Inter-Domain Routing
Internet-Draft
Intended status: Standards Track
Expires: June 24, 2011

P. Marques, Ed.
R. White
Cisco Systems, Inc.

December 21, 2010

Topology-based aggregation
draft-marques-idr-aggregate-00

Abstract

This document defines a mechanism which allows more-specific IP address prefixes to be aggregated when they are topologically equivalent or less preferable than a less-specific advertisement.

It is designed to allow multi-homed sites to use "Provider Aggregatable" (PA) addresses and obtain both redundancy and local traffic optimizations when using multiple service providers.

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 June 24, 2011.

Copyright Notice

Table of Contents

1. Introduction

With the existing inter-domain routing functionality as defined by RFC 4271 [RFC4271], multi-homed sites feel compelled to advertise their individual prefixes to the entire Internet in order to achieve the desired reliability and traffic-engineering behavior.

Multi-homed sites typically advertise "Provider Independent" (PI) prefixes. An alternative approach would be for "Provider Aggregatable" (PA) space to be used along with a set of procedures that allow for route advertisements to be aggregated. This option must retain the functionality that is provided today by PI advertisements.

One assumption made here is that renumbering of a multi-homed site is economically feasible given the increased usage of dynamic host configuration protocols and/or network address translation.

This document is being written at a time when IP addresses are becoming scarse. It is difficult to predict whether Internet address allocation and assignment policies will drift torwards the use of PI space in order to achieve more efficient allocation. Or whether scarcity will make it harder to obtain PI space.

In the latter case, this document define an approach that would allow multi-homed sites a method for using PA addresses without bumping into address space filtering rules that may be in place to limit the growth of the internet table size.

In order to meet the requirements stated above for multi-home site routing, the following is proposed:

- The routing advertisement must be taken out of "Provider Aggregatable" (PA) space.

- The routing advertisement must be leaked through one or more alternate providers, other than the one owning the PA space.

- These more-specific route advertisements shall be automatically aggregated, depending on the network topology.

- If the multi-homed site becomes disconnected from the owner of the address space it must be possible to unsuppress the most-specific adververtisement.

In order to provide topology-dependent aggregation, this document defines a new BGP path attribute, AGGREGATE_INFO, which defines a BGP prefix as being a more specific of a given aggregate prefix. A BGP

speaker that receives such a prefix MUST compare the received prefix with the specified aggregate, if present in its Loc-RIB. The standard path selection algorithm is applied between the paths of the more-specific prefix and the best-path of the aggregate. If the best-path of the aggregate is preferable, the more-specific prefix should be considered as "Inactive". It SHOULD NOT be further re-advertised into External BGP sessions. It MAY BE re-advertised into Internal BGP sessions, if the path-selection criteria between the aggregate and more-specific justifies it.

Conceptually, the aggregate prefix conveys implicit path information that applies to the delegated more-specifics. Path selection occurs between the explicit paths that are present in the routing system and these implicit paths represented by the aggregates.

The AGGREGATE_INFO attribute contains an operational status field. This field is used to indicate the status of the connectivity between the multi-homed site and the provider owning the aggregate. It can be used in a situation of failure in which the customer becomes detached from the service provider originating the PA aggregate.

When the operational status denotes connectivity failure this will result on the more-specific being unsuppressed and attracting traffic through the failover paths. The operational status is used explicitly in order to inform downstreams that the more-specific is temporary and will be removed from the routing system once connectivity is restored.

The operational status field uses three colors: green, yellow and red. Green means full connectivity. Red means no connectivity. Yellow informs the routing system that while the site itself has no direct connectivity to the primary provider, it believes that there is sufficient redundant connectivity in the network that its prefix is still reachable through it.

2. Topology-based aggregation

The intent of this extension is to achieve the same semantics as "Provider Independent" (PI) advertisements, while removing the more specifics from the BGP routing table in locations of the network where the aggregate provides equal or better service to the IP destination prefix in question.

