

Scalable BGP FRR Protection against Edge Node Failure
draft-bashandy-bgp-edge-node-frr-00.txt

Abstract

Consider a BGP free core scenario. Suppose the edge BGP speakers PE1, PE2,..., PEn know about a prefix P/p via the external routers CE1, CE2,..., CEm. If the edge router PEi crashes or becomes totally disconnected from the core, it desirable for a penultimate hop route "P" carrying traffic to the failed edge router PEi to immediately restore traffic by re-tunneling packets originally tunneled to PEi and destined to the prefix P/p to one of the other edge routers that advertised P/p, say PEj, until BGP re-converges. In doing so, it is highly desirable to keep the core BGP-free while not imposing restrictions on external connectivity. Thus (1) a core router should not be required to learn any BGP prefix, (2) the size of the forwarding and routing tables in the core routers should be independent of the number of BGP prefixes, (3) there should be no special router (or group of routers) that handles restoring traffic, and (4) there should be no restrictions on what edge routers advertise what prefixes. For labeled prefixes, (5) the penultimate hop router must swap the label advertised by the failed edge router PEi for the prefix P/p with the label advertised for the same prefix by the edge router PEj before re-tunneling the packet to PEj

Status of this Memo

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

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

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 10, 2012.

Copyright Notice

Copyright (c) 2011 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
1.1. Problem definition.....	4
1.2. Conventions used in this document.....	5
1.3. Terminology.....	5
2. Control Plane Operation.....	6
2.1. Control plane Operation for unlabeled prefixes.....	6
2.1.1. Step 1: Calculation of the Repair PE.....	7
2.1.2. Step 2: Assigning and Advertising the BGP Next-hop...	7
2.1.3. Step 3: Informing Core Routers about the Repair PE...	7
2.1.4. Step 4: How a P router (a core router) Programs its Forwarding Plane.....	8

2.2. Control plane Operation for Labeled Prefixes.....	8
2.2.1. Step 1: Calculation of the Repair PE.....	9
2.2.2. Step 2: Assigning and Advertising the BGP Next-hop...	9
2.2.3. Step 3: Informing core routers about the repair path.	9
2.2.4. Step 4: How a P router (a core router) programs its forwarding plane.....	10
2.3. Rules for a choosing Repair path.....	11
2.3.1. General Rules for Choosing and Programming the Repair Path.....	11
2.3.2. Rules for Choosing the Repair Path for Labeled Prefixes	11
2.4. Forwarding Plane Operation.....	12
2.4.1. For unlabeled prefix.....	12
2.4.2. For Labeled prefix.....	13
3. Example.....	14
4. Security Considerations.....	16
5. IANA Considerations.....	16
6. Conclusions.....	16
7. References.....	16
7.1. Normative References.....	16
7.2. Informative References.....	16
8. Acknowledgments.....	17

1. Introduction

In a BGP free core, where traffic is tunneled between edge routers, BGP speakers advertise reachability information about prefixes. For labeled address families, namely AFI/SAFI 1/4, 2/4, 1/128, and 2/128, an edge router assigns local labels to prefixes and associates the local label with each advertised prefix such as L3VPN [.6], 6PE . [7], and Software [.5]. Suppose that a given edge router is chosen as the best next-hop for a prefix P/p. An ingress router that receives a packet from an external router and destined for the prefix P/p "tunnels" the packet across the core to that egress router. If the prefix P/p is a labeled prefix, the ingress router pushes the label advertised by the egress router before tunneling the packet to the egress router. Upon receiving the packet from the core, the egress router takes the appropriate forwarding decision based on the content of the packet or the label pushed on the packet.

In modern networks, it is not uncommon to have a prefix reachable via multiple edge routers. One example is the best external path [.4]. Another more common and widely deployed scenario is L3VPN [.6] with multi-homed VPN sites. As an example, consider the L3VPN topology depicted in Figure 1.

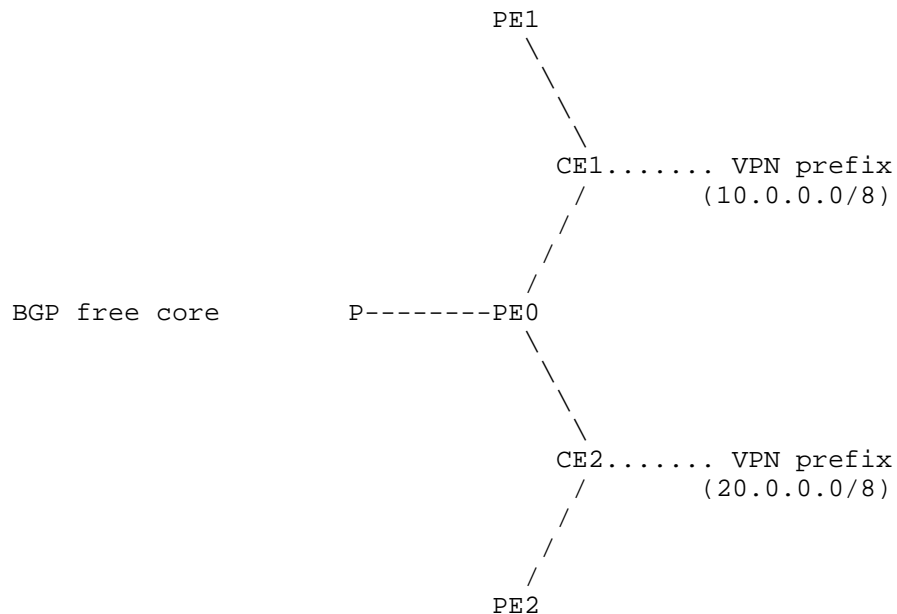


Figure 1 VPN prefix reachable via multiple PEs

As illustrated in Figure 1, the edge router PE0 is the primary NH for both 10.0.0.0/8 and 20.0.0.0/8. At the same time, both 10.0.0.0/8 and 20.0.0.0/8 are reachable through the other edge routers PE1 and PE2, respectively.

1.1. Problem definition

The problem that we are trying to solve is as follows

- o Even though multiple prefixes may share the same egress router, they have different backup edge router. In Figure 1 above, both 10.0.0.0/8 and 20.0.0.0/8 share the same primary next hop PE0, the routing protocol(s) must identify that the node protecting loop free alternate for 10.0.0.0/8 is PE1 while the node protecting loop free alternate for 11.0.0.0/8 is PE2
- o On loosing connection to the edge router, the core router "P" needs to redirect traffic towards the "correct" backup edge router without waiting for IGP or BGP to re-converge and update the routing tables. On the failure of PE0 illustrated in Figure 1, the core router P MUST reroute traffic for 10.0.0.0/8 towards PE1 and traffic for 11.0.0.0/8 towards PE2

- o The core router P MUST NOT be forced to learn about the BGP prefixes on any of the edge routers. The same applies for all core routers.
- o There SHOULD NOT be a need for a special router or group of routers to handle rerouting traffic on edge node failure.
- o The size of the routing table on any core router MUST be independent of the number of BGP prefixes in the network.
- o For labeled prefixes, the core router MUST swap the label advertised by egress edge router (PE0 in Figure 1) with the label advertised by the backup router (PE1 and PE2 in Figure 1).

1.2. 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 [1].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

1.3. Terminology

This section outlines the terms used in this document. For ease of use, we will use terms similar to those used by L3VPN [6].

- o BGP-Free core: A network where BGP prefixes are only known to the edge routers and traffic is tunneled between edge routers
- o Protected prefix: It is a prefix P/p (of any AFI) that a BGP speaker has an external path to. The BGP speaker may learn about the prefix from an external peer through BGP, some other protocol, or manual configuration. The protected prefix is advertised to some or all the internal peers.
- o Primary egress PE: It is an IBGP peer that can reach the protected prefix P/p through an external path and advertised the prefix to the other IBGP peers. The primary egress PE was chosen as the best path by one or more internal peers. In other words, the primary egress PE is an egress PE that will normally be used by some ingress PEs when there is no failure. Referring to Figure 1, PE0 is a primary egress PE.

- o Primary next-hop: It is an IPv4 or IPv6 host address belonging to the primary egress PE. If the prefix is advertised via BGP, then the primary next-hop is the next-hop attribute in the BGP update message [.2].
- [3].
- o CE: It is an external router through which an egress PE can reach a prefix P/p. The routers "CE1" and "CE2" in F .igure 1 are examples of such CE.
 - o Ingress PE: It is a BGP speaker that learns about a prefix through another IBGP peer and chooses that IBGP peer as the next-hop for the prefix.
 - o Repairing P router: A core router that attempts to restore traffic when the primary egress PE is no longer reachable without waiting for IGP or BGP to re-converge. The repairing P router restores the traffic by rerouting the traffic (through a tunnel) towards the pre-calculated repair PE when it detects that the primary egress PE is no longer reachable. Referring to F .igure 1, the router "P" is the repairing P router.
 - o Repair egress PE: It is an egress PE other than the primary egress PE that can reach the protected prefix P/p through an external neighbor. The repair PE is pre-calculated via other PEs prior to any failure. Referring to F .igure 1, PE1 is the repair PE for 10.0.0.0/8 while PE2 is the repair PE for 20.0.0.0/8.
 - o Protected egress PE: Any primary egress PE protected by a repairing P router.
 - o Protected edge router: Any protected egress PE.
 - o Repair path: It is the repair egress PE. If the protected prefix is a labeled prefix, the repair path is the repair egress PE together with the label that will be pushed when the repairing P router reroutes traffic to the repair PE.

2. Control Plane Operation

This section specifies the control plane operation needed to solve the problem mentioned in the Introduction.

2.1. Control plane Operation for unlabeled prefixes

This section specifies the operation of the control plane for AFI/SAFI 1/1, 2/1, 1/2, and 2/2

The control plane operation can be summarized in 4 steps:

1. Calculation of the repair PE
2. Assigning and advertising the next-hop for protected prefixes
3. Informing core routers about repair PEs
4. How a P router (a core router) programs its forwarding plane

2.1.1. Step 1: Calculation of the Repair PE

1. Consider the prefix P/p learnt by an egress edge router PE_i via an external neighbor. The edge router advertises the prefix P/p to some or all of its IBGP peers
2. The edge router PE_i MAY choose a repair PE for the external prefix P/p. Section 2 ..3. specifies the rules for choosing the repair edge router PE_j.
3. In the end, an egress edge route PE_i will have a repair edge router PE_j for some or all prefixes that PE_i has external path(s) to.

2.1.2. Step 2: Assigning and Advertising the BGP Next-hop

1. An edge router PE_i groups the set of prefixes that have a repair PE as follows: Two prefixes belong to the same group if they share the same repair PE
2. For each group, the PE assigns a local next-hop. Thus if a prefix P/p belongs to group G_i, its primary next-hop is NH_i. For example, the PE assigns a different loopback interface address as the next-hop for each of the groups of prefixes.
3. When advertising the prefix and its primary next-hop to its IBGP peers, the PE router uses NH_i as the next-hop attribute of prefixes belonging to the group G_i

2.1.3. Step 3: Informing Core Routers about the Repair PE

1. In step 2 (Section 2 ..1.2.) the egress PE assigns a primary next-hop NH_i for protected prefixes belonging to group G_i.
2. The primary next-hop NH_i is advertised to the core using IGP as usual

3. The repair next-hop is the next-hop attribute advertised for the prefix P/p by the repair edge router PEj. Let's denote the repair next-hop for prefixes belonging to group Gi by "rNHi". Because rNHi is the next-hop advertised by the repair PE, rNHi will also be known to all core routers via IGP
 4. The egress PE MUST advertise the pair (NHi,rNHi) to all directly connected core routers.
 5. The egress PE MAY advertise the pair (NHi,rNHi) to all core routers in the network.
 6. The structure and method of advertising the pair (NHi,rNHi) is beyond the scope of this document. For example, the pair (NHi,rNHi) may be advertised through an ISIS optional TLV.
 7. The semantics of the pair (NHi,rNHi) are: If the next-hop NHi becomes unreachable, then traffic destined to the next-hop NHi SHOULD be re-tunneled to the next-hop rNHi because rNHi can reach prefixes reachable via the primary next-hop NHi.
 8. Because of the previous steps, core routers that are directly connected to the egress PE (and possibly other core routers) are aware of the repair next-hop for protected BGP prefixes reachable via the egress PE. Note that a core router is totally unaware of the BGP prefixes.
- 2.1.4. Step 4: How a P router (a core router) Programs its Forwarding Plane
1. Through usual IGP mechanism, the P router has a prefix matching every BGP next-hop. Let the next-hop NHi match the route Ri
 2. The next-hop of prefix Ri is on the path towards the primary egress PE.
 3. Thus the FIB entry for Ri is programmed as follows:
 - a. Primary next-hop: the next router on the path towards NHi
 - b. Repair next-hop: the next-router on the path towards rNHi
- 2.2. Control plane Operation for Labeled Prefixes

This section specifies the operation of the control plane for AFI/SAFI 1/4, 2/4, 1/128, and 2/128.

2.2.1. Step 1: Calculation of the Repair PE

1. As usual, each PE allocates a local label for each prefix it can reach through an external neighbor CE. A PE may also allocate a repair label is specified in [.11].
2. Each edge router advertises the prefix together with the local label and possibly the repair label [.11] to some or all of its IBGP peers.
3. As a result, an edge router PE_i having an external path to the prefix P/p may learn about the prefix through other IBGP peers.
4. The edge router PE_i MAY choose a repair PE for the external prefix P/p. Rules for choosing the repair PE are specified in Section 2 ..3.
5. The edge router PE_i chooses the one of labels advertised by the other edge router PE_j for the prefix P/p as the "repair label". The algorithm for choosing the repair label is specified in Section 2 ..3.
6. In the end, if the edge router PE_i can reach the prefix P/p through an external path and the prefix P/p is advertised by at least one other PE, the edge router PE_i will have
 - o a primary path towards the CE from which it learnt the prefix, and
 - o a repair path consisting of a repair PE and a repair label advertised by the chosen repair PE.

2.2.2. Step 2: Assigning and Advertising the BGP Next-hop

1. The edge router PE_i groups all BGP prefixes for which PE_i has an external path and a repair path as follows: Two prefixes belong to the same group G_i if they share the same repair PE and repair label
2. The remaining steps are identical to the steps used by unlabeled prefixes in section 2 ..1.2.

2.2.3. Step 3: Informing core routers about the repair path

1. In step 2 (Section 2 ..2.2.) the egress PE as signs a primary next-hop NH_i and a repair path (consisting of a repair next-hop and repair label) for protected prefixes belonging to group G_i.
2. The primary next-hop NH_i is advertised into IGP as usual

3. The repair next-hop is the next-hop advertised for the prefix P/p by the repair edge router PEj. Denote the repair next-hop for prefixes belonging to group Gi by "rNHi". Denote the repair label for prefixes belonging to group Gi by "rLi". Because rNHi is the next-hop attribute advertised by the repair PE, rNHi will also be known to all core routers via IGP
 4. The repairing egress PE MUST advertise the triplet (NHi, rNHi, rLi) to all directly connected core routers.
 5. The repairing egress PE MAY advertise the triplet (NHi, rNHi, rLi) to all core routers in the network
 6. The triplet (NHi,rNHi, rLi) may be advertised through various means, such as ISIS optional TLV. The structure and method of advertising the triplet (NHi,rNHi,rLi) is beyond the scope of this document.
 7. The semantics of the pair (NHi,rNHi,rLi) are: If the next-hop NHi becomes unreachable, then traffic destined to the next-hop NHi should be re-tunneled to the next-hop rNHi and the label pushed by the ingress PE MUST be swapped with the label rLi
 8. Because of the previous steps, core routers that are directly connected to the egress edge router (and possibly other core routers) are aware of the repair path for protected BGP prefixes reachable via the egress edge router. Note that a core router is totally unaware of the BGP prefixes themselves
- 2.2.4. Step 4: How a P router (a core router) programs its forwarding plane
1. Through usual IGP mechanism, the P router has a prefix matching every BGP next-hop. Let the primary next-hop NHi match the route Ri
 2. The next-hop of prefix Ri is on the path towards the protected egress edge router PEi. The next-hop of the prefix Ri is considered the primary path for the prefix Ri
 3. Thus the FIB entry for Ri is programmed as follows
 - a. Primary path: the next router on the path towards NHi
 - b. Repair path:
 - i. Pop label in the packet right under the tunnel header (irrespective of the value of that label)

ii. Push the repair label rLi

iii. Re-tunnel the packet towards the repair next-hop rNH_i

2.3. Rules for a choosing Repair path

This section specifies rules governing how an egress edge router PE_i chooses the repair path. Other than the rules in this section, the method of choosing the repair path is beyond the scope of this document.

2.3.1. General Rules for Choosing and Programming the Repair Path

This section specifies general rules for choosing the repair path for both labeled and unlabeled prefixes.

1. A repair PE MUST be another edge router PE_j that advertises the same prefix to the edge router PE_i via IBGP peering.
2. If the repairing "P" router determines that the path taken by the tunnel from the repairing "P" router to repair edge router PE_j passes through the protected edge router PE_i, then the repairing router "P" SHOULD NOT install the repair path in its forwarding plane. Instead the repair path MAY use a different FRR protection mechanism such as that specified in [.8], [.9], and [.10].

The reason for this rule is that the tunnel to the repair edge router PE_j does not provide protection against the failure of the edge node PE_i. Instead it provides core protection against the failure of the path through the core leading to the protected edge node PE_i. Thus existing core FRR protection mechanisms such as those specified in [.8], [.9], and [.10] can be used

[.8], [.9], and [.10] can be used

2.3.2. Rules for Choosing the Repair Path for Labeled Prefixes

This section specifies additional rules by which an egress edge router PE_i chooses the repair path for an external labeled prefix P/p.

1. A primary edge router PE_i SHOULD only choose the edge router PE_j and the repair label rLi as a repair path for the prefix P/p if label advertised for the prefix P/p by the repair edge router PE_j is allocated on per-VPN or per-CE/per-next-hop basis.

The reason for this is as follows. As mentioned in the abstract and Introduction, the core of the network SHOULD remain BGP-free and the size of the routing table on a core router SHOULD remain independent of the number of BGP prefixes. BGP prefix grouping in section 2

o two

different groups if the labels advertised by the repair PE for the two prefixes are different. Thus if the repair edge router allocates labels on per-prefix basis, protected edge router PE_i will advertise a different primary next-hop for each protected prefix. This is equivalent to having core router "P" knowing about every BGP prefix. In addition the size of the routing table of the "P" router becomes comparable to the number of BGP prefixes.

2. If the repair edge router PE_j advertises a repair label as described in [.11], then the protected edge router PE_i MAY choose the repair label advertised by PE_j as the repair label for the prefix P/p.

Using the repair label specified in [.11] has two advantages:

- o A repairing edge router PE_j need not change the primary label allocation policy (which may be per-prefix) but can be chosen as repair PE if the repair labels are allocated on per-CE or per-VRF basis.
- o As mentioned in [.11], an edge router does NOT repair a packet arriving with a repair label. Hence using the repair label when re-tunneling the packet towards PE_j guarantees loop freedom in case of PE-CE link failure.

2.4. Forwarding Plane Operation

This section specifies the forwarding plane operation on the core router "P" when it detects that the protected edge router PE_i is no longer reachable. We assume that the core router has pre-programmed its forwarding plane according to Sections 2

2.2. .

2.4.1. For unlabeled prefix

The forwarding table for the route R_i is programmed according to section . 2.1.4. As soon as the "P" router detects that the primary next-hop for R_i is not reachable it does the following for any packet destined to the protected edge router PE_i.

