Abstract

This draft describes a framework how the current IPv4 address structure can be extended towards a similar hierarchical numbering structure as used in the Public Switched Telephone Network and bring a new level of hierarchy to the routing architecture of Internet. The hierarchical IPv4 framework is backwards compatible with the current IPv4 framework; it will also discuss a method to decouple the location and identifier functions, future applications can make use of the separation. The framework requires extensions to the existing Domain Name System architecture, the existing IPv4 stack of the endpoints and to routers in the Internet. The framework can be implemented incrementally to the endpoints, databases, and routers.

Table of Contents

1. Requirements Notation.................................3
2. Introduction..............................................3
3. An overview of the hIPv4 framework...................5
4. Definitions of terms.......................................7
5. Mandatory extensions to current architectures (unicast)....9
6. The header of a hIPv4 packet............................10
7. Life of a hIPv4 session..................................13
8. Overlapping Source and Destination ELOC prefixes/ports...16
9. Traceroute considerations................................17
10. Multicast considerations................................17
11. Traffic engineering considerations.....................18
12. Large encapsulated packets.............................20
13. Mobility considerations..................................20
14. Affected Applications and Implications..................21
15. The Future Role of the LSR............................22
16. Transition considerations................................23
17. Security Considerations..................................24
18. IANA Considerations....................................24
19. Conclusion................................................24
20. References................................................25
   20.1. References............................................25
   20.2. Informative References............................26
21. Acknowledgements.......................................27
Appendix A. Future IPv4 address allocation policies..........28
Appendix B. Multi-homing becomes multi-pathing...............30
Appendix C. Mobile site crossing a RIR border..............35
Appendix D. Transition Arguments..........................37
1. 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].

2. Introduction

The hierarchical IPv4 (hIPv4) framework has been developed to address the issues discussed in the [IAB report] from the Routing and Addressing Workshop that was held by the Internet Architecture Board (IAB) on October 18-19, 2006, in Amsterdam, Netherlands.

The current addressing (IPv4) and the future addressing (IPv6) schemes of Internet are single dimensional by their nature. This limitation, i.e. the single level addressing scheme, has created some roadblocks for further growth of Internet. If we compare Internet’s current addressing schemes to other global addressing or location schemes we can notice that the other schemes use several levels in their structures. E.g. the postal system uses street address, city and country to locate a destination. Also to locate a geographical site we are using longitude and latitude in the cartography system.

The other global network, the Public Switched Telephone Network (PSTN), have been build upon a three level numbering scheme that have 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 the IAB report can be solved. 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 as to the country code in PSTN, i.e. the ALOC prefix locates an area in Internet, the area is called an ALOC realm. The Endpoint Locator (ELOC) prefixes in hIPv4 can be compared to the subscriber numbers in PSTN, i.e. the ELOC is regionally unique at the attached ALOC realm - the ELOC can also be attached simultaneously to several ALOCs realms (multi-homing).

By inserting the ALOC and ELOC elements as a shim header (similar as in [MPLS] and [RBridge] architectures) between the IP protocol 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] - outcome is that the current forwarding plane do not need to be upgraded though some minor changes are needed in the control plane (e.g. ICMP extensions).
Another important influence source are 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, an identifier element needs to be added to the hIPv4 framework to provide a path for future applications to be able to remove the current dependency of the underlying network layer addressing scheme (local and remote IP address tuples). Multipath transport protocols, such as [SCTP] and the currently under development Multipath TCP [MPTCP], are the most interesting candidates to enable an identifier functionality for the hIPv4 framework. Especially MPTCP is interesting from hIPv4 point of view – one of the main goals of MPTCP is to provide backwards compatibility with current implementations, hIPv4 share the same goal. MPTCP itself do not provide a host identifier solution as e.g. [HIP] do, 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 layer. Thus the usage of a token/session identifier will provide a mechanism to improve mobility, both in site and endpoint mobility scenarios.

Some of the design goals of this proposal include:

1. The hierarchical IPv4 framework must be backwards compatible with the current IPv4 framework.

2. Minimize introduction of totally new protocols or signaling architectures, instead use well-proven protocols and insert extension to protocols where needed.

3. Create a hierarchical IPv4 addressing structure which enables a more regional allocation of IPv4 address blocks and therefore the routing table size in the "default-free zone" (DFZ) will be reduced.

4. Remove the single IPv4 address space constraint; reuse IPv4 address blocks by a hierarchy.

5. Improve site mobility, i.e. a site wishes to changes its attachment point to Internet without changing its IPv4 address block.
6. Make use of the current forwarding plane (IPv4); introduce a new forwarding plane for only a few routers in an Autonomous System or group of Autonomous Systems.

7. Reduce the amount of Network Address Translation (NAT) need for IPv4-to-IPv4 traffic.

8. Provide a smooth transition path to the hierarchical IPv4 framework.

3. An overview of the hIPv4 framework

In this section we will discuss the roles of the new elements, introduced by the hIPv4 framework, and their dependencies.

As mentioned in the introduction section the role of an Area LOCator (ALOC) prefix is similar as to a country code in PSTN. I.e. the ALOC prefix provides a location functionality of an area within an Autonomous System (AS), or an area spanning over a group of AS, in Internet. An AS can have several ALOC prefixes assigned, e.g. due to traffic engineering requirements. The ALOC prefix will be used for routing and forwarding purposes in Internet, thus 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 AS) are assigned an ALOC prefix the area has the potential to become an ALOC realm. In order to establish an ALOC realm more elements, further than the ALOC prefix, are needed. One or multiple Locator Swap Routers (LSR) must be attached to the ALOC realm. A LSR element is a node capable of swapping the values of the IPv4 header and the new shim header, called the locator header. The swap mechanism of the headers is described in detail in section 7, step 4. Today’s routers do not support the LSR 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 LSR might be a computer with two interface attached to a router, the first interface configured with the prefix of the ALOC and the second interface with any IPv4 prefix. The LSR do not make us of dynamic routing protocols, neither a forwarding information base (FIB) nor a cache is needed - the LSR is producing a service, i.e. swapping headers. The swap mechanism is applied on per packet basis and the information needed to carry out the swap is included in the locator header of the hIPv4 packet. Thus a computer with enough computing and I/O resources is sufficient to take the role as a LSR. Later on, the LSR functionality might be integrated into the forwarding plane of a router. One LSR
can not handle all the incoming traffic designated for an ALOC realm; it would also create a potential single point of failure in the network. Therefore, several LSRs might be installed in the ALOC realm and the LSRs shall use the ALOC prefix as their locator and the routers are announcing the ALOC prefix as an Anycast address within the local ALOC realm. Also, the ALOC prefix is advertised throughout the DFZ by BGP mechanisms. The placement of the LSRs in the network will influence on the ingress traffic to the ALOC realm, the LSR is providing "nearest routing" functionality.

