An Architectural Framework of the Internet for the Real IP World
draft-shyam-real-ip-framework-25.txt

Abstract

This document tries to propose an architectural framework of the internet in the real IP world. It shows how to reorganize the provider network with a large address space. It describes how a three-tier mesh structured hierarchy can be established based on fragmenting the entire space into some regions and some sub regions inside each of them. It addresses issues which could be relevant to this architecture in the context of IPv6.

Table of Contents

1. Introduction

Transition from IPv4 to IPv6 is in the process. Work has been done to upgrade individual nodes (workstations) from IPv4 to IPv6. Also, there are established documents to make routers/switches to work to support IPv4 as well as IPv6 packets simultaneously in order to make the transition possible [1]. CIDR[2] based hierarchical architecture in the existing 32-bit system is supposed to be continued in IPv6 too with a large address space. There are documents/concerns over BGP table entries to become too large in the existing system [3]. There are proposals to upgrade Autonomous System number to 32-bit from 16-bit to support the demand at the same time [4]. The challenge relies on how to make the transition smooth from IPv4 to a real IP world with least changes possible.

The term "real IP environment" is referred to an environment where hosts in a customer network will possess globally unique IP addresses and communicate with the rest of the world without the help of NAT[5].

2. Background

Existing system is in work with Autonomous System (AS) and inter-AS

layer with the approach of CIDR. In order to meet the need within the 32-bit address space, Autonomous Systems of various sizes maintain CIDR based hierarchical architecture. With the help of NAT [5], a stub network can maintain an user ID space as large as a class A network and can meet its useful need to communicate with the rest of the world with very few real IP addresses. With the combination of CIDR and NAT applied in the entire space, most of the part of 32-bit address space gets effectively used as network ID. This is how, 16-bit 'Autonomous System Number' is realized as insufficient in order to meet the need of growing customers. If the same gets continued with a larger network ID, load in the switches will become too high.

With traditional CIDR based hierarchy, a node of higher prefix can be divided into number of nodes with lower prefixes. Each divided node can further be subdivided with nodes of further lower prefixes. This process can be continued till no further division is possible. The point worth noting is at each point the designer of the network has to preconceive the future expansion of the network with the concept in the mind that the resource can not be exhausted at any point of time. This phenomenon leads the designer to allocate resources much higher than whatever is needed which leads to a space of unused address space and the concept of H-D (host-density) ratio comes into play. The problem gets aggravated once resource gets exhausted by any chance. e.g. a node of prefix /16 can be divided with a number of nodes of prefixes /24. If any one of the nodes /24 gets exhausted, resources of other nodes of prefixes /24 can not be used even if they are available.

Transition from private IP to real IP may not appear to be a simple task. This has happened due to the desperate attempt of the service providers to provide internet services with the help of NAT. e.g. a large educational institute meets its current requirement with 4 real IP addresses; one for its mail server, one for its web server, one for its ftp server and another one for its proxy server to provide web based services to all of its users. These four types of services are used by any organization of any size(it may be 400 or even 40000). In the current provider network these organizations are supported their need with 4 IP addresses and the CIDR based tree has been built using these components together. When private IP will be replaced with real IP, each customer network will require IP addresses based on its size and requirement. So, even if CIDR based architecture is maintained with real IP space, existing provider based network needs to be reorganized. The desired approach will be to assign address block that will be proportional to the sizes (bandwidth) of the ports of the switches of the provider network.

3. A Three-tier mesh structured hierarchical network

As Autonomous Systems of various sizes are supported, Autonomous Systems and the nodes inside the Autonomous Systems can be viewed as graphically lying on the same plane within the address apace. If network can be viewed as lying on different planes, routing issues can be made simpler. If network is designed with a fixed length of prefix for the Autonomous System everywhere, routing information for the rest will get confined with the other part of the network prefix. Which means the maximum size of AS gets assigned to all irrespective of their actual sizes. This can be made possible with the advantage of using a large address space and dividing it into number of regions of fixed sizes inside it. Thus entire network can be viewed as a network of inter-AS layer nodes. Each node in the inter-AS layer can act either only as a router in the inter-AS layer or as a router in the inter-AS layer with an Autonomous System attached to it with a single point of attachment or as an Autonomous System with multiple Autonomous System border routers (ASBR) appearing like a mesh. Thus two tier mesh structured hierarchy gets established between AS layer and inter-AS layer with each AS having a fixed length of prefix.

