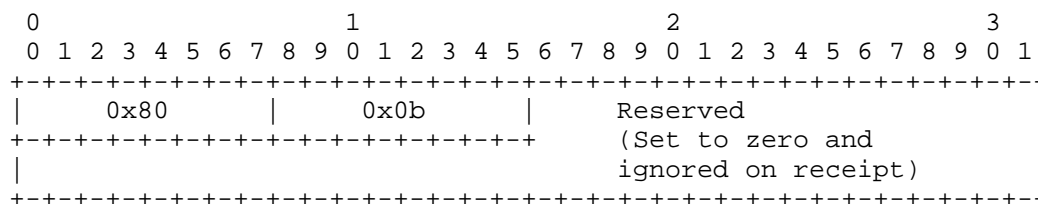
2. Terminology

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].

3. Redirect to IP Extended Community

This document proposes a new BGP extended community called "flow spec redirect-to-IP". IANA is requested to allocate a type value of 0x800b for this purpose. This extended community can be added to any UPDATE message announcing the reachability of one or more flow-spec NLRI. The encoding of the attribute is shown in Figure 1; the 6 bytes of data after the 2-byte type value is a reserved field and should be set to 0 by the originating BGP speaker and ignored by receiving BGP speakers.

The redirect-to-IP extended community is valid with any other set of flow-spec extended communities except if that set includes a redirect-to-VRF extended community (type 0x8008) and in that case the redirect-to-IP extended community should be ignored.



Flow-spec Redirect-to-IP Extended Community

Figure 1

When a BGP speaker receives an UPDATE message with the redirect-to-IP extended community it is expected to create a traffic filtering rule for every flow-spec NLRI in the message that has this path as its best path. The filter entry matches the IP packets described in the NLRI field and forwards them towards the IPv4 or IPv6 address specified in the 'Network Address of Next-Hop' field of the associated MP_REACH_NLRI. More specifically: if an IPv4 [or IPv6] packet with destination address D that is normally forwarded to a next-hop A matches a filter entry of the type described above it MUST instead be forwarded to next-hop B, where B is found by FIB lookup of the IPv4 [or IPv6] address contained in the MP_REACH_NLRI next-hop field.

If an MP_REACH_NLRI containing one or more flow-spec NLRI does not have a valid IPv4 or IPv6 address in its next-hop field, or the length of the next-hop is 0, then the redirect-to-IP extended community, if present, should be ignored.

The scope of application (in terms of router interfaces/contexts) of the filter rules derived from the redirect-to-IP extended community is outside the scope of this specification except for noting that filter rules derived from VPNv4 and VPNv6 flow-spec routes should only be installed in the VRF contexts that import the routes.

The redirect-to-IP extended community is transitive across AS boundaries. When a flow-spec route with this community is advertised to an EBGp peer the next-hop address in the MP_REACH_NLRI SHOULD be reset to an address of the advertising router by default, per normal BGP procedures. Alternatively, the advertising router MAY be configured to keep the next-hop unchanged, if it is known that the destination AS has a valid route to the next-hop address.

The validation check described in [RFC 5575] and revised in [VALIDATE] SHOULD be applied by default to received flow-spec routes with the redirect-to-IP extended community, as it is to all types of flow-spec routes. This means that a flow-spec route with a destination prefix subcomponent SHOULD NOT be accepted from an EBGp peer unless that peer also advertised the best path for the matching unicast route. BGP speakers that support the redirect-to-IP extended community MUST also, by default, enforce the following check when receiving a flow-spec route from an EBGp peer: if the flow-spec route has an IP next-hop X and includes a redirect-to-IP extended community then the BGP speaker SHOULD discard the redirect-to-ip extended community (and not propagate it further with the flow-spec route) if the last AS in the AS_PATH or AS4_PATH attribute of the longest prefix match for X does not match the AS of the EBGp peer. It MUST be possible to disable this additional validation check on a per-EBGP session basis.

4. Security Considerations

A system that originates a flow-spec route with a redirect-to-IP extended community can cause many receivers of the flow-spec route to send traffic to a single next-hop, overwhelming that next-hop and resulting in an inadvertent or deliberate denial-of-service. This is particularly a concern when the redirect-to-IP extended community is allowed to cross AS boundaries. The validation check described in section 3 significantly reduces this risk.

