Network Working Group
Internet-Draft

Mike O'Dell
UUNET Technologies
1996/10/22 05:58:54GMT
Expire in six months

8+8 - An Alternate Addressing Architecture for IPv6

<draft-odell-8+8-00.txt>

1. Status of this Memo

This document is an Internet-Draft. Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as ``work in progress.''

To learn the current status of any Internet-Draft, please check the 1id-abstracts.txt listing contained in the Internet-Drafts Shadow Directories on ftp.is.co.za (Africa) , nic.nordu.net (Europe), munnari.oz.au (Pacific Rim), ds.internic.net (US East Coast), or ftp.isi.edu (US West Coast).

2. Abstract

This document presents an alternative addressing architecture for IPv6 which controls global routing growth with very aggressive topological aggregation. It also includes support for scalable multihoming as a distinguished service while freeing sites and service resellers from the tyranny of CIDR-based aggregation by providing transparent rehoming of both.

3. Introduction

IP version 6 represents a significant advancement in the technology of the Internet. It provides large addresses, many sorely-needed functional capabilities, and was intended to be a platform for the further evolution of the Global Internet. Unfortunately, when IPv6 was created, Route Scaling, which has become the most significant problem for the continued growth of the Internet, was not widely understood to be the forcing function we now know it to be. Because of that, the current IPv6 addressing proposal fails to provide an operationally-scalable scheme for aggressive topological aggregation and the continued scaling of the routing architecture.

The current IPv6 addressing proposals continue to rely almost entirely upon CIDR-style aggregation for route growth control. Unlike IPv4, in IPv6 this mechanism is coupled with support for easier network renumbering which may make so-called "provider-based addressing" a bit more palatable.

In general, the current IPv6 addressing model is inadequate for several reasons. CIDR-style aggregation breaks down in the face of the accelerating growth of multi-homed sites (leaf sites or regional networks). Renumbering to accomplish simple topological rehoming (e.g., changing ISPs) is a problem whose magnitude will only grow over time. It will always be difficult to explain this to customers, increasingly so with decreasing customer sophistication. 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 continues at current levels, this presents a serious problem because of the scaling behavior of 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 uses and supports the dynamic address assignment machinery in IPv6, but makes the exact role of that machinery a local decision with understandable costs and benefits rather than a mandatory mechanism for simple rehoming situations.

The model also identifies the special work done by the global Internet infrastructure to support multihomed sites, isolating it into a specific mechanism which is then traceable to and incurred by only those sites wishing to use this capability. This then makes it possible for sites to make informed cost-benefit decisions about multihoming.

4. Central Concepts

The addressing model proposed here is called "8+8" to distinguish it from the existing proposals which are called "Flat-16" in this document. The first central concept in 8+8 is simple:

> The 16 byte IPv6 address is split into two 8-byte objects stored in the existing 16-byte container.

The lower 8 bytes (least significant) form the "End System Designator," or ESD. The upper 8 bytes (most significant) are called the "Routing Goop", or RG. The ESD designates a computer system and

the RG encodes information about its attachment to the global Internet topology.

As with other schemes distinguishing location from identity, the 8+8 model requires modifying the upper level protocols to consider only the ESD when performing pseudo-header operations meant to identify the end system as opposed to its location in the topology. A few important examples: the TCP checksum pseudo-header would use only the ESDs instead of the Flat-16 addresses; TCP associations would be identified by ESD/Port instead of Flat-16/Port; IPSEC Authentication and ESP header calculations would only consider the ESD and not the RG of the address. Together these allow session-scale state like TCP connections to survive global topology changes without special considerations in the transport protocol.

Note: this proposal does not effect the IPv6 multicast, loopback, or link-local address formats or usage. It is probably necessary to create a new version of the "IPv6 site-local prefix" which uses an ESD as the lower 8 bytes and would be used for within-site sessions (in the exiting IPv6 sense) and for originating external traffic.

The second central concept is:

> Formalize the distinction between "Public Topology" and "Private Topology".

"Public Topology" is structure which must be understood by a number other organizations, especially and specifically transit networks, for constructing global Internet connectivity. "Private Topology" is structure which is of no particular interest outside the containing organization. In particular, general transit service is provided by networks exposed in the Public Topology; networks composed of only Private Topology cannot provide general transit service to the Global Internet.

