Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 12, 2014

C. Camilo Cardona
P. Pierre Francois
IMDEA Networks
S. Ray
K. Patel
P. Paolo Lucente
Cisco Systems
P. Mohapatra
Cumulus Networks
July 11, 2013

BGP Path Marking
draft-bgp-path-marking-00

Abstract

The potential advertisement of non-best paths by a BGP speaker supporting the add-path or the best-external extensions makes it difficult for other BGP speakers to identify the paths that have been selected as best by those who advertise them. This information is required for proper operation of some applications. Towards that end, this document proposes marking the paths using extended communities that encode the path type.

Table of Contents

1. Introduction

When there are multiple paths for a given address prefix, BGP chooses one of the paths as the "best-path" according to the best-path selection rules prescribed in [RFC4271] and installs the best-path in its forwarding table. Classically, each BGP speaker advertises only

the best-path to its peers. So when a BGP speaker receives a path from one of its peers, it is assured that the path is used by the peer for forwarding and all other peers have received the same path from this peer. This leads to consistent routing in a BGP network.

The classical advertisement rule of sending only the best-path does not convey the full routing state of a destination present on a BGP speaker to its peers.

- In order to improve link bandwidth utilization, most BGP implementations choose additional paths, that satisfy certain conditions, as "multi-path", and install them in the forwarding table. Incoming packets for that destination are load-balanced across the best-path and the multi-path(s). I.e., there may be paths installed in the forwarding table that are not advertised to the peers.

- When an Autonomous System (AS) deploys a route-reflector ([RFC4456]) instead of using full IBGP mesh, the BGP speakers receive only the route reflector's best-path and therefore lose information about the best-paths of other IBGP peers.

- If an IBGP path is chosen as the best-path by a non-route-reflector BGP speaker, then the best-path is not sent to its IBGP peers. Thus the IBGP peers learn nothing from this BGP speaker even though it might have other EBGP paths for that destination.

- Even when a BGP speaker selects an EBGP path as the best-path and advertises it to its peers, it may have additional EBGP paths for the destination. Should those paths be advertised a priori, they could be used by the peers in the event of loss of reachability of the best-path resulting in faster convergence.

There are extensions to the classical BGP advertisement rule to provide additional information about the routing state of a destination. A BGP speaker supporting the best-external [I-D.ietf-idr-best-external] extension sends its best external path to its IBGP peers when the best-path is an IBGP path. A BGP speaker supporting the add-path [I-D.ietf-idr-add-paths] extension advertises multiple paths for a given address prefix.

With best-external or add-path extensions in use, when a BGP speaker receives a path from a peer, that path may not be the best-path, or it may not be installed in the peer's forwarding table. In some scenarios, knowledge of the path type - i.e., whether the path is the best-path, or whether the path is installed in the forwarding table - is essential.

For instance, in a typical dual-homed VPN in primary-backup configuration, the backup path is created by advertising the best-external path from the backup PE with worse LOCAL_PREF. However, when the customer adds a site in another AS, the LOCAL_PREF information does not reach that site. As a result, data traffic coming from that site may incorrectly be forwarded over the backup link instead of the primary link.

Similarly when an add-path enabled peer receives multiple paths from a peer, it does not know which one among those paths is the best-path and which ones are installed in the forwarding table. An exogenous monitoring system, e.g., would require that information to properly tweak the policies on the router to effect desired forwarding optimization.

This draft proposes marking the advertised paths by an extended community, called Path Type community, that encodes the path type. The path type provides the necessary information to the BGP speakers about how the path is used by the sender when add-path or best-external extensions are in use.

2. The BGP Path Type Community

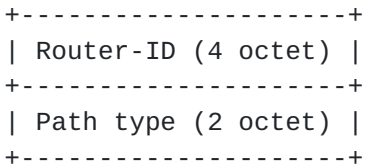
The BGP Path Type Community is an IPv4 Address Extended Community ([RFC4360]) defined as follows:

Type Field:

The value of the high-order octet of the extended Type Field is 0x01, which indicates that it is transitive. The value of low-order octet of the extended type field for this community is TBD.

