Multi6 I. van Beijnum
Internet-Draft June 30, 2003
Expires: December 29, 2003

Provider-Internal Aggregation based on Geography to Support Multihoming in IPv6
draft-van-beijnum-multi6-isp-int-aggr-01

Abstract

Current 6bone backbone routing guidelines prohibit traditional multihoming in IPv6, because current IPv4-style multihoming doesn't scale. This stands in the way of successful adoption of IPv6. The solution outlined in this memo proposes aggregating the routing information for multihomed destinations inside service provider networks based on geography to accomplish scalable multihoming in IPv6 using current protocols and implementations. This solution does not require network operators to increase the density of interconnection; nor does it require significant cooperation or simultaneous adoption.

1. Introduction

Current IPv4 and IPv6 interdomain routing operational practices depend heavily on aggregation in order to reach the necessary scalability. Current aggregation is exclusively service provider based: ISPs (Internet Service Providers) obtain blocks of address space from the Regional Internet Registries (RIRs) and assign their customers addresses from these blocks. Then they announce a single route for each block to other networks. This aggregation makes it possible for millions of

organizations to be connected to the internet while limiting the global routing table to only slightly more than a hundred thousand destination prefixes.

Unfortunately, provider-based aggregation doesn't work for networks connected to the internet over more than one connection ("multi-homed" networks). In the current IPv4 internet, multihoming is typically done by announcing a route for an independent address block to two or more ISPs. The address block may actually be part of a larger PA (provider aggregatable) block, but it must be visible in the global routing table independently from possible aggregates to make multihoming work under all circumstances. This makes it impossible for many millions of networks to multihome: the global routing table would grow beyond what routers can handle.

There are efforts underway to provide in IPv6 the failover and load balancing functionality present in current "IPv4 style" multihoming in different ways that wouldn't increase the size the global routing table. However, all these new multihoming solutions are still on the drawing board and need changes to protocols and implementations. In the mean time, the current 6Bone backbone routing guidelines [RFC2772] don't allow non-aggregated routes in the IPv6 global routing table and thereby make IPv4-style multihoming impossible.

This memo proposes new operational practices that will allow networks to handle a much larger global routing table, so multihoming in IPv6 can be made possible within a very short time frame. However, it is very important to note this isn't a perfect "one size fits all" solution that scales to huge numbers of multihomed networks without any pain or effort. (See the Limitations section later in this document.) But at least this mechanism makes multihoming possible almost immediately, without having to wait for protocols and implementations to be changed or even for network operators to reconfigure their networks. The latter can be done later, and on a per-network basis, as the size of the global routing table becomes problematic for individual networks. The idea is to make multihoming possible now, while providing networks with the means to control the

size of the routing table in their routers later as necessary.

After implementing the necessary filtering mechanisms, growth to several million multihomed networks world wide should be possible without much trouble. In theory, this mechanism can support many hundreds of millions multihomed networks, but this will be hard to accomplish in practice, so work on more advanced multihoming solutions should continue.

NOTE WELL

This mechanism does NOT require networks to announce geographical aggregate routes to anyone; those aggregates are only used internally. In this respect, the mechanism discussed here is very different from earlier geographical aggregation proposals.

The full routing information for all destinations connected to the internet is still present in each network (AS) that doesn't use a default route (in other words, is part of the default-free zone). It's just that this information is distributed over all the routers in the network so each router holds part of the information, rather than being replicated as is done today, where each router holds a full copy of the information.

2. How It Works

To make multihoming (as we know it today) possible, individual routes must be present in the global routing table. But in order to fit the routing table into a router, there must be aggregation. These requirements seem at odds with each other. This is because there is a hidden assumption: the full global routing table must be present in all routers that are part of the default-free zone. Dropping this requirement makes everything much more complex, but it is possible. The global routing table can then be split into several parts, where individual routers all handle one (or a few) of those parts.

