Network Working Group Mike O'Dell
Internet-Draft UUNET Technologies
1997/02/24 01:32:32GMT

GSE - An Alternate Addressing Architecture for IPv6

<draft-ietf-ipngwg-gseaddr-00.txt>

1. Status of this Memo

2. Abstract

This document presents an alternative addressing architecture for IPv6 which controls global routing growth by very aggressive topological aggregation. It includes support for scalable multi-homing as a distinguished service. It provides for future independent evolution of routing and forwarding models with essentially no impact on end systems. Finally, it frees sites and service resellers from the tyranny of CIDR-based aggregation by providing transparent re-homing of both.

3. Introduction

This alternative IPv6 addressing architecture addresses several scalability issues with the current IPv6 addressing proposals.

Scaling of the global route computation

Ease of re-homing (both leaf Sites and upstream Resellers)

Economic scalability of of Multi-homing

The current IPv6 addressing proposals address route and topology aggregation by continuing to rely on CIDR-style "Provider-based Addressing" coupled with a powerful new dynamic address assignment mechanism which is intended to make renumbering more palatable.

However, CIDR-style provider-based aggregation breaks down in the face of the accelerating growth of multi-homed sites (leaf sites or regional networks). Worse, renumbering an entire Site to accomplish a simple topological re-homing such as changing ISPs is a problem whose magnitude can only grow over time. It will remain increasingly difficult to explain this renumbering requirement to customers with the spectre of a complete failure of this aggregation approach a distinct possibility.

While the large IPv6 addresses provide for a huge increase in the number of end systems which can be accommodated, it also portends a huge increase in the number of routes required to reach them. Even if CIDR aggregation were to continue at current levels (maintaining current efficiency is relatively unlikely), this still presents a serious problem for the growth of the the global route computations.

This document presents a new proposal for using the 16 byte IPv6 address which mitigates the route scaling problem and with it a number of collateral issues. This model provides for aggressive topological aggregation while controlling the complexity of flat-routed regions. It exploits and supports the dynamic address assignment machinery in IPv6 but makes the exact role of that machinery a decision local to a Site. It is therefore subject to engineering cost and benefit analysis rather than being mandatory for simple Site re-homing situations.

This new model also identifies the special work done by the global Internet infrastructure on behalf of multi-homed sites. Rather than continuing the current "Tragedy of the Commons", the multi-homing is isolated into a specific mechanism which is then traceable to and incurred by only those sites wishing to subscribe to this capability. Again, this makes it possible for sites to make informed cost-benefit decisions about multi-homing.

4. Central Concepts of the Architecture

The architecture is based upon a few central concepts.

A strong distinction between Public and Private Topology

A strong distinction between system identity and location

GSE - Global, Site, and End-system address elements

The deep similarity of Re-homing and Multi-homing

Rewriting address prefixes at Site boundaries

Very aggressive hierarchical network topology aggregation

Optimizing actual forwarding paths by limited-scope cut-throughs

This model draws a strong distinction between the Public Topology which forms the transit infrastructure of the Global Internet and a "Site" which can contain a rich but strictly private local network topology which cannot "leak" into the global routing machinery. The Site is the fundamental unit of attachment to the Global Internet and is therefore strictly a leaf, even if possibly multi-homed.

This model also draws a very strong distinction between the identity of a computer system and where it attaches to the the Public Topology. In IPv4 and current IPv6 models, these notions of identity and location are deeply co-mingled and this is the fundamental reason why simple topology changes have such wide-ranging impact on address assignment (if aggregation is to be maintained at all).

The 16 byte IPv6 address is split into 3 pieces:

```
  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
|  Routing Goop    | STP| End System Designator |
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
       6+ bytes   ~2 bytes       8 bytes
```

Routing Goop signifies where the Site attaches to the Global Internet. The Site Topology Partition (STP) is Site-private "LAN segment" information. The End System Designator (ESD) specifies an interface on an end-system.

One surprising notion is that re-homing and multi-homing are very deeply related. Multi-homing can be viewed as rather like several simultaneous re-homings happening at once. Achieving both painless re-homing and scalable multi-homing rely on the same set of fundamental mechanisms, each with a few distinct details.

Rewriting IPv6 addresses by Site Border Routers is by far the most controversial, but also most critical part of this proposal. To control the complexity of routing information which must be managed within a Site and to isolate end systems and interior routers from external topology changes, the RG of some addresses is modified by Site Border Routers. Packets exiting a site have the RG for the Site

egress point inserted into source addresses, while packets entering a Site have the RG in all destination addresses replaced with a canonical prefix signifying "within this Site" (the "Site-local prefix").