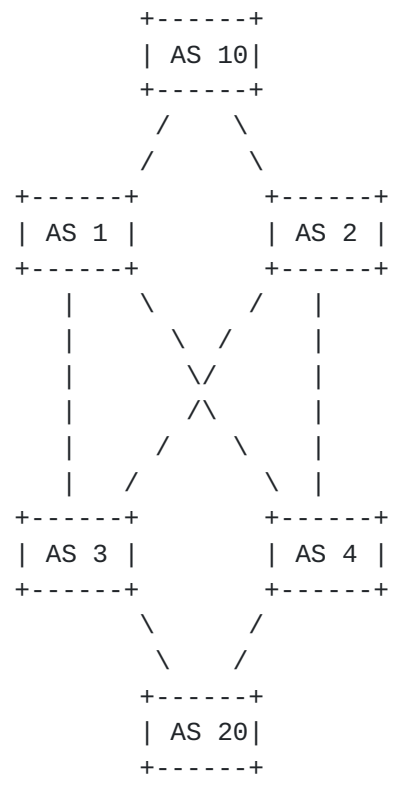

Figure 1

Figure 1 contains an example of the usage of the BGP AGGREGATE_INFO attribute. AS 10 in the example above has been delegated "10.0.1/24" prefix by AS 1. Using this extension, it will advertise the prefix into AS 2, which will likely prefer a customer router over a peer route to AS 1. When AS 2 re-advertises the more-specific "10.0.1/24" to its peers, AS 3 and 4 in this example, the peers will compare the more-specific to the "10.0/16" aggregate received from AS 1. Typically AS 3 will prefer the aggregate (as-path: "1", length 1) over the more-specific (as-path: "2 10", length: 2). When this is the case, the more-specific will be suppressed and no longer propagated in the network. If, for any reason, AS 1 becomes disconnected from AS 3, the more-specific route to "10.0.1/24" will become active again, achieving the required failover protection.

From a traffic-engineering perspective, the more-specific is selected in locations in the network where AS 10 is topologically closer than AS 1.

In the example described above, the aggregate route may have a shorter as-path than the equivalent PI prefix that is in use currently. A PI prefix that is injected by the customer AS (AS 10) would be advertised to AS 3 with an as-path of "1 10". In order to

provide multi-homed sites with equivalent functionality as it is available to them using PI space, the AGGREGATE_INFO BGP attribute allows the originator to specify an AS_PATH attribute to be appended with the path contained in the aggregate route. This allows the customer AS (AS 10) to indicate to AS 3 that the attribute comparison should be performed between the explicitly advertised more-specific with as-path "2 10" and an implicit more-specific path with an as-path of "1 10". This implicit path is derived from the aggregate prefix.

3. BGP AGGREGATE_INFO attribute

The BGP AGGREGATE_INFO attribute is a well-known, transitive attribute with Type Code 129. It contains a list of one or more aggregate target elements. Each aggregate target contains a mandatory part, with the operational status field followed by a route prefix. That may be followed by additional BGP PATH attributes that apply to the specified aggregate target prefix.

The operational status is encoded as a 1-octect field with the following values:

Value	Color	Description
0	Red	No connectivity between customer and provider
1	Yellow	Direct connectivity unavailable
2	Green	Connectivity fully operational

The prefix is encoded as a 2 byte AFI [RFC1700] value, followed by a variable length prefix encoded as a 1 byte prefix-length in bits and the prefix itself padded to a byte boundary. This is the same encoding used for NLRI in BGP UPDATE messages.

The prefix contained in the AGGREGATE_INFO attribute SHOULD be a less-specific prefix containing all the NLRI specified in the BGP UPDATE message that includes this attribute.

Following the route prefix, the encoding allows for one of more BGP path attributes using the encoding specified the BGP [RFC4271] protocol specification. An implementation MAY choose to include an AS_PATH attribute in this optional element.

When an AS_PATH attribute is contained inside an AGGREGATE_INFO attribute, the path segments that it contains shall be appended to the AS_PATH of the implicit path represented by the aggregate prefix.

This implicit path is then compared with the best path of NLRI prefix(es) included in the UPDATE message containing this attribute.