Since the forwarding paradigm of multicast packets is quite different from forwarding unicast packets the multicast functionality will have an impact on the LSR. Also, the multicast LSR (mLSR) functionality is not available on today’s routers, an external device is needed, and later on the functionality might be integrated to the routers. The mLSR shall take the role of an Anycast RP with MSDP and PIM capabilities, but to forward packets a FIB is not required. As with the LSR, the multicast hIPv4 packets are carrying all needed information in their headers in order to apply the swap, for details see section 10.

The ALOC realm is not yet fully constructed, we can now locate the ALOC realm in Internet but to locate the endpoints attached to the ALOC realm a new element is needed, i.e. the Endpoint Locator (ELOC). As mentioned in the introduction section the ELOC prefixes can be considered similar as to the subscriber numbers in PSTN. Actually, the ELOC is not a new element; the ELOC is a redefinition of the current IPv4 address configured at an endpoint. The redefinition is applied because when the hIPv4 framework is fully implemented the global uniqueness of the IPv4 addresses are no longer valid and 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 domain, the ELOC prefix is not used in the intermediate ALOC domains. When a client is establishing a session to a server residing outside the local ALOC realm the destination IP address field in the IPv4 header of an outgoing packet is no longer the remote ELOC prefix - instead the remote ALOC prefix is installed in the destination IP address field of the IPv4 header. Because the destination IP address is an ALOC prefix, the intermediate ALOC realms do not need to carry the ELOC prefixes of other ALOC realms in their RIB - it is sufficient for the intermediate ALOC realms to carry only the ALOC prefixes. Outcome is that the RIB and FIB 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 can't 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 will be forwarded to the nearest LSR, since the destination IPv4 address is the remote ALOC prefix. When the LSR has swapped the hIPv4 header, the destination IP address field in the IPv4 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 13.

Next, how can we locate the remote ELOC (endpoint) and remote ALOC realm in Internet, also how to assemble the header of the hIPv4 packet? Another matter is that the addressing structure is no longer single dimensional; instead a second level has been added on top of the old one. It is obvious that 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 complicated, the proxy needs to listen to DNS messages and a cache solution does have 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, but backwards compatibility must be preserved. Most of the application will not be aware of the extensions - other applications, such as Mobile IP, SIP, IPsec AH and so on (see section 14) will suffer and can not be used outside their ALOC realm when the hIPv4 framework is fully implemented - unless the applications are upgraded. The reason is that these applications are depending upon the underlying network addressing structure to e.g. identify an endpoint.

4. Definitions of terms

Regional Internet Registry (RIR):

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 locator is a name for a point of attachment within the topology at a given layer. Objects that change their point of attachment(s) will need to change their associated locator(s). In the hIPv4 framework two types of locators have been defined, Area Locator (ALOC) and Endpoint Locator (ELOC).

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 Internet. The ALOC is assigned by a RIR to a service provider or a multi-homed enterprise. 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 a RIR to a service provider or to an enterprise. The ELOC block is only unique in a geographical region or globally unique in a business area defined by the RIRs. The final policy of uniqueness shall be defined by the RIRs.

ALOC realm:

An area in the Internet with at least one attached Locator Swap Router (LSR), also an ALOC must be assigned to the ALOC realm. The 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.

Locator Swap Router (LSR):

A router or node which is capable to process the hIPv4 header; once the header is processed the LSR will forward the packet upon the IPv4 destination address. The LSR must have the ALOC assigned as its locator.

Locator Header:

A 4, 8, 12 or 16 byte field, inserted between the IP and transport header.
Identifier:

An identifier is 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.

Session:

Is a semi-permanent interactive information exchange between communicating devices that is established at a certain time and torn down at a later time.

Provider Independent Address Space (PI addresses):

An IPv4 address block which is assigned by a Regional Internet Registries directly to an end-user organization.

Provider Aggregatable Address Space (PA addresses):

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

5. Mandatory extensions to current architectures (unicast)

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

1. The DNS architecture must support a new extension, i.e. an A type Resource Record should be able to carry an ALOC prefix.

2. The hIPv4 capable endpoint shall have information about the local ALOC value; the local ALOC value can be configured manually or provided via a new DHCP option.

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. A multi-homed enterprise might allocate an ALOC prefix from the GLB.

4. ALOC prefixes are announced via current BGP protocol to adjacent service providers and multi-homed enterprises, the ALOC prefixes are installed in the RIB of the DFZ. When the hIPv4 framework is fully implemented only ALOC prefixes are announced between the service providers and multi-homed enterprises.
5. A hIPv4 capable ALOC realm must have one or several LSRs attached to its realm. The ALOC prefix is configured as an Anycast IP address on the LSR. 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 IP address and the remote ALOC prefix as the destination IP address in the IPv4 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.

6. The header of a hIPv4 packet

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Version|  IHL  |Type of Service|          Total Length         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Identification|Flags|   Fragment Offset |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Time to Live |    Protocol   |         Header Checksum       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Source Address                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Destination Address                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Options                    |    Padding    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
[A|P|S|VLB|L| R |    Protocol   |         LH Checksum           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Area Locator                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Endpoint Locator                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  Private Locator Referral                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Version: 4 bits

The Version field is identical to that of RFC 791.
IHL: 4 bits

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 field is identical to that of RFC 791.
ALOC Realm Bit, A-bit: 1 bit

When the source and destination endpoints 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. When the A-bit is set to 0 the source and destination endpoints reside within the same ALOC realm, the Area and Endpoint Locator shall not be used in the locator header.

