Internet Draft   T. Hain
Document: draft-hain-ipv6-pi-addr-use-10.txt   Cisco
Expires: February 2007   August 2006

Application and Use of the IPv6 Provider Independent Global Unicast Address Format

Status of this Memo

Abstract

This document discusses the expected use of the Provider Independent address format discussed in the companion document GEO [1] in the Internet. Several parties have expressed interest in using this approach as a good fit for managing their networks. With the long timeframe that the shim6 effort will take, this approach provides a scalable multi-homing approach for use in the interim. In addition to covering implementations where it adds value in managing the growth of the Internet routing tables, the document will discuss implementations that should be avoided due to the negative impact on the routing tables.

Conventions used in this document

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 RFC-2119 [2].

Introduction

This document discusses the application of the Provider Independent (PI) global unicast address format for the IPv6 address assignment. This address format is based on WGS-84 [3] derived geographic location. Historically there have been many debates about the value of geographic versus provider based allocation schemes. One interesting observation about this debate is the overriding assumption that the Internet will have to be built using one or the other, rather than leveraging the strengths of each in context. The intent here is not to reopen that debate, but to present the case that with the multiple address capabilities of IPv6 the approaches can be used in a complementary manner.

The protracted debate in the RIR policy discussions show that we are no closer to consensus on a Provider Independent (PI) allocation policy than we were five years ago. While the service provider community continues to focus on the threat of routing table meltdown, the global enterprise network managers continue pressing the position that the current IPv6 Provider Aggregate (PA) addressing scheme is unusable by anyone except Internet providers trying to serve the household and small business market. It is clear to all that additional mechanisms will be required to address the needs of large multi-continent organizations, the difficult part is defining who gets a routing slot and who does not in a globally equitable manner.

The current PA focus on route aggregation, deals with the technical problems of memory and CPU resource consumption when site attachments conform to the model, but the other significant business issues related to PI approaches remain open. In any case, a site that is expressing an explicit global routing policy will have full-length prefixes announced. The PI address format discussed here does not remove that issue, instead it offers a PI approach for those

looking for something more than PA yet less than a global routing slot.

Additional work is ongoing in the IETF to separate the roles of an address as the identifier or locator. While this effort might eventually provide alternatives for dealing with multi-homed connections, its outcome is far from certain. Recent charter adjustments have restricted its goals in an effort to achieve any hope of consensus, and even if successful at this scaled back capability the approach will require substantial retrofitting of other aspects of Internet management and security that have been built on the assumption that the roles are combined.

It is frequently assumed that any address format that is not based on provider aggregation will degenerate into the 'swamp' that came to describe pre-CIDR IPv4, with the result that the routing table grows unabated. The goal of this scheme is to allow sites to be independent of any provider, while still allowing aggregation for those who do not require explicit global routing policy. As a result, there will need to be consistently applied rules for when a prefix gets aggregated and when it doesn't. These will be discussed in the recommendations section.

This document will highlight the operational configurations where the PI geographic based addresses provide value in provider independence, as well as those situations where they should be avoided in favor of the provider aggregate mechanism.

History

Provider based address aggregation has its roots in the IPv4 routing table growth limiting effort known as CIDR [4]. The basic premise is that routing entries can be summarized in such a way that a large number of sites, which subscribe to the same service provider, generate a single entry in the global inter-provider routing exchange, also known as the Default Free Zone (DFZ). While this works well when sites connect to a single provider, it is inadequate for providing a site with redundant access through multiple service providers. Sites that expect redundant service through an arbitrary number of service providers currently require the global routing table to carry an explicit entry for the full-length prefix

allocated to the site. Traditionally this was accomplished by having a site acquire an address allocation out of the pre-CIDR range of the IPv4 address space, which remained provider independent. Lately this process has evolved into simply arranging with each of the service providers involved for multiple announcements of the explicit prefix allocated to the site by one of those providers. While the effect on the global routing table is the same in either case, the act of 'punching holes' in provider aggregates increases operational complexity, and makes it very difficult for a site to disconnect from the service provider that allocated the address prefix.

As noted, the prime motivation for CIDR deployment was reduction on the size of the IPv4 routing table. Using BGP, the total size of the table is a direct function of the number of address aggregates within the Internet, where each entry describes a contiguous set of subnets that share a common origin AS and a common reachability policy. The mechanism, of aligning site allocations with the provider they attached to, worked well for this purpose, but at the same time was directly contrary to the needs of the site for provider and routing policy independence. The primary operational motivation for sites to connect to multiple providers and/or regions is resiliency. Other factors that come into play are managing overall cost, and optimizing performance or balancing load.

Collectively these issues drive sites away from the nicely structured single-connection hierarchical model that is the foundation of IPv6 Provider Aggregatable [5] allocation. At the same time, due to the evolving state of infrastructure deployments, the concepts of geographic locality and least-cost locality often don't match. The consequence of the collective situation is that no one approach to address allocation will solve the entire set of route scaling problems.

The goal of the PI address format described in GEO [1] is restoring the integrity of PA prefix format for use by the single homed sites, while simultaneously providing a scaleable approach for the growing number of multi-homed sites. This is accomplished by relating one of the IPv6 address prefixes of the multi-homed sites to an unambiguous geographic reference point in a way that summarizes well over physical distance. This is not an attempt to have routers understand geography. Rather it is simply a mechanism to allocate address prefixes to sites in a way that can be abstracted into a minimum number of routing table entries for routers that are not directly involved in the local topology. This approach has a strong advantage over the IPv4 PI space (which is effectively randomly assigned) in

that there is a clear structure that allows for efficient abstractions when the detail is unnecessary.

Site requirements to be provider independent and multi-home:

Several issues play into the reasons that sites insist on remaining provider independent leading them to multi-home. Beyond the simple uncertainty that any given service provider will still be in business next month, there have been enough widespread outages of various kinds of service over the years to show that trusting any one provider (who is in turn dependent on device suppliers where a single bug can lead to system-wide outages) is problematic. At the same time, the cost of transmission circuits is low enough that it is frequently less expensive to buy Internet access services from two or more providers than to pay any one of them for premium service (history has also shown that even these premium services fail). So in addition to increasing the robustness of the Internet access, these sites frequently end up with more bandwidth for use in the normal case.