5. IANA Considerations

This document requests that IANA allocate a new experimental use extended community type value in the range 0x8000-0x8FFF for the flow spec redirect-to-IP action. The proposed type value is 0x800b.

6. References

6.1. Normative References

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

6.2. Informative References

[RFC5575] P. Marques, N. Sheth, R. Raszuk, B. Greene, J. Mauch, D. McPherson, "Dissemination of Flow Specification Rules", RFC 5575, August 2009.

- [IPV6-FLOW] R. Raszuk, B. Pithawala, D. McPherson,
"Dissemination of Flow Specification Rules for
IPv6", draft-ietf-idr-flow-spec-v6-00, June 2011.
- [VALIDATE] Uttaro, J., Filsfils, C., Mohapatra, P., Smith, D.,
"Revised Validation Procedure for BGP Flow
Specifications", draft-ietf-idr-bgp-flowspec-oid-00,
June 2012.

7. Acknowledgments

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

Authors' Addresses

James Uttaro
AT&T
200 S. Laurel Avenue
Middletown, NJ 07748
USA
Email: jul738@att.com

Pradosh Mohapatra
Cisco
170 W. Tasman Drive
San Jose, CA 95134
USA
Email: pmohapat@cisco.com

David Smith
Cisco
111 Wood Avenue South
Iselin, NJ 08830
USA
E-mail: djsmith@cisco.com

Wim Henderickx
Alcatel-Lucent
Copernicuslaan 50
2018 Antwerp, Belgium
Email: wim.henderickx@alcatel-lucent.be

Adam Simpson
Alcatel-Lucent
600 March Road
Ottawa, Ontario K2K 2E6
Canada
Email: adam.simpson@alcatel-lucent.com

Matthieu Texier
Arbor Networks
38 Rue de Berri
75008 Paris
Email: mtexier@arbor.net

Network working group
Internet Draft
Category: Standard Track

Expires: January 2013

X. Xu
Huawei
K. Lee
China Telecom
July 16, 2012

BGP Tunnel Address Prefix Attribute and Tunnel Address Prefix
Extended Community

draft-xu-idr-tunnel-address-prefix-01

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/ietf/lid-abstracts.txt>.

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

This Internet-Draft will expire on January 16, 2013.

Copyright Notice

Copyright (c) 2009 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

Abstract

This document describes a new BGP attribute referred to as Tunnel Address Prefix Attribute and a new BGP address specific extended community referred to as Tunnel Address Prefix Extended Community, both of which are intended for facilitating the load-balancing of IP/GRE tunneled traffic (e.g., L3VPN-over-GRE traffic) in the core of IP-enabled Packet Switch Networks (PSN).

Conventions used in this document

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].

Table of Contents

1. Introduction	3
2. Terminology	4
3. Tunnel Address Prefix Attribute	4
4. Tunnel Address Prefix Extended Community	5
5. Functionality Description	6
5.1. Egress Routers	6
5.2. Ingress Routers.....	6
5.3. Intermediate Routers	6
6. Applicability	6
7. Security Considerations	7
8. IANA Considerations	7
9. Acknowledgements	7
10. References	7
10.1. Normative References	7
10.2. Informative References	7
Authors' Addresses	8

1. Introduction

Equal Cost Multi-Path (ECMP) and Link Aggregation Group (LAG) are widely used in the core of IP-enabled Packet Switch Networks (PSN) for load-balancing purposes. Most core routers in the IP-enabled PSN are capable of load-balancing IP traffic flows across ECMP paths and/or LAG based on the hash of the five-tuple of UDP/TCP packets (i.e., source IP address, destination IP address, source port, destination port, and protocol) or some fields in the IP header of non-UDP/TCP packets (e.g., source IP address, destination IP address). However, in the L3VPN [RFC4364], L2VPN and Softwire mesh [RFC5565] scenarios, distinct customer traffic flows between a given tunnel endpoint pair (e.g., the PE pair in the L3VPN context) would be encapsulated with the same IP/GRE tunnel header prior to traversing the core of IP PSN. In addition, since the IP/GRE encapsulated traffic is neither TCP nor UDP, core routers therefore could only perform hash calculation on the fields in the IP header of IP/GRE tunnels. As a result, core routers could not achieve an effective load-balancing for these IP/GRE tunneled traffic flows in the core network due to the lack of adequate entropy information.