Private Bit, P-bit: 1 bit

The P-bit is set to 1 if the endpoint is using a private IP address [RFC1918] and has published either the private IP address to the public via DNS or discreetly to partners in order to create a bidirectional session model for NAT. When P-bit is set to 1, the Private Locator Referral field must be used in the locator header.

Swap Bit, S-bit: 1 bit

The initiating endpoint sets the S-bit to 0 of the hIPv4 packet. A LSR will set this bit to 1 when it is swapping the IP source and destination addresses of the IP header with the Area and Endpoint Locator of the locator header.

Valiant Load-Balancing, VLB-bits: 2 bits

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 following:

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

01: First subflow, 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

The initiating endpoint 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
must set the value to 1.

Reserved, R-bits: 2 bits

Reserved, must be zero

Protocol: 8 bits

The Protocol field is identical to that of RFC 791

Locator Header Checksum: 16 bits

A checksum is calculated on the locator header only. The checksum
is computed at the initiating endpoint, recomputed at the LSR and
verified at the destination endpoint. The checksum algorithm is
identical to that of RFC 791.

Area Locator (optional): 32 bits

An IPv4 address, the ALOC is assigned by a RIR to a service
provider or a multi-homed enterprise with an Autonomous System (AS)
number. The ALOC is globally unique because it is allocated from
the GLB.

Endpoint Locator (optional): 32 bits

An IPv4 address, the ELOC block is assigned by a RIR to a service
provider or to an enterprise. The ELOC block is only unique in a
geographical region or globally unique in a business area defined
by the RIRs. The final policy of uniqueness shall be defined by the
RIRs.

Private Locator Referral (optional): 32 bits

A private IPv4 address [RFC1918] in order to create a bidirectional
session model for NAT, i.e. the initiating endpoint can set the PLR
value in its outgoing packets in order to traverse a middlebox that
is installed in front of the destination endpoint that is using a
private IPv4 address as its locator. The middlebox can also contain
a local private-public mapping scheme and thus the PLR field do not
need to be filled with a private IPv4 address.

7. Life of a hIPv4 session

This section provides an example of a hIPv4 session between two hIPv4
endpoints; a client and a server residing in different ALOC realms.
When the hIPv4 stack is assembling the packet for transport the hIPv4 stack shall decide if a legacy IPv4 or a hIPv4 header is used upon the ALOC information received by a DNS reply. If the client’s (local) ALOC value equals the server’s (remote) ALOC value there is no need to use the hIPv4 header for routing purposes, because both the client and server reside in the local ALOC realm. The packet is routed upon the IPv4 header since the packet will not exit the local ALOC realm. When the local ALOC prefix doesn’t 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 or to create a bidirectional session model for NAT. The originating endpoint can add the locator header to the packet and by setting the VLB bits to 01 indicating to the remote endpoint 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, thus the size of the locator header will be 4 bytes - or 8 bytes if the Private Locator Referral is also used.

How a hIPv4 session is established follows:

1. The client queries the DNS server; the hIPv4 stack notice that the local and remote ALOC doesn’t match and therefore must use the hIPv4 header for the session. The hIPv4 stack of the client must assemble the packet in the following way:
   a. set local IP address from API in the source IP address field
   b. set remote IP address from API in the ELOC field
   c. set local ALOC prefix in the ALOC field
   d. set remote ALOC prefix in the destination IP address field
   e. set the hIPv4 protocol number in the protocol field of the IP header and set the transport protocol number in the protocol field of the locator header
   f. set the desired parameters in the A-, P-, S-, VLB-, L-, and R-fields of the locator header
g. Apply checksum IP-, locator- and transport header calculations, transport header calculation do not include the locator header fields. When completed the packet is transmitted.

2. The hIPv4 packet is routed throughout Internet upon the destination IP address of the IPv4 header.

3. The hIPv4 packet will reach the closest LSR of the remote ALOC realm. When the LSR notice that the packet matches the given local ALOC value the LSR must:
   a. verify the received packet that it uses the hIPv4 protocol number in the protocol field of the IP header
   b. verify IP-, locator- and transport header checksums, transport header verification do not include the locator header fields
   c. replace the source address in the IPv4 header with the ALOC value of the locator header
   d. replace the destination address in the IPv4 header with the ELOC value of the locator header
   e. replace the ALOC value in the locator header with the destination IP address of the IPv4 header
   f. replace the ELOC value in the locator header with the source IP address of the IPv4 header
   g. set the S-field to 1
   h. decrease TTL value with one
   i. calculate IP-, locator- and transport header checksums, transport header calculation do not include the locator header fields
   j. forward the packet upon the destination IP address of the IPv4 header

4. The swapped hIPv4 packet is now routed inside the remote ALOC realm upon the new destination IP address of the IPv4 header to the final destination.

5. The server receives the hIPv4 packet
a. The hIPv4 stack must verify the received packet that it uses the hIPv4 protocol number in the protocol field of the IP header

b. Verify IP-, locator- and transport header checksums, transport header verification do not include the locator header fields

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

   a. present source IP address as the remote ALOC
   b. present destination IP address as the local IP address
   c. verify the received ALOC as the local ALOC
   d. present ELOC as the remote IP address

7. The server’s application will respond to the client and the returning packet will take almost the same steps, which are steps 1 to 6, as when the client started the session. In step 1 the server doesn’t need to do a DNS lookup since all information is provided by the packet.

8. Overlapping Source and Destination 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 source and destination ELOC prefixes. But the session is still unique because the two processes communicating over the transport protocol form a logical session which is uniquely identifiable by the five tuples involved, i.e. by the combination of <protocol, local IP address, local port, remote IP address, remote port>.

   The session might no longer be unique when two clients with the same source ELOC prefix residing in two separate ALOC realms are accessing a server locate in a third ALOC realm. In this scenario a possibility exists that the clients 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 clients uses the same values in the five tuples <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 clients are informed by ICMP notification about the "identical session situation".

MPTCP introduces a token, which is locally significant and currently defined as 32 bit long. The token will provide a sixth tuple for future applications to identify and verify the uniqueness of a session - the probability to have an "identical session situation" is further reduced.