This works as long as traffic for a certain subset of the destination networks present in the global routing table is always sent to a router containing that part of the global routing table. The obvious way to accomplish this is for each router to announce an aggregate covering the part of the global routing table it serves. For instance, if a network has four routers and wants to divide routing information for the IPv6 global unicast address space over those routers, it could have router A handle 2000::/5, router B 2800::/5, router C 3000::/5 and router D 3800::/5. So if this network peers with another network that announces 2200:abc::/35 and 3ffe:def::/35, all routers except router A filter out the first route, and all routers except router D filter out the second route. When router C then has a packet for 2200:abc:1:2::1, it sends the packet to router A (because router A announces the 2000::/5 aggregate) and router A delivers the packet to the right peer. Note that this behavior is completely hidden from the peer: the aggregates are only used within the local network, they are not announced to peers. To avoid confusion with regular provider aggregatable routes, the term "pilot routes" will be used for this type of private aggregates.

This practice scales relatively well: by adding more routers, it is possible to accommodate a global routing table of arbitrary size. (These extra routers must be "border routers" that interconnect with other networks.) However, there is a major problem: traffic for certain address ranges must always first be transported to the location of the router handling this address range. So if two end-users in Europe want to communicate, but the address range for one of them is handled in North America by the other's ISP, and the other's address range is handled by a router in Japan, this traffic that has the potential to stay within the region has to circle the globe. This "scenic routing" can be avoided by assigning address space to multihomers in a geographically aggregatable manner. This way, networks can have a range of addresses be handled by a router in the region where the addresses are used. However, this is not a strict requirement. For instance, a network that only has a presence in the US doesn't necessarily have to interconnect with other

networks in Europe or Asia. In practice, it will have routers at the US East Coast (where many European networks are present) handle the European address ranges, and routers at the US West Coast (where many Asian networks are present) handle the Asian address ranges.

3. Operational Details

First of all: more specific routes from customers are usually not filtered. They are announced to peers at all interconnect locations. It is up to the network receiving the routes to filter them. Only when two networks agree on where to exchange routing information for certain geographic aggregates, there may be outbound filtering of more specific multihomed routes.

The aggregation scheme works as follows. The network is divided into zones. The exact way in which this is done depends on the particular topology of the network, and doesn't have to match the layout of other networks. Static pilot routes for all address ranges used within the zone are configured on at least two routers (for redundancy) in that zone (or as close to the zone as is practical). Then both EBGP and IBGP filters are configured per peer. The IBGP filters are applied to all sessions with routers in other zones (not to sessions with other routers within the zone) and filter out the more specific routes falling within the address ranges used in the zone. The EBGP filters do the opposite and allow only more specific routes for destinations within the region. This makes sure more specific multihomed routes are allowed in the routing table within the zone, but aren't announced over IBGP to other zones.

3.1 Interconnection

Since interconnection is not an exact science, there may not be adequate interconnection within the zone with some peer networks. When this is the rule rather than the exception, this indicates the zones are too small. Increasing the zone or merging several zones will make sure there is interconnection with most peer networks within the zone itself. For the few networks for which interconnection within the zone isn't possible, EBGP filters that always allow all more specific routes are used. Also, these routes are tagged with an internal community that prevents them from being filtered in IBGP. As a result, there is no aggregation for these peers, but there is still full connectivity. It should be possible to limit this de-aggregation to a small number of zones rather than the entire network with more sophisticated filtering.