1. Decapsulate tunnel header of the arriving packet

2. Tunnel the packet towards the repair egress PE identified by the repair next-hop rNHj

2.4.2. For Labeled prefix

The forwarding table for the route Ri is programmed according to section . 2.2. . Remember that packets tunneled to the egress edge

PEi have a label under the tunnel encapsulation. As soon as the "P" router detects that the primary next-hop for Ri is not reachable it does the following for any arriving packet destined to the protected edge router PEi

1. Decapsulate the tunnel header to expose the labeled packet
2. Swap the label on the top of the packet (irrespective of the value of that label) with the repair label rLi
3. Tunnel the packet towards the repair egress PE identified by rNHj

3. Example

We will use an LDP core as an example. Consider the diagram depicted in Figure 2 below. We assume that the PEs advertise repair labels as specified in [.11]

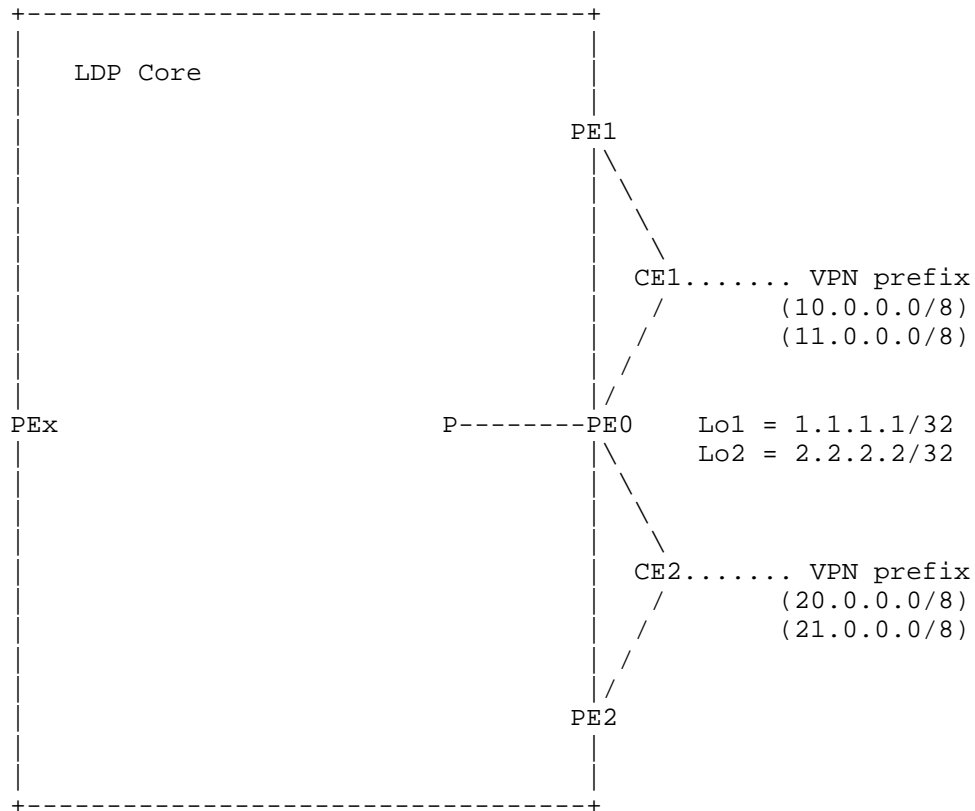


Figure 2 : Edge node BGP FRR in LDP core

1. As we can see, PE0 has 4 prefixes: 10.0.0.0/8, 11.0.0.0/8, 20.0.0.0/8, and 21.0.0.0/8. PE0 may assign a separate label to each prefix. The method and policy of assigning primary labels to each prefixes is irrelevant.
2. PE1 advertises the repair label rL1 for prefixes 10.0.0.0/8 and 11.0.0.0/8
3. PE2 advertises the repair label rL2 for prefixes 20.0.0.0/8 and 21.0.0.0/8

4. As such, PE0 divides its prefixes into two groups
G1 = {10.0.0.0/8, 11.0.0.0/8}
G2 = {20.0.0.0/8, 21.0.0.0/8}
5. When advertising the next-hop to its IBGP peer, PE0 advertises 1.1.1.1 as the next-hop for prefixes belonging to group G1 and 2.2.2.2 as the next-hop for prefixes belonging to group G2.
6. PE0 advertises the prefixes 1.1.1.1/32 and 2.2.2.2/32 using the usual IGP mechanism.
7. When advertising 1.1.1.1/32 into the core, PE0 advertises rL1 and PE1 as a repair path. When advertising 2.2.2.2/32 into the core, PE0 advertises rL2 and PE2 as a repair path. The mechanism by which a repair path is advertised is beyond the scope of the proposal.
8. On the penultimate hop router "P", LDP assigns a different LDP label to 1.1.1.1/32 and 2.2.2.2/32. Core routers other than penultimate hop routers may employ some sort of label aggregation to reduce the number of LDP labels
9. Assume that the penultimate hop router "P" assigns the local LDP label L1 for prefix 1.1.1.1/32 and L2 for prefix 2.2.2.2/32
10. On the penultimate router P, the forwarding entry for L1 will be as follows
Primary path:
 - nexthop is PE0.
 - swap the incoming outer label with the LDP label towards 1.1.1.1Repair path
 - Pop the incoming LDP label
 - Swap the internal label with the repair label rL1
 - Push the LDP label towards PE1
 - Forward the packet
11. On the core router P, the forwarding entry for L2 will be as follows
Primary path: Same as L1
Repair Path
 - Pop the incoming LDP label
 - Swap the internal label with the repair label rL2
 - Push the LDP label towards PE2
 - Forward the packet

12.If the P router detects that PE0 is no longer reachable, it can use the repair path already pre-programmed in the forwarding plane as described above. Because the repair path is pre-programmed as in the case of TE and IP FRR, the P router can re-route traffic very fast

4. Security Considerations

No additional security risk is introduced by using the mechanisms proposed in this document

5. IANA Considerations

No requirements for IANA

6. Conclusions

This document proposes a method that allows fast re-route protection against edge node failure or complete disconnected from the core in a BGP-free core

7. References

7.1. Normative References

- [1] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [2] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006
- [3] Bates, T., Chandra, R., Katz, D., and Rekhter Y., "Multiprotocol Extensions for BGP", RFC 4760, January 2007

7.2. Informative References

- [4] Marques,P., Fernando, R., Chen, E, Mohapatra, P., "Advertisement of the best external route in BGP", draft-ietf-idr-best-external-02.txt, April 2004.
- [5] Wu, J., Cui, Y., Metz, C., and E. Rosen, "Software Mesh Framework", RFC 5565, June 2009.
- [6] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [7] De Clercq, J. , Ooms, D., Prevost, S., Le Faucheur, F., Connecting IPv6 Islands over IPv4 MPLS Using IPv6 Provider Edge Routers (6PE)", RFC 4798, February 2007

- [8] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.
- [9] Shand, S., and Bryant, S., "IP Fast Reroute", RFC5714, January 2010
- [10] Shand, M. and S. Bryant, "A Framework for Loop-Free Convergence", RFC 5715, January 2010.
- [11] Bashandy, P., Pithawala, B., and Hietz J., "Scalable, Loop-Free BGP FRR using Repair Label, "draft-bashandy-idr-bgp-repair-label-02.txt", June 2011

8. Acknowledgments

Special thanks to Keyur Patel, Robert Raszuk, and Eric Rosen for the valuable comments

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

Authors' Addresses

Ahmed Bashandy
Cisco Systems
170 West Tasman Dr, San Jose, CA 95134
Email: bashandy@cisco.com

Network Working Group
Internet Draft
Intended status: Standards Track
Expires: January 2012

A. Bashandy
B. Pithawala
Cisco Systems
Jakob Hietz
Ericsson
July 10, 2011

Scalable, Loop-Free BGP FRR using Repair Label
draft-bashandy-idr-bgp-repair-label-02.txt

Abstract

Consider a BGP free core scenario. Suppose the provider edge BGP speakers PE1, PE2,..., PEn know about a prefix P/p via the external routers CE1, CE2,..., CEm. If the PE router PEi loses connectivity to the primary path, whether it is another PE router or a CE router, it is desirable to immediately restore traffic by rerouting packets arriving to PEi and destined to the prefix P/p to one of the other PE routers that advertised P/p, say PEj, until BGP re-converges. However if the loss of connectivity of PEi to the primary path also resulted in the loss of connectivity between PEj and CEj, rerouting a packet before the control plane converges may result in a loop. In this document, we propose using a repair label for traffic restoration while avoiding loops. We propose advertising the ''repair'' label through BGP.

Status of this Memo

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

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

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 10, 2012.

Copyright Notice

Copyright (c) 2011 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
1.1. Conventions used in this document.....	4
1.2. Terminology.....	4
2. Protocol Operation.....	5
2.1. Control plane Operation.....	5
2.1.1. Additional Rules for allocating and advertising a Repair label.....	6
2.2. Forwarding Plane Operation.....	6
2.3. Example.....	8
3. How to Disseminate Repair Label Information.....	9
3.1.1. Structure of the Repair Label Path Attribute.....	10
3.1.2. Semantics of the Repair Label Attribute.....	11
3.1.3. Additional Rule when Forwarding Advertisements Containing the Repair Path Attribute.....	12
4. Security Considerations.....	12
5. IANA Considerations.....	12
6. Conclusions.....	12

7. References.....	13
7.1. Normative References.....	13
7.2. Informative References.....	13
8. Acknowledgments.....	14

1. Introduction

In a BGP free core, where traffic is tunneled between edge routers and edge routers assign labels to prefixes, BGP speakers advertise reachability information about prefixes and associate a local label with each prefix such as L3VPN [9], 6PE [10], and Softwire [8]. Suppose that a given edge router is chosen as the best next-hop for a prefix P/p. An ingress router that receives a packet from an external router and destined for the prefix P/p pushes the label advertised by the egress edge router and then "tunnels" the packet across the core to that egress router. Upon receiving the labeled packet from the core, the egress router uses the label on the packet to take the appropriate forwarding decision.

In modern networks, it is not uncommon to have a prefix reachable via multiple edge routers. One example is the best external path [7]. Another more common and widely deployed scenario is L3VPN [9] with multi-homed VPN sites. As an example, consider the L3VPN topology depicted in Figure 1.

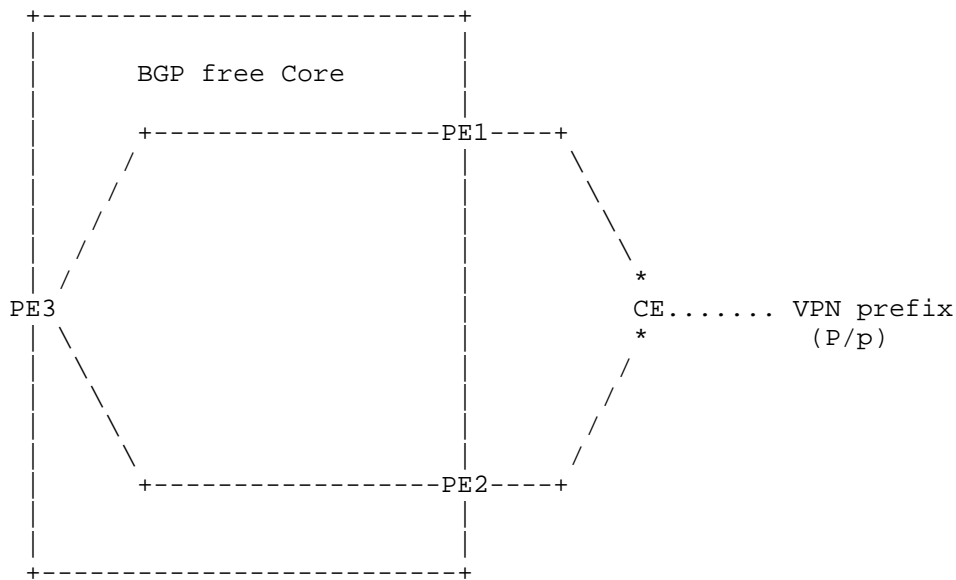


Figure 1 VPN prefix reachable via multiple PEs

PE3 is the ingress PE. PE1 and PE2 are both egress PEs connected to CE. CE advertises one or more VPN prefixes, denoted by P/p. PE1 and PE2 advertise P/p as VPNv4 or VPNv6 routes to all ingress PEs, including PE3, and associates a label with each route.

Suppose that the ingress PE, PE3, chooses PE1 as the next-hop for the prefix P/p. In order to minimize traffic loss, it is highly desirable for PE1 to reroute all traffic destined to P/p to PE2 as soon as the connectivity to CE is lost without waiting for the control plane (whether it is IGP or BGP) to re-converge and compute the new best path. In doing so, PE1 pushes the label advertised by PE2 for the prefix P/p, and then "tunnels" the packet to PE2. However if the loss of PE1-CE connectivity was due to CE crash, then PE2 will also reroute the traffic back to PE1, resulting in a loop. Due to ultra scalability requirements, where there is a need to support thousands of peers and hundreds of thousands of prefixes, there is a need to support quick traffic restoration without waiting for the control plane to converge and without risking loops.

1.1. 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 [1].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

1.2. Terminology

This section outlines the terms used in this document. For ease of use, we will use terms similar to those used by L3VPN [9]

- o Protected prefix: a prefix P/p (of any AFI) that a BGP speaker has an external path to. The BGP speaker may learn about the prefix from an external peer through BGP, some other protocol, or manual configuration. The protected prefix is advertised to some or all the internal peers.
- o Primary egress PE: an IBGP peer that can reach the protected prefix P/p through an external path and advertised the prefix to the other IBGP peers. The primary egress PE was chosen as the best path by one or more internal peers. In other words, the primary egress PE is an egress PE that will normally be used when there is no failure. Referring to Figure 1, PE1 is a primary egress PE.

- o CE: an external router through which an egress PE can reach a prefix P/p. The router "CE" in Figure 1 is an example of such a CE
- o Ingress PE: a BGP speaker that learns about a prefix through another IBGP peer and chooses that IBGP peer as the next-hop for the prefix. PE3 in Figure 1 is an example of an ingress PE
- o Repairing PE: the PE that attempts to restore traffic when the primary path is no longer reachable "without" waiting for BGP to re-converge. The repairing PE restores the traffic by rerouting the traffic (through a tunnel) towards the pre-calculated repair PE when it detects that the primary path is no longer reachable. The primary path may be a CE or another egress PE. Referring to Figure 1, if PE3 chooses PE1 as the primary egress PE and PE1 decides to reroute traffic to PE2 on losing reachability with CE, then PE1 is a repairing PE. If PE3 chooses PE1 as a primary path and PE3 decides to use PE2 as a repair path when it loses reachability to PE2, then PE3 is a repairing PE.
- o Primary label: the label advertised by the primary egress PE to be used for normal traffic forwarding.
- o Repair egress PE: an egress PE other than the primary egress PE that can reach the protected prefix P/p through an external neighbor. The repair PE is pre-calculated via other PEs prior to any failure
- o Repair label: the label that will be pushed on the packet when the repairing PE reroutes the traffic (through a tunnel) towards the repair egress PE. Section 2 discusses how the repair label is used. Section 3 discusses semantics of and the method for disseminating repair label information.
- o Repair path: the repair egress PE and the repair label.
- o internal and external: internal or external to the core.

2. Protocol Operation

This section explains the operation of the control and forwarding planes of routers participating in BGP-free core traffic restoration.

2.1. Control plane Operation

1. As usual, each PE allocates a local label for each prefix it can reach through an external neighbor CE. This is the primary label used for normal traffic forwarding.

2. To provide repair path information to all PEs, the PE also allocates a repair label to the prefix if it can reach that prefix via an external neighbor. Different repair label allocation schemes are proposed in Section 3.
3. The PE advertises both the primary and repair labels to all IBGP peers.
4. When a PE receives the label advertisement from egress PEs, it calculates a primary egress PE and a repair egress PE based on its internal path selection criteria. Note that the method of choosing the repair path is beyond the scope of this document.
5. In the end, for some of the prefixes advertised by more than one PE, a PE will have
 - o a primary path
 - o a repair path consisting of a repair PE and a repair label advertised by the chosen repair PE.
6. A PE "never" protects a repair label. Hence on any PE, a repair label only has paths towards the CE. However a primary label may have a repair path towards a chosen repair PE

2.1.1. Additional Rules for allocating and advertising a Repair label

- o A repair PE MUST NOT advertise a repair label for a prefix if it does NOT have an external path to the prefix
- o A repair PE MUST NOT associate an internal path with a repair label
- o Repair labels SHOULD be advertised with labeled address families only. That is AFI/SAFI 1/4, 2/4, 1/128, and 2/128.

2.2. Forwarding Plane Operation

This section specifies the forwarding plane operation when a PE receives a packet and any of the following two conditions are true:

- o The PE lost the primary path and has not yet calculated another primary path and programmed it in the forwarding plane. The primary path may be external or internal

- o The arriving packet arrived from the core and the PE does not have an external path. It is noteworthy to mention that this condition should be a temporary condition until all ingress PEs converge and stop sending traffic to that PE.

The forwarding plane processes arriving traffic as follows:

1. If the repairing PE is an egress PE, the packet arrives at the repairing PE with the primary label at the top because the packet is "tunneled" from the ingress PE(s). In that case, the repairing PE swaps the incoming label stack with the "repair label stack" advertised by the repair egress PE. Section 3.1.2. specifies all the details
2. If the repairing PE is an ingress PE, it MAY push the "repair label stack" advertised by the repair egress PE. Section 3.1.2. specifies all the details
3. The repairing PE tunnels the packet to the repair PE
4. At the repair PE, the packet arrives with the repair label at the top. The repair PE uses the incoming label stack to take forwarding decisions
5. If the repair egress PE can reach the CE, the repair PE forwards the packet towards the CE.
6. If the repair PE cannot reach the CE, the traffic will be dropped because a PE never protects a repair label

2.3. Example

Consider the L3VPN [9] topology depicted in Figure 2 where two PEs are connected to the same PE. Assume that the core is LDP. We will be using an advertised repair label.

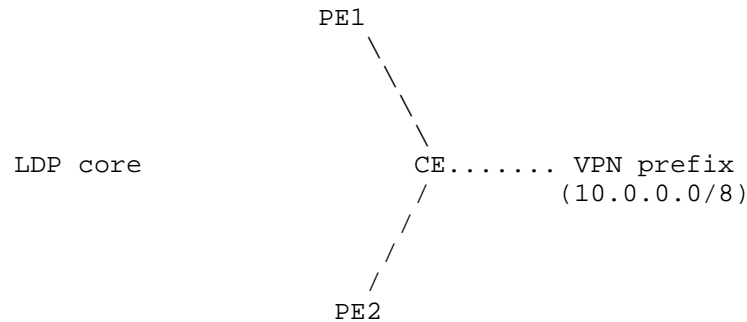


Figure 2 : L3VPN Example

```

PE1: Repairing egress PE
PE2: repair PE
Primary VPN label advertised by PE1 to all PEs: 4000
Repair VPN label advertised by PE1 to all PEs: 5000
Primary VPN label advertised by PE2 to all PEs: 2000
Repair VPN label advertised by PE2 all PEs: 3000

LDP label for PE2 on PE1 is 1234
LDP label for PE1 on PE2 is 4567

Before failure
//////////
PE1 has the following FIB entries

4000 -----> CE (unlabeled)
          -----> PE2, swap 4000 with 3000 and then push 1234
5000 -----> CE (unlabeled)

PE2 has the following
2000 -----> CE (unlabeled)
          -----> PE1, swap 2000 with 5000 and then push 4567
3000 -----> CE (unlabeled)

After the CE crashes
//////////
PE1 has the following entry:
4000 -----> PE2, swap 4000 with 3000 and then push 1234
  
```

5000 -----> Drop

PE2 has the following

2000 -----> PE1, swap 2000 with 5000 and then push 4567

3000 -----> Drop

Because of the above routing entries, any traffic arriving from the core at PE1 and destined for 10.0.0/8, is rerouted towards PE2 using the repair VPN label 3000. PE2 will just drop it instead of looping it back towards PE1.

After the link between PE1 and CE fails (CE did not crash)

////////////////////////////////////

PE1 has the following entry:

4000 -----> PE2, swap 4000 with 3000 and then push 1234

5000 -----> Drop

PE2 has the following

2000 -----> CE (unlabeled)

-----> PE1, swap 2000 with 5000 and then push 4567

3000 -----> CE

Because of the above routing entries, any traffic arriving from the core at PE1 and destined for 10.0.0/8 is rerouted towards PE2 using the repair VPN label 3000. PE2 will forward the traffic towards CE.

3. How to Disseminate Repair Label Information

We propose to advertise the repair label as an optional path attribute. Advertising the repair label as an optional path attributes has some advantages:

- o An egress PE can benefit from a scalable repair label allocation schemes such as per-CE repair label allocation
- o Allows the repairing PE to share the same repair path among multiple protected prefixes. Since the repair path is shared by all labels sharing the path attribute, the repairing PE can optimize its RIB and FIB by sharing the same repair path data structure among a large number of protected prefixes.
- o Reduces the BGP update message size. Instead of having to send additional labels per prefix, multiple prefixes can share the same repair label

- o The number of labels used for traffic restoration does not depend on the number of protected prefixes
- o Allows for incremental deployment because the attribute is optional

The main disadvantage of sharing the same repair path among multiple primary paths is loss of fine grain control. It is not possible to manage, control, or provide differentiated handling to traffic on per prefix basis until the network re-converges. The loss of fine grain control is limited to the BGP re-convergence period.

It is noteworthy to mention that per-CE repair label allocation has some advantages over per-prefix repair label allocation. First it results in using fewer labels. Second it allows for better packing in BGP messages. Third it does not require special handling in the forwarding plane at the repair PE. Fourth it simplifies the forwarding plane while maximizing the packet switching performance because the egress PE can take a forwarding decision with a single FIB lookup.

3.1.1. Structure of the Repair Label Path Attribute

This document defines the repair label attribute as an optional non-transitive path attribute [2] as follows:

Attribute name: REPAIR_LABEL

Type code: TBD

Attribute Flags:

Optional bit: 1

Transitive bit: 0

Partial bit: 0

Extended Length bit: 0

Length of the attribute: length in octets of the attribute

Attribute Value: The attribute value contains a stack of one or more labels. The encoding of the labels is identical to encoding of the "label" field in [4]. The value of the bottom of stack (BOS) bit is determined at traffic restoration time as specified in Section 3.1.2.