One immediate result is that upper-layer protocols must use only the ESD for purposes such as pseudo-header checksums and the like. The ESD is the invariant token, the RG is possibly transient topology information subject to change.

Topology aggregation is accomplished by partitioning the Global Internet into a set of tree-shaped regions anchored by "Large Structures". The Routing Goop in an address specifies a path from the root of the tree (the Large Structure) to a point in the topology; in the terminal case this is a Site. Large Structures are chosen by their ability to aggregate topology and no particular advantage flows from "being one"; actually quite the contrary. Large Structures are responsible for subdividing the space under them and managing that delegation. Large Structures provide a "forwarding token of last resort" which can always be used for selecting a valid next-hop when no other information is available. This significantly limits the minimally-sufficient information required for a "default-free" router. Any additional route information kept is the result of path optimizations from cut-throughs.

While it is useful to think of the Large Structures as trees, the collection is actually a DAG (Directed Acyclic Graph) because the trees can touch each other via cut-throughs. By cross-propagating selected details via a cut-through, a locally-controlled region can learn of alternative paths to some destinations. The distance this optimization information is propagated and the radius of the optimization region advertised are the business of the collaborating regions.

5. The Structure of End System Designators - the ESD

End System Designators denote every computer system in the GSE Internet regardless of whether it is a host, router, or other network element. While a given system can have more than one ESD, each ESD is globally unique. This is critical for their utility to the upper-level protocols. This uniqueness can be induced several ways as will be seen.

A crucial design decision is whether an ESD identifies a system, invariant of its interfaces as in the XNS architecture, or an interface on a system as in the existing IPv4 and IPv6 architecture.

An ESD designates an interface on a computer system and that

interface can be either physical or virtual.

When processing a GSE address, a computer system need only examine the ESD portion of the address to determine whether a packet is destined for that system.

There are circumstances when it is quite useful to have "an address" for a computer system which is independent of any particular physical interface on that system. It has become commonplace in IPv4 practice to use a distinguished virtual interface to provide a system with such an "interface independent identity". This technique affords the same architectural utility of XNS while still allowing the flexibility of the IPv4 "addressed interface" model. This model retains the successful IPv4/IPv6 model.

NOTE: We remain intentionally vague about exactly what constitutes an "interface" and a "computer system". The malleability of those notions in IPv4 has proven manifestly useful in practice.

To summarize the ESD uniqueness characteristics:

(1) an ESD is globally unique
(2) an ESD designates an "interface" on "a computer system"
(3) an Interface may have more than one ESD (current IPv6 already requires implementations to support multiple addresses per interface)
(4) an ESD may not necessarily designate a particular physical computer (Neighbor Discovery continues to provide a level of virtual address translation and considerable cleverness can be disguised therein)

There are two forms of ESD, both 8 bytes long, one a subcase of the other.

It is clear that with the impending onslaught of the IEEE-1394 technology that 8-byte IEEE MAC addresses are simply fait accompli and many devices will be provided with a unique identity in that format at the time of manufacture. The 8-byte IEEE MAC Address format includes the current 6-byte MAC Addresses as a proper subspace. Using the 8-byte IEEE MAC address will be very convenient for many network builders.

There are at least two issues with using *only* the IEEE 8-byte MAC addresses as ESDs: There are point-to-point link interfaces which have no IEEE MAC address assigned for them, and the 8-byte IEEE MAC addresses assigned to the interfaces of a system are essentially random. For some, there is also the issue of whether the IEEE MAC address is "unique enough" for the purposes at hand.

We clearly need a space for generating ESDs for interfaces which don't come equipped with one. Some have also suggested there might be great utility in enabling inverse lookups on just the ESD part of an address. Assigning ESDs in semantic clusters (like current IPv4 addresses) would be a signficant aid to this end. Finally if a network designer decides not to trust the uniqueness of the IEEE MAC addresses, he could always use the Dynamic Numbering machinery of IPv6 to assign ESDs.

We propose that the IETF seek a large (7 bytes or greater) subspace of the IEEE 8-byte MAC space for allocation as IETF-NodeIDs in semantic clusters to provide a pool of addresses which can be used for any of the above reasons, as required. However, it is expected that most network builders will exploit the intrinsic IEEE MAC addresses present in many network interfaces whenever possible.