9. Traceroute considerations

As long as the traceroute is executed inside the local ALOC realm normal IPv4 traceroute mechanism can be used. As soon as the traceroute exits the local ALOC realm the locator header shall be used in the notifications. Therefore extension to ICMP protocol shall be implemented, the extensions shall be compatible with [RFC4884].

10. Multicast considerations

Since source and destination IPv4 prefixes are only installed in the RIB of the local ALOC realm there is a constraint with Reverse Path Forwarding (RPF) which is used to ensure loop-free forwarding of multicast packets. The source IP address of a multicast group (S,G) is used against the RPF check. The source IP address can no longer be used as a RPF checkpoint outside the local ALOC realm.

To enable RPF globally for a (S,G), the multicast enabled LSR (mLSR) must at the source ALOC realm replace the source IP address with the local ALOC prefix for inter-ALOC multicast streams. This can be achieved if the local LSR act also as an Anycast Rendezvous Point with Multicast Source Discovery Protocol (MSDP) and Protocol Independent Multicast capabilities; with these functionalities the LSR becomes a multicast enabled LSR (mLSR). The sender register at the mLSR and a source tree is established between the sender and the mLSR. When an inter-ALOC realm receiver subscribes to the multicast group the mLSR have to swap the IPv4 multicast packet in the following way:

a. verify the received packet that it uses the hIPv4 protocol number in the protocol field of the IP header

b. verify IP- and transport header checksums

c. replace the source address in the IPv4 header with the local ALOC value

d. set the S-field to 1
In order for the mLSR to function as described above the sender must assemble the multicast hIPv4 packet in the following way:

a. set local IP address (S) from API in the source IP address and the ELOC field
b. set remote IP address (G) from API in the destination IP address field
c. set local ALOC value in the ALOC field
d. set the hIPv4 protocol number in the protocol field of the IP header and set the transport protocol number in the protocol field of the locator header
e. set the desired parameters in the A-, P-, S-, VLB-, L-, and R-fields of the locator header
f. Apply checksum IP-, locator- and transport header calculations, transport header calculation do not include the locator header fields. When completed the packet is transmitted.

The downstream routers from the mLSR to the receiver will use the source IP address (which value is the source ALOC prefix after the mLSR) in the IPv4 header for RPF verification. In order for the receiver to create RTCP receiver reports all information is provided in the hIPv4 header of the packet.

Because Source Specific Multicast (SSM) and IGMPv3 uses IP addresses in the payload both protocols needs to be modified to support the hierarchical IPv4 framework.

11. Traffic engineering considerations

When hIPv4 framework is fully implemented ingress load balancing to an ALOC realm can be influenced by the placement of LSRs at the realm; a LSR provides a "nearest routing" scheme. Also, if RIR policies allows, a service provider can have several ALOC assigned, hence traffic engineering and filtering can be done with the help of
ALOC prefixes. E.g. sensitive traffic can be aggregated under one ALOC prefix which is not fully distributed into the DFZ of Internet. If needed an ALOC Traffic Engineering solution between ALOC realms might be developed, i.e. create explicit paths that can be engineered via specific ALOC prefixes. Further studies are needed; first it should be evaluated if there is demand for such a solution.

The usage of multipath enabled transport protocols opens up the possibility to develop a new design methodology of backbone networks, i.e. Valiant Load-Balancing [VLB]. If two single-homed endpoints, using multipath enabled transport protocols and attached to the network with only one interface/IP-address/ELOC-prefix, are communicating over Internet, both subflows will most likely take the shortest path throughout Internet. I.e. 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 – the benefit of multipath is lost. The "subflows-over-same-links" scenario can be avoided if the subflows are traffic engineered to traverse 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 LSR functionality is deployed on a Valiant Load-Balancing enabled backbone node – hereafter called vLSR – and the backbone nodes are interconnected via logical full meshed sessions, Valiant Load-Balancing can be applied for the subflows. When a subflow has the appropriated bits set in the VLB-field of the locator header the first ingress vLSR shall do VLB switching of the subflow. That is, the ingress vLSR is allowed to do VLB switching of the subflow’s packets if the VLB bits are set to 01 or 11, the S-bit is set to 0 and the local ALOC value of the vLSR matches the ALOC-field’s value. If there are no ALOC and ELOC fields in the locator header, but the other fields’ values are set as described above, the vLSR should apply VLB switching as well for the subflow – because it is an inter-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 most likely the subflows will be routed throughout the Internet on different paths. If VLB switching is applied in every ALOC realm it 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 – if the subflows on the first and last mile links 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 on traffic patterns between interconnected VLB [iVLB] backbone networks has been carried out. Nevertheless, more studies are needed around Valiant Load-Balancing scenarios.
12. Large encapsulated packets

Adding the locator header to an IPv4 packet in order to create a hiIPv4 packet will increase the size of it but since the packet is assembled at the endpoint it will not add complications of current Path MTU Discovery (PMTUD) mechanism in the network. The intermediate network between two endpoints will not see any difference in the size of packet; IPv4 and hiIPv4 packet sizes are the same from the network point of view.

13. Mobility considerations

This section will consider two types of mobility solutions, site mobility and endpoint mobility.

Site mobility definition:

a site wishes to changes its attachment point to the Internet without changing its IP address block

