Network Working Group A. Bashandy
Internet Draft B. Pithawala
Intended status: Standards Track Cisco Systems
Expires: August 2011 February 28, 2011

Scalable, Loop-Free BGP FRR using Repair Label
draft-bashandy-idr-bgp-repair-label-00.txt

Abstract

Consider a BGP free core scenario. Suppose the provider edge BGP speaker 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 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 without 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.

This Internet-Draft will expire on August 28, 2011.

Table of Contents

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.

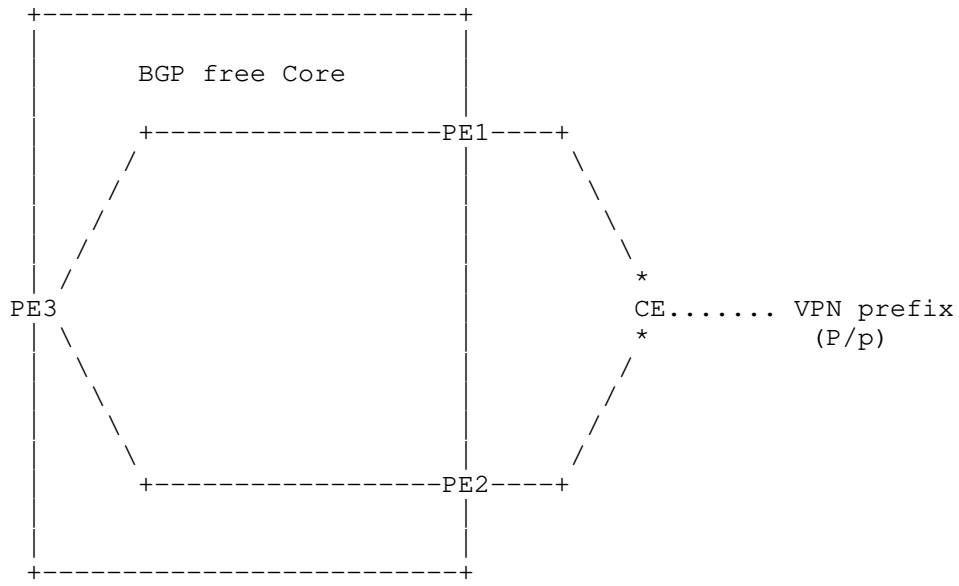


Figure 1 VPN prefix reachable via multiple PEs

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 and without waiting for the control plane (whether it is IGP or BGP) to re-converge and computes new the 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 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]

- 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.

- 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 when there is no failure. Referring to Figure 1, PE1 is a primary egress PE.

- CE: It is 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 CE

- 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. PE3 in Figure 1 is an example of an ingress PE

- Repairing PE: It is 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 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.

- Primary label: It is the label advertised by the primary egress PE to be used for normal traffic forwarding.

- 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

- Repair label: It is 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 methods for disseminating repair label information.

- Repair path: It is the repair egress PE and the repair label.

2. Protocol Operation

This section explains the operation of the control and forwarding planes of routers participating 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. If repair label advertisement is used (Sections 3.1.), 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 prefix advertised by more than one PE, an egress PE will have

    - a primary path

    - a repair path consisting of a repair PE and a repair label advertised by or agreed upon with 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 repairing PE MUST NOT advertise a repair for a prefix if it does NOT have an external path to the prefix
- o A repairing 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 on Losing Primary path Reachability

As soon as a PE loses reachability to the primary path of the protected prefix P/p, the forwarding plane processes arriving traffic as follows:

1. If the repair label is an advertised label
    a. 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 swaps the incoming label stack with the "repair label stack" advertised by the repair egress PE. Section 3.1.2. specifies all the details
    b. If the repairing PE is an ingress PE, it pushes the "repair label stack" advertised by the repair egress PE. Section 3.1.2. specifies all the details
2. If the repair label is an agreed upon service label
    a. If the repairing PE is an egress PE, it swaps the incoming label with the normal label advertised by the repair PE. Otherwise it pushes the primary label advertised by the repair PE.
    b. The repairing PE pushes the repair label on top of the label stack.
    c. Section Error! Reference source not found.specifies 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. If the repair label is a service label, the repair PE pops the service label and uses the rest of the label stack. Otherwise the repair PE uses the incoming label stack

5. If the repair egress PE can reach the CE, the repair PE forwards the packet towards the CE.

6. If the repair CE 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.

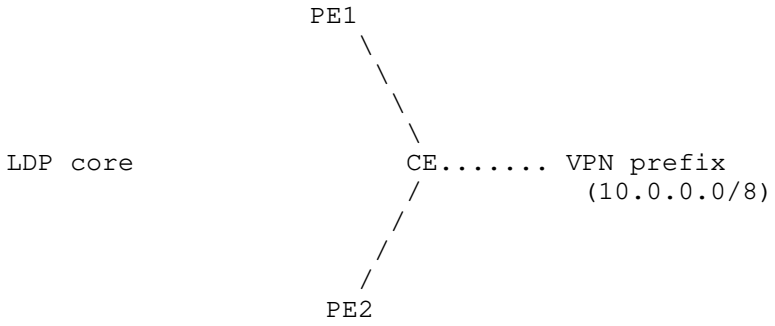


Figure 2 : L3VPN Example

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 PE 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

To ensure maximum flexibility, we specify two approaches to disseminate the repair label:

- Advertise the repair label as an optional path attribute
- An agreed upon service label

3.1. Advertising the repair label as an Optional Path Attribute

Advertising the repair label as an optional path attributes has some advantages:

- An egress PE can benefits from a scalable repair label allocation schemes such as per-CE repair label allocation
- 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.
- Reduces the BGP update message size. Instead of having to send additional labels per prefix, multiple prefixes can share the same repair label
- The number of labels used for traffic restoration does not depend on the number of protected prefixes
- 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 maximizes 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: 129

Attribute Flags:

Optional bit: 1

Transitive bit: 0

Partial bit: 0

Extended Length bit: 0

Length of the attribute: It indicates the 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 ''Lr''. 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 ''L1i'' Suppose PE1 receives a packet with the label ''L1i'' at the top of the stack. 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 has to swap the label ''L1i'' 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 router 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 MUST 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].

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

- Option 1: The BGP speaker MUST NOT advertise the repair label attribute
- 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 ''129'' 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

Network working group
Internet Draft
Intended status: Standards Track
Expires: September 2011

Q. Zeng
J. Dong
Huawei Technologies
J. Heitz
Ericsson Inc.
K. Patel
Cisco Systems
R. Shakir
C&W

One-time Extended Community Based Outbound Route Filter for BGP-4

draft-dong-idr-one-time-ext-community-orf-00.txt

Abstract

This document defines a new Outbound Router Filter (ORF) type for BGP, termed "One-time Extended Community Outbound Route Filter", which would allow a BGP speaker to send to its BGP peer a route refresh request with a set of extended-community-based filters to make the peer re-advertise only the specific routes matching the filters to the speaker. This ORF-type enables a BGP speaker to refresh some specific routes without requiring its peer to re-advertise the whole Adj-RIB-Out, which makes the route refresh operation more efficient and reduces the impact on network stability. This filter does not change the outbound route filters on BGP peers and should only be used for one-time filtering.

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/1id-abstracts.txt.

Copyright Notice

Conventions used in this document

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

Table of Contents

1. Introduction

The Outbound Route Filtering Capability defined in [RFC5291] provides a mechanism for a BGP speaker to send to its BGP peer a set of Outbound Route Filters (ORFs) that can be used by its peer to filter its outbound routing updates to the speaker.

During some network operations, BGP speaker only needs to retrieve some routes with specific extended communities from its peer, but sending plain ROUTE-REFRESH will lead to the peer re-advertising its whole Adj-RIB-Out. Such large amounts of updates include a lot of unnecessary routes which would result in waste of processing resources and bandwidth. With the increase of IPv6 deployment, this problem could be more significant. Even configured with ORF mechanism as defined in [RFC5291], on receipt of a ROUTE-REFRESH message, the peer will re-advertise all the routes matching current outbound route filters, i.e., the whole Adj-Rib-Out for this BGP speaker. Since in this case the BGP speaker does not want to change the outbound route filters on its peer, this requirement cannot be met by current ORF mechanism.

This document defines a new Outbound Router Filter (ORF) type for BGP, termed "One-time Extended Community Outbound Route Filter", which would allow a BGP speaker to send to its BGP peer a route refresh request with a set of Extended Community based filters to make the peer re-advertise only the specific routes matching the filters to the speaker. This ORF-type enables a BGP speaker to retrieve routes with specific Extended Communities without requiring its peer to re-advertise the whole Adj-RIB-Out, which makes such route refresh operation more efficient and also reduces the impact on network stability. This filter does not change the outbound route filters on BGP peers and should only be used for one-time filtering.

One use case of one-time Extended Community ORF would be to refresh routes with specific Route Target (RT) Extended Community. For example, on receipt of routes with specific RTs, according to local policies some attributes of the routes may be changed, or some routes may be discarded. When later such local policies are changed or removed, the routes impacted by such policies need to be refreshed and processed according to the new local policies. With the whole Adj-RIB-Out route refresh it would result in a lot of unnecessary routes being re-advertised, and this would be a waste of the processing resource and bandwidth. In this case, one-time Extended Community ORF would be quite useful to request only routes matching specific RTs to be re-advertised.

2. One-time Extended Community ORF-Type

This document defines a new ORF type: One-time Extended Community ORF. Value of this ORF-Type is to be assigned by IANA.

In the following description, the sending speaker sends a one-time

ORF request and the receiving speaker receives it and sends back the routes to satisfy the request.

As specified in the [RFC5291], an ORF entry is a tuple of the form <AFI/SAFI, ORF-Type, Action, Match, ORF-value> an ORF consists of one or more ORF entries that have a common AFI/SAFI and ORF-Type. An ORF is identified by <AFI/SAFI, ORF-Type>.

The type-specific part consists of a single Extended Community encoded as an eight-octets field.

Since the semantics of this new ORF-Type is "one-time filtering" and has no impact on existing ORFs, the Action field is irrelevant and MUST be ignored on receipt.

The MATCH field of the One-time Extended Community ORF SHOULD be set to PERMIT on the sender and SHOULD be ignored on the receiver. This is the same as defined in Extended-Community ORF [EXT-COMM-ORF].

The ORF entries of this type would only be used as one-time filters that MUST not change any previously installed ORF entry on the receiving speaker.

3. Operation

The capability negotiation of <AFI/SAFI, One-time Extended Community ORF> MUST NOT delay the advertisement of routes with this AFI/SAFI.

The received One-time Extended Community ORF entries SHOULD only be used for one-time route filtering and MUST NOT be saved locally. The received One-time Extended Community ORF entries MUST NOT modify the outbound route filters on the receiving speaker (either locally configured or received from the sending speaker through ORF).

On receipt of ROUTE-REFRESH message with One-time Extended Community ORF entries, the receiving speaker SHOULD re-advertise to the sending speaker the routes from the Adj-RIB-Out associated with the sending speaker which pass the entries carried in the One-time Extended Community ORF as well as the locally saved ORFs (if any) received from the sending speaker.

Since different processing orders may lead to different results, the One-time ORFs and the regular ORFs SHOULD not be encoded in one route-refresh message.

During the period when the receiving speaker is sending updates to satisfy the One-time ORF request, it may experience other routing activity that will require it to send updates unrelated to the One-time ORF request. It is permitted to send these updates before it has completed sending the One-time ORF related updates.

Similarly, if a route that passes the One-time ORF has already been sent and the receiving speaker experiences routing activity that changes this route and the receiving speaker has not yet sent all routes to satisfy the One-time ORF request, it is permitted to send the changed route immediately.

Details about how to interoperate when both One-time ORF Capability and the Enhanced Route Refresh Capability as described in [Enhanced-Refresh] are enabled will be discussed in the next version.

4. Security Considerations

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

5. IANA Considerations

This document specifies a new Outbound Route Filtering (ORF) type, One-time Extended Community ORF. The value of the ORF-type needs to be assigned by IANA.

6. Acknowledgments

The authors would like to thank Robert Raszuk, John Scudder, Susan Hares, Haibo Wang, Jiawei Dong, Yaqun Xiao, Mach Chen for their valuable suggestions and comments to this document.

7. References

7.1. Normative References

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

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

[RFC5291] Chen, E. and Y. Rekhter, "Outbound Route Filtering Capability for BGP-4", RFC 5291, August 2008.

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

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

[RFC4020] Kompella, K. and A. Zinin, "Early IANA Allocation of Standards Track Code Points", BCP 100, RFC 4020, February 2005.

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

7.2. Informative References

[Enhanced-Refresh] K. Patel, E. Chen and B. Venkatachalapathy, "Enhanced Route Refresh Capability for BGP-4", draft-keyur-bgp-enhanced-route-refresh-01.txt, Ocotober 2010

Authors' Addresses

Jie Dong
Huawei Technologies Co.,Ltd.
Huawei Building, No.3 Xinxi Rd.,
Hai-Dian District
Beijing, 100085
P.R. China

Email: jie.dong@huawei.com

Qing Zeng
Huawei Technologies Co.,Ltd.
Huawei Building, No.3 Xinxi Rd.,
Hai-Dian District
Beijing, 100085
P.R. China

Email: zengqing@huawei.com

Jakob Heitz
Ericsson Inc.
100 Headquarters Drive
San Jose CA 95134
USA

Email: jakob.heitz@ericsson.com

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

Email: keyupate@cisco.com

Rob Shakir
Cable&Wireless Worldwide

Email: rob.shakir@cw.com

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

H. Gredler
J. Medved
Juniper Networks, Inc.

March 3, 2011

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

Abstract

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

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

Requirements Language

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

Status of this Memo

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

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

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

This Internet-Draft will expire on September 4, 2011.

Table of Contents

1. Introduction

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

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

A router maintains a database for storing Traffic Engineering related data and link information. The Traffic Engineering Database (TED) is populated by a link-state IGP routing protocol that supports TE extensions: IS-IS or OSPF. The TED can be seen as a protocol-neutral representation of links in the area. Link attributes stored in the TED are: local/remote IP addresses, local/remote interface indices, metric, link bandwidth, reservable bandwidth, per CoS class reservation state, preemption and Shared Risk Link Groups (SRLG). The router’s BGP process can retrieve the TE data from the TED database and distribute it to peer BGP Speakers using the encoding specified in this draft.

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

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

2. Scope

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

3. Transcoding TE Link Information Into a BGP NLRI

The MP_REACH and MP_UNREACH attributes are BGP's containers for carrying opaque information. Each TED NLRI describes a single link anchored by at least a pair of router-IDs. Since there are many Router-IDs formats (32 Bit IPv4 router-ID, 56 Bit ISO Node-ID and 128 Bit IPv6 router-ID) a link may be anchored by more than one Router-ID pair. The anchoring Router-IDs are carried in the Node Anchor TLVs.

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

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

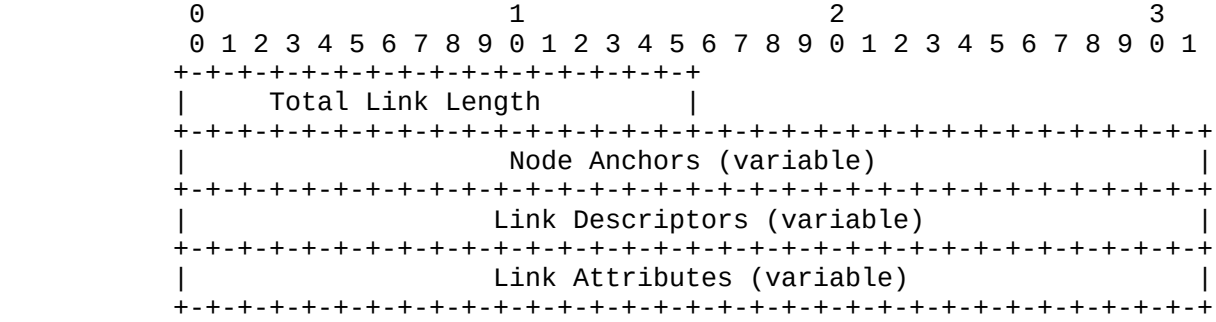


Figure 1: TED SAFI 1 NLRI Format

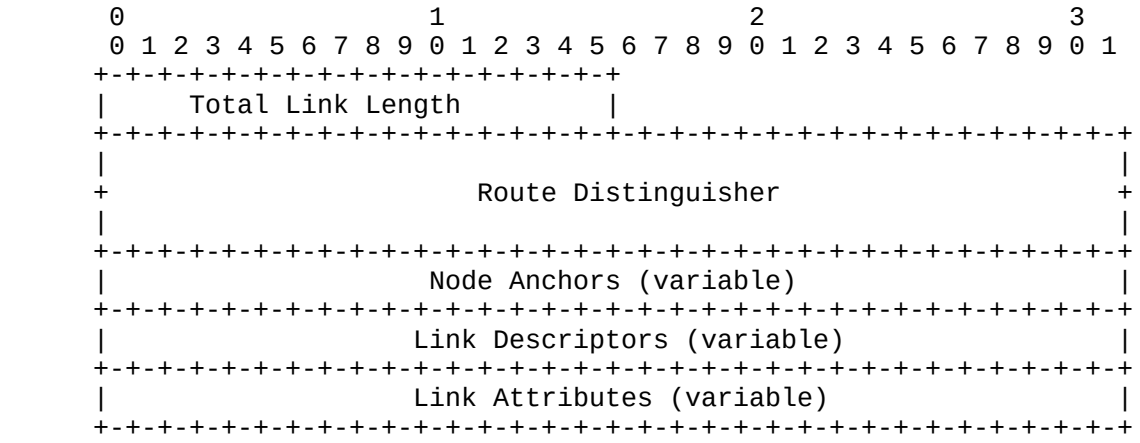


Figure 2: TED SAFI 128 NLRI Format

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

3.1. TLV Format

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

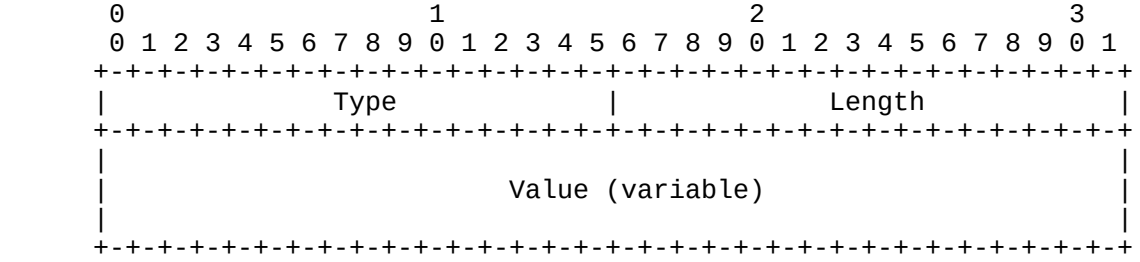


Figure 3: TLV format

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

3.2. Node anchors

The set of Node Anchor TLVs describes which Protocols Router-IDs will be following to "anchor" the link described by the "Link attribute TLVs". There must be at least one "like" router-ID pair per-protocol. If a peer sends an illegal combination in this respect, then this is handled as an NLRI error, described in [RFC4760].

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

Table 1: Node Anchor TLVs

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

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

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

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

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

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

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

are 4 octets wide as described in [RFC4893]. 2-octet AS Numbers shall be expanded to 4-octet AS Numbers by zeroing the two MSB octets.

3.2.1. Router-ID Anchoring Example: ISO Pseudonode

IS-IS Pseudonodes are a good example for the variable Router-ID anchoring. Consider Figure 4. This represents a Broadcast LAN between a pair of routers. The "real" (=non pseudonode) routers have both an IPv4 Router-ID and IS-IS Node-ID. The pseudonode does not have an IPv4 Router-ID. Two unidirectional links (Node1, Pseudonode 1) and (Pseudonode 1, Node 2) are being generated.

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

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

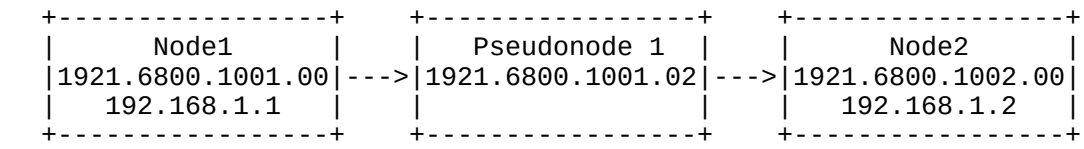


Figure 4: IS-IS Pseudonodes

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

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

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

3.3. Link Descriptors

The 'Link Descriptor' field is a set of Type/Length/Value (TLV) triplets. The format of each TLV is shown in Figure 3. The 'Link descriptor' TLVs uniquely identify a link between a pair of anchor Routers.