The details are being documented more completely in other current works, but an overview of the requirements would include:
Operational reasons:

- Insulation from routing instability striking one upstream provider.
- Insulation from local-loop/fiber cuts, or central office equipment failures.
- Ability to reduce the points at which outages or packet loss can occur.
- Ability to reduce the average number of hops between a network and various important sites.

Business reasons:

- Insulation from being held hostage by the ISP when billing or other disputes occur.
- The negotiating leverage provided when service provider changes are simply circuit installs and don't involve readdressing.
- Risk mitigation to investors and insurers who consider redundant connectivity a business necessity.
- Reducing the overhead of continually changing explicitly configured firewalls for inter-enterprise communication.

Site scale

There are differences between the global enterprise and the small

site / telecommuter in terms of multi-homing needs, but not in their goals for provider independence and resiliency. From the perspective that service providers generally prioritize customer restoration (and sometimes the quality of the engineer working the incident) by the size of the circuit, and it would appear that the lowest speed circuits get the worst service. This leads to a state where those with the smallest demand for bandwidth generally perceive the greatest of need to multi-home for reliability.

Historically many service providers have used access capacity as a rule-of-thumb in distinguishing the difference in multi-homing requirements for these site types, but with the current deployments of gigabit Ethernet over fiber-to-the-home, bandwidth has become an insufficient measure of a multi-homed site's need to express an explicit policy in the DFZ. As a generalization, the small site / telecommuter simply wants to be always available and internally defaults to any available providers, while the global enterprise expresses an explicit prefix policy (for a variety of reasons including traffic management) by participating in the global BGP exchange. This generalization provides a natural (and more accurate) delineation between the types of multi-homed sites, and the PI mechanism described here exploits this operational characteristic to limit table growth.

Current realities

Service provider business issues

During the push to deploy IPv4 CIDR, a disconnect developed within the service provider community between the operational goal of minimizing the table size through enforced aggregation, and the business goal of giving the customer the services they demand. Over the short term the demonstrable realities of the routing collapses in 1994 and 1996 allowed the operations team to wield the upper hand. In the end though, the self-indulgent business interests have overridden the altruistic sentiments of the 'good of the Internet', as organizations eventually realize that bringing money in the door

will always trump the operational desire to limit growth. The effect of this has been documented in recent studies [6], which show the routing table growth returning to an exponential rate after a few years of linear growth.

It is a fairly straightforward process to 'follow the money' and realize that any service provider that wants to survive will

propagate a full-length prefix for a customer site into the DFZ. The fundamental reality is that the site paying for service will refuse to let any service provider dictate the business requirements, and the service provider sales staff will respond to that by selling the service that the customer is demanding (in this case, provider independence).

Further, many service providers consider it a business obligation to supply a full view of the Internet routing table to the customers that request it for load balancing. To accomplish that, the entire set of long prefixes has to be passed everywhere unless the provider resorts to a default route to someone else. This means that with the current routing technology, service providers will be accepting and passing full length site prefixes as long as they are selling the service of a full default-less view to their customers.

Operational issues

In the current Internet, service providers frequently have conflicting operational objectives for handling traffic; in their search to minimize internal costs, they look to hand off traffic as quickly as possible. This is colloquially known as 'hot potato' routing, where the holder of the packet is looking to dump it as early as possible, while the potential receivers are looking to avoid being dumped on as long as possible.

Since the routers understand policy as described through a longest-match algorithm, the 'Dump Early' strategy wants to hear short prefix lengths, while the 'Avoid Being Dumped On' model drives wide distribution of the longest possible prefix. Given this situation it is in the interest of the service provider whose customer is attempting to influence incoming routes to propagate that multi-homed site prefix as far as possible. The result of this is that only traffic for customer sites will transit the boundary. At the same time, the holders of the packets on the other side of the DFZ would prefer to filter any long prefixes so they can simply dump packets as quickly as possible.

The independent policy objective of a global enterprise network then gets injected into this environment of provider conflict. The protocol mechanism for assuring provider independence of a site's specific policies is to distribute the full site prefix list into the DFZ. Since the site's ISP as a receiver is interested in only carrying traffic for that customer, propagating the full length site

prefix is not only self-defensive against dumping, it is aligned with their mutual business interests.

Applicability of the PI format

A fair question to ask is; if short prefixes through proxy aggregation make no business sense, what mechanism will constrain routing table growth? Currently a single routing protocol is expected to sort all of the contradictory policies to arrive at a perceived optimum from every perspective. In this conflicted environment we are left with a single entity, the originating Autonomous System (AS) number, which is the basis of the various mechanisms used to describe a policy as applied to the listed set of prefixes. There are currently around 22,000 AS numbers actively distributed through the global routing system. Of these, ~ 8,800 are the origin AS for a single IPv4 prefix. The set of AS's which are origin for 4 or fewer prefixes is ~ 10,000. This means ~ 15% of the AS's are origin for ~ 85% of the prefixes in the global IPv4 BGP table. On average each AS originating 5 or more prefixes are expressing policy for around 40 IPv4 prefixes. If the IPv4 prefix allocations could be dynamically reclaimed and defragmented completely along provider alignments, the size of the Internet routing table could theoretically be reduced to around 10% of its current size. This would effectively turn back the clock on routing table size concerns by close to 10 years. While not a universal solution, by using the IPv6 PI address format the overwhelming majority of multi-homed organizations could do just that by using a single ::/48 prefix (in effect defragmenting the prefix space), and by taking this approach the number of prefixes with a common origin AS would approach 1. Assuming the goal of constraining the routing table growth is simple stability of the routing protocols; as the average number of prefixes per origin moves from 40 toward 1, aggregation of the explicit 'policy defining' PI addresses in the DFZ becomes unnecessary. Uses of PI addresses that do not attempt to define a global policy will be discussed in the subsequent section on Exchanges.

Accomplishing the goal of limiting table growth would require a slight modification to the registry policy on justification for an AS number. Currently in order to be assigned an ASN, each requesting organization must provide verification that it has one of the following:

- A unique routing policy
- A multi-homed site

This leaves open the opportunity for every multi-homed site (including telecommuters) to express a routing policy by injecting their full prefix into the DFZ. The obvious question is how many sites really want inbound policy control vs. simple path redundancy and fail-over between their attached providers? Since the fundamental requirement for an AS number in a PA context is really a mechanism for expressing policy independent of the provider, the

line about multi-homing becomes an IPv4 artifact and should be removed.

