Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: July 8, 2007

T. Li, Ed.
Cisco Systems, Inc.
R. Fernando, Ed.
Juniper Networks, Inc.
J. Abley, Ed.
Afilias

January 4, 2007

The AS_PATHLIMIT Path Attribute
draft-ietf-idr-as-pathlimit-03

Status of this Memo

Copyright Notice

Abstract

This document describes the 'AS path limit' (AS_PATHLIMIT) 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 graph 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 [RFC4271] will propagate throughout the graph 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.

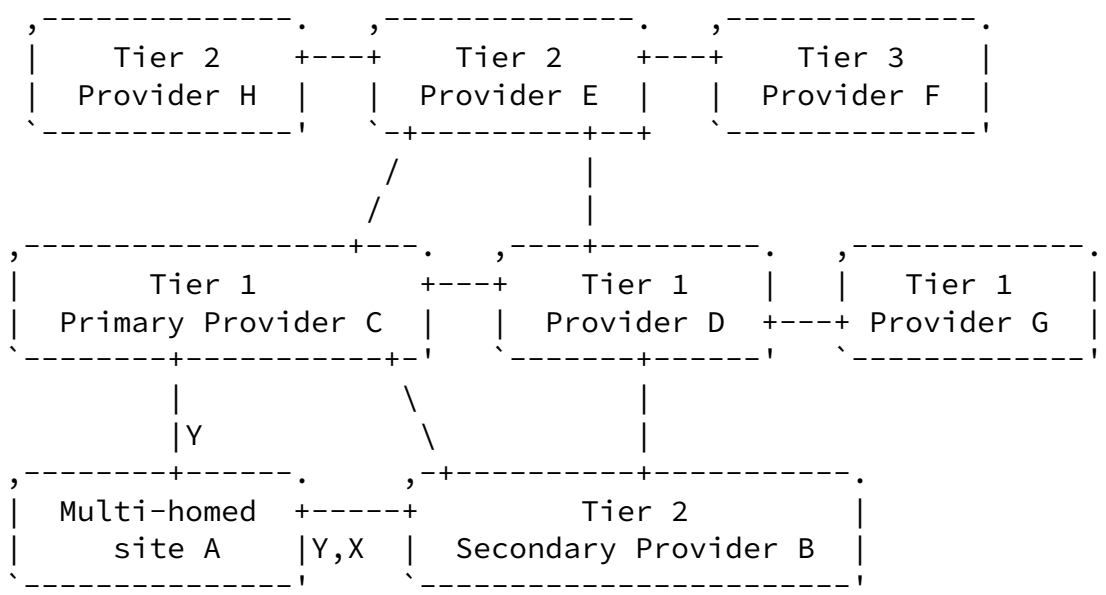
3. Inter-Domain Traffic Engineering

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

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 X and Y 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.



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 neighbouring ASes and yet is also significantly less than the global BGP graph. The solution space for this is 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 neighbouring 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.

Some ASes have created a further mechanism wherein a prefix that is given a particular community will have NO_EXPORT attached to that prefix when the prefix is propagated to a specific AS. This is not a generally deployed mechanism, but is used by some ASes as another means of scope control.

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, scope control mechanism. This coarse mechanism will provide a limited amount of control 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 prefixes provide no value whatsoever. We also observe that most traffic ends up traversing a subset of the network operated by a relatively small number of large market-dominant providers, joined by settlement-free interconnects. If one

looks one AS hop past this subset of the network, it is likely that the longer prefixes and the site aggregate are using the same next hop, and thus the longer prefixes have stopped providing value.

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_PATHLIMIT 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. We propose the creation of a new path attribute that will carry an upper bound on the number of ASes found the AS_PATH attribute. This new path attribute will be called the 'path limit' or AS_PATHLIMIT. For example, if prefix X is advertised with path limit 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 path limit 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 path limit to 3, but this has the unfortunate effect that provider G would also receive prefix X.

Thus, AS_PATHLIMIT is an extremely lightweight mechanism, and achieves a great deal of control. It is easy to imagine more complicated control mechanisms, such IDRP [ISO.10747.1993] distribution lists, but we currently feel that the complexity of such a mechanism is simply not warranted.

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_PATHLIMIT, the AS_PATHLIMIT attribute provides a more flexible and reliable mechanism.

5. The AS_PATHLIMIT Attribute

The AS_PATHLIMIT attribute is a transitive optional BGP path attribute, with Type Code 21. The AS_PATHLIMIT attribute has a fixed length of 5 octets. The first octet is an unsigned number that is the upper bound on the number of ASes in the AS_PATH attribute of the associated paths. One octet suffices because the TTL field of the IP header ensures that only one octet's worth of ASes can ever be traversed. The second thru fifth octets are the AS number of the AS that attached the AS_PATHLIMIT attribute to the NLRI.

