

Network Working Group
Internet Draft
Expiration Date: December 2003
File Name: [draft-grow-bounded-longest-match-04.txt](#)

T. Hardie
R. White
June 2003

Bounding Longest Match Considered
[draft-grow-bounded-longest-match-04.txt](#)

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of [Section 10 of RFC2026](#).

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. Internet Drafts may be updated, replaced, or obsoleted by other documents at any time. It is not appropriate to use Internet Drafts as reference material or to cite them other than as a "working draft" or "work in progress".

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>.

Abstract

Some ASes currently use length-based filters to manage the size of the routing table they use and propagate. This draft explores an alternative to length-based filters which allows for more automatic configuration and which provides for better redundancy.

Rather than use a filter, this draft proposes a method of modifying the BGP longest match algorithm by setting a bound on the prefix lengths eligible for preference. A bound would operate on long prefixes when covering route announcements are available; in certain circumstances it would cause a router to prefer an aggregate over a more specific route announcement.

1. Motivation

Modifying longest match would limit the rate of growth in the routing table seen by many BGP speakers. The current rate of growth and the time to convergence represent threats to the stability to the Internet. In the short term, the IETF is considering efforts to curb these threats while new routing paradigms that attack the fundamental limitations of path vector protocols are developed and deployed.

A number of the practical efforts to limit the rate of growth of the routing table have focused on filter policies, arguing that aggressive filtering will return the Internet to a state in which provider aggregates are a majority of the routes in the routing table[3]. This draft proposes an approach along those same lines, but using a bound on the longest match algorithm rather than a filter policy. The authors believe that this approach can produce a similar (though not identical) effect while retaining full reachability and allowing multi-homing non-transit networks to achieve the main goals which have motivated their becoming independent ASes.

2. Proposed Enhancements

Two enhancements are proposed by this draft: a new community, and a new way of handling overlapping prefixes received from an external peer.

As each prefix is received by a BGP speaker from an external peer, it would be evaluated in the light of other prefixes already received. If two prefixes overlap in space (such as 192.168.0.0/16 and 192.168.1.0/24), the longer prefix would be marked with the NO_EXPORT community, and the local preference set to a very high number so that it would always win in any best path computations within the autonomous system. The longer prefix may also be marked with a new community, NO_INSTALL.

2.1. The NO_INSTALL Community

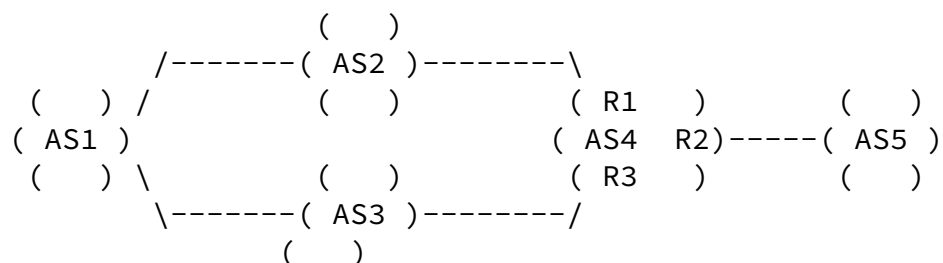
An optional optimization to bounding longer prefixes by marking them with a high Local Preference and the NO_EXPORT community is to also mark them with a new, non-transitive, optional community, NO_INSTALL. The effect of this community would be for any BGP speaker receiving a

prefix with this community set to treat the prefix normally in the BGP bestpath computation, and to forward bestpaths marked as NO_INSTALL to iBGP peers, but to simply not install such prefixes in the local routing table.

This would result in saving some small amount of memory for each prefix not installed in the RIB, and the local forwarding tables built from the RIB. If there are enough prefixes thus marked, the memory and computation savings could be significant. BGP speakers which receive a prefix marked with NO_INSTALL, and do not understand this community, may simply ignore the community.

[2.2.](#) Example of Bounding the Longer Prefix

Assume the following configuration of autonomous systems:



- o AS1 is advertising 192.168.1.0/24 to both AS2 and AS3.
- o AS2 is advertising both 192.168.1.0/24 and 192.168.0.0/16 into AS4
- o AS3 is advertising 192.168.1.0/24 into AS4

When R1 receives both the 192.168.1.0/24 and the 192.168.0.0/16 prefixes, it will mark the 192.168.1.0/24 as NO_EXPORT, and set the local preference to a high value, as described in the section Setting the Local Preference, below, and will then propagate this through AS4.

R3 will receive the longer prefix from AS3, and the iBGP prefix with the high local preference with NO_EXPORT set. Given it does not see

the overlapping prefix, it will compare the default (lower) local preference of the externally learned route with the higher local preference set by the AS2/AS4 border router, and will not advertise the 192.168.1.0/24 prefix into AS4 at all.

R3 border router may also, on detecting the overlap, mark the longer prefix with the NO_INSTALL community.

If the link between AS1 and AS2 fails, the longer length prefix will be withdrawn from AS2, and thus the peering point between AS2 and AS4 will no longer have an overlapping set of prefixes. Within AS4, the border router which peers with AS2 will cease advertising the

192.168.1.0/24 prefix, which allows the AS3/AS4 border router to begin advertising it into AS4, and through AS4 into AS5, restoring connectivity to AS1.

[2.3.](#) Setting the Local Preference

Since there could be multiple points at which an autonomous system may receive the same pair of overlapping prefixes, there must be some way to ensure that one of the longer prefixes wins in the BGP decision algorithm consistently. In practice, this means that each BGP speaker which receives an overlapping set of routes should set the local preference on the set of longer prefixes so there won't be two longer prefixes with matching local preferences.