Overview of the IPv6 PI Address Format

Details of the provider independent address format are provided in the companion document GEO [1]. A high level overview is provided here for local context.

Addresses defined with the Provider Independent format prefix xxxx (IANA assigned) *ARE PORTABLE* between providers. At the same time these addresses are *NOT PORTABLE* across changes in geographic location. Entities that expect identifiers to be portable across physical location moves MUST use alternatives such as Provider Aggregatable prefixes or DNS names.

Provider Independent addresses are constructed using a bit interleave of the WGS-84 based latitude and longitude of a site. While the available 44-bit field allows for resolution of a grid approximately 6.4 meters on a side, addressing privacy concerns may require the allocation to be at 36-bits with the expectation that site assignments in that 100 meter grid will be managed by the smallest appropriate local jurisdiction. Accommodating areas of dense population may require that the grid size be adjusted to allow for more flexible assignments for multi-story buildings and business centers with multiple independent sites per geographic grid. Actual assignments within a geographic grid SHOULD be a local jurisdictional issue (matching scope jurisdiction; building owner, community board, local government, etc.), and independent of any service provider. The only rule is that the allocation point MUST be contained within its natural grid. If a locally administered grid needs to be expanded to accommodate density, it MUST avoid or otherwise coordinate use of any existing values that fall within its new boundaries. One approach to accommodate density would be to annex an uninhabitable adjacent region. It is not clear this will really be necessary since the number of ::/48's available to a multi-story building will typically exceed 1,000, with a minimum of 64 ::/64's per vertical meter of each 6.4 x 6.4m area, or 1.59 ::/64's per cubic meter, 1km deep over the entire surface (see Extended Resolution discussion in [1]). The existing PA registries may choose to play a role in helping to coordinate the actual assignments by providing a public database of the local jurisdictional decisions.

Specification of the WGS-84 reference point SHOULD be the responsibility of the local jurisdiction (who may defer to the layer-1 service provider for process expediency), and SHOULD represent the physical location of the demarcation point of the layer-1 service. In the case where reference points overlap, the local jurisdiction SHOULD provide coordination over the smallest workable scope.

In the case where the local jurisdiction also acts as a local service provider to its tenants (ie: hotel, apartment, or high-rise business complex) it MAY choose to allocate prefix lengths longer than ::/48, as appropriate for the number and needs of its customers. In any case the allocation MUST NOT exceed 64 bits in length to preserve the Interface ID portion of the address, and should be on the nibble boundaries of /52, /56, and /60 to align the management of the dns reverse delegation with the address pool forward delegation.

Constructive implementations

The geographic nature of the Provider Independent address format is designed to allocate addresses to sites which aggregate well in direct proportion to the physical distance from the site. It also allows a locally connected site to easily change providers without impacting the nodes or connections within a site.

In this context, one appropriate use of the Provider Independent address format is a site connecting to multiple providers within a constrained geographic scope such as a city (actual size depends on the local cooperative interconnection between service providers). When used in this way, only those routers providing service within the scope need to know the details about the interconnections, and the global routing table would see a single entry for that routing scope.

Another appropriate use of the Provider Independent address format is when a site will be switching service providers. By preferring the Provider Independent address prefix for a period overlapping the switch, a site may be able to maintain connections while the new service is installed and the old removed.

A third appropriate use would be for an organization providing global content services to provide clients with a proximity hint. The longest match between the list returned from DNS and the PI address of the client should provide the closest physical proximity

(though not necessarily the closest topological proximity). One consideration is that for global load-balancing hints to work, all nodes will need to know their PI address even if they never use it in packets. One way to accomplish this would be by setting the lifetime values in the Router Advertisement for the PI prefix to Valid = Infinite / Preferred = Zero.

A related use recognizes that geography provides distinguished meaning to the term 'home address'. Using a PI address with Mobile-IPv6 [7], where the geographic based PI is 'home', the current provider address would be the care-of address. In this case the nodes are completely independent of provider in both allocation mechanism and packet transport.

Finally, as recognized in RFC 1887 [8], another appropriate use would be for organizations that do not directly connect to or participate in the global Internet (Zero-homed), but do have private links with organizations that are connected. It is necessary for the organization in the middle to be able to differentiate between any privately connected sites and the public Internet, so each of the privately connected sites need to use a unique addresses. This is easiest to accomplish with a globally unique prefix, and since the private site is not connected to an Internet provider, it is unlikely they would be able to do so in a strictly PA environment. While this is not intended to replace [9], by using the PI prefix, there would be no ambiguity.

Troublesome implementations

The PI address format does not provide any benefit to the size of the routing table for sites that require direct connections outside their geographic region. As discussed earlier, these sites will require the full ::/48 prefix to be carried globally, independent of address format type, so if a remote circuit is intended for access to customers of a specific provider, the prefix SHOULD come from the PA space of that provider.

The Provider Independent address allocation mechanism SHOULD NOT be used by a temporary access network (such as a dial point of presence), because scaling routes to single-homed sites attached this way are best addressed through provider based allocations consistent with Provider Aggregatable [5]. The reasons for this are:

1) Temporary access endpoints can not expect to maintain higher-layer connections between physical access events, and therefore

should be using a Provider Aggregatable allocation to minimize their impact on the global routing system as they come and go.

2) The location of the intermittent endpoint is unknown, so the address would have to be based on the access point location. If such an access point were scaled to handle 10,000 attachments it would have to subsume the neighboring addresses for the 2.5 km square it is centered in. Since the currently deployed temporary access points tend to be located in densely populated areas, using them with geographic rather than provider based addresses would have the maximum negative impact.

A site that is multi-homed by fixed and dial-based access will decide between provider and geographic based addresses based on the relationship between the access paths. If the two paths are to the same provider then PA addressing is most appropriate. If the dial-path were to a different provider than the fixed line, it would make more sense to use the PI address, because the site would maintain its prefix and active connections through the routing switch without the need to globally punch a hole in either provider based aggregate.