The encoding of 'Link Descriptor' TLVs, i.e. the Codepoints in

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

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

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

Table 2: Link Descriptor TLVs

3.4. Link Attributes

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

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

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

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

Type	Description	Defined in:
3	Administrative group (color)	[RFC5305], Section 3.1
9	Maximum link bandwidth	[RFC5305], Section 3.3
10	Max. reservable link bandwidth	[RFC5305], Section 3.5
11	Unreserved bandwidth	[RFC5305], Section 3.6
20	Link Protection Type	[RFC5307], Section 1.2
64512	TE Default Metric	Section 3.4.1
64513	IGP Link Metric	Section 3.4.2
64514	Shared Risk Link Group	Section 3.4.3

Table 3: Link Attribute TLVs

3.4.1. TE Default Metric TLV

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

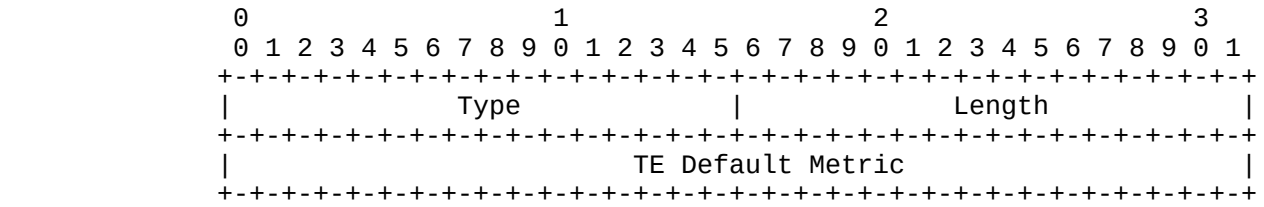


Figure 5: TE Default metric TLV format

3.4.2. IGP Link Metric TLV

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

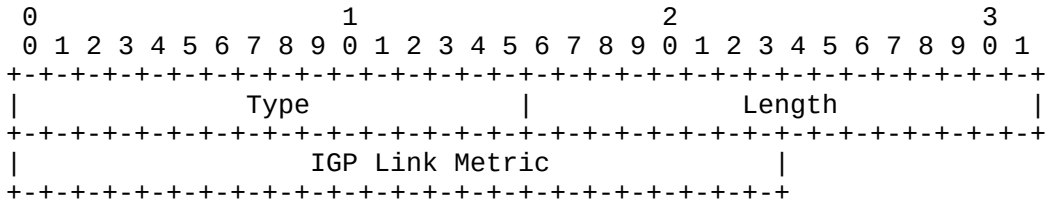


Figure 6: IGP Link Metric TLV format

3.4.3. Shared Risk Link Group TLV

The Shared Risk Link Group (SRLG) TLV (Type 64514) carries the Shared Risk Link Group information (see Section 2.3, "Shared Risk Link Group Information", of [RFC4202]). It contains a data structure consisting of a (variable) list of SRLG values, where each element in the list has 4 octets, as shown in Figure 7. The length of this TLV is 4 * (number of SRLG values).

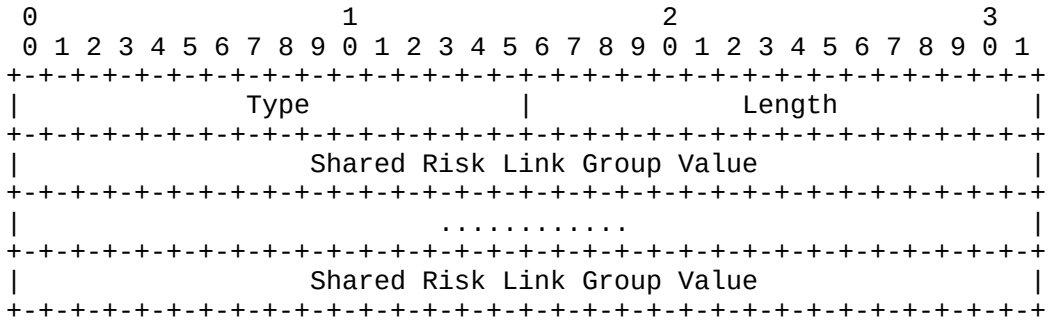


Figure 7: Shared Risk Link Group TLV format

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

3.5. IGP Area Information

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

3.6. Inter-AS Links

The main source of TE information is the IGP, which is not active on inter-AS links. In order to inject a non-IGP enabled link into the traffic-engineering database (TED) an implementation must support configuration of static TE links.

4. Link to Path Aggregation

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

4.1. Example: No Link Aggregation

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

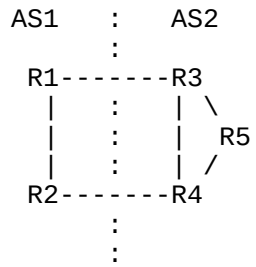


Figure 8: no-link-aggregation

4.2. Example: ASBR to ASBR Path Aggregation

The brief difference between the "no-link aggregation" example and this example is that no specific link gets exposed. Consider Figure 9. The only link which gets advertised by AS2 is an "aggregate" link between R3 and R4. This is enough to tell AS1 that there is a backup path. However the actual links being used are hidden from the topology.

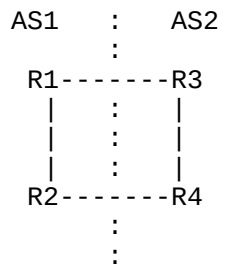


Figure 9: asbr-link-aggregation

4.3. Example: Multi-AS Path Aggregation

Service providers in control of multiple-ASes may even decide to not expose their internal inter-AS links. Consider Figure 10. Rather than exposing all specific R3 to R6 links, AS3 is modeled as a single node which connects to the border routers of the aggregated domain.

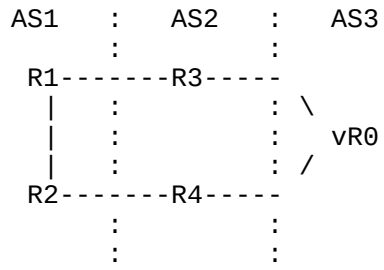


Figure 10: multi-as-aggregation

5. Originating the TED NLRI

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

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

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

6. Receiving the TED NLRI

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

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

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

7. Use Cases

7.1. MPLS TE

If a router wants to compute a MPLS TE path across IGP areas TED lacks visibility of the complete topology. This is an issue for large scale networks that need to segment their core networks into distinct areas because inter-area TE cannot get deployed there. Current solutions for inter area TE only compute the path for the first area. The router only has full topological visibility for the first area along the path, but not for subsequent areas. The best practice is to use a technique called "loose-hop-expansion" which uses the IGP computed shortest path topology for the remainder of the path. Therefore no non-SPF based path setup is possible across areas. This has disadvantages for path protection and path engineering applications, as shown in Figure 11.

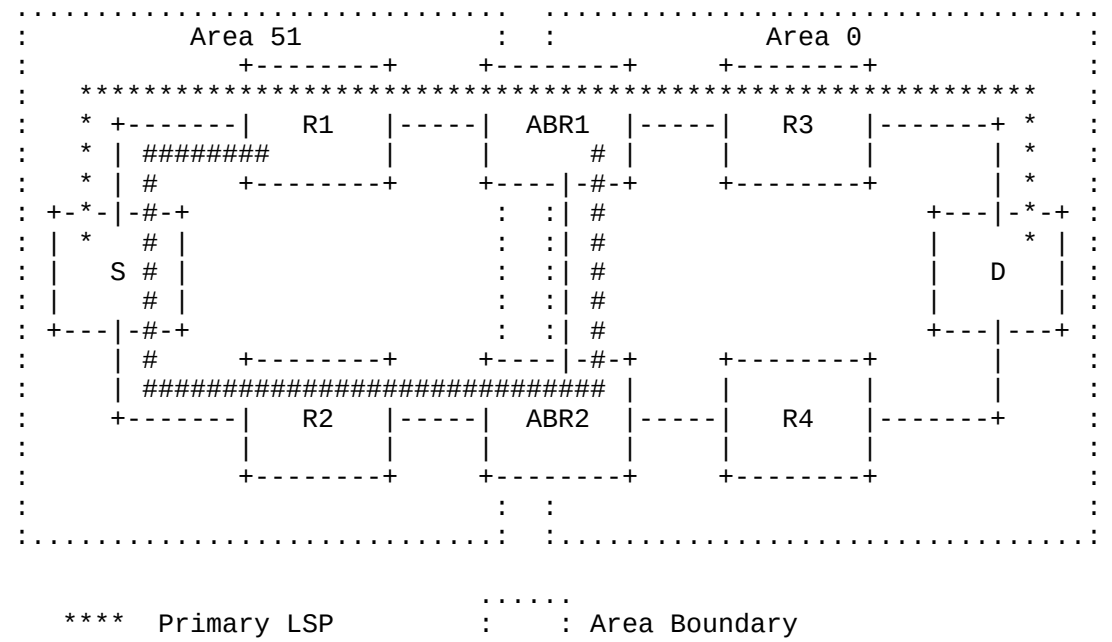


Figure 11: MPLS TE Bypass LSP problem

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

7.2. ALTO Server Network API

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

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

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

7.3. Path Computation Element (PCE) TED Synchronization Protocol

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

8. IANA Considerations

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

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

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

9. Security Considerations

This draft does not affect the BGP security model.

10. Acknowledgements

We would like to thank Alia Atlas, David Ward, John Scudder, Kaliraj Vairavakkalai, Nischal Sheth and Yakov Rekhter from Juniper Networks, Inc. and Richard Woundy from Comcast for their invaluable input and comments.

11. References

11.1. Normative References

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

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

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

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

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

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

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

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

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

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

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

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

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

11.2. Informative References

[I-D.ietf-alto-protocol]
Alimi, R., Penno, R., and Y. Yang, "ALTO Protocol", draft-ietf-alto-protocol-06 (work in progress), October 2010.

Authors' Addresses

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

Email: hannes@juniper.net

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

Email: jmedved@juniper.net

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 5, 2011

K. Patel
R. Fernando
Cisco Systems
J. Scudder
J. Haas
Juniper Networks

February 2011

Graceful Restart Extensions for BGP
draft-keyupate-idr-bgp-gr-extension-00.txt

Abstract

The current BGP Graceful Restart mechanism limits the usage of BGP Graceful Restart to BGP protocol messages other than a BGP NOTIFICATION message. This document defines an extension to the BGP Graceful Restart that permits the Graceful Restart procedures to be performed when the BGP speaker receives a BGP NOTIFICATION Message. This document also defines a new BGP NOTIFICATION Cease Error subcode to prevent BGP speakers supporting the extension defined in this document from performing a Graceful Restart.

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 August 5, 2011.

Copyright Notice

Table of Contents

1. Introduction

For many classes of errors, the BGP protocol must send a NOTIFICATION message and reset the peering session to handle the error condition. The BGP Graceful Restart extension defined in [RFC4724] requires that normal BGP procedures defined in [RFC4271] be followed when a NOTIFICATION message is sent or received. This document defines an extension to BGP Graceful Restart that permits the Graceful Restart procedures to be performed when the BGP speaker receives a NOTIFICATION message. This permits the BGP speaker to avoid flapping reachability and continue forwarding while the BGP speaker restarts the session to handle errors detected in the BGP protocol.

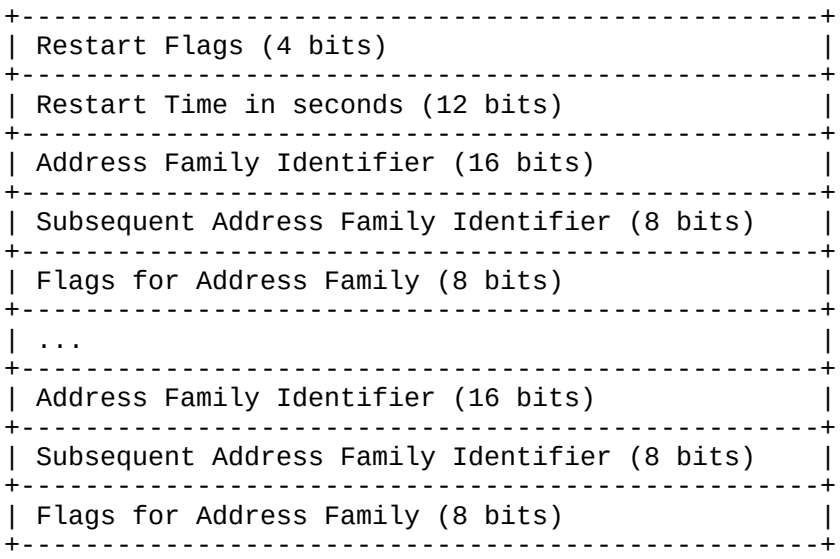
This document defines a BGP NOTIFICATION cease Error subcode for the Cease Error code to prevent BGP speakers supporting the extension defined in this document from performing a Graceful Restart.

1.1. Requirements Language

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

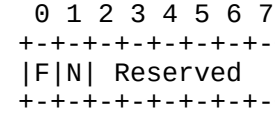
2. Modifications to BGP Graceful Restart Capability

The BGP Graceful Restart Capability is augmented to signal the Graceful Restart support for BGP NOTIFICATION messages. In particular, the flags field for Address Family is augmented as follows:



Flags for Address Family:

This field contains bit flags relating to routes that were advertised with the given AFI and SAFI.



The second most significant bit "N" is defined as a BGP Graceful Notification bit, which is used to indicate the Graceful Restart support for BGP NOTIFICATION messages. BGP speaker indicates the Graceful Restart support for BGP NOTIFICATION messages and its ability to handle the new BGP NOTIFICATION Cease message subcode and the format for a BGP NOTIFICATION Cease message defined in [RFC4486] when the Graceful NOTIFICATION bit is set (value 1).

3. BGP Hard Reset Subcode

A new BGP Cease message subcode is defined known as BGP Hard Reset Subcode. The value of this subcode is 9.

Whenever a BGP speaker receives a NOTIFICATION message with the Cease

Error code and Hard Reset Error subcode, the speaker MUST terminate the BGP session following the standard procedures in [RFC4271].

4. Operation

A BGP speaker that is willing to receive and send BGP NOTIFICATION messages in Graceful mode should advertise the BGP Graceful Notification Flag "N" using the Graceful Restart Capability as defined in [RFC4724].

When a BGP Speaker receives a BGP NOTIFICATION message, it SHOULD follow the standard rules of the receiving speaker mentioned in [RFC4724] for all AFI/SAFIs for which it has announced the BGP Graceful Notification flag. The BGP speaker generating a BGP NOTIFICATION message SHOULD follow the standard rules of the receiving Speaker in [RFC4724] for all AFI/SAFIs that were announced with the BGP Graceful Notification flag.

Once the session is re-established, both BGP speakers MUST set their "Forwarding State" bit to 1 if they want to apply planned graceful restart. The handling of the "Forwarding State" bit should be done as specified by the procedures of the Receiving speaker in [RFC4724] are applied.

As part of this extension, possible consecutive restarts SHOULD NOT delete a route (from the peer) previously marked as stale, until required by rules mentioned in [RFC4724].

5. Acknowledgements

The authors would like to thank Robert Raszuk for the review and comments.

6. IANA Considerations

This document defines a new BGP Cease message subcode known as BGP Hard Reset Subcode. IANA mantains the list of existing BGP Cease message subcodes. This document proposes defining a new BGP Cease message subcode known as BGP Hard Reset Subcode with the value 9.

7. Security Considerations

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

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.

[RFC2842] Chandra, R. and J. Scudder, "Capabilities Advertisement with BGP-4", RFC 2842, May 2000.

[RFC3392] Chandra, R. and J. Scudder, "Capabilities Advertisement with BGP-4", RFC 3392, November 2002.

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

[RFC4486] Chen, E. and V. Gillet, "Subcodes for BGP Cease Notification Message", RFC 4486, April 2006.

[RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, January 2007.

8.2. Informative References

[RFC2858] Bates, T., Rekhter, Y., Chandra, R., and D. Katz, "Multiprotocol Extensions for BGP-4", RFC 2858, June 2000.

Authors' Addresses

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

Email: keyupate@cisco.com

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

Email: rex@cisco.com

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

Email: jgs@juniper.net

Jeff Haas
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
USA

Email: jhaas@juniper.net

L3VPN Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 10, 2011

K. Patel
R. Raszuk
Cisco Systems
M. Djernaes
Juniper Networks
J. Dong
M. Chen
Huawei Technologies Co., Ltd.

March 9, 2011

IPv6 AF Extensions for Route Target Distribution
draft-keyur-bgp-af-specific-rt-constrain-01.txt

Abstract

The current route target distribution specification described in RFC4684 defines Route Target NLRIs of maxiumum length of 12 bytes. The IPv6 specific Route Target extended community is defined in RFC5701 as length of 20 bytes. Since the current specification only supports prefixes of maximum length of 12 bytes, the lack of an IPv6 specific Route Target reachability information may be a problem when an operator wants to use this application in a pure IPv6 environment. This document defines an extension that allows BGP to exchange longer length IPv6 Route Target prefixes.

Status of this Memo

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

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

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

This Internet-Draft will expire on September 10, 2011.

Copyright Notice

Table of Contents

1. Introduction

The current constrained route distribution specification defined in [RFC4684] supports prefixes with a fixed maximum length of 12 bytes. The prefix length needs to be extended to support the IPv6 specific Route Target extended community defined in [RFC5701] which is 20 bytes in length.

This document defines an extension to the current constrained route distribution specification that allows BGP speakers to distribute longer length Route Target prefixes. A new BGP capability known as BGP IPv6 Constrained Route Target capabiltiy is defined as part of extension that allows an exchange of longer length Route Target prefixes. BGP speakers that do not exchange this capability MUST use Route Target NLRIs of maximum length of 12 bytes. In this way, the current extension would preserve the backward compatibility with [RFC4684].

1.1. Requirements Language

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

2. BGP IPV6 Constrained Route Target Capability

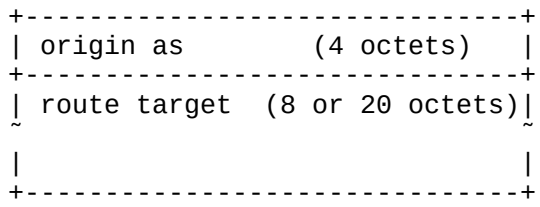
The "BGP IPV6 Constrained Route Target Capability" is a new BGP capability [RFC5492]. The Capability code for this capability is specified in the IANA Considersations section of this document. The Capability length field of this capability is zero.

By advertising this capability to a peer, a BGP speaker conveys to the peer that the speaker support the longer length Route Target prefixes and the related procedures described in this document.

3. IPV6 Constrained Route Target NLRI Advertisements

Route Target membership NLRI is advertised in BGP UPDATE messages using the MP_REACH_NLRI and MP_UNREACH_NLRI attributes as defined in [RFC4760]. The NLRI field in the MP_REACH_NLRI and MP_UNREACH_NLRI is a prefix of 0 to 24 octets, encoded as defined in Section 4 of [5] for all the constrain route distribution.

This prefix is structured as follows:



Except for the default route target, which is encoded as a zero-length prefix, the minimum prefix length is 32 bits.

Route targets can then be expressed as prefixes, where, for instance, a prefix would encompass all route target extended communities assigned by a given Global Administrator [6]. Alternatively, route target prefixes could be aggregated however if done so, then only the Local Administrator field of the Route Target can be aggregated. Route Target Type and the Global Administrator Route Target fields MUST not be aggregated.

The default route target can be used to indicate to a peer the willingness to receive all VPN route advertisements such as, for instance, the case of a route reflector speaking to one of its PE router clients.

4. Acknowledgements

The authors would like to thank Pedro Marques, John Scudder, Alton Lo and Zhengqiang Li for discussions and review.

5. IANA Considerations

This document defined the IPV6 Constrained Route Target Capability for BGP. The Capability code needs to be assigned by the IANA.

6. Security Considerations

This extension to [RFC4684] does not change the underlying security issues inherent in the existing BGP and [RFC4684].

7. References

7.1. Normative References

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

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

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, November 2006.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.

[RFC5701] Rekhter, Y., "IPv6 Address Specific BGP Extended Community Attribute", RFC 5701, November 2009.

