Internet Research Task Force (IRTF)
Request for Comments: 6306
Category: Experimental
ISSN: 2070-1721

P. Frejborg

July 2011

Hierarchical IPv4 Framework

Abstract

This document describes a framework for how the current IPv4 address space can be divided into two new address categories: a core address space (Area Locators, ALOCs) that is globally unique, and an edge address space (Endpoint Locators, ELOCs) that is regionally unique. In the future, the ELOC space will only be significant in a private network or in a service provider domain. Therefore, a 32x32 bit addressing scheme and a hierarchical routing architecture are achieved. The hierarchical IPv4 framework is backwards compatible with the current IPv4 Internet.

This document also discusses a method for decoupling the location and identifier functions -- future applications can make use of the separation. The framework requires extensions to the existing Domain Name System (DNS), the existing IPv4 stack of the endpoints, middleboxes, and routers in the Internet. The framework can be implemented incrementally for endpoints, DNS, middleboxes, and routers.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for examination, experimental implementation, and evaluation.

This document defines an Experimental Protocol for the Internet community. This document is a product of the Internet Research Task Force (IRTF). The IRTF publishes the results of Internet-related research and development activities. These results might not be suitable for deployment. This RFC represents the individual opinion(s) of one or more members of the Routing Research Group of the Internet Research Task Force (IRTF). Documents approved for publication by the IRSG are not a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc6306.

Table of Contents

1. Introduction

A Locator/Identifier Separation Protocol [LISP] presentation from a breakout session at an expo held in January, 2008, triggered a research study; findings from the study are described in this document. Further studies revealed that the routing community at IETF is concerned about the scalability of the routing and addressing system of the future Internet. The Internet Architecture Board (IAB) held a Routing and Addressing workshop on October 18-19, 2006, in Amsterdam. The outcome from the workshop is documented in [RFC4984]. Also, the IRTF had established a Routing Research Group [RRG] in 2007 and created some design guidelines; see [RFC6227].

The author of this document found the LISP approach very interesting because the IP address space is proposed to be separated into two groups: Routing Locators (RLOCs), which are present in the global routing table of the Internet called the Default-Free Zone (DFZ), and Endpoint Identifiers (EIDs), which are only present in edge networks attached to the Internet.

The proposed LISP architecture reduces the routing information in the DFZ, but it also introduces a new mapping system that would require a caching solution at the border routers installed between the edge networks and DFZ. EID prefixes are not needed in the DFZ since a tunneling (overlay) scheme is applied between the border routers. To the author, this seems to be a complex architecture that could be improved by applying lessons learned from similar past architectures -- in the ’90s, overlay architectures were common, deployed on top of Frame Relay and ATM technologies. Cache-based routing architectures have also been tried, for example, Ipsilon’s IP Switching. These architectures have largely been replaced by MPLS [RFC3031] for several reasons -- one being that overlay and caching solutions have historically suffered from scalability issues. Technology has certainly evolved since the ’90s. The scalability issues of overlay and caching solutions may prove to be less relevant for modern hardware and new methods; see [Revisiting_Route_Caching]

Nevertheless, the author has some doubt whether overlay and caching will scale well, based upon lessons learned from past overlay and caching architectures. The hierarchical IPv4 framework proposal arose from the question of whether the edge and core IP addressing groupings from LISP could be used without creating an overlay solution by borrowing ideas from MPLS to develop a peer-to-peer architecture. That is, instead of tunneling, why not swap IP addresses (hereafter called locators) on a node in the DFZ? By introducing a shim header to the IPv4 header and Realm Border Router (RBR) functionality on the network, the edge locators are no longer needed in the routing table of DFZ.

Two architectural options existed regarding how to assemble the packet so that RBR functionality can be applied in the DFZ: the packet was assembled by either an ingress network node (similar to LISP or MPLS) or at the endpoint itself. The major drawback in assembling the packet with a shim header at the endpoint is that the endpoint's stack must be upgraded; however, a significant advantage is that the Path MTU Discovery issue, as discussed in, e.g., LISP, would not exist. In addition, the caching scalability issue is mitigated to the greatest extent possible by pushing caching to the endpoint.

This approach also opened up the possibility of extending the current IP address scheme with a new dimension. In an MPLS network, overlapping IP addresses are allowed since the forwarding plane is leveraging label information from the MPLS shim header. By applying RBR functionality, extending the current IPv4 header with a shim header and assembling the new header at endpoints, an IP network can also carry packets with overlapping edge locators, although the core locators must still be globally unique. The location of an endpoint is also no longer described by a single address space; it is described by a combination of an edge locator and a core locator, or a set of core locators.

Later on, it was determined that the current 32-bit address scheme can be extended to 64 bits -- 32 bits reserved for globally unique core locators and 32 bits reserved for locally unique edge locators.

The new 64-bit addressing scheme is backwards compatible with the currently deployed Internet addressing scheme.

By making the architectural decisions described above, the foundation for the hierarchical IPv4 framework was laid out.

Note that the hierarchical IPv4 framework is abbreviated as hIPv4, which is close to the abbreviation of Host Identity Protocol (HIP) [RFC4423]. Thus, the reader needs to pay attention to the use of the two abbreviations -- hIPv4 and HIP, which represent two different architectures.

Use of the hIPv4 abbreviation has caused much confusion, but it was chosen for two reasons:

- Hierarchical - to emphasize that a hierarchical addressing scheme is developed. A formalized hierarchy is achieved in the routing architecture. Some literature describes today's Internet as already using hierarchical addressing. The author believes that this claim is not valid -- today's Internet uses one flat address space.

It is true that we have hierarchical routing in place. A routing architecture can consist of at least three types of areas: stub area, backbone area, and autonomous system (AS). The current flat address space is summarized or aggregated at border routers between the areas to suppress the size of a routing table. In order to carry out summaries or aggregates of prefixes, the address space must be continuous over the areas.

Thus, the author concludes that the current method is best described as an aggregating addressing scheme since there are address block dependencies between the areas. Dividing addresses into edge and core locator spaces (a formalized hierarchy) opens up a new dimension -- the edge locator space can still be deployed as an aggregating address scheme on the three types of areas mentioned earlier. In hIPv4, the core locators are combined with edge locators, independent from each other -- the two locator space allocation policies are separated and no dependencies exist between the two addressing schemes in the long-term architecture.

A new hierarchical addressing scheme is achieved: a two-level addressing scheme describing how the endpoint is attached to the local network and also how the endpoint is attached to the Internet. This change in the addressing scheme will enable a fourth level, called the Area Locator (ALOC) realm, at the routing architecture.

- IPv4 - to emphasize that the framework is still based upon the IPv4 addressing scheme, and is only an evolution from the currently deployed addressing scheme of the Internet.

While performing this research study, the author reviewed a previous hierarchical addressing and routing architecture that had been proposed in the past, the Extended Internet Protocol (EIP) [RFC1385]. Should the hIPv4 framework ever be developed from a research study to a standard RFC, it is recommended that the hierarchical IPv4 framework name be replaced with Extended Internet Protocol, EIP, since both architectures share similarities, e.g., backwards compatibility with existing deployed architecture, hierarchical addressing, etc., and the hIPv4 abbreviation can be mixed up with HIP.

This document is an individual contribution to the IRTF Routing Research Group (RRG); discussions between those on the mailing list of the group have influenced the framework. The views in this document are considered controversial by the IRTF Routing Research Group (RRG), but the group reached a consensus that the document should still be published. Since consensus was not achieved at RGG regarding which proposal should be preferred -- as stated in

[RFC6115]: "The group explored a number of proposed solutions but did not reach consensus on a single best approach" -- thus, all proposals produced within RRG can be considered controversial.

2. Requirements Notation

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 [RFC2119].

3. Definitions of Terms

This document makes use of the following terms:

Regional Internet Registry (RIR):

> This is an organization overseeing the allocation and registration of Internet number resources within a particular region of the world. Resources include IP addresses (both IPv4 and IPv6) and autonomous system numbers.

Locator:

> A name for a point of attachment within the topology at a given layer. Objects that change their point(s) of attachment will need to change their associated locator(s).

Global Locator Block (GLB):

> An IPv4 address block that is globally unique.

Area Locator (ALOC):

> An IPv4 address (/32) assigned to locate an ALOC realm in the Internet. The ALOC is assigned by an RIR to a service provider. The ALOC is globally unique because it is allocated from the GLB.

Endpoint Locator (ELOC):

> An IPv4 address assigned to locate an endpoint in a local network. The ELOC block is assigned by an RIR to a service provider or to an enterprise. In the intermediate routing architecture, the ELOC block is only unique in a geographical region. The final policy of uniqueness shall be defined by the RIRs. In the long-term routing architecture, the ELOC block is no longer assigned by an RIR; it is only unique in the local ALOC realm.

ALOC realm:

An area in the Internet with at least one attached Realm Border Router (RBR). Also, an ALOC must be assigned to the ALOC realm. The Routing Information Base (RIB) of an ALOC realm holds both local ELOC prefixes and global ALOC prefixes. An ALOC realm exchanges only ALOC prefixes with other ALOC realms.

Realm Border Router (RBR):

A router or node that is able to identify and process the hIPv4 header. In the intermediate routing architecture, the RBR shall be able to produce a service, that is, to swap the prefixes in the IP header and locator header, and then forward the packet according to the value in the destination address field of the IP header. In the long-term routing architecture, the RBR is not required to produce the swap service. Instead, the RBR can make use of the Forwarding Indicator field in the locator header. Once the FI-bits are processed, the RBR forwards the packet according to the value in the destination address of the IP header or according to the value in the ELOC field of the locator header. The RBR must have the ALOC assigned as its locator.

Locator Header:

A 4-byte or 12-byte field, inserted between the IP header and transport protocol header. If an identifier/locator split scheme is used, the size of the locator header is further expanded.

Identifier:

The name of an object at a given layer. Identifiers have no topological sensitivity and do not have to change, even if the object changes its point(s) of attachment within the network topology.

Identifier/locator split scheme:

Separate identifiers used by applications from locators that are used for routing. The exchange of identifiers can occur discreetly between endpoints that have established a session, or the identifier/locator split can be mapped at a public database.

Session:

An interactive information exchange between endpoints that is established at a certain time and torn down at a later time.

Provider Independent Address Space (PI addresses/prefixes):

An IPv4 address block that is assigned by a Regional Internet Registry directly to a user organization.

Provider Aggregatable Address Space (PA addresses/prefixes):

An IPv4 address block assigned by a Regional Internet Registry to an Internet Service Provider that can be aggregated into a single route advertisement.

Site mobility:

A site wishing to change its attachment point to the Internet without changing its IP address block.

Endpoint mobility:

An endpoint moves relatively rapidly between different networks, changing its IP layer network attachment point.

Subflow:

A flow of packets operating over an individual path, the flow forming part of a larger transport protocol connection.

4. Hierarchical Addressing

The current IP addressing (IPv4) and the future addressing (IPv6) schemes of the Internet are unidimensional by their nature. This limitation -- the unidimensional addressing scheme -- has created some roadblocks, for example, breaking end-to-end connectivity due to NAT, limited deployment of Stream Control Transmission Protocol (SCTP) [RFC4960], etc., for further growth of the Internet.

If we compare the Internet’s current addressing schemes to other global addressing or location schemes, we notice that the other schemes use several levels in their structures. For example, the postal system uses street address, city, and country to locate a destination. To locate a geographical site, we use longitude and latitude in the cartography system. The other global network, the Public Switched Telephone Network (PSTN), has been built upon a three-level numbering scheme that has enabled a hierarchical

signaling architecture. By expanding the current IPv4 addressing scheme from a single level to a two-level addressing structure, most of the issues discussed in [RFC4984] can be solved. Also, a hierarchical addressing scheme would better describe the Internet we have in place today.

Looking back, it seems that the architecture of the Internet changed quite radically from the intended architecture with the introduction of [RFC1918], which divides the hosts into three categories and the address space into two categories: globally unique and private address spaces. This idea allowed for further growth of the Internet and extended the life of the IPv4 address space, and it ended up becoming much more successful than expected. RFC 1918 didn't solve the multi-homing requirements for endpoints providing services for Internet users, that is, multi-homed sites with globally unique IP addresses at endpoints to be accessed from the Internet.

Multi-homing has imposed some challenges for the routing architecture that [RRG] is addressing in [RFC6115]. Almost all proposals in the report suggest a core and edge locator separation or elimination to create a scalable routing architecture. The core locator space can be viewed to be similar to the globally unique address space, and the edge locator space similar to the private address space in RFC 1918.

RFC 1918 has already demonstrated that Internet scales better with the help of categorized address spaces, that is, globally unique and private address spaces. The RRG proposals suggest that the Internet will be able to scale even further by introducing core and edge locators. Why not then change the addressing scheme (both IPv4 and IPv6 addressing schemes, though this document is only focusing on IPv4) to better reflect the current and forthcoming Internet routing architecture? If we continue to use a flat addressing scheme, and combine it with core (global) and edge (private) locator (address) categories, the routing architecture will have to support additional mechanisms, such as NAT, tunneling, or locator rewriting with the help of an identifier to overcome the mismatch. The result will be that information is lost or hidden for the endpoints. With a two-level addressing scheme, these additional mechanisms can be removed and core/edge locators can be used to create new routing and forwarding directives.

A convenient way to understand the two-level addressing scheme of the hIPv4 framework is to compare it to the PSTN numbering scheme (E.164), which uses country codes, national destination codes, and subscriber numbers. The Area Locator (ALOC) prefix in the hIPv4 addressing scheme can be considered similar to the country code in PSTN; i.e., the ALOC prefix locates an area in the Internet called an ALOC realm. The Endpoint Locator (ELOC) prefixes in hIPv4 can be

compared to the subscriber numbers in PSTN -- the ELOC is regionally unique (in the future, locally unique) at the attached ALOC realm. The ELOC can also be attached simultaneously to several ALOC realms.

By inserting the ALOC and ELOC elements as a shim header (similar to the MPLS and [RBridge] architectures) between the IPv4 header and the transport protocol header, a hIPv4 header is created. From the network point of view, the hIPv4 header "looks and feels like" an IPv4 header, thus fulfilling some of the goals as outlined in EIP and in the early definition of [Nimrod]. The outcome is that the current forwarding plane does not need to be upgraded, though some minor changes are needed in the control plane (e.g., ICMP extensions).

5. Intermediate Routing Architecture

The intermediate routing architecture is backwards compatible with the currently deployed Internet; that is, the forwarding plane remains intact except that the control plane needs to be upgraded to support ICMP extensions. The endpoint’s stack needs to be upgraded, and middleboxes need to be upgraded or replaced. In order to speed up the transition phase, middleboxes might be installed in front of endpoints so that their stack upgrade can be postponed; for further details, see Appendix D.