Today, classical multi-homing is the most common solution for enterprises that wishes to achieve site mobility. Multi-homing is one of the key findings behind the growth of the DFZ RIB, see the [IAB report], section 2.1 and 3.1.2. The hiIPv4 framework can provide a solution for enterprises to have site mobility without the requirement of implementing a classical multi-homed solution. This new multi-homed solution utilizing PI addresses is depended upon the forthcoming IPv4 address allocation policy which is discussed in appendix A. If the geographical region based IPv4 address block allocation is deployed the enterprise can be concurrently attached to two different Internet service providers (ISP) without the need to implement AS border routing. The ISPs provide their globally unique ALOC prefixes for the enterprise and the IPv4 address block of the enterprise is geographically unique, a PI address block. The enterprise can change on per endpoint basis the local ALOC prefix, i.e. from the previous ISP’s ALOC prefix to the new ISP’s ALOC prefix. Sessions initiated at the enterprise needs to be routed to the correct ISP, i.e. the border router of the enterprise needs to apply policy based routing upon the ALOC prefix in the locator header. For sessions initiated from the Internet the DNS record for an endpoint needs to be updated, also the local ALOC prefix on the endpoint needs be changed to achieve a symmetric path. Since the border router is enforcing policy based routing upon the ALOC prefix of the locator header the server can apply basic session load balancing over the two ISPs based upon from which ISP the session has been initiated, i.e. if the server have two valid DNS records with two different ALOC prefixes. The updating rate of DNS records is
considered slow but recent studies have shown that this is no longer the case, updates at rates as high as once per second can be achieved, see [DynamicDNS]. Conclusion is that a single-homed enterprise can achieve smooth transition from one ISP to another by only changing the ALOC prefix on the endpoints and at DNS records - the local IP addressing (ELOC) scheme remains intact. Also a single-homed enterprise can become multi-homed without implementing AS border routing or to have an own ALOC prefix assigned. If a better session load balancing scheme is required the application should be migrated to a multipath enabled transport protocol such as [SCTP] or [MPTCP]. Multi-homing is discussed in detail in appendix B.

Endpoint mobility definition:

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

Mobile IP [MIP] is used today for IPv4 endpoints in order to provide mobility. Mobile IP is an overlay protocol; it is also an application that uses IP addresses in its payload. It is obvious that IPv4 extensions need to be added to the MIP framework. Another approach is to investigate what [MPTCP] can offer to solve endpoint mobility scenarios. MPTCP introduces a token, which is locally significant and currently defined as 32 bit long. The token will provide a sixth tuple to identify and verify the uniqueness of a session. This sixth tuple - the token - is not depended upon the underlying layer, i.e. the IP layer. The session is identified with the help of the token and thus the application is not aware when the IP parameters are changed, e.g. during a roaming situation - but it is required that the application is not making use of IP addresses. Security issues arise; the token can be capture during the session by e.g. a man-in-the-middle attack. If the application requires protection against man-in-the-middle attacks the user should apply Transport Layer Security [TLS] Protocol for the session.

14. Affected Applications and Implications

There are several applications that are inserting IPv4 address information in the payload of a packet. Some applications use the IPv4 address information to create new sessions or for identification purposes. This section is trying to list the applications that need to be enhanced; however, this is by no means a comprehensive list. The applications can be divided in four main categories:

- Applications based on raw sockets, a raw socket is receiving packets containing the complete header in comparison to the other sockets that only receives the payload.
Applications needed to enable the hIPv4 framework, i.e. DNS and DHCP databases which must be extended to support ALOC prefixes.

Applications that insert IPv4 addresses to the payload and uses the IPv4 address for setting up new sessions or for some kind of identification. The application belonging to this category can not set up sessions to other ALOC domains until extensions have been incorporated. Within the local ALOC domain there are no restrictions since the current IPv4 scheme is still valid. The following applications have been identified:

- SIP; IPv4 addresses are inserted in the SDP, Contact and Via header
- Mobile IP; the mobile node uses several IPv4 addresses during the registration process
- IPsec AH; designed to detect alterations at the IPv4 packet header
- RSVP; RSVP messages are sent hop-by-hop between RSVP-capable routers to construct an explicit path
- ICMP; notifications needs 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 of the sender
- IGMPv3; a source-list is included in the IGMP reports
- Applications related to security, such as firewalls, must be enhanced to support ALOC prefixes
- Applications that will function with FQDN but many uses an IPv4 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.

15. The Future Role of the LSR

The LSR was added to the framework in order to provide a smooth transition from the current IPv4 framework to the hierarchical IPv4 framework, i.e. a major forklift of the current forwarding plane is avoided by the introduction of the LSR element. In the future, the LSR can be left as such in the network, if preferred, or the LSR...
functionality can be expanded towards the edge when routers are upgraded due to their natural lifecycle process. Once an upgrade of a router is required because of e.g. increased demand for bandwidth, the modified forwarding plane might support concurrently IPv4 and hIPv4 forwarding – and the LSR functionality can be pushed towards the edge (the ultimate goal is to have LSR functionality integrated in the endpoints). This is accomplished by adding extension to the current routing protocols, both IGP and BGP. When a LSR receives a hIPv4 packet where the destination IPv4 address matches the local ALOC prefix the LSR shall – contrary to the tasks defined in section 7, step 4 – lookup the ELOC field in the locator header and compare this value against the FIB. If the next-hop entry is LSR capable the packet shall be forwarded upon the ELOC value. If the next-hop is a legacy IPv4 router the LSR must apply the tasks defined in section 7, step 4 and once completed forward the packet upon the new IPv4 destination address.

Once the routers from the first ingress LSR to the final destination endpoint is upgraded to support hIPv4 forwarding there exist no longer a need to implement LSR functionality in the network of the remote ALOC realm, the packet is forwarded as such to the endpoint’s extended stack. The hIPv4 stack must check that the ELOC value matches its local IPv4 address, because the destination IPv4 address matched the local ALOC prefix. Then the hIPv4 stack of the destination must present to the extended IPv4 socket API the following:

a. present source IP address as the remote IP address
b. present the destination IP address field value as the local ALOC
c. present the ALOC field value as the remote ALOC
d. present the ELOC field value as the local IP address

Multicast LSR (mLSR) functionality remains in the network; it is an extension to the Anycast RP with MSDP element. For sessions inside the ALOC domain legacy IPv4 forwarding plane is kept in place.

16. Transition considerations

The hIPv4 framework is not introducing any new protocols that would be mandatory to carry out the transition from IPv4 to hIPv4; instead extensions are added to existing protocols – the hIPv4 framework requires extensions to the current IPv4 stack, databases and to some applications that use IP addresses in the payload but the current
forwarding plane in Internet remains intact apart from that a new forwarding element (the LSR) is required to create an ALOC realm. Extensions to the IPv4 stack, databases, applications that uses IP addresses in the payload and routers can be deployed in parallel with the current IPv4 framework. Even genuine hIPv4 sessions can be established between endpoints though the current single dimensional Internet structure is still present. When will the single dimensional routing architecture then be upgraded to a two level architecture? The author thinks there are two possible tipping points:

- When the RIB of DFZ is getting close to the capabilities of current forwarding plane - 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 why hIPv4 framework will not succeed is the short timeframe before the expected depletion of the IPv4 address space occurs. Also, will enterprise give up their global allocation of the current IPv4 address block they have gained? Another risk is, will the enterprises and residences carry out an upgrade of their endpoints and security nodes? Transitions arguments are discussed in appendix D.