7.2. Informative References

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

Authors' Addresses

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

Email: keyupate@cisco.com

Robert Raszuk
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: raszuk@cisco.com

Martine Djernaes
Juniper Networks
1194 N. Mathilda Avenue
Sunnyvale, CA 94089
USA

Email: mdjernaes@juniper.net

Jie Dong
Huawei Technologies Co., Ltd.
KuiKe Building, No.9 Xinxi Rd.
Hai-Dian District, Beijing 100085
P.R. China

Email: dongjie_dj@huawei.com

Mach(Guoyi) Chen
Huawei Technologies Co., Ltd.
KuiKe Building, No.9 Xinxi Rd.
Hai-Dian District, Beijing 100085
P.R. China

Email: mach@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 3, 2011

P. Mohapatra
A. Sreekantiah
K. Patel
A. Lo
Cisco Systems

March 02, 2011

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

Abstract

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

Status of this Memo

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

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

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

This Internet-Draft will expire on September 3, 2011.

Copyright Notice

Table of Contents

1. Introduction

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

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

1.1. Requirements Language

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

2. Basic Idea

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

3. Detailed Operation

3.1. Legacy PE Behavior

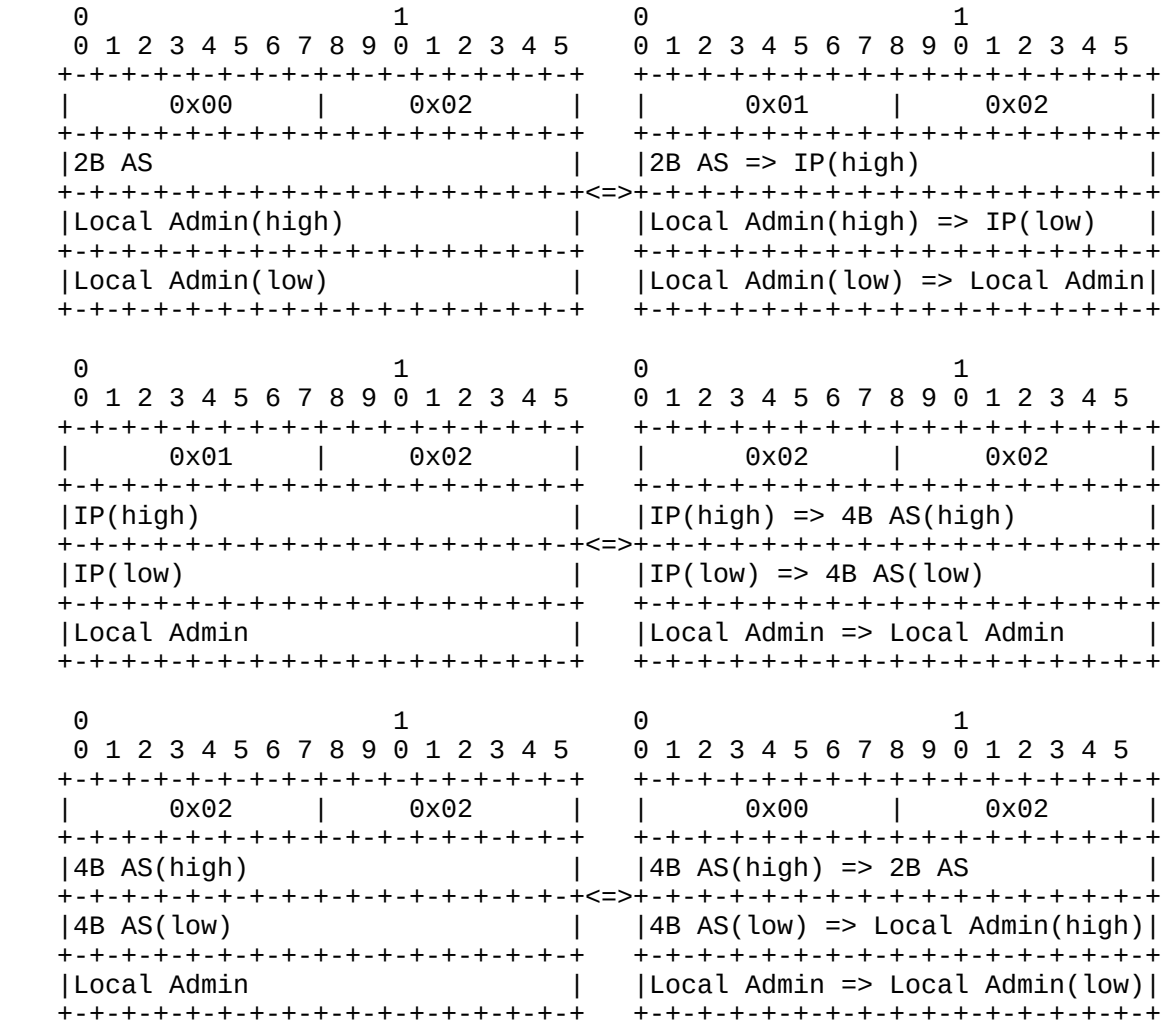
The following simple steps are performed on the legacy PE device:

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

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

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

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



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

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

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

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

3.2. RR behavior

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

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

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

4. ROUTE_FILTER community

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

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

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

5. Deployment Considerations

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

6. Contributors

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

7. Acknowledgements

8. IANA Considerations

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

9. Security Considerations

None.

10. Normative References

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

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

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

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

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

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

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

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

Authors' Addresses

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

Email: pmohapat@cisco.com

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

Email: asreekan@cisco.com

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

Email: keyupate@cisco.com

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

Email: altonlo@cisco.com

IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 12, 2011

R. Raszuk
E. Chen
Cisco Systems
B. Decraene
France Telecom

March 11, 2011

BGP Diagnostic Message
draft-raszuk-bgp-diagnostic-message-02

Abstract

BGP protocol lacks self diagnostic tools which would allow for monitoring and detection of any possible bgp state database differences between BGP_RIB_Out of the sender and BGP_RIB_In of the receiver over BGP peering session. It also lacks of build in mechanism to inform peer about subset of prefixes received over session which experienced some errors and which per protocol specification either resulted in attribute drop or "treat-as-withdraw" action.

The intention of this document is to start a new class of work which will make BGP protocol and therefor assuring services constructed with the help of BGP protocol to become much more reliable and robust.

Status of this Memo

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

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

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

This Internet-Draft will expire on September 12, 2011.

Copyright Notice

Table of Contents

1. Introduction

In this document we will first define a new diagnostic communication channel in the form of new BGP message then construct the set of basic message encoding to be used for simple diagnostic self test routines periodically exchanged between BGP speakers. We will also define set of other TLVs which can be very useful in precise description of prefixes affected by various cases of BGP session malfunctions.

The goal of this document is to provide the background which will in turn allow for very easy extensibility once new needs and new BGP diagnostic ideas surface.

2. Applications

Authors would like to propose four main applications which BGP Diagnostic 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 organized in the following application groups:

General TLVs used for operational purposes of the described mechanism.

Set of TLVs designed to carry information about BGP state across BGP peers that include per neighbor counters and global counters. There are two modes this functionality can be used - on demand by explicit query as well as periodic in an automated mode. The scope of messages is to be able to operate both on the iBGP as well as eBGP boundaries. It is in the control of the operator to decide which set of information would be send to a given set of peers.

Messages which operate in an automated push mode (as long as peer negotiated listen capability for them) and are designed to inform BGP peer on the list of impacted NLRIs which were received along with malformed attribute or within malformed update message.

Following recommendation from MP-BGP4 RFC4760 next group of messages are used to indicate which AFI/SAFIs were disabled for any further processing by BGP peer due to detection of an incorrect attribute present in the BGP Update message.

In number of troubleshooting efforts in real networks it is often very helpful to verify state of a given prefix in the neighboring

router's BGP database. This is particularly useful on the EBGP boundaries where there is no CLI/SNMP access to the router. Authors define a new way of query peer's BGP for the state of particular prefix.

Last set of messages is an attempt to allow for intra-domain better analysis of the BGP best path selection tie break decisions.

3. BGP diagnostic message

When defining any self test tool the critical element is to find a right separation balance between the test object and testing instruments.

For the vast majority of real BGP issues found in the life production networks authors believe that the right balance is the definition of new BGP message which could be exchanged along with any negotiated AFI/SAFI between those BGP speakers which will during initial OPEN message exchange new BGP diagnostic message capability.

The two extreme alternatives which were considered were the definition of new BGP attribute which may inherit and share potential issues of given BGP address family it is designed to diagnose and on the other extreme to build a separate and independent network diagnostic protocol. The use of BGP message seems to provide sufficient isolation from any service address family and is much easier to deploy then enabling an entire new intra and inter-domain protocol. Another very important issue with using any other protocol for detection of potential differences of BGP databases state is lack of synchronization with BGP UPDATE messages. This alone in the continuously churning BGP environment would not allow for any benefit.

3.1. BGP DIAGNOSTIC Message Encoding

BGP message as defined in RFC 4271 consists of a fixed-size header followed by two octet length field and one octet of type value. RFC 4271 limits maximum message size to 4096 octets. As one of the applications of BGP Diagnostic message is to be able to carry entire potentially malformed BGP message this specification extends the maximum size of BGP Diagnostic message to be always 128 octets bigger then any other BGP Message. Considering the current RFC 4271 maximum BGP message size to be 4096 octets maximum size of BGP diagnostic message would be 4224 octets.

For the purpose of diagnostic message information encoding we will

use one or more Type-Length-Value containers where each TLV will have the following format:

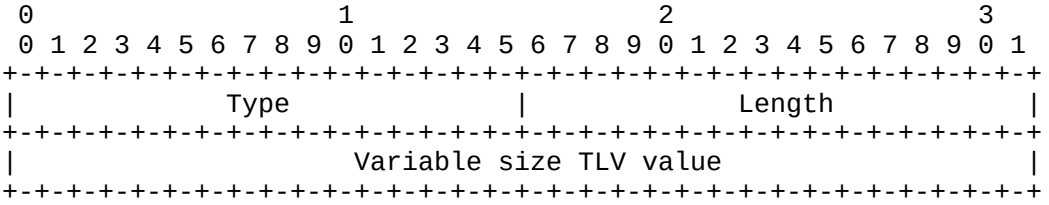


Figure 1: DIAGNOSTIC message TLV Format

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 issue some types of TLVs will need to contain a sequence number to correlate request with associated to it replies. The sequence number will consist of 8 octets and will be of form: 4 octet bgp_router_id + local 4 octet number. When local 4 octet number reaches 0xFFFF it should restart from 0x0000.

Typical application scenario for use of sequence number is to include it in the diagnostic request message and during reply to copy it into reply messages triggered by such request message.

3.2. BGP DIAGNOSTIC Message TLVs

This document defines the following diagnostic TLV types:

* Operational TLVs

* BGP database counters exchange

* Diagnostics for encoding errors in BGP messages

* AFI/SAFI signaling when malformed update

* Prefix specific BGP debugging

* Intra-domain bgp decision monitoring

* Exchange of Route Target filters

* Errors and warnings detected when validating BGP paths and prefixes

3.2.1. Operational TLVs

Type 1 - Diagnostic Message Periodic Request
Length - 2 octets - variable value

Value (N x 2 octets):
TLV type - 2 octets

Use: To indicate the request to periodically receive listed TLV information. TLV type of 0xFFFF indicates request to receive all available diagnostic TLVs from the peer.

Type 2 - Max frequency permitted
Length - 2 octets - variable value

Value (N x 4 octets):
TLV type - 2 octets
Frequency value in seconds two octets 0..65535

Special values:
0 - never send given diagnostic TLV
65535 - no TLV inter-gap minimum set

Use: To indicate in seconds the maximum frequency given TLV may be periodically sent to the bgp speaker

Type 3 - Diagnostic Message Query
Length - 2 octets - variable value
Sequence number - 8 octets

Value (N x 2 octets):
TLV type - 2 octets

Use: To interactively (during debugging/troubleshooting) request to receive listed TLV information. TLV type of 0xFFFF indicates request to receive all available diagnostic TLVs from the peer. TLV of type 0x0000 indicates request to receive a list of all enabled and available diagnostic TLV types from the peer towards querying BGP speaker. The support of this TLV type is mandatory.

Type 4 - Counter’s reset request
Length - 2 octets - variable value

Value (N x 2 octets):
TLV type - 2 octets - List of TLVs subject to counter’s reset.

Use: To request rest of per neighbor counters of a given TLV type. TLV type of 0xFFFF indicates request to zero all per neighbor counters.

```
   Type 5 - Not supported TLV reply
   Length - 2 octets - variable value

     Value (N x 3 octets):
       TLV type - 2 octets - TLV that is not supported by the peer
                  but where part of TLV Request or TLV Query message
       Error Code - 1 octet - Error code

           Error codes:

           0x01 - Wrong TLV value
           0x02 - TLV not supported for this peer
           0x03 - Max query frequency exceeded
           0x04 - Administratively disabled

     Use: To indicate to the peer that the TLV he has requeste
          either in TLV Request or in TLV Query message is not
          supported. The support of this TLV type is mandatory.


   Type 6 - Enabled and supported TLV types
   Length - 2 octets - variable value

     Value (N x 2 octets):
       TLV type - 2 octets - TLV that is enabled and supported
                             by the peer

     Use: To indicate to the peer that the enclosed list of TLVs
          can be requested either in TLV Request or in TLV Query
          messages. The support of this TLV type is mandatory.
```

3.2.2. BGP database counters exchange

Type 7 - Number of Reachable Prefixes Transmitted/Received
Length - 2 octets - variable value
Sequence number - 8 octets

Value (N x 11 octets):
AFI/SAFI - 3 octets
Number of prefixes transmitted - 4 octets
Number of prefixes received - 4 octets

Use: To indicate number of reachable prefixes exchanged for a given AFI/SAFI between two bgp speakers. This message can be sent only based on the remote query Type 3 which contains the query sequence number to be placed in the reply.

Type 8 - Number of prefixes in BGP_RIB_Out
Length - 2 octets - variable value

Value (N x 7 octets):
AFI/SAFI - 3 octets
Number of prefixes 4 octets

Use: To indicate number of prefixes kept in BGP_RIB_Out between bgp speakers for a given AFI/SAFI between two bgp speakers.

Type 9 - Number of paths in BGP_RIB_Out
Length - 2 octets - variable value

Value (N x 6 octets):
AFI/SAFI - 3 octets
Number of paths 4 octets

Use: To indicate number of paths kept in BGP_RIB_Out between bgp speakers for a given AFI/SAFI between two bgp speakers.

Type 10 - Number of prefixes present in BGP_RIB
Length - 2 octets - variable value

Value (N x 6 octets):
AFI/SAFI - 3 octets
Number of prefixes 4 octets

Use: To indicate number of prefixes kept in BGP RIB for a given AFI/SAFI.

Type 11 - Number of paths present in BGP_RIB
Length - 2 octets - variable value

Value (N x 7 octets):
AFI/SAFI - 3 octets
Number of prefixes 4 octets

Use: To indicate number of paths kept in BGP RIB for a given AFI/SAFI.

3.2.3. Diagnostics for encoding errors in BGP messages

Type 12 - Reachable prefixes present in dropped attribute UPDATE msg
Length - 2 octets - variable value

Value (N octets):
AFI/SAFI - 3 octets
1 .. M - List of prefixes

Use: To list reachable prefixes present in the update message where optional transitive attribute with partial bit set was malformed and has been removed from the update message. Prefix encoding should follow given AFI/SAFI definition.

Type 13 - Unreachable prefixes present in dropped attribute UPDATE msg
Length - 2 octets - variable value

Value (N octets):
AFI/SAFI - 3 octets
1 .. M - List of prefixes

Use: To list unreachable prefixes present in the update message where optional transitive attribute with partial bit set was malformed and has been removed from the update message. Prefix encoding should follow given AFI/SAFI definition.

Type 14 - Reachable prefixes present in malformed UPDATE msg
Length - 2 octets - variable value

Value (N octets):
AFI/SAFI - 3 octets
1 .. M - List of prefixes

Use: To list reachable prefixes present in the malformed update message which were subject to "treat-as-withdraw" behaviour. Prefix encoding should follow given AFI/SAFI definition.

Type 15 - Entire malformed update message enclosure
Length - 2 octets - variable value
Sequence number - 8 octets

Value:
Malformed message

Use: Propagate the malformed message to the peer upon it’s request or at the event of error detection. That includes propagation of messages which had malformed attribute, unparsable content or any other abnormal encoding. If more then a single message has been determined as malformed the subsequent replies will contain the same sequence number and should not be treated as an override.

3.2.4. AFI/SAFI signaling when malformed update

Type 16 - List of ignored AFI/SAFIs by the peer over given session
Length - 2 octets - variable value

Value (N octets):
1..M AFI/SAFI - 3 octets each

Use: To list those AFI/SAFIs which were detected to be malformed by the peer and while session is up were transitioned to IGNORE state.

Such case is inline with Multiprotocol Extensions RFC 4760 as per it's section 7 Error Handling:

"For the duration of the BGP session over which the UPDATE message was received, the speaker then SHOULD ignore all the subsequent routes with that AFI/SAFI received over that session".

3.2.5. Prefix specific BGP debugging

Type 17 - Prefix specific BGP query
Length - 2 octets - variable value

Value (N octets):
AFI/SAFI - 3 octets
Prefix under query
Prefix mask (optional)

Use: To query peer for the status of prefix under examination. When prefix mask is present the request is for exact match. When prefix mask is not present the request is for the longest match. Prefix encoding should follow given AFI/SAFI definition.

```
   Type 18 - Prefix specific BGP response
   Length - 2 octets - variable value

     Value (N octets):
       AFI/SAFI - 3 octets
       Prefix under query
           Prefix mask (optional)
           Prefix status (1 octet)

         Status:
           0x01 - prefix not found in BGP table
           0x02 - prefix in BGP table and active (in FIB)
           0x03 - prefix in BGP table and not-active (not in FIB)
                   0x04 - administratively disabled

     Use: To inform peer querying about the status of particular
          prefix status. Prefix encoding should follow given AFI/SAFI
          definition.


   Type 19 - BGP attribute based prefix query
   Length - 2 octets - variable value

     Value (N octets):
       AFI/SAFI - 3 octets
           Query Parameters - 1 octet
       BGP Attribute TLV

       Defined Query Parameters:
         Bit 0 - value 0 - Exact match
         Bit 0 - value 1 - Partial match

     Use: To query peer for the list of prefixes which paths contain
          given BGP attribute. BGP attribute encoding should follow
              given attribute's specification.
```

```
   Type 20 - BGP attribute based prefix reply
   Length - 2 octets - variable value

     Value (N octets):
       AFI/SAFI - 3 octets
           Query Parameters - 1 octet
       1 .. M - List of prefixes

            Defined Query Parameters:
            Bit 0 - value 0  - Exact match
            Bit 0 - value 1  - Partial match

     Use: To inform bgp peer about presence of set of prefixes
          which contain with exact or partial match the BGP
          Attribute as specified in the query. Prefix encoding
          should follow given AFI/SAFI definition.
```

3.2.6. Intra-domain bgp decision monitoring

```
   Type 21 - Number of IGP metric best path tie breaks executed
   Length - 2 octets - variable value

     Value (N x 7 octets):
       AFI/SAFI - 3 octets
       Number of tie breaks 4 octets

     Use: To indicate number of prefixes with their best path selected
          by tie break of IGP metric to their BGP next hop distance
          step of BGP best path selection algorithm.


   Type 22 - Number of BGP best path tie breaks in each selection step
   Length - 2 octets - variable value

     Value (N x 7 octets):
       AFI/SAFI - 3 octets
           Best path selection step N - Number of tie breaks 4 octets

     Use: To indicate number of cases where in BGP best path selection
          algorithm given step has been used as a tie break during
          overall best path selection process for a given prefix.
```

3.2.7. Exchange of installed Route Target filters

```
Type 23 - Request for reception of route target filters
          installed towards given peer by RFC4684
Length - 2 octets - variable value
Sequence number - 8 octets

  Value (N x 7 octets):
    AFI/SAFI - 3 octets
    BGP Router ID of the peer - 4 octets

  Use: To request reception of full table of route target
       fiters installed towards listed BGP peer for a requested
       AFI/SAFI. Single request may contain multiple pairs of
       AFI/SAFIs and/or BGP Router IDs.


Type 24 - Reply containing all route target filters installed
          towards given peer
Length - 2 octets - variable value
Sequence number - 8 octets

  Value (7 + N * 12 or 24 octets):
    AFI/SAFI - 3 octets
    BGP Router ID of the peer - 4 octets
        List of route targets - each 12 or 24 octets

  Use: Allows for troubleshooting purposes to share list of
       route targets installed for a given AFI/SAFI towards
       indicated BGP peer. In the event that RT filtering
       table size will not fit in single BGP Diagnostic
       Message reply the subsequent reply should include
       the same sequence number.
```

4. Operation

BGP implementation which supports DIAGNOSTIC message can support all or subset of defined diagnostic types. The range of supported TLV types will be signaled in the new BGP capability message during BGP connection establishment phase.