5.1. Overview

As mentioned in previous sections, the role of an Area Locator (ALOC) prefix is similar to a country code in PSTN; the ALOC prefix provides a location functionality of an area within an autonomous system (AS), or an area spanning over a group of ASes, in the Internet. An area can have several ALOC prefixes assigned, e.g., for traffic engineering purposes such as load balancing among several ingress/egress points at the area. The ALOC prefix is used for routing and forwarding purposes on the Internet, and so the ALOC prefix must be globally unique and is allocated from an IPv4 address block. This globally unique IPv4 address block is called the Global Locator Block (GLB).

When an area within an AS (or a group of ASes) is assigned an ALOC prefix, the area has the potential to become an ALOC realm. In order to establish an ALOC realm, more elements, more than just the ALOC prefix, are needed. One or multiple Realm Border Routers (RBRs) must be attached to the ALOC realm. An RBR element is a node capable of swapping the prefixes of the IP header and the new shim header, called the locator header. The swap service is described in detail in Section 5.2, step 3.

Today's routers do not support this RBR functionality. Therefore, the new functionality will most likely be developed on an external device attached to a router belonging to the ALOC realm. The external RBR might be a server with two interfaces attached to a router, the first interface configured with the prefix of the ALOC and the second with any IPv4 prefix. The RBRs do not make use of dynamic routing protocols, so neither a Forwarding Information Base (FIB) nor a cache is needed -- the RBR performs a service, swapping headers.

The swap service is applied on a per-packet basis, and the information needed to carry out the swap is included in the locator header of the hIPv4 packet. Thus, a standalone device with sufficient computing and I/O resources to handle the incoming traffic can take the role as an RBR. Later on, the RBR functionality might be integrated into the forwarding plane of a router. It is expected that one RBR will not be able to handle all the incoming traffic designated for an ALOC realm and that having a single RBR would also create a potential single point of failure in the network. Therefore, several RBRs might be installed in the ALOC realm and the RBRs shall use the ALOC prefix as their locator, and the routers announce the ALOC prefix as an anycast locator within the local ALOC realm. The ALOC prefix is advertised throughout the DFZ by BGP mechanisms. The placement of the RBRs in the network will influence the ingress traffic to the ALOC realm.

Since the forwarding paradigm of multicast packets is quite different from forwarding unicast packets, the multicast functionality will have an impact on the RBR. Because the multicast RBR (mRBR) functionality is not available on today's routers, an external device is needed -- later on the functionality might be integrated into the routers. The mRBR shall take the role of an anycast Rendezvous Point with the Multicast Source Discovery Protocol (MSDP) [RFC3618] and Protocol Independent Multicast (PIM) [RFC4601] capabilities, but to swap headers neither a FIB nor a cache is required. As with the RBR, the multicast hIPv4 packets are carrying all needed information in their headers in order to apply the swap service; for details, see Section 10.5.

The ALOC realm is not yet fully constructed. We can now locate the ALOC realm on the Internet, but to locate the endpoints attached to the ALOC realm, a new element is needed: the Endpoint Locator (ELOC). As mentioned in the previous section, the ELOC prefixes can be considered similar to the subscriber numbers in PSTN. The ELOC is not a new element but a redefinition of the current IPv4 address configured at an endpoint. The term redefinition is applied because when the hIPv4 framework is fully implemented, the global uniqueness of the IPv4 addresses is no longer valid. A more regional address

allocation policy of IPv4 addresses can be deployed, as discussed in Appendix A. The ELOC prefix will only be used for routing and forwarding purposes inside the local and remote ALOC realms, and it is not used in the intermediate ALOC realms.

When an initiator is establishing a session to a responder residing outside the local ALOC realm, the value in the destination address field of the IP header of an outgoing packet is no longer the remote destination address (ELOC prefix); instead, the remote ALOC prefix is installed in the destination address field of the IP header. Because the value in the destination address field of the IP header is carrying an ALOC prefix, the intermediate ALOC realms do not need to install the ELOC prefixes of other ALOC realms in their routing tables. It is sufficient for the intermediate ALOC realms to carry only the ALOC prefixes.

The outcome is that the routing tables at each ALOC realm will be reduced when the hIPV4 framework is fully implemented. The ALOC prefixes are still globally unique and must be installed in the DFZ. Thus, the service provider cannot control the growth of the ALOC prefixes, but she/he can control the amount of local ELOC prefixes in her/his local ALOC realm.

When the hIPv4 packet arrives at the remote ALOC realm, it is forwarded to the nearest RBR, since the value in the destination address field of the IP header is the remote ALOC prefix. When the RBR has swapped the hIPv4 header, the value in the destination address field of the IP header is the remote ELOC; thus, the hIPv4 packet will be forwarded to the final destination at the remote ALOC realm. An endpoint using an ELOC prefix can be attached simultaneously to two different ALOC realms without the requirement to deploy a classical multi-homing solution; for details, see Section 12 and Appendix B.

Understanding that the addressing structure is no longer unidimensional and that a second level of hierarchy has been added, it is important to solve the problems of locating the remote ELOC (endpoint) and remote ALOC realm on the Internet, as well as determining where to assemble the header of the hIPv4 packet. The hierarchical IPv4 framework relies upon the Domain Name System needs to support a new record type so that the ALOC information can be distributed to the endpoints. To construct the header of the hIPv4 packet, either the endpoint or an intermediate node (e.g., a proxy) should be used. A proxy solution is likely to prove suboptimal due to a complication induced by the proxy’s need to listen to DNS messages, and a cache solution has scalability issues.

A better solution is to extend the current IPv4 stack at the endpoints so that the ALOC and ELOC elements are incorporated at the endpoint's stack; however, backwards compatibility must be preserved. Most applications will not be aware of the extensions while other IP-aware applications, such as Mobile IP, SIP, IPsec AH and so on (see Section 10.3) will suffer and cannot be used outside their ALOC realm when the hIPv4 framework is fully implemented, unless they are upgraded. The reason is that the IP-aware applications depend upon the underlying network addressing structure, e.g., to identify an endpoint.

Note that the applications used inside the local ALOC realm (e.g., enterprise's private network) do not need to be upgraded -- neither in the intermediate nor in the long-term routing architecture. The classical IPv4 framework is preserved in that only IP-aware applications used between ALOC realms need to be upgraded to support the hIPv4 header.

Figure 1 shows a conceptual overview of the intermediate routing architecture. When this architecture is in place, the ELOC space is no longer globally unique. Instead, a regional allocation policy can be implemented. For further details, see Appendix A. The transition from the current routing architecture to the intermediate routing architecture is discussed in Appendix D.

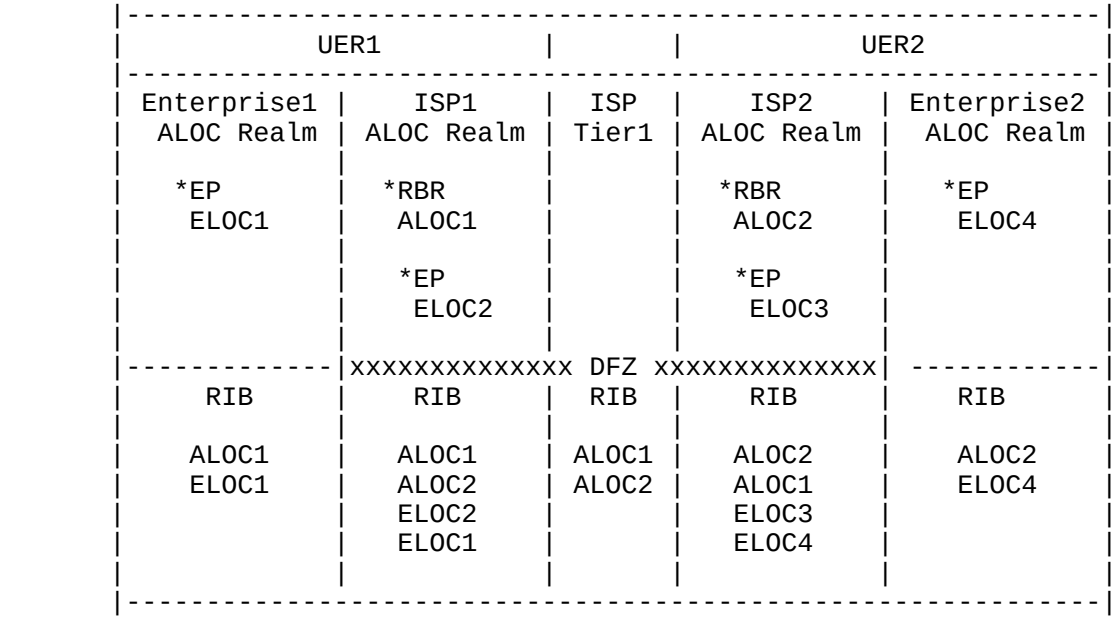

Figure 1: Intermediate routing architecture of hIPv4

5.2. Life of a hIPv4 Session

This section provides an example of a hIPv4 session between two hIPv4 endpoints: an initiator and a responder residing in different ALOC realms.

When the hIPv4 stack is assembling the packet for transport, the hIPv4 stack shall decide if a classical IPv4 or a hIPv4 header is used based on the ALOC information received by a DNS reply. If the initiator’s local ALOC prefix equals the responder’s ALOC prefix, there is no need to use the hIPv4 header for routing purposes, because both the initiator and responder reside in the local ALOC realm. The packet is routed according to the prefixes in the IP header since the packet will not exit the local ALOC realm. When the local ALOC prefix does not match the remote ALOC prefix, a hIPv4 header must be assembled because the packet needs to be routed to a remote ALOC realm.

A session between two endpoints inside an ALOC realm might use the locator header -- not for routing purposes, but to make use of Valiant Load-Balancing [VLB] for multipath-enabled transport protocols (see Section 11.1) or to make use of an identifier/locator split scheme (see Section 7). When making use of VLB, the initiator adds the locator header to the packet and by setting the VLB-bits to 01 or 11, indicating to the responder and intermediate routers that VLB is requested for the subflow. Because this is an intra-ALOC realm session, there is no need to add ALOC and ELOC fields to the locator header, and thus the size of the locator header will be 4 bytes.

If an identifier/locator split scheme is applied for the session (intra-ALOC or inter-ALOC), the initiator must set the I-bit to 1 and make use of the Locator Header Length field. Identifier/locator split scheme information is inserted into the locator header after the Locator Header Length field.

How a hIPv4 session is established follows:

1. The initiator queries the DNS server. The hIPv4 stack notices that the local and remote ALOCs do not match and therefore must use the hIPv4 header for the session. The hIPv4 stack of the initiator must assemble the packet by the following method:

   a. Set the local IP address from the API in the source address field of the IP header.

   b. Set the remote IP address from the API in the ELOC field of the locator header.

   c. Set the local ALOC prefix in the ALOC field of the locator header.

   d. Set the remote ALOC prefix in the destination address field of the IP header.

   e. Set the transport protocol value in the protocol field of the locator header and set the hIPv4 protocol value in the protocol field of the IP header.

   f. Set the desired parameters in the A-, I-, S-, VLB-, and L-fields of the locator header.

   g. Set the FI-bits of the locator header to 00.

h. Calculate IP, locator, and transport protocol header checksums. The transport protocol header calculation does not include the locator header fields. When completed, the packet is transmitted.

2. The hIPv4 packet is routed throughout the Internet based on the value in the destination address field of the IP header.

3. The hIPv4 packet will reach the closest RBR of the remote ALOC realm. When the RBR notices that the value in the destination address of the IP header matches the local ALOC prefix, the RBR must:

   a. Verify that the received packet uses the hIPv4 protocol value in the protocol field of the IP header.

   b. Verify IP, locator, and transport protocol header checksums. The transport protocol header verification does not include the locator header fields.

   c. Replace the source address in the IP header with the ALOC prefix of the locator header.

   d. Replace the destination address in the IP header with the ELOC prefix of the locator header.

   e. Replace the ALOC prefix in the locator header with the destination address of the IP header.

   f. Replace the ELOC prefix in the locator header with the source address of the IP header.

   g. Set the S-field to 1.

   h. Decrease the Time to Live (TTL) value by one.

   i. Calculate IP, locator, and transport protocol header checksums. The transport header calculation does not include the locator header fields.

   j. Forward the packet according to the value in the destination address field of the IP header.

4. The swapped hIPv4 packet is now routed inside the remote ALOC realm based on the new value in the destination address field of the IP header to the final destination.

5. The responder receives the hIPv4 packet.

   a. The hIPv4 stack must verify that the received packet uses the hIPv4 protocol value in the protocol field of the IP header.

   b. Verify IP, locator, and transport protocol header checksums. The transport protocol header verification does not include the locator header fields.

6. The hIPv4 stack of the responder must present the following to the extended IPv4 socket API:

   a. The source address of the IP header as the remote ALOC prefix.

   b. The destination address of the IP header as the local IP address.

   c. Verify that the received ALOC prefix of the locator header equals the local ALOC prefix.

   d. The ELOC prefix of the locator header as the remote IP address.

The responder's application will respond to the initiator and the returning packet will take almost the same steps, which are steps 1 to 6, as when the initiator started the session. In step 1, the responder does not need to do a DNS lookup since all information is provided by the packet.

6. Long-Term Routing Architecture

The long-term routing architecture is established once the forwarding planes of private ALOC realms or service providers ALOC realms containing subscribers are upgraded. The forwarding planes of transit DFZ routers do not need to be upgraded. Why then would private network or service provider administrators upgrade their infrastructure? There are two incentives:

- The overlay local ALOC exit routing topology (as discussed in Section 11) can be replaced by a peer-to-peer local ALOC exit routing topology, which is simpler to operate, thus decreasing operational expenditures.

- Locator freedom: Once the local ALOC realm is upgraded, the enterprise or service provider can use the full 32-bit ELOC address space to remove address space constraints and to design a well-aggregated routing topology with an overdimensioned ELOC allocation policy.

When an enterprise or service provider upgrades the forwarding plane in their ALOC realm, the previous PI or PA address space allocation is released back to the RIR to be used for ALOC allocations in the GLB.

6.1. Overview

The swap service at the RBR was added to the framework in order to provide a smooth transition from the current IPv4 framework to the hIPv4 framework; a major upgrade of the current forwarding plane is avoided by the introduction of the swap service. In the future, the swap service can be left "as is" in the ALOC realm, if preferred, or the swap service can be pushed towards the edge of the ALOC realm when routers are upgraded in their natural lifecycle process.