In the current IPv4 Internet, the distinction between Public and Private Topology exists as a side-effect but it is not used to any significant advantage beyond that which arises naturally from CIDR-style aggregation. A current example of private topology is the subnet structure used by the topology within a site as applied to the CIDR block for the entire site. No one else outside the site particularly cares about the internal structure of the site so there is no real need to carry any routing information about it other than the CIDR block describing it as a whole.

The 8+8 model elevates this observation to a major architectural component providing an explicit notion of a "Site". A "Site" is the

simplest unit attachment to the Global Internet and is also the unit of Private Topology. Within a Site, the ESD of a system is sufficient for reaching it across the Private Topology as well as globally identifying the system outside the confines of the Site. This site-internal reachability can be accomplished by either flat-routing on the ESD with a site (whether this is called "LAN Switching" or something else is irrelevant), or by using a structured ESD within the site. Both of these solutions are supported by the structure of the ESD and each has identifiable and understandable costs and benefits. These will be discussed at length later.

The "Public Topology" is the transit infrastructure which carries traffic from one Site to another. It is composed of the various carrier, reseller, and regional networks which we know today. The Routing Goop portion of an 8+8 address is a locator which encodes information about the way a Site (containing Private Topology) is connected to the Public Topology of the transit networks. As will be explained later, Routing Goop compactly encodes topology information with very high degrees of aggregation while still affording the opportunity to carry local detail for optimizing regional routes without sacrificing global aggregation. Again, this will be discussed later.

The third central concept is:

> Dynamic insertion of Routing Goop into source addresses by Site Boundary Routers when a packet leaves a Site and enters the Public Topology.

This is one of the most radical parts of this proposal and was not included in earlier versions of this document, but discussions with various people convinced the author that it solves a sufficiently compelling number of problems with one simple mechanism that it was adopted. It too will be discussed later.

5. The Structure of End System Designators - the ESD

End System Designators designate every computer system in the 8+8 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.

An interesting question is whether an ESD identifies a system, possibly as in the XNS architecture, or an interface, as in the existing IPv4 and IPv6 architecture. The answer is that an ESD designates an interface on a computer system and that interface can

be either physical or virtual.

When processing an 8+8 address, a computer system need only examine the ESD portion of the address to determine whether a packet is destined for that system.

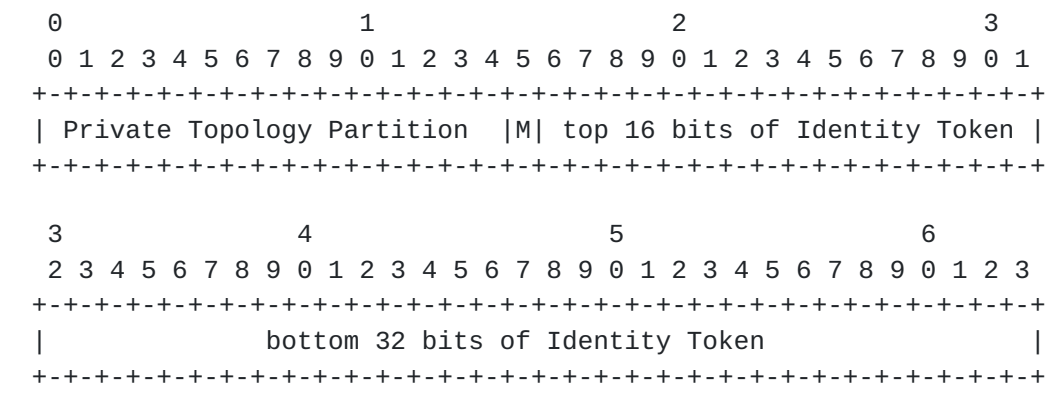
There are circumstances when it is quite useful to have "an address" for a computer system which is independent of any particular physical interface on that system. It has become commonplace in IPv4 practice to use a distinguished virtual interface to provide a system with such an "interface independent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

NOTE: We specifically avoid being pedantic about exactly what constitutes an "interface" and a "computer system" as 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 great cleverness can be contained therein)

The following describes the 8 bytes of the currently-defined ESD structures.



