Route Scaling and Multihoming

Martin Mc Court

Masters of Science in Computing 2006

Institute of Technology, Sligo

Contents

Route Scaling and Multihoming ..... **i**
Contents ..... **i**
List of Figures ..... **iv**
List of Tables ..... **vi**
Abstract **1**
Acknowledgment **2**
Chapter 1 – Introduction **3**
Chapter 2- IPv6 Format **10**
Field Descriptions 11
Version 11
Traffic Class 11
Flow Label 11
Payload Length 12
Next Header 13
Hop Limit 13
Source Address 13
Destination Address 13
Extension Headers 14
Hop-by-Hop Options Header 16
Routing Header 16
Fragment Header 17
Destination Options Header 18
IPv6 Addressing **20**
IPv6 Address Types 20
Address Notation 21
Text Representation of Address Prefix 22
Format Prefixes 22
IPv6 Address Allocation 25
Recommendations on IPv6 Address Allocations to Sites 25
Address Assignment examples 26

Chapter 3 – Border Gateway Protocol **27**
Local Preference 29
Multi-Exit Discriminator (MED) 29
Origin Attribute 30
AS Path Attribute 30
Next Hop Attribute. 30
Community Attribute 31
Chapter 4 - Route Scaling Techniques in IPv4 ..... **32**
Route Scaling 32
Classless Interdomain Routing (CIDR) 33
Route Aggregation 34
Multihoming 35
Existing Multihoming Techniques 36
Auto Route Injection BGP-4 [RFC2260] 36
Non Direct eBGP Peering BGP-4 [RFC2260] 37
Solution 3 Provider Independent Addressing [RFC1518] 38
Solution 4 Prefix Filtering [RFC1518] 38
Address Space Fragmentation 41
Chapter 5 - Address Scaling Techniques in IPv4 ..... **43**
Network Address Translation 43
Chapter 6 - Multihoming with NAT **46**
DNS with Bidirectional NAT 47
Externally Originated Connection 47
Conclusion 49
Chapter 7 - Financial Incentives for Route Aggregation **51**
Chapter 8 - Multihoming in IPv6 ..... **54**
Chapter 9 - Comparison of Multihoming Techniques **58**
Auto Route Injection 58
Non Direct Peering 60
Network Address Translation 61
Financial Incentives for Route Aggregation 62
Multihoming in IPv6 63
Conclusion 65

Bibliography **66**

List of Figures

Figure 1 Active BGP Entries (www.potaroo.net) 5
Figure 2 Active BGP Entries 2004-2005 (www.potaroo.net) 5
Figure 3 Average Span of BGP Advertisements (www.potaroo.net) 7
Figure 4 Specific Prefix Advertisements (www.potaroo.net) 8
Figure 5 IPv6 Header format 10
Figure 6 TLV Format 14
Figure 7 Use of Extension headers 15
Figure 8 Hop-by-Hop Options header 16
Figure 9 Routing header 17
Figure 10 Fragment Header 18
Figure 11 Destination Options Header 18
Figure 12 IPv6 address format 20
Figure 13 Link & Site local address format 23
Figure 14 Global Unicast address format 24
Figure 15 Address allocation hierarchy [RIPE] 25
Figure 16 Internet Routing 27
Figure 17 Internet Backbone 28
Figure 18 Local Preference 29
Figure 19 AS Path Attribute 30
Figure 20 Next Hop Attribute 31
Figure 21 Active BGP entries Aug 2004 [potaroo.net] 32
Figure 22 Route Aggregation 34
Figure 23 Multihoming 35
Figure 24 Internal Peering 36
Figure 25 Indirect Peering 37
Figure 26 Portion of BGP routing table 41
Figure 27 Network Address Translation 44
Figure 28 NAT Table 45
Figure 29 Bidirectional NAT 46
Figure 30 Multihoming with NAT and DNS 48
Figure 31 Dissemination of Routes 52

Figure 32 Auto Route Injection.....58
Figure 33 BGP Link Failure.....59

List of Tables

Table 1 Next Header values 12
Table 2 Assigned prefixes 22
Table 3 Reserved Addresses 38
Table 4 RIPE Guidelines for Filtered Prefixes 39

Abstract

The hyperexponential growth of Internet routing tables will soon exceed the capabilities of existing hardware adding to increased costs and significant degradation in routing performance.