Fundamental Issues

There are ongoing debates as to the fundamental problems created by unconstrained routing table growth as the Internet topology flattens. Some of the issues raised in these debates include:

- Memory size for holding the ever-expanding table
- Memory to CPU bandwidth for accessing the table contents
- CPU speed in processing table updates
- Convergence time as each event results in a burst of processing
- Inconsistent inter-provider announcement policies
- The volume of information stored with each prefix
- Management of a distributed database with insufficient concurrency controls

While it is clear there are many potential problems, any solutions need to balance these against the costs of equipment capable of solving them. Most service providers will say they want all of these problems solved, but when it comes down to paying for hardware they frequently compromise long-term growth in favor of short-term cost control. As a result, any mechanism or policy needs to take the inconsistency of hardware capabilities into account.

Routing issues

As noted earlier, the unstated business motivation of the service provider is to push the longest possible prefix as far as possible. The primary impact of this on routing becomes one of dealing with the long prefixes of the set of sites expressing global policy. At the same time the routing system needs to be capable of aggregating all the multi-homed connections where the only policy is 'stay connected within a region'.

While the basic mechanism described in GEO [1] is a bit interleave of the WGS-84 latitude and longitude values, the prefix length used by the routing protocols MAY be established on any bit boundary. At the same time, the operational choices will naturally be limited by the requirement for all service providers at that short prefix boundary to have some mechanism for interconnect with all others for traffic delivery. The result is that at some point in the hierarchy all service providers for a scope have to agree on the boundary, then share routes and traffic. It becomes an engineering tradeoff between the size of the routing table, and the cost for maintaining a large number of points where these interconnections occur.

From a site perspective

- on resiliency, there is a single address block that allows connections to survive any shifts in routing due to provider outages.
- on traffic management, the set of address blocks may influence incoming behavior.

From a service provider perspective

- each site is identified in a subset of its PA space, as well as the PI prefix.
- on the DFZ, there will be multiple paths to any full length ::/48 PI prefix (the intent should be that multiple exist, else PA is the right choice), and the shorter prefixes will identify regional interconnections.

The 'hot potato' routing policies will assume a short prefix represents a contiguous interconnection of providers in a given region. To simplify the relationships between providers, it may be necessary to separate the transit service between regions function from the local service delivery function. This will help to contain the longer prefixes to their region of applicability.

Example 1:

Network DEF provides transit services within Europe. For global connectivity it subscribes to provider ABC. It has local transit agreements with competitive service providers GHI and JKL. The company XYZ is a customer of both DEF and JKL with offices in Paris and Moscow. XYZ policy is to prefer the internal network to the public network.

```
                 -------
                |  ABC  |
                 -------
                /   |
         ------  ------  ------
        | DEF |-| GHI |-| JKL |
         ------  ------  ------
           \  '-----------'  /
           ------      ------
          |XYZ-P|-----|XYZ-M|
           ------      ------
```

Route announcement between:
XYZ-P & XYZ-M - full PI and PA ::/48 of the each site
XYZ-P & DEF - full PI ::/48 of this site up; DEF explicit customers ::/0 down
XYZ-M & JKL - full PI ::/48 of this site up; JKL explicit customers ::/0 down
DEF & GHI explicit customers + xAE2:: of XYZ-P
DEF & JKL explicit customers (which includes XYZ)
JKL & GHI explicit customers + xBAC:: of XYZ-M
DEF & ABC xAE2:: up; x::/4 down
GHI & ABC xBAC:: up; x::/4 down
ABC & peers xA00::/7 out; explicit ::/16s from each

Nodes in the Paris office of XYZ would use the xAE2::/16 prefix when talking to sites in the Moscow region, and conversely nodes in the Moscow office would use the xBAC::/16 prefix when talking to sites around Paris.

If XYZ opens an office in New York, it would announce each of that site's ::/48 prefixes to the other two sites, but that should not extend beyond to DEF or JKL. Nodes within the XYZ network SHOULD NOT use the US prefix to talk to nodes in Europe unless the internal connection across the Atlantic is unavailable. In that case, only the New York office nodes would be receiving the local PI prefix so they might choose to use it. If the provider serving the New York office were acquiring its allocation from ABC, the address selection

longest match would lead hosts to select PA.

Example 2:

ISP-1 prefers connections to region 2 via ISP-2, and accepts the short aggregate over that path. ISP-3 has an arrangement with ISP-1 to provide service for its customers over a direct connection between them, and announces it's PA prefix along with the PI specifics of its customers.

Sub-scenario 1:

The Site uses its Provider Independent address. Customers of ISP-1 would use the path via ISP-3 due to the longer prefix announcement. If the link between the Site and ISP-3 failed, ISP-3 would reroute via ISP-4 due to the intra-region announcements. ISP-3 may choose to stop announcing the Site prefix in this case, which would cause ISP-1 to route via ISP-2 due to the short region prefix announcement. Connections between ISP-1's customers and the Site would remain intact during these rerouting events.

Sub-scenario 2:

For cost reasons the Site prefers ISP-4. Implementing the site's preference would require them to use the prefixes from each provider, and then via local policy order the selection rules appropriately. Customers of ISP-1 would not be aware of the site's preferences, and would use their own local policies when initiating connections to decide between the values returned by DNS. Connections between ISP-1's customers and the Site would drop if the connection from ISP-4 to the Site, or ISP-2, failed.

```
              ------       |      ------
             |ISP-1 |------|-----|ISP-2 |-
              ------       |      ------  \
                     \     |         |     \
                      \    |      ------    \ ------
                       ----|-----|ISP-3 |----|ISP-4 |
                           |      ------      ------
                           |           \      /
                           |            ------
                         1 | 2         | Site |
                           |            ------
                    Region Boundary
```


Example 3:

Site-2 connects to ISP's 3 & 6, which announce the short scope prefix to ISP's 2 & 5. None of the ISP's beyond 3 or 6 are explicitly aware of Site-2.

```
                               |
        ISP-1 ---- ISP-2 --|- ISP-3 -
          |   \_     |     |/   |    \
          |     \_   |   _/|    |     Site-2
          |       \  |  /  |    |    /
        ISP-4 ---- ISP-5 --|- ISP-6 
       /                   |
  Site-1                   |->
                     Scope Boundary
```

If the link (or regional exchange) between ISP-3 & ISP-6 failed causing a partition of the scope, specifics announced via ISP-5 could be used to heal.

Exchange Issues

Historically, exchange points have been used to optimize the number and size of circuits needed to reach a group of peer networks. As more of them are deployed, they also provide a degree of traffic localization.