Bits 0-14: 15-bit Private Topology Partition (PTP)
Provides for 32768 distinct partitions in the
Private Topology
Bit 15: Identity Token Mode Indicator
0 => 48-bit Identity Token
1 => Mode in upper bits of Identity Token
Bits 16-63: 48-bit Identity Token

Identity Tokens are formed as follows:

Mode 0 ESDs: (Bit 15: 0)
Identity Token is 48-bits of IEEE MAC Address
Bits 16-63: IEEE 48-bit MAC Address

Mode 1 ESDs: (Bits 15-18: 1001)
Identity Token is 45 bit "IETF NodeID" integer which are
assigned densely starting with 1.
Bits 19-63: IETF NodeID

Mode 2 ESDs: (Bits 15-18: 1010)
Identity Token is 32 bit officially-assigned public IPv4
address (i.e., NOT an RFC-1918 private-use address),
zero padded
Bits 19-31: must be zero
Bits 32-63: valid IPv4 Address

Mode 3 through Mode 7 ESDs (Bits 15-18: 1011 - 1111)
RESERVED

For interfaces with IEEE-assigned 48-bit MAC addresses, a Mode-0 ESD is the most natural ESD for that particular interface. On the other hand, a point-to-point interface with no other naturally-occurring MAC address could be labeled using a Mode-1 ESD. Mode-2 ESDs provide for exploiting an already widely-deployed identifier space for easing the transition to 8+8. Links with MAC addresses larger than 6 bytes

can use Mode-2 ESDs and IPv6 dynamic configuration support with Neighbor Discovery.

The IETF NodeID in the Mode-1 ESD is a 45-bit unsigned integer which starts at one (1) and is incremented, assigning the numbers as densely as possible. There is no particular need to delegate on bit boundaries as powers-of-2 don't matter. The numbers merely must be assigned uniquely to requesters. We leave the actual assignment strategy and any potential delegation to the purview of the IANA.

A few comments on "global uniqueness" are in order because in previous discussions, some people seem to think that unless "uniqueness" can be accomplished with absolute and complete mathematical perfection any scheme using the concept is unworkable. This complete and utter nonsense and is rendered patently false by multiple counter-example:

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.

Yes, it is true that sometimes accidents happen and an IPv4 prefix is misconfigured and it can be troublesome to track down. But the problem is quite manageable. Moreover, even with its extreme rarity, it is much more common than two Ethernet interfaces having the same MAC address. The author believes that the IEEE MAC address assignment machinery coupled with the job the manufacturers do is the closest approximation to "global uniqueness" which any significant human enterprise can achieve, and it is more than adequate to the task at hand. The IETF NodeIDs will be assigned at least as well as IPv4 addresses, and IPv4 seems to work well enough for the Global Internet to function with incredibly few problems arising from this particular source.

6. The Structure of a Site

The 8+8 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 ESD structure provides for a rich set of organizational alternatives with different cost-benefit tradeoffs.

ESDs are globally unique but can also carry internal structure. The global uniqueness is provided by the Identity Token while the internal structure is carried in the Private Topology Partition. The ESD structure provides for 32768 distinct Private Topology Partitions (PTPs) within a Site. This is the equivalent of EVERY Site having been assigned a CIDR block of 128 Class-B addresses subnetted down to a Class-C. The difference is that in an ESD, the subnet population is limited strictly by the link-level (LAN) technology and not by the 253 host limit of the Class-C subnet. This allows an extremely rich topology to be contained within a Site without it exporting complexity into the global routing structure which must then be concealed by tricks like CIDR aggregation.