3.1.2. Semantics of the Repair Label Attribute

This document specifies the semantics of the repair label attribute when the attribute carries one repair label only. The semantics of more than one repair label is beyond the scope of this document.

Suppose a BGP speaker PE1 receives an update message with a repair label attribute containing the label "Lr2" from the IBGP peer PE2. Suppose the NLRI in the MP_REACH_NLRI attribute [3] contains the prefixes R1, R2, ..., Rn each bound to a label L21, L22, ..., L2n, respectively. This means the following:

1. PE2 will never attempt to repair a packet arriving with the label "Lr2". Hence PE2 will either forward the packet to an external CE or drop the packet
2. PE2 expects the following from PE1:
 - a. Case a: The route Ri on PE1 is bound to a local label "Lli". Suppose PE1 receives a packet with the label "Lli" at the top of the stack. If the PE1 loses the primary path for a prefix Ri or PE1 receives a packet from the core while not having an external path, and PE1 decides that PE2 is the repair PE for the prefix Ri, then PE1 has to swap the label "Lli" on the packet with the repair label "Lr2" and then tunnel the packet to PE2. The bottom of stack (BOS) bit MUST be copied from the label arriving on the packet to the label "Lr2"
 - b. Case b: The route Ri on PE1 is not bound to any local label. If the PE1 loses the primary path for a prefix Ri and PE1 decides that PE2 is the repair PE for the prefix Ri, then PE1 MAY push the label "Lr2" and then tunnel the packet to PE2. The bottom of stack (BOS) bit in "Lr2" MUST be set as specified in [5].
 - c. Case c: The route Ri on PE1 is bound to an aggregate label (e.g. per-vrf label). In that case, PE1 has to perform more than one route lookup to determine the primary path. Eventually, there will either be an IP lookup or a label lookup that points to the primary path:
 - i. A label lookup points to the primary path: In that case, PE1 handles the packet as described in item 2.a above.
 - ii. An IP lookup points to the primary path: In that case, PE1 handles the packet similar described in item 2.b above.

3.1.3. Additional Rule when Forwarding Advertisements Containing the Repair Path Attribute

As specified in Section 3.1.1, the repair label attribute is a non-transitive attribute. However there may be cases, such as inter-AS option (b)[9], route reflectors [11], or confederation [12], where a router may replace the advertised next-hop with its own before forwarding an advertisement. If a BGP speaker replaces the next-hop attribute with its own and the advertisement contains a repair label attribute with label stack "Sr", there are two options

- o Option 1: The BGP speaker MUST NOT advertise the repair label attribute
- o Option 2: The BGP speaker MUST replace the repair label stack "Sr" with a locally allocated label stack "Sr1" before advertising the route and then advertise the stack "Sr1" in the repair label attribute. For the forwarding plane, the BGP speaker MUST install a swap forwarding entry such that if the BGP speaker receives a packet with the label stack "Sr1", it swaps "Sr1" with the stack "Sr".

Note that advertising the repair label attribute by the router depends on whether the router understands the semantics of and supports the repair label attribute at the time of receiving an advertisement containing the repair label attribute.

4. Security Considerations

No additional security risk is introduced by using the mechanisms proposed in this document

5. IANA Considerations

This document defines a new BGP path attribute. IANA maintains a list of the current BGP attribute typecodes in [6]. This document proposes defining a new typecode value of "TBD" for the REPAIR_LABEL path attribute

6. Conclusions

This document proposes using a repair label to allow restoring traffic prior to BGP convergence while avoiding loops

7. References

7.1. Normative References

- [1] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [2] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006
- [3] Bates, T., Chandra, R., Katz, D., and Rekhter Y., "Multiprotocol Extensions for BGP", RFC 4760, January 2007
- [4] Rosen, E., Rekhter, Y., "Carrying Label Information in BGP-4", RFC 3107, May 2001
- [5] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T. and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.

7.2. Informative References

- [6] BGP Parameters, <http://www.iana.org/assignments/bgp-parameters/bgp-parameters.xhtml>
- [7] Marques, P., Fernando, R., Chen, E., Mohapatra, P., "Advertisement of the best external route in BGP", draft-ietf-idr-best-external-02.txt, April 2004.
- [8] Wu, J., Cui, Y., Metz, C., and E. Rosen, "Softwire Mesh Framework", RFC 5565, June 2009.
- [9] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [10] De Clercq, J. , Ooms, D., Prevost, S., Le Faucheur, F., "Connecting IPv6 Islands over IPv4 MPLS Using IPv6 Provider Edge Routers (6PE)", RFC 4798, February 2007
- [11] Bates, T., Chen, E., and Chandra, R., "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006
- [12] Traina, P., McPherson, P., and Scudder, J., "Autonomous System Confederations for BGP", RFC 5065, August 2007

8. Acknowledgments

Special thanks to Keyur Patel, Robert Raszuk, and Eric Rosen for the valuable comments

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

Authors' Addresses

Ahmed Bashandy
Cisco Systems
170 West Tasman Dr, San Jose, CA 95134
Email: bashandy@cisco.com

Burjiz Pithawala
Cisco Systems
170 West Tasman Dr, San Jose, CA 95134
Email: bpithaw@cisco.com

Jakob Heitz
Ericsson
100 Headquarters Drive, San Jose, CA, 95134
Email: jakob.heizt@ericsson.com

IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 1, 2012

D. Freedman
Claranet
R. Raszuk
Cisco Systems
R. Shakir
C&W
June 30, 2011

BGP OPERATIONAL Message
draft-frs-bgp-operational-message-00

Abstract

The BGP Version 4 routing protocol (RFC4271) is now used in many ways, crossing boundaries of administrative and technical responsibility.

The protocol lacks an operational messaging plane which could be utilised to diagnose, troubleshoot and inform upon various conditions across these boundaries, securely, during protocol operation, without disruption.

This document proposes a new BGP message type, the OPERATIONAL message, which can be used to effect such a messaging plane for use both between and within Autonomous Systems.

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 1, 2012.

Copyright Notice

Copyright (c) 2011 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. Applications	4
3. BGP OPERATIONAL message	5
3.1. BGP OPERATIONAL message capability	5
3.2. BGP OPERATIONAL message encoding	5
3.3. PRI Format	6
3.4. BGP OPERATIONAL message TLVs	9
3.4.1. ADVISE TLVs	9
3.4.2. STATE TLVs	10
3.4.3. DUMP TLVs	11
3.4.4. CONTROL TLVs	13
4. On the use of STATE and DUMP TLVs	16
5. On the use of ADVISE TLVs	17
6. Error Handling	19
7. Security considerations	20
8. IANA Considerations	21
9. Acknowledgements	23
10. References	24
10.1. Normative References	24
10.2. Informative References	24
Authors' Addresses	26

1. Introduction

In this document, a new BGP message type, the OPERATIONAL message is defined, creating a communication channel over which messages can be passed, using a series of contained TLV elements.

The messages can be human readable, for the attention of device operators or machine readable, in order to provide simple self test routines, which can be exchanged between BGP speakers.

A number of TLV elements will be assigned to provide for these message types, along with TLV elements to assist with description of the message data, such as describing precisely BGP prefixes and encapsulating BGP UPDATE messages to be sent back for inspection in order to troubleshoot session malfunctions.

The use of OPERATIONAL messages will be negotiated by BGP Capability [RFC5492], since the messages are in-band with the BGP session, they can be assumed to either be authenticated as originating directly from the BGP neighbor.

The goal of this document is to provide a simple, extensible framework within which new messaging and diagnostic requirements can live.

2. Applications

The authors would like to propose three main applications which BGP OPERATIONAL TLVs are designed to address. New TLVs can be easily added to enhance further current applications or to propose new applications.

The set of TLVs is organised in the following four functional groups comprising the three applications and some control messaging:

- o ADVISE TLVs, designed to convey human readable information to be passed, cross boundary to operators, to inform them of past or upcoming error conditions, or provide other relevant, in-band operational information. The "Advisory Demand Message" ADM (Section 3.4.1.1) is an example of this.
- o STATE TLVs, designed to carry information about BGP state across BGP neighbors, including both per-neighbor and global counters.
- o DUMP TLVs, designed to describe or encapsulate data to assist in realtime or post-mortem diagnostics, such as structured representations of affected prefixes / NLRI and encapsulated raw UPDATE messages for inspection.
- o CONTROL TLVs, designed to facilitate control messaging such as replies to requests which can not be satisfied.

Means concerning the reporting of information carried by these TLVs, either in reply or request processing are implementation specific but could include methods such as SYSLOG.

3. BGP OPERATIONAL message

3.1. BGP OPERATIONAL message capability

A BGP speaker that is willing to exchange BGP OPERATIONAL Messages with a neighbor should advertise the new OPERATIONAL Message Capability to the neighbor using BGP Capabilities advertisement [RFC5492] . A BGP speaker may send an OPERATIONAL message to its neighbor only if it has received the OPERATIONAL message capability from them.

The Capability Code for this capability is specified in the IANA Considerations section of this document.

The Capability Length field of this capability is 2 octets.

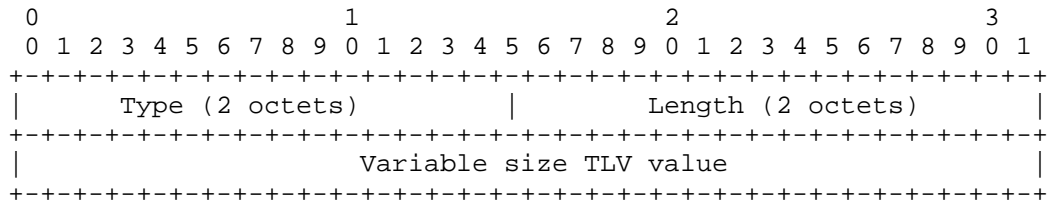
```
+-----+
| Capability Code (1 octet) |
+-----+
| Capability Length (1 octet) |
+-----+
```

OPERATIONAL message BGP Capability Format

3.2. BGP OPERATIONAL message encoding

The BGP message as defined [RFC4271] consists of a fixed-size header followed by two octet length field and one octet of type value. The RFC limits the maximum message size to 4096 octets. As one of the applications of BGP OPERATIONAL message (through the MUD (Section 3.4.3.3) message) is to be able to carry an entire, potentially malformed BGP UPDATE, this specification mandates that when the neighbor has negotiated the BGP OPERATIONAL message capability, any further BGP message which may be subject enclosure within a BGP OPERATIONAL message must be sent with the maximum size reduced to accommodate for the potential need of additional wrapping header size requirements. This is applicable to both the current BGP maximum message size limit or for any future modifications.

For the purpose of the OPERATIONAL message information encoding we will use one or more Type-Length-Value containers where each TLV will have the following format:



OPERATIONAL message TLV Format

TYPE: 2 octet value indicating the TLV type

LENGTH: 2 octet value indicating the TLV length in octets

VALUE: Variable length value field depending on the type of the TLVs carried.

To work around continued BGP churn issues some types of TLVs will need to contain a sequence number to correlate a request with associated replies. The sequence number will consist of 8 octets and will be of the form: (4 octet bgp_router_id) + (local 4 octet number). When the local 4 octet number reaches 0xFFFF it should restart from 0x0000. The sequence number is only used if the TLV requires sequencing else it is not included.

The typical application scenario for use of the sequence number is for it to be included in a request TLV to be copied into associated reply messages in order to correlate requests with their associated replies.

3.3. PRI Format

Prefix Reachability Indicators (PRI) are used to represent prefix NLRI and BGP attributes in a request and only prefix NLRI in a response, in this draft.

Each PRI is encoded as a 3-tuple of the form <Flags, Payload Type, Payload> whose fields are described below:

b) Payload Type:

This one octet type specifies the type and geometry of the payload.

ba) Type 0 - NLRI:

The payload contains (perhaps multiple) NLRI, the format of each NLRI is as defined in the base specification of such NLRI appropriate for the AFI/SAFI.

bb) Type 1 - Next Hop:

The payload contains a Next Hop address, appropriate for the AFI/SAFI. When used in an SSQ (Section 3.4.2.7) message the response is expected to contain prefixes from the selected RIBs which contain this next-hop in their next-hop attribute.

bc) Type 2 - AS Number:

The payload contains a 16 or 32 bit AS number (as defined in [RFC4893]), when used in an SSQ message the response is expected to contain prefixes from the selected RIBs which contain this AS number in their AS_PATH or AS4_PATH (as appropriate) attributes.

bc) Type 3 - Standard Community:

The payload contains a standard community (as defined in [RFC1997]), when used in an SSQ message the response is expected to contain prefixes from the selected RIBs which contain this standard community in their communities attribute.

bd) Type 4 - Extended Community:

The payload contains an extended community (as defined in [RFC4360]), when used in an SSQ message the response is expected to contain prefixes from the selected RIBs which contain this standard community in their extended communities attribute.

be) Types 5-65535 - Reserved:

Types 5-65535 are reserved for future use.

c) Payload:

Contains the actual payload, as defined by the payload type, the payload is of variable length, to be calculated from the remaining TLV length.

PRI are used for both request and response modes, a response MUST only contain an NLRI (type 0) payload but a request MAY contain payloads specifying a type to search for, an implementation MUST validate all PRI it receives in a request against the type of request which was made.

An implementation MUST NOT send a PRI in response with no NLRI (type 0) payload, this is considered to be invalid. If the implementation wishes to signal that a request did not yield any valid results an implementation MAY respond with an NS TLV (Section 3.4.4.2), using the "Not Found" subcode, for example.

3.4. BGP OPERATIONAL message TLVs

3.4.1. ADVISE TLVs

ADVISE TLVs convey human readable information to be passed, cross boundary to operators, to inform them of past or upcoming error conditions, or provide other relevant, in-band operational information.

3.4.1.1. Advisory Demand Message (ADM)

TYPE: 1 - ADM

LENGTH: 3 Octets(AFI+SAFI) + Variable value (up to 2K octets)

USE: To carry a message, on demand, comprised of a string of UTF-8 characters (up to 2K octets in size), with no null termination. Upon reception, the string SHOULD be reported to the host's administrator.

Implementations SHOULD provide their users the ability to transmit a free form text message generated by user input.

3.4.1.2. Advisory Static Message (ASM)

TYPE: 2 - ASM

LENGTH: 3 Octets(AFI+SAFI) + Variable value (up to 2K octets)

USE: To carry a message, on demand, comprised of a string of UTF-8 characters, with no null termination. Upon reception, the string SHOULD be stored in the BGP neighbor statistics field within the router. The string SHOULD be accessible to the operator by executing CLI commands or any other method (local or remote) to obtain BGP neighbor statistics (e.g. NETCONF, SNMP).

The expectation is that the last ASM received from a BGP neighbor will be the message visible to the operator (the most current ASM).

Implementations SHOULD provide their users the ability to transmit a free form text message generated by user input.

3.4.2. STATE TLVs

STATE TLVs reflect, on demand, the internal state of a BGP neighbor as seen from the other neighbor's perspective.

3.4.2.1. Reachable Prefix Count Request (RPCQ)

TYPE: 3 - RPCQ

LENGTH: 3 Octets(AFI+SAFI) + Sequence Number

USE: Sent to the neighbor to request that an RPCP (Section 3.4.2.2) message is generated in response.

3.4.2.2. Reachable Prefix Count Reply (RPCP)

TYPE: 4 - RPCP

LENGTH: 3 Octets(AFI+SAFI) + Sequence Number + 4 Octet RX Prefix Counter (RXC) + 4 Octet TX Prefix Counter (TXC)

USE: Sent in reply to an RPCQ (Section 3.4.2.1) message from a neighbor, RXC is populated with the number of reachable prefixes accepted from the peer and TXC with the number of prefixes to be transmitted to the peer for the AFI/SAFI.

3.4.2.3. Adj-Rib-Out Prefix Count Request (APCQ)

TYPE: 5 - APCQ

LENGTH: 3 Octets(AFI+SAFI) + Sequence Number

USE: Sent to the neighbor to request that an APCP (Section 3.4.2.4) message is generated in response.

APCQ can be used as a simple mechanism when an implementation does not permit or support the use of RPCQ.

3.4.2.4. Adj-Rib-Out Prefix Count Reply (APCP)

TYPE: 6 - APCP

LENGTH: 3 Octets(AFI+SAFI) + Sequence Number + 4 Octet TX Prefix Counter (TXC)

USE: Sent in reply to an APCQ (Section 3.4.2.3) message from a neighbor, TXC is populated with the number of prefixes held in the Adj-Rib-Out for the neighbor for the AFI/SAFI.

3.4.2.5. BGP Loc-Rib Prefix Count Request (LPCQ)

TYPE: 7 - LPCQ

LENGTH: 3 Octets(AFI+SAFI) + Sequence Number

USE: Sent to the peer to request that an LPCP (Section 3.4.2.6) message is generated in response.

3.4.2.6. BGP Loc-Rib Prefix Count Reply (LPCP)

TYPE: 8 - LPCP

LENGTH: 3 Octets(AFI+SAFI) + Sequence Number + 4 Octet Loc-Rib Counter (LC)

USE: Sent in reply to an LPCQ (Section 3.4.2.5) message from a neighbor, LC is populated with the number of prefixes held in the entire Loc-Rib for the AFI/SAFI.

3.4.2.7. Simple State Request (SSQ)

TYPE: 9 - SSQ

LENGTH: 3 Octets(AFI+SAFI) + Sequence Number + Single request PRI (Variable)

USE: Using a PRI as a request form (See Section 3.3), an implementation can be asked to return information about prefixes found in various RIBs.

A single, simple PRI is used in the request, containing a single NLRI or attribute as the PRI payload. RIB response filtering may take place through the setting of the I, O and L bits in the PRI Flags field.

An implementation MAY respond to an SSQ TLV in with an SSP (See Section 3.4.3.4) TLV (containing the appropriate data). An implementation MAY also respond to an SSQ with an NS TLV (with the appropriate subcode set) indicating why there will not be an SSP TLV in response. An implementation MAY also not respond at all (See Section 7).

3.4.3. DUMP TLVs

DUMP TLVs provide data in both structured and unstructured formats in response to events, for use in debugging scenarios.

3.4.3.1. Dropped Update Prefixes (DUP)

TYPE: 10 - DUP

LENGTH: 3 Octets(AFI+SAFI) + Variable number of dropped UPDATE Prefix Reachability Indicators (PRI) (See Section 3.3)

USE: To report to a neighbor a structured set of prefix reachability indicators retrievable from the last dropped UPDATE message, sent in response to an UPDATE message which was well formed but not accepted by the neighbor by policy.

For example, an UPDATE which was dropped and the rescued NLRI concerned a number of both reachable and unreachable prefixes, the DUP would encapsulate two PRI, one with the R-Bit (reachable) set, housing the rescued reachable NLRI and the other with the R-Bit cleared (unreachable), housing the rescued unreachable NLRI as payload.

3.4.3.2. Malformed Update Prefixes (MUP)

TYPE: 11 - MUP

LENGTH: 3 Octets(AFI+SAFI) + Variable number of dropped update Prefix Reachability Indicators (PRI) (See Section 3.3) due to UPDATE Malformation.

USE: To report to a neighbor a structured set of prefix reachability indicators retrievable from the last UPDATE message dropped through malformation, sent in response to an UPDATE message which was not well formed and not accepted by the neighbor, where a NOTIFICATION message was not sent. A MUP TLV may accompany a MUD (Section 3.4.3.3) TLV.

See the example from Section 3.4.3.1.

3.4.3.3. Malformed Update Dump (MUD)

TYPE: 12 - MUD

LENGTH: 3 Octets(AFI+SAFI) + Variable length representing retrievable malformed update octet stream.

USE: To report to a peer a copy of the last UPDATE message dropped through malformation, sent in response to an UPDATE message which was not well formed and not accepted by the neighbor, where a NOTIFICATION message was not sent. A MUD TLV may accompany a MUP (Section 3.4.3.2) TLV.

3.4.3.4. Simple State Response (SSP)

TYPE: 13 - SSP

LENGTH: 3 Octets(AFI+SAFI) + Sequence Number + Single Response PRI (Variable)

USE: Using a PRI as a response form (See Section 3.3), an implementation uses the SSP TLV to return a response to an SSQ (See Section 3.4.2.7) TLV which should contain information about prefixes found in various RIBs. These RIBs should be walked to extract the information according to local policy.

A single, simple PRI is used in the response, containing multiple NLRI. The I, O and L bits in the PRI Flags field should be set indicating which RIBs the prefixes were found in.

An implementation MAY respond to an SSQ TLV in with an SSP TLV (containing the appropriate data). An implementation MAY also respond to an SSQ with an NS TLV (with the appropriate subcode set) indicating why there will not be an SSP TLV in response. An implementation MAY also not respond at all (See Section 7).

If no data is found to satisfy a query which is permitted to be answered, an implementation MAY respond with an NS TLV with the subcode "Not Found" to indicate that no data was found in response to the query. An implementation MUST NOT send a PRI in response with no NLRI payload, this is considered to be invalid.