17. Security Considerations

Hijacking of a single ELOC prefix by longest match from another ALOC realm is no longer possible since the prefixes are separated by a locator, the ALOC. To apply a hijack of a certain ELOC prefix the whole ALOC realm must be routed via a bogus ALOC realm. Studies should be carried out with the Secure Inter-Domain Routing (SIDR) workgroup if the ALOC prefixes can be protected from hijacking.

18. IANA Considerations

TBD

19. Conclusion

This document provides a high level overview of the hierarchical IPv4 framework which could be build in parallel with the current single dimensional Internet by implementing extensions at several architectures. Implementation of the hIPv4 framework will not require a major service window break in the Internet, neither at the internal networks of enterprises. Basically, the hIPv4 framework is an
evolution of the current IPv4 framework. For sessions inside an ALOC realm the IPv4 framework can be used in the future and for inter-ALOC realm sessions the hIPv4 framework is needed. Though there is a long journey ahead and many things that need to be sorted out the hierarchical IPv4 framework looks promising. The transition can be attractive for the enterprises since the hIPv4 framework doesn’t create a catch-22 situation, it introduces functionalities (related to site and endpoint mobility) that could better serve their business models, introduce less expensive multi-homing solutions, it slows down the expected growth of Internet’s carbon footprint and it is inline with the Corporate Social Responsibility programs that many enterprises have implemented. The framework should also be interesting for the service providers, when the transition phase is completed the growth of DFZ will be controlled by the service providers and only the service providers - multi-homed enterprises might not influence on the RIB size of the DFZ anymore. After the transition the RIB size of the DFZ will be reduced, which should have a decreasing effect on the expected cost structure of future DFZ routers, both operating and capital expenditure.

20. References

20.1. References


20.2. Informative References

[RBridge] Perlman, R., "RBridges, Transparent Routing", Infocomm, 2004

[DynamicDNS] Pappas, A., Hailes, S., Giaffreda, R., "Mobile Host Location Tracking through DNS", University College London and BTexact Technologies, 2002


[MPTCP] Ford, A., Raiciu, C., Handley, M., Barre, S., "TCP Extensions for Multipath Operation with Multiple Addresses", IETF TSV Area, July 2009

[VLB] Zhang-Shen, R., McKeown, N., "Designing a Predictable Internet Backbone with Valiant Load-Balancing", Stanford University

[iVLB] Babaioff, M., Chuang, J., "On the Optimality and Interconnection of Valiant Load-Balancing Networks", University of California at Berkeley


21. Acknowledgements

The author would like to acknowledge Aki Anttila, Antti Jarvenpaa and Robin Whittle for giving helpful feedback. Also the active participants at the Routing Research Group [RRG] mailing-list are acknowledged. The participants have provided ideas, proposals and discussions that have influenced on the architecture of the hIPv4 framework.
Appendix A. Future IPv4 address allocation policies

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

When the hIPv4 framework is fully implemented every ALOC realm can have a full IPv4 address space - except the GLB - to allocate ELOC prefix blocks from. There are some implications though. In order for an enterprise to achieve site mobility, i.e. to change service provider without changing its IP addressing scheme, the enterprise should implement an Autonomous System (AS) solution with ALOC prefix at the attachment point to the service provider. Larger enterprises do have the resources to implement AS border routing; most of the large enterprises have already implemented multi-homing solutions. The small and midsize enterprises (SME) may not have the resources to implement AS border routing, or the implementation introduces unnecessary costs for the SME. Also if every SME needs to have an ALOC prefix it will have an impact on the RIB at the DFZ, the RIB will be populated with a huge amount of ALOC prefixes.

It is clear that a compromise is needed. A SME is usually single-homed and the SME should be able to reserve a PI address block from the RIR without the need to be forced to create an ALOC realm, i.e. implement a LSR solution and AS border routing. The PI address block is no longer globally unique, the SME can only reserve the PI address block for the country or countries where it is active or has it attachment point to Internet. The attachment point rarely changes to another country; therefore it is sufficient that the PI address block is regionally unique. When the SME is replacing its Internet service provider the SME do not have to change its ELOC addressing scheme - only the local ALOC prefix at the endpoints are changed. The internal traffic at an SME does not use the ALOC prefix, the internal routing is applied by the IPv4 header and thus the internal routing and addressing architectures are preserved.

Mergers and acquisitions of SME can cause IP address conflicts, because the PI address block is hereafter only regionally unique. If a SME in region A overtakes a SME in region B there is a slight chance that both SME can have overlapping IPv4 addresses. An idea to address this scenario is to categorize the SME upon their business areas. It is highly unlikely that a SME in, e.g. agriculture business area will ever acquire a SME in the medical business area, or vice versa. When a SME applies for a PI address block the RIR could verify to which business area the SME operates in and do a global check in
order to give the SME a globally unique IP address block from that business area. Large enterprises also merge, but since the large enterprises are usually multi-homed the merger of networks can be rapidly carried out with the help of an ALOC realm. During a merger usually the infrastructure of a company is slowly incorporated to the other company’s infrastructure, this integration usually requires redesign of the network architecture and therefore in most cases the ALOC realm routing can be removed.

Finally, residential users will receive only PA addresses. When a residential user changes a service provider the residential user has to replace the IP addresses. A PA address block is no longer globally unique, every Internet service provider can use the PA address blocks at their ALOC realms - the PA addresses becomes kind of private IP addresses for the service providers.