The operation of this extension can be realized on a pool/query based or push based principles. An implementation may provide, a timer to periodically send selected Diagnostic types TLVs to the peer or to the management station.

Similarly BGP peer may periodically or by manual cli request the reception of selected or all of the defined diagnostic TLV types.

The received values are then compared against local counters. When discrepancy is found operator is alarmed and further analysis should follow. The repair actions is out of scope of this document.

Example:

Under some situations when determined that the discrepancy is detected an automated or manual Route Refresh message can be triggered with it’s extension for Start_of_Refresh and End_of_Refresh markers . That would allow for purge of any stalled data across two BGP databases.

An important point which needs to be discussed is the exchange of counter’s values in light of continued BGP churn presence. As BGP is never stable it is expected that any sort of described counters will also be subject to continues value change making any comparison of their values questionable.

There are three classes of counters defined in this document: sent counters, received counters and current table state counters.

Only "sent" counters can be used for not correlated comparison and problem detection between any two BGP speakers. They are not subject to BGP churn issue due to the fact that DIAGNOSTIC messages would be exchanged inline with BGP UPDATE messages on a given session. An implementation must be able to freeze the received counters when comparing or displaying the received "sent" counters from BGP peer.

Received counters send in the Diagnostic messages are only meaningful in the context of explicit request trigger situation generated by the BGP speaker. BGP speaker should stop transmitting any BGP message of a given AFI/SAFI or freeze corresponding counter after sending diagnostic message request to the peer and before reception of actual diagnostic message reply. In order to correlate diagnostic message requests with associated replies use of build in sequence numbers is provided.

Table state counters (for example number of BGP RIB entries) are exchanged only for informational reasons and they should not be subject to comparison with any local counter values.

5. Capability negotiation

A BGP speaker that is willing to send or receive the BGP DIAGNOSTIC

Messages from its peer should advertise the new DIAGNOSTIC Messages Capability to the peer using BGP Capabilities advertisement [BGP-CAP]. A BGP speaker may send a DIAGNOSTIC message to its peer only if it has received the DIAGNOSTIC message capability from its peer.

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. The Capability Value field consists of reserved flags field.

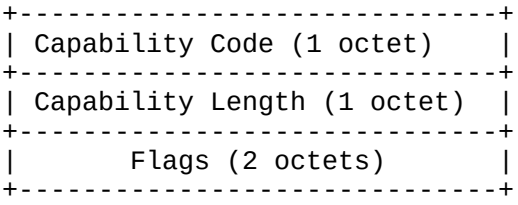


Figure 2: DIAGNOSTIC message BGP Capability Format

6. Security considerations

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

7. IANA Considerations

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

This document requests IANA to define and maintain a new registry named: "DIAGNOSTIC Message Type Values". The reserved types are: 0x0000 0xFFFF. The allocation policy is on a first come first served basis.

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

Type 1 - Diagnostic Message TLV(s) Request
Type 2 - Max frequency permitted
Type 3 - Diagnostic Message TLV(s) Query
Type 4 - Counter's reset request
Type 5 - Not supported TLV
Type 6 - Enabled and supported TLV types

Type 7 - Number of Reachable Prefixes Transmitted/Received
Type 8 - Number of prefixes in BGP_RIB_Out
Type 9 - Number of paths in BGP_RIB_Out
Type 10 - Number of prefixes present in BGP_RIB
Type 11 - Number of paths present in BGP_RIB

Type 12 - Reachable prefixes present in dropped attribute message
Type 13 - Unreachable prefixes present in dropped attribute message
Type 14 - Reachable prefixes present in malformed UPDATE message
Type 15 - Entire malformed update message enclosure

Type 16 - List of ignored AFI/SAFIs by the peer over given session

Type 17 - Prefix specific BGP query
Type 18 - Prefix specific BGP response
Type 19 - BGP attribute based prefix query
Type 20 - BGP attribute based prefix reply

Type 21 - Number of IGP metric best path tie breaks executed
Type 22 - Number of BGP best path tie breaks in each selection step

Type 23 - Request for reception of route target filters
Type 24 - Reply containing all route target filters installed

Type 25 - 65534 Free for future allocation.
Type 65535 - Reserved

8. Acknowledgments

Authors would like to thank Alton Lo for his valuable input.

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.

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

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.

9.2. Informative References

[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

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

Email: raszuk@cisco.com

Enke Chen
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US

Email: enkechen@cisco.com

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

Email: bruno.decraene@orange-ftgroup.com

IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 12, 2011

R. Raszuk
C. Cassar
Cisco Systems
E. Aman
TeliaSonera
B. Decraene
France Telecom
March 11, 2011

BGP Optimal Route Reflection (BGP-ORR)
draft-raszuk-bgp-optimal-route-reflection-01

Abstract

[RFC4456] asserts that, because the Interior Gateway Protocol (IGP) cost to a given point in the network will vary across routers, "the route reflection approach may not yield the same route selection result as that of the full IBGP mesh approach." One practical implication of this assertion is that the deployment of route reflection may thwart the ability to achieve hot potato routing. Hot potato routing attempts to direct traffic to the closest AS egress point in cases where no higher priority policy dictates otherwise. As a consequence of the route reflection method, the choice of exit point for a route reflector and its clients will be the egress point closest to the route reflector - and not necessarily closest to the RR clients.

Section 11 of [RFC4456] describes a deployment approach and a set of constraints which, if satsified, would result in the deployment of route reflection yielding the same results as the iBGP full mesh approach. Such a deployment approach would make route reflection compatible with the application of hot potato routing policy.

As networks evolved to accommodate architectural requirements of new services, tunneled (LSP/IP tunneling) networks with centralized route reflectors became commonplace. This is one type of common deployment where it would be impractical to satisfy the constraints described in Section 11 of [RFC4456]. Yet, in such an environment, hot potato routing policy remains desirable.

This document proposes two new solutions which can be deployed to facilitate the application of closest exit point policy centralized route reflection deployments.

Status of this Memo

This Internet-Draft is submitted in full conformance with the

provisions of BCP 78 and BCP 79.

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

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

This Internet-Draft will expire on September 12, 2011.

Table of Contents

1. Introduction

There are three types of BGP deployments within Autonomous Systems today: full mesh, confederations and route reflection.

BGP route reflection is the most popular way to distribute BGP routes between BGP speakers belonging to the same administrative domain. Traditionally route reflectors have been deployed in the forwarding path and carefully placed on the POP to core boundaries. That model of BGP route reflector placement has started to evolve. The placement of route reflectors outside the forwarding path was triggered by applications which required traffic to be tunneled from AS ingress PE to egress PE: for example L3VPN.

This evolving model of intra-domain network design has enabled deployments of centralized route reflectors. Initially this model was only employed for new address families e.g. L3VPNs, L2VPNs etc

With edge to edge MPLS or IP encapsulation also being used to carry internet traffic, this model has been gradually extended to other BGP address families including IPv4 and IPv6 Internet routing. This is also applicable to new services achieved with BGP as control plane for example 6PE.

Such centralized route reflectors can be placed on the POP to core boundaries, but they are often placed in arbitrary locations in the core of large networks.

Such deployments suffer from a critical drawback in the context of best path selection. A route reflector with knowledge of multiple paths for a given prefix will pick the best path and only advertise that best path to the the route reflector clients. If the best path for a prefix is selected on the basis of an IGP tie break, the best path advertised from the route reflector to its clients will be the exit point closest to the route reflector. But route reflector clients will be in a place in the network toplogy which is different from the route reflector. In networks with centralized route reflectors, this difference will be even more acute. It follows that the best path chosen by the route reflector is not necessarily the same as the path which would have been chosen by the client if the client considered the same set of candidate paths as the route reflector. Furthermore, the path chosen by the client might have been a better path from that chosen by the route reflector for traffic entering the network at the client. The path chosen by the client would have guaranteed the lowest cost and delay trajectory through the network.

Route reflector clients switch packets using routing information

learnt from route reflectors which are not on the forwarding path of the packet through the network even in the absence of end-to-end encapsulation. In those cases the path chosen as best and propagated to the clients will often not be the optimal path chosen by the client given all available paths.

Eliminating the IGP distance to the BGP nexthop as a tie breaker on centralized route reflectors does not address the issue. Ignoring IGP distance to the BGP next hop results in the tie breaking procedure contributing the best path by differentiating between paths using attributes otherwise considered less important than IGP cost to the BGP nexthop.

One possible valid solution or workaround to this problem requires sending all domain external paths from the RR to all its clients. This approach suffers the significant drawback of pushing a large amount of BGP state to all the edge routers. In many networks, the number of EBGP peers over which full Internet routing information is received would correlate directly to the number of paths present in each ASBR. This could easily result in tens of paths for each prefix.

Notwithstanding this drawback, there are a number of reasons for sending more than just the single best path to the clients. Improved path diversity at the edge is a requirement for fast connectivity restoration, and a requirement for effective BGP level load balancing. Protocol extensions like add-paths [I-D.ietf-idr-add-paths] or diverse-path [I-D.ietf-grow-diverse-bgp-path-dist] allow for such improved path diversity and can be used to address the same problems addressed by the mechanisms proposed in this draft. In practical terms, add/ diverse path deployments are expected to result in the distribution of 2, 3 or n (where n is a small number) 'good' paths rather than all domain external paths. While the route reflector chooses one set of n paths and distributes those same n paths to all its route reflector clients, those n paths may not be the right n paths for all clients. In the context of the problem described above, those n paths will not necessarily include the closest egress point out of the network for each route reflector client. The mechanisms proposed in this document are likely to be complementary to mechanisms aimed at improving path diversity.

2. Proposed solutions

This document proposes two simple solutions to the problem described above. Both of these solutions make it possible for route reflector clients to direct traffic to their closest exit point in hot potato

routing deployments, without requiring further state to be pushed out to the edge. These solutions are primarily applicable in deployments using centralized route reflectors, which are typically implemented in devices without a capable forwarding plane.

The two alternatives are:

"Best path selection for BGP hot potato routing from client's IGP network position"

"Angular distance approximation for BGP warm potato routing"

Both solutions rely upon all route reflectors learning all paths which are eligible for consideration for hot potato routing. In order to satisfy this requirement, path diversity enhancing mechanisms such as add paths/diverse paths may need to be deployed between route reflectors.

In both of these solutions the route reflector selects and distributes a route to each client based on what would be optimal from the client's perspective. In the respective solutions the choice is made either factoring in IGP costs or the configured angular distance to the next hop. The route reflector makes different decisions for different clients only in the case where the tie breaker for path selection would have been the IGP distance to the BGP nexthop (as in hot potato routing).

A signficant advantage of this approach is that the RR clients do not need to run new software or hardware.

3. Best path selection for BGP hot potato routing from customized IGP network position

This section describes a method for calculating the order of preference of BGP paths from the point of view of each separate route reflector client. More specifically, the route relflector will compute the IGP metric to the BGP nexthop from the position of the client to which the resulting path will be distributed, if the IGP metric is the tie breaker applied to a set of possible paths. In the subsequent model authors will propose virtual reflector placement at operator's selected IGP location.

In the case of a hierarchical IGP deployment where the client is in a different level in the hierarchy to the route reflector, the route reflector will compute IGP distance to the BGP nexthop from the Area Border Routers (ABR) leading to the client in lieu of the route reflector client itself, and use the shortest distance from these

ABRs to the nexthop. This provides an approximation to the desired functionality. Rather than a client picking the closest path, the client would be picking the exit point closest to the client region as defined by area or level. In cases where one or more nexthops are in the same region as the client, one of those nexthops would be preferred, with tie breaking within those nexthops performed from the route reflector's position in the network.

It is assumed that reachability through a set of ABRs is always advertised through identical prefixes from those ABRs. If a nexthop is reachable through multiple ABRs but the ABRs advertise reachability through prefixes of different length, then only the ABR advertising the longest prefix will be considered as a viable path to the nexthop.

BGP best path selection and its distribution has a natural consequence of limiting the amount of state in the network. That is not in itself a drawback. BGP speakers will rarely need to receive all available BGP paths. In network deployments with multiple upstream peerings or with very dense peering schemes, the number of available BGP paths for a given BGP prefix can be high. Real network deployments with the number of paths for a prefix ranging from 10s to 100s have been observed. It would be wasteful to propagate all of those paths to all clients, such that each client can select paths according to the position of the nexthop relative to the client.

Whenever a BGP route reflector would need to decide what path or paths need to be selected for advertisement to one of its clients, the route reflector would need to virtually position itself in its client IGP network location in order to choose the right set of paths based on the IGP metric to the next hops from the client's perspective.

This technique applies in deployments with or without diverse paths or the various path selection modes contemplated in add-paths.

3.1. Client's perspective best path selection algorithm

For each centralized route reflector the proposal assumes that the route reflector participates in a common IGP with its clients. There are two scenarios to consider - flat versus hierarchical IGP network.

3.1.1. Flat IGP network

Reflectors run SPF from the client IGP node point of view such that the cost of BGP nexthops from the client can be determined if necessary. For the purpose of BGP path selection the interesting product of this calculation is the ability to determine the IGP

distance from a client to a BGP next hop. This distance to a nexthop would be interesting in cases where that next hop is for a path which is contending with otherwise equally preferred paths. This approach works in tunneled as well as conventional hop-by-hop IP forwarding cores.

When the path selection tie breaker for a prefix is the IGP metric to the BGP nexthops of the contending paths, then the route reflector will determine the order of preference of the contending paths by considering the distance from the client to the path nexthops in order to decide what path/s to advertise to a client (or group of clients where feasible). It should be noted that an operator may wish to provide a distance tolerance value, such that beyond a certain granularity, differences between IGP metric are invisible to the path selection algorithm. This will allow a route reflector some leeway in selecting between paths such that rather than pick one path over another on the basis of a difference in distance which is operationally irrelevant, the route reflector can choose to optimise for update generation grouping. Furthermore, this tolerance will reduce the likelihood of generation of BGP updates when the IGP topology changes in a way which is not operationally relevant. In the case that a path is selected from a set for a given prefix while ignoring differences in distance within the tolerance figure, then that same path must always be preferred for all clients where the paths are within the tolerance figure

3.1.2. Hierarchical IGP network

Hierarchy introduces two challenges:

The first challenge is that the RR IGP view may differ from a client IGP view by virtue of one or the other having a summarised view versus the other. Summarisation, by its nature, loses information. Consider the example where a client within a PoP sees two prefixes with two metrics for two egress points within the PoP, but where the RR only sees a single summary covering reachability to both nexthops as injected by the ABR. However it needs to be observed that inter area networks running LDP are required to disable summarization of all FEC advertised in LDP (typically all loopbacks) unless [RFC5283] is deployed. Such deployments are not likely to suffer summarisation difficulties.

The second challenge is that in cases where the client is in a different level of hierarchy from the RR, the RR can not build a Shortest Path First (SPF) tree with the client node as root, simply because the toplogy derived by the IGP will not include the client node. It will instead only include reachability to the

client from one or more ABRs. In order to overcome this problem, the RR could compute an SPF tree from the ABRs in the area. The RR would then determine the shortest distance from a client which lives behind the ABRs, to a nexthop, by adding the advertised distances from an ABR to the client and the distance from from the ABR to a nexthop, for each ABR, and picking the minimum. This assumes that IGP metrics on links are symmetric; i.e. that the distance from the ABR to the client or nexthop is equal to the distance from the client or nexthop to the ABR.

There are cases where the above approach does not help. If RR is trying to arbitrate amongst a set of paths for a client which is in the same hierarchy as some of those paths, and in a different hierarchy to the RR, the opaqueness of the region containing the client at the RR defeats the selection process. It is impossible to determine the relative position of the RR client and the paths within the client region.

The solution for hierarchical IGP networks also assumes that if RRs are present and are responsible for calculation of BGP best path to clients they are either placed in each local area coinciding with area containing clients or they are placed in the core (area 0/level 2) of the network.

3.2. Aside: Configuration-based flexible route reflector placement

The ability to exploit topology information available in the IGP in ways described above can also be used to virtually place the RR at different points in the network for purposes other than hot potato routing.

A route reflector can be globally configured to "pretend" its logical location is one of any of the other nodes within a given IGP area/ level flooding scope regardless of its physical connectivity.

Such flexibility provides a useful tool for reflector virtualization, and supports moving or replacing physical route reflectors without any effect on routing. Such a change can be permanent or it could be performed during network maintenance in order to minimize network impact.

A possible variation would allow the virtual placement of RR to be effected on a per-AF or AF plus update/peer group granularity. It should be noted that this approach provides for splitting one centralized route reflector such that it is virtually positioned at various network locations, with the network location depending upon of address family or address family plus update/peer group.

Virtual slicing of a centralized route reflector relaxes the need to propagate all BGP paths between RRs in a alternative conventional distributed RR deployment. It is expected that such RRs would be deployed in redundant sets, and that those RRs would not need to be physically colocated, while still benefiting from the possibility of being logically colocated, and therefore not compromising any of the best path selection symmetry.

3.3. Route reflector client grouping

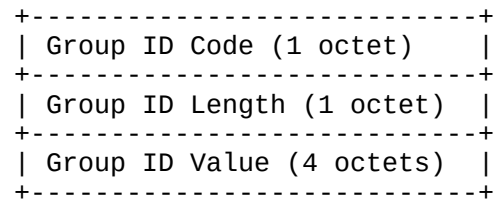
It may be appropriate to allow the operator, or the route reflector itself, to group clients together using IGP distance between clients to determine grouping. All the operation discussed above which relied upong computing best path for each client, and measuring distances from each client to different nexthops, would instead be performed for each group of clients. A configurable thresholds can be used to determine which IGP metric changes should be visible to BGP, and trigger best paths recomputation. The latter would be beneficial in existng BGP RR code too.

Alternatively route reflector client grouping could be accomplished statically by the operator by coloring clients belonging to a common group (for example being part of the same POP). In order to accomplish such marking it is proposed that BGP OPEN message be augmented with an optional paramater indicating the Group ID given peer belongs to.

3.3.1. Route Reflector Client Group ID

This is an Optional Parameter in BGP OPEN message that is used by a BGP speaker to convey to its route reflectors the Group ID value. Such value will allow automatic and predictable peer grouping on the route reflectors as deemed necessary from operator’s network architecture.

The parameter contains precisely one set of [Group_ID Code, Group_ID Length, Group_ID Value] encoded as shown below:



The use and meaning of these fields are as follows:

Group ID Code:

Group ID Code is a one octet field that identifies Group ID optional parameter of BGP OPEN message. Value TBD by IANA Recommended value: 3.

Group ID Length:

Group ID Length is a one octet field that contains the length of the Group ID Value field in octets. It is fixed and equals to 4.

Group ID Value:

Group ID Value is a fixed length field of size equal to four octets that contains the numerical value of group given BGP speaker should be part of on the route reflector.

Two special values are reserved:

0x00000000 - No grouping preference
0xFFFFFFFF - Do not group this BGP speaker

An implementation may allow automatic population of GROUP_ID value using IGP area identifier.

Route reflectors or EBGP speakers receiving such Group IDs from their respective BGP peers as part of the BGP OPEN procedure MAY use them when constructing update or peer groups in addition to any of the existing grouping mechanism already available. An implementation may allow operator to explicitly allow or disallow honoring such grouping or provide means for manual overwrite via explicit configuration.

3.4. Discussion

This is not the first instance where a router participating in an IGP is required to build the SPF tree using a root other than itself. Determination of loop free alternate paths as described in [RFC5714] is one such example.

Determining the shortest path and associated cost between any two arbitrary points in a network based on the IGP topology learned by a router is expected to add some extra cost in terms of CPU resource. However SPF tree generation code is now implemented efficiently in a number of implementations, and therefor this is not expected to be a major drawback. The number of SPTs computed in the general non-hierarchical case is expected to be of the order of the number of clients of an RR whenever a topology change is detected. Advanced optimisations like partial and incremental SPF may also be exploited. By the nature of route reflection, the number of clients can be split arbitrarily by the deployment of more route reflectors for a given number of clients. While this is not expected to be necessary in existing networks with best in class route reflectors available today, this avenue to scaling up the route reflection infrastructure would be available. If we consider the overall network wide cost/benefit factor, the only alternative to achieve the same level of optimality would require significanly increasing state on the edges of the network, which, in turn, will consume CPU and memory resources on all BGP speakers in the network. Building this client perspective into the route reflectors seems appropriate.