3.4.4. CONTROL TLVs

CONTROL TLVs satisfy control mechanism messaging between neighbors, they are used for such functions as to refuse messages and dynamically signal OPERATIONAL capabilities to neighbors during operation.

3.4.4.1. Max Permitted (MP)

TYPE: 65534 - MP

LENGTH: 3 Octets(AFI+SAFI) + 2 Octet Value

USE: The Max Permitted TLV is used to signal to the neighbor the maximum number of OPERATIONAL messages that will be accepted in a second of time (see Section 7, Security Considerations), an implementation MUST, on receipt of an MP TLV, ensure that it does not exceed the rate specified in the MP TLV for sending OPERATIONAL messages to the neighbor, for the duration of the session.

An implementation MAY send subsequent MP TLVs during the session's lifetime, updating the maximum acceptable rate

MP TLVs MAY be rate limited by the receiver as part of OPERATIONAL rate limiting (see Section 7, Security Considerations).

3.4.4.2. Not Satisfied (NS)

TYPE: 65535 - NS

LENGTH: 3 Octets(AFI+SAFI) + Sequence Number + 2 Octet Error Subcode

USE: To respond to a query to indicate that the implementation can or will not answer this query. The following subcodes are defined:

0x01 - Request TLV Malformed: Used to signal to the neighbor that the request was malformed and will not be processed. A neighbor on receiving this message MAY re-transmit the request but MUST increment the sequence number. Implementations SHOULD ensure that the same request is not retransmitted excessively when repeatedly receiving this Error Subcode in response.

0x02 - TLV Unsupported for this neighbor: Used to signal to the neighbor that the request was unsupported and will not be processed. A neighbor on receiving this message MUST NOT retransmit the request for the duration of the session.

0x03 - Max query frequency exceeded: Used to signal to the neighbor that the request has exceeded the rate at which the neighbor finds acceptable for the implementation to transmit requests at, see Section 3.4.4.1 (MP TLV) and Section 7 and (Security Considerations) for more information.

0x04 - Administratively prohibited: Used to signal to the neighbor that the request was administratively prohibited and will not be processed. A neighbor on receiving this message MUST NOT retransmit the request for the duration of the session.

0x05 - Busy: Used to signal to the neighbor that the request will not be replied to, due to lack of resources estimated to satisfy the request. It is suggested that, on receipt of this error subcode a message is logged to inform the operator of this failure as opposed to automatically attempting to re-try the previous query.

0x06 - Not Found: Used to signal to the neighbor that the request would have been replied to but does not contain any data (i.e the data was not found). An implementation MUST NOT send a PRI response with no NLRI payload, this is considered to be invalid.

NS TLVs MAY be rate limited by the receiver as part of OPERATIONAL rate limiting (see Section 7, Security Considerations).

4. On the use of STATE and DUMP TLVs

The STATE TLVs use three classes of counters, defined in this document: sent counters (TXC), received counters (RXC) and current table state counters (LC). The table state counters (for example number of BGP RIB entries) are exchanged only for informational purposes and they should not be subject to comparison with any local counter values.

Where a query of the neighbor's RXC is required to be correlated, the local TXC coupled with the sequence number SHOULD be stored and used to perform such a correlation. If a discrepancy is detected, an automated or manual Route Refresh message can be triggered (utilising Start_of_Refresh and End_of_Refresh markers) that would allow for purge of any stalled data across two BGP databases.

It is important to note that, as BGP is never stable it is expected that the counters will also be subject to continues value change making any comparison of their values questionable.

The DUMP TLVs report information back to an operator about messages which were not accepted, from machine-readable rescued UPDATE NLRI to an entire copy of the malformed UPDATE message. These can be used for troubleshooting purposes when such a message is transmitted and the implementation gracefully continues (such as treat-as-withdraw).

5. On the use of ADVISE TLVs

The BGP routing protocol is used with external as well as internal neighbors to propagate route advertisements. In the case of external BGP sessions, there is typically a demarcation of administrative responsibility between the two entities. While initial configuration and troubleshooting of these sessions is handled via offline means such as email or telephone calls, there is gap when it comes to advising a BGP neighbor of a behaviour that is occurring or will occur momentarily. There is a need for operators to transmit a message to a BGP neighbor to notify them of a variety of types of messages. These messages typically would include those related to a planned or unplanned maintenance action. These ADVISE messages could then be interpreted by the remote party and either parsed via logging mechanisms or viewed by a human on the remote end via the CLI. This capability will improve operator NOC-to-NOC communication by providing a communications medium on an established and trusted BGP session between two autonomous systems.

The reason that this method is preferred for NOC-to-NOC communications is that other offline methods do fail for a variety of reasons. Emails to NOC aliases ahead of a planned maintenance may have ignored the mail or may have not recorded it properly within an internal tracking system. Even if the message was recorded properly, the staff that are on-duty at the time of the maintenance event typically are not the same staff who received the maintenance notice several days prior. In addition, the staff on duty at the time of the event may not even be able to find the recorded event in their internal tracking systems. The end result is that during a planned event, some subset of eBGP peers will respond to a session/peer down event with additional communications to the operator who is initiating the maintenance action. This can be via telephone or via email, but either way, it may result in a sizeable amount of replies inquiring as to why the session is down.

The result of this is that the NOC responsible for initiating the maintenance can be inundated with calls/emails from a variety of parties inquiring as to the status of the BGP session. The NOC initiating the maintenance may have to further inquire with engineering staff (if they are not already aware) to find out the extent of the maintenance and communicate this back to all of the NOCs calling for additional information. The above scenario outlines what is typical in a planned maintenance event. In an unplanned maintenance event (the need for and immediate router upgrade/reload), the number of calls and emails will dramatically increase as more parties are unaware of the event.

With the ADVISE TLV set, an operator can transmit an OPERATIONAL

message just prior to initiating the maintenance specifying what event will happen, what ticket number this event is associated with and the expected duration of the event. This message would be received by BGP peers and stored in their logs as well as any monitoring system if they have this capability. Now, all of the BGP peers have immediate access to the information about this session, why it went down, what ticket number this is being tracked under and how long they should wait before assuming there is an actual problem. Even smaller networks without the network management capabilities to correlate BGP events and OPERATIONAL messages would typically have an operator login to a router and examine the logs via the CLI.

This draft specifies two types of ADVISE TLV, a DEMAND message (ADM) and a STATIC message (ASM), it is anticipated that the DEMAND message will be used to send a message, on demand to the BGP neighbor, to inform them of realtime events. The STATIC message can be used to provide continual, "Sticky" information to the neighbor, such as a contact telephone number or e-mail address should there be a requirement to have continual access to this information.

6. Error Handling

An implementation MUST NOT send an OPERATIONAL message to a neighbor in response to an erroneous or malformed OPERATIONAL message. Any erroneous or malformed OPERATIONAL message received SHOULD be logged for the attention of the operator and then MAY be discarded.

7. Security considerations

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

Where a request type is not supported or allowed by an implementation for some reason, the implementation MAY send an NS (Section 3.4.4.2) TLV in response, the Error subcode of this TLV SHOULD be set according to the reason that this request will not be responded to.

Implementations MUST rate-limit the rate at which they transmit and receive OPERATIONAL messages. Specifically, an implementation MUST NOT allow the handling of OPERATIONAL messages to negatively impact any other functions on a router such as regular BGP message handling or other routing protocols.

Although an NS error subcode is provided to indicate that a request was rate-limited, an implementation need not reply to a request at all, this is the suggested course of action when rate-limiting the sending of responses to a neighbor.

An implementation MAY send an MP (Section 3.4.4.1) TLV to indicate the maximum rate at which it will accept OPERATIONAL messages from a neighbor, upon receipt of this TLV the sender MUST ensure it does not transmit above this rate for the duration of the session.

An implementation, considering a request to be too computationally expensive, MAY reply with the "Busy" NS error subcode to indicate such, though the implementation need not reply to the request.

Implementations MUST provide a mechanism for preventing access to information requested by SSR (Section 3.4.2.7) messages for the operator. Implementations SHOULD ensure that responses concerning the Loc-RIB (PRI with L-Bit set or responses which would set the L-Bit) are filtered in the default configuration.

8. IANA Considerations

IANA is requested to allocate a type code for the OPERATIONAL message from the BGP Message Types registry, as well as requesting a type code for the new OPERATIONAL Message Capability negotiation from BGP Capability Codes registry.

This document requests IANA to define and maintain a new registry named: "OPERATIONAL Message Type Values". The allocation policy is on a first come first served basis.

This document makes the following assignments for the OPERATIONAL Message Type Values:

ADVISE:

- * Type 1 - Advisory Demand Message (ADM)
- * Type 2 - Advisory Static Message (ASM)

STATE:

- * Type 3 - Reachable Prefix Count Request (RPCQ)
- * Type 4 - Reachable Prefix Count Response (RCPQ)
- * Type 5 - Adj-RIB-Out Prefix Count Request (APCQ)
- * Type 6 - Adj-RIB-Out Prefix Count Response (APCP)
- * Type 7 - Loc-Rib Prefix Count Request (LPCQ)
- * Type 8 - Loc-Rib Prefix Count Response (LPCP)
- * Type 9 - Simple State Request (SSQ)

DUMP:

- * Type 10 - Dropped Update Prefixes (DUP)
- * Type 11 - Malformed Update Prefixes (MUP)
- * Type 12 - Malformed Update Dump (MUD)
- * Type 13 - Simple State Response (SSP)

CONTROL:

- * Type 65534 - Max Permitted (MP)
- * Type 65535 - Not Satisfied (NS)

9. Acknowledgements

This memo is based on existing works [I-D.ietf-idr-advisory] and [I-D.raszuk-bgp-diagnostic-message] which describe a number of operational message types documented here. The authors would like to thank Enke Chen, Bruno Decraene, Alton Lo, Tom Scholl, John Scudder and Richard Steenbergen for their valuable input.

10. References

10.1. Normative References

- [RFC1997] Chandrasekeran, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, August 1996.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006.
- [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.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.

10.2. Informative References

- [I-D.ietf-idr-advisory]
Scholl, T., Scudder, J., Steenbergen, R., and D. Freedman, "BGP Advisory Message", draft-ietf-idr-advisory-00 (work in progress), October 2009.
- [I-D.jasinska-ix-bgp-route-server]
Jasinska, E., Hilliard, N., Raszuk, R., and N. Bakker, "Internet Exchange Route Server", draft-jasinska-ix-bgp-route-server-02 (work in progress), March 2011.
- [I-D.nalawade-bgp-inform]
Nalawade, G., Scudder, J., and D. Ward, "BGPv4 INFORM message", draft-nalawade-bgp-inform-02 (work in progress), August 2002.
- [I-D.nalawade-bgp-soft-notify]
Nalawade, G., "BGPv4 Soft-Notification Message", draft-nalawade-bgp-soft-notify-01 (work in progress), July 2005.

[I-D.raszuk-bgp-diagnostic-message]

Raszuk, R., Chen, E., and B. Decraene, "BGP Diagnostic Message", draft-raszuk-bgp-diagnostic-message-02 (work in progress), March 2011.

[I-D.retana-bgp-security-state-diagnostic]

Retana, A. and R. Raszuk, "BGP Security State Diagnostic Message", draft-retana-bgp-security-state-diagnostic-00 (work in progress), March 2011.

[I-D.shakir-idr-ops-reqs-for-bgp-error-handling]

Shakir, R., "Operational Requirements for Enhanced Error Handling Behaviour in BGP-4", draft-shakir-idr-ops-reqs-for-bgp-error-handling-01 (work in progress), February 2011.

Authors' Addresses

David Freedman
Claranet
London
UK

Email: david.freedman@uk.clara.net

Robert Raszuk
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US

Email: raszuk@cisco.com

Rob Shakir
Cable&Wireless Worldwide

Email: rob.shakir@cw.com

Inter-Domain Routing
Internet-Draft
Intended status: Standards Track
Expires: January 12, 2012

H. Gredler
J. Medved
Juniper Networks, Inc.
S. Previdi
Cisco Systems, Inc.
July 11, 2011

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

Abstract

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

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

Requirements Language

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

Status of this Memo

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

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

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any

time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on January 12, 2012.

Copyright Notice

Copyright (c) 2011 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. Scope	5
3. Transcoding Link State Information into a BGP NLRI	5
3.1. NLRI format	5
3.2. TLV Format	7
3.3. Node Descriptors	7
3.3.1. Local Node Descriptors	8
3.3.2. Remote Node Descriptors	8
3.3.3. Node Descriptor Sub-TLVs	9
3.3.4. Router-ID Anchoring Example: ISO Pseudonode	9
3.3.5. Router-ID Anchoring Example: OSPFv2 to IS-IS Migration	10
3.4. Link Descriptors	10
3.5. Link Attributes	11
3.5.1. MPLS Protocol TLV	12
3.5.2. TE Default Metric TLV	12
3.5.3. IGP Link Metric TLV	13
3.5.4. Shared Risk Link Group TLV	13
3.5.5. OSPF specific link attribute TLV	14
3.5.6. IS-IS specific link attribute TLV	14
3.6. Node Attributes	15
3.6.1. Node Flag Bits TLV	15
3.6.2. OSPF Specific Node Properties TLV	15
3.6.3. IS-IS Specific Node Properties TLV	16
3.7. IGP Area Information	16
3.8. Inter-AS Links	17
4. Link to Path Aggregation	17
4.1. Example: No Link Aggregation	17
4.2. Example: ASBR to ASBR Path Aggregation	18
4.3. Example: Multi-AS Path Aggregation	18
5. Originating the TED NLRI	18
6. Receiving the TED NLRI	19
7. Use Cases	19
7.1. MPLS TE	19
7.2. ALTO Server Network API	20
7.3. Path Computation Element (PCE) TED Synchronization Protocol	21
8. IANA Considerations	21
9. Security Considerations	21
10. Acknowledgements	21
11. References	22
11.1. Normative References	22
11.2. Informative References	23
Authors' Addresses	23

1. Introduction

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

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

A router maintains one or more databases for storing link-state information about nodes and links in any given area. Link attributes stored in these databases are: local/remote IP addresses, local/remote interface indices, metric, link bandwidth, reservable bandwidth, per CoS class reservation state, preemption and Shared Risk Link Groups (SRLG). The router's BGP process can retrieve topology from one of the link-state databases and distribute it to peer BGP Speakers using the encoding specified in this draft.

A BGP Speaker may distribute the real physical topology from the Link State database or the Traffic Engineering database, or create an abstracted topology, where virtual, aggregated nodes are connected by virtual paths. Aggregated nodes can be created, for example, out of multiple routers in a POP. Abstracted topology can also be a mix of physical and virtual nodes and physical and virtual links.

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

2. Scope

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

3. Transcoding Link State Information into a BGP NLRI

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

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

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

3.1. NLRI format

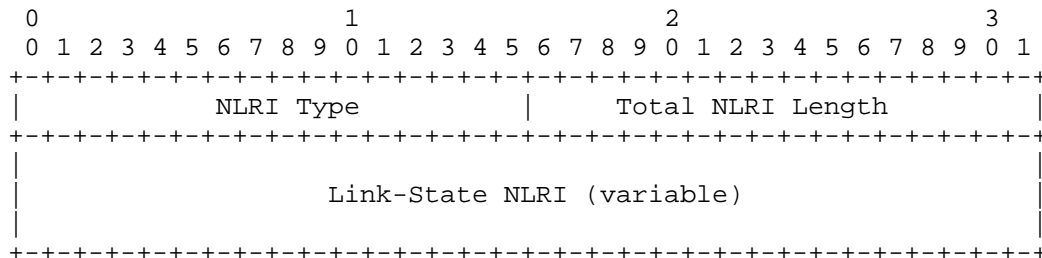


Figure 1: Link State SAFI 1 NLRI Format

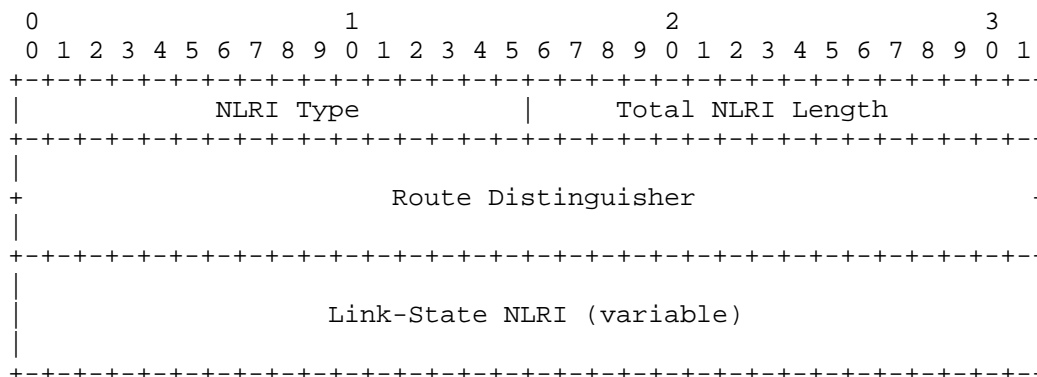


Figure 2: Link State SAFI 128 NLRI Format

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

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

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

Type = 2: Node NLRI, contains node attributes

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

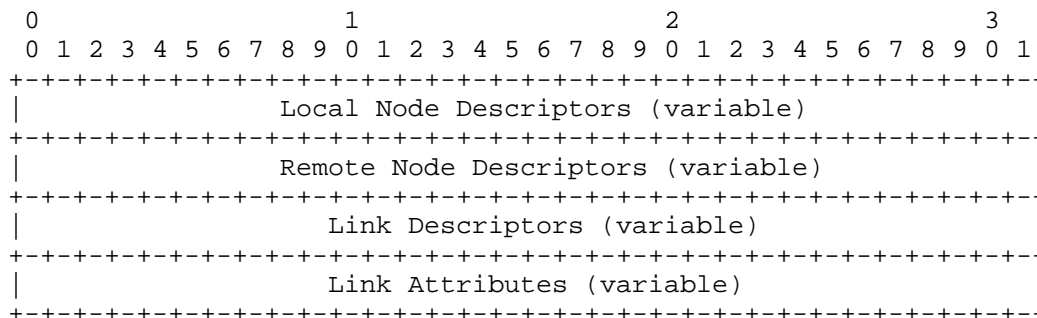


Figure 3: The Link NLRI format

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

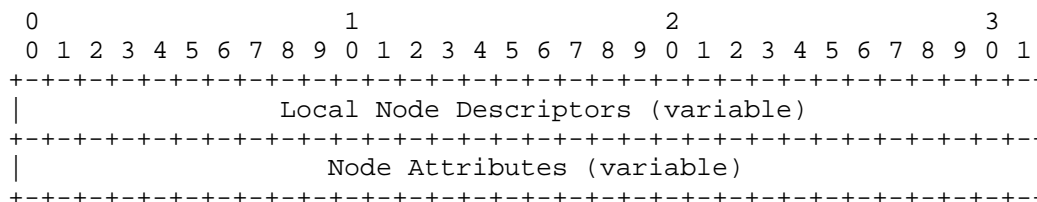


Figure 4: The Node NLRI format

3.2. TLV Format

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

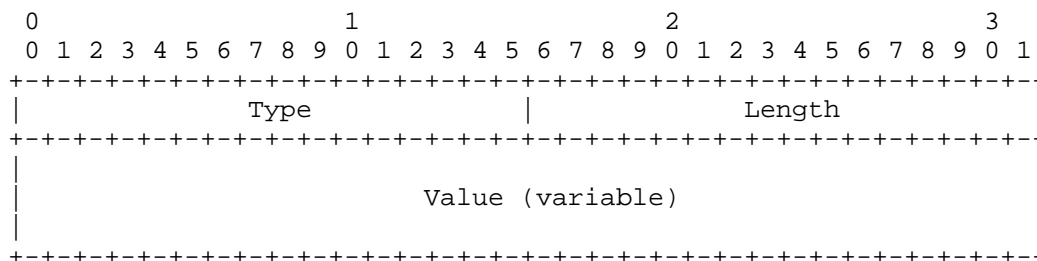


Figure 5: TLV format

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

3.3. Node Descriptors

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

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

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

3.3.1. Local Node Descriptors

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

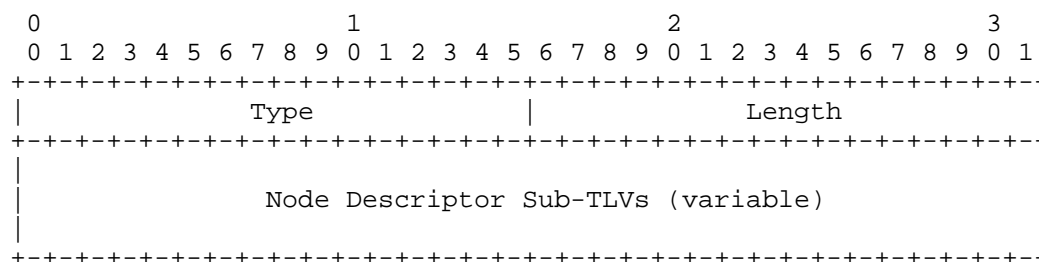


Figure 6: Local Node Descriptors TLV format

3.3.2. Remote Node Descriptors

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

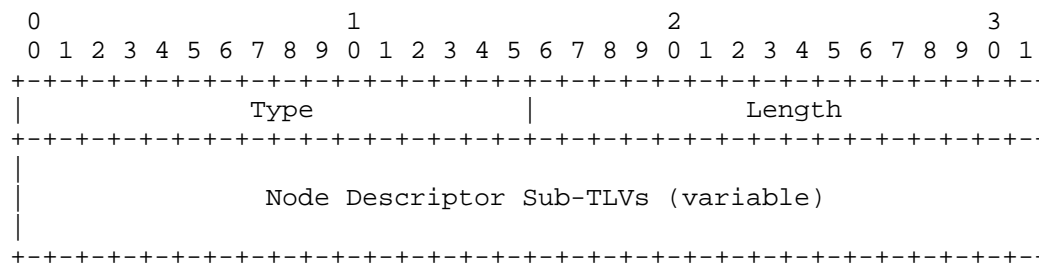


Figure 7: Remote Node Descriptors TLV format

3.3.3. Node Descriptor Sub-TLVs

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

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

Table 1: Node Descriptor Sub-TLVs

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

Autonomous System: opaque value (32 Bit AS ID)