The IETF-NodeID space should be partitioned into two regions - one exactly isomorphic to the existing IPv4 address space to provide instant grandfathering of IPv4 addresses, and another space which is simply larger but allocated in a similar manner.

A few comments on "global uniqueness" are in order because in previous discussions, some have asserted that unless "uniqueness" can be accomplished with absolute and complete mathematical perfection, any scheme using the concept is unworkable. This extreme view inconsistent with mass-market experience.

IEEE MAC addresses are globally unique by nature of the delegation process where they are assigned to interfaces by the manufacturers. Both XNS and IPX rely on this uniqueness and it works very well in practice. IETF-NodeID values will be globally unique by nature of the same kind of assignment mechanism. IPv4 addresses must be globally unique for the Internet to function, and it does, mostly, by nature of exactly the same kind of assignment mechanism.

While accidents and manufacturing defects do occasionally violate the uniqueness of IEEE MAC address assignment, humans routinely make errors in assigning IPv4 addresses to systems with equally mystifying results. Given the reliance of IEEE-1394 Firewire interconnects on these unique MAC addresses, it is likely that the frequency of these occurence (relative to the total number of objects with assigned addresses) will only decrease. The economic pressure to insure this will be intense.

6. The Structure of a Site

The GSE global routing architecture ultimately views a Site as a leaf

of the topology and doesn't concern itself with the interior of this private topology. However, the internal topology of a Site is extremely important to the management and operation of the Site so the GSE address architecture provides for a rich set of organizational alternatives with different cost-benefit tradeoffs.

The GSE address structure provides for 16384 distinct Site Topology Partitions (STPs) within a Site. This is the number of SEGMENTS in the internal topology, not hosts. The number of attached hosts is limited strictly by available local network technology, and the Site's ability to buy enough machines to exhaust the available IEEE 8-byte MAC address space, or the available 7-byte IETF-NodeID space.

Using this structure, a single Site can develop an internal topology which is a very significant fraction of the total CIDR routes in the IPv4 Global Internet.

An organization is not constrained to being structured as a single Site. The trade-off is that the inter-Site topology must then be part of the Public Topology. While the individual Sites can retain considerable independence in topological structure and attachment to the Global Internet, they must be aware of changes between the constituent Sites and that re-homing of constituent Sites will potentially impact long-running sessions. That is the cost of exploiting the routing machinery available to the Public Topology.

Given the generous flexibility available for organizing a Site, it is worthwhile to examine a few examples. Note that none of these organizational approaches is exclusive. A large Site might well mix these approaches to good effect and indeed the goal is to provide the designer of private Site topology with a broad spectrum of design alternatives.

The simplest structure to imagine is a Site using all IEEE MAC Addresses with all the systems connected in a single Private Topology Partition (i.e., all the GSE addresses carry the same STP value which is assigned by the local network administration). Given the sophistication of current LAN-switching technology, a Site like this could be both large and internally complex yet have simple IPv6 addressing. The complexity is absorbed into the LAN infrastructure and it appears to be only one partition from the GSE Site Topology view. This structure has one very significant advantage: long-running TCP sessions will will survive arbitrary changes in the local topology. This works, of course, because the single STP is a virtual topology with the real topology hidden by the LAN Switching machinery.

The second Site model is like the one just described, except it would

have multiple STPs with routers moving traffic between the segments. This is very close to the common IPv4 structure of a CIDR block being subnetted to assign a prefix to each STP. This approach has the advantage of familiarity, but it has the disadvantage that long-lived TCP connections don't necessarily survive arbitrary changes to the private topology. This arises because even though the ESD is invariant, reachability will fail because a change in the STP of one of the system doesn't get injected into the protocol stack of the communicating systems when they move. The existing IPv6 dynamic address assignment machinery will serve to make such internal changes much less painful than with IPv4, however.

One point worth noting is that even with multiple STPs routed within a Site, a "Private Topology Partition" need not correspond to a "physical" LAN cable. The STP values could be used to label larger organizational structures like "Engineering" or "Finance". This could reduce the likelihood that common internal topology changes break long-lived connections.