Acknowledgment

I would like to thank Aidan Mc Donald, Cork Institute of Technology, for all his assistance and advice in preparing this dissertation.

Chapter 1 – Introduction

Papers suggest that the primary reason for designing IPv6 was because the current version of IP was beginning to exhaust its address space, but is the address limitation of IPv4 really an issue any more? With the introduction of Network Address Translation and other life extending techniques, the demand for legitimate IP addresses has been significantly reduced.

Both LAN and WAN technologies have changed dramatically since the introduction of IPv4 over twenty years ago and as a result so too have applications. We now want to deliver voice and video and other real time applications across the Internet. These applications require and demand a certain quality of service (QOS). This was to be one of the built-in features of IPv6. QOS, however, is also natively supported by ATM, while on a legacy LAN QOS is supported by the RSVP protocol. Likewise, mobile IP is supported in both IPv4 and IPv6. Has IPv6 really got anything to offer when compared to the "workaround" solutions and protocols that can be used with IPv4?

Many articles, both subjective and objective, have been written comparing the two protocols. Huston (2003) presents a rather interesting article that compares IPv6 features to those of IPv4. There is, however, no doubt that an altruistic approach by everyone to embrace IPv6 would be immensely beneficial to the entire Internet community. Undoubtedly many will also feel that the upheaval involved in upgrading to IPv6 would not be justified. But has IPv6 got much more to offer other than an abundance of addresses, QOS and mobility?

For as long as I can remember the Internet has always being running out of address space and absolute address depletion was only around the corner. No matter when a paper was published, or what new procedures, policies and protocols were in place, complete address depletion always seemed immanent. It's a bit like the proverbial frog always jumping half of his remaining distance and never actually getting there. So is the IPv4 address depletion just like our frog? In many respects IPv4 addresses are just like our frog and we will never actually run out of addresses and the sky will never fall in on the Internet from this regard. Although we will never be able to tell for certain, it does seem that the predictions about

eventual address depletion were probably correct. We have never run out of address space simply because plenty was done about it in a timely manner.

In the early nineties Internet scaling issues were studied and RFC 1287 was produced in December 1991. This RFC identified four major issues regarding scalability. These were:

- *The Internet will run out of certain classes of IP network addresses, e.g., B addresses.*
- *The Internet will run out of the 32-bit IP address space altogether, as the space is currently subdivided and managed.*
- *The total number of IP network numbers will grow to the point where reasonable routing algorithms will not be able to perform routing based upon network numbers.*
- *There will be a need for more than one route from a source to a destination, to permit variation in TOS and policy conformance.*

A number of solutions were produced on foot of this RFC. NAT was introduced as a short term solution to eventual address depletion. Classless Interdomain Domain Routing (CIDR) and BGP4 were introduced to curb the excessive growth in advertised routes. A new IP protocol IPv6 was also produced that could "apparently" solve all the critical issues raised in RFC 1287, while, at the same time introducing many new enhancements. It will be shown later, however, that IPv6 must have left the party early as we are still left with the problem of exponential growth of advertised routes on the Internet, although this time the cause of the problem is quite different to what it was in the nineties.