Practical requirements for exchanges include, proximity to the physical cabling infrastructure, diversity of its own physical location and the interconnect capacity between those parts, as well as appropriate scaling to the number and types of customers in the region. As a general rule, an exchange fabric at layer-2 is the most flexible, but the exchange service may also want to provide a layer-3 peering aggregator to reduce the number of participants in an N-way mesh.

The general point is that the transit providers interconnecting the metro areas only need to know the aggregates. To accomplish that there needs to be a common structured exchange point, or subset of routers which know the interconnect detail. As the number of full length prefixes (::/48) grows, the convergence time of the routing protocol rises. It is assumed that simply for reasons of physical infrastructure scale, before the list of advertisements grows too long, additional exchanges will be established using longer prefix subsets of the established exchange.

Care must be given to the fact that when an area scope is too large, it may become partitioned by natural terrain based boundaries. In these cases, either the more specific prefix values must be advertised, or the providers involved must carry the specifics necessary to heal the partition.

Note: exchanges for a scope don't have to be physically located in the scope of interest; they are simply required to have service provider agreement about aggregation and traffic exchange.

One concern that has been raised is that the majority business model of the current exchange points is focused on being an interconnect fabric rather than acting as a service provider. There is nothing in the PI prefix mechanism that requires that to change, as sites could multi-home through local ISP's rather than direct attachments to the exchange. It is also true that any exchange which provides direct service to sites could use a PA prefix like any other local service provider. This means that the exchange business model is not a factor in the allocation of geo based PI prefixes.

What the exchange brings to the PI mechanism is a focal point and simplified relationships to help ensure that the infrastructure of the short prefix scope is contiguous. While it is technically possible to operate without an exchange fabric (and for performance reasons some interconnects within a scope will choose this), the inter-provider relationship matrix becomes more complex without one.

Placement

With an expansion in the number of multi-homed sites, additional exchanges may need to be built. The decision to do so will be a clear engineering driven decision based on the acceptable size of the local routing table (driven by the number of multi-homed sites) and the circuit costs providers will have for connecting. Operational experience shows that over time service providers have deployed exchanges with 40 – 600 km spacing loosely based on connected population density [10] (2-1991 -> 200-2002 -> 220-2006).

One reason for the current set of exchanges is the reality that costs have been significant when national boundaries are crossed. While minimizing the size of the table for any given router would drive deployment of exchanges with ever-closer spacing, the continuing circuit cost for connecting to multiple exchanges will act as a natural counter balance to prevent an excessive number of them from being created. The costs for these additional exchanges should be directly mapped back to the multi-homed sites that create the need. Punching holes in PA space leads to a situation where it is difficult to associate the site that creates the routing table growth problem with the point where the pain is felt (the DFZ); but distribution of the PI format prefix creates a mechanism where the providers could point at a specific local cost (the exchange) which

supports the goals of a site, and the site could in turn see explicit value for the additional cost. Replacing the current arbitrary inter-provider filtering arrangement with a clear architecture around exchange points will make it easier to explain the costs.

While cost pressure is going to push back and discourage a massive number of small exchanges, there will be a clear benefit to exchanges covering a large expanse. Even if economics only justifies exchanges for the 22 dense population centers listed below, over 300 million people are covered (~5% of global population). Taking the example out to 12 bits (::/16) provides additional granularity for those regions where several large population centers already support multiple exchanges, and may simplify operations. Couple this with the likelihood that significant geographic areas are connected through these population centers and there is little immediate need to add additional exchanges.

Population centers (~10 M people)		Prefix	Example Current Exchanges
Sydney	-	x128::/12	AUSIX
Tokyo, Osaka	-	x2D3::/12	NSPIXP-2, JPIX
Seoul	-	x2C9::/16	IX, DACOM
Beijing	-	x29D::/16	Terramark
Shanghai	-	x2C0::/16	SHIX
Manila	-	x24A::/12	PHIX, PHNET CORE, HKIX
Jakarta	-	x0B8::/12	Napsindo, Sing Tel, KLIX
Delhi, Calcutta	-	xB7A::/12	THIX
Mumbai	-	xB67::/16	EMIX
Karachi, Teheran	-	xB69::/16	Karachi NAP
Moscow	-	xBAC::/12	M9-IX, MPIX
Cairo	-	xB80::/16	AIX, CyIX
Istanbul	-	xADD::/16	TIX
London	-	xAB7::/12	LINX
Paris	-	xAE2::/12	PARIX, AMS-IX
Sao Paulo	-	x5C7::/12	Diveo NAP
NY	-	x798::/12	MAE-East, NYIIX
Mexico City	-	x673::/12	Chicago NAP
LA	-	x6C2::/12	MAE-West, MAE-LA

Some neighboring regions may find it advantageous to leak full prefix lengths between themselves. While a region has a flat routing table, providers in that region can ignore the detail in the majority of the global table. The interconnection robustness at the scale of this example is fairly high, so there is potential for

significant gain. Within the vast regions, it becomes a mater of local politics and business practice as to how much further the plan can go, or how additional existing exchanges might be leveraged. Certainly evolving to a structured interconnect model will be more difficult in either Europe or the US than all the others combined, but if those PI regions are initially written off as hopelessly intertwined, there is still an opportunity for significant gain when the rest of the world is able to ignore the details of that interconnection mess.

A further reduction is possible by starting with three groupings.
London/Amsterdam : Tokyo/Osaka : Chicago/New York

90.00000 s	90.00000 e	Tokyo	xC00::/10
60.00000 s	90.00000 e	Tokyo	x000::/6
90.00000 n	90.00000 e	Tokyo	xE00::/10
90.00000 s	180.00000 e	Chicago	xC40::/10
60.00000 s	210.00000 e	Chicago	x400::/6
90.00000 n	180.00000 e	Chicago	xE40::/10
90.00000 s	0.00000 e	London	xD40::/10
60.00000 s	330.00000 e	London	x800::/6
90.00000 n	0.00000 e	London	xF40::/10

(Getting the polar sections to map to even binary units requires dividing 360 by 2^n. Given the land mass alignments, it makes sense for n to be 2, with 0 – 90, 90 – 180, & 180 – 360 groupings.)

```
>-------------------global service providers------------------<
      |                        |                        |
      |                        |                        |
  Asia IX                  Amer IX                 Euro IX
      |                        |                        |
      |                        |                        |
 -----------      regional service providers      -----------
  /   |   \              /     |     \             /   |   \
 /    |    \            /      |      \           /    |    \
  Local IX's               Local IX's              Local IX's
      |                        |                        |
      |                        |                        |
 -----------        local service providers       -----------
```

Engineering considerations

Many private-peer connections exist to avoid the performance limitations of a shared interconnect. These limitations include both the interconnect fabric, and the access paths between the fabric and the provider network. While not as simple to operate as exchange interconnections, these peering points are an engineering necessity for scale. Fortunately, both interconnect strategies work with the PI address format, as long as the scope of the advertised PI prefix is contiguous (ie: there is a path between the private interconnect and the shared fabric when the prefix applies to both).