The third Site model uses IETF-NodeID ESDs based on existing IPv4 address assignments. In this case, all the IPv4-style ESDs could be placed in a single STP and then routed internally on the IPv4 address in the lowest 4 bytes of the ESD. It must be emphasized that the IPv4 addresses used in IPv4-style ESD must be an officially-registered, public-use IPv4 address and NOT an RFC-1918 private-use address. Using an RFC-1918 private-use address violates the global uniqueness properties required of an ESD.

In all of the multi-segment cases, an IETF-NodeID ESD could be used to designate any point-to-point link endpoint, the loopback addresses in routers, or any other IP-accessible network elements which don't naturally have IEEE MAC address for forming an ESD. And in all of the cases, an IETF-NodeID ESDs could be used universally, although it is more appropriate to use IEEE ESD form whenever possible.

In all of the cases where the real topology is not completely virtualized by the LAN technology, there will be "Internal Renumbering" events caused by moving systems between infrastructure segments (STPs). This will have the effect of killing long-running off-Site connections unless provisions are made to allow the systems (and the routing infrastructure) to carry the previous ESDs as synonyms for a while. Given that most significant topology moves involve powering off the end system in question, this is hardly a hardship. However, the powerful renumbering support already developed for IPv6 can make those other moves considerably less impacting.

Most importantly, external re-homing of a Site to the global

infrastructure can be made completely transparent.

7. Dynamic Address Re-writing by Site Border Routers

A critical component of this architecture is the modification of addresses when packets leave or enter a Site. Re-writing source addresses to insert appropriate Routing Goop at the Site egress point was part of the 8+8 proposal, but this proposal extends this to re-writing destination addresses when inbound packets arrive at a Site Border Router.

The reasons for both re-writings are the same: to insulate the interior of the Site from external topology changes and egress policy details.

When a Site Border Router inserts the correct RG in the source address of outbound packets, it frees the end-systems in the Site from having to know the RG for the Site. This is especially important if the site is Multi-homed and the Site implements a complex egress selection policy.

In the case of inbound packets, if the destination address were not converted to a canonical form, the Site interior routers would have to be aware of all the different RG which could be used to reach the site, essentially creating aliasing of the destination addresses. In the singly-homed case, this doesn't seem like a significant issue, but in complex Multi-homing scenarios there could be a significant problem managing this information.

This symmetric re-writing essentially isolates the Site from the Global Internet just as the hard boundary between RG and STP components insulates the Global Internet from the Site topology.

8. The Structure of Routing Goop

Routing Goop, or "RG" is the upper 6+ bytes of a GSE address. This somewhat non-technical term was chosen because all the other alternatives seem to have various degrees of conceptual baggage which would be as much work to neutralize as the new notions are to explain in the first place.

Fundamentally, RG is a Locator. It encodes the topological connectivity of the Site containing the computer system identified by the ESD in the lower 8 bytes. In the case of a singly-homed Site, re-homing to a new attachment to the Public Topology will change ONLY the RG in full GSE addresses for computer systems at that Site. One example of such a re-homing would be a change of the Site's Internet Service Provider. This change-over can be made essentially

completely transparent to users both inside and outside the Site, although it does involve a practical limit on the transition duration relating to how long the departing ISP is willing to extend transitional courtesies. During a changeover, though, all new connections will be initiated via the new ISP connection.

This brings up the deep structure of the topology information carried in RG and how it is encoded. More specifically, RG is a hierarchical locator which is a rooted path-expression of flat-routed regions which are tangent. Each element in the path-expression includes only enough detail to negotiate the flat-routed region.

It has been observed before that the graph of the Global Internet is not obviously a hierarchy so how can this work?

We start with the observation that every connected graph has at least one labeling which forms a spanning tree covering the nodes. The hierarchy is induced by a labeling function which partitions the global graph into regions and recursively into subregions. This function is only globally visible at the top-level where an initial partitioning of the graph is used to form the first level of what will become the hierarchy. Within each partition there is a local sub-partition function which assigns labels, and we proceed recursively. The nested recursions directly induce the hierarchy.

This decomposition of the Global Internet produces a recursive graph where each level is composed of a set of subgraphs which are explicitly connected (i.e., explicitly routed between the subgraphs) while the structure within each subgraph is assumed to be flat-routed (at least as seen at that level).

From an abstract viewpoint, a hierarchical partitioning can be induced with an arbitrary choice of labeling function (as long as the function produces the minimally-required partitioning). However, we desire the partitions to have several important properties which effects the choice of labeling function.

The general goal is to produce a global labeling which represents the topology as compactly as possible, yet allows rich connectivity while bounding the complexity of the discrete regions which are flat-routed.