Once an upgrade of a router is required because of, for example, increased demand for bandwidth, the modified forwarding plane might concurrently support IPv4 and hIPv4 forwarding -- and the swap service can be pushed towards the edge and in the future removed at the ALOC realm. This is accomplished by adding an extension to the current routing protocols, both IGP and BGP. When an RBR receives a hIPv4 packet where the value of the destination address field in the IP header matches the local ALOC prefix, the RBR will -- contrary to the tasks defined in Section 5.2, step 3 -- look up the ELOC field in the locator header and compare this prefix against the FIB. If the next-hop entry is RBR-capable, the packet will be forwarded according to the ELOC prefix. If the next-hop is a classical IPv4 router, the RBR must apply the tasks defined in Section 5.2, step 3 and, once completed, forward the packet according to the new value in the destination address field of the IP header.

When all endpoints (that need to establish sessions outside the local ALOC realm) and infrastructure nodes in an ALOC realm are hIPv4-capable, there is no need to apply swap service for unicast sessions. Forwarding decisions can be based on information in the IP and locator headers. In the local ALOC realm, packets are routed to their upstream anycast or unicast ALOC RBR according to the ALOC prefix in the locator header; local ALOC exit routing is applied against the local ALOC FIB. Remote ELOC approach routing is applied against the ELOC FIB in the remote ALOC realm.

Note that IP and transport protocol headers will remain intact (except for TTL values, since the RBR is a router); only FI and LH checksum values in the locator header will alternate in local ALOC exit routing mode and remote ELOC approach routing mode.

Figure 2 shows a conceptual overview of the long-term hIPv4 routing architecture.

Legend: *attachment point in the ALOC realm
UER=Unique ELOC region
EP=Endpoint
aRBR=anycast RBR
uRBR=unicast RBR

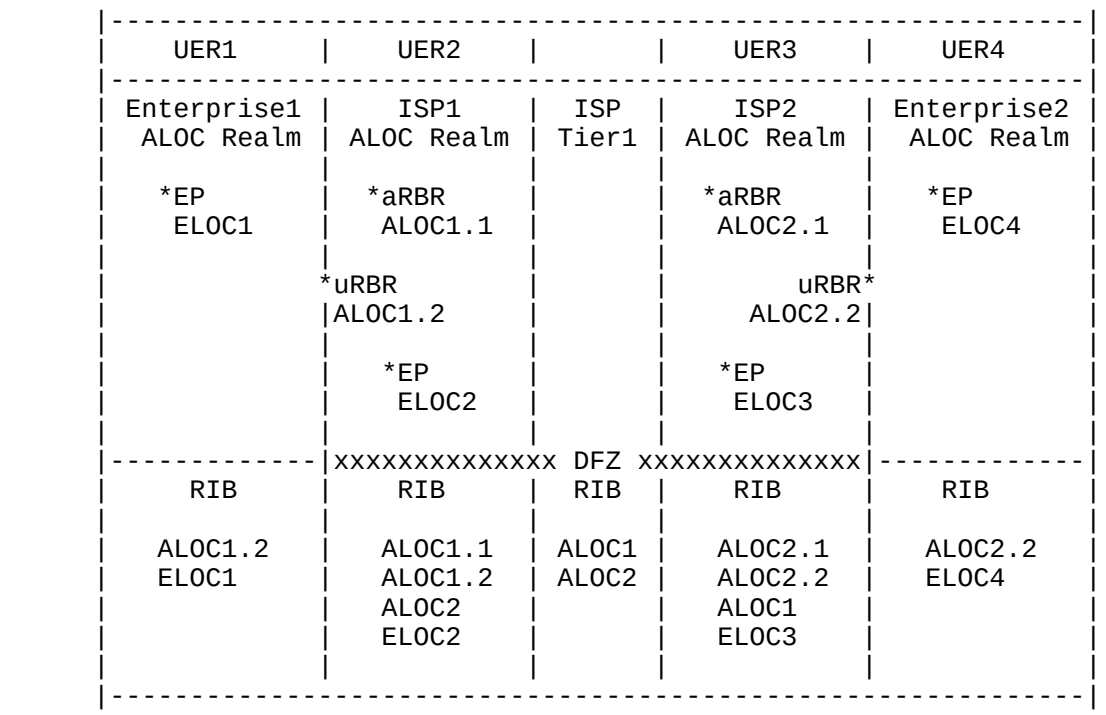

Figure 2: Long-term routing architecture of hIPv4

Also, the swap service for multicast can be removed when the forwarding planes are upgraded in all consequent ALOC realms. The source's ALOC RBR sets the FI-bits to 11, and a Reverse Path Forwarding (RPF) check is hereafter applied against the ALOC prefix in the locator header. Here, IP and transport protocol headers will not alternate.

A long-term evolution will provide a 32x32 bit locator space. The ALOC prefixes are allocated only to service providers; ELOC prefixes are only significant at a local ALOC realm. An enterprise can use a 32-bit locator space for its private network (the ALOC prefix is

rented from the attached ISP), and an ISP can use a 32-bit ELOC space to provide Internet connectivity services for its directly attached customers (residential and enterprise).

6.2. Exit, DFZ, and Approach Routing

This section provides an example of a hIPv4 session between two hIPv4 endpoints: an initiator in an ALOC realm where the forwarding plane has been upgraded to support the hIPv4 framework, and a responder residing in a remote ALOC realm with the classical IPv4 forwarding plane.

When the forwarding plane at the local ALOC realm has been upgraded, the endpoints must be informed about it; that is, extensions to DHCP are needed or the endpoints are manually configured to be notified that the local ALOC realm is fully hIPv4 compliant.

How a hIPV4 session is established follows:

1. The initiator queries the DNS server. The hIPv4 stack notices that the local and remote ALOCs do not match and therefore must use the hIPv4 header for the session. The hIPv4 stack of the initiator must assemble the packet as described in Section 5.2, step 1, except for the following:

   g. Set the FI-bits of the locator header to 01.

2. The hIPv4 packet is routed throughout the local ALOC realm according to the ALOC prefix of the locator header; local ALOC exit routing is applied.

3. The hIPv4 packet will reach the closest RBR of the local ALOC realm. When the RBR notices that the packet’s ALOC prefix of the locator header matches the local ALOC prefix and the FI-bits are set to 01, the RBR must:

   a. Verify that the received packet uses the hIPv4 protocol value in the protocol field of the IP header.

   b. Verify the IP and locator header checksums.

   c. Set the FI-bits of the locator header to 00.

   d. Decrease the TTL value by one.

   e. Calculate IP and locator header checksums.

   f. Forward the packet according to the value in the destination address field of the IP header.

4. The hIPv4 packet is routed to the responder as described in Section 5.2, steps 2 to 6. DFZ routing is applied.

5. The responder's application responds to the initiator and the returning packet takes almost the same steps as described in Section 5.2 except for:

6. The hIPv4 packet will reach the closest RBR of the initiator's ALOC realm. When the RBR notices that the value in the destination address field of the IP header matches the local ALOC prefix and the FI-bits are set to 00, the RBR must:

   a. Verify that the received packet uses the hIPv4 protocol value in the protocol field of the IP header.

   b. Verify the IP and locator header checksums.

   c. Set the FI-bits of the locator header to 10.

   d. Decrease the TTL value by one.

   e. Calculate IP and locator header checksums.

   f. Forward the packet according to the ELOC prefix of the locator header.

7. The hIPv4 packet is routed throughout the initiator's ALOC realm according to the ELOC prefix of the locator header. Remote ELOC approach routing is applied.

8. The hIPv4 stack of the responder must present the following to the extended IPv4 socket API:

   a. The source address of the IP header as the remote IP address.

   b. The destination address of the IP header as the local ALOC prefix.

   c. The ALOC prefix of the locator header as the remote ALOC prefix.

   d. The ELOC prefix of the locator header as the local IP address.

7. Decoupling Location and Identification

The design guidelines and rationale behind decoupling the location from identification are stated in [RFC6227]. Another important influence source is the report and presentations from the [Dagstuhl] workshop that declared "a future Internet architecture must hence decouple the functions of IP addresses as names, locators, and forwarding directives in order to facilitate the growth and new network-topological dynamisms of the Internet".

Therefore, identifier elements need to be added to the hIPv4 framework to provide a path for future applications to be able to remove the current dependency on the underlying network layer addressing scheme (local and remote IP address tuple).

However, there are various ways to apply an identifier/locator split, as discussed in an [ID/loc_Split] presentation from the MobiArch workshop at Sigcomm 2008. Thus, the hIPv4 framework will not propose or define a single identifier/locator split solution; a split can be achieved by, for example, a multipath transport protocol or by an identifier/locator database scheme such as HIP. A placeholder has been added to the locator header so identifier/locator split schemes can be integrated into the hIPv4 framework. But identifier/locator split schemes may cause privacy inconveniences, as discussed in [Mobility_&_Privacy].

Multipath transport protocols, such as SCTP and the currently under development Multipath TCP (MPTCP) [RFC6182], are the most interesting candidates to enable an identifier/locator split for the hIPv4 framework. MPTCP is especially interesting from hIPv4’s point of view; one of the main goals of MPTCP is to provide backwards compatibility with current implementations: hIPv4 shares the same goal.

MPTCP itself does not provide an identifier/locator database scheme as HIP does. Instead, MPTCP is proposing a token -- with local meaning -- to manage and bundle subflows under one session between two endpoints. The token can be considered to have the characteristics of a session identifier, providing a generic cookie mechanism for the application layer and creating a session layer between the application and transport layers. Thus, the use of a session identifier will provide a mechanism to improve mobility, both in site and endpoint mobility scenarios.

Since the session identifier improves site and endpoint mobility, routing scalability is improved by introducing a hierarchical addressing scheme, why then add an identifier/locator database scheme to the hIPv4 framework? Introducing an identifier/locator database

scheme, as described in HIP, Identifier/Locator Network Protocol [ILNP] and Name-Based Sockets [NBS], might ease or remove the locator renumbering dependencies at firewalls that are used to scope security zones, but this approach would fundamentally change the currently deployed security architecture.

However, combining an identifier/locator database scheme with DNS Security (DNSSEC) [RFC4033] is interesting. Today, security zones are scoped by using locator prefixes in the security rule sets. Instead, a Fully Qualified Domain Name (FQDN) could be used in the rule sets and the renumbering of locator prefixes would no longer depend upon the security rule sets in firewalls. Another interesting aspect is that an FQDN is and needs to be globally unique. The ALOC prefix must be globally unique, but ELOC prefixes are only regionally unique and in the long-term only locally unique. Nevertheless, combining identifier/locator database schemes with security architectures and DNSSEC needs further study.

In order to provide multi-homing and mobility capabilities for single path transport protocols such as TCP and UDP, an identifier/locator database scheme is needed. This scheme can also be used to create a bidirectional NAT traversal solution with a locator translation map consisting of private locator prefixes and public identifiers at the border router.

The hIPv4 routing architecture provides only location information for the endpoints; that is, the ELOC describes how the endpoint is attached to the local network, and the ALOC prefixes describe how the endpoint is attached to the Internet. Identifier/locator split schemes are decoupled from the routing architecture -- the application layer may or may not make use of an identifier/locator split scheme.

8. ALOC Use Cases

Several ALOC use cases are explored in this section. As mentioned in Section 5.1, ALOC describes an area in the Internet that can span several autonomous systems (ASes), or if the area is equal to an AS you can say that the ALOC describes an AS. When the ALOC describes an area, it is hereafter called an anycast ALOC.

The ALOC can also be used to describe a specific node between two ALOC realms, e.g., a node installed between a private and an ISP ALOC realm, or between two private ALOC realms. In this use case the ALOC describes an attachment point, e.g., where a private network is attached to the Internet. This ALOC type is hereafter called a unicast ALOC.

The main difference between anycast and unicast ALOC types is:

- In an anycast ALOC scenario, ELOC routing information is shared between the attached ALOC realms.

- In a unicast ALOC scenario, no ELOC routing information is shared between the attached ALOC realms.

Unicast ALOC functionalities should not be deployed between private and ISP ALOC realms in the intermediate routing architecture -- it would require too many locators from the GLB space. Instead, unicast ALOC functionality will be used to separate private ALOC realms.

ALOC space is divided into two types, a globally unique ALOC space (a.k.a. GLB) that is installed in DFZ, and a private ALOC space that is used inside private networks. Private ALOCs use the same locator space as defined in [RFC1918]; a private ALOC must be unique inside the private network and not overlap private ELOC prefixes. Only ISPs should be allowed to apply for global ALOC prefixes. For further discussion, see Appendix A. The ISP should aggregate global ALOC prefixes as much as possible in order to reduce the size of the routing table in DFZ.

When a user logs on to the enterprise’s network, the endpoint will receive the following locator prefixes via provisioning means (e.g., DHCP or manually configured):

- One ELOC prefix for each network interface.

- One private ALOC prefix due to

  - The enterprise has recently been merged with another enterprise and overlapping ELOC spaces exist.

- Several private ALOC prefixes due to

  - The enterprise network spans high-speed long-distance connections. It is well-known that TCP cannot sustain high throughput for extended periods of time. Higher throughput might be achieved by using multiple paths concurrently.

- One or several global ALOC prefixes. These ALOCs describe how the enterprise network is attached to the Internet.

As the user establishes a session to a remote endpoint, DNS is usually used to resolve remote locator prefixes. DNS will return ELOC and ALOC prefixes of the remote endpoint. If no ALOC prefixes are returned, a classical IPv4 session is initiated to the remote

endpoint. When ALOC prefixes are returned, the initiator compares the ALOC prefixes with its own local ALOC prefixes (that are provided via DHCP or manually configured).

- If the remote ALOC prefix is from the private ALOC space, the initiator will use the given private ALOC prefix for the session.

Two use cases exist to design a network to use private ALOC functionality. The remote endpoint is far away, leveraging high-speed long-distance connections, and in order to improve performance for the session a multipath transport protocol should be used.

The other use case is when the remote endpoint resides in a network that recently has been merged and private ELOC [RFC1918] spaces overlap if no renumbering is applied. One or several unicast ALOC solutions are needed in the network between the initiator and responder. For long-distance sessions with no overlapping ELOC prefixes, anycast or unicast ALOC solutions can be deployed.

A third use case follows; again the initiator compares returned ALOC prefixes from DNS with its own local ALOC prefixes:

- If the remote ALOC prefix is from the global ALOC space and the remote ALOC doesn't match the given global ALOC prefix, the initiator will use the given global ALOC prefix for the session.