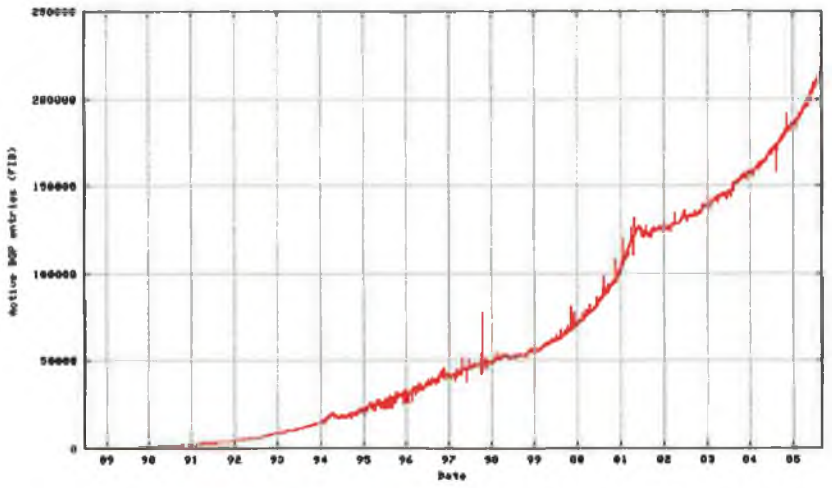

Figure 1 Active BGP Entries (www.potaroo.net)

Figure 1 above shows a graph of active BGP entries plotted against time. This graph is taken from data collected on the BGP routers in AS1221. BGP data is collected on an hourly basis and is publicly available on the website www.potaroo.net. As can be seen from the graph the number of advertised routes is growing at an alarming rate. Reference to figure 2 below puts this exponential growth into perspective by showing a plot of the number of increased routes in the past twelve months (27 August 2004 to 27 August 2005).

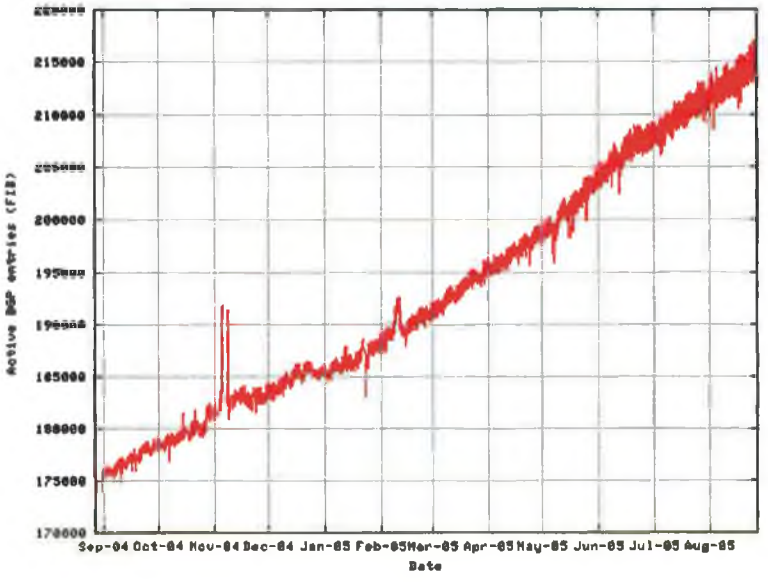

Figure 2 Active BGP Entries 2004-2005 (www.potaroo.net)

As can be seen, in the past twelve months alone, the number of routes advertised has increased from about 175,000 to over 215,000.

So what impact will this growth have on us if left unabated? Does it mean that we are running out of address space yet again? The growth in routing table size can only be loosely linked to the increased consumption of IP addresses as more and more people and businesses "get connected". The real problems here are manifold.

In the first place, can routing hardware keep up the pace? Li [1] predicts that the "hyperexponential growth of routing tables will eventually outgrow Moore's law". Moore's law, although 40 years old still holds true today even for routers. Loosely put, Moore's law states that hardware performance will double about every eighteen months. If this hyperexponential growth continues not only will costs be driven up, but the impact on performance will also be immense. For example, how long will it take for routers to converge? How long will it take for a router to index through its routing tables for a particular network? This can only be bad news as more and more users embrace VOIP and other such delay sensitive applications. Even most home users now pump data into the Internet at DSL rates making real time multimedia applications a normal occurrence on most domestic computers.

And what about IPv6 with its massive address space? Items like domestic appliances can now be connected to the Internet and controlled remotely from a web browser. With an almost limitless amount of available addresses in IPv6, will this new protocol be part of the problem or part of the solution?