The top level objects in the GSE graph hierarchy are called "Large Structures". These are objects chosen for their ability to naturally represent significant topological aggregation of substructure (not geographical, political, or geometric). The number of Large Structures is explicitly limited to bound the complexity at the top level of the aggregation graph.

Within Large Structures, the (sub-)partition function is a trade-off between the flat-routing complexity within a region and minimizing total depth of the substructure. This is driven by the internal topology of a Large Structure and the choices in different Large Structures will not necessarily be the same. This is why Routing Goop only has one hard bit boundary; Large Structures are free to internally subdivide as they chose. They are only required to encapsulate a significant portion of the Public Topology.

One obvious candidate for Large Structures is large networks which already represent considerable aggregation based on existing CIDR deployment. Another good candidate might be "Exchange Points". The GSE model can accommodate both of these simultaneously, allowing IPv6-style "Network-anchored Prefixes" and "Exchange-anchored Prefixes" like that proposed by some to coexist and be subsumed into a unified notion of "Aggregator-anchored Prefixes." Of course, these aren't prefixes strictly in the IPv4 CIDR sense, but the left-anchored substrings of the Routing Goop are intuitively quite similar.

Large Structures are assigned a Large Structure Identifier, known as an LSID. The total number of LSIDs is intentionally limited as we assume the paths between Large Structures are only flat-routed.

Two consenting Large Structures remain free to share a tangency below the top level and exchange routes so as to provide for improved routing between the two of them (formalizing cut-throughs in the natural hierarchy). The goal is to provide for manageable complexity of the ultimate default-free zone (the top level of the global hierarchy) while allowing for controlled circumvention of the natural hierarchical paths.

Bit-level structure of Routing Goop:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| xxx | 13 Bits of LSID         |      Upper 16 bits of Goop    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

 3               4                   5                   6
 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Bottom 18 bits of Routing Goop    | 14 bits of Site Topology  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

NOTE: The Routing Goop structure above assumes that the GSE proposal is designated by a 3-bit type of IPv6 address. If a GSE address is identified by two upper bits, the LSID would expand to 14 bits. If identified by one bit, the LSID would stay at 14 bits and the Upper 16 bits of Goop would expand to 17 bits.

Routing between two interior points of two different Large Structures is always possible based solely on the LSID. This provides a "forwarding strategy of last resort" for a router running "default-free". From one point of view, the LSID partitions the Global Internet into a set of regions such that an interior router only need carry a "per-LSID default" pointing at an appropriate boundary router which knows how to to handle traffic bound outside the containing Large Structure for a point in the other Large Structure.

If two Large Structures share a tangency somewhere below the top level, then some interior routers of both Large Structures will share routes to exploit the tangency for optimizing paths. How this cut-through information is distributed within the two Large Structures is not revealed elsewhere in the global topology. The exact "shape" of the optimization region is controlled by the decisions about which routes to advertise across the cut-through. These decisions are made by the collaborators and the optimized region need not be symmetric with respect to the cut-through. The size of the optimization area is controlled by how far routes learned via the cut-through are propagated within the sub-graphs tangent via the cut-through. Again, this is a matter of engineering choices made by the collaborators operating the cut-through.

While the LSID is may appear similar to the Autonomous System Number currently used in IPv4 policy-based routing machinery, the LSID is quite distinct from the AS number and the two identifiers play very different roles. AS Numbers will continue be used for policy routing information exchange and must remain distinct from the LSID space.

9. The "Flow" of Routing Goop

It is intuitively useful to think about Routing Goop as "flowing downhill" through the hierarchy from the topmost Large Structures, through the intermediate levels of the Public Topology, and ultimately down to the Site. As the RG propagates downward, the prefix extends to the right, just like in IPv4 CIDR, with each extension navigating the nested flat-routed subgraphs, eventually terminating at the Site, which then descends invisibly into the Private Topology of that Site.

The nested flat-routed areas correspond to transit subnetworks of the Large Structure. One very important example of such subnets is the "reseller" or "wholesale transit customer" of a Large Structure. (Note that whether the Large Structure is a network or an exchange point doesn't matter.) The reseller network provides transit for Sites, so must be part of the Public Topology and appears as a substring within the Routing Goop, usually the right-most extension unless the reseller has further reseller customers. In that case, the next level reseller will have his own extension to record his place in the Public Topology and to provide for navigating through it as well.