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

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

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

3.3.4. Router-ID Anchoring Example: ISO Pseudonode

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

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

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

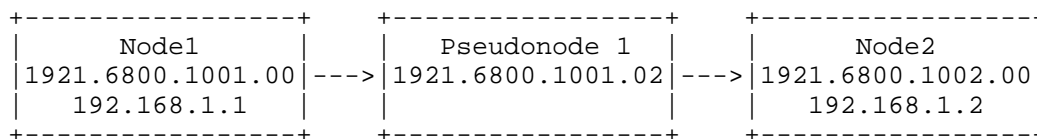


Figure 8: IS-IS Pseudonodes

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

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

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

3.4. Link Descriptors

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

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

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

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

Table 2: Link Descriptor TLVs

3.5. Link Attributes

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

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

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

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

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

Table 3: Link Attribute TLVs

3.5.1. MPLS Protocol TLV

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

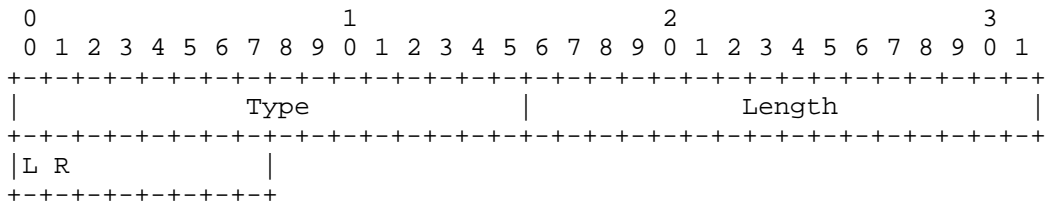


Figure 9: MPLS Protocol TLV

The following bits are defined:

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

Table 4: MPLS Protocol TLV Codes

3.5.2. TE Default Metric TLV

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

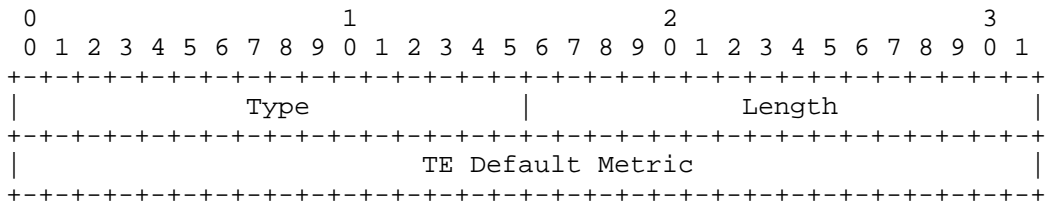


Figure 10: TE Default metric TLV format

3.5.3. IGP Link Metric TLV

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

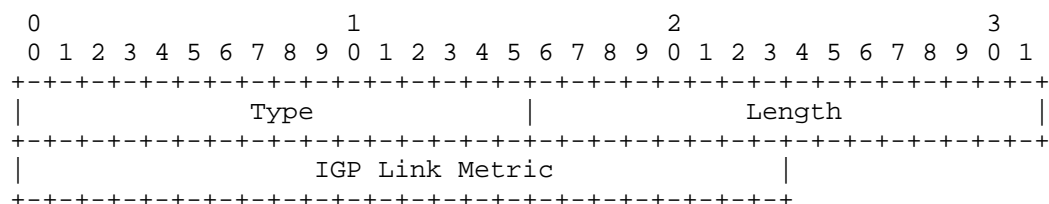


Figure 11: IGP Link Metric TLV format

3.5.4. Shared Risk Link Group TLV

The Shared Risk Link Group (SRLG) TLV (Type 64514) carries the Shared Risk Link Group information (see Section 2.3, "Shared Risk Link Group Information", of [RFC4202]). It contains a data structure consisting of a (variable) list of SRLG values, where each element in the list has 4 octets, as shown in Figure 12. The length of this TLV is 4 * (number of SRLG values).

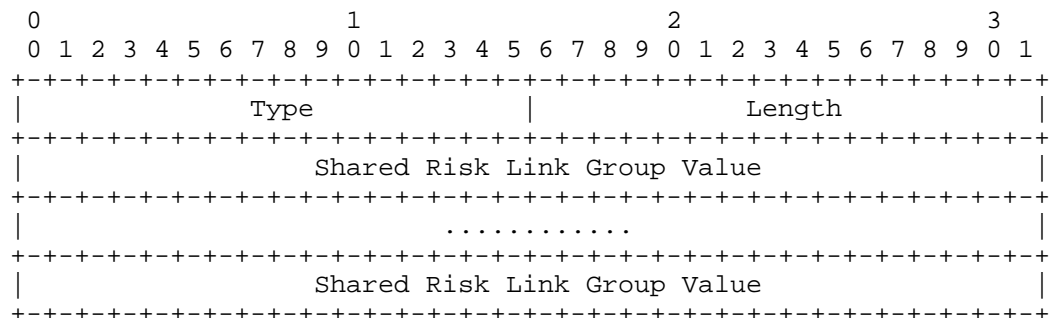


Figure 12: Shared Risk Link Group TLV format

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

3.5.5. OSPF specific link attribute TLV

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

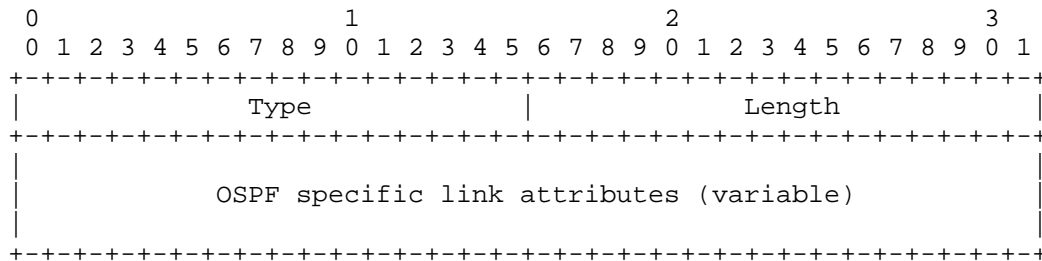


Figure 13: OSPF specific link attribute format

3.5.6. IS-IS specific link attribute TLV

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

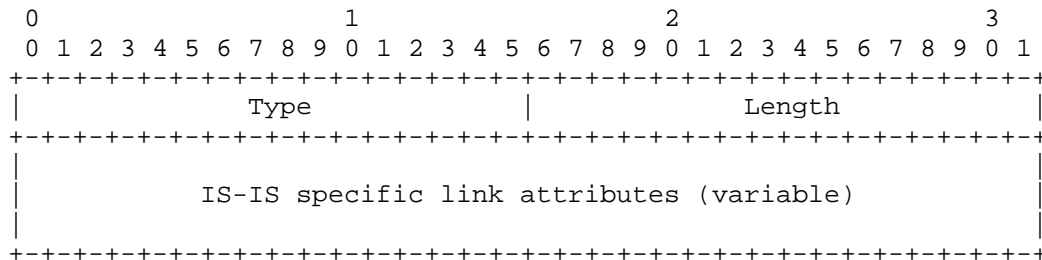


Figure 14: IS-IS specific link attribute format

3.6. Node Attributes

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

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

Table 5: Node Attribute TLVs

3.6.1. Node Flag Bits TLV

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

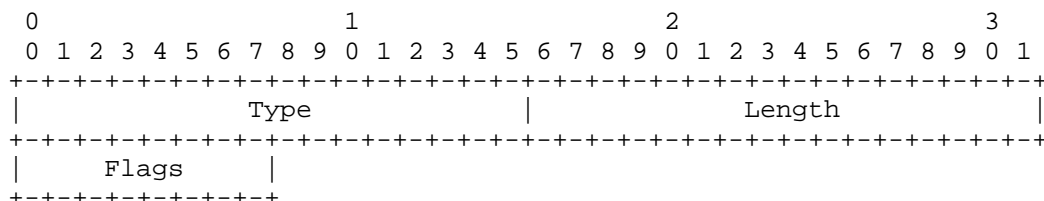


Figure 15: Node Flag Bits TLV format

The bits are defined as follows:

Bit	Description	Reference
0	Overload Bit	[RFC1195]
1	Attached Bit	[RFC1195]
2	External Bit	[RFC2328]
3	ABR Bit	[RFC2328]

Table 6: Node Flag Bits Definitions

3.6.2. OSPF Specific Node Properties TLV

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

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

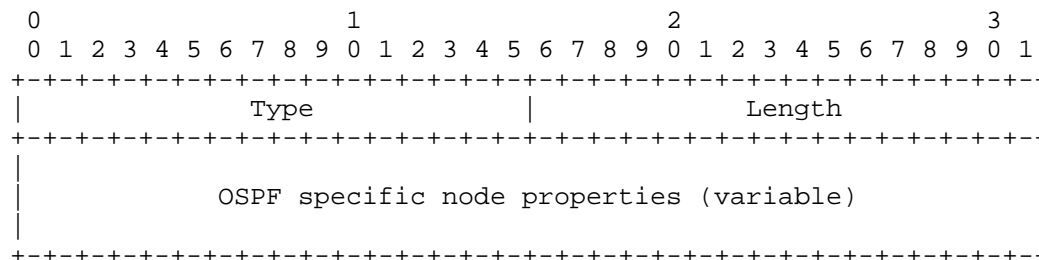


Figure 16: OSPF specific Node property format

3.6.3. IS-IS Specific Node Properties TLV

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

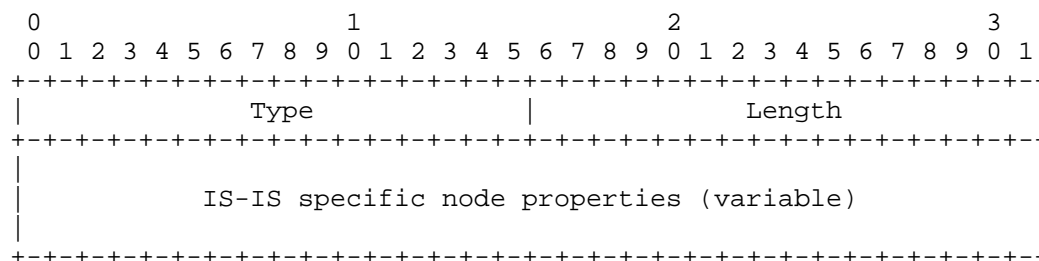


Figure 17: IS-IS specific Node property format

3.7. IGP Area Information

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

3.8. Inter-AS Links

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

4. Link to Path Aggregation

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

4.1. Example: No Link Aggregation

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

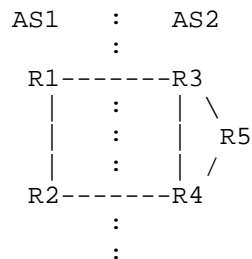


Figure 18: no-link-aggregation

4.2. Example: ASBR to ASBR Path Aggregation

The brief difference between the "no-link aggregation" example and this example is that no specific link gets exposed. Consider Figure 19. The only link which gets advertised by AS2 is an "aggregate" link between R3 and R4. This is enough to tell AS1 that there is a backup path. However the actual links being used are hidden from the topology.

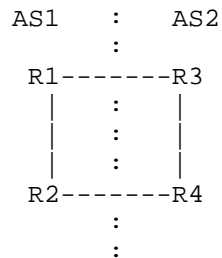


Figure 19: asbr-link-aggregation

4.3. Example: Multi-AS Path Aggregation

Service providers in control of multiple-ASes may even decide to not expose their internal inter-AS links. Consider Figure 20. Rather than exposing all specific R3 to R6 links, AS3 is modeled as a single node which connects to the border routers of the aggregated domain.

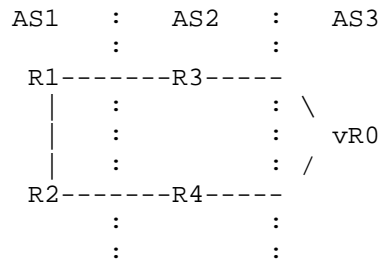


Figure 20: multi-as-aggregation

5. Originating the TED NLRI

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

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

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

6. Receiving the TED NLRI

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

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

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

7. Use Cases

7.1. MPLS TE

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

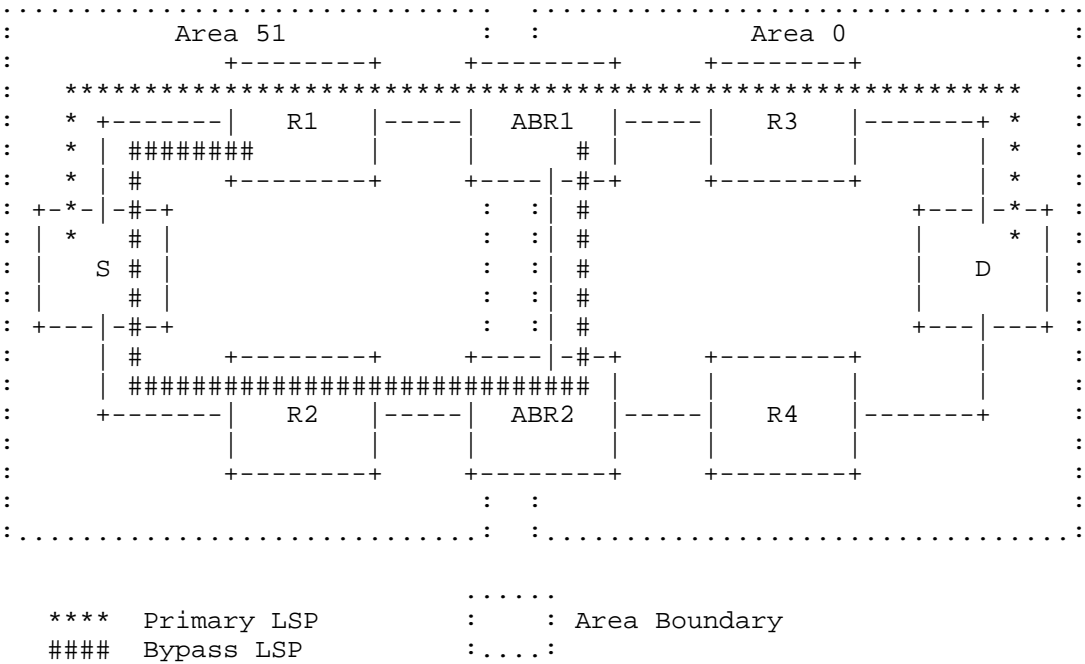


Figure 21: MPLS TE Bypass LSP problem

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

7.2. ALTO Server Network API

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

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

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

7.3. Path Computation Element (PCE) TED Synchronization Protocol

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

8. IANA Considerations

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

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

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

9. Security Considerations

This draft does not affect the BGP security model.

10. Acknowledgements

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

Systems, and Richard Woundy from Comcast for their comments.

11. References

11.1. Normative References

- [IANA-ISIS] "IS-IS TLV Codepoint, Sub-TLVs for TLV 22", <<http://www.iana.org/assignments/isis-tlv-codepoints/isis-tlv-codepoints.xml#isis-tlv-codepoints-3>>.
- [RFC1195] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, December 1990.
- [RFC1918] Rekhter, Y., Moskowitz, R., Karrenberg, D., Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, February 1996.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [RFC4202] Kompella, K. and Y. Rekhter, "Routing Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4202, October 2005.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.
- [RFC4893] Vohra, Q. and E. Chen, "BGP Support for Four-octet AS Number Space", RFC 4893, May 2007.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.
- [RFC5307] Kompella, K. and Y. Rekhter, "IS-IS Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 5307, October 2008.
- [RFC6119] Harrison, J., Berger, J., and M. Bartlett, "IPv6 Traffic Engineering in IS-IS", RFC 6119, February 2011.

11.2. Informative References

- [I-D.ietf-alto-protocol] Alimi, R., Penno, R., and Y. Yang, "ALTO Protocol", draft-ietf-alto-protocol-08 (work in progress), May 2011.
- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC4970] Lindem, A., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 4970, July 2007.
- [RFC5073] Vasseur, J. and J. Le Roux, "IGP Routing Protocol Extensions for Discovery of Traffic Engineering Node Capabilities", RFC 5073, December 2007.

Authors' Addresses

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

Email: hannes@juniper.net

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

Email: jmedved@juniper.net

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

Email: sprevidi@cisco.com

Network Working Group
Internet Draft
Intended status: Standards Track
May 25, 2011
Expires: Nov 25, 2011

J. Uttaro
AT&T
V. Van den Schrieck
P. Francois
UCLouvain
R. Fragassi
A. Simpson
Alcatel-Lucent
P. Mohapatra
Cisco Systems

Best Practices for Advertisement of Multiple Paths in IBGP
draft-ietf-idr-add-paths-guidelines-01.txt

Status of this Memo

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

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

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

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/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 Nov 25, 2011.

Copyright Notice

Copyright (c) 2010 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 BSD License.

Abstract

Add-Paths is a BGP enhancement that allows a BGP router to advertise multiple distinct paths for the same prefix/NLRI. This provides a number of potential benefits, including reduced routing churn, faster convergence and better loadsharing.

This document provides recommendations to implementers of Add-Paths so that network operators have the tools needed to address their specific applications and to manage the scalability impact of Add-Paths. A router implementing Add-Paths may learn many paths for a prefix and must decide which of these to advertise to peers. This document analyses different algorithms for making this selection and provides recommendations based on the target application.

Table of Contents

1. Introduction.....	4
2. Terminology.....	4
3. Add-Paths Applications.....	5
3.1. Fast Connectivity Restoration.....	5
3.2. Load Balancing.....	7
3.3. Churn Reduction.....	7
3.4. Suppression of MED-Related Persistent Route Oscillation...	7
4. Implementation Guidelines.....	8
4.1. Capability Negotiation.....	8
4.2. Receiving Multiple Paths.....	9
4.3. Advertising Multiple Paths.....	9
4.3.1. Path Selection Modes.....	11
4.3.1.1. Advertise All Paths.....	11
4.3.1.2. Advertise N Paths.....	12
4.3.1.3. Advertise All AS-Wide Best Paths.....	12
4.3.1.4. Advertise ALL AS-Wide Best and Next-Best Paths (Double AS Wide).....	13
4.3.2. Derived Modes from Bounding the Number of Advertised Paths.....	14
5. Deployment Considerations.....	14
5.1. Introducing Add-Paths into an Existing Network.....	14
5.2. Scalability Considerations.....	16
5.3. Routing Consistency Considerations.....	17
5.4. Consistency between Advertised Paths and Forwarding Paths	17

5.5. Routing Churn.....	18
6. Security Considerations.....	18
7. IANA Considerations.....	18
8. Conclusions.....	18
9. References.....	19
9.1. Normative References.....	19
9.2. Informative References.....	19
10. Acknowledgments.....	19
Appendix A. Other Path Selection Modes.....	20
A.1. Advertise Neighbor-AS Group Best Path.....	20
A.2. Best LocPref/Second LocPref.....	20
A.3. Advertise Paths at decisive step -1.....	21

1. Introduction

The BGP Add-Paths capability enhances current BGP implementations by allowing a BGP router to exchange with its BGP peers more than one path for the same destination/NLRI. The base BGP standard [RFC 4271] does not provide for such a capability. If a BGP router learns multiple paths for the same NLRI (from multiple peers), it selects only one as its best path and advertises the best path to its peers. The primary goal of Add-Paths is to increase the visibility of paths within an iBGP system. This has the effect of improving robustness in case of failure, reducing the number of BGP messages exchanged during such an event, and offering the potential for faster re-convergence. Through careful selection of the paths to be advertised, Add-Paths can also prevent routing oscillations.