One engineering consideration is that the size and location of an exchange has no mandatory relationship to the prefix lengths exchanged there. For example, assume there is a massive exchange in London with hundreds of providers participating covering all of the UK, and nearby another exchange, say Moscow, where there may only be tens of providers, but they cover all of Russia from a single exchange point. The fact that one has more participants, or covers a region approximately 10 degrees square, while the other covers a region 20 x 150 degrees has nothing to do with the number of multi-homed sites supported. A single exchange in each may be inadequate,

or may be overkill for the required service. The requirement is that all participants agree on the set of prefixes to be exchanged, and that set will almost assuredly contain multiple lengths to avoid overlapping with a neighboring exchange. The existing Regional Registries for PA format addresses already have the appropriate constituency and fora to act as a catalyst for the necessary agreement on prefixes at each exchange point.

It should be noted that when a site is directly connected to an exchange, the exchange becomes the logical customer of the transit service providers that tie the exchanges together. In this context the exchange itself appears to be one of the service providers for the regional aggregate. While the current set of exchanges are not likely to scale to support millions of multi-homed sites for a specific scope, in the long-run the location and number of exchanges will evolve to meet the engineering cost/benefit analysis. The design of the PI mechanism allows for the creation of exchanges at the scope that makes local engineering sense, without impact on any other scopes.

Observations and Considerations

Address Selection

IPv6 defines that interfaces will have multiple addresses, so having

a PI set as well as potentially several PA sets should not present any particular concerns to the end nodes. The primary technical issue will be limitations in the size of a DNS response packet. Using both the PI and PA prefixes, multi-site networks SHOULD internally advertise all of the appropriate natural prefixes for the connections the network manager is willing to use, then let the host address selection rules [11] sort it out. Due to longest match selection the default rules would result in systems using a source address that most closely matches one for the destination. When the destination is single-homed return traffic would naturally be directed toward the site boundary closest to the destination site (ie: traffic would traverse the public network as little as possible). If this is not the desired behavior, local policy may establish an appropriate set of rules to reorder the end system preferences.

While broad advertisement of available prefixes provides the most robust infrastructure to the end systems, managers of large multi-national organization networks should exercise operational care to administer the distribution scope of any prefix. It is unlikely that nodes in a 10,000-seat office complex would be expected to use the local Internet access provided for a 3-person office halfway around the world. When this policy is true, the small-office prefix SHOULD NOT be propagated beyond that local office, because doing so would only clutter and slow the address selection process for the larger segment of the organization's network.

The longest match algorithm will automatically select between PA and PI prefixes. If the two sites share some part of the provider hierarchy and some degree of geographic locality, it will become a case-by-case issue as to which one is longer. On one hand they may be geographic neighbors using different providers with no relationship in the PA based allocation (longest match rule would cause hosts to select PI based prefix). In the contrasting case, they share the same provider but are on opposing sides of the globe (longest match rule would cause hosts to select PA prefix). While the hosts have no direct access to current topology information, the simple longest match rule for address selection would appear to bias connections toward the most appropriate path. In any case, once the packets are sent, traffic flow will follow the inter-provider policy perspective of the optimal route.

In the case where one site is single-homed (therefore using a PA prefix), and the other is multi-homed using PI, the routing system would not particularly care because these are both global unicast

prefixes and will be handled appropriately. (Creating this situation presumes that the multi-homed site is not informing its hosts or DNS about any PA prefixes it may have, or has a local policy overriding the default selection rules.) In fact this may be a useful case for a content provider trying to do global load distribution, though it would require the PA node to be aware of its PI prefix, even if it was never used in a packet.

Path Selection

A frequently asked question is how a source selects the correct first hop when more than one exists? This is actually a multipart question since it involves both the address selection as well as the first hop router selection.

Many arguments about address selection revolve around the host's knowledge (or lack thereof) about the technically optimum path for the metrics of bit-rate, loss-rate, delay, and jitter, but they generally avoid the topic of actual access cost, which is all the site usually knows. Address selection was dealt with in the previous section, and lacking local policy to the contrary, will be based on longest match between the source and destination.

A fundamental characteristic of IPv6 hosts is that they will always choose one of the available routers, and expect to be redirected by the routers which actually know at least part of the optimal path. This set of routers for a site will be managed according to local policy and will forward or redirect in that context. While many discussions assume the destination route announcement determines the source's routing; the reality is the holder of the packet always decides based on its perspectives of cost. The source policy has always determined at least the first hop, and any intermediate policy may bias the route at any point by ignoring any announced destination policy.

Partitioning

One of the concerns that aggregation through an exchange raises is the potential for a portion of the local topology to partition. This would effectively create a black hole for sites that are only attached to the disconnected partition. While this is clearly a problem for single-homed sites, those sites should be using PA space and not be subject to the aggregation of PI prefixes. For those multi-homed sites using PI in a metro-aggregate context, their exposure to partitioning occurs when all of their local providers

partition from the set of transit providers at the same time. The potential for simultaneous partition raises the case that any metro interconnection topology could create a single point of failure, which further leads to a strong recommendation that these metro interconnects actually consist of 2 or more interconnected fabrics per scope. The routing implications of this are that the number of BGP speakers will increase in proportion to the number of fabrics, but as long as the set of prefixes match they will appear to be one logical exchange point. In any case partitions can be locally healed with explicit routing entries in the interconnecting providers, and the rest of the world does not need to be aware.

Renumbering

Even though this address format is derived from geographic information, renumbering is not required as devices move within a network. The only time renumbering becomes a concern is when the layer 1 demarcation for the network changes. In this case all of the attached devices would renumber together, just as they would for a change of providers when using the PA prefix model.

