Network Working Group
Internet-Draft
Expires: April 3, 2006

T. Li, Ed.
Portola Networks, Inc.
R. Fernando, Ed.
Cisco Systems, Inc.
J. Abley, Ed.
Internet Systems Consortium

September 30, 2005

The AS_HOPCOUNT Path Attribute
draft-li-as-hopcount-03.txt

Status of this Memo

Copyright Notice

Abstract

This document describes the AS hopcount path attribute for BGP. This is an optional, transitive path attribute that is designed to help limit the distribution of routing information in the Internet.

By default, prefixes advertised into the BGP mesh are distributed

freely, and if not blocked by policy will propagate globally. This is harmful to the scalability of the routing subsystem since information that only has a local effect on routing will cause state creation throughout the default-free zone. This attribute can be attached to a particular path to limit its scope to a subset of the Internet.

Table of Contents

1. Requirements notation

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

2. Introduction

A prefix that is injected into BGP [RFC1771] will propagate throughout the mesh of all BGP speakers unless it is explicitly blocked by policy configuration. This behavior is necessary for the correct operation of BGP, but has some unfortunate interactions with current operational procedures. Currently, it is beneficial in some cases to inject longer prefixes into BGP to control the flow of traffic headed towards a particular destination. These longer prefixes may be advertised in addition to an aggregate, even when the aggregate advertisement is sufficient for basic reachability. This particular application is known as "inter-domain traffic engineering" and is a well-known phenomenon that is contributing to growth in the size of the global routing table[RFC3221]. The mechanism proposed here allows the propagation of those longer prefixes to be limited, allowing some traffic engineering problems to be solved without such global implications.

Another application of this mechanism is concerned with the distribution of services across the Internet using anycast. Allowing an anycast address advertisement to be limited to a subset of ASes in the network can help control the scope of the anycast service area.

3. Inter-Domain Traffic Engineering

To perform traffic engineering, a multi-homed site advertises its prefix to all of its neighbors and then also advertises more specific prefixes to a subset of its neighbors. The longest match lookup algorithm then causes traffic for the more specific prefixes to prefer the subset of neighbors with the more specific.

Figure 1 shows an example of traffic engineering and its impact on the network. The multi-homed site (A) has a primary provider (C) and a secondary provider (B). It has a prefix, Y, that provides reachability to all of A, and advertises this to both B and C. In addition, due to the internal topology of end-site A, it wishes that all incoming traffic to subset X of its site enter through provider B. To accomplish this, A advertises the more specific prefix, X, to provider B. Longest match again causes traffic to prefer X over Y if the destination of the traffic is within X.

Assuming that there are no policy boundaries involved, BGP will propagate both of these prefixes A and X throughout the entire AS-level topology. This includes distant providers such as H, F and G. Unfortunately, this adds to the amount of overhead in the routing subsystem. The problem to be solved is to reduce this overhead and thereby improve the scalability of the routing of the Internet.

```
     ,--------------.   ,--------------.   ,--------------.
     |    Tier 2    +---+    Tier 2    +---+    Tier 3    |
     |  Provider H  |   |  Provider E  |   |  Provider F  |
     `--------------'   `-+---------+--+   `--------------'
                         /          |
                        /           |
    ,------------------+---.   ,----+---------.   ,-------------.
    |       Tier 1         +---+    Tier 1    |   |   Tier 1    |
    |  Primary Provider C  |   |  Provider D  +---+ Provider G  |
    `--------+-----------+-'   `-------+------'   `-------------'
             |            \            |
             |Y            \           |
    ,--------+------.     ,-+----------+-----------.
    |  Multi-homed  +-----+         Tier 2         |
    |     site A    |Y,X  |  Secondary Provider B  |
    `---------------'     `------------------------'
```

The longer prefix X traverse a core and then coincides with the less-specific, covering prefix Y.

Figure 1

3.1. Traffic Engineering on a Diet

What is needed is one or more mechanisms that an AS can use to distribute its more specific routing information to a subset of the network that exceeds its immediate neighboring ASes and yet is also significantly less than the global BGP mesh. The solution space for this is fully unbounded, as the limits that a source AS may wish to apply to its more specific routes could be a fairly complicated manifestation of its routing policies. One can imagine a policy that restricts more specifics to ASes that only have prime AS numbers, for example.

We already have one mechanism for performing this type of function. The BGP NO_EXPORT community string attribute [RFC1997] can be attached to more specific prefixes. This will cause the more specifics not to be advertised past the immediate neighboring AS. This is effective at helping to prevent more specific prefixes from becoming global, but it is extremely limited in that the more specific prefixes can only propagate to adjacent ASes.