Of course, 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 retain considerable independence in topological structure and attachment to the Global Internet, they must be aware of changes between the constituent Sites and that rehoming 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 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 Mode-0 ESDs with all the systems connected in a single Private Topology Partition (i.e., all the ESDs carry the same PTP 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, but the complexity is absorbed into the LAN infrastructure and it appears to be only one partition from the 8+8 Private Topology view. This structure has one very significant advantage: rehoming a system within this structure will not change the ESD and TCP sessions (for example) will survive arbitrary changes in the private topology. This works, of course, because the single PTP 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 PTPs with routing carrying 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 PTP. 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. 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, though, is that even with multiple PTPs routed within a Site, a "Private Topology Partition" need not correspond to a "physical" LAN cable. The PTP 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 Mode-2 ESDs based on existing IPv4 address assignments. In this case, all the IPv4 Identity Tokens could be placed in a single PTP and then routed internally on the IPv4 address in the lowest 4 bytes of the Identity Token. This has the advantage of significant familiarity, but also can induce externally-visible changes if ESDs must be reassigned because of private topology requirements. Again, it must be emphasized that the IPv4 addresses used in a Mode-2 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, a Mode-1 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 a Mode-0 ESD. And in all of the cases, Mode-1 ESDs could be used universally, although it is more appropriate to use Mode-0 whenever possible; no sense wasting Identity Tokens when it isn't necessary.

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 (PTPs). This will have the effect of killing long-running off-Site connections unless provisions are made to allow the systems 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.

But most importantly, external rehoming of a Site to the global infrastructure can be made completely transparent in almost every case.

7. The Structure of Routing Goop

Routing Goop, or "RG" is the upper 8 bytes of an 8+8 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, rehoming to a new attachment to the Public Topology will change ONLY the RG in full 8+8 addresses for computer systems at that Site. One example of such a rehoming 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 can be viewed as a rooted path-expression of flat-routed regions which are tangent. Each element in the path-expression contains 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 8+8 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.

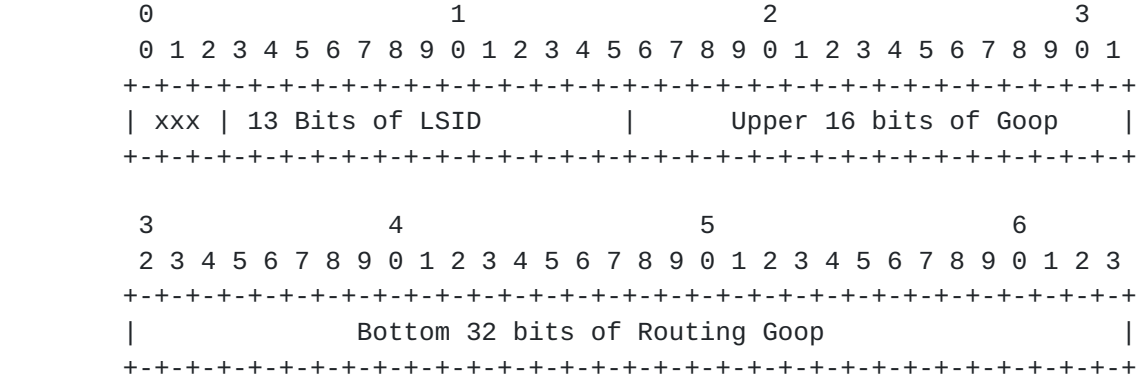
Within Large Structures, the (sub-)partition function is a trade-off between the flat-routing complexity within a region and minimizing total depth of the substructure. This is driven by the internal topology of a Large Structure and the choices in different Large Structures will not necessarily be the 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 8+8 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:



NOTE: The Routing Goop structure above assumes that the 8+8 proposal is designated by a 3-bit type of IPv6 address. If an 8+8 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 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.

We note that while the LSID is intuitively 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 will remain distinct.

8. 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 8+8 model.

While the 8+8 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.

9. 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.

9.1 RG for Source Addresses

The 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 Site Boundary Router inserts the correct RG in the source address based on the path the destination should use to return a packet to the sender. Except in very unusual circumstances this will be the RG which corresponds to the attachment path of the Site Boundary Router to the Global Internet.

If the Site is Mulithomed 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 Multihomed via more than one Site Boundary Router, which router the packet leaves by is purely local policy and which RG gets applied is likewise local policy.

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

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

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

(3) It means that more complex Multihoming 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 Mobile IP is dramatically simplified over the current model, but we postpone that discussion to another day.

(5) 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 without it, thereby enforcing tracability. 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.

This simple mechanism solves a number of problems and actually simplifies the operation and deployment of this architecture so is well worth the implications it has for Site Border Routers.

