Global Routing Operations Working Group
Internet-Draft
Intended status: Informational
Expires: April 24, 2015

I. van Beijnum
Institute IMDEA Networks

October 21, 2014

Controlled IPv6 deaggregation by large organizations
draft-van-beijnum-grow-controlled-deagg-00

Abstract

The use of IPv6 addresses by large organizations doesn't fit the commonly used PA/PI dichotomy. Such organizations may hold a large address block which is deaggregated into subprefixes that are advertised by subunits of the organization. This document proposes a set of best practices to allow this deaggregation to be controlled through filtering so that on the one hand, the size of the IPv6 global routing table isn't unduly inflated, while on the other hand organizations that seek to deaggregate a large IPv6 address block don't see their reachability limited by remote filters.

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 April 24, 2015.

Table of Contents

1. Introduction

Generally, two classes of global unicast address prefixes are recognized: provider aggregatable (PA) and provider independent (PI). PA prefixes are the prefixes advertised into the global routing table by ISPs, covering the addresses used by multiple customers of that ISP. PI prefixes are the address blocks used by a single organization.

However, there is a third class of addresses: the addresses used by large organizations with subunites that independently connect to the internet. An example are multinational corporations. Another are national governments. Such organizations often desire a single IPv6 prefix so the addresses used by subunits are easily recognized as being part of the larger organization in firewalls and router filters. As such, many of these organizations become "enterprise LIRs" (local internet registries) at one or more of the five regional internet registries (RIRs) that distribute IP addresses. However, unlike regular LIRs (ISPs), they are not in the business of moving IP packets between locations, and as such different locations or subunits advertise deaggregates (subprefixes) of the organization's LIR PA prefix, often to different ISPs. This advertisement of deaggregates would be unexpected from regular LIRs, and as such, the deaggregates may be filtered.

Currently, the IPv6 global routing table is small and in no immediate danger of growing beyond what today's routers can handle. However, without some of the limitations that are present in IPv4, the IPv6 routing table could conceivably grow at a high rate for decades to come, and would then at some point become hard to manage.

This document proposes two mechanisms that will allow organizations that seek to deaggregate an enterprise LIR prefix to enjoy the same level of connectivity as users of PI and PA space while at the same time limiting the impact of this practice on the IPv6 global routing table. The first mechanism is the establishment of an "aggregate of last resort" (AoLR), the second mechanism is a set of communities that allow deaggregates to be filtered in some parts of a network without loss of reachability.

This document is meant to start a discussion. As such, it may be split into several documents, and/or the venue for discussion and eventual publication is subject to change.

2. The aggregate of last resort service

The assumption is that an enterprise LIR allocates addresses from a single block to different organizational subunits, and that these subunits advertise those smaller blocks to the ISPs they use to connect to the internet, where different subunits use different ISPs. For reasons of cost and routing efficiency it's not possible or desired to use an internal network between the subunits or locations to transport traffic to/from the internet from one organizational subunit to another.

One way to run such a network would be for the enterprise to advertise its aggregate in a small number of locations. The traffic is then delivered to those locations, and then from there sent back to an ISP that has a path to the subunit in question. However, this has the downside that traffic has to pass through one of the locations advertising the aggregate, using up additional bandwidth and possibly incurring long detours. For instance, if an organization advertises its prefix in Europe then a third party in the US that sends traffic to one of the organization's offices in the US may see its traffic cross the Atlantic twice.

The solution is to ask one or more ISPs to advertise the aggregate--preferably ones with a large geographic footprint. By default, networks hand over traffic to a remote network as soon as possible ("hot potato" routing), so in this case the traffic just has to flow to the closest location where the ISP in question has a presence. If that ISP then connects to an ISP serving the organizational subunit in question, the traffic can be handed over between the two ISPs at the nearest location where they interconnect.

This way, deaggregates only have to be carried by ISPs providing the aggregate of last resort service and the ISPs connecting subunits of the organization. Because the organization has customer - service

provider relationships with each, presumably those ISPs will not filter the deaggregates.

3. Geographic communities

BGP supports a community mechanism that allows a router to tag a prefix with additional information that may be interpreted by other routers. This document proposes a set of communities that encode the geographic origin of a deaggregated prefix. This allows network operators to filter prefixes that are covered by an aggregate. Additionally, such filtering may be applied selectively.

For instance, a network that operates in the APNIC region may want to filter out deaggregates originated in other regions, but allow the ones originated in the APNIC region. Or a North American network may want to carry European deaggregates only at the US East Coast, where it interconnects with European networks, and only carry Asian deaggregates at the US West Coast, where it interconnects with Asian networks.

An objection against encoding geographic information in the routing system is that topology doesn't follow geography. Strictly speaking, this is of course true. In theory, a user in Tokyo could connect to the internet in Madrid. In practice, this is is exceedingly rare. And in the case where this happens and BGP is in the position to make decisions, having this information available is even more useful than in in routine situations: when that user in Tokyo connects to the internet in Madrid and Hong Kong, users outside Europe would do well to avoid the route through Madrid. A geographic community would allow them to do exactly that.