Referring again to our example, A can advertise X with NO_EXPORT to provider B. However, this will cause provider B not to advertise X to the remainder of the network, and providers C, D, and G will not have the longer prefixes and will thus send all of A's traffic via provider C. This is not what A hoped to accomplish with advertising a longer prefix and demonstrates why this NO_EXPORT mechanism is not sufficiently flexible.

Instead of attempting to provide an infinitely flexible and complicated mechanism for controlling the distribution of prefixes, we propose a single, coarse control mechanism. This coarse mechanism will provide a limited amount of control but at a very low cost and address most of the evils associated with performing traffic engineering through route distribution.

We observe that traffic engineering via longer prefixes is only effective when the longer prefixes have a different next hop from the less specific prefix. Thus, past the point where the next hops become identical, the longer prefix no longer provides any value whatsoever. We also observe that the many of the domains that are practising traffic engineering are connected to multiple highly-reliable providers, typically at Tier 1 or Tier 2. Most traffic for a multi-homed site will traverse these core providers and thus, the traffic will encounter a longer prefix in one of these network. If one looks one AS hop past these domains, it is very likely that the longer prefixes and the site aggregate are using the same next hop, and thus the longer prefixes have stopped providing value after they have traveresed the center of the network.

We can see this clearly in our example. Provider F sees that both prefix X and prefix Y will lead all traffic through provider E. There is no point in F carrying and propagating the more specific prefix X. Similarly, providers G and H need not carry prefix X.

3.2. AS_HOPCOUNT as Control

To accomplish this, we propose to add information that will limit the radius of propagation of more specific prefixes. If we attach a count of the ASes that may be traversed by the more specific prefix, we gain much of the control that we hope to achieve. For example, if prefix X is advertised with hopcount 1, then only provider B has the information and we get an effect that is identical to NO_EXPORT. If prefix X is advertised with hopcount 2, then only B, C and D will carry it. This is an interesting compromise as traffic for X will now flow consistently through provider B, as desired.

However, this is not identical to fully distributing X. Consider, for example that provider E in this circumstance will not receive prefix X and is likely to prefer provider C for all A destinations. This causes traffic for X to flow from E to C to B. If provider E did have prefix X, it may choose to prefer provider D instead, resulting in a different path. This second result can be achieved by increasing the hopcount to 3, but this has the unfortunate effect that provider G would also receive prefix X.

Thus, AS_HOPCOUNT is an extremely lightweight mechanism, and achieves a great deal of control. It is easy to imagine more complicated control mechanisms, such IDRP [IDRP] distribution lists, but we currently find that the complexity of such a mechanism is simply not warranted.

3.3. AS_HOPCOUNT and NO_EXPORT

Further control can be achieved by considering the implications of using both AS_HOPCOUNT and NO_EXPORT simultaneously. Since NO_EXPORT is widely deployed, understood by almost all implementations, and since AS_HOPCOUNT is not deployed, we can make use of the overlap in their semantics to provide a powerful transition mechanism.

Systems that receive NLRI with only the AS_HOPCOUNT attribute but which do not implement AS_HOPCOUNT will ignore the attribute. This will provide the current, existing behavior and the NLRI will propagate according to normal BGP rules.

Systems that receive NLRI with both an AS_HOPCOUNT and NO_EXPORT and which do implement AS_HOPCOUNT will ignore the NO_EXPORT community and propagate the NLRI.

Systems that receive NLRI with both an AS_HOPCOUNT and NO_EXPORT but which do not implement AS_HOPCOUNT will recognize and operate according to NO_EXPORT semantics. This will cause them not to forward the NLRI to other ASes.

Thus, an AS that chooses to attach the AS_HOPCOUNT attribute can control how their NLRI will be processed by other ASes. If the NLRI should be dropped by ASes that do not support AS_HOPCOUNT, then NO_EXPORT can be attached. If the NLRI should propagate by default, then NO_EXPORT should not be attached.

4. Anycast Service Distribution

A growing number of services are being distributed using anycast, by advertising a route which covers one or more addresses for a service which is provided autonomously at multiple locations.

For some services, it is useful to restrict the peak possible service load, to avoid overloading local connectivity or service infrastructure capabilities; it may be a better failure mode for service to be retained only for a small community of surrounding networks than for a single node to fail under a global load of queries.

Although to some degree this policy can be accomplished through negotiation and judicious use of NO_EXPORT without AS_HOPCOUNT, the AS_HOPCOUNT attribute provides a more flexible and reliable mechanism.

5. The AS_HOPCOUNT Attribute

The AS_HOPCOUNT attribute is a transitive optional BGP path attribute, with Type Code XXXX. The AS_HOPCOUNT attribute has a fixed length of 5 octets. The first octet is an unsigned number that is the hopcount of the associated paths. The second thru fifth octest are the AS number of the AS that attached the AS_HOPCOUNT attribute to the NLRI.