[RFC5640] describes a method for improving the load-balancing in Softwire mesh networks [RFC5565]. However, this method requires core routers to be able to perform hash calculation on the fields including the specific "load-balancing" field contained in the L2TPv3 or GRE tunnel header. [Entropy-Label] proposes to use the "entropy labels" for achieving a better load-balancing for MPLS traffic flows in the core of MPLS-enabled PSN. Although the entropy label could be inserted in the "Key" field of the GRE header by ingress PE routers in the case where the PSN is IP enabled rather than MPLS enabled, it still requires core routers to be capable of performing hash calculation on the "entropy label" contained in the GRE tunnel header. Any of the above two load-balancing methods requires a change to the data plane of core routers.

This document describes an alternative load-balancing method suitable for the above scenarios, which is backwards compatible to those already-deployed core routers that could only perform hash calculation on the fields in the IP header in the case of IP/GRE tunneled traffic flows. The basic idea of this method is: a given (tunnel) egress router signals to (tunnel) ingress routers a special prefix called "tunnel address prefix" via BGP and any addresses beginning with that prefix would be used by those ingress routers as tunnel destination addresses when tunneling traffic towards that egress router. Therefore distinct traffic flows between that tunnel endpoint pair could be encapsulated with as many different tunnel

destination addresses as possible. In this way, core routers could achieve a better load-balancing for those IP/GRE tunneled traffic through performing hash calculation just on the fields in the IP header.

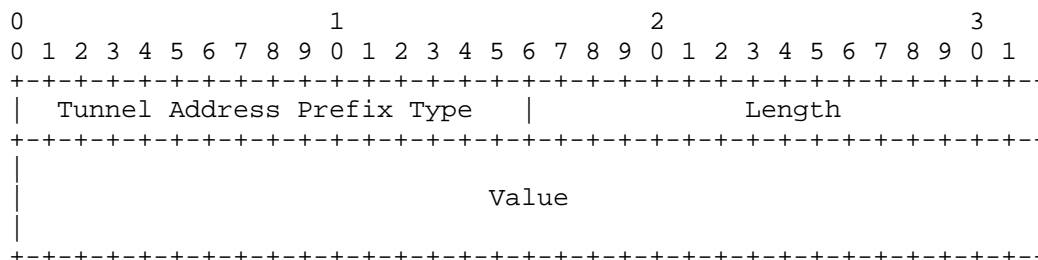
2. Terminology

This memo makes use of the terms defined in [RFC4364] and [RFC5565].

3. Tunnel Address Prefix Attribute

For a given BGP router to tell remote BGP routers what tunnel destination addresses could be used when tunneling traffic flows to it, a new BGP attribute contained in the Encapsulation SAFI [RFC5512], referred to as "Tunnel Address Prefix Attribute", could be used to indicate the available tunnel destination addresses to be used.

The Tunnel Address Prefix attribute is an optional transitive attribute. The TLV is structured as follows:



- Tunnel Address Prefix Type (2 octets): indicates the Value field of such TLV contains the tunnel address prefix information in the form of IP address and subnet mask pair.

- Length (2 octets): indicates the total number of octets of the value field. If the AFI of the Encapsulation SAFI is IPv4, the length value is set to 64; otherwise if the AFI is IPv6, the length value is set to 256.

- Value (variable): contains the tunnel address prefix information in the form of IP address and subnet mask pair.

4. Tunnel Address Prefix Extended Community

Here, we also define an address specific extended community referred to as Tunnel Address Prefix extended community that can be attached to BGP UPDATE messages to indicate the available tunnel destination addresses to be used when tunneling traffic flows from an ingress router to an egress router. This extended community is useful in the case where the Encapsulation SAFI capability is not supported between BGP routers or one really wants to specify different tunnel destination address prefixes for distinct sets of traffic flows. For example, one wants to assign two different tunnel address prefixes for traffic flows destined for prefix X and Y respectively.