The purpose of this document is to provide the necessary recommendations to the implementers of Add-Paths so that network operators have the tools needed to address their specific applications and to manage the scalability impact of Add-Paths while maintaining routing consistency. A router implementing Add-Paths may learn many paths for a prefix and must decide which of these to advertise to peers. This document analyses different algorithms for making this selection and provides recommendations based on the target application.

2. Terminology

In this document the following terms are used:

Add-Paths peer: refers a peer with which the local system has agreed to receive and/or send NLRI with path identifiers

Primary path: A path toward a prefix that is considered a best path by the BGP decision process [RFC 4271] and actively used for forwarding traffic to that prefix. A router may have multiple primary paths for a prefix if it implements multipath.

Diverse path: A BGP path associated with a different BGP next-hop and BGP router than some other set of paths. The BGP router associated with a path is inferred from the ORIGINATOR_ID attribute or, if there is none, the BGP Identifier of the peer that advertised the path.

Backup path: A diverse path with respect to the primary paths toward a prefix. The backup path can be used to forward traffic to the destination if the primary paths fail.

Optimal backup path: The backup path that will be selected as the new best path for a prefix when all primary paths are removed/withdrawn.

AS-Wide preferred paths: All paths that are considered as best when applying rules of the BGP decision process up to the IGP tie-break.

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. Add-Paths Applications

[draft-pmohapat] presents the applications that would benefit from multiple paths advertisement in iBGP. They are summarized in the following subsections.

3.1. Fast Connectivity Restoration

With the dissemination of backup paths, fast connectivity restoration and convergence can be achieved. If a router has a backup path, it can directly select that path as best upon failure of the primary path. This minimizes packet loss in the dataplane. Sending multiple paths in iBGP allows routers to receive backup paths when path visibility is not sufficient with classical BGP. This is especially useful when Route Reflection is used.

Consider a network such as the one depicted in Figure 1 and suppose that none of the routers support Add-Paths. AS1 receives from AS3 2 paths (A and B) to a particular destination XYZ. Suppose path A is preferred over path B due to path A having a lower MED (multi-exit discriminator).

AS1 uses a route reflector RR1 to reduce the scale of its iBGP mesh. If the routers in AS1 are not configured for best-external then RR1 knows about only path A during steady state because router B suppresses/withdraws its advertisement of path (B) to RR1. If the routers in AS1 do support best-external then RR1 may have both paths in its Adj-RIB-IN, but regardless of the best-external configuration RR1 can only advertise its best path A to its peers, including router D.

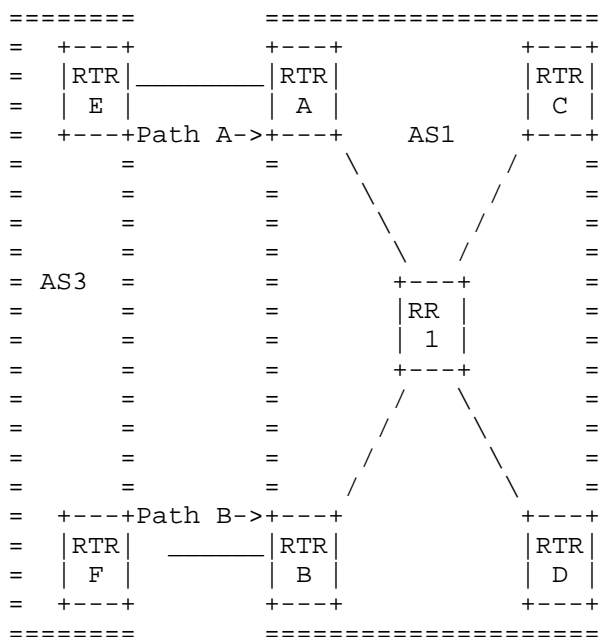


Figure 1: Example Topology

Under these circumstances consider the steps required to restore traffic from router D to destination XYZ when the link between Router A and Router E fails. (Assume that router A set next-hop to self when advertising path A and that router B is not configured for best-external).

1. Router A sends a BGP UPDATE message withdrawing its advertisement of path (A).
2. RR1 receives the withdrawal, and propagates it to its other client peers, routers B, C and D.
3. When router B receives the withdrawal of path (A) it reruns its decision process and selects path (B) as its new best path. Router B advertises path (B) to RR1.
4. RR1 reruns its decision process and selects path (B) as its new best path. RR1 advertises path (B) to client peers A, C and D.

5. Router D reruns its decisions process, determines path (B) to be the best path, and updates its forwarding table. After this step traffic from router D to destination XYZ is restored (the traffic path has changed from A to B).

With the use of Add-Paths, the convergence time for the above path failure example can be reduced considerably. The main reason for the improvement is that Add-Paths allows router D to be aware of more than one path to destination XYZ prior to the failure of the best path (A). In steady-state (with no failures) router B decides, as before, that path (A) is its best path but because of its Add-Paths (or best-external) configuration it also advertises path (B) to RR1. Using Add-Paths RR1 can advertise both learned paths to its IBGP peers, including router D. Now consider again the scenario where the link between Router A and Router E fails. In this case, with Add-Paths, fewer steps are required to achieve re-convergence:

1. Router A sends a BGP UPDATE message withdrawing its advertisement of path (A).
2. RR1 receives the withdrawal, and propagates it to its other client peers, routers B, C and D.
3. Router D receives the withdrawal, reruns the decision process and updates the forwarding entry for destination XYZ.

3.2. Load Balancing

Increased path diversity allows routers to install several paths in their forwarding tables in order to load balance traffic across those paths.

3.3. Churn Reduction

When Add-Paths is used in an AS, the availability of additional backup paths means failures can be recovered locally with much less path exploration in iBGP and therefore less updates disseminated in eBGP. When the preferred backup path is the post-convergence path, churn is minimized.

3.4. Suppression of MED-Related Persistent Route Oscillation

As described in [oscillation], Add-Paths is a valuable tool in helping to stop persistent route oscillations caused by comparison of paths based on MED in topologies where route reflectors or the confederation structure hide some paths. With the appropriate path selection algorithm Add-Paths stops these route oscillations because

the same set of paths are consistently advertised by the route reflector or the confederation border router and the routers receiving this set of paths make stable routing decisions about the best path.

4. Implementation Guidelines

This section discusses recommendations for the implementation of Add-Paths. The following topics are addressed:

- . Considerations related to Add-Paths capability negotiation
- . Receiving BGP routes from Add-Paths peers
- . Advertising BGP routes to Add-Paths peers. This section discusses various path selection algorithms, which are the procedures available to an Add-Paths speaker for deciding which set of paths to advertise to an Add-Paths peer for particular prefixes.

4.1. Capability Negotiation

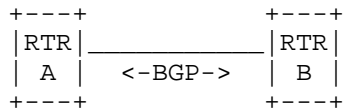


Figure 2: BGP Peering Example

In Figure 2, in order for a router A to receive multiple paths per NLRI from peer B, for a particular address family (AFI=x, SAFI=y), the BGP capabilities advertisements during session setup must indicate that peer B wants to send multiple paths for AFI=x, SAFI=y and that router A is willing to receive multiple paths for AFI=x, SAFI=y. Similarly, in order for router A to send multiple paths per NLRI to peer B, for a particular address family (AFI=x, SAFI=y), the BGP capabilities advertisements must indicate that router A wants to send multiple paths for AFI=x, SAFI=y and peer B is willing to receive multiple paths for AFI=x, SAFI=y. Refer to [Add-Paths] for details of the Add-Paths capabilities advertisement.

The capabilities of the local router **MUST** be configurable per peer and per address family, and **SHOULD** support the ability to configure send-only operation or receive-only operation. The default mode of operation shall be to both send and receive.

4.2. Receiving Multiple Paths

Currently, per standard BGP behavior, if a BGP router receives an advertisement of an NLRI and path from a specific peer and that peer subsequently advertises the same NLRI with different path information (e.g. a different NEXT_HOP and/or different path attributes) the new path effectively overwrites the existing path.

When Add-Paths has been negotiated with the peer, the newly advertised path should be stored in the RIB-IN along with all of the paths previously advertised (and not withdrawn) by the peer.

When an Add-Paths speaker has negotiated to receive multiple paths for (AFIx, SAFIy) from a peer all advertisements and withdrawals of NLRI within that address family from that peer MUST include a path identifier, as described in [Add-Paths]. The path identifiers have no significance to the receiving peer. If the combination of NLRI and path identifier in an advertisement from a peer is unique (does not match an existing route in the RIB-IN from that peer) then the route is added to the RIB-IN. If the combination of NLRI and path identifier in a received advertisement is the same as an existing route in the RIB-IN from the peer then the new route replaces the existing one. If the combination of NLRI and path identifier in a received withdrawal matches an existing route in the RIB-IN from the peer then that route shall be removed from the RIB-IN.

A BGP UPDATE message from an Add-Paths peer may advertise and withdraw more than one NLRI belonging to one or more address families. In this case Add-Paths may be supported for some of the address families and not others. In this situation the receiving BGP router should not expect that all of the path identifiers in the UPDATE message will be the same.

4.3. Advertising Multiple Paths

[Add-Paths] specifies how to encode the advertisement of multiple paths towards the same NLRI over an iBGP session, but provides no details about which set of multiple paths should be advertised. In this section, four path selection algorithms are described and compared with each other. These 4 algorithms are considered to be the most useful across the widest range of deployment scenarios. The list of possible path selection algorithms is much larger and for the interested reader Appendix A provides information about other path selection modes that were considered in historical versions of this document.

In comparing any two path selection algorithms the following factors should be taken into account:

Control Plane Load: When a router receives multiples paths for a prefix from an iBGP client it has to store more paths in its Adj-Rib-Ins.

Control Plane Stress: Coping with multiple iBGP paths has two implications on the computation that a router has to handle. First, it has to compute the paths to send to its peers, i.e. more than the best path. Second, it also has to handle the potential churn related to the exchange of those multiple paths.

MED/IGP oscillations: BGP sometimes suffers from routing oscillations when the physical topology differs from the logical topology, or when the MED attribute is used. This is due to the limited path visibility when a single path is advertised and Route Reflection is used. Increasing the path visibility by advertising multiple paths can help solve this issue.

Path optimality: When a single path is advertised, border routers do not always receive the optimal path. As an example, Route Reflectors typically send a single path chosen based on their own IGP tie-break (although modifications to this are proposed in [BGP-ORR]). Increasing path visibility would also help routers to learn the path that is best suited for them w.r.t. the IGP tie-break.

Backup path optimality: Multiple paths advertisement gives routers the opportunity to have a backup path. However, some backup paths are better than others. Indeed, when a link failure occurs, if a router already knows its post-convergence path, the BGP re-convergence is straightforward and traffic is less impacted by the transient use of non-best forwarding paths.

Convergence time: Advertising multiple paths in iBGP has an impact on the convergence time of the BGP system. More paths need to be exchanged, but on the other hand, the routing information is propagated faster. With an increased path visibility, there is less path exploration during the convergence. Also, with the availability of backup paths, convergence time in case of failure is also reduced.

Target application: Depending on the application type, the number of paths to advertise for a prefix will vary. For example, for fast connectivity restoration, it may be sufficient to advertise only 2 paths to a peer so that it will have the best path and the optimal backup path. For load balancing purposes, it may be desirable to advertise more paths, but inclusion of the optimal backup path in the

set may be less critical. For route oscillation elimination, it is required to advertise all group-best paths for a prefix.

4.3.1. Path Selection Modes

The following subsections describe the 4 main path selection modes considered in this draft. Each mode is considered either MANDATORY or OPTIONAL. A MANDATORY mode MUST be supported by any implementation that claims compliance with this document. An OPTIONAL mode may be supported by some but not all implementations.

The path selection mode and any parameters applicable to the mode MUST be configurable per AFI/SAFI and per peer and SHOULD be configurable per prefix. To illustrate the value of this flexibility, consider a prefix P that belongs to an address family F requiring path IDs to be included with every NLRI (e.g. due to the Add-Paths capability negotiation with the peer). If P is one of a number of prefixes that would not benefit from the advertisement of multiple paths then it is perfectly valid to send only the best path.

4.3.1.1. Advertise All Paths

A simple rule for advertising multiple paths in iBGP is to simply advertise to iBGP peers all received paths, provided they pass export filters. This solution is easy to implement, but the counterpart is that all those paths need to be stored by all routers that receive them, which can be quite expensive. If a path to a prefix P is advertised to N border routers, with a Full Mesh of iBGP sessions, all routers have N paths in their Adj-RIB-Ins. If Route Reflection is used and each client is connected to 2 Route Reflectors, it may learn up to 2*N paths.

This solution gives a perfect path visibility to all routers, thus limiting churn and losses of connectivity in case of failure. Indeed, this allows routers to select their optimal primary path, and to switch on their optimal backup path in case of failure.

However, as more paths are exchanged, the number of BGP messages disseminated during the initial iBGP convergence can be high, and convergence may be slower.

Routing oscillations are prevented with this rule, because a router won't need to withdraw a previously advertised path when its best path changes.

This path selection mode is OPTIONAL.

4.3.1.2. Advertise N Paths

Another solution is for a router to advertise a maximum of N paths to iBGP peers. Here, the computational cost is the selection of the N paths. Indeed, there must be a ranking of the paths in order to advertise the most interesting ones. A way for a router to select N paths is to run N times its decision process. At each iteration of the process only those paths not selected during a previous iteration and those with a different NEXT_HOP and BGP Identifier (or Originator ID) combination from previously-selected paths are eligible for consideration. The memory cost is bounded: a router receives a maximum of N paths for each prefix from each peer. With N equal to 2, all routers know at least two paths and can provide local recovery in case of failure. If multipath routing is to be deployed in the AS, N can be increased to provide more alternate paths to the routers.

Path optimality and backup path optimality are not guaranteed, i.e. it is possible that the optimal path of a router (w.r.t. IGP tie-break) is not contained in the set of paths advertised by its Route Reflector. However, as the number of paths that it receives is higher than without Add-Paths, it is possible that the chosen nexthop is closer to the router in terms of IGP cost than the nexthop that would have been chosen without Add-Paths.

This solution helps to reduce routing oscillations, but not in all cases. Indeed, path visibility is still constrained by the maximum number of paths, and configurations with routing oscillations still exist.

This path selection mode is MANDATORY. The default value of N MUST be 2. The value of N MUST be configurable and MAY be upper bounded by an implementation.

The default value of 2 ensures the availability of a backup path (if 2 or more paths have been received) while maintaining minimum impact to memory and churn. If Add-N with N equal to 2 is insufficient to meet another objective (e.g. loadsharing or MED/IGP oscillation) there is always a large enough value of N that can be selected, if N is configurable, to meet that objective.

4.3.1.3. Advertise All AS-Wide Best Paths

Another choice is to advertise all paths with the same AS-wide preference [Basu-ibgp-osc], i.e. the paths that all routers would select based on the rules of the decision process that are not router-dependent (i.e. Local-preference, ASPath length and MED

rules). Thus, for a given router, those paths only differ by the IGP cost to the nexthop or by the tie-breaking rules.

The computational cost is reduced, as a router only has to send the paths remaining before applying the IGP tie-breaking rule. However, it is difficult to predict how many paths will be stored, as it depends on the number of eBGP sessions on which this prefix is advertised with the best AS-wide preference.

With this rule, the routing system is optimal: all routers can choose their best path (or best paths if multipath is used) based on their router-specific preferences, i.e. the IGP cost to the nexthop. Hot potato routing is respected. Also, MED oscillations are prevented, because the path visibility among the AS-wide preferred paths is total.

The existence of a backup path is not guaranteed. If only one path with the AS-wide best attributes exists, there is no backup path disseminated. However, if such a path exists, it is optimal as it has the same AS-wide preference as the primary

This path selection mode is OPTIONAL.

4.3.1.4. Advertise ALL AS-Wide Best and Next-Best Paths (Double AS Wide)

This variant of "Advertise All AS Wide Best Paths" trades-off the number of paths being propagated within the iBGP system for post-convergence alternate paths availability and routing stability. A BGP speaker running this mode will select for advertisement its AS Wide Best paths, plus all the AS Wide Best paths obtained when removing the first ones from consideration.

Under this mode, a BGP speaker knows multiple AS-Wide best paths or the AS-Wide best path and all the second AS-Wide best paths, so that routing optimality and backup path availability are ensured. Note that the post-convergence paths will be known by each BGP node in an AS supporting this mode.

The computation complexity of this mode is relatively low as it requires to run the usual BGP Decision Process up to and including the MED rule. The set of paths remaining after that step form the AS-Wide best paths. Next, a best path selection algorithm is run up to and including the MED rule, based on the paths that are not in the set of AS-Wide best paths.

The number of paths for a prefix p , known by a given router of the AS, is the number of AS-Wide best and second AS-Wide best paths found at the Borders of the AS.

MED Oscillations are avoided by this mode, both for the primary and alternate paths being picked under this mode.

This path selection mode is OPTIONAL.

4.3.2. Derived Modes from Bounding the Number of Advertised Paths

For some of the modes discussed in section 4.3.1 the number of paths selected by the algorithm (M) is not predictable in advance, and depends on factors such as network topology. For such modes, implementations MAY support the ability to limit the number of advertised paths to some value N that is less than M .

It must be noted that the resulting derivative mode may no longer meet the properties stated in section 4.3.1 (which assumes $N=M$). This is particularly true for the MED oscillation avoidance property. The use of such bounds thus needs to be considered carefully in deployments where MED oscillation avoidance is a key goal of deploying Add-path. If fast recovery is the main objective then it is reasonable and sufficient to set N to 2. If the main goal is improved load-balancing then limiting N to number of ECMP paths supported by the forwarding planes of the receiving routers is also a reasonable practice.

5. Deployment Considerations

This section proposes a potential strategy for introducing Add-Paths into an existing network and discusses considerations related to scalability, routing consistency and routing churn.

5.1. Introducing Add-Paths into an Existing Network

There are many possible ways that Add-Paths can be introduced into an existing deployed network. It is not a practical goal for this document to list all of these options and discuss the pros and cons for each one. It is however valuable to consider an example migration strategy that may be relatively common among layer 3 service providers that currently use route reflectors for scaling. This example migration strategy is attractive for several reasons:

1. It involves incremental steps that allow the impact of Add-Paths to be carefully evaluated before proceeding to the next step.

2. It recognizes the fact that many routers will require at least a software upgrade to support Add-Paths, and it will not be practical to upgrade all of these routers all at once.
3. It reduces convergence time (in stages) with a relatively moderate increase in router memory and CPU demands.

The example migration strategy assumes a starting point of a deployed network with one or more RR clusters. None of the routers in the network support Add-Paths without an upgrade, but some do support best-external. Two of the clusters in this network are shown in Figure 3. In cluster 2, PE1, PE2, RRY and RRz are configured for best-external. This makes RRY and RRz aware of all external paths received by PEs in cluster 2 and ensures that RRY and RRz can advertise a path to the RRs in cluster 1 if it happens that the best overall route is learned from cluster 1. It doesn't however allow other clusters to be aware of more than one path per prefix learned by cluster 2.

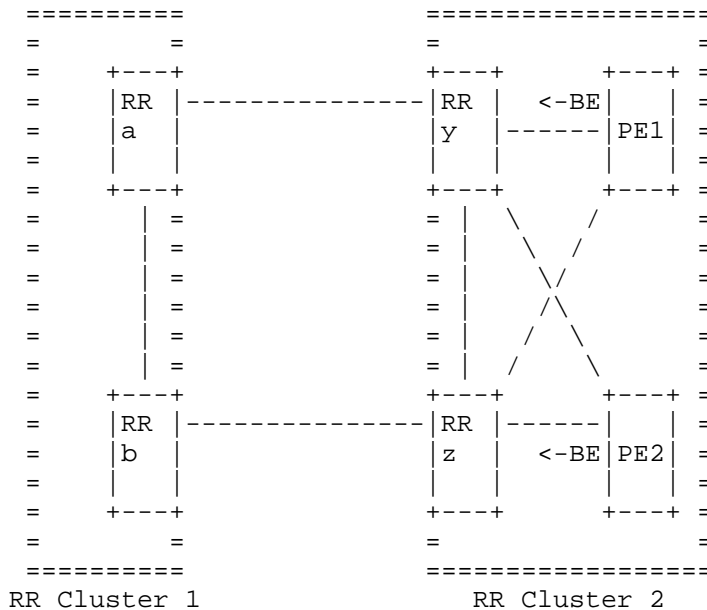


Figure 3: RR Cluster Before Add-Paths

The following sequence of steps occurs in the example migration strategy:

1. The route reflectors are upgraded in each cluster, one by one, to support Add-Paths. This allows the intra- and (eventually) inter-cluster RR-to-RR sessions to start using Add-Paths. All RRs are configured to use the Add-N, N=2 path selection algorithm. The effect of this step is to slightly reduce convergence time when the best and second-best paths for a prefix are learned by a single cluster (such as cluster 2 in Figure 3).
2. The clients are upgraded in each cluster, one by one, to support Add-Paths. On the RRs Add-Paths is configured to use the Add-N, N=2 path selection algorithm towards upgraded client peers. At this step clients are configured in the receive-only Add-Paths mode. This means that best-external continues to operate as before in the client-to-RR direction. The effect of this step is to ensure that all clients have two paths per prefix for ECMP or fast failover, assuming at least 2 paths are available.
3. The clients are re-configured to use Add-Paths in the transmit direction towards their RR peers. This causes Add-Paths to replace the best-external behavior. The effect of this step is to free up CPU and memory resources related to the storage of paths that are third best or worse. If a cluster such as the one in Figure 3 had 50 clients, and 10 of these learned an external route for the same prefix, then the RRs in that cluster would need to store up to 12 paths for that prefix. This would be true even if the 2 best overall paths came from another cluster. Contrast this with the use of Add-Paths in the client-to-RR direction. For the same case the route reflectors need only store the 2 paths learned from non-client peers.