But what exactly is causing the exponential growth of BGP routes? Was CIDR not introduced as part of the solution to curb such growth? Referring back to figure 1 we can see that CIDR had a very positive impact on growth rates between 1994 and 1998. A straight linear increase can be seen between these years. However, from 1998 onward the exponential increase resumes yet again.

As mentioned previously, a small part of the problem is the increased uptake in Internet connections by end users and, no doubt, the associated increase in the number of ISPs. The

Potaroo website [2], makes available all sorts of data pertaining to BGP statistics. Figure 3 shows a very interesting graph that plots the average number of addresses advertised by each BGP announcement. As is clearly evident from the graph, the address span for each advertisement continues to decrease. For example in 2000, on average, each BGP advertisement represented just over 15,000 addresses. Today each BGP advertisement on average represents less than 7000 addresses. One can easily infer from this that an increase in address consumption will cause an increase in BGP advertisements beyond the control of CIDR route aggregation. As will be explained in detail later, the main benefit of CIDR is to allow the aggregation of subnets into just one route advertisement. So, for example, all subnets beginning with 193.1.x.x will be advertised as just one route under network 193.1.0.0. Therefore, one would expect the opposite of what is shown in figure 2, that is, a larger address span meaning that each route advertisement represents a larger number of addresses.

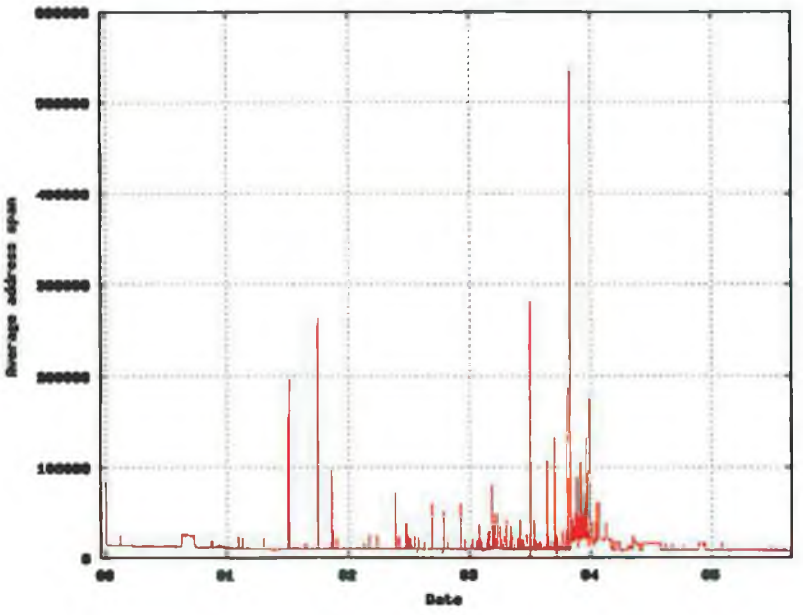

Figure 3 Average Span of BGP Advertisements (www.potaroo.net)

Evidently this is not the case. Each route advertisement represents fewer and fewer addresses which means that much more specific routes (networks with a longer subnet prefix) are now being advertised. Apart from increased table size, the impact of such specific advertisements is an increase in "the rate of dynamic path recomputations that occur in the wake of

announcements and withdrawals" as these more specific networks go up and down, Huston (2001). This again begs the question as to whether the Internet hardware will be able to support this increased strain.

In an article written by Borthic (2001), the author states that we were quite fortunate that the Dot Com bubble burst when it did, as the Internet infrastructure simply just couldn't have handled such explosive growth.