In this use case the remote endpoint resides outside the enterprise's private network, and the global remote ALOC prefixes indicate how the remote network is attached to the Internet. When a multipath transport protocol is used, the subflows can be routed via separate border routers to the remote endpoint -- both at the local and remote sites, if both are multi-homed. The initiator's egress packets in the local ALOC realm can be identified by the protocol value in the IP header, routed to an explicit path (e.g., MPLS LSP, L2TPv3 tunnel, etc.) based on the ALOC prefix in the locator header. A local ALOC overlay exit routing scheme can be designed. In the long-term routing architecture the overlay, the tunnel mechanism, can be removed; see Section 6.2.

Figure 3 shows a conceptual diagram with two endpoints having a multipath session over a VPN connection and over the Internet (in the intermediate routing architecture).

Legend: *attachment point in the ALOC realm
UER=Unique ELOC region
EP=Endpoint
aRBR=anycast RBR
uRBR=unicast RBR
BR=Border Router

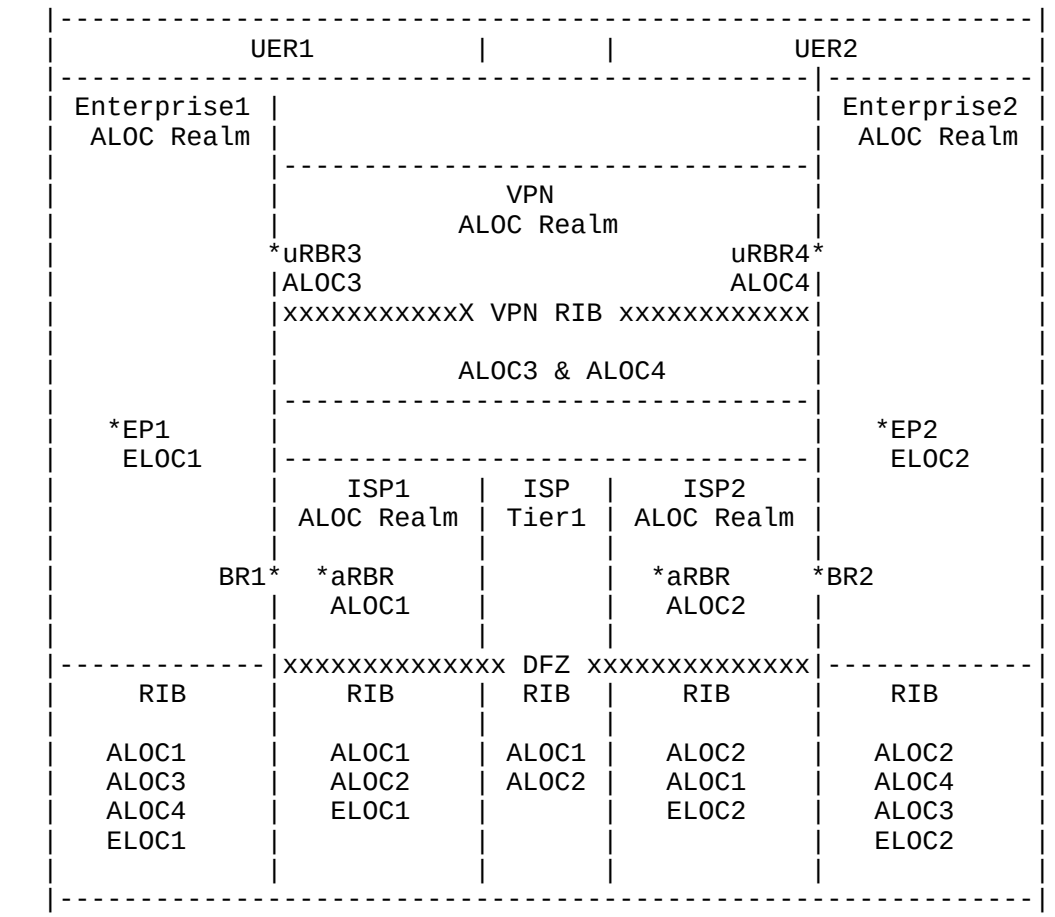

Figure 3: Multi-pathing via VPN and the Internet

The first subflow is established from the initiator (EP1) via uRBR3 and uRBR4 (both use a private unicast ALOC prefix) to the responder (EP2). Normal unicast forwarding is applied; ALOC prefixes of uRBR3 and uRBR4 are installed in the routing tables of both the local and remote ALOC realms. A second subflow is established via the Internet, that is, via BR1->BR2 to EP2. 0/0 exit routing is used to enter the Internet at both ALOC realms.

Note that ELOC prefixes can overlap since the local and remote ALOC realms reside in different ELOC regions and are separated by private unicast ALOC prefixes.

The fourth use case is to leverage the private and global ALOC functionalities to be aligned with the design and implementation of [Split-DNS] solutions.

The fifth use case is for residential users. A residential user may use one or several ALOC prefixes, depending upon the service offer and network design of the ISP. If the ISP prefers to offer advanced support for multipath transport protocols and local ALOC exit routing, the residential user is provided with several ALOC prefixes. The ALOC provided for residential users is taken from the GLB space and anycast ALOC functionality is applied.

9. Mandatory Extensions

9.1. Overview

To implement the hierarchical IPv4 framework, some basic rules are needed:

1. The DNS architecture must support a new extension; an A type Resource Record should be able to associate ALOC prefixes.

2. An endpoint upgraded to support hIPv4 shall have information about the local ALOC prefixes; the local ALOC prefixes can be configured manually or provided via provisioning means such as DHCP.

3. A globally unique IPv4 address block shall be reserved; this block is called the Global Locator Block (GLB). A service provider can have one or several ALOC prefixes allocated from the GLB.

4. ALOC prefixes are announced via current BGP to adjacent peers. They are installed in the RIB of the DFZ. When the hIPV4 framework is fully implemented, only ALOC prefixes are announced between the BGP peers in the DFZ.

5. An ALOC realm must have one or several RBRs attached to it. The ALOC prefix is configured as an anycast IP address on the RBR. The anycast IP address is installed to appropriate routing protocols in order to be distributed to the DFZ.

6. The IPv4 socket API at endpoints must be extended to support local and remote ALOC prefixes. The modified IPv4 socket API must be backwards compatible with the current IPv4 socket API. The outgoing hIPv4 packet must be assembled by the hIPv4 stack with

the local IP address from the socket as the source address and the remote ALOC prefix as the destination address in the IP header. The local ALOC prefix is inserted in the ALOC field of the locator header. The remote IP address from the socket API is inserted in the ELOC field of the locator header.

9.2. DNS Extensions

Since the hierarchical IPv4 framework introduces an extended addressing scheme and because DNS serves as the "phone book" for the Internet, it is obvious that DNS needs a new Resource Record (RR) type to serve endpoints that are upgraded to support hIPv4. Future RR types must follow the guidelines described in [RFC3597] and [RFC5395] with the following characteristics:

- Associated with the appropriate Fully Qualified Domain Name (FQDN), inserted in the NAME field.
- Assigned a new integer (QTYPE) in the TYPE field, to be assigned by IANA.
- The CLASS field is set to IN.
- The RDATA field is of an unknown type as defined in [RFC3597] and shall have the following format:
  - Preference subfield: A 16-bit integer that specifies the preference given to this RR among others associated with a FQDN. Lower values are preferred over higher values.
  - ALOC subfield: A 32-bit integer that specifies the Area Locator of the associated FQDN.

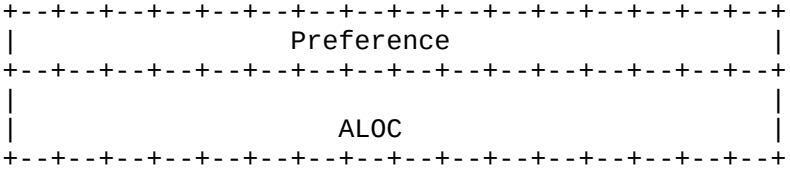

Figure 4: RDATA format of the ALOC RR

Only endpoints that have been upgraded to support hIPv4 shall make use of the new ALOC RR. Also, there is no need to define a new ELOC RR because the A RR is used for that purpose when the ALOC RR is returned.

9.3. Extensions to the IPv4 Header

Figure 5 shows how the locator header is added to the current IPv4 header, creating a hIPv4 header.

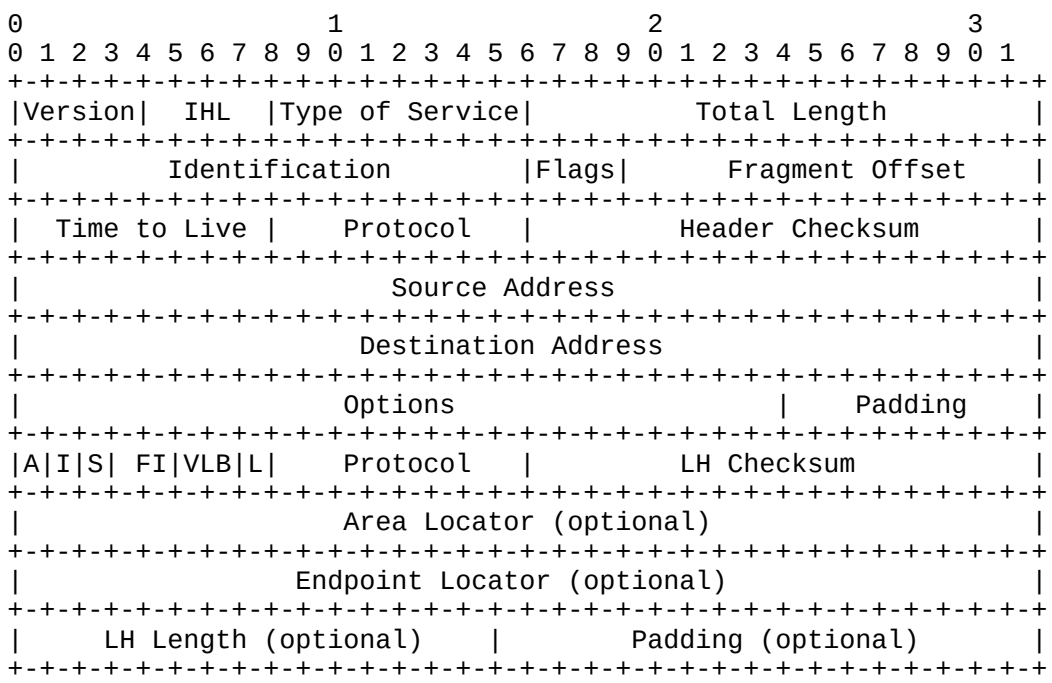

Figure 5: hIPv4 header

Version: 4 bits

The Version field is identical to that of RFC 791.

IHL: 4 bits

The Internet Header Length field is identical to that of RFC 791.

Type of Service: 8 bits

The Type of Service is identical to that of RFC 791.

Total Length: 16 bits

The Total Length field is identical to that of RFC 791.

Identification: 16 bits

The Identification field is identical to that of RFC 791.

Flags: 3 bits

The Flags field is identical to that of RFC 791.

Fragment Offset: 13 bits

The Fragment Offset field is identical to that of RFC 791.

Time to Live: 8 bits

The Time to Live field is identical to that of RFC 791.

Protocol: 8 bits

A new protocol number must be assigned for hIPv4.

Header Checksum: 16 bits

The Header Checksum field is identical to that of RFC 791.

Source Address: 32 bits

The Source Address field is identical to that of RFC 791.

Destination Address: 32 bits

The Destination Address field is identical to that of RFC 791.

Options and Padding: Variable length

The Options and Padding fields are identical to that of RFC 791.

ALOC Realm Bit, A-bit: 1 bit

When the initiator and responder reside in different ALOC realms, the A-bit is set to 1 and the Area and Endpoint Locator fields must be used in the locator header. The size of the locator header is 12 bytes. When the A-bit is set to 0, the initiator and responder reside within the same ALOC realm. The Area and Endpoint Locator shall not be used in the locator header. The size of the locator header is 4 bytes.

Identifier Bit, I-bit: 1 bit

The identifier bit is set to 1 if the endpoint is using an identifier/locator split scheme within the locator header. The identifier/locator split scheme must indicate by how much the size of the locator header is increased. The Locator Header Length field is also added to the locator header.

Swap Bit, S-bit: 1 bit

The initiator sets the swap bit to 0 in the hIPv4 packet. An RBR will set this bit to 1 when it is swapping the source and destination addresses of the IP header with the ALOC and ELOC prefixes of the locator header.

Forwarding Indicator, FI-bits: 2 bits

The purpose of the Forwarding Indicator (FI) field is to provide a mechanism for a future forwarding plane to identify which Forwarding Information Base (FIB) should be used for inter-ALOC realm sessions. The new forwarding plane will remove the swap functionality of IP and locator header values for both unicast and multicast sessions. The outcome is that the IP and transport protocol headers will remain intact and only FI and LH checksum values in the locator header will alternate. The following values are defined:

01: Local ALOC exit routing mode. The initiator shall set the FI-bits to 01 and the ALOC prefix in the locator header is used to forward the packets to the RBR that is the owner of the local ALOC prefix. The RBR shall change the FI-bits to 00.

00: DFZ routing mode. The local ALOC RBR shall forward the packets according to the value in the destination address field of the IP header. The DFZ routers shall forward the packets based on the value in the destination address field of the IP header unless the destination address matches the local ALOC prefix. When this situation occurs, the packet enters the remote ALOC realm and the remote RBR shall change the FI-bits to 10.

10: Remote ELOC approach routing mode. The remote ALOC RBR and following routers shall forward the packets based on the ELOC prefix in the locator header.

11: Inter-ALOC RPF check mode. The local ALOC RBR changes the FI-bits to 11 and the following inter-ALOC routers on the shared tree shall apply the RPF check against the ALOC prefix in the locator header.

Valiant Load-Balancing, VLB-bits: 2 bits (optional, subject for further research)

The purpose of the Valiant Load-Balancing field is to provide a mechanism for multipath-enabled transport protocols to request explicit paths in the network for subflows, which are component parts of a session between two endpoints. The subflow path request can be set as follows:

00: Latency-sensitive application. Only one single subflow (multipath not applied), the shortest path through the network is requested.

01: First subflow. The shortest path or Valiant Load-Balancing might be applied.

11: Next subflow(s). Valiant Load-Balancing should be applied

Load-Balanced, L-bit: 1 bit (optional, subject for further research)

The initiator must set the L-bit to zero. A Valiant Load-Balancing-capable node can apply VLB switching for the session if the value is set to zero; if the value is set to 1, VLB switching is not allowed. When VLB switching is applied for the session, the node applying the VLB algorithm must set the value to 1.