The overall picture can now be drawn as a forest of trees distributing Routing Goop down to the Sites, with each tree being a Large Structure and the Large Structures connected arbitrarily at the top level. This structure will be mirrored by the actual machinery for distributing Routing Goop to the Sites as will be discussed a bit later, but this mental image of the prefixes "flowing" from the anchoring Large Structures is critical to understanding fundamental self-organizing abilities in the GSE model.

While the GSE machinery is intended to be adequate for almost completely automated self-organization with respect to the construction and propagation of Routing Goop on an Internet-wide basis, we proceed for now closely following current practice (admitting manual configuration of certain information like Routing Goop) because of the additional complexity of the self-organization functions. Initial deployment following current practice would not preclude eventual deployment of a fully self-organizing Global Internet.

10. The Distribution of Routing Goop

There are two cases to consider for how Routing Goop gets distributed: source addresses and destination addresses. In both cases RG is part of the address, one way or another, so we show how a full 16-byte address with the right RG gets created in these two

cases.

10.1 RG for Source Addresses

The initial RG of a source address is almost always the Site-local prefix. If the destination address is not within the Site, the packet will leave the Site via one of possibly several Site Boundary Routers. The egress Site Border Router will insert the correct RG in the source address based on the path the destination should use to return a packet to the sender. Except in unusual circumstances this will be the RG which corresponds to the attachment path of that egress Site Boundary Router to the Global Internet.

If the Site is multi-homed via just one Site Boundary Router, then the router is free to apply whatever local policy suits. It simply must fill in a valid RG path which leads back to a Site Boundary Router for that Site. If the Site is multi-homed via more than one Site Boundary Router, which router provides egress is purely local policy and which RG gets applied is likewise local policy.

The dynamic insertion of RG upon Site egress accomplishes a number of things.

(1) It means that for most purposes, a computer system at a Site need not concern itself with egress policy matters which can be particularly tricky in Multi-homed Sites.

(2) It means that computer systems are essentially not impacted at all by topological re-homing of the Site.

(3) It means that more complex multi-homing scenarios with multiple Site Boundary Routers each with multiple connections to the Global Internet can execute arbitrarily complex path recovery policy without concern for how it might impact a computer system doing source address selection.

(4) It means that while a computer systems might forge the ESD in a source address, it CANNOT forge the point of injection into the Public Topology. This is not strong authentication down to the particular computer system, but it is probably a strong deterrent to certain obnoxious activities due to the dramatically improved traceability. We also note that the first-hop attachment router in the Public Topology is free to insert or override the RG if somehow an errant packet escapes a Site carrying invalid RG, thereby enforcing traceability. Of course, the Public first-hop router could always just drop a packet carrying inappropriate source RG as well. But to make it very clear, we put the burden of inserting correct RG in exiting source addresses squarely and solely on the Site and the

Site Border Router. Any other location of the task has bad performance scaling.

The Site Border Router acquires the necessary RG from the first-hop attachment router in the Public Topology. Alternately, as an initial mechanism the RG could be statically configured, but the real goal is completely automated propagation down the tree so that an entire complex subtree can be rehomed without human intervention or service disruption.

10.2 RG for Destination Addresses

Currently, an IPv6 address lookup for a DNS name returns the information in a "AAAA" record which is the full 16 bytes of the IPv6 address.

The GSE design proposes synthesizing the 16 bytes of information in a query response from two different sources: an "AAA" record and an "RG" record. The "AAA" record carries the 8-byte ESD + ~2 byte STP for the DNS name in question and the "RG" record carries 6+ bytes of the appropriate Routing Goop.

One interesting question is how the AAA record gets paired with an RG record in a given nameserver. One simpleminded implementation would be to pair an RG record with a zone, but that has the problem of requiring all the systems in that zone to use the same Routing Goop and hence be in the same Site.

A better scheme is to carry an "RG Name" in the "AAA" record which would allow a nameserver to concatenate an arbitrary RG prefix to the ESD+STP producing the full 16 byte response. The "RG Name" would be a full DNS name which could be recursively translated (and the result cached). Structured as an "upward delegation" with an appropriate Time-to-Live, a Site could import the Routing Goop information from their service provider completely automatically. This capability will be used to great advantage in the discussions of re-homing which follows. [Interactions between RG TTL and zone TTL is an issue to be explored more.]