5.2. Scalability Considerations

In terms of scalability, we note that advertising multiple paths per prefix requires more memory and state than the current behavior of advertising the best path only. A BGP speaker that does not implement Add-Paths maintains send state information in its prefix data structure per neighbor as a way to determine that the prefix has been advertised to the neighbor. With Add-Paths, this information has to be replicated on a per path basis that needs to be advertised. Mathematically, if "send state" size per prefix is 's' bytes, number of neighbors is 'n', and number of paths being advertised is 'p', then the current memory requirement for BGP "send state" = $n * s$ bytes; with Add-Paths, it becomes $n * s * p$ bytes. In practice, this value may be reduced with implementation optimizations similar to

attribute sharing. Receiving multiple paths per prefix also requires more memory and state since each path is a separate entry in the Adj-RIB-In.

5.3. Routing Consistency Considerations

As discussed in previous sections Add-Paths can help routers select more optimal paths and it can help deal with certain route oscillation conditions arising from incomplete knowledge of the available paths. But depending on the path selection algorithm and how it is used Add-Paths is not immune to its own cases of routing inconsistencies. If the BGP routers within an AS do not make consistent routing decisions about how to reach a particular destination, route oscillations may occur and these route oscillations may result in traffic loss.

Optimizing an Add-Paths deployment for scalability may run counter to routing consistency goals, and in these circumstances operators have to decide the correct tradeoff for their particular deployment. For example the Advertise All Paths mode, if applied to many prefixes, is far from ideal from a scalability perspective but it does guarantee routing consistency and correctness. A path selection mode that allows better control over scalability is the Advertise N paths mode, but this is susceptible to routing inconsistency. First, if the N paths do not include the best path from each neighbor AS group then route oscillation cannot be precluded. Second, if the advertising router (e.g. an RR) advertises N paths to peer_n and M paths to peer_m, and $N < M$, care must be exercised to ensure that all paths advertised to peer_n are included in the paths advertised to peer_m. This can be assured as long as the advertising router has strictly ordered all of its paths.

5.4. Consistency between Advertised Paths and Forwarding Paths

When using Add-Paths, routers may advertise paths that they have not selected as best, and that they are thus not using for traffic forwarding. This is generally not an issue if encapsulation is used in the AS as described in [RFC4364] and all forwarding decisions, including by the tunnel egress router, are based on label information - i.e. if only the ingress router performs an IP FIB lookup. In this situation the dataplane path followed by the packets is the one intended by the ingress router, and corresponds to the control plane path it selected.

On the other hand, if Add-Paths is used in a network without encapsulation, some scenarios can result in forwarding deflection or loops. Such forwarding anomalies already occur without Add-Paths,

when the routers on the forwarding path do not have a synchronized view of the best path. They will deflect the traffic to their own local view of the best path, and, when multiple deflections occur, forwarding loops can occur. With Add-Paths, the issue can be exacerbated due to routers advertising non-best paths. As discussed above, encapsulation can help with this issue, but only to the extent that it allows downstream routers to forward without an IP FIB lookup.

A first example of such issue is when the Local-Pref of non-primary paths received over iBGP sessions is modified. The ingress router may thus select as best a path non-preferred by the egress, and the egress router will thus deflect the traffic.

Another example is when the best path is selected based on tie-breaking rule. When the ingress and the egress base their path selection on the router-id of the neighbor that advertised the path to them, the result may be different for each of them. This specific issue is described and solved in [draft-pmohapat].

5.5. Routing Churn

As noted in section 3.3 using Add-Paths between iBGP peers can help to reduce routing churn with eBGP peers. This benefit does however come at the cost of potentially increased churn between the iBGP Add-Paths peers. In a non Add-Paths deployment a change in the preference order of non-best paths requires no updates to be sent to peers. But when a router has Add-Paths peers changes in non-best path preference may no longer be invisible and increased route churn may be observable. Choosing the right path selection mode and parameters - for example not setting N unnecessarily large in the Add-N mode, is important to minimizing this additional churn.

6. Security Considerations

TBD

7. IANA Considerations

TBD

8. Conclusions

TBD

9. References

9.1. Normative References

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

9.2. Informative References

- [Add-Paths] Walton, D., Retana, A., Chen E., Scudder J., "Advertisement of Multiple Paths in BGP", February 6, 2010.
- [draft-pmohapat] Mohapatra, P., Fernando, R., Filsfils, C., and R. Raszuk, "Fast Connectivity Restoration Using BGP Add-path", draft-pmohapat-idr-fast-conn-restore-00.txt (work in progress), September 2008.
- [oscillation] Walton, D., Retana, A., Chen, E., Scudder, J., "BGP Persistent Route Oscillation Solutions", draft-walton-bgp-route-oscillation-stop-03.txt, May 10, 2010.
- [Basu-ibgp-osc] Basu, A., Ong, C., Rasala, A., Sheperd, B., and G. Wilfong, "Route oscillations in iBGP with Route Reflection", Sigcomm 2002.
- [BGP-ORR] Raszuk, R., Cassar, C., Aman, E., Decraene, B., "BGP Optimal Route Reflection", draft-raszuk-bgp-optimal-route-reflection-01, March 11, 2011.
- [RFC4271] Rekhter, Y., Li, T., Hares, S., "A Border Gateway Protocol 4 (BGP-4)", January 2006.

10. Acknowledgments

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

Appendix A. Other Path Selection Modes

A.1. Advertise Neighbor-AS Group Best Path

[walton-osc] proposes that a router groups its paths based on the neighbor AS from which it was learned, and to advertise the best path in each of those groups.

The control plane stress induced by this solution is the computation of the per-neighbor path group, and the application of the decision process to each of them. The Control-Plane load is bounded by the number of neighboring ASes advertising a prefix, which cannot be known a-priori.

Path optimality and backup path optimality are not guaranteed, as the paths advertised are not all the AS-wide preferred paths. Backup path availability is not guaranteed. Indeed, if only one AS advertises this prefix, even on multiple eBGP sessions, only one of the paths may be selected and advertised.

A.2. Best LocPref/Second LocPref

This selection method consists in grouping the paths by Local Preference. A router sends to its peers all paths with the highest Local Preference. If there is only a single path with the highest Local Preference, it also sends all paths with the second best Local Preference.

This method ensures that all routers know all paths with the best local preference. As local preference are often related to the type of peering of the peer the path comes from, this ensures that in case of failure, routers have a backup path of equivalent quality. This prevents for example that a router switches temporarily on a peer path while an alternate path from a customer is available but hidden at the border of the AS. Such a situation could result in a temporary withdrawal of the prefix on some eBGP sessions when the router selects the path via the peer.

The advertisement of the Second Local Preference occurs when there is no alternate path with the same quality as the best path. This way, fast convergence is still ensured. Backup path is optimal, as it has the second AS-Wide preference, which becomes the AS-wide best preference upon failure of the primary one.

Sending all the paths with a given Local Preference also has a positive impact on routing optimality. Indeed, this allows border

routers to have an increased path visibility and to choose their best path based on their own criteria.

The computational cost of this solution is reduced when there are several paths with the best local preference. In this case, it is sufficient to stop the decision process after the first rule to have the set of paths to be advertised. When it is necessary to advertise the paths with second local-preference, the additional cost is to apply a second time the first rule of the decision process, which is still reasonable. The memory cost depends on the number of paths with the best local preference.

A.3. Advertise Paths at decisive step -1

When the goal is to provide fast recovery by advertising candidate post-reconvergence paths, one can choose to stop the decision process just before the step where only one path remains. If the decision process comes to IGP tie-break, all remaining paths are advertised. This way, routers advertise as many paths as possible with a quality as similar as possible.

This path selection is an intermediary solution between the two preceding ones. Here, instead of stopping the decision process at the local preference step or the IGP step, we stop it before the rule that removes the best potential backup paths. This way, we minimize the number of paths to advertise while guaranteeing the presence of a backup path. Primary and backup path optimality is ensured, as all paths with the same AS-wide preference as the best paths are included in the set of paths advertised.

Authors' Addresses

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

Virginie Van den Schrieck
UCLouvain
Place Ste Barbe, 2
Louvain-la-Neuve 1348 BE
Email: virginie.vandenschrieck@uclouvain.be
URI: <http://inl.info.ucl.ac.be/vvandens>

Pierre Francois
UCLouvain
Place Ste Barbe, 2
Louvain-la-Neuve 1348 BE
Email: pierre.francois@uclouvain.be
URI: <http://inl.info.ucl.ac.be/pfr>

Roberto Fragassi
Alcatel-Lucent
600 Mountain Avenue
Murray Hill, New Jersey
Email: roberto.fragassi@alcatel-lucent.com

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

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

Network Working Group
Internet Draft
Intended Status: Standards Track
Expiration Date: Dec 30, 2011

K. Patel
E. Chen
R. Fernando
Cisco Systems
J. Scudder
Juniper Networks
June 29, 2011

Accelerated Routing Convergence for BGP Graceful Restart
draft-keyur-idr-enhanced-gr-00.txt

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on December 30, 2011.

Copyright Notice

Copyright (c) 2011 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

In this document we specify extensions to BGP graceful restart in order to avoid unnecessary transmission of the routing information preserved across a session restart, thus accelerating the routing convergence.

1. Introduction

Currently the BGP graceful restart (GR) mechanism specified in [RFC4724] requires a complete re-advertisement of the routing information across a session restart, even though partial or complete routing information is usually preserved. For example, as described in [RFC4724], the "Receiving Speaker" temporarily maintains the routes received from its neighbor with the GR Capability. In addition, the "Restarting Speaker" may also be able to preserve partial or full routing information across a BGP restart by checkpointing routing information to a standby or secondary facility.

Clearly the routing re-convergence post a session restart would be faster if we can avoid unnecessary transmission of the routing information preserved across a session restart. That is the goal of this document.

In this document we specify extensions to BGP graceful restart in order to avoid unnecessary transmission of the routing information preserved across a session restart, thus accelerating the routing convergence. More specifically, we describe a "version number" based mechanism for keeping track of the routing information across a session restart. A new BGP message type, UPDATE-VERSION, is introduced for checkpointing the update version maintained for a neighbor. We also introduce the Enhanced Graceful Restart Capability, and specify procedures for handling routing update across a session restart.

1.1. Specification of Requirements

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

2. Version Numbers for Routing Entities

In order to avoid unnecessary transmission of the routing information preserved across a session restart, a BGP speaker will need to identify exactly "what" has been preserved by a remote speaker.

The approach described here is "version number" (or "sequence number") based, and it consists of (a) assigning a unique, monotonically increasing number as the version number for each routing entity (e.g., route or message) when it is created or modified; and (b) maintaining an update version (for each neighbor) calculated as the maximum of the version numbers of all the routing entities that have been sent to the neighbor.

A BGP speaker can tell whether a given routing entity has been sent to a neighbor by comparing the version number of the entity with the update version for the neighbor. Thus by checkpointing the update version for a neighbor across a session restart, a BGP speaker would be able to identify exactly "what" has been preserved by a remote speaker, and also "what" remains to be sent.

In this document a version number is a 8-octet unsigned integer. Value 0 is used to indicate the beginning (or "epoch") of the update generation. The version number is not expected to wrap. However, in the unlikely scenario that it does wrap, the sender MUST maintain its internal consistency, and also MUST perform a route refresh [RFC2918, EH-RR] toward the receiver.

The number space for the version numbers should be AFI/SAFI [RFC4760] specific. Version numbers are also assigned (from the same number space) to other AFI/SAFI specific, non-update information (such as ROUTE-REFRESH [RFC2918]), and are included in the calculation of the update version for a neighbor.

3. UPDATE-VERSION Message

The UPDATE-VERSION message is a new BGP message type with type code <TBD>. In addition to the fixed-size BGP header [RFC4271], the UPDATE-VERSION message contains the following fields:

```
+-----+
| Address Family Identifier (2 octets) |
+-----+
| Subsequent Address Family Identifier (1 octet) |
+-----+
| Message Subtype (1 octet) |
+-----+
```

```

+-----+
| Version (8 octets) |
+-----+

```

The "Address Family Identifier" (AFI) field and the "Subsequent Address Family Identifier" (SAFI) field are the same as the ones used in [RFC4760].

The "Message Subtype" field indicates whether the sender is (a) sending an update version (value 1), (b) acknowledging the receipt of an update version (value 2), or (c) requesting updates from the very last update version the sender has acknowledged (value 3).

The Version field contains an update version associated with the message subtypes 1 and 2. The value of this field is irrelevant for the message subtype 3. This value of the field is opaque to the receiver.

As detailed in the Operation section, the UPDATE-VERSION message can be used by a BGP speaker to either carry an update version, or acknowledge the receipt of an update version, or request updates from the very last update version acknowledged.

4. Enhanced Graceful Restart Capability

The Enhanced Graceful Restart (GR) Capability is a new BGP capability [RFC5492]. The Capability Code for this capability is specified in the IANA Considerations section of this document. The Capability Length field of this capability is 0.

By advertising the Enhanced GR Capability to a peer, a BGP speaker conveys to the peer that the speaker is capable of receiving and properly handling the UPDATE-VERSION message from the peer, as well as recognizing the two new bit flags defined below for the GR Capability.

The two new bit flags for the "Flags for Address Family" field of the GR Capability are defined as follows:

```

  0 1 2 3 4 5 6 7
+---+---+---+---+
| | |R|T| | |
+---+---+---+---+

```


The third most significant bit (R) is defined as the "RX Routing State", which is used to indicate whether during the previous session restart the routes of the given AFI/SAFI that were received have indeed been preserved up to the update version acknowledged by the speaker previously. When set (value 1), the bit indicates that the routes have been preserved.

The fourth most significant bit (T) is defined as the "TX Routing State", which is used to indicate whether the speaker has indeed preserved enough state to resume advertising routes of the given AFI/SAFI from the update version acknowledged by the neighbor previously. When set (value 1), the bit indicates that the state has been preserved.

5. Operation

In order for a BGP speaker to be able to resume sending routing information for an AFI/SAFI from the last update version that was previously acknowledged by a peer, the speaker **MUST** maintain enough state for all the routing information that has been sent until their acknowledgment is received by the speaker. The routing information includes reachable / unreachable information as well as other AFI/SAFI specific, non-update information. Furthermore, the route advertisement state needs to be maintained properly in order to minimize spurious route withdraws across a session restart.

An implementation **SHOULD** impose an upper bound on how much state it would maintain in the case that a receiver ("slow peer") is not able to generate an acknowledgment in a timely manner. The upper bound might be based on a number of factors such as the number of pending unacknowledged withdraws or more generally, the volume of unacknowledged state, and a timer. Once the acknowledgment from a peer is not received within the specified upper bound, and the maintained state is compromised, then the speaker **MUST** clear the "TX Routing State" in the GR Capability to be advertised to the peer in the next session restart.

A BGP speaker **MAY** advertise the Enhanced GR Capability to its peer if the speaker is capable of receiving and properly handling the UPDATE-VERSION message from the peer, and also recognizing the two new bit flags in the GR Capability. If the GR Capability is to be sent by the speaker, the "RX Routing State" for an AFI/SAFI in the GR Capability **SHOULD** be set if the speaker has preserved the routing information from the peer up to the update version that the speaker acknowledged previously. In addition, the "TX Routing State" for an AFI/SAFI in the GR Capability **SHOULD** be set if the speaker has preserved enough routing state to resume sending messages from the

update version acknowledged by the peer previously.

When both the GR Capability and the Enhanced GR Capability are to be included in an OPEN message, it is RECOMMENDED (though not required) that the Enhanced GR Capability be placed ahead of the GR Capability.

In processing the GR Capability in an OPEN message from a peer, a BGP speaker MUST NOT examine the two new bit flags defined in this document for the GR Capability unless the Enhanced GR Capability is also present in the OPEN message.

A BGP speaker MAY send an UPDATE-VERSION message to a peer only if the Enhanced GR Capability is received from the peer.

Once a BGP speaker receives the Enhanced GR Capability from its peer, the speaker SHOULD send an UPDATE-VERSION message carrying the update version after sending significant amount of routing information (including non-UPDATE messages) for an AFI/SAFI. This SHALL continue as long as routing information is being sent. To reduce the overhead by excessive number of UPDATE-VERSION messages, we highly recommend the "batching" approach, that is, use one UPDATE-VERSION message to cover a number of routing updates, and/or a meaningful duration of time.

When a BGP speaker receives an UPDATE-VERSION message carrying an update version, if the AFI/SAFI carried by the message does not match any AFI/SAFI that the speaker is willing to receive from the peer, the UPDATE-VERSION message SHALL be ignored. Otherwise, the speaker MUST send an UPDATE-VERSION message back promptly acknowledging the receipt of the update version. The UPDATE-VERSION messages carrying the acknowledgments MUST be sent in the same order as the received UPDATE-VERSION messages carrying the update versions.

When a BGP speakers receives an UPDATE-VERSION message acknowledging an update version, the speaker MUST record this latest update version being acknowledged for future use.

Consider the case that both the GR Capability and the Enhanced GR Capability are exchanged between Speaker A and Speaker B, and for an AFI/SAFI the "TX Routing State" is set in the GR advertised by A, and the "RX Routing State" is also set in the GR received from B. Then Speaker A SHALL send routing information from the last update version that was previously acknowledged by Speaker B. Note that it may be advantageous for Speaker B to send an UPDATE-VERSION message acknowledging the most recent update version immediately after the session is established. Also, Speaker B MUST not follow the procedures described in [RFC4724] for purging stale routes. If the conditions specified in this paragraph are not satisfied, then the

procedures described in [RFC4724] remain unchanged.

During the lifetime of an established session, if needed, a BGP speaker MAY use the UPDATE-VERSION message to request updates from the last update version that was previously acknowledged as long as the speaker has received the Enhanced GR Capability from its peer.

When a BGP speaker receives such a request, it SHALL try to send routing information from the last acknowledged update version that the speaker has recorded. If the speaker is unable to do so for some reason (e.g., "slow peer"), then it SHOULD perform a route refresh using mechanism defined in [EH-RR] if possible. Otherwise, the BGP speaker SHOULD reset the session.

6. Error Handling

This document defines a new NOTIFICATION error code:

Error Code	Symbolic Name
TBD	UPDATE-VERSION Message Error

The following error subcodes are defined as well:

Subcode	Symbolic Name
1	Invalid Message Length
2	Invalid Message Subtype

If a BGP speaker detects an error while processing an UPDATE-VERSION message, it MUST send a NOTIFICATION message with Error Code UPDATE-VERSION Message Error. The Data field of the NOTIFICATION message MUST contain the complete UPDATE-VERSION message.

If the Length field for the UPDATE-VERSION message is incorrect, then the error subcode is set to "Invalid Message Length".

If the Message Subtype in the UPDATE-VERSION message is not any of the defined value, then the error subcode is set to "Invalid Message Subtype".

7. IANA Considerations

This document introduces the Enhanced Graceful Restart Capability. The capability code needs to be assigned by IANA per [RFC5492].

This document introduces a new BGP message type, UPDATE-VERSION. The type code needs to be assigned by IANA.

In addition, this document defines an NOTIFICATION error code and several error subcodes for the UPDATE-VERSION message. They need to be registered with the IANA.

8. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing BGP [RFC4271, RFC4724].

9. Acknowledgments

TBD.

10. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2918] Chen, E., "Route Refresh Capability for BGP-4", RFC 2918, September 2000.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC4724] Sangli, S., E. Chen, R. Rernando, J. Scudder, and Y. Rekhter, "Graceful Restart Mechanism for BGP", January 2007.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.
- [EH-RR] Patel, K., E. Chen and B. Venkatachalapathy, "Enhanced

Route Refresh Capability for BGP-4", work in progress.