Value Field:

The Value field contains two sub-fields, described below:



The Router-ID field contains the BGP identifier of the BGP speaker that adds the Path Type community to a path.

The Path type field contains a bitfield where each bit encodes a specific role of the path. Multiple bits may be set when a path is used in multiple roles.

Value	Path type
0x0000	Unknown
0x0001	Best-path
0x0002	Best-external path
0x0004	Multi-path
0x0008	Backup path
0x0010	Uninstalled path
0x0020	Unreachable path

Table 1: Path Type Values

The best-path is defined in [RFC4271] and the best-external path is defined in [I-D.ietf-idr-best-external].

A multi-path is not the best-path but installed in the forwarding table and used for forwarding packets. We use the convention that the best-path is not considered a multi-path.

A backup path is installed in the forwarding table, but it is not used for forwarding until all multipath(s) and the best-path become unreachable. Backup paths are used for fast convergence in the event of failures.

All other reachable paths are marked as 'Uninstalled'.

Lastly, all paths that are considered unreachable are marked as 'Unreachable'. Unreachable paths may be sent only in special cases (such as to a monitoring application).

3. Rules

- A BGP speaker MAY add the Path Type community to an originated path.

- When a BGP speaker receives a path from a peer and propagates it without changing the NEXT_HOP to self:

  * If the path contained a Path Type community, it MUST be retained in the propagated path.

    * If the path did not contain a Path Type community, the speaker MAY add a Path Type community with 'Unknown' value.

- When a path received from a peer is propagated after changing the NEXT_HOP to self:

    * If the path did not contain a Path Type community, the Path Type community indicating the path role MAY be added.

    * If the path contained a Path Type community:

        + If data traffic entering the router for the given destination may be forwarded over other paths (e.g., for doing load balancing), then the existing Path Type community MUST be removed. The BGP speaker MAY add its own Path Type community.

        + If data traffic entering the router for the given destination is forwarded only along the given path, then the existing Path Type community MAY be retained.

In all cases, when a BGP speaker adds its own Path Type community, it sets its own router-id in the community. Note that BGP router-id need not be unique across ASes.

The above rule-set prevents a route reflector from modifying the Path Type community set by its client (unless the route reflector is changing the NEXT_HOP to self).

When a peer is capable of sending only one path for a given address prefix and it sends the path without any Path Type community, the path MAY be considered as the best-path of the peer. In all other cases, a path without any Path Type community SHOULD be considered to have an 'Unknown' Path type.

A local policy might modify the above rules. For instance, if a monitoring application peers with a BGP speaker with add-path capability for the sole purpose of learning its paths and their types, then the speaker may always add its own Path Type community when it advertises the paths to that peer even if it does not change the NEXT_HOP to self. Such overriding policies should be used with caution if the advertised paths may impact forwarding decisions in the network.

4. Operational Considerations

If a speaker receives a path with a Path Type community with an invalid combination of bits (e.g., both 'Multi-path' and 'Backup'

bits are set), the path MUST NOT be considered invalid. Such error cases SHOULD be logged through other means.

An implementation SHOULD provide a configurable option for the user to indicate whether a path should be readvertised when its type is changed. If the user does not configure the option, the BGP speaker MUST NOT readvertise a path just to update its Path Type community (e.g., when a path type changes from 'Multi-path' to 'Uninstalled' due to a change in IGP metric).

An implementation SHOULD provide a configurable option for removing Path Type communities from paths that are advertised to untrusted peers.

An implementation SHOULD mark all paths for a given address prefix consistently. If one of the paths is marked, then all other paths SHOULD be marked.

An implementation MAY modify its best-path selection algorithm to take path type information into account. For instance, paths with type 'Best-path' MAY be preferred over paths of other types. Similarly, paths of type 'Best-external' MAY be considered ineligible for being a multipath.

5. Applications

In this section, we illustrate some applications that benefit from the Path Type community proposed in this draft.