Based on the definition of Autonomous System, it is a small area within the entire network that maintains its own independent identity that communicates with the rest of the world through some specific border routers. In the similar manner, if a larger area (say region or state) can be considered as network of Autonomous Systems, that can maintain its own identity by communicating with the rest of the world through some border routers (say, state border router), mesh structured hierarchy can be established within the inter-AS layer. The inter-AS layer will be split into inter-AS-top and inter-AS-bottom. To maintain this hierarchy, each node of inter-AS-top needs to have multiple regional or state border routers (say, SBR) through which each one will communicate with the rest of the world in the similar manner an Autonomous System maintains ASBR. Thus, entire network will appear as a network of nodes of inter-AS-top layer. To maintain hierarchy, each node of the inter-AS-top needs to have a fixed length of prefix. i.e. each node of the inter-AS top will be assigned a maximum (fixed) number of nodes of Autonomous Systems.

Thus, with three-tier mesh structured hierarchy in the network layer, network ID can be viewed as A.B.C. If pA, pB and pC be the prefix lengths of inter-AS-top, inter-AS-bottom and AS layers respectively, there will be 2^pA nodes at the topmost layer, 2^pB at the inter-AS-bottom layer and 2^pC nodes at the AS layer. Thus the entire space gets divided into a fixed number of regions and each region gets divided into fixed number of sub regions. This division is supposed to be made based on geography, population density and their demands and related factors.

Let nMaxInterASTopNodes be the possible maximum number of nodes assigned at the top most layer and nMaxInterASBottomNodes be that at the inter-AS-bottom layer and nMaxASNodes at the AS layer. Where nMaxInterASTopNodes <= 2^pA and nMaxInterASBottomNodes <= 2^pB and nMaxASNodes <= 2^pC.

3.1. Route propagation

With hierarchy established, routing information that gets established inside a node of inter-AS-top, does not need to be propagated to another node of inter-AS-top. Entire routing information of inter-AS-top layer needs to be propagated to inter-AS-bottom layer. So, each router of inter-AS layer will have two tables of information, one for the inter-AS-top and another for the inter-AS-bottom of the inter-AS-top node that it belongs to. BGP (with little modification) will work very well with a trick applied at the SBRs. Each SBR will not propagate the routing information of inter-AS-bottom layer of its domain to another SBR of neighboring domain. i.e. SBR of one top layer node will propagate routing information only of inter-AS-top layer to SBR of another top layer node. Inside a node of inter-AS-top, routing information of inter-AS-top and inter-AS-bottom need to be propagated from one ASBR to another neighboring ASBR. Inside a top layer node A, routing information of another top layer node B will have two parts; one for the list of SBRs through which a packet will traverse from top layer node A to B and another for the list of ASBRs through which the packet will traverse from one AS to another inside A. In terms of BGP, AS_PATH attribute will be split into two parts; one for the information of the top layer and another for the bottom layer. Within the same node A routing information of one AS to another AS will not have any top layer information. i.e. the top layer information will be set to as NULL.

Similarly, each node of the AS layer will have three tables of routing entries. One for the inter-AS-top, one for the inter-AS-bottom and another for the routing information inside the Autonomous System itself.

Introduction of hierarchy at the inter-AS layer reduces the size of the routing table substantially. With the availability of hardware resources if flat address space is maintained at each layer, problems related to CIDR can be avoided. With flat address space, no hierarchical relationship needs to be established between any two nodes in the same layer. So, all the nodes inside each layer can be used till they get exhausted. With flat address space (i.e. without prefix reduction), BGP tables will have maximum nMaxInterASTopNodes + nMaxInterASBottomNodes entries.

IGP like OSPF has got provision to divide AS into smaller areas. OSPF

hides the topology of an area from the rest of the Autonomous System. This information hiding enables a significant reduction in routing traffic. With the support of subnetting, OSPF attaches an IP address mask to indicate a range of IP addresses being described by that particular route. With this approach it reduces the size of the routing traffic instead of describing all the nodes inside it, but introduces another level of hierarchy. If subnetting concept can be avoided from the AS layer(with the additional overhead of computation inside the SPF tree), each area can be configured from a free pool of addresses based on its requirement dynamically. So, an AS can be divided into number of areas of heterogeneous sizes with the nodes from a free pool of address space.

Similarly, the concept of area can be introduced in the inter-AS-bottom layer the way it works in OSPF. The area border routers in the inter-AS-bottom layer have to behave exactly in the similar manner the way an ABR behaves in OSPF. i.e. an area border router will hide the topology inside an area to the rest of the world and will distribute the collected information inside the area to the rest. It will distribute the collected routing information from outside to the nodes inside as well. In order to implement this, protocol running in the inter-AS layer (say BGP) will have to introduce a 'cost' factor. This cost factor can be interpreted as the cost of propagation of a packet from one AS to another. The protocols running inside AS layer (RIP/OSPF, etc) will have to the supply the cost information for a packet to travel from one ASBR to another. All the protocols must behave in unison for supplying this information. The cost factor is needed for a remote node while sending a packet to a node inside an area while more than one area border routers are equidistant from that remote node. Thus inter-AS-bottom layer (i.e. one inter-AS-top level node) can be divided into number of areas of heterogeneous sizes with nodes of AS from a free pool of address space. BGP adopts a technique called route aggregation. Along with route aggregation it reduces routing information within a message. In the similar manner, introduction of area inside inter-AS-bottom layer will not only reduce the complexity of the protocol, but will reduce the size of a BGP packet substantially.