Example encoding for prefix 10.0/16, as-path "10":

Attr Flags = 0x40, Attr Code = 0x81, Attr Length = 0x0e

OpStatus=0x2, AFI = 0x00 0x01, Prefix Length = 0x10, Prefix Data = 0x0a 0x00

Attr Flags = 0x40, Attr Code = 0x02, Attr Length = 0x04, Data = 0x02 0x01 0x00 0x0a

In the example given above, an AS_PATH segment of "10" in the aggregate-info attribute and an aggregate path with an AS_PATH of "1" would result in a as-path of "1 10", of length 2.

When multiple aggregate target prefixes are present in a AGGREGATE_INFO attribute, the most significant prefix present in the Loc-Rib is used to generate the implicit path used in path selection.

Multiple targets can be used when prefix assignment and delegation happens at more than one level.

As an example, a provider X may have a /16 out of which it delegates to Y a specific /22 block. Y then allocates a /24 to a specific multi-homed customer Z. If Y itself is using aggregation its prefix may be suppressed. Where Z to originate a route with a single aggregation-target (/22), that prefix would not be aggregated in regions of the network where the /22 had itself be aggregated.

For this mechanism to behave as expected one would have to ensure that if Y's prefix has been suppress then Z's has also been suppressed. Otherwise if Z's prefix is present, its aggregation target of Y will be ignored.

Since this condition cannot be guaranteed, the protocol allows the originator of the more-specific prefix (Z) to include multiple aggregation targets (Y and X) in its route advertisement. Whenever Y is present in the Loc-Rib of BGP speaker, Y is used as source of the implicit aggregation path. Otherwise X is used if present.

The choice of explicitly listing the aggregation targets rather than automatically deriving the parent is designed to avoid situations in which the less-specific is being artificially generated such as, for instance, the default route.

4. BGP extension deployment

BGP speakers that support the extensions described in this document SHALL use the Capability Advertisement [RFC5492] BGP extension to advertise that support to its BGP peers.

Compliant implementations should advertise the BGP Capability Code TBD. The capability data should contain a 1-byte value which is interpreted as the version of this specification. It should contain the value 1.

When a BGP route is placed in the Out-RIB for a given external BGP peer and the peer in question doesn't support this capability, if the path in the Loc-Rib contains the AGGREGATE_INFO attribute this should result in the prefix being suppressed. If a previous path was advertised to this peer that path shall be withdrawn.

If the peer in question is an internal BGP peer which doesn't support this capability an implementation MAY choose to replace this attribute with the NO_EXPORT [RFC1997] BGP community attribute, rather than suppress the path.

This mechanism assures that a path that originated with an AGGREGATE_INFO attribute is not used by a router without being compared to the respective aggregate. This is intended to facilitate the incremental deployment of this functionality.

5. Path selection criteria

A BGP implementation shall run its path selection algorithm unmodified between all the paths for a given prefix. If the selected best-path contains the BGP AGGREGATE_INFO attribute, this path shall be compared with the best-path of the aggregate prefix indicated by the attribute in question.

The AGGREGATE_INFO attribute represents an implicit path for the more-specific prefix (the NLRI containing that attribute). The BGP path attributes of this implicit prefix are the attributes of the best-path of the aggregate prefix. If the AGGREGATE_INFO contains an optional AS_PATH attribute, the AS_PATH segments in that attribute shall be appended to the AS_PATH of the aggregate prefix best-path before comparison.

When the Operational Status of the specified aggregate target is "Red" the corresponding implicit path is considered to be unreachable. When the Operational Status is "Yellow" the originating AS of the aggregate target prefix MUST treat the implicit path as

unreachable also and use the more-specific. Autonomous-systems further downstream MAY choose whether to ignore or use the aggregation information.

The "Yellow" state represents that the originator of the prefix believes that there is a path between the primary and backup providers for the site such that this path always prefers the more-specific advertisement. This is often the case if both providers have a direct peering relationship.