5.1. Operations

A BGP speaker attaching the AS_HOPCOUNT attribute to an NLRI MUST encode its AS number in the second thru fifth octets. The encoding is described in [4B AS]. This information is intended to aid debugging in the case where the AS_HOPCOUNT attribute is added by an AS other than the originator of the NLRI.

A BGP speaker receiving a route with an associated AS_HOPCOUNT attribute from an EBGP neighbor MUST examine the value of the attribute. If the attribute value is zero, the path MUST be ignored without further processing. If the attribute value is non-zero, then the BGP speaker MAY process the path.

When a BGP speaker propagates a route with an associated AS_HOPCOUNT attribute, which it has learned from another BGP speaker's UPDATE message, it MUST modify the route's AS_HOPCOUNT attribute based on the location of the BGP speaker to which the route will be sent:

a. When a given BGP speaker advertises the route to an internal peer, the advertising speaker SHALL NOT modify the AS_HOPCOUNT attribute associated with the route.

b. If the BGP speaker chooses to advertise the route to an external peer, then the BGP speaker MUST advertise an AS_HOPCOUNT attribute of one less than the value received.

If a BGP speaker receives a route with both the AS_HOPCOUNT attribute and the NO_EXPORT community string attribute, then the normal semantics of NO_EXPORT do not apply and the route should be processed as if NO_EXPORT was not present.

BGP requires that a BGP speaker that advertises a less specific prefix, but not a more specific prefix that it is using, must advertise the less specific prefix with the ATOMIC_AGGREGATE attribute. BGP speakers that do not advertise a more specific prefix based on the AS_HOPCOUNT must comply with this rule and advertise the less specific prefixes with the ATOMIC_AGGREGATE attribute. To help ensure compliance with this, sites that choose to advertise the

AS_HOPCOUNT path attribute should advertise the ATOMIC_AGGREGATE attribute on all less specific covering prefixes.

5.2. Proxy Control

An AS may attach the AS_HOPCOUNT attribute to a path that it has received from another system. This is a form of proxy aggregation and may result in routing behaviors that the origin of the path did not intend. Further, if the overlapping prefixes are not advertised with the ATOMIC_AGGREGATE attribute, adding the AS_HOPCOUNT attribute may cause defective implementations to advertise incorrect paths. Before adding the AS_HOPCOUNT attribute an AS must carefully consider the risks and consequences outlined here.

6. Security Considerations

This new BGP attribute creates no new security issues. For it to be used, it must be attached to a BGP route. If the router is forging a route, then this attribute limits the extent of the damage caused by the forgery. If a router attaches this prefix to a route, then it could have just as easily have used normal policy mechanisms to filter out the route.

7. IANA Considerations

IANA is hereby requested to allocate a code point from the BGP path attribute Type Code space for the AS_HOPCOUNT path attribute. Please replace 'XXXX' in the text above with the newly allocated code point value.

8. Acknowledgements

The editors would like to acknowledge that they are not the original initiators of this concept. Over the years, many similar proposals have come our way, and we had hoped that self-discipline would cause this type of mechanism to be unnecessary. We were overly optimistic.

The names of those who originally proposed this are now lost to the mists of time. This should rightfully be their document. We would like to thank them for the opportunity to steward their concept to fruition.

9. References

[4B AS]   Vohra, Q. and E. Chen, "BGP support for Four-octet AS Number Space", Sept. 2005, <http://www.ietf.org/internet-drafts/draft-ietf-idr-as4bytes-11.txt>.

[IDRP]   ISO/IEC, "Information Processing Systems - Telecommunications and Information Exchange between Systems - Protocol for Exchange of Inter-domain Routeing Information among Intermediate Systems to Support Forwarding of ISO 8473 PDUs", IS 10747, 1993, <http://www.acm.org/sigcomm/standards/iso_stds/IDRP/10747.TXT>.

[RFC1771] Rekhter, Y. and T. Li, "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, March 1995.

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

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

[RFC3221] Huston, G., "Commentary on Inter-Domain Routing in the Internet", RFC 3221, December 2001.

Authors' Addresses

T. Li (editor)
Portola Networks, Inc.

Email: tony.li@tony.li

R. Fernando (editor)
Cisco Systems, Inc.
170 W. Tasman Dr.
San Jose, CA 95134-1706
US

Phone: +1 408 525-1253
Email: rex@cisco.com

J. Abley (editor)
Internet Systems Consortium
950 Charter Street
Redwood City, CA 94023
US

Phone: +1 650 423 1317
Email: jabley@isc.org

Intellectual Property Statement

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.

Disclaimer of Validity

Copyright Statement

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.