With this architecture, each node(router) inside an AS is represented as A.B.C. Each node may or may not be attached with a network which acts as a leaf node (i.e. a network will not act as a transit). In order to make use of user-id space properly and to support customer networks of heterogeneous sizes, the user-ID space needs to be divided as subnet-ID and user-ID. Profoundly, a VLSM (variable length subnet mask) type of approach has to be adopted at each node of an AS. So, each node of the AS layer will act as the root of a tree whose leaves are independent small customer networks which will act as stub. As the routing information of inter-AS layer as well as AS

layer need not be passed inside any node of the VLSM tree, each router inside the tree should maintain default route for any address outside of its network. With this approach, load on each router of the service providers will become negligible. Protocols that supports VLSM with MPLS/VPN has to be implemented inside the tree (inside the VLSM tree, all the physical ports of a switch have to be configured with the subnet mask. So, mere MPLS on top of static routing table should do the rest).

The fundamental assumptions based on which this architecture lies can be summarized as follows:

i) Entire network can be viewed as a network of regions or states where each region or state can have its own identity by communicating with the rest of the world through some state border routers. Each region or state is a network of Autonomous Systems. Each region as well as each Autonomous System inside them will have a fixed (maximum) length of prefix.

ii) Availability of hardware resources is such that flat address space can be maintained at the inter-AS layer.

Introduction of mesh-structured hierarchy will have several advantages:

- o Load at each router will get reduced substantially.
- o Concept of CIDR style approach and complexity related to prefix reduction can be easily avoided.
- o Mesh structured hierarchy will make traffic evenly distributed.
- o Physical cable connection can be optimized.
- o Administrative issues will become easier.

3.2. Determination of prefix lengths

With this architecture, IP address can be described as A.B.C.D where the D part represents the user id. Each router in the inter-AS layer will have two tables of information, one for the inter-AS-top and another for the inter-AS-bottom of the inter-AS-top node that it belongs to. Whereas, each node of the AS layer will have three tables of routing entries; one for the inter-AS-top, one for the inter-AS-bottom and another for the routing information inside the Autonomous System itself. In the worst case. a node inside an AS needs to maintain nMaxInterASTopNodes + nMaxInterASBottomNodes + nMaxASNodes entries in its routing table.

The dynamic nature of allocating an area from a free pool of address space is more frequent at the AS layer than at the inter-AS-bottom layer. As OSPF supports all the features needed, it can be considered

as default choice in the AS layer. Existing implementation of OSPF (Version 2) supports subnetting, by which an entire area can be represented as a combination of network address and subnet mask. With this approach, entire routing table gets reduced substantially. With the removal of subnetting, all the nodes inside an area will have an entry inside the routing table (OSPF Version 1). So the deterministic factor is what is the maximum number of nodes inside an AS OSPF can support once subnetting support gets removed. So the prefix length of AS layer will be determined by this factor of OSPF.

With the introduction of hierarchy in the inter-AS layer, number of entries in the BGP routing table will get reduced substantially. Even if pA and pB both are selected as 16, number of routing entries come within the admissible range of existing BGP protocol. But, it is the responsibility of IANA to come out with a scheme how nMaxInterASTopNodes and nMaxInterASBottomNodes are to be selected. Each top level node will have nMaxInterASBottomNodes nodes. It will be a waste of address space if each country gets assigned a top level nodes (e.g. china has got a population of 1,306,313,800 people where as Vatican City has got only 920 according to a census of 2006). So a moderate value of nMaxInterASBottomNodes is desirable, with which larger countries will have a number of top level nodes. e.g. each state of USA can be assigned a top level node. With the introduction of area in the inter-AS-bottom layer, each top level node can be divided into number of areas of heterogeneous sizes. So, a group of neighboring countries with less population can share the address space of a top level node. Similarly, user-id space has to be decided based on the largest area VLSM tree should be spanned through. All these issues are completely geo political and have to be decided by IANA.