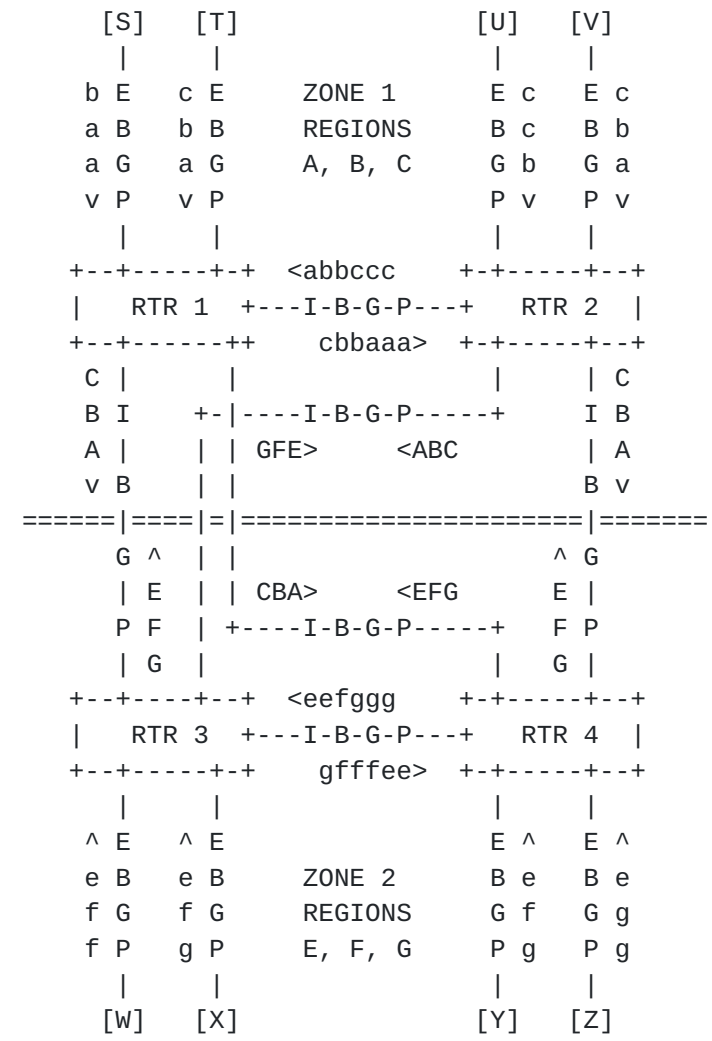
3.2 Zone Partitioning

It is important that regions are never partitioned, because when this happens, packets for certain destinations will loop. The router inside the zone will route them outside the zone because of the more specifics pointing to the other partition of the zone over a router that isn't part of the zone, and the first router outside the zone will route the packets back into the region to the closest router

announcing the pilot route.

3.3 Example Picture

The following picture represents an AS with four routers and eight peers, divided into two zones that each handle routing for three regions:



[S], [T], [U], [V], [W], [X], [Y], [Z] Peer EBGP routers
RTR 1, RTR 2 Routers in zone 1
RTR 3, RTR 4 Routers in zone 2
A, B, C, E, F, G Pilot (aggregate) routes
a, b, c, e, f, g Individual /48 routes for end-user networks
<, > ^, v The direction of the routing information flow

Figure 1: Geographic aggregation example

4. Migration

Migration from a regular, non-aggregated setup to full geographical aggregation doesn't have to be immediate. The process can be carried out is several steps:

1. The border router handling most of the traffic to a specific geographical destination or aggregate of several destinations is promoted to "designated router" for the matching address range. The designated router is configured to announce a pilot route over IBGP and with filters that don't allow more specifics for the destinations covered by the pilot route to be announced over IBGP to non-border routers. Now only border routers have the more specific routes.

2. Border routers are configured with EBGP filters to filter out incoming more specific routes covered by pilot routes announced by far away designated routers. (For instance, routers in Europe are configured to filter out American more specifics for which an American router announces a pilot route.) The designated router is configured to no longer send these more specifics over IBGP to the routers that now filter those same routes on EBGP sessions. (For the American routers, their European IBGP neighbors now essentially become part of the group "non-border routers".) Now each border router only has a subset of all multihomed more specifics in its routing table.

Step 1 can be implemented on individual routers one at a time, and, barring configuration mistakes, doesn't pose any risks. There is only one pilot route, and only more specific routes announced by the same router as the pilot route are suppressed. Since both the new pilot route and the now suppressed more specific routes point to the same border router, the way packets are routed through the network is completely identical and there is no risk of loops. If different a router than the designated router has the preferred external route for a more specific, this more specific route will be announced as before, since only the designated router is configured to filter out these more specifics.

When the designated router is the one holding the best external route, non-border routers won't see any more specific routes for this destination. The designated router has a filter, and the other border routers don't announce the route over IBGP because they aren't the ones holding the best route. To aid aggregation, the designated router can be configured to increase the IBGP Local Preference attribute for the more specifics it acts as designated router for. This way, the route over the designated router is always preferred, even if another router has a matching more specific with a shorter AS

path or better Multi Exit Discriminator metric.

When the designated router becomes unreachable or loses its external routes, there will be automatic de-aggregation: more specific routes are announced by other routers.