Protocol: 8 bits

The Protocol field is identical to that of RFC 791.

Locator Header Checksum: 16 bits

A checksum is calculated for the locator header only. The checksum is computed at the initiator, recomputed at the RBR, and verified at the responder. The checksum algorithm is identical to that of RFC 791.

Area Locator (optional): 32 bits

The Area Locator is an IPv4 address assigned to locate an ALOC realm in the Internet. The ALOC is assigned by an RIR to a service provider. The ALOC is globally unique because it is allocated from the GLB.

Endpoint Locator (optional): 32 bits

The Endpoint Locator is an IPv4 address assigned to locate an endpoint in a local network. The ELOC block is assigned by an RIR to a service provider or to an enterprise. In the intermediate routing architecture the ELOC block is only unique in a geographical region. The final policy of uniqueness shall be defined by the RIRs. In the long-term routing architecture the ELOC block is no longer assigned by an RIR; it is only unique in the local ALOC realm.

Locator Header Length (optional): 16 bits

The Locator Header Length is the total length of the locator header. Locator Header Length is applied when the identifier bit is set to 1. Identifier/locator split scheme parameters are inserted into the locator header after this field.

Padding (optional): variable

The locator header padding is used to ensure that the locator header ends on a 32-bit boundary. The padding is zero.

10. Consequences

10.1. Overlapping Local and Remote ELOC Prefixes/Ports

Because an ELOC prefix is only significant within the local ALOC realm, there is a slight possibility that a session between two endpoints residing in separate ALOC realms might use the same local and remote ELOC prefixes. But the session is still unique because the two processes communicating over the transport protocol form a logical session that is uniquely identifiable by the 5-tuple involved, by the combination of <protocol, local IP address, local port, remote IP address, remote port>.

The session might no longer be unique when two initiators with the same local ELOC prefix residing in two separate ALOC realms are accessing a responder located in a third ALOC realm. In this scenario, the possibility exists that the initiators will use the same local port value. This situation will cause an "identical session situation" for the application layer.

To overcome this scenario, the hIPv4 stack must accept only one unique session with the help of the ALOC information. If there is an "identical session situation", i.e., both initiators use the same values in the 5-tuple <protocol, local IP address, local port, remote IP address, remote port>, the hIPv4 stack shall allow only the first

established session to continue. The following sessions must be prohibited and the initiator is informed by ICMP notification about the "identical session situation".

MPTCP introduces a token that is locally significant and currently defined as 32 bits long. The token will provide a sixth tuple for future applications to identify and verify the uniqueness of a session. Thus, the probability to have an "identical session situation" is further reduced. By adding an identifier/locator database scheme to the hIPv4 framework, the "identical session situation" is completely removed.

10.2. Large Encapsulated Packets

Adding the locator header to an IPv4 packet in order to create a hIPv4 packet will increase the size of it, but since the packet is assembled at the endpoint it will not add complications of the current Path MTU Discovery (PMTUD) mechanism in the network. The intermediate network between two endpoints will not see any difference in the size of packets; IPv4 and hIPv4 packet sizes are the same from the network point of view.

10.3. Affected Applications

There are several applications that insert IP address information to the payload of a packet. Some applications use the IP address information to create new sessions or for identification purposes. Some applications collect IP address information to be used as referrals. This section tries to list the applications that need to be enhanced; however, this is by no means a comprehensive list. The applications can be divided into five main categories:

- Applications based on raw sockets - a raw socket receives packets containing the complete header, in contrast to the other sockets that only receive the payload.

- Applications needed to enable the hIPv4 framework, such as DNS and DHCP databases, which must be extended to support ALOC prefixes.

- Applications that insert IP addresses into the payload or use the IP address for setting up new sessions or for some kind of identification or as referrals. An application belonging to this category cannot set up sessions to other ALOC realms until extensions have been incorporated. Within the local ALOC realm there are no restrictions since the current IPv4 scheme is still valid. The following applications have been identified:

- SIP: IP addresses are inserted in the SDP offers/answers, XML body, Contact, Via, maddr, Route, Record-Route SIP headers.

- Mobile IP: the mobile node uses several IP addresses during the registration process.

- IPsec AH: designed to detect alterations at the IP packet header.

- RSVP: Resource Reservation Protocol (RSVP) messages are sent hop-by-hop between RSVP-capable routers to construct an explicit path.

- ICMP: notifications need to be able to incorporate ALOC information and assemble the hIPv4 header in order to be routed back to the source.

- Source Specific Multicast: the receiver must specify the source address.

- IGMPv3: a source-list is included in the IGMP reports.

o Applications related to security, such as firewalls, must be enhanced to support ALOC prefixes.

o Applications that will function with FQDN, but many use IP addresses instead, such as ping, traceroute, telnet, and so on. The CLI syntax needs to be upgraded to support ALOC and ELOC information via the extended socket API.

At first glance, it seems that a lot of applications need to be re-engineered and ported, but the situation is not all that bad. The applications used inside the local ALOC realm (e.g., an enterprise's private network) do not need to be upgraded, neither in the intermediate nor in the long-term architecture. The classical IPv4 framework is preserved. Only IP-aware applications used between ALOC realms need to be upgraded to support the hIPv4 header. IPv6 has the definitions in place of the applications mentioned above, but the migration of applications from IPv4 to IPv6 can impose some capital expenditures for enterprises, especially if the applications are customized or homegrown; see [Porting_IPv4].

As stated earlier, hIPv4 does not require to port applications used inside a private network. The conclusion is that, whatever next generation architecture is deployed, some applications will suffer, either during the transition period or when being re-engineered in order to be compatible with the new architecture.

10.4. ICMP

As long as the ICMP request is executed inside the local ALOC realm, the normal IPv4 ICMP mechanism can be used. As soon as the ICMP request exits the local ALOC realm, the locator header shall be used in the notifications. Therefore, extensions to the ICMP shall be implemented. These shall be compatible with [RFC4884] and support ALOC and ELOC information.

10.5. Multicast

Since local ELOC prefixes are only installed in the routing table of the local ALOC realm, there is a constraint with Reverse Path Forwarding (RPF) that is used to ensure loop-free forwarding of multicast packets. The source address of a multicast group (S,G) is used against the RPF check. The address of the source can no longer be used as a RPF checkpoint outside the local ALOC realm.

To enable RPF globally for an (S,G), the multicast-enabled RBR (mRBR) must at the source’s ALOC realm replace the value of the source address field in the IP header with the local ALOC prefix for inter-ALOC multicast streams. This can be achieved if the local RBR acts also as an anycast Rendezvous Point with MSDP and PIM capabilities. With these functionalities the RBR becomes a multicast-enabled RBR (mRBR). The source registers at the mRBR and a source tree is established between the source and the mRBR. When an inter-ALOC realm receiver subscribes to the multicast group, the mRBR has to swap the hIPv4 header in the following way:

a. Verify that the received packet uses the hIPv4 protocol value in the protocol field of the IP header.

b. Verify IP, locator, and transport protocol header checksums.

c. Replace the source address in the IP header with the local ALOC prefix.

d. Set the S-field to 1.

e. Decrease the TTL value by one.

f. Calculate IP, locator, and transport protocol header checksums. Transport protocol header calculations do not include the locator header fields.

g. Forward the packet to the shared multicast tree.

In order for the mRBR to function as described above, the source must assemble the multicast hIPv4 packet in the following way:

a. Set the local IP address (S) from the API in the source address field of the IP header and in the ELOC field of the locator header.

b. Set the multicast address (G) from the API in the destination address field of the IP header.

c. Set the local ALOC prefix in the ALOC field of the locator header.

d. Set the transport protocol value in the protocol field of the locator header and the hIPv4 protocol value in the protocol field of the IP header.

e. Set the desired parameters in the A-, I-, S-, VLB-, and L-fields of the locator header.

f. Set the FI-bits of the locator header to 00.

g. Calculate IP, locator, and transport protocol header checksums. Transport protocol header calculations do not include the locator header fields. When completed, the packet is transmitted.

The downstream routers from the mRBR towards the receiver will use the source address (which is the source's ALOC prefix after the mRBR) in the IP header for RPF verification. In order for the receiver to create Real-time Transport Control Protocol (RTCP) receiver reports, all information is provided in the hIPv4 header of the packet.

Because Source Specific Multicast (SSM) and IGMPv3 use IP addresses in the payload, both protocols need to be modified to support the hIPv4 framework.

11. Traffic Engineering Considerations

When the intermediate phase of the hIPv4 framework is fully implemented, ingress load balancing to an ALOC realm can be influenced by the placement of RBRs at the realm; an RBR provides a shortest path scheme. Also, if RIR policies allow, a service provider can have several ALOCs assigned. Hence, traffic engineering and filtering can be done with the help of ALOC prefixes. For example, sensitive traffic can be aggregated under one ALOC prefix that is not fully distributed into the DFZ.

If needed, an ALOC traffic engineering solution between ALOC realms might be developed, to create explicit paths that can be engineered via specific ALOC prefixes. For example, develop a mechanism similar to the one described in [Pathlet_Routing]. Further studies are needed; first it should be evaluated whether there is demand for such a solution.

Ingress load balancing to a private remote ALOC realm (remote site) is influenced by how many attachment points to the Internet the site uses and where the attachment points are placed at the site. In order to apply local ALOC exit routing, e.g., from a multi-homed site, some new network nodes are needed between the initiator and the border routers of the site.

In the intermediate routing architecture this is achieved by using overlay architectures such as MPLS LSP, L2TPv3 tunnels, etc. The new network node(s) shall be able to identify hIPv4 packets, based on the protocol field in the IP header, and switch the packets to explicit paths based on the ALOC prefix in the locator header. In the long-term routing architecture the overlay solution is replaced with a new forwarding plane; see Section 6.2.

Together with a multipath transport protocol, the subflows can be routed via specific attachment points, that is, border routers sitting between the private local/remote ALOC realms (multi-homed sites) and the Internet. Multi-homing becomes multi-pathing. For details, see Appendix B.

11.1. Valiant Load-Balancing

The use of multipath-enabled transport protocols opens up the possibility to develop a new design methodology of backbone networks, based on Valiant Load-Balancing [VLB]. If two sites that are connected with a single uplink to the Internet, and the endpoints are using multipath-enabled transport protocols and are attached to the network with only one interface/ELOC-prefix, both subflows will most likely take the shortest path throughout the Internet. That is, both subflows are established over the same links and when there is congestion on a link or a failure of a link, both subflows might simultaneously drop packets. Thus, the benefit of multi-pathing is lost.

The "subflows-over-same-links" scenario can be avoided if the subflows are traffic engineered to traverse the Internet on different paths, but this is difficult to achieve by using classical traffic engineering, such as IGP tuning or MPLS-based traffic engineering. By adding a mechanism to the locator header, the "subflows-over-same-links" scenario might be avoided.

If the RBR functionality is deployed on a Valiant Load-Balancing enabled backbone node -- hereafter called vRBR -- and the backbone nodes are interconnected via logical full meshed connections, Valiant Load-Balancing can be applied for the subflows. When a subflow has the appropriate bits set in the VLB-field of the locator header, the first ingress vRBR shall do VLB switching of the subflow. That is, the ingress vRBR is allowed to do VLB switching of the subflow's packets if the VLB-bits are set to 01 or 11, the L-bit is set to 0, and the local ALOC prefix of the vRBR matches the ALOC-field's prefix. If there are no ALOC and ELOC fields in the locator header, but the other fields' values are set as described above, the vRBR should apply VLB switching as well for the subflow -- because it is an intra-ALOC realm subflow belonging to a multipath-enabled session.

With this combination of parameters in the locator header, the subflow is VLB switched only at the first ALOC realm and the subflows might be routed throughout the Internet on different paths. If VLB switching is applied at every ALOC realm, this would most likely add too much latency for the subflows. The VLB switching at the first ALOC realm will not separate the subflows on the first and last mile links (site with a single uplink). If the subflows on the first and last mile link need to be routed on separate links, the endpoints should be deployed in a multi-homed environment. Studies on how Valiant Load-Balancing is influencing traffic patterns between interconnected VLB [iVLB] backbone networks have been done. Nevertheless, more studies are needed regarding Valiant Load-Balancing scenarios.

12. Mobility Considerations

This section considers two types of mobility solutions: site mobility and endpoint mobility.

Site mobility:

Today, classical multi-homing is the most common solution for enterprises that wish to achieve site mobility. Multi-homing is one of the key findings behind the growth of the DFZ RIB; see [RFC4984], Sections 2.1 and 3.1.2. The hIPv4 framework can provide a solution for enterprises to have site mobility without the requirement of implementing a classical multi-homed solution.

One of the reasons to deploy multi-homing is to avoid renumbering of the local infrastructure when an upstream ISP is replaced. Thus, today, PI-address blocks are deployed at enterprises. In the intermediate routing architecture, an enterprise is allocated a regional PI ELOC block (for details, see Appendix A) that is used for internal routing. The upstream ISP provides an ALOC prefix that

describes how the enterprise's network is connected to the Internet. If the enterprise wishes to switch to another ISP, it only changes the ALOC prefix at endpoints, from the previous ISP's ALOC prefix to the new ISP's ALOC prefix, without connectivity interruptions in the local network since the ALOC prefix is only used for Internet connectivity -- several ALOCs can be used simultaneously at the endpoints; thus, a smooth migration from one ISP to another is possible. In the long-term routing architecture, when the forwarding plane is upgraded, the regional PI ELOC block is returned to the RIR and the enterprise can use a full 32-bit ELOC space to design the internal routing topology.

An enterprise can easily become multi-homed or switch ISPs. The local ELOC block is used for internal routing and upstream ISPs provide their ALOC prefixes for Internet connectivity. Multi-homing is discussed in detail in Appendix B.

Endpoint mobility:

As said earlier, MPTCP is the most interesting identifier/locator split scheme to solve endpoint mobility scenarios. MPTCP introduces a token, which is locally significant and currently defined as 32 bits long. The token will provide a sixth tuple to identify and verify the uniqueness of a session. This sixth tuple -- the token -- does not depend upon the underlying layer, the IP layer. The session is identified with the help of the token and thus the application is not aware when the locator parameters are changed, e.g., during a roaming situation, but it is required that the application is not making use of ELOC/ALOC information. In multi-homed scenarios, the application can make use of ELOC information, which will not change if the endpoint is fixed to the location.

Security issues arise: the token can be captured during the session by, for example, a man-in-the-middle attack. These attacks can be mitigated by applying [tcpcrypt], for example. If the application requires full protection against man-in-the-middle attacks, the user should apply the Transport Layer Security Protocol (TLS) [RFC5246] for the session.