3.2.1. A pseudo optimal distribution of prefixes in a 64bit architecture

In order to have optimal use of cable connections, length of the VLSM tree is expected to be as short as possible. Also any single organization may prefer to have its user id space to be under the same network id. So, a 16bit user-id may become insufficient for places like large university campus, where as 32bit will become too large. Hence, 24bit user-id will be a moderate one which is the class A address space in IPv4 (also used as the space for private IP). As published in 1998 [6], OSPF can support an area with 1600 routers and 30K external LSAs. So, 11 bits are needed to support this space. With the assumption that OSPF can support much more address space with the advancement of hardware technology as well as to keep the space open for future expansions, 12 bits are assigned for the AS layer. 16 bits are assigned for the inter-AS-bottom layer. So, if on the average, 16bit equivalent space gets used within the user-id space (i.e. one out of 256) and 8bit equivalent nodes gets used inside an AS (16% of

1600), for a top level node (with 16bit equivalent AS nodes), it will generate 2^40 IP addresses, which will give 8629 IP addresses per person in Japan (with a population of 127417200; Japan is at the 10th position from the top in the population list of the world). So, even if all the countries with population less than or equal to Japan are assigned a top level node and all the provinces/states of countries with larger population are assigned a top level node each, total number of nodes will come well under 1024. If a number of neighboring countries with lesser population shares a top level node, total number of top level nodes will come down further. This suggests that 62 bit equivalent (10(pA)+16(pB)+12(pC)+24(user-id)) space will be good enough for unicast addresses. This distribution expects OSPF to support 65K (64K+1K) external LSAs.

64bit address space may be divided into two 63bit blocks as follows:

i. Global unicast addresses with the most significant bit set to 0. In order to separate out router address space from the host computers of customer networks, routers may be assigned a prefix 01 whereas the host computers will have prefix 00. With three-tier hierarchy, network ID is represented as A.B.C. Any router inside the VLSM tree including the root will have an address 01A.B.C.router-id. Where as a host interface inside a customer network will be represented as 00A.B.C.uid.

As the number of nodes representing routers in the provider network will be way too less than the user-id space for the customer networks, in order to keep more space for unicast addresses of customer networks as well as to keep the option open for future expansion, entire 63 bit address space with the MSB set to 0 has been assigned to customer networks for unicast addresses. So, the distribution will look like 10(pA)+17(pB)+12(pC)+24(user-id). Router address space will be assigned from the address space with the MSB set to 1.

One can think of a larger size for the VLSM tree. It has to be compensated with a smaller size for the inter-AS space. Say the distribution may look like 10(pA)+15(pB)+12(pC)+26(user-id). As the size of the user-id space (or the VLSM tree) is fixed, larger the size of the tree, larger will be the waste. This factor can be decided based on the data supplied (or suggested) by the service providers.

ii. Address space with the MSB set to 1 will be distributed within the rest. Each of them will have a fixed prefix which will be determined with the consultation with IANA. This distribution will be based on the requirements and the work that have already been done in connection to IPv6 along with the following requirements:

a) Router address space: Any node in the router address space will be designated with a prefix followed by A.B.C.router-id.

b) Address space for multicasting:

c) Address space for private IP: A 32 bit address space should be good enough for private IP.

d) Provider independent address space: This space will be used for the customers who would like to retain their number even after changing their providers. With this architecture, addressing is based on the routing topology i.e. all unicast addresses will be based on the provider assigned address space. So, each of these provider independent addresses has to be mapped with an address from the global unicast address space. Section 4 describes issues related to PI addressing and IP mobility in detail.

In order to provide support of IP mobility as well as provider independent addressing, each customer network has to be assigned some extra space along with their usual need. The actual amount of space to be reserved has to be determined by IANA.

3.2.2. Whether to go for a two-tier or three-tier hierarchy

Establishment of hierarchy in the inter-AS layer reduces the size of BGP entries to a great extent, but leads to an improper use of address space due to geo-political reason. If hierarchy in the inter-AS space gets removed, entire 26bit (10+16) space will be available for a single layer and use of inter-AS space will be true to its sense, but will increase external LSA (and/or number of entries in the BGP table) dramatically. So, it depends on to what extent OSPF can support external LSAs. BGP expects the packet length to be limited to 4096 bytes. BGP manages to make it work with this limitation with the concept of prefix reduction in the CIDR based environment. As the number of inter-AS nodes increases, BGP has to change this limit in order to make it work in flat address space. The alternate will be to divide the inter-AS space into number of areas as defined in section 2.1. The area border routers will advertise the aggregated information to the rest of the world. BGP may have to incorporate both the options at the same time. As the number of nodes in the inter-AS layer increases, in order to reduce the number of entries in the routing table, inter-AS space has to be split into two separate planes. So, two-tier hierarchy can be considered as an interim state to go for three-tier hierarchy. If it so happen that current available data is good enough to support the present need, it will be worth to look for to what extent it can support in the future. Assignment of inter-AS nodes in two-tier hierarchy should be based on the geographical distribution as if it is part of three-tier

hierarchy. Otherwise, introduction of three-tier hierarchy in the future will become another difficult task to go through. Based on the report of year 2011, BGP supports ~400,000 entries in the routing table. With this growing trend, BGP may have to change the limit of packet length even in a CIDR based environment. With the introduction of two-tier hierarchy, number of entries in the routing table will come down drastically and with the three-tier approach, it will come down further.