3.5. Advantages

The solution described provides a model for integrating the client perspective into the best path computation for RRs. More specifically, the choice or BGP path factors in the IGP metric between the client and the nexthop, rather than the distance from the RR to the nexthop. The documented method does not require any BGP or IGP protocol changes as required changes are contained within the RR implementation.

This solution can be deployed in traditional hop-by-hop forwarding networks as well as in end-to-end tunneled environments. In the networks where there are multiple route reflectors and unencapsulated hop-by-hop forwarding, such optimisations should be enabled on all route reflectors. Otherwise clients may receive an inconsistent view of the network and in turn lead to intra-domain forwarding loops.

With this approach, an ISP can effect a hot potato routing policy even if route reflection has been moved from the forwarding plane to the core and hop-by-hop switching has been replaced by end to end

MPLS or IP encapsulation.

As per above, the approach reduces the amount of state which needs to be pushed to the edge in order to perform hot potato routing. The memory and CPU resource required at the edge to provide hot potato routing using this approach is lower than what would be required in order to achieve the same level of optimality by pushing and retaining all available paths (potentially 10s) per each prefix at the edge.

The proposal allows for a fast and safe transition to BGP control plane route reflection without compromising an operator's closest exit operational principle. Hot potato routing is important to most ISPs. The inability to perform hot potato routing effectively stops migrations to centralized route reflection and edge-to-edge LSP/IP encapsulation for traffic to IPv4 and IPv6 prefixes.

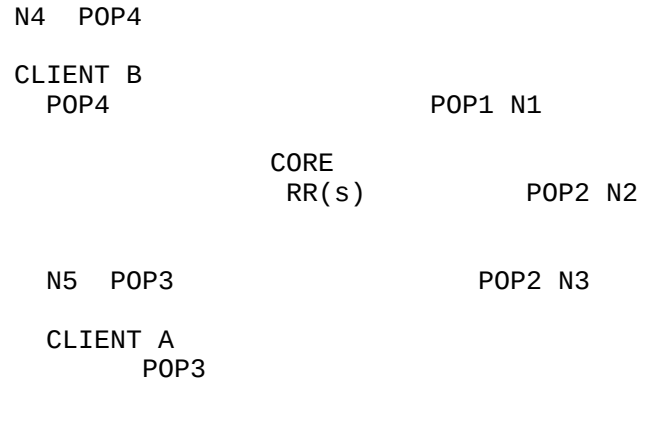
4. Angular distance approximation for BGP warm potato routing

This section describes an alternative solution to the use of IGP topology information to virtually position the RR at the client location in the network. This solution involves modelling the network topology as a set of elements (regions, PoPs or routers) arranged in a circle. Route reflector clients and inter-domain exit points would then be statically assigned to those elements such that one can compute the angular distance between route-reflector clients and the various exit points in order to infer the distance between any two elements. This measure of distance can be used as an effective alternative to the IGP distance as a tie breaker in the path selection algorithm if necessary.

4.1. Problem statement

This solution addresses the problem described in earlier sections, while attempting to minimise computational overhead. The aim of the proposed solution is to enable a route reflector to provide a route reflector client with an exit point for a prefix which is 'closest' to the client rather than the route-reflector, without having to distribute all paths to that client, or having to derive each client's view of the network topology. The measure of closest is based on a simplistic description of network topology provided by the operator.

Consider the following example of an ISP network topology drawn to reflect the location of the nodes and POPs:



N - represents the different exit points for a given prefix. POP2 is a geographically large PoP with two paths; N2 and N3.

In a deployment where the centralized RRs tie break on the basis of their IGP-based view of the network, N1 above would be advertised to all clients on the basis that it is closest to the RR. Path N4 would be a more appropriate choice for client B. Similarly, N5 would be more appropriate for client A since path N5 is closer to client A then path N1.

4.2. Proposed solution

The proposed solution revolves around the operator establishing the angular position of the route-reflector clients and inter-domain exit points in the network. The route reflector then picks the path to advertise to a client based on the client's angular position versus the angular position of the inter-domain exit points originating the paths. The operator can choose the granularity of angular position appropriate to the desired goals. On one hand, the coarseness of the angular position will effect the operator overhead; versus the optimality of routing on the other. The finest granularity possible will be the relative position of originating clients.

Note that this solution has nothing to do with actual IGP link metrics and resulting topology in the network.

It can be shown that for each network topology, elements such as AS exit points can be mapped on to a circle. By putting POPs, Regions or individual clients onto the hypothetical circle we can identify an angular location for each element relative to some fixed direction; for example defining the angular north of the circle at 0 degress.

The angular position of elements in the network can be conveyed to a route reflector in a number of ways:

Assignment of angular position of each RR client through configuration on the route reflector itself; per client configuration on RR

Assignment of angular position of an RR client at each client, then propagating it to RRs.

The proposed angular distance approximation is compatible with both flat and hierarchical IGP deployments.

In the example illustrated above the route reflector might learn or be configured with the following set of paths and corresponding angular positions:

Prefix X/Y	N1	N2	N3	N4	N5
Location in degrees	60	85	120	290	260

If the absolute angular position of clients A and B were as follows:

Client A: 260 degrees

Client B: 290 degrees

Then the corresponding angular distances for those clients versus the exit points can be calculated as follows:

Prefix X/Y	N1	N2	N3	N4	N5
Client A	200	175	140	30	0
Client B	230	205	170	0	30

With an RR running the BGP best path algorithm modified to use the angular distance from the client to the nexthops, rather than its IGP distance to the nexthops as tie breaker, each client is provided with its closest path with the measure of closeness reflecting the angular position as configured by the operator.

The model used by the operator in order to determine the angular position of a client or exit point, might involve grouping elements together by region or PoP, or might involve no grouping at all.

Implementations should allow the operator to pick the appropriate granularity.

4.3. Centralized vs distributed route reflectors

In an environment where the RR clusters are distributed (yet centralized enough to make hot potato routing hard), and each RR cluster serves a subset of clients, it becomes necessary to propagate the angular position of the clients between route reflectors. This can be achieved as follows:

Deploy add-paths between route reflectors in order to maximise path diversity within the cluster.

A non AS transitive BGP community of type (TBA by IANA) can be used to encode and propagate angular position between 0 and 359 of a client. This community is only relevant to the route reflectors of a given BGP domain and should be stripped either at the ASBR boundary or when propagating updates to BGP peers which are not route reflectors.

The angular position marking could also be added by clients and advertised to the route reflector. This would require some configuration effort.

5. Deployment considerations

The solutions are primarily intended for end-to-end tunneled environments, i.e. where traffic is label switched or IP tunneled across the core. If unencapsulated hop-by-hop forwarding is used, either misconfigurations or conflicts between these optimizations and classical BGP path selection rules could lead to intra-domain forwarding loops. Under certain circumstances the solutions can also be deployable without end-to-end tunneling. In particular the best path selection based on the client’s IGP best-path selection is guaranteed not to cause any forwarding loops (other than micro loops associated with reconvergence) when deployed in a flat IGP area provided that no distance tolerance value is used so that the path choice is truly made on a per-client basis.

It should be self evident that this solution does not interfere with policies enforced above IGP tie breaking in the BGP best path algorithm.

The solution applies to NLRIs of all address families which can be route reflected and which can be tie broken by IGP distance to the nexthop.

It should be noted that customized per-client or group of clients best path selection is already in use today in the context of Internet Exchange Point (IXP) route servers. In an IXP route server the client best path is selected as a result of different policies rather than IGP metric distance to BGP next hop.

A possible scalability impact of optimising path selection to take account of the RR client position is that different RR clients receive different paths, and therefore update/peer group efficiency diminishes. This cost is imposed by the requirement given the requirement is to optimise the egress path from the client's perspective. It is also not unlikely that groups of clients will end up receiving the same best path/s, in which case, inefficiency of update generation will be minimised. It should be noted that in the cases described under flexible router placement where placement is determined on a per update/peer group basis or per route reflector, the scale benefits of peer groupings are retained.

6. Security considerations

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

7. IANA Considerations

IANA is requested to allocate a type code for the Standard BGP Community to be used for inter cluster propagation of angular position of the clients.

IANA is requested to allocate a new type code from BGP OPEN Optional Parameter Types registry to be used for Group_ID propagation.

8. Acknowledgments

Authors would like to thank Eric Rosen, Clarence Filsfils and Mike Shand for their valuable input.

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.

[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.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.

9.2. Informative References

[I-D.ietf-grow-diverse-bgp-path-dist] Raszuk, R., Fernando, R., Patel, K., McPherson, D., and K. Kumaki, "Distribution of diverse BGP paths.", draft-ietf-grow-diverse-bgp-path-dist-03 (work in progress), January 2011.

[I-D.ietf-idr-add-paths] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", draft-ietf-idr-add-paths-04 (work in progress), August 2010.

[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.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.

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

[RFC5283] Decraene, B., Le Roux, JL., and I. Minei, "LDP Extension for Inter-Area Label Switched Paths (LSPs)", RFC 5283, July 2008.

[RFC5668] Rekhter, Y., Sangli, S., and D. Tappan, "4-Octet AS Specific BGP Extended Community", RFC 5668, October 2009.

[RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework",

RFC 5714, January 2010.

Authors' Addresses

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

Email: raszuk@cisco.com

Christian Cassar
Cisco Systems
10 New Square Park
Bedfont Lakes, FELTHAM TW14 8HA
UK

Email: ccassar@cisco.com

Erik Aman
TeliaSonera
Marbackagatan 11
Farsta, SE-123 86
Sweden

Email: erik.aman@teliasonera.com

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

Email: bruno.decraene@orange-ftgroup.com

IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 12, 2011

R. Raszuk
B. Pithawala
Cisco Systems
D. McPherson
Verisign, Inc.

March 11, 2011

Dissemination of Flow Specification Rules for IPv6
draft-raszuk-idr-flow-spec-v6-01

Abstract

Dissemination of Flow Specification Rules [RFC5575] provides a protocol extension for propagation of traffic flow information for the purpose of rate limiting or filtering. The [RFC5575] specifies those extensions for IPv4 protocol data packets.

This specification extends the current [RFC5575] and defines changes to the original document in order to make it also usable and applicable to IPv6 data packets.

Status of this Memo

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

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

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

This Internet-Draft will expire on September 12, 2011.

Copyright Notice

Table of Contents

1. Introduction

The growing amount of IPv6 traffic in private and public networks requires the extension of tools used in the IPv4 only networks to be also capable of supporting IPv6 data packets.

In this document authors analyze the differences of IPv6 [RFC2460] flows description from those of traditional IPv4 packets and propose subset of new encoding formats to enable Dissemination of Flow Specification Rules [RFC5575] for IPv6.

This specification should be treated as an extension of base [RFC5575] specification and not its replacement. It only defines the delta changes required to support IPv6 while all other definitions and operation mechanisms of Dissemination of Flow Specification Rules will remain in the main specification and will not be repeated here.

2. IPv6 Flow Specification encoding in BGP

The [RFC5575] defines a new SAFIs (133 for IPv4) and (134 for VPNv4) applications in order to carry corresponding to each such application flow specification.

This document will redefine the [RFC5575] SAFIs in order to make them AFI specific and applicable to both IPv4 and IPv6 applications.

The following changes are defined:

> "SAFI 133 for IPv4 dissemination of flow specification rules" to now be defined as "SAFI 133 for IP dissemination of flow specification rules"
>
> "SAFI 134 for VPNv4 dissemination of flow specification rules" to now be defined as "SAFI 134 for L3VPN dissemination of flow specification rules"

For both SAFIs the indication to which address family they are referring to will be recognized by AFI value (AFI=1 for IPv4 or VPNv4, AFI=2 for IPv6 and VPNv6 respectively). Such modification is fully backwards compatible with existing implementation and production deployments.

It needs to be observed that such choice of proposed encoding is compatible with filter validation against routing reachability information as described in section 6 of RFC5575. Validation tables will now be performed according to the following rules.

Flow specification received over AFI/SAFI=1/133 will be validated against routing reachability received over AFI/SAFI=1/1

Flow specification received over AFI/SAFI=1/134 will be validated against routing reachability received over AFI/SAFI=1/128

Flow specification received over AFI/SAFI=2/133 will be validated against routing reachability received over AFI/SAFI=2/1

Flow specification received over AFI/SAFI=2/134 will be validated against routing reachability received over AFI/SAFI=2/128

3. IPv6 Flow Specification types changes

The following component types are redefined or added for the purpose of accommodating new IPv6 header encoding. Unless otherwise stated all other types as defined in RFC5575 apply to IPv6 packets as is.

Type 1 - Destination IPv6 Prefix

Encoding: <type (1 octet), prefix length (1 octet), prefix offset (1 octet), prefix>

Defines the destination prefix to match. Prefix offset has been defined to allow for flexible match on the part of the IPv6 address where we want to skip (don't care) of N first bits of the address. This can be especially useful where part of the IPv6 address consists of embedded IPv4 address and match needs to happen only on the part of embedded IPv4 address. The default value for prefix offset is 0x00 (match on all bits as indicated by prefix length). Otherwise prefixes are encoded as in BGP UPDATE messages, a length in bits is followed by enough octets to contain the prefix information.

Type 2 - Source IPv6 Prefix

Encoding: <type (1 octet), prefix length (1 octet), prefix offset (1 octet), prefix>

Defines the source prefix to match. Prefix offset has been defined to allow for flexible match on the part of the IPv6 address where we want to skip (don't care) of N first bits of the address. This can be especially useful where part of the IPv6 address consists of embedded IPv4 address and match needs to happen only on the part of embedded IPv4 address. The default value for prefix offset is 0x00 (match on all bits as indicated by prefix length). Otherwise prefixes are encoded as in BGP UPDATE

messages, a length in bits is followed by enough octets to contain the prefix information.

Type 3 - Next Header

Encoding: <type (1 octet), [op, value]+>

Contains a set of {operator, value} pairs that are used to match the last Next Header value octet in IPv6 packets. The operator byte is encoded as specified in component type 3 of [RFC5575].

While IPv6 allows for more then one Next Header field in the packet the main goal of Type 3 flow specification component is to match on the subsequent IP protocol value. Therefor the definition is limited to match only on last Next Header field in the packet.

Type 11 - Traffic Class

Encoding: <type (1 octet), [op, value]+>

Contains a set of {operator, value} pairs that are used to match the Traffic Class 8-bit field [RFC2460] encoded in a single octet.The operator byte is encoded as specified in component type 3 of [RFC5575].

Type 12 - Fragment - Removed

This type is removed for IPv6 flow specification as in IPv6 fragmentation does not happen in the network.

Type 13 - Flow Label - New type

Encoding: <type (1 octet), [op, value]+>

Contains a set of {operator, value} pairs that are used to match the 20-bit Flow Label field [RFC2460].The operator byte is encoded as specified in the component type 3 of [RFC5575].

4. IPv6 Flow Specification Traffic Filtering Action changes

One of the traffic filtering actions which can be expressed by BGP extended community is defined in [RFC5575] as traffic-marking. This extended community type is of value: 0x8009.

For the purpose of making it compatible with IPv6 header action expressed by presence of this extended community has been modified to

read:

Traffic Marking: The traffic marking extended community instructs a system to modify the Traffic Class bits of a transiting IPv6 packet to the corresponding value. This extended community is encoded as a sequence of 5 zero bytes followed by the 8 bit Traffic Class value encoded in the 6th byte.

5. Security considerations

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

6. IANA Considerations

IANA is requested to rename currently defined SAFI 133 and SAFI 134 per [RFC5575] to read:

133 Dissemination of flow specification rules
134 L3VPN dissemination of flow specification rules

IANA is requested to create and maintain a new registry entitled: "Flow Spec IPv6 Component Types". The following component types have been registered:

Type 1 - Destination IPv6 Prefix
Type 2 - Source IPv6 Prefix
Type 3 - Next Header
Type 4 - Port
Type 5 - Destination port
Type 6 - Source port
Type 7 - ICMP type
Type 8 - ICMP code
Type 9 - TCP flags
Type 10 - Packet length
Type 11 - Traffic Class
Type 12 - Reserved
Type 13 - Flow Label

7. Acknowledgments

Authors would like to thank Pedro Marques and Hannes Gredler for their valuable input.

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.

[RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.

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

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.

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

8.2. Informative References

[RFC5095] Abley, J., Savola, P., and G. Neville-Neil, "Deprecation of Type 0 Routing Headers in IPv6", RFC 5095, December 2007.

Authors' Addresses

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

Email: raszuk@cisco.com

Burjiz Pithawala
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US

Email: bpithaw@cisco.com

Danny McPherson
Verisign, Inc.

Email: dmcpherson@verisign.com

IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 19, 2011

R. Raszuk
Cisco Systems
J. Haas
Juniper Networks
R. Steenbergen
nLayer Communications, Inc.
B. Decraene
France Telecom
P. Jakma
DCS, Uni. of Glasgow

October 16, 2010

Wide BGP Communities Attribute
draft-raszuk-wide-bgp-communities-01

Abstract

Communicating various routing policies via route tagging plays an important role in external BGP peering relations. 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 realized with the use of BGP communities.

Such information is important for the BGP speakers to perform some mutually agreed actions without the need to maintain a separate offline database for each pair of prefix and an associated with it requested 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 brings over currently defined BGP communities is the ability to specify, carry as well as use for execution operator’s defined set of parameters. Specification 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/.

This Internet-Draft will expire on April 19, 2011.

Table of Contents

1. Introduction

RFC 1997 [RFC1997] defines a BGP Community Attribute to be used as a tool to contain in BGP update message various additional information about routes which may help to automate peering administration. As defined in RFC 1997 [RFC1997] BGP Communities Attribute consists of one or more sets of four octet values, where each one of them specifies a different community. Except two reserved ranges the encoding of community values mandates that first two octets are to contain the Autonomous System number followed by next two octets containing 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 encoding of 4 octet AS numbers are being addressed by [draft-ietf-idr-as4octet-extcomm-generic-subtype] neither the base BGP communities (both standard or extended) nor as4octet-extcomm-generic document define sufficient level of encoding freedom which could be of practical use. Authors believe that defining a new BGP Path Attribute which will provide ability to contain locally defined parameters will enhance current level of network policies as well as simplify BGP policy management. Proposed simple encoding will also enable to deliver a set of new network services without a need to define a new BGP extension each time.

While defining a new type of any tool there is always a unique opportunity to specify a subset of well recognized behaviors. List of the most commonly used today 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

For the purposes of encoding for Wide BGP Communities a new BGP Path Attribute has been defined. The attribute type code is of the value (TBC by IANA).

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

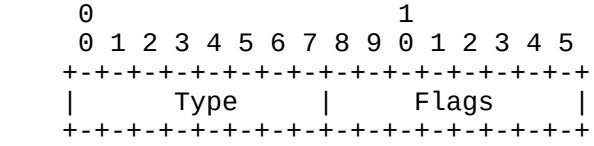
The attribute contains a number of typed containers, which are either fixed or variable in size. Any given container type may appear multiple times, unless that container type’s definition says

otherwise.

3. Wide BGP Community Attribute Containers

Two container templates are defined for carrying BGP community information, to hold fixed or variably sized data. All container definitions MUST conform with one of these two templates.

Containers always start with the following header:



Container header

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

```
  Bit 0: 0 => local community value
         1 => registered community value
      1: 0 => do not decrement TTL field across confederation
              boundaries
         1 => decrement TTL across confederation boundaries
  2...7:   => ignored, preserve or set to zero.
```

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 propagation scope of given community across confederation boundaries. When not set (value of 0) TTL field is not consider at the sub-AS boundaries. When set (value of 1) sub-AS border router follows the same procedure reg handling TTL field as applicable to ASBR at the domain boundary.