5.1. Avoiding suboptimal routing in Inter-AS VPN

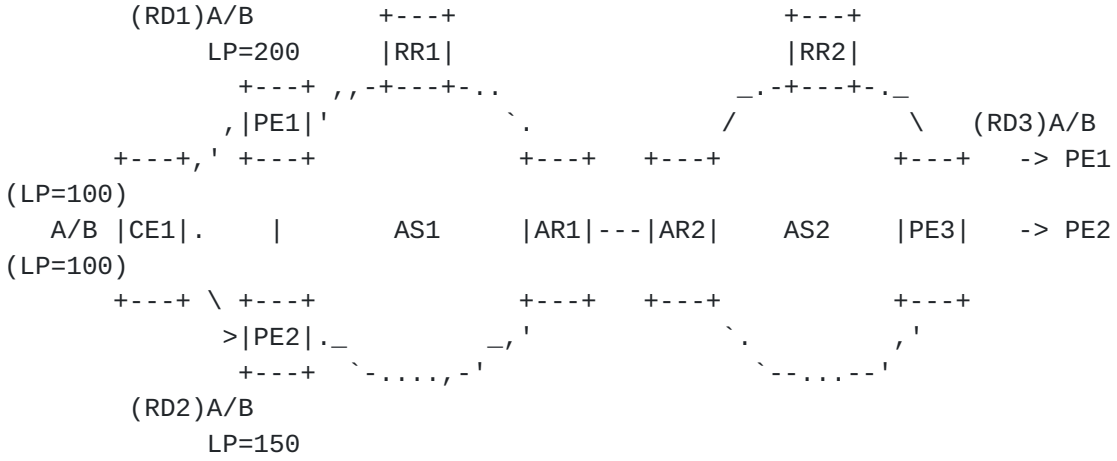


Figure 1: Inter-AS VPN scenario

Figure 1 depicts an L3VPN network that spans two ASes: AS1 and AS2.

The ASes may be connected using either Option-B or Option-C

techniques [RFC4364]. A customer site with equipment CE1 is dual-homed in AS1, connected to PE1 and PE2. For prefix A/B, the customer prefers to use the link between CE1 and PE1. This routing preference is expressed by setting the LOCAL_PREF of the prefix advertised by PE1 to a higher value than that of the prefix advertised by PE2. This causes PE2 to use PE1's route as the best-path and its own EBGP path becomes the best-external path. PE2 is configured to advertise its best-external path. Therefore, both PEs continue to advertise their own EBGP path. The provider uses unique route-distinguishers for its VPNs. So PE1 and PE2 advertises different VPN prefixes: (RD1)A/B and (RD2)A/B. Both these prefixes are advertised to PE3 in AS2. PE3 imports both paths to its own VPN with route-distinguisher RD3.

Existing behavior:

> Since LOCAL_PREF is not sent across AS boundary, both paths on PE3 have the default LOCAL_PREF of 100. As a result the best-path selection on PE3 may boil down to tie breaking steps and the path towards PE2, which is the best-external path, may be chosen. Alternately, the path from PE2 may be chosen as the multipath and may be used for load-balancing. Therefore, some or all data traffic entering PE3 would reach CE1 via PE2, which is not what the customer desired.

Behavior with Path Type Community:

> When PE2 advertises its path, it adds the best-external Path Type community. This community is preserved across AS boundary. If option C is used, then RR1 or RR2 does not change the NEXT_HOP and hence the community is preserved according to the rule-set (Section 3). If option B is used, then the community reaches AR1 since RR1 does not change the NEXT_HOP. At AR1, (RD2)A/B has only one path and forwarding traffic entering AR1 from AR2 for this destination (determined by the outer label) would use this path. Therefore, AR1 retains the Path Type community set by PE2. The same applies to AR2. So at PE3, the path to PE2 has the best-external Path Type community and therefore PE3 can choose to not use this path for forwarding.

If the best-path algorithm takes the Path Type community values into account, it eliminates the need for setting LOCAL_PREF to deprefer the bext-external path even within a single AS. This simplifies the network design and management.