3.3. Issues related to Satellite communications

Establishment of hierarchy in the inter-AS layer expects the only way any two autonomous systems in two different top level nodes communicate is through their SBRs. If two autonomous systems inside the same top level node communicate through satellite, it will be considered as a direct link between them. Whenever autonomous system 'ASa' of top level node 'A' communicates with autonomous system 'ASb' of top level node 'B' through satellite, they have to go through their state border routers. i.e. satellite port inside 'A' that communicates with a satellite port inside 'B' will be considered as state border router. If multiple such ports exists inside node 'A', all of them will be equidistant from any port inside 'B'. Which expects any satellite port inside 'B' to have prior knowledge of list of autonomous systems that will be under the purview of any port inside 'A'. So, all the satellite ports of 'A' have to exchange such group of information with all the satellite ports of 'B' and vice versa. These group of autonomous systems can be considered as a cluster of autonomous systems inside an area of a top level node. If number of such ports is small, some heuristics can be applied while assigning AS numbers in order to reduce the processing time during the circuit establishment phase. It will become difficult to maintain such heuristics once the number of such ports becomes large. So, in case of satellite communication, the advantage of establishing hierarchy inside inter-AS layer diminishes as the number of satellite ports increases. If any private corporate maintains its own satellite channel to communicate between its offices at distant locations, all of these offices are going to be considered as under the user-id space of its network. Service providers that provide satellite services to the end-site customers, can operate in the usual manner as they will provide connection to customer networks which will act as stub.

4. Issues related to PI addressing and IP mobility

As far as implementation is concerned, provider independent addressing will be a costly affair. First of all in order to resolve the currently mapped location, there has to be a mechanism which is to some extent similar to the DNS entry resolution. Inside a customer

network which is based on the provider assigned address space, routing of IP packets will be based on the provider assigned addresses. So, for every IP packet that is destined to a PI address will have a stack of addresses; the mapped address (or the care-of address) and the PI address. While initiating communication with a PI address, the mapped address has to be resolved first and then both the PI address as well as the mapped address has to be passed down to the transport layer. Transport layer needs to form a stack of addresses while filling up the IP packet. The above complexities can be avoided if the entire customer network is assigned a contiguous set of PI addresses. So, for the entire system, provider independent addressing has to be supported either based on the individual customer basis or on the entire customer network basis but not both. Customers who would like to have mobility support, the mapped address can be considered as the "Home Address" of the mobile node as defined in the specification of "IP Mobility Support"[7]. Once a node with PI address moves to a co-located care-of address[7], system needs to make decision based on PI address, its mapped address as well as the co-located care-of address. So, provider independent address with mobility support will be the costliest operation.

Assignment of contiguous block of PI address space to an entire customer network apparently do not make much sense. This is just equivalent to assigning PA address space to a customer network. So, assignment of PI address space to an entire customer network has to be avoided unless there is a real need that can not be solved (or avoided) by using PA address space. PI address assignment always have to be burdened with the look up procedure to resolve the mapped address even if an entire customer network gets assigned PI addresses.

Assignment of PI addresses has to be restricted to a limited number of users. This limit has to be decided by IANA. As the number of users with PI addresses increases, complexities within the entire system increases proportionately.

4.1 IP address aliasing

An interface of a customer network may have several IP addresses (e.g. for a multihomed customer site, each interface will have multiple global unicast addresses also it may have a private address). This phenomenon is commonly known as IP address aliasing.

A second type of aliasing is required to support IP mobility and provider independent addressing. For a mobile node that has been moved to a customer network which get services from two service providers and maintains private IP addresses, will have at least four IP addresses; provider one assigned unicast address, provider two

assigned unicast address, private address and its permanent "Home Address". The "Home Address" will be aliased with one of the provider assigned addresses (i.e. the co-located care-of address). Similarly for a node with provider independent address will have four IP addresses. The interface address holding the PI address will be aliased with one of the provider assigned addresses as its mapped address. If the node with PI address moved to a foreign site, will have a care-of address. The mapped address will be treated as the "Home Address". So the interface structure needs to have two additional fields to hold the values of care-of address and mapped address. The PCB structure will have two additional fields 'lmpiaddr' and 'lcladdr' to hold these information. In case a PI node that has not been moved, both 'lcladdr' and 'lmpiaddr' will have the same value. So 'lcladdr' will have the current provider assigned address that a foreign node needs to use for communication. The field 'laddr' that is used to hold the vale of local address will hold the value of PI address for a node with PI address; it will hold the value of "Home Address" of a mobile node in case it does not have a PI address.