11. Authors' Addresses

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

Email: keyupate@cisco.com

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

EMail: enkechen@cisco.com

Rex Fernando
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: rex@cisco.com

John Scudder
Juniper Networks

Email: jgs@juniper.net

IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 12, 2012

R. Raszuk, Ed.
Cisco Systems
J. Haas, Ed.
Juniper Networks
S. Amante
Level 3 Communications, LLC
R. Steenbergen
nLayer Communications, Inc.
B. Decraene
France Telecom
P. Jakma
Uni. of Glasgow
July 11, 2011

Wide BGP Communities Attribute
draft-raszuk-wide-bgp-communities-02

Abstract

Route tagging plays an important role in external BGP relations, in communicating various routing policies between peers. It is also a very common best practice among operators to propagate various additional information about routes intra-domain. The most common tool used today to attach various information about routes is through the use of BGP communities.

Such information is important to allow BGP speakers to perform some mutually agreed upon actions without the need to maintain a separate offline database for each tuple of prefix and associated set of action entries.

This document defines a new encoding which will enhance and simplify what can be accomplished today with the use of BGP communities. The most important addition this specification makes over currently defined BGP communities is the ability to specify, carry as well as use for execution an operator's defined set of parameters. It also provides an extensible platform for any new community encoding needs in the future.

Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-

Drafts is at <http://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on January 12, 2012.

Copyright Notice

Copyright (c) 2011 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. Wide BGP Community Attribute	4
2.1. Wide BGP Community Attribute Container Header	5
2.2. Wide Community Atoms	6
3. Container Type 1: Wide Community	7
3.1. Community Value	8
3.2. Source AS Number	8
3.3. Target AS Number	8
3.4. Wide Community Target(s) TLV	8
3.5. Wide Community Parameter(s) TLV	9
3.6. Usage	10
4. Well Known Standard BGP Communities	10
5. Operational considerations	10
6. Example	10
6.1. Example Wide Community Definition	11
6.2. Example Wide Community Encoding	11
7. Security considerations	13
8. IANA Considerations	13
9. Change History	14
10. Contributors	14
11. Acknowledgments	15
12. References	15
12.1. Normative References	15
12.2. Informative References	16
Authors' Addresses	16

1. Introduction

RFC 1997 [RFC1997] defines the BGP Community Attribute. This attribute is used as a tool to carry additional information in BGP routes which may help to automate peering administration. The BGP Communities Attribute consists of one or more sets of four octet values, where each specifies a different community. Except for two reserved ranges, the encoding of community values mandates that the first two octets are to contain the Autonomous System number, with the next two octets containing some locally defined value.

With the introduction of 4-octet Autonomous System numbers by RFC 4893 [RFC4893] it became obvious that BGP Communities as specified in RFC 1997 will not be able to accommodate new AS encoding. In fact RFC 4893 explicitly recommends use of four octets AS specific extended communities as a way to encode new 4 octet AS numbers.

While the encoding of 4 octet AS numbers is being addressed by [draft-ietf-idr-as4octet-extcomm-generic-subtype], neither the base BGP communities (standard or extended) nor as4octet-extcomm-generic document define a sufficient level of encoding freedom which could be of practical use. The authors believe that defining a new BGP Path Attribute, with the ability to contain locally defined parameters will enhance the current level of network policies, as well as simplify BGP policy management. The proposed simple encoding will also facilitate the delivery of new network services without a need to define a new BGP extension each time.

When defining any new type of tool there is always a unique opportunity to specify a subset of well recognized behaviors. Lists of the current most commonly used BGP communities, as well as provision for a new registry for future definitions will be contained in a separate document.

2. Wide BGP Community Attribute

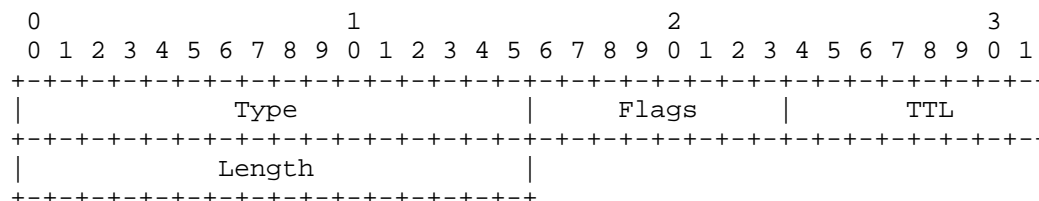
This document defines a new BGP Path Attribute, the Wide BGP Community. The attribute type code is (TBA by IANA).

The Wide BGP Community Attribute is an optional, transitive BGP attribute, and may be present only once in the update message.

The attribute contains a number of typed containers. Any given container type may appear multiple times, unless that container type's definition specifies otherwise.

2.1. Wide BGP Community Attribute Container Header

Containers always start with the following header:



Bit	Value	Meaning
0	0	Local community value.
	1	Registered community value.
1	0	Do not decrement TTL field across confederation boundaries.
	1	Decrement TTL field across confederation boundaries.
2..7	-	SHOULD be zero.

Flags are defined globally, to apply to all wide community container types.

Table 1: Flags

Bit 0 set (value 1) indicates that the given container carries a Wide BGP Community which is registered with IANA. When not set (value 0) it indicates that community value which follows is locally assigned with a local meaning. Ignored bits SHOULD be preserved in any received containers, or set to 0 otherwise.

Bit 1 is used to manage the propagation scope of a given Wide BGP Community across confederation boundaries. When not set (value of 0), the TTL field is not considered at the sub-AS boundaries. When set (value of 1), sub-AS border router follows the same procedure regarding the handling of the TTL field as applicable to ASBR at the domain boundary.

The TTL field represents the forwarding radius, in units of AS hops, for the given Wide BGP Community. A TTL value of zero indicates that this wide community must not cross any further AS boundaries. At each AS boundary, when propagating a given wide community over an EBGp session, the TTL field MUST be decremented by the sending EBGp

speaker.

The exact same decrement procedures described above apply also to sub-confederation boundaries when the global C flag is set to 1.

The special value of 0xFF indicates that the enclosed community may always be propagated over an EBGp boundary. A TTL value of 0xFF MUST NOT be decremented during propagation.

The length represents the total lengths of a given container in octets.

2.2. Wide Community Atoms

Wide BGP communities will act on and hence need to encode some distinct atoms of data. These are encoded as Sub-TLVs, where each Sub-TLV has the following format:

```

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

The Sub-Type field contains a value of 0-254. The value 255 is reserved for future use. The sub-TLV types are to be assigned and maintained by IANA registry.

The length represents the total length of a given sub-TLV in octets.

The value field contains the sub-TLV value.

Supported format of the sub-TLVs can be:

3.1. Community Value

Community Value: 4 octets

The Wide BGP Community value indicates what set of actions a router is requested to take upon reception of a route containing this community. The semantics of this value depend on whether this is a private/local community (when R is 0) or registered (when R is 1).

3.2. Source AS Number

Source Autonomous System Number: 4 octets

The Autonomous System number which indicates the originator of this Wide BGP Community.

When the Autonomous System is a two octet number the first two octets of this 4 octet value MUST be filled with zeros.

3.3. Target AS Number

Target Autonomous System Number: 4 octets

The Autonomous System number that indicates the context of the Registered/Local Value. When the value is a Registered Value (and thus registered with IANA), this field MUST be 0.

When the wide community is locally registered, the Target Autonomous System Number indicates the AS that defines the format of this wide community for the given Local Value. (In other words, value 1 will likely refer to different formats for AS 1 vs. AS 2.)

3.4. Wide Community Target(s) TLV

Type: 1

The Wide Community Target(s) TLV has the same format as a Wide Community atom.

Wide Community Targets define the matching criteria for the community. A given wide community may have a number of targets that it applies to. The semantics of these targets will vary on a per community basis. Depending on the definition of the community, targets may be optional.

The value field of the Wide Community Target(s) TLV is a series of

Wide Community Atom TLVs. The semantics of any given atom TLV MUST be part of the definition of a given Wide Community.

Typically, Wide Community Targets consist of a series of atoms that have "match any" semantics. Thus, if any given target matches per the semantics of that atom for the community, the community is considered to match and the action defined by the community should be executed.

The Grouping Container atom permits a set of atoms with semantics defined by the community to be nested. The Grouping Container atom is considered to be a matching target if, and only if, all of its contained atoms match per the semantics of the community.

If the semantics of a given atom is undefined for the community in question, it MUST be ignored. If an atom with undefined semantics is part of a Grouping Container, the entire container MUST be ignored.

When no targets are required by the definition of a given Wide Community, the Wide Community Target TLV SHOULD NOT be encoded in the community. Implementations MUST be prepared to accept a Wide Community Target TLV with an empty value field.

3.5. Wide Community Parameter(s) TLV

Type: 2

The Wide Community Parameter(s) TLV has the same format as a Wide Community atom.

A given wide community may have parameters which are used as inputs for executing actions defined for that community. These parameters, and any constraints implied by the parameters, MUST be defined by the wide community definition. Parameters consist of an ordered set of atom sub-TLVs. The semantics of any specific positional instance of an atom MUST be defined by the wide community.

If it is the case that a parameter for a given community is of an unexpected type, the community MUST be ignored.

If it is the case that there are too many or too few parameters for a given community, the community MUST be ignored.

When no parameters are required by the definition of a given Wide Community, the Wide Community Parameters TLV SHOULD NOT be encoded in the community. Implementations MUST be prepared to accept a Wide Community Parameter TLV with an empty value field.

3.6. Usage

The detailed interpretation of the targets or parameters SHALL be provided when describing given community type in a separate document or when locally defined by an operator.

4. Well Known Standard BGP Communities

According to RFC 1997, as well as IANA's Well-Known BGP Communities registry, the following BGP communities are defined to have global significance:

0xFFFF0000	planned-shut	[draft-francois-bgp-gshut]
0xFFFFFFFF01	NO_EXPORT	[RFC1997]
0xFFFFFFFF02	NO_ADVERTISE	[RFC1997]
0xFFFFFFFF03	NO_EXPORT_SUBCONFED	[RFC1997]
0xFFFFFFFF04	NOPEER	[RFC3765]

This document recommends for simplicity as well as for avoidance of backward compatibility issues the continued use of BGP Standard Community Attribute type 8 as defined in RFC 1997 to distribute non Autonomous System specific Well-Known BGP Communities.

For the same reason, this document does not intended to obsolete the currently defined and deployed BGP Extended Communities.

5. Operational considerations

Having two different ways to propagate locally assigned BGP communities, one via the use of Standard BGP Communities and the other one via the use of Wide BGP Communities, may seem to potentially cause problems when considering propagation of conflicting actions. However, even at present, an operator may append Standard BGP Communities with conflicting information. It is therefore recommended that any implementation, in supporting both standard and Wide BGP communities, allow for their easy inbound and outbound processing. The actual execution of all communities should be treated as a union and, if supported by an implementation, their execution permissions are to be a local configuration matter.

6. Example

6.1. Example Wide Community Definition

An operator wishes to locally define a Wide Community with the semantics of permitting AS_PATH prepending with targets that include AS numbers of peer ASes and peers who have been marked with a set of defined "color" strings.

Target semantics:

Grouping containers MAY be used.

The Autonomous System Number atom refers to the target peer AS Number.

The UTF-8 String atom refers to a peer "color". The values are constrained to the strings "red", "green" or "blue".

The semantics of all other atoms are undefined for this community.

Parameter semantics:

The parameter TLV shall consist of exactly one integer value that is constrained to have a value of 2..8.

6.2. Example Wide Community Encoding

AS_PATH prepend 4 TIMES TO AS 2424, AS 8888, to peers marked as "red" or to peers marked "blue" AND AS 1111.

Use TTL 0 to request the receiving router to not propagate this wide community.

Locally community value (flag bit 0 = 0).

Do not decrement TTL field across confederation boundaries (0)

Local community 1 for sample AS 64512.


```

      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
+-----+-----+-----+-----+
|      Container Type 1 (1)      |
+-----+-----+-----+-----+
|0 0 0 0 0 0 0 0 0|
+-----+-----+-----+-----+
|      TTL: 0      |
+-----+-----+-----+-----+
|      Length: 34      |
+-----+-----+-----+-----+
|      Community: LOCAL PREPEND ACTION CATEGORY 1      |
+-----+-----+-----+-----+
|                               Own ASN                               |
+-----+-----+-----+-----+
|      Target ASN# 64512 (0x0000FC00)      |
+-----+-----+-----+-----+
| Target TLV (1)|      Length: 23      |
+-----+-----+-----+-----+
| Type ASN (1) |      Length: 4      |
+-----+-----+-----+-----+
|      Target ASN# 2424 (0x00000978)      |
+-----+-----+-----+-----+
| Type ASN (1) |      Length: 4      |
+-----+-----+-----+-----+
|      Target ASN# 8888 (0x000022B8)      |
+-----+-----+-----+-----+
| Type Str (5) |      Length: 3      |
+-----+-----+-----+-----+
|      Peer color "red"      |
+-----+-----+-----+-----+
| Target Grp (7)|      Length: 12      |
+-----+-----+-----+-----+
| Type Str (5) |      Length: 4      |
+-----+-----+-----+-----+
|      Peer color "blue"      |
+-----+-----+-----+-----+
| Type ASN (1) |      Length: 4      |
+-----+-----+-----+-----+
|      Target ASN# 1111 (0x00000457)      |
+-----+-----+-----+-----+
| Param TLV (2) |      Length: 3      |
+-----+-----+-----+-----+
| Type INT (4) |      Length: 1      | Prepend #: 4      |
+-----+-----+-----+-----+

```

7. Security considerations

All the security considerations for BGP Communities as well as for BGP RFCs apply here.

8. IANA Considerations

This document defines a new BGP Path Attribute called Wide BGP Community Attribute. For this new type IANA is to allocate a new value in the corresponding registry:

Registry Name: BGP Path Attributes

This document makes the following assignments for the optional, transitive Wide BGP Communities Attribute:

Name	Type Value
----	-----
Wide BGP Community Attribute	TBA

This document requests IANA to define and maintain a new registry named: "Wide BGP Communities Attribute Container Types".

The pool of: 0x0000-0xFFFF has been defined for its allocations. The allocation policy is on a first come, first served basis.

This document makes the following assignments for the Wide BGP Communities Attribute Types values:

Name	Type Value
----	-----
Reserved	0x0000
Type 1	0x0001
Types 2-1023 to be allocated using IETF Consensus	
Types 1024-64511 to be allocated first come, first served	
Types 64512-65534 are reserved for experimental use	
Reserved	0xFFFF

This document requests IANA to define and maintain a new registry named: "Wide BGP Communities sub-TLV types". The pool of 0x0000-0xFFFF has been defined for its allocations. This document defines type 1. Types 2-1024 are to be allocated using an IETF Consensus policy. Types 1024-64511 are to be allocated on a first come, first served basis. Types 64512-65534 are to be reserved for experimental

use.

This document makes the following assignments for the Wide BGP Communities sub-TLV type values:

Name	Type Value
----	-----
Reserved	0x00
AS Number	0x01
IPv4 Prefix	0x02
IPv6 Prefix	0x03
Integer	0x04
UTF-8 string	0x05
IEEE Floating Point Value	0x06
Container Group	0x07
Reserved	0xFF

9. Change History

Changes since from -01 to -02:

The Type field has been expanded to 2 octets.

The Length field has been moved to the common header.

Changed format to use TLVs.

Added atom TLV to define well defined syntactic items.

Added TLVs to distinguish targets from parameters.

Various editorial changes to language.

10. Contributors

The following people contributed significantly to the content of the document:

Shintaro Kojima
OTEMACHI 1st. SQUARE EAST TOWER, 3F
1-5-1, Otemachi,
Chiyoda-ku, Tokyo 100-0004
Japan
Email: koji@mfeed.ad.jp

Juan Alcaide
Cisco Systems
Research Triangle Park, NC
United States
Email: jalcaide@cisco.com

Burjiz Pithawala
Cisco Systems
170 West Tasman Dr
San Jose, CA
United States
Email: bpithaw@cisco.com

Saku Ytti
TDC Oy
Mechelininkatu 1a
00094 TDC
Finland
Email: ytti@tdc.net

11. Acknowledgments

Authors would like to thank Enke Chen, Pedro Marques and Alton Lo for their valuable input.

12. References

12.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended

Communities Attribute", RFC 4360, February 2006.

12.2. Informative References

- [RFC1997] Chandrasekeran, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, August 1996.
- [RFC1998] Chen, E. and T. Bates, "An Application of the BGP Community Attribute in Multi-home Routing", RFC 1998, August 1996.
- [RFC4384] Meyer, D., "BGP Communities for Data Collection", BCP 114, RFC 4384, February 2006.
- [RFC4893] Vohra, Q. and E. Chen, "BGP Support for Four-octet AS Number Space", RFC 4893, May 2007.
- [RFC5668] Rekhter, Y., Sangli, S., and D. Tappan, "4-Octet AS Specific BGP Extended Community", RFC 5668, October 2009.

Authors' Addresses

Robert Raszuk (editor)
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US

Email: raszuk@cisco.com

Jeffrey Haas (editor)
Juniper Networks
1194 N.Mathilda Ave
Sunnyvale, CA 94089
US

Email: jhaas@juniper.net

Shane Amante
Level 3 Communications, LLC
1025 Eldorado Blvd
Broomfield, CO 80021
US

Email: shane@level3.net

Richard A Steenbergen
nLayer Communications, Inc.
209 W Jackson Blvd
Chicago, IL 60606
US

Email: ras@nlayer.net

Bruno Decraene
France Telecom
38-40 rue du General Leclerc
Issi Moulineaux cedex 9 92794
France

Email: bruno.decraene@orange-ftgroup.com

Paul Jakma
University of Glasgow
School of Computing Science
Sir Alwyn Williams Building
University of Glasgow
Glasgow G12 8QQ
UK

Email: paulj@dcsc.gla.ac.uk

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 5, 2012

Q. Zeng
J. Dong
Huawei Technologies
July 4, 2011

Maximum Transmission Unit Extended Community for BGP-4
draft-zeng-idr-bgp-mtu-extension-00

Abstract

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

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

Requirements Language

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

Status of this Memo

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

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

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

This Internet-Draft will expire on January 5, 2012.

Copyright Notice

Copyright (c) 2011 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. BGP LSP MTU Discovery	3
2.1. Definitions	3
2.2. MTU Extended Community	3
2.3. Signaling	4
3. Applicability Considerations	4
4. IANA Considerations	4
5. Security Considerations	4
6. Contributors	5
7. Acknowledgements	5
8. References	5
8.1. Normative References	5
8.2. Informative References	5
Authors' Addresses	6

1. Introduction

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

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

2. BGP LSP MTU Discovery

2.1. Definitions

BGP LSP Path MTU: The Path MTU of the LSP from given BGP LSR to a specific prefix. It is carried as a Extended Community with the BGP labeled IPv4 (or IPv6) route.

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

2.2. MTU Extended Community

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

0										1										2										3																			
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1																		
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+																			
MTU extended community Type																				Reserved																													
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+																			
										Reserved																				MTU Value																			
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+										+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+																			

The MTU extended community type is to be assigned by IANA. The first four octets of the value field should be reserved, and the MTU value

is carried in the following two octets of the value field.

2.3. Signaling

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

- A. If BGP speaker A is originator of the labeled IPv4 (or IPv6) route, A sets its BGP LSP Path MTU to the maximal value, advertises the labeled IPv4 (or IPv6) route with the MTU Extended Community to its BGP Peer (its upstream BGP LSR).
- B. BGP speaker B receives the labeled IPv4 (or IPv6) route with BGP LSP Path MTU from its BGP peer.
 - a) B SHOULD compute the BGP LSR Link MTU to the Next Hop of the received message, then sets its BGP LSP Path MTU to the minimum of the received BGP LSP Path MTU and the BGP LSR Link MTU.
 - b). If B distributes the route with the Next Hop attribute unchanged, it MUST keep the MTU Extended Community unchanged when advertising the message to its upstream BGP LSRs.
 - c). If B would change the Next Hop attribute to itself in the subsequent advertisement, it SHOULD set the MTU Extended Community in the message with its BGP LSP Path MTU obtained through step (a).

3. Applicability Considerations

The BGP MTU Extended Community is applicable to the BGP LSP defined in [RFC3107].

4. IANA Considerations

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

5. Security Considerations

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

6. Contributors

The following individuals contributed to this document:

Haibo Wang

Haijun Xu

7. Acknowledgements

TBD

8. References

8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, May 2001.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006.

8.2. Informative References

- [RFC1191] Mogul, J. and S. Deering, "Path MTU discovery", RFC 1191, November 1990.
- [RFC3988] Black, B. and K. Kompella, "Maximum Transmission Unit Signalling Extensions for the Label Distribution Protocol", RFC 3988, January 2005.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC4659] De Clercq, J., Ooms, D., Carugi, M., and F. Le Faucheur, "BGP/MPLS IP Virtual Private Network (VPN) Extension for IPv6 VPN", RFC 4659, September 2006.

Authors' Addresses

Qing Zeng
Huawei Technologies
Huawei Building, No.3 Xinxu Rd
Beijing 100085
China

Email: zengqing@huawei.com

Jie Dong
Huawei Technologies
Huawei Building, No.3 Xinxu Rd
Beijing 100085
China

Email: jie.dong@huawei.com