Instead of using Path Type communities, it is possible to use policies on the border routers (AR1 and AR2 for option B, or RR1 and

RR2 for option C) to recreate the LOCAL_PREF in AS2 (e.g., by matching on the RD and the prefix). However, the recreated LOCAL_PREF may interfere with the local policies set in AS2 (e.g., if there are other paths in AS2 for A/B that the customer wants to use as secondary paths). In addition, such policies are error-prone and complex to manage, especially when the customer is allowed to change the primary/backup relationships between PE1 and PE2 on its own. The standardized mechanism of Path Type community is free from such drawbacks.

5.2. Monitoring applications

A modern Service Provider (SP) network may contain thousands of BGP routers. For planning, proper engineering and operation of a backbone, it is a good practice to continuously monitor the routers' states and perhaps keep a history. Many Network Management Systems (NMS) establish IBGP sessions with BGP speakers to collect the paths the speaker has. When the speaker supports add-path (or best-external), the NMS receives non-best-paths. There are also monitoring protocols such as BMP [I-D.ietf-grow-bmp] that similarly receives all paths from a speaker.

When an NMS receives multiple paths for a destination, it is important for its operation to know which path is the best-path, which paths are installed in forwarding table, which path is used as a backup, etc. The NMS system may run the best-path algorithm on those paths on its own. However, its information, especially on IGP metric, local policies, etc., may be incomplete and hence its own calculations may not match that of the router's. It is also noted that even if the NMS system collected additional information to run the best-path algorithm from the point-of-view of the router, it would have to do so for every router in the network, which would impose a very high computational burden on the NMS.

When Path Type community is in use, the router provides the required information directly, thus avoiding computational load on the NMS as well as potential discrepancies between the point-of-view of the router and that of the NMS.

5.3. SDN applications

Similar to the monitoring applications, a "Software Defined Networking" application monitors the routing state and based on it, may change the policies on the router, or inject additional paths, to influence the forwarding. When a BGP speaker supports Path Type communities and add-path, an SDN application can simply peer with the router to receive its routing state in real-time even if the router does not provide vendor-specific APIs for doing the same.

5.4. Selective Best-path

When the classical BGP advertisement rule is followed, all paths a BGP speaker considers for best-path are already installed in the forwarding table of the peer. However, when add-path, or best-external extensions are used, that no longer holds. If the BGP speakers support the Path Type communities, then the classical behavior can be reinstated by considering only those paths in the best-path algorithm that are marked as best-path or multi-path. Detailed discussions on the rules and benefits of such an approach are outside the scope of this draft.

6. IANA Considerations

Section 2 defines an IPv4 Address specific transitive extended community called the Path Type extended community. IANA is requested to assign a sub-type value for the Path Type extended community. The last 2 bytes of the value field of the Path Type extended community contains a bitfield that encodes the type of the advertised path. IANA is expected to maintain a registry for these bits. Section 2 defines 6 of those bits. The rest of the bits are to be assigned by IANA using the "IETF Consensus" policy defined in [RFC2434].

7. Security Considerations

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

8. Contributors

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

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

9. Acknowledgments

We would like to thank Bruno Decraene for his feedback on this work.

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.

[RFC2434] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 2434, October 1998.

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

10.2. Informative References

[I-D.ietf-grow-bmp] Scudder, J., Fernando, R., and S. Stuart, "BGP Monitoring Protocol", draft-ietf-grow-bmp-07 (work in progress), October 2012.

[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-08 (work in progress), December 2012.

[I-D.ietf-idr-best-external] Marques, P., Fernando, R., Chen, E., Mohapatra, P., and H. Gredler, "Advertisement of the best external route in BGP", draft-ietf-idr-best-external-05 (work in progress), January 2012.

[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.

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

Authors' Addresses

Camilo Cardona
IMDEA Networks
Avenida del Mar Mediterraneo
Leganes 28919
Spain

Email: juancamilo.cardona@imdea.org

Pierre Francois
IMDEA Networks
Avenida del Mar Mediterraneo
Leganes 28919
Spain

Email: pierre.francois@imdea.org

Saikat Ray
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: sairay@cisco.com

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

Email: keyupate@cisco.com

Paolo Lucente
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: plucente@cisco.com

Pradosh Mohapatra
Cumulus Networks
140 C. Whisman Rd.
Mountain View, CA 94041
USA

Email: pmohapat@cumulusnetworks.com