In order to support multihoming, an outgoing IP packet needs to be forwarded based on its source address [8]. In order to support this, an outgoing packet from a mobile node or a node with PI address needs to be stacked with the associated care-of address. A client application program needs to call 'getsrcaddr'[8] to get the source address based on the destination address. The client program needs to bind this address before communicating with its peer. The 'bind' system call needs to go through the interface list and fetch the associated structure to check whether the source address is aliased or not and needs to fill the values of 'lcladdr' and 'lmpiaddr' of PCB accordingly. Protocol output routines like 'tcp_output' and 'udp_output' need this information while filling up the IP packet.

Similarly, PCB needs to introduce two more fields 'fmpiaddr' and 'fcladdr' to support the destination address to be provider independent and/or mobile. If foreign address is stationary and provider independent, both 'fmpiaddr' and 'fcladdr' will have the same value. The existing field 'faddr' which is used to address a foreign address will hold the value of PI address for a node with PI address. Similarly it will hold the value of "Home Address" of the mobile node if it is not provider independent.

If destination address is provider independent, client applications need to resolve the mapped address before communicating with its peer. There could be several approaches to resolve the mapped address for a PI address. This issue needs to be discussed in a separate document. A new system level routine needs to be introduced to get the mapped address.

```
struct in_addr getmappedaddr(struct in_addr *piaddr);
```

Once mapped address is resolved, it needs to be registered with the PCB.

A client application needs to call 'getsrcaddr'[8] to get the source address based on the destination address before communicating with its peer. If destination address is provider independent, 'getsrcaddr' needs to call 'getmappedaddr' to resolve the mapped address and register it with the PCB at the beginning. It needs to resolve source address based on the destination address afterwords.

IP address stacking can be performed with the approach introduced in section 6.4 of RFC6275[9]. RFC6275 talks about the stacking of IP addresses for a destination address (Let us call it as type 0 stacking). Two more types of stacking need to be introduced; type 1 stacking where only source address will appear in the stack and type 2 stacking where both source address and destination address will appear in the stack with a particular type of ordering.

Protocol output routine like 'tcp_output' or 'udp_output' needs to fill the IP packet in the following manner.

If the socket contains a valid 'lcladdr', use 'lcladdr' as the source address and 'laddr' will appear in the stack. If the socket contains a valid 'fcladdr' use 'fcladdr' as the destination address and 'faddr' will appear in the stack. If only 'fcladdr' contains a valid address where as 'lcladdr' is NULL, use type 0 stacking. If only 'lcladdr' contains a valid address where as 'fcladdr' is set as NULL, use type 1 stacking. If both 'lcladdr' and 'fcladdr' contains valid addresses, use type 2 stacking.

Protocol input routine like 'tcp_input' or 'udp_input' needs to process the packet in the reverse order based on the type of stacking. For type 0 stacking, use the address in the stack as the destination address; for type 1 stacking, use the address in the stack as the source address; for type 2 stacking use both source address and destination address from the stack.

When TCP receives a SYN for connection establishment, it allocates a PCB and assigns the values for 'laddr', and related fields. During this phase, TCP also needs to check whether the local address is aliased or not and needs to fill the values of 'lcladdr' and 'lmpiaddr' accordingly. Similarly if destination address is found to be aliased, based on the stacking type, it needs to fill up the field 'fcladdr'.

4.2. Changes expected with the specifications related to IP mobility

RFC6275 demands correspondent node binding from mobile nodes for route optimization. This binding is required when a connection gets established as well as when the mobile node changes it address space. There are application like HTTP which opens up multiple connections on the run time which are very short lived. If mobile nodes need to send binding messages for all the connections, network will be unnecessarily congested. This congestion can be avoided with the establishment of binding at the time of connection establishment itself. So, if TCP server happens to be mobile, it will set the value of 'lcladdr' in the stack while sending SYN+ACK. TCP client which initiates communication through 'connect' needs to set 'fcladdr' field on receiving TCP+ACK. With this approach correspondent node binding messages need to be sent only when a mobile node changes its position from one address space to another.

Route optimization is not applicable to applications which are of multicast type. In these cases packets need to be forwarded with the mechanism of reverse tunneling with the approach of "IP Encapsulation within IP" as defined in RFC 2003. In order to support packet delivery with route optimization method as well as with "Encapsulating Delivery Style" based on the application type the protocol control block needs to introduce another field 'hagentaddr' to hold the address of the home agent of the mobile node. The interface structure also needs to have same field. The 'bind' system call needs to go through the interface list to fetch 'hagentaddr' to the PCB along with 'lcladdr' and 'lmpiaddr' as described earlier. So, protocol output routines like 'tcp_output', 'udp_output' need to fill up the packets based on the application type. In "Encapsulating Delivery Style" packets need to be formed in the following manner.

```
The inner IP header will contain
   if ('lmpiaddr' == NULL && 'fcladdr' == NULL) {
      Source Address: Home address of the mobile node (i.e. 'laddr')
      Destination address: Address of the correspondent node (i.e.
      'faddr')
   } else {
      If ('lmpiaddr' != NULL), use 'lmpiaddr' as the source address
      and 'laddr' will appear in the stack.  If ('fcladdr' != NULL),
      use 'fcladdr' as the destination address and 'faddr' will
      appear in the stack.

      If ('fcladdr' != NULL && 'lmpiaddr' == NULL), use type 0
      stacking.  If ('lmpiaddr' != NULL && 'fcladdr' == NULL), use
      type 1 stacking.  If ('both 'lmpiaddr' != NULL && 'fcladdr' !=
      NULL), use type 2 stacking.
   }
```