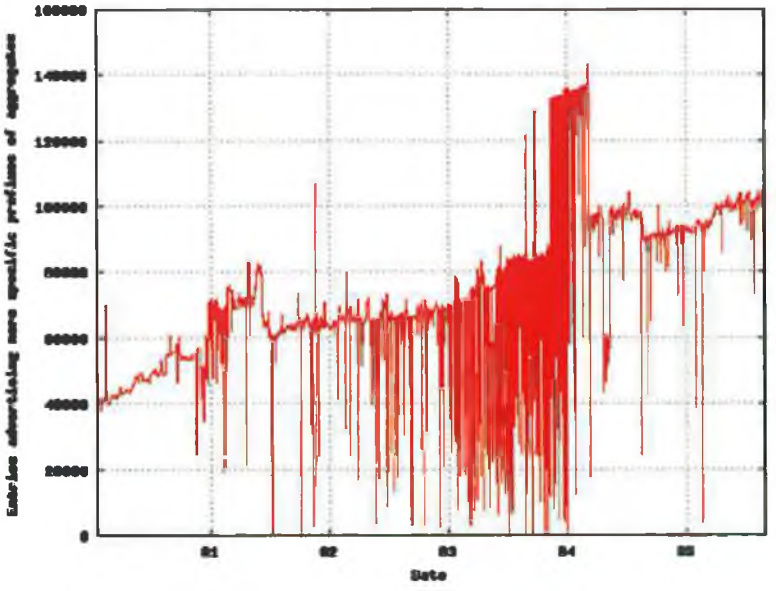

Figure 4 Specific Prefix Advertisements (www.potaroo.net)

Looking at figure 4 above we can see the exponential growth of specific prefix advertisements. What this graph shows is that at present there are almost 140,000 specific network advertisements that could be, but are not, being advertised under their own larger aggregate address.

So what are the reasons for specific advertisements, and why can't CIDR aggregate these smaller specific advertisements into significantly fewer but larger aggregate route advertisements?

There are many causes for specific advertisements. For example, if a business connects to an ISP and is allocated a block of addresses under the ISP's own aggregate address block, then in the event that the business should move to another ISP, this address block would move with the business. This would leave a hole in the original ISP's addresses and would also cause the new ISP to advertise an address block not within its own aggregate address range.

This type of address is referred to as a Provider Independent Address (PIA). There are two main advantages for subscribers using PIA. In the first place there is no need for site renumbering, the benefits of this are obvious. Secondly, because the address block is not part of the ISP's address block, then it must be advertised as a specific route. The advantage of advertising a specific route is that routers will always select this path in preference to a path that advertises an aggregate announcement.

The main reason, however, for more specific advertisements is the use of multihoming. Multihoming is the practice of networks connecting to more than one ISP. The main advantages of multihoming are resilience, load balancing and traffic engineering. When a site is multihomed, for instance, to two ISPs, it will have a separate address block allocated by each ISP. However, in the event that a connection to one of the ISPs should fail, then the other ISP will have to advertise the address block belonging to the failed connection. This address block will essentially punch a hole in the ISP's routing advertisement since again it is not part of the ISP's own address space and therefore can not be aggregated. This is looked at in greater detail in subsequent chapters.

This dissertation therefore sets out to examine route scaling techniques that can and are being used to curb the growth of route advertisements in BGP tables. Since multihoming has been identified as the main cause for the exponential growth of more specific route advertisements, the main focus will be on multihoming techniques. Given that, however, IPv6 was developed mainly out of address and route scaling necessity, a particular emphasis will be placed on IPv6.

Chapter 2- IPv6 Format

In order to fully understand and appreciate the arguments set out in this dissertation, it is essential to first look at the detail of the IPv6 protocol. IPv6 was designed as an evolutionary step from IPv4 as opposed to a complete upgrade and revamp. The header itself is actually a streamlined version of the IPv4 header and is specified in RFC 2460.

The frame format is shown below in figure 5.