Step 2 can also be implemented one router at a time. The new EBGP filters should be installed first, after which the designated router can be configured to no longer announce more specifics to the border routers with the new EBGP filters. If this is done the other way around, more specifics will leak over IBGP and there will be non-optimal routing. Without step 2, there is no aggregation in border routers: they need to hear the designated router announce a "better" more specific, or they will start to announce their own over IBGP. Introducing step 2 introduces the risk that certain destinations become unreachable when there is an outage. For instance, when European routers no longer see American more specifics, and the European and American parts of the network become partitioned, it is no longer possible for the European routers to send traffic to American destinations, even if there is peering in Europe that would have made this possible before. This step should only be taken if the risk of network partitions is negligible.

5. Limitations

Since this scheme depends on geography for aggregation, it only works well for organizations that connect to the internet in locations that are close together. An organization with a network spanning multiple countries and connecting to the internet in all those countries isn't geographically aggregatable, and neither is an organization connecting to ISPs very far way, for instance by means of a satellite circuit.

These types of organizations must choose address space falling within a geographic area that doesn't (fully) fit if they elect to use the type of address space this aggregation scheme uses. This choice will have consequences on routing efficiency, and when the infrastructure changes, the organization may need to adopt a new address range to minimize the routing efficiencies created by the change.

Although the notion of geographic aggregation has been discussed many times within the IETF over the past eight years or so, this approach is generally believed to be flawed since the topology of the networks that make up the internet and the interconnection between those networks doesn't align well with geography. This is indeed a problem, but it doesn't automatically make geographical aggregation useless, it only makes it less effective. Since network topology is under constant revision, and as networks get faster, the main disadvantage of "scenic routing" (the increased speed of light delay) becomes more acute as bandwidth increases, it is certainly not unthinkable for the topology of the internet to align itself more with geography over time.

Additionally, while the past decade or so the trend among high speed IP backbones was to run IP as directly over the physical infrastructure as possible, today this trend seems to be reversing with the adoption of MPLS and switched optical services. This allows two routers to communicate directly at the IP level without the need for a direct physical connection, making it possible for two routers in the same aggregation area that don't share a physical connection to exchange packets without the need for routers outside the aggregation area to know routing information for the area. Instead, a direct virtual lower layer connection is used so the traffic can pass through areas where the routing information isn't known, enabling aggregation to become largely independent from the physical network topology.

The same can be achieved through tunneling.

6. Route Visibility for Customers

In order to be able to do traffic engineering for outbound traffic, multihomed customers need to receive a consistent view of the global routing table from all their ISPs. If the aggregation levels of different ISPs used by a multihomed customer don't match, because of the longest match first rule, most of the traffic will flow over the ISP doing the least aggregation. To avoid this, ISPs are strongly encouraged to provide their customers with a full, unaggregated view of the global routing table. If an ISP aggregates internally, such a view could be obtained by the customer by having an EBGP (multihop if necessary) session with one or more route servers, in addition to the regular EBGP session to the next hop router.

ISPs should also provide their customers with pilot routes at all aggregation levels, even if the ISPs themselves don't (yet) aggregate. This makes it possible for customers to filter out more specifics and still maintain a consistent view of the global routing table. If an ISP can't do this immediately (adding a large number of pilot routes is a lot of work) the ISP should establish a time frame for implementing the necessary pilot routes and communicate this to existing and potential customers. A reasonable time frame would be six months to implement continent/country/province/state level pilot routes for the whole world, a year to implement metropolitan area pilot routes for the regions the ISP is active, and 18 months to implement world wide metropolitan area pilot routes, starting from the moment a geographically aggregatable address allocation mechanism is implemented.

7. Traffic Flow

Larger ISP and ISP-like networks that interconnect with other networks in more than one location must have a policy on how to select the interconnect location used for traffic to those other networks. At present, the most widely adopted policy is "early exit" or "hot potato": packets are routed to the closest interconnect location where the other network is present and delivered to the destination network there. As a result, packets travel most of the way over the destination network. If both networks use the early exit policy, traffic in one direction will travel most of the way over one network, and traffic in the other direction most of the way over the other network, so the policy is "fair" as long as the traffic volumes are fairly equal in both directions. This policy is implemented by not changing the default behavior for the most widely available BGP implementations.