When comparing the more-specific path with its implicit path (represented by the aggregate), the following changes to the standard path selection algorithm should be taken into account:

- The Origin attributes of both paths are not comparable. This is step b) in the path selection algorithm and should be bypassed.

- If the paths in question are equal upto step d) of path selection algorithm, if both paths are EBGP paths, the less-specific (aggregate) should be preferred. This replaces the step in path selection where the oldest EBGP path is preferred [RFC5004].

- If both paths are iBGP paths, the less-specific (aggregate) should be preferred in case where the paths are equal up-to the router-id comparison step of path selection.

When the aggregate path is considered to be preferable over the more-specific, the more-specific should be considered inactive and should not be installed in the FIB or subsequently advertised to other peers.

6. Network deployment

The objective of this document is to provide multi-homed sites with the resilience to failures and limited traffic-engineering capabilities without the need to recurse to PI advertisements.

Instead of using a PI prefix, a multi-homed site can choose to address its network with PA prefix from one service provider which it then advertises through a secondary provider. Or it may choose to dual address its hosts and/or NAT appliances.

In order for a multi-homed site to achieve the required resilience it should be allowed by other service providers to inject the more-specifics that have been delegated to it with the BGP AGGREGATE_INFO attribute.

The AGGREGATE_INFO attribute should only be added to a BGP path by the originator of the route advertisement. This rule is intended to ensure that there aren't instances of the same BGP path information flowing through the Internet routing system with and without the specified attribute.

In order to maintain the loop free properties of BGP one must ensure that when suppressing a more-specific this doesn't result in traffic being forwarded in a way which results in a loop.

For this to occur, the following conditions would be necessary:

A transit AS (X) prefers the more-specific route.

Another AS (Y) receives both aggregate and more-specific from X and prefers the former.

Y is in the transit path for the more-specific.

The last condition cannot occur since Y, by definition prefers the aggregate path and will not advertise the more-specific.

7. Acknowledgements

There have been several prior proposals to reduce routing information used in muli-homing scenarios. For instance, using BGP communities [I-D.white-bounded-longest-match] and AS hops [I-D.ietf-idr-as-hopcount].

The current document builds upon the previous work and proposes the use of standard BGP path selection using both implicit and explicit paths in order limit information to parts of the network where it is useful.

8. Contributors

Central parts of the protocol operation where defined by Robert Raszuk and Keyur Patel. Russ White, Enke Chen, Dave Meyer and Vince Fuller provided essential input in the early stages of the proposal.

9. IANA Considerations

This memo requests IANA to allocate a BGP attribute type code value, for the BGP aggregate-info attribute defined herein. It also requests IANA to allocate a Capability Code according to the

procedures defined in RFC 5492 [RFC5492].

10. Security Considerations

The BGP aggregate-info attribute in itself doesn't create a new security threat. This attribute can only lead to the route being suppressed.

The presence of more-specifics in the routing system makes a stronger case for the usefulness of performing origin authentication of route advertisements.

11. References

11.1. Normative References

[RFC1700] Reynolds, J. and J. Postel, "Assigned Numbers", RFC 1700, October 1994.

[RFC1997] Chandrasekeran, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, August 1996.

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

[RFC5004] Chen, E. and S. Sangli, "Avoid BGP Best Path Transitions from One External to Another", RFC 5004, September 2007.

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

11.2. Informative References

[I-D.ietf-idr-as-hopcount]
Li, T., "The AS_HOPCOUNT Path Attribute", draft-ietf-idr-as-hopcount-00 (work in progress), December 2005.

[I-D.white-bounded-longest-match]
Hares, S., "Bounding Longer Routes to Remove TE", draft-white-bounded-longest-match-02 (work in progress), July 2008.

Authors' Addresses

Pedro Marques (editor)
Cisco Systems, Inc.
170 W. Tasman Dr.
San Jose, CA 94040
US

Phone: +1 408 853 1193
Email: roque@cisco.com

Russ White
Cisco Systems, Inc.
7025 Kit Creek Road
Research Triangle Park, NC 27709
US

Email: riw@cisco.com