Relationship to telephony addressing model

It has been noted that the PI format shares characteristics with the global telephone address plan (an alternative PI aggregation scheme, discussed in GAPI [12], is a closer match to the traditional telephony model of allocations). While the distribution of prefixes to specific geographic areas would appear to be similar, the telephone environment address space was divvied up in a pseudo-random way where the resulting provider boundaries aligned with the political notion of geography at one point in time. The PI address format is devoid of any political context (beyond agreement on WGS-84 as the reference tool), and allows for structured aggregation at any bit boundary. Unfortunately the cost models for circuits still align with political notions of geography. This situation is expected to ease as the telephony system continues its efforts at deregulation and privatization. The one place where there is a strong similarity between the addressing models is the perspective that some providers operate within a geographic area (routing full-length prefixes), while other providers tie the diverse areas together (routing short area aggregates). Thus the common

characteristic is the fundamental model that allows local detail to be abstracted at a distance.

General Considerations

One concern raised by enterprise managers is that a ::/48 might not be enough for some large organizations. Using the PI format, a large organization will almost assuredly have multiple ::/48's to work with. For example, if their facilities covered a contiguous 100x100 meter lot they would have an entire ::/36 to work with, and the types of organizations that will need more than a single ::/48 are also likely to have multiple 100x100 meter lots just to house that number of end systems.

While the IPv6 PI address format is designed to support exchange-based aggregation in the context of various scope sizes, it is not dependent on exchanges as a fabric for its overall route aggregation properties. It will provide efficient route aggregation in a global context when providers in a given scope interconnect by whatever means (ie: common third party providing transit services), such that only the shorter prefix is announced outside that scope. Any provider (including a traditional exchange point route server) announcing a short prefix MUST be able to deliver packets to anywhere in the scope, either directly or through specific arrangements. In the case where a service provider does not interconnect with others in its scope it MUST advertise the longer prefixes associated with its customers.

It is not likely to happen soon, but there is a concern that eventually a few regions may exist with extreme densities (greater than 1M independent multi-homed sites per area 6.5 km/side). When the density of independently multi-homed subnets exceeds 64 per vertical meter, of 6.4 x 6.4m horizontal surface, in 1km tall buildings, an alternative allocation mechanism will be required.

One characteristic that is frequently overlooked is that geography provides distinguished meaning to the term 'home address'. So a node using Mobile-IPv6 [7] with PI addresses as the home address and the current value from the intermittent access provider for the care-of-address could expect to maintain connections across access events. Note this does not mean the geographic address is allocated or even known to the intermittent access point. The routing system doesn't need to know the binding for the geographic address since packets are routed according to the PA care-of-address. The home-agent would need a way to inject its Provider Independent prefix and current binding. This could be a form of tunnel-broker within a region.

When used in conjunction with RFC3306, it would be possible for governments to establish a regional notification multicast service. While they could do this with PA addresses, the burden of connecting to the right multicast would fall on the end user. Using the PI format, a hierarchy of groups could be defined, where very targeted

messages could be multicast efficiently without need for the consumer to understand the technical details.

Recommendations

Attempting to balance the conflicting policies of the service provider operations staff, their business staff, each end site, and any additional service providers will require a clear policy that everyone can rely on for consistent packet treatment. As it is not the purview of the IETF to establish operational policy for independent operators, the closest fit is a recommendation in the form known as Best Current Practice. In that context, this document recommends:

1) that all providers filter out and ignore any announcements that include 5 or more PI prefixes longer than ::/28, originating from a common AS, where the AS path length is longer than 2.

2) an AS number be restricted to those who require injecting explicit policy into the DFZ.

3) metro interconnects actually consist of 2 or more interconnected fabrics per scope.

Policy 1 allows global enterprise sites which need to inject global policy into the DFZ, to inject up to 4 long prefixes if they can justify to the registries they require an AS number. At the same time it removes the small business and telecommuter announcements from the DFZ because those would have an origin AS from a provider that would most likely be sourcing more than 5 long prefixes. While it removes those small types of sites from the DFZ, it still allows them a degree of provider independence and resiliency in the metro context.

Policy 2 removes multi-homing as an independent requirement to acquire an AS number. If the AS number becomes more difficult to acquire through a change in policy, and service providers employ a filter (either at the protocol level ::/28, or by charging extra per prefix) on AS paths longer than 2 where 5 or more PI prefixes share the same origin AS, the growth of the routing table will be slowed to at most 4x the growth in AS allocations. This change in policies would allow the global enterprise to manage its own policies, while avoiding the table explosion that would happen if every small business or telecommuter appeared in the DFZ. This would also allow neighboring service providers in a region to share detailed information about customers using the PI prefix.

Policy 3 removes the potential for a single point of failure that

would be contrary to the goals of multi-homing resiliency.

It has been noted that the existing inter-provider relationships and settlement models do not align precisely with the concept of

regional aggregation that is recommended here. While this is undoubtedly true, the situation is partially due to the lack of a structured address plan to align with. Other factors that play into the situation are that the perceived costs are not a strict function of distance, and that the industry lacks a rate structure for packet settlement. While the fundamental cost of physical media does have a distance component, the current pricing realities often ignore this. Compounding this situation is the fact that the Internet providers have gone out of their way to avoid hierarchy on one hand, while looking for someone to dump packets to on the other. The resulting relationship complexity could be simplified through packet accounting, but that model runs counter to the current culture that is best described a 'shared pain'.

For the IPv4 Internet, service providers have attempted technological restraint systems through routing filters to varying degrees of success. For the IPv6 Internet the PI address format looks to provide a reasonable tool for aggregation, while allowing well-defined exceptions. Given this environment, economics and human nature will align the interconnect strategies of the service providers over time.

In any case, deploying a new approach will require a significant number of service providers and sites to agree that these recommendations result in a sustainable business model, which actually lowers overall costs. To reach that goal, the PI address model explicitly trades address consumption for simplicity in the derivation and routing, as well as trades maximal routing efficiency for end-to-end system level efficiency.

RFC Editor Considerations

The format prefix x in the examples needs to be replaced by the value assigned by IANA.