3.1. Fixed size container template

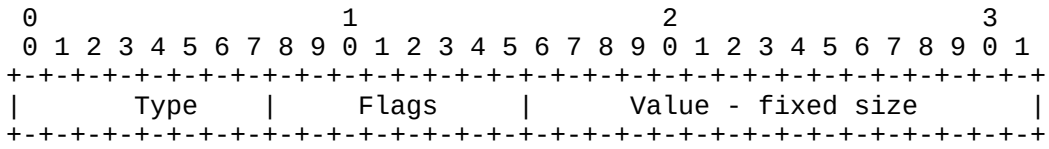


Figure 3: Fixed size type container

3.2. Variable size container template

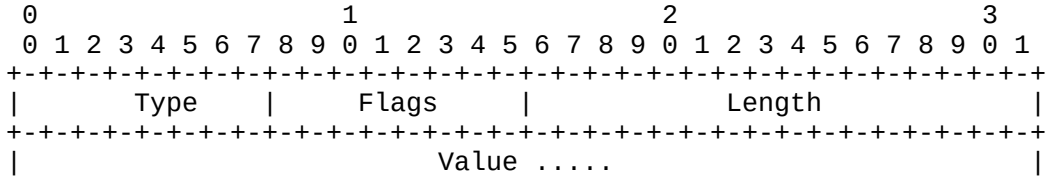
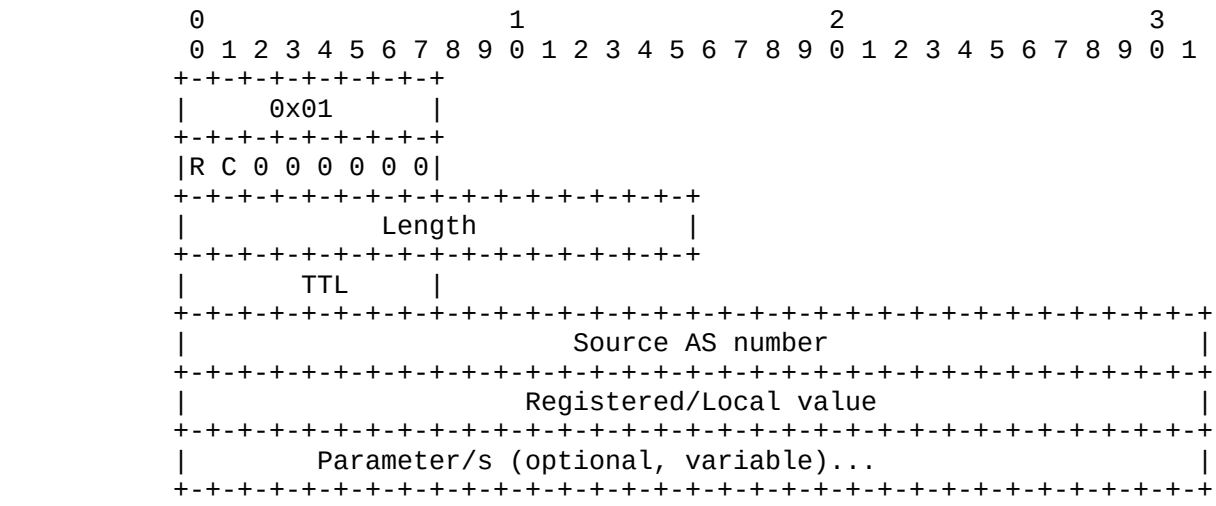


Figure 4: Variable size type container (TLV Format)

4. Container Type 1: Wide Community

Wide BGP Community Type 1 container is of variable size and is encoded as follows:



R is the value of the registered/local bit. C is the value indicating how to treat TTL field across confederation boundaries.

Figure 4: Wide BGP Community Type 1

4.1. Container Type 1 - TTL

TTL: 1 octet

This field represents the forwarding radius in the unit of AS hops for given Wide BGP Community. At each AS boundary when propagating given community over an EBGP session the TTL field must be decremented by value of 1 by the sending EBGP speaker. TTL with value of zero received to the ASBR over IBGP session indicates that this community must not cross an AS boundary.

The special value of 0xFF indicates that the enclosed community may be always propagated over EBGP boundary. Value of 0xFF must not be decremented during propagation.

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

4.2. Container Type 1 - Length

The length represents the total lengths of a given container in octets. The minimum length when no optional parameters are attached is 13 octets.

4.3. Container Type 1 - Community Value

Community Value: 2 octets

The Wide BGP Community value encoded in this field indicates private/local or registered Wide BGP Community type which defines what set of actions a router is requested or recommended to take upon reception of routes with such BGP communities.

4.4. Container Type 1 - Source AS number

Source Autonomous System number: 4 octets

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

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

4.5. Container Type 1 - Community Parameters

Parameters: variable size

Community parameter are defined to contain additional data for execution of given BGP community.

Community parameter field could consist of an autonomous system number(s) which should be conditionally compared when executing given community, AS PATH prepend count to be added, local preference value to be inserted under some conditions, markers indicating number of BGP speakers traversed, cumulative IGP metrics to be used for transparent redistribution, etc...

For consistent Autonomous System treatment all encoded AS numbers SHOULD be encoded as 4 octet values. When such AS is a two octet number the first two octets of this 4 octet value are to be filled with zeros.

Two special values are reserved in the Parameter Autonomous System number field: 0x00000000 - to indicate "None of Autonomous Systems" and value of 0xFFFFFFFF - to indicate "All of Autonomous Systems".

The detailed interpretation of each set of parameters will be provided when describing given community type in a separate document or when locally defined by an operator.

5. Well Known Standard BGP Communities

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

0xFFFF0000	planned-shut	[draft-francois-bgp-gshut]
0xFFFFFF01	NO_EXPORT	[RFC1997]
0xFFFFFF02	NO_ADVERTISE	[RFC1997]
0xFFFFFF03	NO_EXPORT_SUBCONFED	[RFC1997]
0xFFFFFF04	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 the described registry does not intended to obsolete BGP Extended Community Attribute and any already defined and already deployed extended communities.

6. Operational considerations

Having two different ways to propagate locally assigned BGP communities, one via use of Standard BGP Community attribute and the other one via use of Wide BGP Community may seem to potentially cause problems when considering propagation of conflicting actions.

However it needs to be noticed and pointed out that today even within Standard BGP Communities operator or operators may append similar conflicting information to already existing community propagation tool set.

It is therefor recommended that any implementation when supporting both standard and wide BGP communities will allow for their easy inbound and outbound policing. For the actual execution all communities should be treated as union and if supported by an implementation their execution permission are to be a local configuration matter.

When advertising as well as during insertion of Wide BGP Communities

which are predefined as range of values - only use of one value of selected range is allowed.

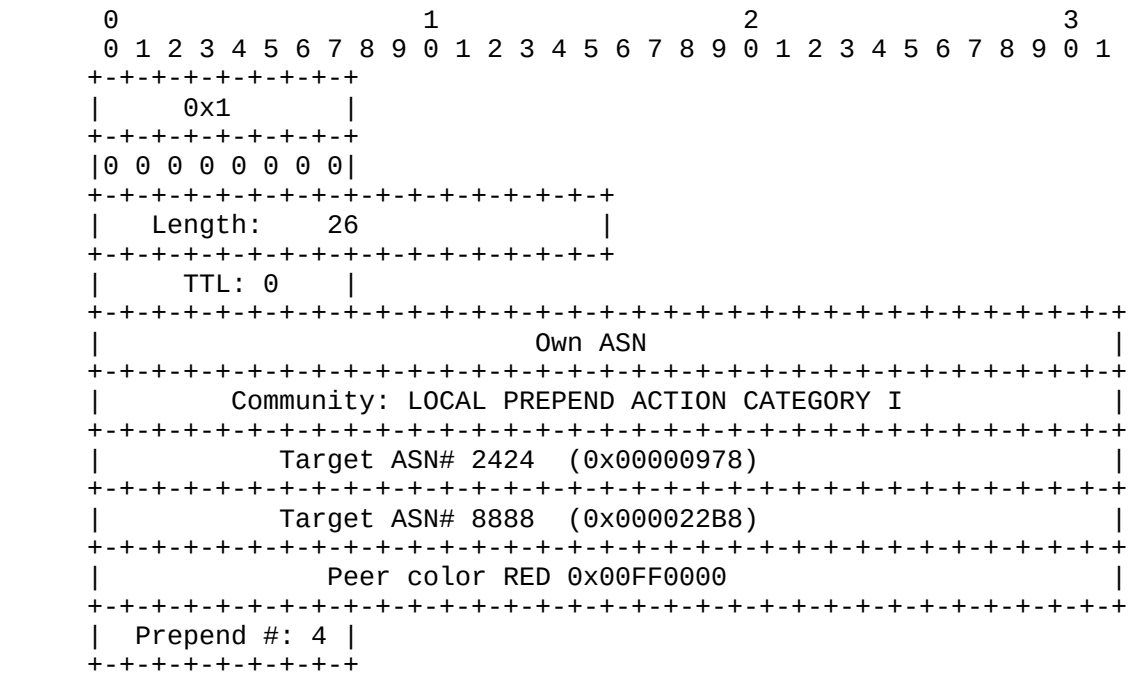
7. Example

An operator wishes to tag incoming routes with a policy specifying that during their advertisement to two peering ASes 2424 and 8888 or during their advertisement to peers marked as RED (0xFF0000) the routes carrying such community will be advertised with AS_PREPEND equal to 4.

That can be easily accomplished by locally defining by an operator a new wide community value using type 1 proposed encoding as below:

PREPEND 4 TIMES TO AS 2424 or 8888 or to peers marked as RED

```
TTL - 0x00
LENGTH - 26 octets
VALUE - 01 / 0x12
PARAMETERS - 2 x 4 octets AS number
             1 x class of peers
             1 octet prepend’s number
```



8. Security considerations

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

9. IANA Considerations

This document defines a new BGP Path Attribute called Wide BGP Communities Attribute. For this new type IANA is to allocate new type 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	27

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

The pool of: 0x00-0xFF 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	0x00
Type 1	0x01
Types 2-254 to be allocated on FCFS basis	
Reserved	0xFF

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

Email: raszuk@cisco.com

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

Email: jhaas@pfrc.org

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
School of Computing Science, Uni. of Glasgow
Sir Alwyn Williams Building
University of Glasgow
Glasgow G1 5AE
UK

Email: paulj@dcs.gla.ac.uk

Inter-Domain Routing
Internet-Draft
Intended status: Standards Track
Expires: September 15, 2011

A. Retana
R. White
Cisco Systems, Inc.
March 14, 2011

BGP Custom Decision Process
draft-retana-bgp-custom-decision-01

Abstract

The BGP specification defines a Decision Process for installation of routes into the Loc-RIB. This process takes into account an extensive series of path attributes, which can be manipulated to indicate preference for specific paths. It is cumbersome (if at all possible) for the end user to define policies that will select, after partial comparison, a path based on subjective local (domain and/or node) criteria.

This document defines a new Extended Community, called the Cost Community, which may be used in tie breaking during the best path selection process. The end result is a local custom decision process.

Table of Contents

1. Introduction

There are a number of metrics available within the BGP decision process [RFC4271] which can be used to determine the exit point for traffic, but there is no metric, or combination of metrics, which can be used to break a tie among generally equal paths.

- LOCAL_PREF: The LOCAL_PREF is an absolute tie breaker near the beginning of the decision process. There is no way to configure the LOCAL_PREF such that the MED, IGP metric, and other metrics are considered before breaking a tie.

- MED: The MULTI_EXIT_DISC is an indicator of which local entrance point an AS would like a peering AS to use; MED isn't suitable to break the tie between two equal cost paths learned from two peer ASes. MED is also compared before the IGP metric; there is no way to set the MED so a path with a higher IGP metric is preferred over a path with a lower IGP metric.

- IGP Metric: It is possible, using the IGP metric, to influence individual paths with otherwise equal cost metrics, but only by changing the next hop towards each path, and configuring the IGP costs of reaching each next hop. This method is cumbersome, and prone to confusion and error.

The BGP specification defines a Decision Process for installation of routes into the Loc-RIB. This process takes into account an extensive series of path attributes, which can be manipulated to indicate preference for specific paths. It is cumbersome (if at all possible) for the end user to define policies that will select, after partial comparison, a path based on subjective local (domain and/or node) criteria.

This document defines a new Extended Community, called the Cost Community, which may be used in tie breaking during the best path selection process. The end result is a custom decision process.

2. Requirements Language

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

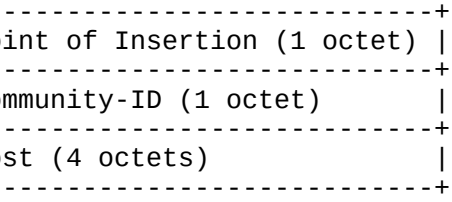
3. The BGP Cost Community

The BGP Cost Community is an Opaque Extended Community [RFC4360]

defined as follows:

Type Field:
The value of the high-order octet of this Opaque Extended Community is 0x03 or 0x43. The value of the low-order octet of the extended type field for this community is 0x01.

Value Field:
The Value field contains three distinct sub-fields, described below:



The Point of Insertion sub-field contains the value of the path attribute *after* which this community MUST be considered during the best path selection process.

The BGP decision process includes some steps that do not correspond to any path attribute; the following values are defined:

128 ABSOLUTE_VALUE - Indicates that the Cost Community MUST be considered as the first step in determining the Degree of Preference of a path.

129 IGP_COST - Indicates that the Cost Community MUST be considered after the interior (IGP) distance to the next-hop has been compared.

130 EXTERNAL_INTERNAL - Indicates that the Cost Community MUST be considered after the paths advertised by BGP speakers in a neighboring autonomous system (if any) have been selected.

131 BGP_ID - Indicates that the Cost Community MUST be considered after the BGP Identifier (or ORIGINATOR_ID [RFC4456]) has been compared.

The Community-ID sub-field contains an identifier to distinguish between multiple instances of the Cost Community.

The Cost sub-field contains a value assigned by the network administrator and that is significant to the local autonomous system. The lower cost MUST be preferred. The default value is 0x7FFFFFFF (half the maximum value).

4. Operation

The network administrator may use the Cost Community to assign a value to a path originated or learned from a peer in any part of the local domain. The Point of Insertion may also be specified using the values assigned by IANA (Section 6) or this document.

If a BGP speaker receives a path that contains the Cost Community, it SHOULD consider its value at the Point of Insertion specified, when calculating the best path [RFC4271].

If the Point of Insertion is not valid for the local best path selection implementation, then the Cost Community SHOULD be silently ignored. Paths that do not contain the Cost Community (for a valid, particular Point of Insertion) MUST be considered to have the default value.

Multiple Cost Communities may indicate the same Point of Insertion. In this case, the Cost Community with the lowest Community-ID is considered first. In other words, all the Cost Communities for a specific Point of Insertion MUST be considered, starting with the one with the lowest Community-ID.

If a range of routes is to be aggregated and the resultant aggregates path attributes do not carry the ATOMIC_AGGREGATE attribute, then the resulting aggregate SHOULD have an Extended Communities path attribute which contains the set union of all the Cost Communities from all of the aggregated routes. If multiple Cost Communities for the same Point of Insertion (and with the same Community-ID), then only the ones with the highest Cost SHOULD be included.

If the non-transitive version of a Cost Community is received across an Autonomous System boundary, then the receiver SHOULD strip it off the BGP update, and ignore it when running the selection process.

5. Deployment Considerations

The mechanisms described in this document may be used to modify the BGP path selection process arbitrarily. It is important that a consistent path selection process be maintained across the local Autonomous System to avoid potential routing loops. In other words,

if the Cost Community is used, all the nodes in the AS that may have to consider this new community at any Point of Insertion SHOULD be aware of the mechanisms described in this document.

6. IANA Considerations

IANA is asked to assign the type values indicated in Section 3 to the Cost Community in the BGP Opaque Extended Community registry [BGP_EXT].

Section 3 also defines a series of values to be used to indicate steps in the best path selection process that do not map directly to a path attribute. IANA is expected to maintain a registry for the Cost Community Point of Insertion values. Values 1 through 127 are to be assigned using the "Standards Action" policy or the Early Allocation process [RFC4020]. Values 128 through 191 are to be assigned using the "IETF Consensus" policy. Values 192 through 254 are to be assigned using the "First Come First Served" policy. Values 0 and 255 are reserved for future use and SHOULD NOT be used. All the policies mentioned are documented in [RFC5226].

Some of the values in this new registry match the values assigned in the BGP Path Attributes registry [BGP_PAR]. It is RECOMMENDED that an effort be made to assign the same values in both tables when applicable. The table in Appendix A shows the initial allocations for the new Cost Community Point of Insertion registry.

7. Security Considerations

This document introduces no new security concerns to BGP or other specifications referenced in this document.

8. Acknowledgements

The authors would like to thank Chris Whyte, Khamsa Enaya, John Scudder, Tom Barron, Eric Rosen, Barry Friedman, Gargi Nalawade, Ruchi Kapoor, Chandra Appanna, Keyur Patel and Pradosh Mohapatra for their comments and suggestions. We would like to also thank Dan Tappan for the Opaque Extended Community type.

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.

[RFC4020] Kompella, K. and A. Zinin, "Early IANA Allocation of Standards Track Code Points", BCP 100, RFC 4020, February 2005.

[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.

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

9.2. Informative References

[BGP_EXT] Internet Assigned Numbers Authority, "BGP Extended Communities", 2010, <http://www.iana.org/assignments/bgp-extended-communities>.

[BGP_PAR] Internet Assigned Numbers Authority, "BGP Parameters", 2010, <http://www.iana.org/assignments/bgp-parameters/>.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.

Appendix A. Cost Community Point of Insertion Registry

The tables below document the initial Cost Community Point of Insertion Registry

Range	Registration Procedure
0	Reserved
1-127	Standards Action
128-191	IETF Consensus
192-254	First Come First Served
255	Reserved

Registration Procedure

Value	Code	Reference
1	ORIGIN	RFC4271
2	AS_PATH	RFC4271
3	Unassigned	
4	MULTI_EXIT_DISC	RFC4271
5	LOCAL_PREF	RFC4271
6-25	Unassigned	
26	AIGP	draft-ietf-idr-aigp
27-127	Unassigned	
128	ABSOLUTE_VALUE	draft-retana-bgp-custom-decision
129	IGP_COST	draft-retana-bgp-custom-decision
130	EXTERNAL_INTERNAL	draft-retana-bgp-custom-decision
131	BGP_ID	draft-retana-bgp-custom-decision

Point of Insertion Codes

Appendix B. Changes from version -00

The changes with respect to version -00 of this draft are as follow:

- Defined a transitive type. (Section 3)
- Updated the IANA Considerations (Section 6) to create a Cost Community Point of Insertion Registry. (Appendix A)
- Miscellaneous Updates: updated format, refreshed references, updated acknowledgements, minor edits.

Authors' Addresses

Alvaro Retana
Cisco Systems, Inc.
7025 Kit Creek Rd.
Research Triangle Park, NC 27709
USA

Phone: +1 919 392 2061
Email: aretana@cisco.com

Russ White
Cisco Systems, Inc.
7025 Kit Creek Rd.
Research Triangle Park, NC 27709
USA

Phone: +1 919 392 3139
Email: russwh@cisco.com

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: August 24, 2011

R. Shakir
C&W

February 20, 2011

Operational Requirements for Enhanced Error Handling Behaviour in BGP-4
draft-shakir-idr-ops-reqs-for-bgp-error-handling-01

Abstract

BGP-4 is utilised as a key intra- and inter-Autonomous System routing protocol in modern IP networks. The failure modes as defined by the original protocol standards are based on a number of assumptions around the impact of session failure. Numerous incidents both in the global Internet routing table and within Service Provider networks have been caused by strict handling of a single invalid UPDATE message causing large-scale failures in one or more Autonomous Systems.

This memo describes the current use of BGP-4 within Service Provider networks, and outlines a set of requirements for further work to enhance the mechanisms available to a BGP-4 implementation when erroneous data is detected. Whilst this document does not provide specification of any standard, it is intended as an overview of a set of enhancements to BGP-4 to improve the protocol’s robustness to suit its current deployment.

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 August 24, 2011.

Copyright Notice

Table of Contents

1. Introduction

Where BGP-4 [RFC4271] is deployed in the Internet and Service Provider networks, numerous incidents have been recorded due to the manner in which [RFC4271] specifies errors in routing information should be handled. Whilst the behaviour defined in the existing standards retains utility, the deployments of the protocol have changed within modern networks, resulting in significantly different demands for protocol robustness. Whilst a number of Internet Drafts have been written to begin to enhance the behaviour of BGP-4 in terms of the handling of erroneous messages, this draft intends to define a set of requirements for ongoing work. These requirements are considered from the perspective of a Network Operator, and hence this draft does not intend to define the protocol mechanisms by which such error handling behaviour is to be implemented.