The easiest way to ensure this within an autonomous system is to set the local preference for longer prefixes based on some unique number assigned to each BGP speaker. Given the router ID and the local preference are both 32 bit numbers, an ideal solution appears to be to simply set the local preference to the router ID of the BGP speaker. The primary problem with this is that in some cases, the router ID of the device may be lower than some standard Local Preference, perhaps even lower than a standard Local Preference used by default throughout a network.

To alleviate this problem, the local preference of longer prefixes which overlap with shorter prefixes should be set to the router ID of the BGP speaker, and then the high order bit of the Local Preference should be set, so the setting will be guaranteed to be at least above

64,000.

[2.4.](#) Implications for Load Sharing

Since the goal of this proposal is to reduce the number of paths stored within local tables, and to reduce the amount of information passed through to neighboring autonomous systems, the implementation of this draft as described above would have a negative impact on the ability to load share between multiple paths to the same destination.

[3.](#) An Alternative Implementation Using ADD PATH

An implementation which supports [\[ADD-PATH\]](#) could optionally use this capability to block the overlapping prefixes into neighboring autonomous systems, and preserve local load sharing.

- o Any router receiving a pair of overlapped routes from its external peers would mark the longer prefix with the NO_EXPORT community, and propagate the overlapped prefix using the technique described in [\[ADD-PATH\]](#).
- o Any router receiving a pair of overlapped routes, with the longer prefix learned from an external peer, and the shorter prefix learned from an internal peer, would mark the longer prefix with the NO_EXPORT community, and propagate the prefix normally to its internal peers.

[4.](#) Benefits and Risks

The benefits and risks associated with this proposal are discussed in

the sections below.

[4.1.](#) Advantages to the Service Provider

AS4, in each of the situations, reduces the number of prefixes carried through the autonomous system by the number of longer prefixes that overlap with aggregates of those prefixes. While one copy of the prefix continues to be carried through the autonomous system, this entry can be marked with the optional NO_INSTALL community, so it is not placed in the forwarding table, nor is it propagated outside the autonomous system.

AS5 receives one prefix instead of two (or possibly more).

[4.2.](#) Advantages to the Customer

In this case, the customer is represented as AS1. The customer will continue to receive some amount of traffic over both peering sessions, and dual homing through two Service Providers is still effective. If the customer's primary link fails, the alternate link through AS3 will take over receiving all inbound traffic automatically. With most other schemes presented to this point, the customer loses all impact of dual-homing into the Internet, unless both connections are through one Service Provider.

[4.3.](#) Advantages to the Internet

Beyond the second AS hop, aggregation is preserved in all cases. While this would not reduce the backbone routing table by the dramatic amounts that other methods might, the advantages to the community are great, and at greatly reduced risk to customers.

[4.4.](#) Implications for Router processing

This proposal clearly adds to the work which needs to be done during overall BGP processing. Because a check needs to be done for both covered and covering routes, some part of this work is required for routes of lengths on either side of the bound. Should this become common, however, the rate of growth in the number of routes should be

smaller and a balance should be struck between the extra processing per route and the number of routes.

[4.5.](#) Implications for Traffic engineering

The implementation of a bound risks magnifying or removing the effect of certain widely deployed traffic engineering methods. If, for example, an AS chose to prepend its own route to an announcement in order to alter the preference for that route, a BGP neighbor using a bounded longest match might now see that route as eligible for discard in favor of an aggregate. While it is fairly easy to code around that particular problem, to avoid this class of problems it might be preferable to allow this to apply to specific AS Sets as well as to all BGP neighbors.

[4.6.](#) Implications for Propagation delay and increased convergence time.

If the route to the AS providing the route to the aggregate should be lost, the more-specific must propagate into the ASes which had formerly heard only the aggregate. This increases convergence time and may create situations in which reachability is temporarily compromised. Unlike the filter case, however, normal BGP behavior should restore reachability without changes to the router configuration. There is also a risk that during a pathological event the increased processing required by this change will degrade propagation times during those events. This depends on both the speed of specific implementations and the character of the topology.

[5.](#) Security Considerations

This document presumes that the implementation of bounded longest match is a knob inside a router config. Since the use of the knob affects route announcements not originating within the router's AS or its direct neighbors, the new behavior may result in surprises to the announcing AS. It is possible that this behavior might be considered a denial of service or mistaken for a denial of service by systems

designed to detect black-holing on behalf of the origin AS.

6. Acknowledgements

Cengiz Alaentinioglu, Alvaro Retana, Daniel Walton, Danny McPherson, and Barry Greene gave valuable comments on this draft. A number of colleagues also gave the author valuable comments on the white board markings that gave rise to this paper; among them are Lane Patterson, Ian Cooper, Gerd Besch, Bill Norton, Diarmuid Flynn, and Sean Donelan.

7. References

- [1] Huston, Geoff. <http://www.telstra.net/ops/bgp/index.html>
- [2] Ahuja, Abha. <http://www.merit.edu/~ahuja/ptomaine-bof/ahuja-ietf-ptomaine/index.htm>
- [3] Bush, Randy. Plenary, IETF 51. Eventually at:
<http://www.ietf.org/proceedings/01aug/>
- [ADD-PATH]
Walton, D, et al, "Advertisement of Multiple Paths in BGP," [draft-walton-bgp-add-paths-00.txt](#)

8. Authors' Addresses

Ted Hardie
Ted.Hardie@nominum.com

Russ White
Cisco Systems, Inc.
7025 Kit Creek Rd.
Research Triangle Park, NC 27709
EMail: riw@cisco.com