Since the aggregation scheme described in this document requires traffic to be transported to a location where more specific routing information is known, and this location is presumably close to the destination of the packet, adoption of this scheme leads to a "late exit" routing policy for multihomed traffic. Assuming early exit is still used for single homed traffic, there are four possible permutations for the traffic flow between any two hosts:

1. Hosts A and B both single homed: both early exit = "fair"

2. Host A single homed, host B multihomed: traffic is exchanged close to host B = host A's network does most of the work

3. Host A multihomed, host B single homed: traffic is exchanged close to host A = host B's network does most of the work

4. Hosts A and B both multihomed: both late exit = "fair"

Since networks can control the level of late exit routing by (selectively) de-aggregating and many interconnection (peering) agreements call for equal traffic volumes in both directions, the potential for changes in the flow of traffic should not adversely affect existing networks.

8. Geographical Address Allocation

This section establishes an address allocation framework for Geographically Aggregatable Provider Independent (GAPI) IPv6 addresses for the purpose of multihoming. A /16 is divided in a hierarchical manner over geographical entities such as parts of continents, countries, states, provinces and metropolitan areas, with each receiving one or more /32 allocations from which end-user assignments can be made. The number of /32 allocations for a geographical entity depends on the current population.

Note that this section was previously a separate draft.

The geographical aggregation scheme splits the global routing domain into a number of smaller regional ones, where flat routing happens in each region. Ideally, outside the region only aggregates are visible. For simplicity and to allow efficient implementation, the framework presented here requires "areas" where flat routing takes place to have a fixed size: a /32 holding up to 65536 (2^16) fixed sized end-user /48 assignments. The maximum number of these /32 areas is also 65536. Areas are grouped in CIDR-like fashion if a geographic region has a population that warrants allocating more than a single /32. The highest level of aggregation is the subcontinent or "zone" level. There are 13 entities at this level, in order of population:

1. China
2. Continental Asia
3. India
4. Northern Africa
5. Asian Islands
6. Western Europe
7. North America
8. South America
9. Eastern Europe
10. Middle East
11. Southern Africa
12. Central America

13. Oceania

The next level is the country level. Every country is assigned a range of /32 blocks, depending on population. Countries that are medium-sized or larger may be subdivided according to existing administrative boundaries, such as by state or province. The allocation size per state or province must match the population relative to the country and other states or provinces. The lowest level of aggregation is the metropolitan level. Cities of sufficient size are allocated one or more "metro areas". Assignments to end-users in, or very close to, a city are drawn from one of the metro area /32s allocated to the city. Addresses for end-users in small cities or rural areas are drawn from one of the /32 areas allocated to the country (if not subdivided), state or province (a country/state/province or "CSP" area).

8.1 Allocation policy

The goals of the allocation policy are:

1. Be completely neutral, fair and unbiased, in order to minimize the potential for political complications

2. Good aggregation at all levels

3. Reasonable flexibility

4. Ease of implementation

8.2 Country Allocations

Each independent country is allocated at least one /32 area. The allocation size depends on the country's population figure for the year 2001. This is divided by the number D1, which equals 131072. The result of the division is rounded up to the next power of two.

This is the number of /32s constituting the country's allocation.

8.3 Zone Allocations

The subdivision of the globe in 13 zones is relatively arbitrary. However, this division fits current and expected future Regional Internet Regions well, and limits the population per zone somewhat over a strict by-continent subdivision. Zone allocations are chosen such that they are large enough to hold the country allocations for all countries located within the geographic bounds of the zone. If for any of the zones that encompass more than a single country, the

number of /32s not allocated to countries is less than 25%, the zone allocation size is doubled.

8.4 Subdivision of Large Countries

When a country has an allocation of 32 or more /32s, this address space may be distributed over the country by allocating blocks of /32s to existing sub-entities such as provinces or states. The exact geographic bounds of these sub-entities must be clear to the general public and not subject to any controversy. The size of each allocation is determined by dividing the population of the sub-entity by the number D2, which is twice D1.

At least 40% of the country allocation must remain unallocated. If necessary, a higher value than D2 may be used as a divisor in this country to reach this objective. The average number of /32 areas per state or province must be at least 4.