1.1. Role of BGP-4 in Service Provider Networks

BGP was designed as an inter-Autonomous System (AS) routing protocol and hence many of the error handling mechanisms within the protocol specification are designed to be conducive to this role. In general, this consideration as an inter-AS routing propagation mechanism results in the view that a BGP session propagates a relatively small amount of network-layer reachability information (NLRI) between two ASes. In this case, it is the expectation of session resilience for those adjacencies that are key to routing continuity (for example, it is expected that two networks peering via BGP would connect multiple times in order to safeguard equipment or protocol failure). In addition, there is some expectation of multiple paths to a particular NLRI being available - it would be expected that a network can fall back to utilising alternate, less direct, paths where a failure of a more direct path occurs.

Traditional network architectures would deploy an Interior Gateway Protocol (IGP) to carry infrastructure and customer prefixes, with an Exterior Gateway Protocol (EGP) such as BGP being utilised to propagate these prefixes to other Autonomous Systems. However, with the growth of IP-based services, this is no longer considered best practice. In order to ensure that convergence is within acceptable time bounds, the amount of routing information carried within the IGP is significantly reduced - and tends to be only infrastructure prefixes. iBGP is then utilised to propagate both customer, and external prefixes within an AS. As such, BGP has become an IGP, with traditional IGPs acting as a means by which to propagate the routing information which is required to establish a BGP session, and reach the egress node within the local routing domain. This change in role presents different requirements for the robustness of BGP as a routing protocol - with the expectation of similar level of

robustness to that of an IGP being set.

Along with this change in role, the nature of the IP routing information that is carried has changed. BGP has become a ubiquitous means by which service information can be propagated between devices. For instance, BGP is utilised to carry routing information for IP/MPLS VPN services as described in [RFC4364]. Since there is an existing deployment of the protocol between PE devices in numerous networks, it has been adapted to propagate this routing information, as its use limits number of routing protocols required on each device. This additional information being propagated represents a large change in requirement for the error handling of the protocol - where session failure occurs, it is likely a complete service outage for at least a subset of a network's customers is experienced where an erroneous packet may have occurred within a different sub-topology or even service (a different address family for example). For this reason, there is a significant demand to avoid service affecting failures that may be triggered by routing information within a single sub-topology or service.

Both within Internet and multi-service routing architectures, a number of BGP sessions propagate a large proportion of the required routing information for network operation. For Internet routing, these are typically BGP sessions which propagate the global routing table to an AS - failure of these sessions may have a large impact on network service, based on a single erroneous update. In an multi-service environment, typical deployments utilise a small number of core-facing BGP sessions, typically towards route reflector devices. Failure of these sessions may also result in a large impact to network operation. Clearly, the avoidance of conditions requiring these sessions to fail is of great utility to any network operator, and provides further motivation for the revision of the existing behaviour.

Whilst the behaviour in [RFC4271] is suited to ensuring that BGP messages with erroneous routing information in are limited in scope (by means of session reset), with the above considerations, it is clear that this mechanism is not suited to all deployments. It should, however, be noted that the change in scope affects the handling only of errors occurring after BGP session establishment. There is no current operational requirement to amend the means by which error handling in session establishment, or liveliness detection, are performed.

1.2. Overview of Operator Requirements for BGP-4 Error Handling

It is the intention of this document to define a set of criteria for the manner in which a revised error handling mechanism in BGP-4 is

required to conform. The motivation for the definition of these requirements can be summarised based on certain behaviour currently present in the protocol that is not deemed acceptable within current operational deployments, or where there is a short-fall in the tool set available to an operator. These key requirements can be summarised as follows:

- It is unacceptable within modern deployments of the BGP-4 protocol that a single erroneous UPDATE packet affects prefixes that it does not carry. This requirement therefore requires some modification to the means by which erroneous UPDATE packets are handled, and reacted to - with a particular focus on avoiding the use of the NOTIFICATION message.

- It is recognised that some error conditions may occur within the BGP-4 protocol may not always be handled gracefully, and may result in conditions whereby an implementation cannot recover. In these (and similar) cases, it is unacceptable for an operator that this reset of the BGP-4 session results in interruption to forwarding packets (by means of withdrawing prefixes installed by BGP-4 into a device's RIB, and subsequently FIB). To this end, there is a requirement to define a session reset mechanism which provides session re-initialisation in a non-destructive manner.

- Further to the requirements to provide a more robust protocol, the current visibility into error conditions within the BGP-4 protocol is extremely limited - where further modifications to this behaviour are to be made, complexity is likely to be added. Thus, to ensure that BGP-4 is manageable, there are requirements for mechanisms by which the protocol can be examined and monitored.

This document describes each of these requirements in further depth, along with an overview of means by which they are expected to be achieved. In addition, the mechanism by which the enhancements meeting these requirements are to interact is discussed.

2. Avoiding use of NOTIFICATION

The error handling behaviour defined in RFC4271 is problematic due to the limited options that are available to an implementation. When an erroneous BGP message is received, at the current time, the implementation must either ignore the error, or send a NOTIFICATION message, after which it is mandatory to terminate the BGP session. It is apparent that this requirement is at odds with that of protocol robustness.

There is significant complexity to this requirement. The mechanism defined in [I-D.chen-ebgp-error-handling] describes a means by which no NOTIFICATION message is generated for all cases whereby NLRI can be extracted from an UPDATE. The NLRI contained within the erroneous UPDATE message is considered as though the remote BGP speaker has provided an UPDATE marking it as withdrawn. This results in a limit in the propagation of the invalid routing information, whilst also ensuring that no traffic is forwarded via a previously-known path that may no longer be valid. This mechanism is referred to as "treat-as-withdraw".

Whilst this behaviour results in avoiding a NOTIFICATION message, keeping other routing information advertised by the remote BGP speaker within the RIB, it may result in unreachability for a sub-set of the NLRI advertised by the remote speaker. Two cases should be considered - that where the entry for a prefix in the Adj-RIB-In of the neighbour propagating an erroneous packet is utilised, and that where the prefix installed in the device’s RIB is learnt from another BGP speaker. In the former case, should the identified NLRI not be treated as withdrawn, the original NLRI is utilised within the global RIB. However, this information is potentially now invalid (i.e. it no longer provides a valid forwarding path), whilst an alternate (valid) path may exist in another Adj-RIB-In. By continuing to utilise the NLRI for which the UPDATE was considered invalid, traffic may be forwarded via an invalid path, resulting in routing loops, or black-holing. In the second case, no impact to the forwarding of traffic, or global RIB, is incurred, yet where treat-as-withdraw is implemented, possibly stale routing information is purged from the Adj-RIB-In of the neighbour propagating errors.

Whilst mechanisms such as "treat-as-withdraw" are currently documented, the proposals are limited in their scope - particularly in terms of restrictions to implementation only on eBGP sessions. This limitation is made based on the view that the BGP RIB must be consistent across an autonomous system. By implementing treat-as-withdraw for a iBGP session, one or more routers within the Autonomous System may not have reachability to a prefix, and hence blackholing of traffic, or routing loops, may occur. It should,

however, be considered if this view is valid, in light of the manner in which BGP is utilised within operator networks. Inconsistency in a RIB based on a single UPDATE being treated as withdrawn may cause a inconsistency in a single sub-topology (e.g. Layer 3 VPN service), or a service not operating completely (in the case of an UPDATE carrying service membership information). Where a NOTIFICATION and teardown is utilised this is destructive to all sub-topologies in all address family identifiers (AFIs) carried by the session in question. Even where mechanisms such as multi-session BGP are utilised, a whole AFI is affected by such a NOTIFICATION message. In terms of routing operation, it is therefore far less costly to endure a situation where a limited sub-set of routing information within an AS is invalid, than to consider all routing information as invalid based on a single trigger.

It is considered that, if extended to cover iBGP, the mechanisms described in [I-D.chen-ebgp-error-handling] and [I-D.ietf-idr-optional-transitive] provide a means to avoid the transmission of a NOTIFICATION to a remote BGP speaker based on a single erroneous message, where at all possible, and hence meet this requirement. The failure cases whereby NLRI cannot be extracted from the UPDATE message represent a case whereby the receiving system cannot handle the error gracefully based on this mechanism.

3. Recovering RIB Consistency

The recommendations described in Section 2 may result in the RIB for a topology within an AS being inconsistent across the AS' internal routers. Alternatively, where such mechanisms are deployed at an AS boundary, interconnects between two ASes may be inconsistent with each other. There are therefore risks of traffic blackholing, due to missing routing information, or forwarding loops. Whilst this is deemed an acceptable compromise in the short term, clearly, it is suboptimal. Therefore, a requirement exists to provide mechanisms by which a BGP speaker is able to recover the consistency of the Adj-RIB-In for a particular neighbour.

It is envisaged that during such routing inconsistencies, the local BGP speaker is aware that some routing information was not able to be processed - due to the fact that an UPDATE message was not parsed correctly. If the 'treat-as-withdraw' mechanism described within Section 2 is utilised, it is also possible for the local BGP speaker to have determined the set of NLRI for which an erroneous UPDATE message was received. In this scenario, by utilising targeted mechanisms to re-request the specific NLRI that was unreachable, this routing information can be re-transmitted from the remote BGP speaker. Such a request requires extension to the existing BGP-4 protocol, in terms of specific UPDATE generation filters with a transient lifetime. It is envisaged that the work within [I-D.zeng-one-time-prefix-orf] provides a mechanism allowing targeted elements of the Adj-RIB-In for a BGP neighbour to be recovered.

In addition to such cases where specific routing information is known to be erroneous, the more general case where either a large amount of the Adj-RIB-In is contained in UPDATE messages subject to treat-as-withdraw, or the specific prefixes are unknown to the local BGP speaker must be considered. In this case, there is a requirement for a BGP speaker to re-request the entire RIB advertised by a remote neighbour. In this case, where such re-advertisement is required, it is envisaged that a ROUTE-REFRESH as per the description in [RFC2918] is utilised. [I-D.keyur-bgp-enhanced-route-refresh] provides a means by which the ROUTE-REFRESH mechanism can be extended in order to meet this requirement.

It is of particular note for both means of recovering RIB consistency described that these are effective only when considering transitive errors within an implementation - for instance, should an RFC interpretation error within an implementation be present, regardless of the number of times a specific UPDATE is generated, it is likely that this error condition will persist. For this reason, there is an requirement to consider the means by which such consistency recovery mechanisms are utilised. It is not advisable that a transitive

filter and advertisement mechanism is triggered by all error handling events due to the load this is likely to place on the neighbour receiving such a request. Where this BGP speaker is a relatively centralised device - a route reflector (as described by [RFC4456]) for example - the act of generation of UPDATE messages with such frequency is likely to cause disproportionate load. It is therefore an operational requirement of such mechanisms that means of request dampening be required by any such extension.

4. Reducing the Impact of Session Reset

Even where protocol enhancements allow errors in the BGP-4 protocol to cease to trigger NOTIFICATION messages, and hence reset a BGP session, it is clear that some error conditions may not be exited. In particular, errors due to existing state, or memory structures, associated with a specific BGP session will not be handled. It is therefore important to consider how these error conditions are currently handled by the protocol. It should be noted that the following discussion and analysis considers only those NOTIFICATION messages generated in response to errors in UPDATE messages (as defined by Section 6.3 in [RFC4271]).

The existing NOTIFICATION behaviour triggers a reset of all elements of the BGP-4 session, as described in Section 6 of [RFC4271]. It is expected that session teardown requires an implementation to re-initialise all structures and state required for session maintenance. Clearly, there is some utility to this requirement, as error conditions in BGP are, in general, exited from. However, this definition is responsible for the forwarding outages within networks utilising BGP for route propagation when each error is experienced. The requirement described in Section 2 is intended to reduce the cases whereby a NOTIFICATION is required, however, any mechanism implemented as a response to this requirement by definition cannot provide a session reset to the extent of that achieved by the current behaviour.

In order to address this, there is a requirement for a means by which a BGP speaker can signal that an unhandled error condition in an UPDATE message occurred - requiring a session reset - yet also continue to utilise the paths advertised by the neighbour that are currently in use within the RIB. In this case, the Adj-RIB-In received from the neighbour is not considered invalid, despite a NOTIFICATION, and session reset, being required. This set of requirements is akin to those answered by the BGP Graceful Restart mechanism described in [RFC4724]. Since the operational requirement in this case is to provide a means to achieve a complete session restart without disrupting the forwarding path of those prefixes in use within a BGP speaker’s RIB, it is expected that utilising a procedure similar to the Graceful Restart mechanism meets the error handling requirement. By responding to an error condition (repeated or otherwise) with a message indicating that an error that cannot be handled has occurred, forcing session reset, whilst retaining forwarding information within the RIB allows forwarding to all prefixes within a system’s RIB to continue, whilst the session restarts. By placing a time bound on the restart lifetime, should an error condition not be transient - for example, should an error have occurred with the BGP process, rather than a specific of the BGP

session - the remote BGP speaker is still detected as an invalid device for forwarding.

It should, however, be noted that a protocol enhancement meeting this requirement is not able to solve all error conditions - however, a complete restart of the BGP and TCP session between two BGP speakers implements an identical recovery mechanism to that which is achieved by the existing behaviour. Where an error condition such as memory or configuration corruption has occurred in a BGP implementation, it is expected that a mechanism meeting this requirement continues to detect this, by means of a bound on time for session restart to occur. Whilst there may be some consideration that packets continue to be forwarded through a device which can be in an failure mode of this nature for a longer period, due to this requirement, the architecture of modern IP routers should be considered. A divided forwarding and control plane is common in many devices, as well as process separation for software-based devices - corruption of a specific protocol daemon does not necessarily imply forwarding is affected. Indeed, where forwarding behaviour of a device is affected, it is envisaged that a failure detection mechanism (be it Bidirectional Forwarding Detection, or indeed BGP KEEPALIVE packets) will detect such a failure in almost all cases, with the symptomatic behaviour of such a failure being an invalid UPDATE message in very few other cases.

5. Operational Toolset for Monitoring BGP

A significant complexity that is introduced through the requirements defined in this document is that of monitoring BGP session status for an operator. Although the existing error handling behaviour causes a disproportionate failure, session failure is extremely visible to most operational personnel within a Network Operator due to both existing definitions of SNMP trap mechanisms for BGP, along with the forwarding impact typically caused by such a failure. By introducing mechanisms by which errors of this nature are not as visible, this is no longer the case. There is a requirement that where subsets of the RIB on a device are no longer reachable from a BGP speaker, or indeed an AS, that some mechanism to determine the cause is available to an operator. Whilst, to some extent, this can be solved by mandating a sub-requirement of each of the aforementioned requirements that a BGP speaker must log where such errors occur, and are hence handled, this does not solve all cases. In order to clarify this requirement, the example of the transmission of an erroneous Optional Transitive attribute can be considered. Since, by definition, there is no requirement for all BGP speakers to parse such an attribute, a receiving router may treat NLRI as withdrawn based on an erroneous attribute not examined by its neighbour. In this case, the upstream device or network, propagating the UPDATE, has no visibility of this error. Operationally, however, it is of interest to the upstream router operator that such invalid information was propagated.

The requirement for logging of error conditions in transmitted BGP messages, which are visible to only the receiver, cannot be achieved by any existing BGP message, or capability. It is envisaged that each erroneous event should be transmitted to the remote peer - including the information as to the set of NLRI that were considered invalid. Whilst with some mechanisms this is achieved by default (for example, One-Time Prefix ORF [I-D.zeng-one-time-prefix-orf] (Outbound Route Filtering) will transmit the set of prefixes that are required), the operator requirement is to know which prefixes may have been unreachable in all cases. It is envisaged that an extension to meet this requirement will allow for such information to be transmitted between peers, and hence logged. Such a mechanism may provide further utility as a either a diagnostic, or logging toolset.

It should be noted that numerous work items within the IETF exist at the time of writing that begin to solve this requirement. Within the IDR working group both [I-D.raszuk-bgp-diagnostic-message] and [I-D.ietf-idr-advisory] provide mechanisms by which such information can be propagated in-band to an existing BGP session. Transmitting such diagnostic information in-band is considered the optimal means by which to propagate details of errors present in UPDATE messages, due to the fact that no additional protocols (and hence security and

trust concerns) must be configured between two Autonomous Systems (where the errors occur at an AS boundary), and the load on each BGP speaker is increased only due to an additional capability, rather than an additional code base, and protocol. Clearly, any mechanism implemented in-band to a BGP session is required to be relatively lightweight, since the information provided over the session is an enhancement to the operational visibility of the protocol, and should not disrupt core protocol operations. Other, out-of-band, mechanisms - such as that proposed in [I-D.ietf-grow-bmp] are likely to provide mechanisms by which further insight into BGP operation can be achieved. The fact that such a protocol is implemented independently of the BGP protocol results in further flexibility to provide detailed protocol data, without introducing further complexity to the BGP protocol itself.

6. Operational Complexities Introduced by Altering RFC4271

The existing NOTIFICATION and subsequent teardown of a BGP session upon encountering an error has the advantage that a consistent approach to error handling is required of all implementations of the BGP-4 protocol. This is of operational advantage, as it provides a clear expectation of the behaviour of the protocol. The requirements defined herein add further complexity to the error-handling within BGP, and hence are liable to compromise the existing deterministic protocol behaviour. It is therefore deemed that there is a further requirement to provide a clear method by which an erroneous UPDATE should be reacted to, in order that all protocol implementations provide a consistent means by which recovery is achieved. A further complexity is introduced due to the disparate nature of the work items altering the BGP error handling behaviour - since all items are likely to be implemented as a BGP capability [RFC5492], situations are likely to occur between devices (especially those with different BGP implementations), where some of the mechanisms referenced are unsupported. This adds further barriers to a standard definition of the BGP-4 error handling behaviour.

In general, the approach considered ideal upon encountering an erroneous UPDATE message can be divided into two cases - those where the NLRI can be determined from the message, and those where it cannot be. The latter case is the simpler of the two. In this case, there is a requirement for the implementation to reset the BGP session, utilising the reduced-impact approach, described in Section 4. In the case where the remote BGP speaker is in a transient error condition related to specific peer data structures, or state, a single instance of this behaviour is likely to exit the error condition. In the case of implementation errors, it is possible that the BGP session in question may enter a continuous loop of being reset, with a partial RIB being held by one or more of the BGP speakers due to an non-deterministic order of UPDATE propagation. It is therefore a requirement that within this reduced-impact procedure any subsequent UPDATE messages that would result in further session resets are ignored. Whilst this results in a condition where an undetermined amount of the RIB is inconsistent, partial reachability is maintained. In this case, the operational toolsets discussed in Section 5 is likely to provide mechanisms by which this condition can be brought to the attention of the relevant operators. This requirement to accept a partial RIB, which results in potential invalid traffic forwarding is a direct result of the deployments of BGP-4, as described in Section 1.1.

The case where NLRI can be determined from an erroneous UPDATE provides further complexities. In this case, a BGP speaker is aware of the sub-set of the RIB which have been identified as being

contained within invalid UPDATE messages. This allows a local BGP speaker to re-request single prefixes, utilising a mechanism such as "one-time prefix ORF". However, a similar result is achieved by re-requesting the entire RIB - albeit with greater resource requirements. It is therefore expected that the process of recovery utilises a staged set of mechanisms to attempt to restore consistency of the RIB:

1. Where available, a mechanism capable of requesting only the NLRI determined to have been contained within a invalid UPDATE should be utilised. However, since it is possible that such an error condition can be transient in nature, it is likely that more than one request is to be transmitted (assuming the first does not return a valid UPDATE message). In order to allow a deterministic process, there is a requirement for a limit on the number of specific requests transmitted to be defined.

2. Where a specific refresh mechanism is not available, a peer should re-request the entire RIB. Again, there is a requirement to limit the number of complete RIB requests that should be sent via an implementation, in order to provide a bound both on the expected level of load a device may experience, and on the time for which the RIB may be inconsistent.

3. Finally, a session reset should be performed, as per the reduced-impact NOTIFICATION requirement defined in Section 4. At this point, a similar challenge to that discussed above exists, should the error condition persist. In this case, as defined above, there is a requirement to ignore those UPDATE messages that continue to be erroneous.

It is envisaged that where limits are required, these will be defined on a per memo-basis, or within a further revision of the requirements described herein.

Whilst the approach described above provides a standard means by which error recovery may be handled on a per UPDATE basis, further complexities are raised where multiple errors occur. Clearly, following this procedure causes control-plane load on both the BGP speakers - for this reason, consideration of how repeated use of the mechanisms discussed in this document is required. It is notable that errors may not occur with UPDATE messages relating to only a single NLRI, independent errors in multiple NLRIs may be experienced. For this reason, it is required that an implementation rate limits the number of error handling events sourced towards a particular neighbour. It is expected that such rate limiting, or event suppression is achieved on a per-session basis, where state information is already held, rather than on a per-prefix basis as it

is envisaged that such behaviour presents significant scaling problems, and introduces further state requirements for an implementation of the protocol. It is recommended that where a flag indicative of erroneous behaviour is implemented, the state of such a value is maintained independently of session establishment.