5.1. Operations

A BGP speaker attaching the AS_PATHLIMIT attribute to an NLRI MUST encode its AS number in the second thru fifth octets. The encoding is described in [I-D.ietf-idr-as4bytes]. This information is intended to aid debugging in the case where the AS_PATHLIMIT attribute is added by an AS other than the originator of the NLRI.

A BGP speaker sending a route with an associated AS_PATHLIMIT attribute to an EBGP neighbour MUST examine the value of the attribute and the associated AS_PATH to be advertised. If the number of ASes found in the AS_PATH exceeds the AS_PATHLIMIT value, then the route SHOULD NOT be sent.

For the purposes of this attribute, private AS numbers [RFC1930] and confederation AS members [RFC3065] found in the AS_PATH are not counted. AS numbers found within an AS_SET are not counted and an entire AS_SET is counted as a single AS. Each instance of an AS number that appears multiple times in an AS_PATH is counted.

If the AS_PATHLIMIT attribute is attached to a prefix by a private AS, then when the prefix is advertised outside of the parent AS, the AS number contained in the AS_PATHLIMIT attribute should be replaced by the AS number of the parent AS.

Similarly, if the AS_PATHLIMIT attribute is attached to a prefix by a member of a confederation, then when the prefix is advertised outside of the confederation boundary, then the AS number of the confederation member inside of the AS_PATHLIMIT attribute should be replaced by the confederation's AS number.

A BGP speaker receiving a route with an associated AS_PATHLIMIT attribute from an EBGP neighbour MUST examine the value of the attribute. If the number of ASes in the AS_PATH exceeds the value of the AS_PATHLIMIT attribute, then the route MUST be ignored without

further processing.

When a BGP speaker propagates a route with an associated AS_PATHLIMIT attribute, which it has learned from another BGP speaker's UPDATE message, it MUST NOT modify the route's AS_PATHLIMIT attribute. It may remove the AS_PATHLIMIT in its entirety. It may also attach a new AS_PATHLIMIT attribute that encodes its own AS number.

To ensure loop prevention, BGP requires that all aggregate routes with AS paths that omit any AS number from the AS_PATHs being aggregated to be originated with the ATOMIC_AGGREGATE attribute. To help ensure compliance with this, sites that choose to advertise the AS_PATHLIMIT path attribute SHOULD advertise the ATOMIC_AGGREGATE on all less specific covering prefixes as well as the more specific prefixes.

5.2. Proxy Control

An AS may attach the AS_PATHLIMIT attribute to a route that it has received from another AS. This is a form of proxy aggregation and may result in routing behaviors that the origin of the route did not intend. Further, if the overlapping prefixes are not advertised with the ATOMIC_AGGREGATE attribute, adding the AS_PATHLIMIT attribute may cause defective implementations to advertise incorrect paths. Before adding the AS_PATHLIMIT 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. This may be used by attackers to limit the scope and thus the visibility of their attacks. Presently, the same approach can be applied with the use of the NO_EXPORT community, but just as the AS_PATHLIMIT attribute gives network operators more granularity in the distribution of prefixes, it also gives attackers more granularity in their attacks. If a router fraudulently attaches the AS_PATHLIMIT attribute to a route, then it could have just as easily have used normal policy mechanisms to filter out the route completely. Thus, the AS_PATHLIMIT attribute does not enable new attacks, but it does give an attacker the ability to create more subtle attacks that only affect a subset of the entire network.

7. IANA Considerations

This document has no actions for IANA. IANA has already allocated a code point for the AS_PATHLIMIT attribute under the Early IANA Allocation process.

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

9.1. Normative References

[I-D.ietf-idr-as4bytes]
Vohra, Q. and E. Chen, "BGP Support for Four-octet AS Number Space", draft-ietf-idr-as4bytes-12 (work in progress), November 2005.

[RFC1930] Hawkinson, J. and T. Bates, "Guidelines for creation, selection, and registration of an Autonomous System (AS)",

BCP 6, RFC 1930, March 1996.

[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.

[RFC3065] Traina, P., McPherson, D., and J. Scudder, "Autonomous System Confederations for BGP", RFC 3065, February 2001.

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

9.2. Informative References

[ISO.10747.1993]
International Organization for Standardization, "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", ISO Standard 10747, 1993.

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

Authors' Addresses

T. Li (editor)
Cisco Systems, Inc.

425 East Tasman Drive
San Jose, CA 95134

Phone: +1 408 525 1254
Email: tli@cisco.com

R. Fernando (editor)
Juniper Networks, Inc.
1194 North Mathilda Avenue
Sunnyvale, CA 94089-1206
US

Phone: +1 888 586 4737
Email: rex@juniper.net

J. Ableу (editor)
Afilias Canada, Inc.
4141 Yonge Street, Suite 204
Toronto, ON M2P 2A8
CA

Phone: +1 416 673 4176
Email: jabley@ca.afilias.info

Full Copyright Statement

Intellectual Property

Acknowledgment

Funding for the RFC Editor function is provided by the IETF Administrative Support Activity (IASA).