The outer IP header will contain
Source Address: co-located care of address of the mobile node (i.e. 'lcladdr')
Destination Address: Address of the home agent of the mobile node (i.e. 'hagentaddr')
Protocol field: IP in IP

5. Refinements over existing IPv6 specification

As IPv6 was envisioned long before some of the newer technologies e.g. MPLS came into picture, some refinements can be made over the existing specification. These considerations are related to bandwidth usages and performance inside switches. Experimental results show that smaller packet size gives better result for the processing of RT packets. So, it is desirable to have IP packet header to be as small as possible.

As described earlier, evaluation of the parameters nMaxInterASTopNodes, nMaxInterASBottomNodes and nMaxASNodes is geo-political and have to be decided by IANA. Once these parameters are determined with mutual agreements, values of pA, pB, pC and prefix length of user id can be determined. With 64bit address space, IP header will be reduced by 16 bytes.

The 'flow label' field of IPv6 packet header may not be of any use with MPLS is in use. ATM used to have 4 priority classes. The first specification of IPv6 RFC-1883 used a 4bit type of service field along with a 24bits flow label field. These two were modified to a 8bit type of service field and a 20bit flow label field in the current spec RFC-2460. Too many priority classes may increase complexities to process inside switches. If type of service field of IPv6 header may be reduced to be of 4bit length as it was stated in RFC-1883 and 'flow label' field gets removed, another three bytes may be reduced from the IPv6 header.

The field 'Hop Limit' has got a 8bit value in the existing spec. The role of this field needs to be discussed properly with a large address space.

RFC4862[10] introduces the concept of "Stateless auto configuration" with the goal in mind that no manual configuration is required by individual machines before connecting them to the network. It generates a link local address with a link-local prefix and the link address (e.g. Ethernet/E.164 for ISDN) first. This link local address is used to configure global unicast address and any other configurable parameters based on router advertisement. Global unicast addresses are generated by the prefix supplied by the router advertisement and the link specific interface identifier. This

identifier can be as large as 64 bit length. So irrespective of the size of the network (it may be 10000 or 100 or even less than that) every customer network will consume a 64bit equivalent addresses. This seems to be a huge blunder. What is expected is the length of the interface identifier is equivalent to support the number of nodes supported by that subnet. In order to achieve this the router itself or a server in that subnet needs to maintain a storage which will generate the interface identifier based on the request from individual hosts. It may be desirable that interface identifiers are generated from DHCP servers. With the option of generating interface identifier through DHCP, changes in the auto configuration process can be looked at as follows:

From the point of view of a host, it can be considered as a two step process. Host needs to send Router Solicitations message to find out the presence of a router. Router Advertisement message should include an option field which will inform whether prefix information should be configured through Router Advertisement or through DHCP. Host needs to send a request message to get the interface identifier. If both the information needs to be obtained from a DHCP server they can be obtained through a single message.

From the server's point of view, it needs to maintain a database for a mapping of the link-layer address and subnet specific interface identifier. Lifetime of an interface identifier has to be processed in the usual manner the way existing DHCP implementation treats IP addresses.

There seem to be another possible danger to obtain prefix information through Router Advertisement. As the Router Advertisement comes in the form of ICMP messages, once it is received by the ICMP layer, it looses information from which interface the message has been received (This problem arises for hosts that are having multiple interfaces and not all of them are attached to the same subnet). So, auto configuration of a host has to be performed one interface at a time by making all other interfaces disabled. Once configuration of all the interfaces are done, all of them have to be enabled.

If it is expected that hosts should reconfigure their addresses dynamically based on Router Advertisement message, Router Advertisement needs to generate a special message for a certain amount of time that needs to include old prefix and the corresponding new prefix in the message.

In order to support multihoming[8], prefix information needs to include the fields 'default router' and 'next hop address' to reach the default router for each of the prefixes.

In a 64bit architecture, link-local address can be formed with a link-local prefix and link-layer address in a suitable manner; say it can be formed with a 16bit link-local prefix followed by a 48bit link-layer address. For hardware that supports more than 48bit addressing (say E.164), the least significant 48bits may be considered to generate link-local addresses.