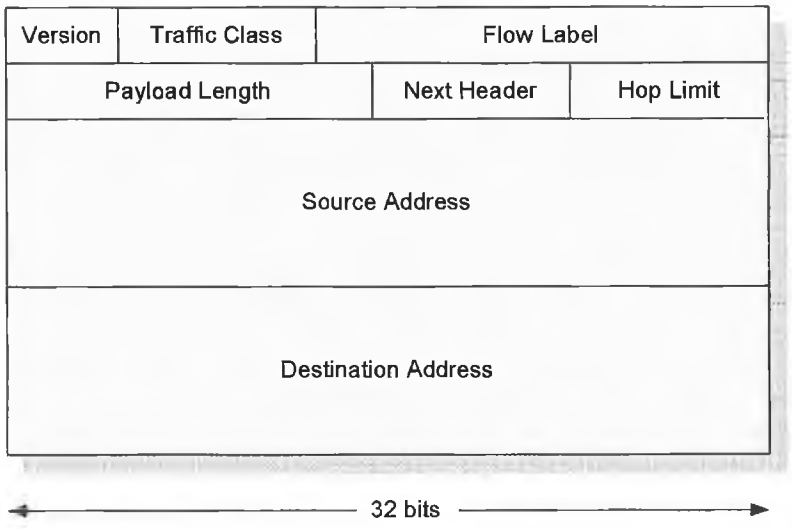

Figure 5 IPv6 Header format

From looking at figure 5, the most obvious changes are the size of the addresses, the removal of the fragment fields and the introduction of a flow label.

In IPv4 the minimum header length is 20 bytes but can be extended up to 60 bytes by adding on options in 32 bit quantities. IPv6, on the other hand, has a fixed length header of 40 bytes. Even though this is twice as long as an IPv4 header, 32 of these bytes are for source and destination addresses, leaving only 8 bytes for general header information.

With IPv4 the header length is variable and is used to indicate whether or not options have been included. The header length cannot exceed 60 bytes in total. This in itself limits the development of new protocols. With IPv6, on the other hand, an endless amount of new protocols can be catered for through the use of the Next Header field which is explained later.

The designers of IPv6 were able to remove the fragment fields, namely ID, FLAGS, and OFFSET because with IPv6 only the end nodes can actually fragment the datagram. Enroute routers can not fragment an IPv6 datagram as in IPv4. Hosts must determine the path MTU through a procedure known as Path MTU Discovery. This procedure is quite straight forward in that a host sends a large datagram to a remote host. Since routers can not fragment an IPv6 datagram, the first router along the path that discovers fragmentation is required will send an ICMP error message back to the source indicating that fragmentation is required. If a host needs to fragment a datagram, the "Next Header" field is used to indicate this; as will be described later. In IPv4, if the datagram is not fragmented, the fragment fields still need to be included and processed.

Field Descriptions

Version

The version field is functionally identical to that used in IPv4. This 4-bit field indicates the version number of the IP datagram and will contain a value of 6_{10}.

Traffic Class

This 1-byte field supersedes IPv4's Type of Service field. This field is used for Differentiated Services (Diffserv) on the Internet and is used to give priority to certain types of data. It allows for the handling of real time data and is described in RFC 2474. This field is the same as the newer DS field used in IPv4.

Flow Label

This 20-bit field allows for the labelling of datagrams that are required to be handled in the same way. This means that routers along the path don't have to examine the headers of subsequent packets belonging to the same flow. A flow is considered to be the labelling of packets that require non default quality of service and special handling.

Payload Length

This 2-byte field indicates the length of the payload in bytes. Because this field is 16 bits long, the maximum payload is 64 KB. Unlike IPv4, however, the payload length does not include the header but does include extension headers. The payload therefore begins immediately after the destination address.

Value	Description
0	Hop by Hop option
1	ICMPv4 support
2	IGMPv4 support
4	IP in IP
6	TCP
8	EGP
9	IGP
17	UDP
41	IPv6
43	Routing header
44	Fragmentation header
45	IDRP
46	RSVP
50	Encrypted security payload header
51	Authentication header
58	ICMPv6
59	No Next Header for IPv6
60	Destination Options header
88	EIGRP
89	OSPF
108	IP payload compression protocol
115	Layer 2 tunnelling Protocol
132	Stream Control Transmission Protocol
134-254	Unassigned
255	Reserved

Table 1 Next Header values

Next Header

This field indicates the payload type. If the IPv6 datagram is carrying a TCP or UDP payload, the next header field will contain the value 6 or 17 respectively just like IPv4. If options are present in the header, this will be indicated with an appropriate value in the next header field. These values are listed in Table 1 above.