IPv4 Tunnel Address Prefix extended community is as follows:

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	2	3	4	5	6	7	8	9																								
0x01																Sub-Type																Global Administrator																															
Global Administrator (cont.)																Local Administrator																																															

The value of the high-order octet of the extended type field is 0x01, which indicates it is transitive across ASes. The value of the low-order octet of the extended type field is to be defined. The Global Administrator sub-field contains an IPv4 unicast address while the Local Administrator sub-field contains the corresponding Prefix Length.

IPv6 Tunnel Address Prefix extended community is as follows:

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	2	3	4	5	6	7	8	9																								
0x00																Sub-Type																Global Administrator																															
Global Administrator (cont.)																																																															
Global Administrator (cont.)																																																															
Global Administrator (cont.)																																																															
Global Administrator (cont.)																Local Administrator																																															

The value of the high-order octet of the extended type field is 0x00, which indicates it is transitive across ASes. The value of the low-order octet of the extended type field is to be defined. The Global Administrator sub-field contains an IPv6 unicast address while the Local Administrator sub-field contains the corresponding Prefix Length.

5. Functionality Description

5.1. Egress Routers

An egress router could attach the above BGP Tunnel Address Prefix attribute or extended community to BGP UPDATE messages in order to indicate the available tunnel destination addresses to be used by any ingress routers when tunneling traffic flows to it. In addition, it SHOULD create the corresponding loopback interface for each IP address within that prefix and accordingly advertise a route for that prefix via IGP. As such, it could receive and process those IP/GRE tunneled traffic flows destined for any of those addresses beginning with that prefix.

5.2. Ingress Routers

For an ingress router receiving the above BGP Tunnel Address Prefix attribute or extended community announced by a given egress router, it could use any addresses beginning with the tunnel address prefix, in addition to the BGP next-hop address contained in the MP_REACH_NLRI attribute, as tunnel destination addresses when tunneling traffic flows towards that egress router.

5.3. Intermediate Routers

There is no special requirement on Intermediate Routers (i.e., core routers). In other words, they could perform load-balancing of the IP/GRE tunneled traffic on basis of the hash of the fields in the IP headers as normal.

6. Applicability

The load-balancing approach described in this document is suitable for many scenarios including but not limited to L3VPN [RFC4364], 6PE [RFC4798], Software mesh [RFC5565], BGP free core and L2VPN including VPLS [RFC4761, RFC4762] and E-VPN [E-VPN]. In the existing VPLS case where BGP is used for auto-discovery, the above BGP Tunnel Address Prefix attribute or extended community would be attached to the BGP update messages as well. Once a customer MAC address is learnt against a given BGP next-hop address, any addresses beginning

with the Tunnel Address Prefix which is associated with that BGP next-hop address could be used as tunnel destination addresses when tunneling MAC frames destined for that MAC address.

7. Security Considerations

TBD.

8. IANA Considerations

The type code of the Tunnel Address Prefix Attribute needs to be allocated by IANA. Meanwhile, a new Sub-Type of the Address Specific BGP Extended Communities of IPv4 and IPv6 respectively SHOULD also be assigned by IANA.

9. Acknowledgements

Thanks to Robert Raszuk for his valuable comments on this document.

10. References

10.1. Normative References

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

10.2. Informative References

[RFC5640] Filsfils, C., Mohapatra, P and C. Pignataro, "Load-Balancing for Mesh Softwires", RFC 5640, August 2009.

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

[RFC5701] Rekhter, Y., "IPv6 Address Specific BGP Extended Communities Attribute", RFC5701, November 2009.

[RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, April 2009.

[RFC5565] Wu, J., Cui, Y., Metz, C., and E. Rosen, "Softwire Mesh Framework", RFC 5565, June 2009.

[RFC4364] "BGP/MPLS IP VPNs", Rosen, Rekhter, et. al., February 2006

[Entropy-Label] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", draft-ietf-mpls-entropy-label-01, work in progress, October, 2011.

[RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

[RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

[E-VPN] Aggarwal et al., "BGP MPLS Based Ethernet VPN", draft-ietf-12vpn-evpn-00.txt, work in progress, February, 2012.

Authors' Addresses

Xiaohu Xu
Huawei Technologies,
Beijing, China

Phone: +86-10-60610041
Email: xuxiaohu@huawei.com

Kai Lee
China Telecom,
Beijing, China.