The most common endpoint mobility use case today is that the responder resides in the fixed network and the initiator is mobile. Thus, MPTCP will provide roaming capabilities for the mobile endpoint, if both endpoints are making use of the MPTCP extension. However, in some use cases, the fixed endpoint needs to initialize a session to a mobile responder. Therefore, Mobile IP (MIP) [RFC5944] should incorporate the hIPv4 extension -- MIP provides a rendezvous service for the mobile endpoints.

Also, many applications provide rendezvous services for their users, e.g., SIP, peer-to-peer, Instant Messaging services. A generic rendezvous service solution can be provided by an identifier/locator database scheme, e.g., HIP, ILNP, or NBS. If desired, the user (actually the application) can make use of one of these rendezvous service schemes, such as extended MIP, some application-specific rendezvous services, or a generic rendezvous service -- or some combination of them.

The hIPv4 framework will not define which identifier/locator split solution should be used for endpoint mobility. The hIPv4 framework is focusing on routing scalability and supports several identifier/locator split solutions that can be exploited to develop new services, with the focus on endpoint mobility.

13. Transition Considerations

The hIPv4 framework is not introducing any new protocols that would be mandatory for the transition from IPv4 to hIPv4; instead, extensions are added to existing protocols. The hIPv4 framework requires extensions to the current IPv4 stack, to infrastructure systems, and to some applications that use IP address information, but the current forwarding plane in the Internet remains intact, except that a new forwarding element (the RBR) is required to create an ALOC realm.

Extensions to the IPv4 stack, to infrastructure systems, and to applications that make use of IP address information can be deployed in parallel with the current IPv4 framework. Genuine hIPv4 sessions can be established between endpoints even though the current unidimensional addressing structure is still present.

When will the unidimensional addressing structure be replaced by a hierarchical addressing scheme and a fourth hierarchy added to the routing architecture? The author thinks there are two possible tipping points:

- When the RIB of DFZ is getting close to the capabilities of current forwarding planes. Who will pay for the upgrade? Or will the service provider only accept ALOC prefixes from other service providers and avoid capital expenditures?

- When the depletion of IPv4 addresses is causing enough problems for service providers and enterprises.

The biggest risk and reason why the hIPv4 framework will not succeed is the very short time frame until the expected depletion of the IPv4 address space occurs -- actually the first RIR has run out of IPv4

addresses during the IESG review process of this document (April 2011). Also, will enterprises give up their global allocation of the current IPv4 address block they have gained, as an IPv4 address block has become an asset with an economical value.

The transition requires the upgrade of endpoint's stack, and this is a drawback compared to the [CES] architectures proposed in [RFC6115]. A transition to an architecture that requires the upgrade of endpoint's stack is considerably slower than an architecture that requires only upgrade of some network nodes. But the transition might not be as slow or challenging at it first seems since hIPv4 is an evolution of the current deployed Internet.

- Not all endpoints need to be upgraded; the endpoints that do not establish sessions to other ALOC realms can continue to make use of the classical IPv4 framework. Also, legacy applications that are used only inside a local ALOC realm do not need to be ported to another framework. For further details, see Appendix C.

- Upgrading endpoint's stack, e.g., at critical or complicated systems, will definitely take time; thus, it would be more convenient to install a middlebox in front of such systems. It is obvious that the hIPv4 framework needs a middlebox solution to speed up the transition; combining CES architectures with the hIPv4 framework might produce such a middlebox. For further details, see Appendix D.

- The framework is incrementally deployable. Not all endpoints in the Internet need to be upgraded before the first IPv4 block can be released from a globally unique allocation status to a regionally unique allocation status. That is, to achieve ELOC status for the prefixes used in a local network in the intermediate routing architecture, see Appendix D. An ALOC realm that wishes to achieve local unique status for its ELOC block in the long-term routing architecture does not need to wait for other ALOC realms to proceed to the same level simultaneously. It is sufficient that the other ALOC realms have achieved the intermediate routing architecture status. For further details, see Section 6.

14. Security Considerations

Because the hIPv4 framework does not introduce other network mechanisms than a new type of border router to the currently deployed routing architecture, the best current practices for securing ISP networks are still valid. Since the DFZ will no longer contain ELOC prefixes, there are some benefits and complications regarding security that need to be taken into account.

The hijacking of a single ELOC prefix by longest match from another ALOC realm is no longer possible because the prefixes are separated by a locator, the ALOC. To carry out a hijack of a certain ELOC prefix, the whole ALOC realm must be routed via a bogus ALOC realm. Studies should be done with the Secure Inter-Domain Routing (SIDR) working group to determine whether the ALOC prefixes can be protected from hijacking.

By not being able to hijack a certain ELOC prefix, there are some implications when mitigating distributed denial-of-service (DDoS) attacks. This implication occurs especially in the long-term routing architecture, e.g., when a multi-homed enterprise is connected with unicast ALOC RBRs to the ISPs.

One method used today to mitigate DDoS attacks is to inject a more specific prefix (typically host prefix) to the routing table so that the victim of the attack is "relocated", i.e., a sinkhole is created in front of the victim. The sinkhole may separate bogus traffic from valid traffic or analyze the attack. The challenge in the long-term routing architecture is how to reroute a specific ELOC prefix of the multi-homed enterprise when the ELOC prefix cannot be installed in the ISP’s routing table.

Creating a sinkhole for all traffic designated to an ALOC realm might be challenging and expensive, depending on the size of the multi-homed enterprise. To have the sinkhole at the enterprise’s ALOC realm may saturate the connections between the enterprise and ISPs, thus this approach is not a real option.

By borrowing ideas from a service-centric networking architecture [SCAFFOLD], a sinkhole service can be created. An example of how a distributed sinkhole service can be designed follows:

a. A firewall (or similar node) at the victim’s ALOC realm discovers an attack. The security staff at the enterprise realizes that the amount of the incoming traffic caused by the attack is soon saturating the connections or other resources. Thus, the staff informs the upstream ISPs of the attack, also about the victim’s ALOC prefix X and ELOC prefix Y.

b. The ISP reserves the resources for the sinkhole service. These resources make use of ALOC prefix Z; the resources are programmed with a service ID and the victim’s X and Y prefixes. The ISP informs the victim’s security staff of the service ID. The ISP applies a NAT rule on their RBRs and/or hIPv4-enabled routers. The NAT rule replaces the destination address in the IP header of packets with Z when the destination address of the IP header matches X and the ELOC prefix of the locator header

matches Y. Also, the service ID is inserted to the locator header; the service ID acts as a referral for the sinkhole. It is possible that the sinkhole serves several victims; thus, a referral is needed. PMTUD issues must be taken into account.

c. The victim's inbound traffic is now routed at the RBRs and/or hIPv4-enabled routers to the sinkhole(s); the traffic is identified by the service ID. Bogus traffic is discarded at the sinkhole, for valid traffic the value of the destination address in the IP header Z is replaced with X. By using a service ID in the analyzed packets, the enterprise is informed that the packets containing service ID are valid traffic and allowed to be forwarded to the victim. It might be possible that not all upstream ISPs are redirecting traffic to the distributed sinkholes. Thus, traffic that does not contain the agreed service ID might be bogus. Also, by inserting a service ID to the valid packets, overlay solutions between the routers, sinkholes, and victim can be avoided. In case the valid packet with a service ID traverses another RBR or hIPv4-enabled router containing the same NAT rule, that packet is not rerouted to the sinkhole. The enterprise shall ensure that the victim does not use the service ID in its replies -- if the attacker becomes aware of the service ID, the sinkhole is disarmed.

Today, traffic is sent to sinkholes by injecting host routes into the routing table. This method can still be used inside an ALOC realm for intra-ALOC attacks. For attacks spanning over several ALOC realms new methods are needed; one example is described above. It is desirable that the RBR and hIPv4-enabled routers are capable of applying NAT rules and inserting service ID to selected packets in the forwarding plane.

15. Conclusions

This document offers a high-level overview of the hierarchical IPv4 framework that could be built in parallel with the current Internet by implementing extensions at several architectures. Implementation of the hIPv4 framework will not require a major service window break in the Internet or at the private networks of enterprises. Basically, the hIPv4 framework is an evolution of the current IPv4 framework.

The transition to hIPv4 might be attractive for enterprises since the hIPv4 framework does not create a catch-22 situation, e.g., when should an application used only inside the private network be ported from IPv4 to IPv6? Also, what is the business justification for porting the application to IPv6? Another matter is that when an

IPv4/v6 dual-stack solution is used it could impose operational expenditures, especially with rule sets at firewalls -- both in front of servers and at clients.

If an enterprise chooses to deploy hIPv4, however, the legacy applications do not need to be ported because hIPv4 is backwards compatible with the classical IPv4 framework. This means lower costs for the enterprise, and an additional bonus is the new stack's capabilities to better serve mobility use cases.

But the enterprise must take the decision soon and act promptly, because the IPv4 address depletion is a reality in the very near future. If the decision is delayed, IPv6 will arrive, and then, sooner or later, the legacy applications will need to be ported.

However, though this document has focused only on IPv4, a similar scheme can be deployed for IPv6 in the future, that is, creating a 64x64 bit locator space. But some benefits would have been lost at the time this document was written, such as:

- Backwards compatibility with the current Internet and therefore no smooth migration plan is gained.

- The locator header, including ALOC and ELOC prefixes, would have been larger, 160 bits versus 96 bits. And the identifier (EUI-64) would always have been present, which can be considered as pros or cons, depending upon one's view of the privacy issue, as discussed in [RFC4941] and in [Mobility_& _Privacy].

If an enterprise prefers hIPv4 (e.g., due to gaining additional IPv4 addresses and smooth migration capabilities), there is an unintentional side effect (from the enterprise's point of view) on the routing architecture of the Internet; multi-homing becomes multi-pathing, and an opportunity opens up for the service providers to create an Internet routing architecture that holds less prefixes and generates less BGP updates in DFZ than the current Internet.

The hIPv4 framework is providing a new hierarchy in the routing subsystem and is complementary work to multipath-enabled transport protocols (such as MPTCP and SCTP) and service-centric networking architectures (such as SCAFFOLD). End users and enterprises are not interested in routing issues in the Internet; instead, a holistic view should be applied on the three disciplines with a focus on new service opportunities and communicated to the end users and enterprises. Then perhaps the transition request to a new routing architecture will be accepted and carried out. However, more work is needed to accomplish a holistic framework of the three disciplines.

16. References

16.1. Normative References

[RFC1385] Wang, Z., "EIP: The Extended Internet Protocol", RFC 1385, November 1992.

[RFC1812] Baker, F., Ed., "Requirements for IP Version 4 Routers", RFC 1812, June 1995.

[RFC1918] Rekhter, Y., Moskowitz, B., Karrenberg, D., de Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, February 1996.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

[RFC4033] Arends, R., Austein, R., Larson, M., Massey, D., and S. Rose, "DNS Security Introduction and Requirements", RFC 4033, March 2005.

[RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.

[RFC4884] Bonica, R., Gan, D., Tappan, D., and C. Pignataro, "Extended ICMP to Support Multi-Part Messages", RFC 4884, April 2007.

[RFC5246] Dierks, T. and E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.2", RFC 5246, August 2008.

[RFC5944] Perkins, C., Ed., "IP Mobility Support for IPv4, Revised", RFC 5944, November 2010.

16.2. Informative References

[CES] Jen, D., Meisel, M., Yan, H. Massey, D., Wang, L., Zhang, B., Zhang, L., "Towards A New Internet Routing Architecture: Arguments for Separating Edges from Transit Core", 2008, http://conferences.sigcomm.org/hotnets/2008/papers/18.pdf.

[Dagstuhl] Arkko, J., Braun, M.B., Brim, S., Eggert, L., Vogt, C., Zhang, L., "Perspectives Workshop: Naming and Addressing in a Future Internet", 2009, http://www.dagstuhl.de/de/programm/kalender/semhp/?semnr=09102.

[ID/loc_Split] Thaler, D., "Why do we really want an ID/locator split anyway?", 2008, http://conferences.sigcomm.org/sigcomm/2008/workshops/mobiarch/slides/thaler.pdf.

[ILNP] Atkinson, R., "ILNP Concept of Operations", Work in Progress, February 2011.

[iVLB] Babaioff, M., Chuang, J., "On the Optimality and Interconnection of Valiant Load-Balancing Networks", 2007, http://people.ischool.berkeley.edu/~chuang/pubs/VLB-infocom07.pdf.

[LISP] Farinacci, D., Fuller, V., Meyer, D., and D. Lewis, "Locator/ID Separation Protocol", Work in Progress, June 2011.

[Mobility_&_Privacy] Brim, S., Linsner. M., McLaughlin, B., and K. Wierenga, "Mobility and Privacy", Work in Progress, March 2011.

[NBS] Ubillos, J., Xu, M., Ming, Z., and C. Vogt, "Name-Based Sockets Architecture", Work in Progress, September 2010.

[Nimrod] Chiappa, N., "A New IP Routing and Addressing Architecture", 1991, http://ana-3.lcs.mit.edu/~jnc/nimrod/overview.txt.

[Pathlet_Routing] Godfrey, P.G., Shenker, S., Stoica, I., "Pathlet Routing", 2008, http://conferences.sigcomm.org/hotnets/2008/papers/17.pdf.

[Porting_IPv4] DeLong, O., "Porting IPv4 applications to dual stack, with examples", 2010, http://www.apricot.net/apricot2010/program/tutorials/porting-ipv4-apps.html.

[RBridge] Perlman, R., "RBridges, Transparent Routing", 2004, http://www.ieee-infocom.org/2004/Papers/26_1.PDF.

[Revisiting_Route_Caching]
Kim, C., Caesar, M., Gerber, A., Rexford, J., "Revisiting Route Caching: The World Should Be Flat", 2009, http://www.springerlink.com/content/80w13260665v2013/.

[RFC3597] Gustafsson, A., "Handling of Unknown DNS Resource Record (RR) Types", RFC 3597, September 2003.

[RFC3618] Fenner, B., Ed., and D. Meyer, Ed., "Multicast Source Discovery Protocol (MSDP)", RFC 3618, October 2003.

[RFC4423] Moskowitz, R. and P. Nikander, "Host Identity Protocol (HIP) Architecture", RFC 4423, May 2006.

[RFC4941] Narten, T., Draves, R., and S. Krishnan, "Privacy Extensions for Stateless Address Autoconfiguration in IPv6", RFC 4941, September 2007.

[RFC4960] Stewart, R., Ed., "Stream Control Transmission Protocol", RFC 4960, September 2007.

[RFC4984] Meyer, D., Ed., Zhang, L., Ed., and K. Fall, Ed., "Report from the IAB Workshop on Routing and Addressing", RFC 4984, September 2007.