The hIPv4 framework will provide re-usage of IPv4 address blocks, the globally unique reservation of IPv4 address block shall be replaced by a geographical region and/or business area specific allocation. The biggest challenge is when merger and acquisitions are carried out, there is a possibility that overlapping of IP addresses could still happen but the hIPv4 framework reduces this problem to a minimum compared what is seen today during mergers with the usage of private IP addresses [RFC1918].
Appendix B. Multi-homing becomes multi-pathing

When the transition of the hIPv4 framework is fully completed the RIB of an ISP, that has created an ALOC realm, will have the following entries:

- the PA addresses (ELOC) of directly attached customers (e.g. residential and enterprises)
- the PI addresses (ELOC) of directly attached customers (e.g. enterprises)
- the globally unique ALOC prefixes, received from other service providers and enterprises using classical multi-homing (i.e. PI addresses, AS-number and BGP) with an assigned ALOC prefix

The ISP will not carry in its RIB any PA or PI addresses (ELOC) from other service providers. In order to do routing and forwarding of packets between ISPs only ALOC information of other ISPs is needed. So the ALOC is a sort of a super-aggregate, locating the ALOC realm of a service provider in Internet and thus reducing the RIB size in the DFZ.

But this approach will not help that much in classical multi-homing scenarios, i.e. if the enterprise is also assigned an ALOC prefix for the multi-homed network. The classical multi-homing is causing the biggest impact on growth of the size of the RIB in the DFZ - replacing a /20 IPv4 prefix with a /32 ALOC prefix will not reduce the size of the RIB in the DFZ.

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

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

By implementing this kind of ALOC allocating policy it will reduce the RIB size in the DFZ quite well, multi-homing will no longer increase the RIB size of the DFZ. But this policy will have some impact on the resilience behavior, by compressing routing information
we will lose visibility in the network. In today’s multi-homing solutions the network always know where the remote destination resides, in case of a link or network failure a backup path is calculated and an alternative path is found - 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 of what failures (almost) happen in the network. And with the growth of multi-homed networks the routers in the DFZ have been forced to carry a greater workload, perhaps close to their limits - the workload between the network and endpoints is not in balance. Conclusion is that the endpoints should take more responsibilities for their sessions and that way off-loads the workload in the network. How, lets walk through an example:

An enterprise has been given a PI address block 192.168.1.0/24 (ELOC) that is either via static routing or BGP announced to the upstream service providers. The upstream service providers are providing the ALOC information for the enterprise, i.e. 10.1.1.1 and 10.2.2.2. A server has been installed, it has been given ELOC 192.168.1.1 - the ELOC is a locator defining where the server is attached to the local network. The server has been assigned ALOCs 10.1.1.1 and 10.2.2.2 - the ALOC have no forwarding functionality within the local network - an ALOC is a locator defining the attachment point of the local network to Internet.

My client, which has ELOC 172.16.1.1 and ALOC prefixes 10.3.3.3 & 10.4.4.4, has established a session by using source ALOC 10.3.3.3 to the server at ELOC 192.168.1.1 and ALOC 10.1.1.1, i.e. both networks 192.168.10/24 and 172.16.1.0/24 are multi-homed. ALOC are not available in 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. Next a network failure occurs, the link between the server and service provider that owns ALOC 10.1.1.1 goes down. My border router 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 the border router in front of the server will notice the link failure; the border router could replace the ALOC field in the locator header for this session, i.e. from 10.1.1.1 to 10.2.2.2 and send the packets to the second service provider. Now the session between my client using ELOC 172.16.1.1, ALOC 10.3.3.3 and the server using ELOC 192.168.1.1, ALOC 10.2.2.2 remains intact because the five tuples at the IP stack API do not change - only the ALOC value of the server has changed and this information is not shown to the application. An assumption here is that the hIPv4 stack
does accept changes of ALOC values on the fly (more about this later).

If the network link between my border router and ISP using ALOC 10.3.3.3 fails, my border router could replace the ALOC value in the locator header and route the packets via ISP using ALOC 10.4.4.4 - and the sessions stays up. If there is a failure somewhere in the network the border routers might receive an ICMP destination unreachable message and thus try to switch to session over to the other ISP by replacing the ALOC values 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 five API tuples remains intact.

If border routers or one of the endpoint changes the ALOC value without a negotiation with the remote endpoint security issues arises. Can the endpoint(s) trust the remote endpoint when ALOC value(s) 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. Two types of identifiers have been discussed on the [RRG] mailing-list, session and host identifiers. The host identifier has the characteristics of a Public Key Infrastructure certificate solution. PKI solutions has been developed and deployed, thus it is recommended that PKI solutions should be used when an endpoint needs to be authenticated. When the ALOC value changes the PKI solution might need to re-authenticate the endpoints, it is up to the security experts to evaluate the risks and threats. When the security requirements are lower, e.g. browsing a web-site, a less complicated identification mechanism is preferable - it should be less complex to deploy and maintain. 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. If the application requires protection against man-in-the-middle attacks the user should apply Transport Layer Security [TLS] Protocol 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 redundant paths to the other endpoint via different ISPs
true dynamic load-balancing, the endpoints do not participate in any routing protocols

only a single NIC on the endpoints is required, but several NICs can be used

in case of a border router or ISP failure, the transport protocol will provide resilience

By adding more intelligence at the endpoints, i.e. multipath enabled transport protocols, the workload of the network is off-loaded and can take less responsibility for providing visibility of destination prefixes in Internet – i.e. prefix compression in the DFZ can be applied and only the attachment points of a local network needs 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, the rest is only regionally unique as discussed in appendix A. Outcome is that the current multi-homing solution can migrate towards a multi-pathing environment that will have the following characteristics:

AS number is not mandatory

regional PI address space is mandatory

BGP protocol is not mandatory at the enterprise’s border routers, static routing with Bidirectional Failure Detection [BFD] is an option

allocation of ALOC for the enterprise is not mandatory, upstream ISPs are providing the ALOC prefixes for the enterprise

MPTCP provides dynamic load-balancing without using routing protocols, several paths can be simultaneously used and thus resilience is achieved

provide low growth of RIB entries at the DFZ

when static routing is used between the enterprise and ISP:

- the RIB size at the enterprise’s border routers are not depended upon the size of the RIB in DFZ nor adjacent ISPs
- the enterprise’s border router can not cause BGP churn in the DFZ or in the adjacent ISPs’ RIB

when dynamic routing is used between the enterprise and ISP:
o the RIB size at the enterprise’s border routers are depended upon the size of the RIB in DFZ and adjacent ISPs

o the enterprise’s border router can cause BGP churn for the adjacent ISPs but not in the DFZ

o the cost of border router is less expensive than in today’s multi-homing solution
Appendix C. Mobile site crossing a RIR border

Discussions regarding Network Address Translation, NAT, have been taking place on the [RRG] mailing-list. The outcome of the discussions are that NAT has become a de-facto part of the current Internet architecture - NAT has been so widely deployed that NAT can no longer be ignored as a temporary solution and thus NAT needs to be taken into account in the research work of a new routing architecture. Though the hIPv4 framework has the capabilities to reduce the usage of NAT, hIPv4 will not make NAT to be totally obsolete in the future. In the future there will still be use cases where NAT might be required, e.g. mobile vehicles that are crossing RIR boundaries and the vehicle (e.g. aircraft, train, ferry etc) carries a local network. If the RIR are setting up an IP address allocation policy as discussed in appendix A, there are no longer globally unique IP addresses, except one block (GLB) that is reserved to create the foundation of the DFZ. IP addresses from the GLB block can not be used for networks at mobile vehicles, nor might PI addresses be used if the vehicle crosses a RIR boundary. Enterprises could reserve a PI address block in every region and that way create a globally unique IP-address block, again this scenario is depended upon the forthcoming RIR policies.

Thus, most likely, a private IP address block [RFC1918] needs to be assigned for a LAN enabled vehicle that is crossing regional borders. With this requirement in mind, mechanisms to ease the inbound NAT traversal challenges - e.g. sessions initiated from Internet to an endpoint, using a private IP address [RFC1918], which is attached to a private network - is needed, i.e. the hIPv4 framework must provide a scalable bidirectional session model for NAT. Therefore, a private locator referral (PLR) mechanism has been added to the hIPv4 framework. The PLR mechanism is a local static global-private locator mapping relationship in a middlebox, sitting on the border between a private network and Internet. The mapping relationship can be published to the general public via DNS or only published discreetly to partners for e.g. business-to-business sessions.

When DNS is used to publish the PLR a new type of DNS record is required. When an endpoint receives the value of the new DNS record it shall copy the value into PLR field of the locator header for the appropriate session - the A-record will contain the public IP address of the middlebox. The middlebox, which is sitting in front of the remote endpoint, must have a mapping scheme, i.e. a table of private locator referral values that are associated with appropriate private IP addresses of the endpoints inside the private network. Since the PLR field is 32-bit the private IP address can be published as such
and no local mapping scheme is required on the middlebox, the private IP address is carried within the PLR-field during the session.

The middlebox must also multipath capable, i.e. using multipath transport protocol to apply the transition of the session from one ALOC realm to another ALOC realm. The server onboard the mobile site doesn’t necessary need to make use of a multipath enabled transport protocol; the middlebox will act as a multipath proxy in front of the server. Also the client doesn’t need to make use of a multipath enabled transport protocol - if the DNS server is not on the mobile site and the middlebox can cache DNS messages on behalf of the client. It might become complicated, thus it is recommended that the client make use of multipath enabled transport protocols. During the transition the ELOC values for a session will not change, as discussed in appendix B, only ALOC value changes. Neither the client nor the server at the mobile site need to setup new subflows during the transition phase, the middlebox needs to setup the subflows since it will discover when there is a new attachment point to Internet available - unless the middlebox informs the client and servers of the new attachment point, for that, a new protocol or an extension to ICMP is needed.
Appendix D. Transition Arguments

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

- Not all endpoints needs to be upgraded, only those endpoints that are directly attached to the Internet. These kinds of endpoints are portable laptops, smart mobile phones, proxies, and DMZ/frontend servers. But the most critical servers, the backend servers where enterprises keep their most critical business applications do not need to be upgraded; the backend servers should not be reached at all from Internet - only from the Intranet - and this functionality can be achieved with the hIPv4 framework, since it is backwards compatible with the current IPv4 stack.

- Mobility, it is estimated that the demand for applications that performs well over the wireless access network will increase. Introduction of MPTCP opens up a new possibility to create new solutions and applications that are optimized for mobility. The hIPv4 framework requires an upgrade of the endpoints' stack; if possible the hIPv4 stack should also contain MPTCP features. Applications designed for mobility could bring competitive benefits for the enterprises.

- The intermediate routers in the network do not need to be upgraded (hardware), the current forwarding plane can still be used and the hIPv4 packet is capable to traverse most of the current NAT implementations. The benefit is that the current network equipment can be preserved at the service providers, enterprises and residences. That 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.
o The migration from IPv4 to IPv6 (current defined architecture) will increase the RIB and FIB throughout DFZ, will it require a new upgrade of the forwarding plane as discussed in the IAB report 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 is reserving one or several RIB/FIB entries on every router in the DFZ it means that the enterprise is increasing the power consumption of Internet, thus increasing the carbon footprint. And many enterprises are committed to green programs - if hIPv4 gets deployed, the power consumption of Internet will not grow as much as compared in an IPv4 to IPv6 transition scenario.

o Another issue, if the migration from IPv4 to IPv6 (current defined architecture) occurs, is that the routers in the DFZ most likely need to be upgraded to more expensive routers - as discussed in the IAB report. In the wealthy part of the world, where a large penetration of Internet users is already present, the service provider can pass along more easily the costs of the upgrade to their subscribers - with a "wealthy/high penetration" ratio the cost will not grow that much that the subscribers would abandon 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 that easily be covered - a "less wealthy/low penetration" ratio could have a dramatic increase on 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 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, also other arguments that the general public is interested in and concerned about can be found. Such arguments as that the Internet becomes greener and more affordable for everyone than compared to current forecast of the evolution of Internet. These non-technical values need to be communicated to the general public, you could ask them - do you care about the Planet, People and Internet? If you do, please upgrade the stack on your Internet enabled device.
Author’s Address

Patrick Frejborg
Email: pfrejborg@gmail.com