6. Distributed processing and Multicasting

With the inherent hierarchy involved in this architecture, distributed applications can also be structured in a suitable manner. Say, for a commonly used web based application a master level server will be there at every top level node. Any change that might happen in the application, has to be synchronized within these master level servers first. There might be servers at the middle layer (inside each inter-AS-bottom) inside each top level node. Once the changes get reflected at the master node, all the servers at the middle layer needs to update themselves with their master level node. This will reduce network traffic substantially. Inherent hierarchy in the architecture will also help establishing multicast tree in the similar manner. Work on these issues can be progressed only after this architecture gets approved.

7. Transition to real IP from private IP

Both CIDR based hierarchy and Mesh structured hierarchy expects a VLSM tree at the bottom. In VLSM, in real IP space with provider assigned (PA) addresses, assignment of network resources has to be associated with the address space to be used with the type of service. Within a typical switch supporting multiple types of ports, a line card of strength OC48 can be replaced with 4 line cards of strength OC12. An OC12 card may also be replaced with 4 OC3 cards. An OC12 card may be attached to another switch with DS3 ports and so on. When it reaches to the customer network port density of a switch has to be directly proportional to the address block that a customer network will be assigned to. i.e. each customer network has to be assigned a block of address space (say, 128, 256, 512, 1K, 2K etc). Within the switch these ports have to be assigned net address/net mask the way VLSM works.

In IPv4 environment, providers have provided services in terms of bandwidth of the ports say, 2 Mbps/4 Mbps/1 Gbps line etc. If these ports were assigned addresses based on the number of users of the customer network, transition from private IP to real IP is simple. Consider a switch that has supplied 2 Mbps line to a set of customers with number of users within 1K to 2k, each of them will be assigned a block of 2K each. But if number of users are not proportional to the bandwidth used, say same 2 Mbps line were used to customers of sizes

1K, 2K 10K and 16K respectively reorganization will be needed if possible. This rearrangement may be possible within the switch itself or by connecting ports of appropriate sizes from different switch, otherwise each of them has to be assigned an address block of 16K each creating a block of unused address space. So, address block assignment in the VLSM tree has to grow in a bottom up approach.

Thus transition of existing provider network without reorganization to a real IP space with CIDR based approach is apparently not a difficult job. In a CIDR based approach, sizes of the VLSM trees are heterogeneous that leads to number of routing entries to be very high. Mesh structured hierarchy is convenient to reduce the routing overhead as well as for distribution of network resources in a suitable manner in the long run. To covert CIDR based approach to Mesh structured hierarchy requires reorganization mainly in the routing domain and by splitting trees of very large sizes (>24 bit address space) at the top.

Section 3.2.1 reveals that in Mesh structured hierarchy a 64bit architecture will be good enough for our need in a provider assigned (PA) address space; the same is true for CIDR based approach as well.

8. IANA Consideration

This is a first level draft for proposed standard. Hence, IANA actions should come into play at a later stage, if needed.

9. Security Consideration

This document does not include any security related issues.

10. Acknowledgments

The author would like to thank to Professor Amitava Datta of University of Western Australia for his review and constructive comments.

11. Normative References

[1] Nordmark, E. and R. Gilligan, "Basic Transition Mechanisms for IPv6 Hosts and Routers", RFC 4213, October 2005.

[2] Fuller V., Li. T., "Classless Inter-Domain Routing (CIDR): The Internet Address Assignment and Aggregation Plan", RFC 4632, August 2006.

[3] Huston, G., "Commentary on Inter-Domain Routing in the Internet", RFC 3221, December 2001.

[4] Q. Vohra, E. Chen., "BGP Support for Four-octet AS Number Space", RFC 4893, May 2007.

[5] Srisuresh, P. and K. Egevang, "Traditional IP Network Address Translator (Traditional NAT)", RFC 3022, January 2001.

[6] J. Moy., OSPF Standardization Report, RFC 2329, April 1998

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

[8] S. Bandyopadhyay, "Solution for Site Multihoming in a Real IP Environment", <draft-shyam-site-multi-38> work in progress.

[9] C. Perkins, Ed., D. Johnson, J. Arkko, "Mobility Support in IPv6" RFC 6275, July 2011.

[10] S. Thomson, T. Narten, T. Jinmei, "IPv6 Stateless Address Autoconfiguration", RFC 4862, September 2007.

12. Informative References

[11] Postel, J., "Internet Protocol", STD 5, RFC 791, September 1981.

[12] Rekhter, Y., and T., Li, "A Border Gateway Protocol 4 (BGP-4)",RFC 1771, March 1995.

[13] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification, RFC 1883, December 1995.

[14] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.

[15] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.

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

13. Author's Address

Shyamaprasad Bandyopadhyay
HL No 205/157/7, Kharagpur 721305, India
Phone: +91 3222 225137
e-mail: shyamb66@gmail.com