[RFC5395] Eastlake 3rd, D., "Domain Name System (DNS) IANA Considerations", RFC 5395, November 2008.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.

[RFC6115] Li, T., Ed., "Recommendation for a Routing Architecture", RFC 6115, February 2011.

[RFC6182] Ford, A., Raiciu, C., Handley, M., Barre, S., and J. Iyengar, "Architectural Guidelines for Multipath TCP Development", RFC 6182, March 2011.

[RFC6227] Li, T., Ed., "Design Goals for Scalable Internet Routing", RFC 6227, May 2011.

[RRG] RRG, "IRTF Routing Research Group Home Page", http://tools.ietf.org/group/irtf/trac/wiki/RoutingResearchGroup.

[SCAFFOLD] Freedman, M.J., Arye, M., Gopalan, P., Steven Y. Ko, S.Y., Nordstrom, E., Rexford, J., Shue, D. "Service-Centric Networking with SCAFFOLD", September 2010 http://www.cs.princeton.edu/research/techreps/TR-885-10.

[Split-DNS] BIND 9 Administrator Reference Manual, http://www.bind9.net/manual/bind/9.3.1/Bv9ARM.ch04.html#AEN767.

[tcpcrypt] Bittau, A., Hamburg, M., Handley, M., Mazi‘eres, D., Boneh, D., "The case for ubiquitous transport-level encryption", 2010, http://tcpcrypt.org/tcpcrypt.pdf.

[VLB] Zhang-Shen, R., McKeown, N., "Designing a Predictable Internet Backbone with Valiant Load-Balancing", 2004, http://conferences.sigcomm.org/hotnets/2004/HotNets-III%20Proceedings/zhang-shen.pdf.

17. Acknowledgments

The active participants at the Routing Research Group [RRG] mailing list are acknowledged. They have provided ideas, proposals, and discussions that have influenced the architecture of the hIPv4 framework. The following persons, in alphabetical order, have provided valuable review input: Aki Anttila, Mohamed Boucadair, Antti Jarvenpaa, Dae Young Kim, Mark Lewis, Wes Toman, and Robin Whittle.

Also, during the IRSG and IESG review process, Rajeev Koodli, Wesley Eddy, Jari Arkko, and Adrian Farrel provided valuable review input.

Lastly, a special thanks to Alfred Schwab from the Poughkeepsie ITSO for his editorial assistance.

Appendix A. Short-Term and Future IPv4 Address Allocation Policy

In this section, we study how the hIPv4 framework could influence the IPv4 address allocation policies to ensure that the new framework will enable some reusage of IPv4 address blocks. It is the Regional Internet Registries (RIRs) that shall define the final policies.

When the intermediate routing architecture (see Figure 1) is fully implemented, every ALOC realm could have a full IPv4 address space, except the GLB, from which to allocate ELOC blocks. There are some implications, however. In order for an enterprise to achieve site mobility, that is, to change service provider without changing its ELOC scheme, the enterprise should implement an autonomous system (AS) solution with an ALOC prefix at the attachment point to the service provider.

Larger enterprises have the resources to implement AS border routing. Most large enterprises have already implemented multi-homing solutions. Small and midsize enterprises (SMEs) may not have the resources to implement AS border routing, or the implementation introduces unnecessary costs for the SME. Also, if every enterprise needs to have an allocated ALOC prefix, this will have an impact on the RIB at the DFZ -- the RIB will be populated with a huge number of non-aggregatable ALOC prefixes.

It is clear that a compromise is needed. An SME site usually deploys a single uplink to the Internet and should be able to reserve a PI ELOC block from the RIR without being forced to create an ALOC realm, that is, implement an RBR solution and AS border routing. Since the PI ELOC block is no longer globally unique, an SME can only reserve the PI ELOC block for the region where it is active or has its attachment point to the Internet. The attachment point rarely changes to another country; therefore, it is sufficient that the PI ELOC block is regionally unique.

When the enterprise replaces its Internet service provider, it does not have to change its ELOC scheme -- only the local ALOC prefix at the endpoints is changed. The internal traffic at an enterprise does not make use of the ALOC prefix. The internal routing uses only the ELOC prefixes, and thus the internal routing and addressing architectures are preserved.

Mergers and acquisitions of enterprises can cause ELOC conflicts, because the PI ELOC block is hereafter only regionally unique. If an enterprise in region A acquires an enterprise in region B, there is a slight chance that both enterprises have overlapping ELOC prefixes.

If overlapping of ELOC prefixes occurs, the private unicast ALOC solution can be implemented to separate them -- if all affected endpoints support the hIPv4 framework.

Finally, residential users will receive only PA locators. When a residential user changes a service provider, she/he has to replace the locators. Since a PA ELOC block is no longer globally unique, every Internet service provider can use the PA ELOC blocks at their ALOC realms; the PA locators become kind of private locators for the service providers.

If the forwarding planes and all hosts that establish inter-ALOC realm sessions are upgraded to support the hIPv4 framework, that is, the long-term routing architecture (see Figure 2) is implemented, several interesting possibilities occur:

- The regional allocation policy for PI ELOC spaces can be removed, and the enterprise can make use of the whole IPv4 address space that is globally unique today. The ELOC space is hereafter only significant at a local ALOC realm.

- In case of mergers or acquisitions of enterprises, the private unicast ALOC solution can be used to separate overlapping ELOC spaces.

- The GLB space can be expanded to make use of all 32 bits (except for the blocks defined in RFC 1918) for anycast and unicast ALOC allocations; only ISPs are allowed to apply for GLB prefixes.

- The global anycast ALOC solution can be replaced with the global unicast ALOC solution since the ISP and enterprise no longer need to share ELOC routing information. Also, there is enough space in the GLB to reserve global unicast ALOC prefix(es) for every enterprise.

- Residential users will still use global anycast ALOC solutions, and if they change service providers, their locators need to be replaced.

The result is that a 32x32 bit locator space is achieved. When an enterprise replaces an ISP with another ISP, only the ALOC prefix(es) is replaced at endpoints and infrastructure nodes. Renumbering of ALOC prefixes can be automated by, for example, DHCP and extensions to IGP.

Appendix B. Multi-Homing becomes Multi-Pathing

When the transition to the intermediate routing architecture (see Figure 1) is fully completed, the RIB of an ISP that has created an ALOC realm will have the following entries:

- The PA ELOC blocks of directly attached customers (residential and enterprises)
- The PI ELOC blocks of directly attached customers (e.g., enterprises)
- The globally unique ALOC prefixes, received from other service providers

The ISP will not carry any PA or PI ELOC blocks from other service providers in its routing table. In order to do routing and forwarding of packets between ISPs, only ALOC information of other ISPs is needed.

Then, the question is how to keep the growth of ALOC reasonable? If the enterprise is using PI addresses, has an AS number, and is implementing BGP, why not apply for an ALOC prefix?

Classical multi-homing is causing the biggest impact on the growth of the size of the RIB in the DFZ -- so replacing a /20 IPv4 prefix with a /32 ALOC prefix will not reduce the size of the RIB in the DFZ.

Most likely, the only way to prevent this from happening is to impose a yearly cost for the allocation of an ALOC prefix, except if you are a service provider that is providing access and/or transit traffic for your customers. And it is reasonable to impose a cost for allocating an ALOC prefix for the non-service providers, because when an enterprise uses an ALOC prefix, it is reserving a FIB entry throughout the DFZ; the ALOC FIB entry needs to have power, space, hardware, and cooling on all the routers in the DFZ.

Implementing this kind of ALOC allocating policy will reduce the RIB size in the DFZ quite well, because multi-homing will no longer increase the RIB size of the DFZ. But this policy will have some impact on the resilience behavior because by compressing routing information we will lose visibility in the network. In today’s multi-homing solutions the network always knows where the remote endpoint resides. In case of a link or network failure, a backup path is calculated and an alternative path is found, and all routers in the DFZ are aware of the change in the topology. This functionality has off-loaded the workload of the endpoints; they only need to find the closest ingress router and the network will deliver

the packets to the egress router, regardless (almost) of what failures happen in the network. And with the growth of multi-homed prefixes, the routers in the DFZ have been forced to carry greater workloads, perhaps close to their limits -- the workload between the network and endpoints is not in balance. The conclusion is that the endpoints should take more responsibility for their sessions by offloading the workload in the network. How? Let us walk through an example.

A remote enterprise has been given an ELOC block 192.168.1.0/24, either via static routing or BGP announced to the upstream service providers. The upstream service providers provide the ALOC information for the enterprise, 10.1.1.1 and 10.2.2.2. A remote endpoint has been installed and given ELOC 192.168.1.1 -- the ELOC is a locator defining where the remote endpoint is attached to the remote network. The remote endpoint has been assigned ALOCs 10.1.1.1 and 10.2.2.2 -- an ALOC is a locator defining the attachment point of the remote network to the Internet.

The initiator (local endpoint) that has ELOC 172.16.1.1 and ALOC prefixes 10.3.3.3 and 10.4.4.4 has established a session by using ALOC 10.3.3.3 to the responder (remote endpoint) at ELOC 192.168.1.1 and ALOC 10.1.1.1. That is, both networks 192.168.10/24 and 172.16.1.0/24 are multi-homed. ALOCs are not available in the current IP stack’s API, but both ELOCs are seen as the local and remote IP addresses in the API, so the application will communicate between IP addresses 172.16.1.1 and 192.168.1.1. If ALOC prefixes are included, the session is established between 10.3.3.3:172.16.1.1 and 10.1.1.1:192.168.1.1.

Next, a network failure occurs and the link between the responder border router (BR-R1) and service provider that owns ALOC 10.1.1.1 goes down. The border router of the initiator (BR-I3) will not be aware of the situation, because only ALOC information is exchanged between service providers and ELOC information is compressed to stay within ALOC realms. But BR-R1 will notice the link failure; BR-R1 could rewrite the ALOC field in the locator header for this session from 10.1.1.1 to 10.2.2.2 and send the packets to the second service provider via BR-R2. The session between the initiator 10.3.3.3:172.16.1.1 and the responder 10.2.2.2:192.168.1.1 remains intact because the legacy 5-tuple at the IP stack API does not change. Only the ALOC prefix of the responder has changed and this information is not shown to the application. An assumption here is that the hIPv4 stack does accept changes of ALOC prefixes on the fly (more about this later).

If the network link between the BR-I3 and ISP providing ALOC 10.3.3.3 fails, BR-I3 could rewrite the ALOC prefix in the locator header and route the packets via BR-I4 and the session would stay up. If there is a failure somewhere in the network, the border routers might receive an ICMP destination unreachable message (if not blocked by some security functionality) and thus try to switch the session over to the other ISP by replacing the ALOC prefixes in the hIPv4 header. Or the endpoints might try themselves to switch to the other ALOCs after a certain time-out in the session. In all session transition cases the legacy 5-tuple remains intact.

If border routers or one of the endpoints changes the ALOC prefix without a negotiation with the remote endpoint, security issues arise. Can the endpoints trust the remote endpoint when ALOC prefixes are changed on the fly -- is it still the same remote endpoint or has the session been hijacked by a bogus endpoint? The obvious answer is that an identification mechanism is needed to ensure that after a change in the path or a change of the attachment point of the endpoint, the endpoints are still the same. An identifier needs to be exchanged during the transition of the session.

Identifier/locator split schemes have been discussed on the [RRG] mailing list, for example, multipath-enabled transport protocols and identifier database schemes. Both types of identifiers can be used to protect the session from being hijacked. A session identifier will provide a low-level security mechanism, offering some protection against hijacking of the session and also provide mobility. SCTP uses the verification tag to identify the association; MPTCP incorporates a token functionality for the same purpose -- both can be considered to fulfill the characteristics of a session identifier. [tcpcrypt] can be used to further mitigate session hijacking. If the application requires full protection against man-in-the-middle attacks, TLS should be applied for the session. Both transport protocols are also multipath-capable. Implementing multipath-capable transport protocols in a multi-homed environment will provide new capabilities, such as:

- Concurrent and separate exit/entry paths via different attachment points at multi-homed sites.

- True dynamic load-balancing, in which the endpoints do not participate in any routing protocols or do not update rendezvous solutions due to network link or node failures.

- Only a single Network Interface Card (NIC) on the endpoints is required.

- In case of a border router or ISP failure, the multipath transport protocol will provide resilience.

By adding more intelligence at the endpoints, such as multipath-enabled transport protocols, the workload of the network is offloaded and can take less responsibility for providing visibility of destination prefixes on the Internet; for example, prefix compression in the DFZ can be applied and only the attachment points of a local network need to be announced in the DFZ. And the IP address space no longer needs to be globally unique; it is sufficient that only a part is globally unique, with the rest being only regionally unique (in the long-term routing architecture, locally unique) as discussed in Appendix A.

The outcome is that the current multi-homing solution can migrate towards a multi-pathing environment that will have the following characteristics:

- An AS number is not mandatory for enterprises.
- BGP is not mandatory at the enterprise’s border routers; static routing with Bidirectional Forwarding Detection (BFD) [RFC5880] is an option.
- Allocation of global ALOC prefixes for the enterprise should not be allowed; instead, upstream ISPs provide the global ALOC prefixes for the enterprise.
- MPTCP provides dynamic load-balancing without using routing protocols; several paths can be used simultaneously and thus resilience is achieved.
- Provides low growth of RIB entries at the DFZ.
- When static routing is used between the enterprise and the ISP:
  - The RIB size at the enterprise’s border routers does not depend upon the size of the RIB in the DFZ or in adjacent ISPs.
  - The enterprise’s border router cannot cause BGP churn in the DFZ or in the adjacent ISPs’ RIB.
- When dynamic routing is used between the enterprise and the ISP:
  - The RIB size at the enterprise’s border routers depends upon the size of the RIB in the DFZ and adjacent ISPs.

- The enterprise’s border router can cause BGP churn for the adjacent ISPs, but not in the DFZ.

- The cost of the border router should be less than in today’s multi-homing solution.

Appendix C. Incentives and Transition Arguments

The media has announced the meltdown of the Internet and the depletion of IPv4 addresses several times, but the potential chaos has been postponed and the general public has lost interest in these announcements. Perhaps it could be worthwhile to find other valuable arguments that the general public could be interested in, such as:

- Not all endpoints need to be upgraded, only those that are directly attached to the Internet, such as portable laptops, smart mobile phones, proxies, and DMZ/frontend endpoints. But the most critical endpoints, the backend endpoints where enterprises keep their most critical business applications, do not need to be upgraded. These endpoints should not be reached at all from the Internet, only from the private network. And this functionality can be achieved with the hIPv4 framework, since it is backwards compatible with the current IPv4 stack. Therefore, investments in legacy applications used inside an ALOC realm are preserved.