8.5 Metro Allocations

All cities with a population of at least D2 within the city limits are allocated a block of /32s. The population for small cities or municipalities that do not qualify for an address block of their own is added to the closest city that qualifies, if there is one within 40 kilometers. (Distance measured center to center.) The size of the address block for a city and its surroundings is determined by taking the total populace, dividing it by D2 and rounding down to the next power of two.

8.6 Reserved for Future Use

The first 1/64th of each allocation at the country/state/province level is reserved for future special uses and must not be allocated to lower aggregation levels and not be assigned to end-users.

8.7 Subsequent Allocations

Whenever allocated address space gets close to running out, the IANA, Regional Internet Registry or other organization managing (part of) the address space should draw new allocations from the next higher level. New blocks of address space may be allocated in a way that is different from what is outlined here, if analysis of the coordinates for current assignments warrants this.

8.8 End-user Address Assignments

Per country or state/province, only one /32 block is used initially. A new block is used when the first is exhausted, and so on. The /32s

allocated to a metropolitan area may be put into use concurrently, if there is a reason to do so.

When requesting IPv6 GAPI addresses, an organization should provide justification for the use of GAPI space, and information that makes it possible to assign addresses from the right geographic area, in addition to the information required by current assignment policies. Multihoming is justification for the use of GAPI space. Geographic information should consist of the longitude and latitude of the primary location where the addresses will be used. This information should be accurate within 2 kilometers, as long as any inaccuracies don't make the organization appear to be at the other side of an administrative border or natural barrier (such as a river). Preferably, the requesting organization should also include the longitude and latitude of the ISP locations they connect to. However, this information may be omitted.

The minimum assignment size is always /48. Future multihoming solutions may not support the longest match first rule.

9. IANA Considerations

If this scheme is adopted, the number of networks requiring an Autonomous System number will rise beyond what can be accommodated using the current 16-bit AS number space. There is a draft proposing the use of 32-bit AS numbers [32bitAS]. Since having a universally recognized AS number is less important for a multihomed "leaf" network than for a transit network, it is recommended that the 32-bit AS number capability be implemented as soon as possible. All multihomed networks requesting an AS number that are capable of using a 32-bit AS number should be assigned an AS number higher than 65535, so 16-bit compatible AS numbers remain available for transit networks.

The IANA is requested to allocate /16 worth of IPv6 address space for GAPI, and the Regional Internet Registries are asked to further assign this address space to end-user organizations. The Regional Internet Registries should take the requester's geographic location into consideration when assigning address space.

10. Security Considerations

Having addresses that are closely tied to an organization's location may be undesirable in certain situations. Organizations requesting address space should consider the consequences of using GAPI address space, and are encouraged to use provider aggregatable address space if and when they want to avoid disclosing their location.

Some organizations may be uncomfortable with providing very accurate longitude and latitude information when requesting address space. They may introduce a 2 kilometer inaccuracy to avoid exact pinpointing, as described in section 6. In addition, the Regional Internet Registry or other organization responsible for assigning address space must not make location information public. Specifically, this information should not appear as a result of whois queries. Registries are encouraged to provide aggregated location information for policy development purposes, but only as long as this information is anonymized and can't be tied to a single organization or small group of organizations.

This aggregation scheme doesn't propose any changes to protocols or implementations, so it doesn't introduce any new protocol or implementation risks. However, there is one problem: since routing information is removed from large parts of the network, it is no longer possible to use the routing table to do ingress filtering [RFC2267] using the "unicast RPF" feature implemented by several router vendors. The alternative, having statically configured filter lists, doesn't scale. This leaves networks implementing this aggregation scheme with no protection against incoming packets with falsified source addresses, so it is highly recommended that network operators make sure they don't generate or accept from customers packets with falsified source addresses and that vendors implement mechanisms to trace back the source of these falsified packets.

Author's Address

Iljitsch van Beijnum
Karel Roosstraat 95
2571 BG The Hague
NL

EMail: iljitsch@muada.com

Appendix A. References

[RFC2267] RFC 2267, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing"

[RFC2772] RFC 2772, "6Bone Backbone Routing Guidelines"

[32bitAS] "BGP support for four-octet AS number space", work in progress

Intellectual Property Statement

Full Copyright Statement

Acknowledgement

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