4. Encoding of geographic information

There are currently two types of communities defined for BGP: the original community attribute ([RFC1997]), which encodes 32-bit values, and extended community attribute ([RFC4360]), which supports subtypes of various lengths. Regular communities are widely supported and are typically displayed in the form ddddd:ddddd, where ddddd are both 16-bit values displayed in decimal, such as 702:120.

Defining a new extended community subtype has the advantage that it would be possible to specify a new syntax and new semantics tailored to the needs of the new community, but the disadvantage is that it would take a lot of time for this to be implemented by router vendors. As such, geographical information will be encoded into a set of communities within the numbering space of the existing [RFC1997] system. Router vendors are encouraged to recognize these

communities and handle them appropriately as outlined later in this document.

There are many ways to encode geographic information, such as the ISO 3166-1 alpha-2 two-letter country code, the ITU E.164 one-to-three-digit international phone dialing numbers and the ISO 3166-1 three-digit numeric code. The only one that is well-known in numeric form are international phone dialing numbers. However, the size difference in population/area served between the different country codes (and area codes in the North American Numbering Plan) is very large, and the numbers don't lend themselves to easily identifying a geographic region bigger than a metropolitan area but smaller than a country.

To avoid these issues, this document specifies that geographic communities encode latitude/longitude information. This encoding avoids interpretation and contention. By rounding to whole degrees, a reasonable tradeoff between precision and location privacy is achieved.

A geographic community consists of two 16-bit values in decimal notation. In the first value, the least significant bits indicate north or south and east or west, respectively. In the second value, the upper two digits indicate the latitude and the lower three digits indicate the longitude, each rounded to the nearest degree. For example:

Berlin, DE; 52 deg 31 min N, 13 deg 23 min E:

xxxx1:53013

Chicago, US; 41 deg 50 min N, 87 deg 41 min W:

xxxx0:42088

Mumbai, IN; 18 deg 58 min N, 72 deg 49 min E:

xxxx1:19073

Rio de Janeiro, BR; 22 deg 54 min S, 43 deg 11 min W:

xxxx2:23043

Saint Petersburg, RU; 59 deg 57 min N, 30 deg 18 min E:

xxxx1:60030

Locations further than 64 degrees north or south are encoded differently: the upper two digits of the second community value encode the upper two digits of the longitude, the next two digits encode the latitude, and the last digit encodes the lower digit of the longitude:

Spitsbergen, NO; 78 deg 45 min N, 16 deg 00 min E:

xxxx1:00790

McMurdo Station, Antarctica; 77 deg 51 min S 166 deg 40 min E:

xxxx3:16787

This format is somewhat human-readable. However, router vendors are encouraged to recognize these communities and display the values as follows:

xxxx1:53013
53N13E

xxxx0:42088
42N88W

xxxx1:19073
19N73E

xxxx2:23043
23S43W

xxxx1:60030
60N30E

xxxx1:790
79N0E

xxxx3:16787
78S167E

Furthermore, it would be helpful if filters could specify areas in the form 53N3E-50NE8. (This encompasses the Netherlands in its entirety, although it also covers parts of the neighboring countries.)

Although they don't immediately serve the purpose of this draft, two additional forms of geographic communities are specified. This makes for three different sets of geographic communities:

Covered:
The presence of a geographic community of this type indicates that the prefix is covered by an aggregate and can therefore safely be filtered without loss of reachability. The location encoded in the community is the location of the ISP side of circuit that connects the site using the prefix to the internet. If an indication that the prefix is covered by an aggregate is

desired, but not the encoding of a location, then the community xxxx0:999 may be used.

Uncovered:
The presence of a geographic community of this type DOES NOT indicate that a covering aggregate is present. The location encoded in the community is the location of the ISP side of circuit that connects the site using the prefix to the internet and may be presented in order to facilitate best path selection.

Seen-at:
The presence of a geographic community of this type DOES NOT indicate that a covering aggregate is present. The location encoded in the community is a location where the prefix was seen. For instance, the location where a network learned the prefix over EBGP. Multiple instances of this type of geographical community may be present.

5. IANA considerations

IANA is requested to register the following 16-bit ranges of community values out of the subset of community value space that maps to private AS numbers:

Covered origin NW

Covered origin NE

Covered origin SW

Covered origin NE

Uncovered origin NW

Uncovered origin NE

Uncovered origin SW

Uncovered origin NE

Seen-at NW

Seen-at NE

Seen-at SW

Seen-at NE

6. Security considerations

It would be possible for any router along the AS path to rewrite a geographic community and claim a false geographic origin and/or falsely claim that a prefix is covered by an aggregate.

7. Contributors

None at this time.

8. Acknowledgements

None at this time.

9. References

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

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006.

Author's Address

Iljitsch van Beijnum
Institute IMDEA Networks
Avda. del Mar Mediterraneo, 22
Leganes, Madrid 28918
Spain

Email: iljitsch@muada.com