- Mobility - it is estimated that the demand for applications that perform well over the wireless access network will increase. Introduction of MPTCP and identifier/locator split schemes opens up new possibilities to create new solutions and applications that are optimized for mobility. The hIPv4 framework requires an upgrade of the endpoint’s stack; if possible, the hIPv4 stack should also contain MPTCP and identifier/locator split scheme features. Applications designed for mobility could bring competitive benefits.

- The intermediate routers in the network do not need to be upgraded immediately; the current forwarding plane can still be used. The benefit is that the current network equipment can be preserved at the service providers, enterprises, and residences (except middleboxes). This means that the carbon footprint is a lot lower compared to other solutions. Many enterprises do have green programs and many residential users are concerned with the global warming issue.

- The migration from IPv4 to IPv6 (currently defined architecture) will increase the RIB and FIB throughout DFZ. Whether it will require a new upgrade of the forwarding plane as discussed in [RFC4984] is unclear. Most likely an upgrade is needed. The

outcome of deploying IPv4 and IPv6 concurrently is that the routers need to have larger memories for the RIB and FIB -- every globally unique prefix is installed in the routers that are participating in the DFZ. Since the enterprise reserves one or several RIB/FIB entries on every router in the DFZ, it is increasing the power consumption of the Internet, thus increasing the carbon footprint. And many enterprises are committed to green programs. If hIPv4 is deployed, the power consumption of the Internet will not grow as much as in an IPv4 to IPv6 transition scenario.

- o Another issue: if the migration from IPv4 to IPv6 (currently defined architecture) occurs, the routers in the DFZ most likely need to be upgraded to more expensive routers, as discussed in [RFC4984]. In the wealthy part of the world, where a large penetration of Internet users is already present, the service providers can pass the costs of the upgrade along to their subscribers more easily. With a "wealthy/high penetration" ratio the cost will not grow so much that the subscribers would abandon the Internet. But in the less wealthy part of the world, where there is usually a lower penetration of subscribers, the cost of the upgrade cannot be accepted so easily -- a "less wealthy/low penetration" ratio could impose a dramatic increase of the cost that needs to be passed along to the subscribers. And thus fewer subscribers could afford to get connected to the Internet. For the global enterprises and the enterprises in the less wealthy part of the world, this scenario could mean less potential customers and there could be situations when the nomads of the enterprises can’t get connected to the Internet. This is also not fair; every human being should have a fair chance to be able to enjoy the Internet experience -- and the wealthy part of the world should take this right into consideration. Many enterprises are committed to Corporate Social Responsibility programs.

Not only technical and economical arguments can be found. Other arguments that the general public is interested in and concerned about can be found, for example, that the Internet becomes greener and more affordable for everyone, in contrast with the current forecast of the evolution of the Internet.

Appendix D. Integration with CES Architectures

Because the hIPv4 framework requires changes to the endpoint’s stack, it will take some time before the migration of the current IPv4 architecture to the intermediate hIPv4 routing architecture is fully completed. If a hIPv4 proxy solution could be used in front of

classical IPv4 endpoints, the threshold for early adopters to start to migrate towards the hIPv4 framework would be less questionable and the migration phase would also most likely be much shorter.

Therefore, it should be investigated whether the hIPv4 framework can be integrated with Core-Edge Separation [CES] architectures. In CES architectures the endpoints do not need to be modified. The design goal of a CES solution is to minimize the PI-address entries in the DFZ and to preserve the current stack at the endpoints. But a CES solution requires a new mapping system and also introduces a caching mechanism in the map-and-encapsulate network nodes. Much debate about scalability of a mapping system and the caching mechanism has been going on at the [RRG] list. At the present time it is unclear how well both solutions will scale; research work on both topics is still in progress.

Since the CES architectures divide the address spaces into two new categories, one that is installed in the RIB of the DFZ and one that is installed in the local networks, there are to some degree similarities between CES architectures and the hIPv4 framework. Actually, the invention of the IP and locator header swap functionality was inspired by [LISP].

In order to describe how these two architectures might be integrated, some terminology definitions are needed:

CES-node:

A network node installed in front of a local network that must have the following characteristics:

- Map-and-encapsulate ingress functionality
- Map-and-encapsulate egress functionality
- Incorporate the hIPv4 stack
- Routing functionality, [RFC1812]
- Being able to apply policy-based routing on the ALOC field in the locator header

The CES-node does not include the MPTCP extension because it would most likely put too much of a burden on the CES-node to signal and maintain MPTCP subflows for the cached hIPv4 entries.

Consumer site:

A site that is not publishing any services towards the Internet, that is, there are no entries in DNS for this site. It is used by local endpoints to establish outbound connectivity -- endpoints are initiating sessions from the site towards content sites. Usually such sites are found at small enterprises and residences. PA-addresses are usually assigned to them.

Content site:

A site that is publishing services towards the Internet, and which usually does have DNS entries. Such a site is used by local endpoints to establish both inbound and outbound connectivity. Large enterprises use PI-addresses, while midsize/small enterprises use either PI- or PA-address space.

The CES architectures aim to reduce the PI-address entries in the DFZ. Therefore, map-and-encapsulate egress functionality will be installed in front of the content sites. It is likely that the node containing map-and-encapsulate egress functionality will also contain map-and-encapsulate ingress functionality; it is also a router, so the node just needs to support the hIPv4 stack and be able to apply policy-based routing using the ALOC field of the locator header to become a CES-node.

It is possible that the large content providers (LCPs) are not willing to install map-and-encapsulate functionality in front of their sites. If the caching mechanism is not fully reliable or if the mapping lookup delay does have an impact on their clients' user experience, then most likely the LCPs will not adopt the CES architecture.

In order to convince a LCP to adopt the CES architecture, it should provide a mechanism to mitigate the caching and mapping lookup delay risks. One method is to push the CES architectures to the edge -- the closer to the edge you add new functionality, the better it will scale. That is, if the endpoint stack is upgraded, the caching mechanism is maintained by the endpoint itself. The mapping mechanism can be removed if the CES architecture's addressing scheme is replaced with the addressing scheme of hIPv4 when the CES solution is integrated at the endpoints. With this approach, the LCPs might install a CES-node in front of their sites. Also, some endpoints at the content site might be upgraded with the hIPv4 stack.

If the LCP faces issues with the caching or mapping mechanisms, the provider can ask its clients to upgrade their endpoint's stack to ensure a proper service level. At the same time, the LCP promotes the migration from the current routing architecture to a new routing architecture, not for the sake of the routing architecture but instead to ensure a proper service level -- you can say that a business model will promote the migration of a new routing architecture.

The hIPv4 framework proposes that the IPv4 addresses (ELOC) should no longer be globally unique; once the transition is completed, a more regional allocation can be deployed. But this is only possible once all endpoints (that are establishing sessions to other ALOC realms) have migrated to support the hIPv4 framework. Here the CES architecture can speed up the re-usage of IPv4 addresses; that is, once an IPv4 address block has become an ELOC block it can be re-used in the other RIR regions, without the requirement that all endpoints in the Internet must first be upgraded.

As stated earlier, the CES architecture aims to remove PI-addresses from the DFZ, making the content sites more or less the primary target for the roll-out of a CES solution. At large content sites a CES-node most likely will be installed. To upgrade all endpoints (that are providing services towards the Internet) at a large content site will take time, and it might be that the endpoints at the content site are upgraded only within their normal lifecycle process. But if the size of the content site is small, the administrator either installs a CES-node or upgrades the endpoint's stack -- a decision influenced by availability, reliability, and economic feasibility.

Once the content sites have been upgraded, the PI-address entries have been removed from the DFZ. Most likely also some endpoints at the consumer sites have been upgraded to support the hIPv4 stack -- especially if there have been issues with the caches or mapping delays that have influenced the service levels at the LCPs. Then, the issue is how to keep track of the upgrade of the content sites -- have they been migrated or not? If the content sites or content endpoints have been migrated, the DNS records should have either a CES-node entry or ALOC entry for each A-record. When the penetration of CES solutions at content sites (followed up by CES-node/ALOC records in DNS) is high enough, the ISP can start to promote the hIPv4 stack upgrade at the consumer sites.

Once a PA-address block has been migrated it can be released from global allocation to a regional allocation. Why would an ISP then push its customers to deploy hIPv4 stacks? Because of the business model -- it will be more expensive to stay in the current

architecture. The depletion of IPv4 addresses will either cause more NAT at the service provider's network (operational expenditures will increase because the network will become more complex) or the ISP should force its customers to migrate to IPv6. But the ISP could lose customers to other ISPs that are offering IPv4 services.

When PA-addresses have been migrated to the hIPv4 framework, the ISP will have a more independent routing domain (ALOC realm) with only ALOC prefixes from other ISPs and ELOC prefixes from directly attached customers. BGP churn from other ISPs is no longer received, the amount of alternative paths is reduced, and the ISP can better control the growth of the RIB at their ALOC realm. The operational and capital expenditures should be lower than in the current routing architecture.

To summarize, the content providers might find the CES+hIPv4 solution attractive. It will remove the forthcoming IPv4 address depletion constraints without forcing the consumers to switch to IPv6, and thus the content providers can continue to grow (reach more consumers).

The ISP might also find this solution attractive because it should reduce the capital and operational expenditures in the long term. Both the content providers and the ISPs are providing the foundation of the Internet. If both adopt this architecture, the consumers have to adopt. Both providers might find business models to "guide" the consumers towards the new routing architecture.

Then, how will this affect the consumer and content sites? Residential users will need to upgrade their endpoints. But it doesn't really matter which IP version they use. It is the availability and affordability of the Internet that matters most.

Enterprises will be affected a little bit more. The edge devices at the enterprises' local networks need to be upgraded -- edge nodes such as AS border routers, middleboxes, DNS, DHCP, and public nodes -- but by installing a CES-node in front of them, the upgrade process is postponed and the legacy nodes can be upgraded during their normal lifecycle process. The internal infrastructure is preserved, internal applications can still use IPv4, and all investment in IPv4 skills is preserved.

Walkthrough of use cases:

1. A legacy endpoint at a content site establishes a session to a content site with a hIPv4 upgraded endpoint.

When the legacy endpoint resolves the DNS entry for the remote endpoint (a hIPv4 upgraded endpoint), it receives an ALOC record in the DNS response. The legacy endpoint ignores the ALOC record. Only the A-record is used to establish the session. Next, the legacy endpoint initializes the session and a packet is sent towards the map-and-encapsulate ingress node, which needs to do a lookup at the CES mapping system (the assumption here is that no cache entry exists for the remote endpoint). The mapping system returns either a CES-node prefix or an ALOC prefix for the lookup -- since the requested remote endpoint has been upgraded, the mapping system returns an ALOC prefix.

The CES-node will not use the CES encapsulation scheme for this session. Instead, the hIPv4 header scheme will be used and a /32 entry will be created in the cache. A /32 entry must be created; it is possible that not all endpoints at the remote site are upgraded to support the hIPv4 framework. The /32 cache entry can be replaced with a shorter prefix in the cache if all endpoints are upgraded at the remote site. To indicate this situation, a subfield should be added for the ALOC record in the mapping system.

The CES-node must execute the following steps for the egress packets:

a. Verify IP and transport header checksums.

b. Create the locator header and copy the value in the destination address field of the IP header to the ELOC field of the locator header.

c. Replace the destination address in the IP header with the ALOC prefix given in the cache.

d. Insert the local CES-node prefix in the ALOC field of the locator header.

e. Copy the transport protocol value of the IP header to the protocol field of the locator header and set the hIPv4 protocol value in the protocol field of the IP header.

f. Set the desired parameters in the A-, I-, S-, VLB-, and L-fields of the locator header.

g. Set the FI-bits of the locator header to 00.

h. Decrease the TTL value by one.

i. Calculate IP, locator, and transport protocol header checksums. Transport protocol header calculations do not include the locator header fields. When completed, the packet is transmitted.

j. Because the size of the packet might exceed MTU due to the insertion of the locator header, and if MTU is exceeded, the CES-node should inform the source endpoint of the situation with an ICMP message, and the CES-node should apply fragmentation of the hIPv4 packet.

2. A hIPv4-upgraded endpoint at a consumer/content site establishes a session to a content site with a CES-node in front of a legacy endpoint.

   The hIPv4 upgraded endpoint receives, in the DNS response, either an ALOC record or a CES-node record for the resolved destination. From the requesting hIPv4 endpoint’s point of view, it really doesn’t matter if the new record prefix is used to locate RBR-nodes or CES-nodes in the Internet -- the CES-node will act as a hIPv4 proxy in front of the remote legacy endpoint. Thus the hIPv4 endpoint assembles a hIPv4 packet to initialize the session, and when the packet arrives at the CES-node it must execute the following:

   a. Verify that the received packet uses the hIPv4 protocol value in the protocol field of the IP header.

   b. Verify IP, locator, and transport protocol header checksums. Transport protocol header verification does not include the locator header fields.

   c. Replace the protocol field value of the IP header with the protocol field value of the locator header.

   d. Replace the destination address in the IP header with the ELOC prefix of the locator header.

   e. Remove the locator header.

   f. Create a cache entry (unless an entry already exists) for returning packets. A /32 entry is required. To optimize the usage of cache entries, the CES-node might ask the CES mapping node whether all endpoints at the remote site are upgraded or not. If upgraded, a shorter prefix can be used in the cache.

   g. Decrease the TTL value by one.

h. Calculate IP and transport protocol header checksums.

i. Forward the packet according to the destination address of the IP header.

3. A hIPv4-enabled endpoint with a regionally unique ELOC at a consumer site establishes a session to a consumer site with a legacy endpoint.

   In this use case, the sessions will fail unless some mechanism is invented and implemented at the ISPs' map-and-encapsulate nodes. The sessions will work inside an ALOC realm since the classical IPv4 framework is still valid. Sessions between ALOC realms will fail. Some applications establish sessions between consumer sites. The most common are gaming and peer-to-peer applications. These communities have historically been in the forefront of adopting new technologies. It is expected that they either develop workarounds to solve this issue or simply ask their members to upgrade their stacks.

4. A legacy endpoint at a consumer/content site establishes a session to a content site with a CES-node in front of a legacy endpoint.

   Assumed to be described in CES architecture documents.

5. A hIPv4-enabled endpoint at a consumer/content site establishes a session to a content site with a hIPv4-enabled endpoint.

   See Section 5.2.

Author's Address

Patrick Frejborg
EMail: pfrejborg@gmail.com