Email: Leekai@ctbri.com.cn

Network Working Group
Internet-Draft
Updates: RFC 4724 (if approved)
Intended status: Standards Track
Expires: January 7, 2013

H. Zhang
HangZhou H3C Co. Limited
A. Retana
Hewlett-Packard Co.
July 6, 2012

Transitive BGP Graceful Restart
draft-zhang-idr-transitive-gr-00

Abstract

This document defines an extension to BGP Graceful Restart that reduces the negative impact of multiple inter-connected routers restarting. The proposed mechanism does not require any changes to the BGP protocol.

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 7, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
2. Requirements Language	3
3. Proposed Solution	4
4. Security Considerations	4
5. IANA Considerations	4
6. Acknowledgements	5
7. References	5
7.1. Normative References	5
7.2. Informative References	5
Authors' Addresses	5

1. Introduction

The BGP Graceful Restart [RFC4724] process defines a mechanism that a restarting router can use with its non-restarting peers. The existence of other restarting routers results in the use of the base route exchange mechanism [RFC4271] with them, even if the forwarding state has indeed been preserved for (and by) those peers during the restart. As a result, traffic forwarding between restarting routers is disrupted.

This document defines an extension to BGP Graceful Restart that reduces the negative impact of multiple inter-connected restarting routers. The proposed mechanism does not require any changes to the BGP protocol.

The current process [RFC4724] states that routes from restarting peers are to be removed from the local forwarding state when the non-restarting peers converge (the End-of-RIB marker is received from all of them). Assuming a simple topology:

NR1 - R2 - R3 - NR4

where NRx are non-restarting routers, Rx are restarting routers and the lines between them represent BGP sessions.

There are two types of routes affected (from R2's point of view) by the current process:

1. Routes that are only reachable through R3. These routes will be removed from the forwarding table when the non-restarting routers converge, and installed back in when the convergence with R3 is done.
2. Routes that are reachable through both R3 and NR1. These routes will first change to NR1 when the non-restarting routers converge, and later back to R3 (assuming that is in fact still the preferred path).

Both types can clearly cause disruption in traffic forwarding, micro-loops, traffic loss, etc.

2. 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 [RFC2119].

3. Proposed Solution

The extension proposed to BGP Graceful Restart to accommodate for multiple restarting routers, when the forwarding state has been preserved between them, is simply to delay sending the End-of-RIB marker to non-restarting routers.

Specifically, to allow a restarting router the ability to reduce the impact due to other restarting routers, the following paragraph is added as the fifth one in section 4.1 (Procedures for the Restarting Speaker) [RFC4724]:

Before updating the corresponding forwarding states, the BGP speaker MAY advertise the Adj-RIB-Out to the remaining peers (ones with the "Restart State" bit set in the received capability and ones that do not advertise the graceful restart capability), including the End-of-RIB marker, and MAY wait for the corresponding End-of-RIB marker from the restarting ones.

During the recovery period of multiple restarting routers, a BGP speaker may advertise routing information that is not being used at the time. Because the forwarding state of the speakers remains unchanged (from that at the restart), it is clear that this transitive property of sharing routing information between restarting routers doesn't cause any issues in the actual forwarding of traffic. Furthermore, it has the advantage of avoiding further disruptions in the forwarding of traffic through the restarting routers.

4. Security Considerations

This document proposes an extension to an existing mechanism. The same security considerations explained there apply to this extension.

The propagation of routing information that is not in use may cause forwarding loops and an inconsistent state in a network. However, the risk in this document is mitigated by the fact that the information is validated by all peers once the convergence process completes.

5. IANA Considerations

This document has no IANA actions.

6. Acknowledgements

The authors would like to thank Enke Chen for his feedback.

7. References

7.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, January 2007.

7.2. Informative References

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

Authors' Addresses

Haifeng Zhang
HangZhou H3C Co. Limited
310 Liuhe Road, Zhijiang Science Park
Hangzhou
P.R. China

Email: zhanghf@h3c.com

Alvaro Retana
Hewlett-Packard Co.
2610 Wycliff Road
Raleigh, NC 27607
USA

Email: alvaro.retana@hp.com