Alternately, one special case for an RG record could be a delegation to a Site Border Router which could supply the correct RG automatically, at least in single-homed cases, and possibly in multi-homed cases.

The result of this structure is that individual zone entries for individual nodes (AAA records) do NOT change when a Site rehomes. The only thing which changes (logically) is the RG information which is composed with the node's AAA record to produce a full 16-byte

response. This means the general Dynamic DNS machinery is NOT required to support Site re-homing.

One implication of the special Site-local Prefix RG for intra-Site traffic is that Sites will have to provide at least two "faces" on their nameservice - one that returns Site-local as the RG for queries from inside the site, and another that returns full RG responses for requests originating outside the Site. This can be readily accomplished by inspecting the source address - if the source address contains the Site-local Prefix as RG, then return the same. Otherwise, return a fully-general RG-based response (possibly based on egress-path selection policy).

10. Re-homing A Site

When a Site changes its point of attachment to the Global Internet, it is said to "rehome". One of the significant criticisms of IPv4 CIDR and IPv6 "Provider-based Addressing" is the requirement to "renumber" a Site when it rehomes. One of the explicit goals of the GSE architecture is to eliminate, or at least mitigate, the impact of this.

It is important to reiterate the notion that the Routing Goop of a GSE address is not just a Locator, but that it encodes a PATH from the top level of the global hierarchy down to the Site. Changing that path is what makes Re-homing and Multi-homing essentially equivalent operations. We proceed with the simple case first.

When a Site wishes to rehome, it must establish a new attachment point to the Global Internet, and hence establish a new access path. Then it must start using that new path before the old path is removed. The procedure is as follows:

A Site establishes a connection with a new ISP and it becomes able to carry the traffic. At that point, the Site alters the upward delegation of the DNS RG records. Henceforth, all new connections made with the new translations will follow the new path to the Site. The new connection path is then made the preferred egress path and source addresses in packets exiting the Site immediately start being marked with the new return path. The old connection should be maintained for some administratively determined grace period to allow DNS timeouts to transition new sessions to the new path and for long-running sessions to terminate.

At first blush, it might appear that when the egress path for the Site switches over to the new path and the Site Border Router starts marking packets with the new RG, the return path for long-running sessions would automatically switch over to the new path. Alas, this

is not so because a long-running session will be using destination address containing the old RG acquired when the session first started.

Consideration was given to providing some kind of "path redirect" which would allow the other end to deal with "flying cutovers" of a running session, but the security implications of this mechanism are too far-reaching to consider as part of initial deployment. If at some later point it becomes clear how to accomplish this safely, then it could be added. But the complexity, security risks, and the magnitude of the added value do not seem worthwhile at present (although the author would love to be convinced otherwise).

Alternately, the Site could request a "Re-homing Courtesy" from their old ISP which would effectively make it a multi-homed Site for some period of time. After multi-homing was established, the old connection could be taken down and the long-running sessions would continue to survive as long as the Site was multi-homed by way of the Re-homing Courtesy.

Note that at no time did the re-homing effect anything internal to the Site's Private Topology. The only change was the attachment to the Public Topology and the Routing Goop which records that attachment location.

11. Multi-homing a Site

One of the curiosities of IPv4 is that the network does a lot more work for a multi-homed site but it is very hard to pin it down so that the instigator of the effort can compensate the workers.

In the GSE model, Multi-homing is an explicit service which is performed for a Site by the agents of the Public Topology which provide the access for the Site. This mechanism can be made more sophisticated, but the notion is most readily explained by considering a Site which is dual homed to two different ISPs and hence has two distinct access paths represented by two distinct blobs of Routing Goop.

The Site is attached to each ISP via some link and we postulate some kind of keep-alive protocol which determines when reachability to the Site's border router is lost. The ISP routers serving the dual-homed Site are identified to each other (via static configuration information in the simplest case or a dynamic protocol in the more general case), and when a link to the Site is lost, the ISP router anchoring the dead link simply tunnels any traffic destined for the Site via the other ISP router.

This approach clearly requires coordination between the two serving ISPs. This is not a new constraint - multi-homing already requires considerable coordination between the Site and is providers. Of course, creating a protocol for dynamically creating a "homing group" is probably a very worthwhile investment but it is not absolutely necessary at the outset.

It should be obvious now that the "Re-homing Courtesy" in the previous section is simply doing the router-pair coordination with the new ISP for some period of time.

[Note: Yakov and Bates are working on a draft for a Site-side implementation of aggregation-efficient multi-homing which may simplify this even further.]