The Site Border Router gets 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.

9.2 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 8+8 design proposes synthesizing the 16 bytes of information in a query response from two different sources: an "AA" record and an "RG" record. The "AA" record carries the 8-byte ESD for the DNS name in question and the "RG" record carries 8 bytes of the appropriate Routing Goop.

One interesting question is how the AA 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 "AA" record which would allow a nameserver to concatenate an arbitrary RG prefix to the ESD 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 rehoming 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

multihomed cases.

The result of this structure is that individual zone entries for individual nodes (AA records) do NOT change when a Site rehomes. The only thing which changes (logically) is the RG information which is composed with the nodes AA record to produce a full 16-byte response. This means the general Dynamic DNS machinery is NOT required to support Site rehoming.

It also gives rise to significant potential for "smart nameservers" which examine the source address of a query to provided a more topologically appropriate translation for a given DNS query. This isn't perfect, but it is much more detail than current nameservers have available without processing a full BGP routing table to ascertain IPv4 prefix/AS correspondence.

10. Rehoming 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 8+8 architecture is to eliminate, or at least mitigate, the impact of this.

It is important to reiterate the notion that the Routing Goop of an 8+8 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 Rehoming and Multihoming 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 exit 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 exit 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 depolyment. If at some later point it becomes clear how to accomplish this safely, then it could be added downstream. But the complexity, security risks, and the mangnitude of the added value do not make it worthwhile at present, although the author would love to be convinced otherwise.

Alternately, the Site could request a "Rehoming Courtesy" from their old ISP which would effectively make it a multihomed Site for some period of time. After multihoming 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 Rehoming Courtesy.

Note that at no time did the rehoming 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. Multihoming a Site

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

In the 8+8 model, multihoming 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 - multihoming 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 "Rehoming Courtesy" in the previous section is simply doing the router-pair coordination with the new ISP for some period of time.

12. Rehoming a Reseller

Rehoming a Reseller is a slightly more general case of rehoming 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 rehoming and should be completely transparent except for long-lived sessions which do not terminate by the end of the grace period.

13. Multihoming a Reseller

There are two parts to multihoming a Reseller - one part similar to the Multihomed 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 multihomed tunneling services from his two access point routers to provide alternate path service just like a multihomed 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 multihomed 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

Discussions of this proposal raised the question of what it means for Network Address Translation (NAT) boxes. On the one hand, the 8+8 model allows a NAT box to modify the Routing Goop during forwarding without impeding end-to-end TCP checksums which only rely upon the ESDs. On the other hand, it isn't very clear what purpose of a NAT box would have given the 8+8 model.

Typically a NAT box is cited as a way to have private topology within a site (note lower case) which is then attached to the Public Topology via the NAT box without revealing anything about that private topology. The basic structure of the 8+8 model accomplishes exactly this goal - providing genuine Private Topology within local purview while providing independence of attachment point to the Public Topology. The broad conclusion is that pure NAT boxes don't have much of a future given the 8+8 model. More general application gateways performing firewall functions or "intranet bridges" providing crypto-tunnels between the protected interior of two Sites, however, are altogether another matter.

15. General Comments

While some of 8+8 is something of 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. All of these will need a tweak or two based on this proposal and it is beyond this author to do all the analysis required to identify every such tweak needed, so it will be up to the community to analyze this proposal

and if embraced, look at all the related machinery which is touched in some subtle manner.

This document has presented both an outline and the deep ideas behind an 8+8 proposal, and the author believes it has addressed the "hard problems" to the point it can convince the reader of the viability, and indeed the merits of this approach. The routing scaling problems going forward require the kind of flexibility afforded by this approach. Once the 8+8 partitioning of the address is accomplished, we are freed to tinker with the routing and forwarding machinery in ways which cannot be achieved nearly as readily as with a monolithic 16-byte address.

16. Closing Comments

This document presents a 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 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. While it is impossible to mention all of them, a few deserve special mention as having provided comments on drafts or otherwise have significantly influence the thinking contained herein: 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. And a special thanks to all those folks in the IPng working groups who contributed to the foundation which is IPv6.

17. Security Considerations

Almost certainly lots of them.

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