7. IANA Considerations

This memo includes no request to IANA.

8. Security Considerations

The requirements outlined in this document provide mechanisms by which erroneous BGP messages may be responded to with limited impact to forwarding operation. This is of benefit to the security of a BGP speaker in general. Where UPDATE messages may have been propagated by a single malicious Autonomous System or router within a network (or the Internet default free zone - DFZ), which are then propagated to all devices within the same routing domain, all other NLRI available over the same session become unreachable. This mechanism may provide means by which an Autonomous System can be isolated from required routing domains (such as the Internet), should the relevant UPDATE messages be propagated via specific paths. By reducing the impact of such failures, it is envisaged that this possibility may be constrained to a specific set of NLRI, or a specific topology.

Some mechanisms meeting the requirements specified in this document, particularly those within Section 5 may provide further security concerns, however, it is envisaged that these are addressed in per-enhancement memos.

9. Acknowledgements

The author would like to thank Rob Evans, David Freedman, Tom Hodgson, Sven Huster, Jonathan Newton, Neil McRae, Thomas Mangin, Tom Scholl and Ilya Varlashkin for their review and valuable feedback.

10. References

10.1. Normative References

[I-D.chen-ebgp-error-handling]
Chen, E., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP Updates from External Neighbors", draft-chen-ebgp-error-handling-00 (work in progress), September 2010.

[I-D.ietf-grow-bmp]
Scudder, J., Fernando, R., and S. Stuart, "BGP Monitoring Protocol", draft-ietf-grow-bmp-05 (work in progress), December 2010.

[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.ietf-idr-optional-transitive]
Scudder, J. and E. Chen, "Error Handling for Optional Transitive BGP Attributes", draft-ietf-idr-optional-transitive-03 (work in progress), September 2010.

[I-D.keyur-bgp-enhanced-route-refresh]
Patel, K., Chen, E., and B. Venkatachalapathy, "Enhanced Route Refresh Capability for BGP-4", draft-keyur-bgp-enhanced-route-refresh-01 (work in progress), October 2010.

[I-D.raszuk-bgp-diagnostic-message]
Raszuk, R., Chen, E., and B. Decraene, "BGP Diagnostic Message", draft-raszuk-bgp-diagnostic-message-00 (work in progress), October 2010.

[I-D.zeng-one-time-prefix-orf]
Zeng, Q. and J. Dong, "One-time Address-Prefix Based Outbound Route Filter for BGP-4", draft-zeng-one-time-prefix-orf-01 (work in progress), October 2010.

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

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

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

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.

[RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, January 2007.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.

10.2. Informational References

[RFC5881] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)", RFC 5881, June 2010.

Author's Address

Rob Shakir
Cable&Wireless Worldwide

Email: rob.shakir@cw.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 11, 2011

K. Patel
Cisco Systems
D. Ward
Juniper Networks
R. Bush
Internet Initiative Japan
March 10, 2011

Extended Message support for BGP
draft-ymbk-bgp-extended-messages-01

Abstract

The current BGP specification mandates a maximum BGP message size of 4096 octets. As BGP is extended to support newer AFI/SAFIs, there is a need to extend the maximum message size beyond 4096 octets. This draft provides an extension for BGP to extend its current message size for BGP messages from 4096 octets to 65535 octets.

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 September 11, 2011.

Table of Contents

1. Introduction

The current BGP specification [RFC4271] mandates a maximum BGP message size of 4096 octets. As BGP is extended to support newer AFI/SAFIs and newer capabilities (e.g., [I-D.lepinski-bgpsec-overview], there is a need to extend the maximum message size beyond 4096 octets. This draft provides an extension for BGP to extend its current message size for BGP messages from 4096 octets to 65535 octets.

2. Extended message Capability for BGP

To advertise BGP Extended Message Capability to a peer, a BGP speaker uses BGP Capabilities Advertisement [RFC3392]. By advertising the BGP Extended message Capability to a peer, a BGP speaker conveys to that peer that the speaker is capable of receiving and properly handling BGP Extended Messages.

This is an asymmetric capability. I.e. one speaker could signal the capability and the other not, so that extended messages could flow only in the direction toward the speaker which advertised the capability.

The BGP Extended Message Capability is a new BGP Capability [RFC3392] defined with Capability code TBD and Capability length 0.

3. Operation

A BGP speaker that is willing to receive BGP Extended Messages from its peer should advertise the BGP Extended Message Capability to its peer using BGP Capabilities Advertisement [RFC3392]. A BGP speaker may send extended messages to its peer only if it has received the Extended Message Capability from its peer.

All BGP extended messages have maximum message size of 65535 octets. The smallest message that may be sent consists of a BGP header without a data portion (19 octets). All multi-octet fields are in network byte order.

4. Acknowledgements

The authors thank John Scudder for his input.

5. IANA Considerations

This document defines the Extended Message Capability for BGP. The new Capability code needs to be assigned by IANA.

6. Security Considerations

This extension to BGP does not change BGP's underlying security issues.

7. References

7.1. Normative References

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

[RFC3392] Chandra, R. and J. Scudder, "Capabilities Advertisement with BGP-4", RFC 3392, November 2002.

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

7.2. Informative References

[I-D.lepinski-bgpsec-overview] Lepinski, M. and S. Turner, "An Overview of BGPSEC", draft-lepinski-bgpsec-overview-00 (work in progress), March 2011.

Authors' Addresses

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

Email: keyupate@cisco.com

Dave Ward
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
USA

Email: dward@juniper.net

Randy Bush
Internet Initiative Japan
5147 Crystal Springs
Bainbridge Island, Washington 98110
US

Phone: +1 206 780 0431 x1
Email: randy@psg.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 8, 2011

C. Pelsser
R. Bush
IIJ
K. Patel
P. Mohapatra
Cisco Systems
O. Maenel
Loughborough University
March 7, 2011

Making Route Flap Damping Usable
draft-ymbk-rfd-usable-00

Abstract

Route Flap Damping (RFD) was first proposed to reduce BGP churn in routers. Unfortunately, RFD was found to severely penalize sites for being well-connected because topological richness amplifies the number of update messages exchanged. Many operators have turned RFD off. This document recommends adjusting a few RFD algorithmic constants and limits, to reduce the high risks with RFD, with the result being damping a non-trivial amount of long term churn without penalizing well-behaved prefixes' normal convergence process.

Requirements Language

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

Status of this Memo

This Internet-Draft will expire on September 8, 2011.

Table of Contents

1. Suggested Reading

It is assumed that the reader understands BGP, [RFC4271] and Route Flap Damping, [RFC2439]. This work is based on the measurements in the paper [pelsser2011].

2. Introduction

Route Flap Damping (RFD) was first proposed (see [ripe178] and [RFC2439]) and subsequently implemented to reduce BGP churn in routers. Unfortunately, RFD was found to severely penalize sites for being well-connected because topological richness amplifies the number of update messages exchanged, see [mao2002]. Subsequently, many operators turned RFD off, see [ripe378]. This document recommends adjusting a few RFD algorithmic constants and limits, with the result being damping of a non-trivial amount of long term churn without penalizing well-behaved prefixes' normal convergence process.

Very few prefixes are responsible for a large amount of the BGP messages received by a router, see [huston2006] and [pelsser2011]. For example, [pelsser2011] showed that only 3% of the prefixes were responsible for 36% percent of the BGP messages at a router with real feeds from a Tier-1 and an Internet Exchange Point during a one week experiment. Only these very frequently flapping prefixes should be damped. The values recommended in Section 6 achieve this. Thus, RFD can be enabled, and some churn reduced.

The goal is to, with absolutely minimal change, ameliorate the danger of current RFD implementations and use. It is not a panacea, nor is it a deep and thorough approach to flap reduction.

3. Suppress Threshold Versus Churn

By turning RFD back on with the values recommended in Section 6 churn is reduced. Moreover, with these values, prefixes going through normal convergence are generally not damped.

[pelsser2011] estimates that, with a suppress threshold of 6,000, the BGP update rate is reduced by 19% compared to a situation without RFD enabled. With this 6K suppress threshold, 90% fewer prefixes are damped compared to use of a 2K threshold. I.e. far fewer well-behaved prefixes are damped.

Setting the suppress threshold to 12K leads to very few damped prefixes (1.7% of the prefixes damped with a threshold of 2K, in the experiments in [pelsser2011] yielding an average hourly update

reduction of 11% compared to not using RFD.

Suppress Threshold	Damped Instances	Update Rate (one hour bins)
2k	43342	53.11%
4k	11253	74.16%
6k	4352	81.03%
8k	2104	84.85%
10k	1286	87.12%
12k	720	88.74%
14k	504	89.97%
16k	353	91.01%
18k	311	91.88%
20k	261	92.69%

Damped Prefixes Versus Churn

Table 1

4. RFD Parameters

The following RFD parameters are common to all implementations. Some may be adjusted by the operator, some not.

Parameter	Tunable?	Cisco	Juniper
Withdrawal	No	1000	1000
Re-Advertisement	No	0	1000
Attribute Change	No	500	500
Suppress Threshold	Yes	2000	3000
Half-Life (min)	Yes	15	15
Reuse Threshold	Yes	750	750
Max Suppress Time (min)	Yes	60	60

RFD Paramaters of Juniper and Cisco

Table 2

5. Maximum Penalty

It is important to understand that the parameters shown in Table 2, and the implementation's sampling rate, impose an upper bound on the penalty value, which we can call the 'computed maximum penalty'.

In addition, BGP implementations have an internal constant which we will call the 'maximum penalty' which the current computed penalty may not exceed.

6. Recommendations

The following changes are recommended:

Router Maximum Penalty: The internal constant for the maximum penalty value MUST be raised to at least 50,000.

Default Configurable Parameters: In order not to break existing operational configurations, BGP implementations SHOULD NOT change the default values in Table 2.

Minimum Suppress Threshold: Operators wishing damping which is much less destructive than current, but still somewhat aggressive SHOULD configure the Suppress Threshold to no less than 6,000.

Conservative Suppress Threshold: Conservative operators SHOULD configure the Suppress Threshold to no less than 12,000.

Calculate But Do Not Damp: Implementations MAY have a test mode where the operator could see the results of a particular configuration without actually damping any prefixes. This will allow for fine tuning of parameters without losing reachability.

7. Security Considerations

It is well known that an attacker can generate false flapping to cause a victim's prefix(es) to be damped.

As the recommendations merely change parameters to more conservative values, there should be no increase in risk.

In fact, the parameter change to more conservative values should slightly mitigate the false flap attack.

8. IANA Considerations

This document has no IANA Considerations.

9. Acknowledgments

Nate Kushman initiated this work some years ago. Seiichi Kawamura and Erik Muller contributed useful suggestions.

10. References

10.1. Normative References

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

[RFC2439] Villamizar, C., Chandra, R., and R. Govindan, "BGP Route Flap Damping", RFC 2439, November 1998.

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

[mao2002] Mao, Z. M., Govidan, R., Varghese, G., and Katz, R., "Route Flap Damping Excacerbates Internet Routing Convergence", In Proceedings of SIGCOMM , August 2002, <http://www.acm.org/sigcomm/sigcomm2002/papers/routedampening.pdf>.

[pelsser2011] Pelsser, C., Maennel, O., Mohapatra, P., Bush, R., and Patel, K., "Route Flap Damping Made Usable", Passive and Active Measurement (PAM), March 2011, <http://archive.psg.com/110103.pam-rfd.pdf>.

[ripe378] Panigl, P. and Smith, P., "RIPE Routing Working Group Recommendations On Route-flap Damping", 2006, <http://www.ripe.net/ripe/docs/ripe-378>.

10.2. Informative References

[huston2006] Huston, G., "BGP Extreme Routing Noise", RIPE 52 , 2006, <http://meetings.ripe.net/ripe-52/presentations/ripe52-plenary-bgp-review.pdf>.

[ripe178] Barber, T., Doran, S., Karrenberg, D., Panigl, C., and

Schmitz, J., "RIPE Routing-WG Recommendation for Coordinated Route-flap Damping Parameters", 2001, <http://www.ripe.net/ripe/docs/ripe-178>.

Authors' Addresses

Cristel Pelsser
Internet Initiative Japan, Inc.
Jinbocho Mitsui Buiding, 1-105
Kanda-Jinbocho, Chiyoda-ku, Tokyo 101-0051
JP

Phone: +81 3 5205 6464
Email: cristel@iij.ad.jp

Randy Bush
Internet Initiative Japan, Inc.
5147 Crystal Springs
Bainbridge Island, Washington 98110
US

Phone: +1 206 780 0431 x1
Email: randy@psg.com

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

Email: keyupate@cisco.com

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

Email: pmohapat@cisco.com

Olaf Maennel
Loughborough University
Department of Computer Science - N.2.03
Loughborough
UK

Phone: +44 115 714 0042
Email: o@maennel.net

Network working group
Internet Draft
Intended status: Standards Track
Expires: September 2011

Q. Zeng
J. Dong
Huawei Technologies
J. Heitz
Ericsson Inc.
K. Patel
Cisco Systems
R. Shakir
C&W
Z. Huang
China Telecom

March 7, 2011

One-time Address-Prefix Based Outbound Route Filter for BGP-4

draft-zeng-idr-one-time-prefix-orf-00.txt

Abstract

This document defines a new Outbound Router Filter (ORF) type for BGP, termed "One-time Address Prefix Outbound Route Filter", which would allow a BGP speaker to send to its BGP peer a route refresh request with a set of address-prefix-based filters to make the peer re-advertise only the specific routes matching the filters to the speaker. This ORF-type enables a BGP speaker to replay or recover some specific "problematic" routes without requiring its peer to re-advertise the whole Adj-RIB-Out of a specific address family, which makes the trouble shooting operation (such as packets tracking) more efficient and reduces the impact on network stability. This filter does not change the outbound route filters on BGP peers and should only be used for one-time filtering.

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."

This Internet-Draft will expire on September 7, 2011.

Conventions used in this document

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

Table of Contents

1. Introduction

The Outbound Route Filtering Capability defined in [RFC5291] provides a mechanism for a BGP speaker to send to its BGP peer a set

of Outbound Route Filters (ORFs) that can be used by its peer to filter its outbound routing updates to the speaker.

During some network maintenance, BGP speaker only needs to retrieve some specific "problematic" routes from its peer if the routes are possibly lost or contain some problematic attributes for some reason, but send ROUTE-REFRESH will lead to the peer re-advertising its whole Adj-RIB-Out. Such large numbers of updates include a lot of unnecessary routes which would make trouble shooting operation (such as packets tracking) more difficult, and is a waste of processing resources and bandwidth. With the increase of IPV6 deployment, this problem could be more significant. Even configured with ORF mechanism as defined in [RFC5291], on receipt of a ROUTE-REFRESH message, the peer will re-advertise all the routes matching current outbound route filters, i.e., the whole Adj-Rib-Out for this BGP speaker. Since in this case the BGP speaker does not want to change the outbound route filters on its peer, this problem cannot be solved by current ORF mechanism.

This document defines a new Outbound Router Filter (ORF) type for BGP, termed "One-time Address Prefix Outbound Route Filter", which would allow a BGP speaker to send to its BGP peer a route refresh request with a set of address-prefix-based filters to make the peer re-advertise only the specific routes matching the filters to the speaker. This ORF-type enables a BGP speaker to replay or recover some specific "problematic" routes without requiring its peer to re-advertise the whole Adj-RIB-Out of specific address family, which makes the trouble shooting operation (such as packets tracking) more efficient and reduces the impact on network stability. This filter does not change the outbound route filters on BGP peers and should only be used for one-time filtering.

Consider the following scenario: In an Inter-AS environment, if ASBR-A received a malformed UPDATE from ASBR-B and treated it as withdraw. For Operator-A, the log on the ASBR-A was not enough to judge whether the UPDATE was incorrectly sent by ASBR-B or incorrectly processed by ASBR-A. A good method is to replay and debug the packets. One-time Prefix ORF is a low impact way to refresh the UPDATE.

2. One-time Address Prefix ORF-Type

This document defines a new ORF type: One-time Address Prefix ORF.

In the following description, the sending speaker sends a one-time

ORF request and the receiving speaker receives it and sends back the routes to satisfy the request.

As specified in the [RFC5291], an ORF entry is a tuple of the form <AFI/SAFI, ORF-Type, Action, Match, ORF-value> an ORF consists of one or more ORF entries that have a common AFI/SAFI and ORF-Type. An ORF is identified by <AFI/SAFI, ORF-Type>.

The format of One-time Address Prefix ORF-Type entry is the same as the encoding of Address Prefix ORF in [RFC5292], the specific fields are defined as follows:

Since the semantics of this new ORF-Type is always "one-time filtering" and has no impact on existing ORFs, the Action field MUST be ignored.

The matching rules of the One-time Address Prefix ORF are the same as defined in Address-Prefix-Based ORF [RFC-5292].

The ORF entries of this type are used as one-time filters that MUST not change any previously installed ORF entry on the receiving speaker.

3. Operation

The capability negotiation of <AFI/SAFI, One-time Address Prefix ORF> MUST NOT delay the advertisement of routes with this AFI/SAFI.

The received One-time Address Prefix ORF entries SHOULD only be used for one-time route filtering and MUST NOT be saved locally. The received One-time Address Prefix ORF entries MUST NOT modify the outbound route filters on the receiving speaker (either locally configured or received from the sending speaker through ORF).

On receipt of ROUTE-REFRESH message with One-time Address Prefix ORF entries, the receiving speaker SHOULD re-advertise to the sending speaker the routes from the Adj-RIB-Out associated with the sending speaker which pass the entries carried in the One-time Address Prefix ORF as well as the locally saved ORFs (if any) received from the sending speaker.

Since different processing orders may lead to different results, the One-time ORFs and the regular ORFs SHOULD not be encoded in one route-refresh message.

During the period when the receiving speaker is sending updates to satisfy the One-time ORF request, it may experience other routing

activity that will require it to send updates unrelated to the One-time ORF request. It is permitted to send these updates before it has completed sending the One-time ORF related updates.

Similarly, if a route that passes the One-time ORF has already been sent and the receiving speaker experiences routing activity that changes this route and the receiving speaker has not yet sent all routes to satisfy the One-time ORF request, it is permitted to send the changed route immediately.

Details about how to interoperate when both One-time ORF Capability and the Enhanced Route Refresh Capability as described in [Enhanced-Refresh] are enabled will be discussed in the next version.

4. Security Considerations

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

5. IANA Considerations

This document specifies a new Outbound Route Filtering (ORF) type, One-time Address-Prefix ORF. The value of the ORF-type needs to be assigned by the IANA.

6. Acknowledgments

The authors would like to thank Enke Chen, Susan Hares, Haibo Wang, Jiawei Dong, Yaqun Xiao and Mach Chen for their valuable suggestions and comments to this document.

7. References

7.1. Normative References

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

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

[RFC5291] Chen, E. and Y. Rekhter, "Outbound Route Filtering Capability for BGP-4", RFC 5291, August 2008.

[RFC5292] Chen, E. and S. Sangli, "Address-Prefix-Based Outbound Route Filter for BGP-4", RFC 5292, August 2008.

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

[RFC4020] Kompella, K. and A. Zinin, "Early IANA Allocation of Standards Track Code Points", BCP 100, RFC 4020, February 2005.

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

7.2. Informative References

[Enhanced-Refresh] K. Patel, E. Chen and B. Venkatachalapathy, "Enhanced Route Refresh Capability for BGP-4", draft-keyur-bgp-enhanced-route-refresh-01.txt, Ocotober 2010

Authors' Addresses

Qing Zeng
Huawei Technologies Co.,Ltd.
Huawei Building, No.3 Xinxi Rd.,
Hai-Dian District
Beijing, 100085
P.R. China

Email: zengqing@huawei.com

Jie Dong
Huawei Technologies Co.,Ltd.
Huawei Building, No.3 Xinxi Rd.,
Hai-Dian District
Beijing, 100085
P.R. China

Email: jie.dong@huawei.com

Jakob Heitz
Ericsson Inc.
100 Headquarters Drive
San Jose CA 95134
USA

Email: jakob.heitz@ericsson.com

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

Email: keyupate@cisco.com

Rob Shakir
Cable&Wireless Worldwide

Email: rob.shakir@cw.com

ZhiLan Huang

China Telecom
109 West Zhongshan Ave,
Tianhe District, Guanghou, 510630, P.R.C

Email: huangzl@gsta.com