Security Considerations

While there may be concerns about location privacy raised by the geographic scheme, there is nothing inherent in this address format that would raise any more security considerations than any other global addressing format. If location privacy were an issue it would

be wise to avoid this mechanism in favor of location independent mechanisms such as provider based allocations.

Relationship to the IETF Multi-6 WG multi-homing requirements

The multi-homing requirements for IPv6, consistent with current IPv4 usage, are detailed in [13]. The capability of the Provider Independent address format to deal with each of the points in that document will be addressed here. Since the goal for a short-term approach to deal with multi-homed sites was specifically taken off

the table as the IETF moves between the Multi-6 and extended lifetime architectural change of the Shim WG, the proposed GEO approach is clearly not an overlapping mechanism.

Redundancy

The Provider Independent address format is designed to provide redundancy between attached providers. By having the site prefix independent of all service providers, link and routing failures in one provider should be completely transparent to the site. The primary case where things may break is a link or routing failure in any part of the path that lacks physical redundancy.

Load Sharing

This recommendation for specific applications of the Provider Independent address format will allow sites to manage outbound traffic without concern for undue filtering in the routing system. It also allows for load sharing on inbound traffic by large enterprises that can justify expression of policy in the DFZ.

Performance

The Provider Independent address format allows traffic to arrive from a variety of sources over the set of available paths, but does not explicitly provide for remote flow control. A site may exercise some course level of remote traffic flow management by arrangements for service from multiple providers. At a minimum, traffic from the other customers of an attached provider would follow the site's path through that provider, and not transit any other provider.

Policy

Traffic class alignment as policy routing is not an IP routing issue, and even using PA addresses can only be accomplished by announcing explicit subnet or host routes. As such the Provider Independent address format will not offer any additional explicit support. Achieving the goal of this bullet is probably best met with

a mix of Provider Independent and Provider Aggregatable prefix announcements where the hosts respond to the specific address/port mappings according to local policy.

Simplicity

The target of the Provider Independent address format is simplicity, both in the method of allocation, and in the routing expectations. From the site perspective, an allocation independent of provider is what they are after (ie: PI format). From the service provider perspective, handling GEO [1] type PI prefixes is as simple as IPv4 PI prefixes. The potential increase in complexity over current IPv4 deployments is understanding the impact when site chooses to use both PI and PA prefixes.

Transport-Layer Survivability

The Provider Independent address format explicitly deals with transport-layer survivability by isolating the session from the intervening providers. As long as the routing system converges

within the timeout period of the transport-layer, any active connections will survive.

Impact on DNS

There are no modifications required for the Provider Independent address format discussed here. There is a potential issue with the size of a response packet if a multi-homed site chooses to include all of the applicable addresses. Use of a single PI rather than multiple PA prefixes would reduce this concern, while retaining the immunity-to-provider-failure characteristic.

Packet Filtering

The Provider Independent address format does not preclude filtering inappropriate packets, and may facilitate such filtering since the location of the demarcation point helps reduce any ambiguity.

Scalability

No one approach will solve all scalability concerns. An appropriate mix of Provider Independent and Provider Aggregatable address use will solve most concerns without undue complexity in either the host or the routing system.

Impact on Routers

The impact on routers outside a region is a significantly smaller routing table, both from the reaggregation of the provider prefixes and from the ability to further abstract geographically distant sites. Within a scope, the full routes need to be carried, but this

is no worse than the holes punched in provider aggregates, and can be managed through interconnecting at smaller scopes.

Impact on Hosts –

Hosts may have an additional address to select from if the site chose to use advertise both the Provider Independent and Provider Aggregatable formats. Using longest-match rules should easily sort between Provider Independent and Provider Aggregatable prefixes. Hosts may also want to choose to use this as a distinguished form of 'Home' address for mobile applications.

Interaction between hosts & routing system –

Routers providing for IPv6 auto-configuration through announcement of the site prefixes may have an additional one in the list, or may simply choose to announce only the Provider Independent prefix.

Operations and Management –

The mechanism for deriving the Provider Independent address is specifically designed to simplify this operations issue by using the globally ubiquitous WGS84 system of measurement.

Cooperation between Transit Providers –

The Provider Independent address mechanism does not require cooperation between service providers specifically for a given multi-homed site. It does require all service providers for a given

scope to agree on the boundaries for the scope and any traffic exchange point that might be necessary.

Multiple Solutions –

The Provider Independent address mechanism does not preclude other forms of multi-homing. It does provide a complimentary service to the Provider Aggregatable prefixes for single-homed use, and scales much better than punching holes in those for multi-homed sites.

References

1 Geo, draft-hain-ipv6-PI-addr-10.txt, Hain, T., "An IPv6 Provider-Independent (Geographic Reference) Global Unicast Address Format", August 2006.

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

3 http://www.wgs84.com/files/wgsman24.pdf

4 RFC-1518, Rekhter & Li, "An Architecture for IP Address Allocation with CIDR", September 1993

5 RFC-3587, Hinden, B., Nordmark, E., Deering, S., "IPv6 Global Unicast Address Format", RFC 3587, August 2003.

6 http://kahuna.telstra.net/bgp2/as1221/ , G. Huston

7 RFC-3775, Johnson, D., Perkins, C., "Mobility Support in IPv6", June 2004

8 RFC-1887, Rekhter, Y., Li, T. "IPv6 Unicast Address Allocation Architecture", December 1995.

9 RFC-4193, Hinden, R., Haberman, B. "Unique Local IPv6 Unicast Addresses", October 2005.

10 http://www.ep.net/

11 RFC-3484, Draves, R., "Default Address Selection for Internet Protocol version 6 (IPv6)", RFC 3484, February 2003.

12 GAPI, M.Py, I. Beijnum, A Geographically Aggregatable Provider Independent Address Space to Support Multihoming in IPv6, draft-py-multi6-gapi-00.txt, October 2002

13 RFC-3582, Black, et. al., "Goals for IPv6 Site-Multihoming Architectures", RFC 3582, August 2003

Acknowledgments

Discussion threads on the NANOG and IETF/Multi-6 mail lists provided many of the perspectives presented here. Early feedback was provided by Iljitsch van Beijnum, Brian Carpenter, Sean Doran, Geoff Huston, and Pekka Savola.

Author's Addresses

Tony Hain
Cisco Systems
500 108th Ave. N.E. Suite 400

Bellevue, Wa. 98004
Email: alh-ietf@tndh.net