12. Re-homing a Reseller

Re-homing a Reseller is a slightly more general case of re-homing a Site, primarily characterized by more lead time, a longer grace period, and some necessary coordination with customer Sites to insure that the Routing Goop propagates correctly.

The Reseller will establish a new connection which will not only result in a new path for the Reseller's topology, but for that of his customer Sites. When the Reseller alters his upward delegation of Routing Goop, it will ripple downward to his customer Sites by nature of their upward delegations. The downward ripple of Routing Goop via the upward delegations should cause the Site zone TTLs to be reduced appropriately to insure caches expire well within the dual-homed transition grace period for the Reseller.

This essentially rehomes all the Reseller's customer Sites all at the same time the Reseller's infrastructure is re-homing and should be completely transparent except for long-lived sessions which do not terminate by the end of the grace period.

13. Multi-homing a Reseller

There are two parts to multi-homing a Reseller - one part similar to the multi-homed Site case above, and one part which is quite different.

For this discussion, assume a Reseller which is dual-homed and hence has two different Routing Goop prefixes (remember that each path to the top level of the hierarchy has a distinct prefix). The reseller can solicit multi-homed tunneling services from his two access point routers to provide alternate path service just like a multi-homed Site. Why traffic is coming to any particular router, though, is

influenced entirely by what routes are advertised out that particular connection via BGP5 (or IDRP). This is rather different from the multi-homed Site case where the ESD is the object of interest and the RG simply gets the traffic to the Site boundary.

The question arises, however, as to which prefix gets used for extending downward to his customer Sites. The answer in the simplest case is to pick one and use it, making the Sites "natural" in the chosen prefix. The alternate prefix can, of course, be advertised out the alternate path if desired. But this work can be ascribed to the instigator and the superior attachment points can charge for this service. (This is somewhat akin to charging for routes, but only routes which create a discontinuity in the routing space.)

15. A Comment on NAT Boxes

In discussions about requiring destination address re-writing for inbound packets, Brian Carpenter remarked that with the advent of symmetric re-writing (both inbound and outbound), the GSE architecture is essentially "NAT that works." To some, this would be the ultimate insult, but I think it is essentially correct. NAT Boxes provide for isolating a Site from topology changes but severely compromise the end-to-end model. GSE affords very similar operational topological isolation but without violating the end-to-end model, at least not nearly as much. If a Site wishes the additional isolation afforded by NAT Boxes, a firewalls will accomplish that task.

15. General Comments

While some of GSE is a radical departure from IPv6 as we currently know it, in general it relies deeply on all the IPv6 underpinnings which contribute so much to the attractiveness of IPv6: Neighbor Discover, all the dynamic configuration machinery designed to make renumbering palatable even using "provider-based addressing", and the flexibility of the "salami headers" which make tunneling and security attractive. The general forwarding operations based on longest-match-under-prefix-mask and the policy-based routing machinery of BGP5/IDRP are also simply assumed.

16. Closing Comments and Acknowledgments

This document presents a revision of the "8+8" addressing model which has been under construction by the author since before Fall of 1995, at least. Conversations with a great many people have contributed to the design presented in this document. A skeletal version of this proposal first appeared in some email from Dave Clark of MIT who planted the seed and provided the original monicker "8+8". A great

many others have contributed ideas and observations, all of which went into the stew pot for the synthesis contained here.

The original "8+8" draft cited the following individuals for a special thank-you: Vadim Antonov, Ran Atkinson, Scott Bradner, Brian Carpenter, Noel Chiappa, Steve Deering, Sean Doran, Joel Halpern, Christian Huitema, Tony Li, Peter Lothberg, Louis Mamakos, Radia Perlman, Yakov Rekhter, Paul Traina.

This draft has benefited greatly from conversations with Masataka Ohta, who convinced the author of the importance of the IETF-NodeID in addition to the 8-byte IEEE MAC addresses, as well as Brian Carpenter, Scott Brander, Ran Atkinson, all the people who so graciously provided invaluable comments on the original "8+8" draft, and of course Steve Deering, Bob Hinden, and the IPng Working Group.

17. Security Considerations

More than can be imagined.

18. Author's Address

Mike O'Dell
UUNET Technologies, Inc.
3060 Williams Drive
Fairfax, VA 22031
voice: 703-206-5890
fax: 703-206-5471
email: mo@uu.net