Hop Limit

This 1-byte field is similar to the Time-to-Live (TTL) field in IPv4. The difference, however, is that this field contains the number of hops remaining for this datagram and not the time left in seconds. Each router along the path will decrement this value by one.

Source Address

This 16-byte field contains the IP address of the source host.

Destination Address

This 16-byte field will normally contain the destination IP address of the remote host with the exception of the presence of a Routing Extension Header, in which case it will contain the address of the next hop.

Extension Headers

When dealing with IPv4, the presence of an options field is indicated by the value in the header length field. With IPv6 the Next Header field is used to indicate the presence of options. If no options are included then the Next Header field will normally contain the number 6 for TCP, 17 for UDP, or any other value just as with IPv4. These payloads will immediately follow the destination address. The current IPv6 specification (RFC 2460) defines the following six extension headers.

- Hop-by- Hop Options Header
- Routing Header
- Fragment Header
- Destination Options Header
- Authentication Header
- Encrypted Security Payload Header

The Hop-by Hop options header and the Destination options header carry a variable number of type-length-value (TLV) encoded options, which adhere to the following format.

Figure 6 TLV Format

Option Type:	8-bit identifier of the type of option
Option Data Length:	8-bit unsigned integer, indicating the length of the data field in octets.
Option Data:	Variable length field and carrying type specific data.

There can be zero, one, or more extension headers, with each extension header being identified by the Next Header field in the preceding header as shown in Figure 7.

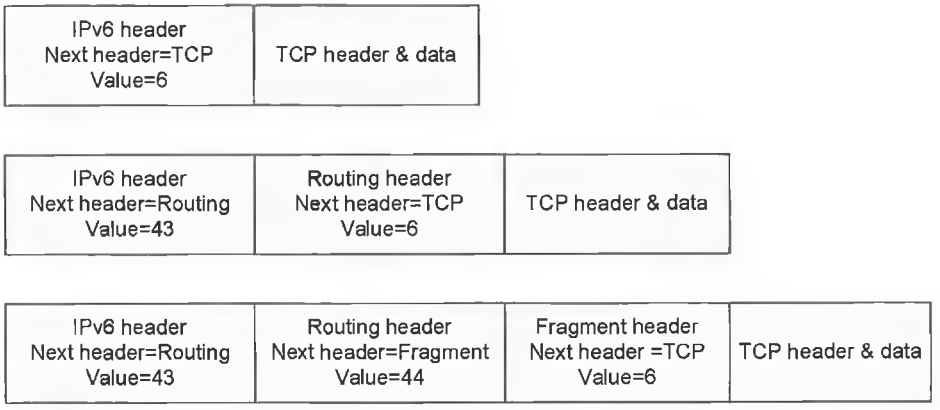

Figure 7 Use of Extension headers

Extension headers are only examined by the host whose address is indicated in the destination address field. The exception to this is when the extension header is a Hop-by-Hop header, in which case the information it carries must be examined and processed by every node along the path of the datagram. Extension headers must be processed in the order they appear in the datagram. Where there is more than one extension header the following header order should be used.

1. IPv6 header
2. Hop-by-Hop header
3. Destination options header
4. Routing header
5. Fragment header
6. Authentication header
7. Encapsulating Security Payload header
8. Destination Options header
9. Upper Layer header.

Hop-by-Hop Options Header

This header contains data that must be examined by every node along the path of the datagram. In IPv4 a router has to examine some of the upper layer data in order to determine whether or not it needs to examine the datagram. This is quite inefficient. With IPv6, however, the absence of a Hop-by-Hop extension header means that the router does not have to process router specific information and can pass the datagram directly on to the next hop. The format of this header is shown below.

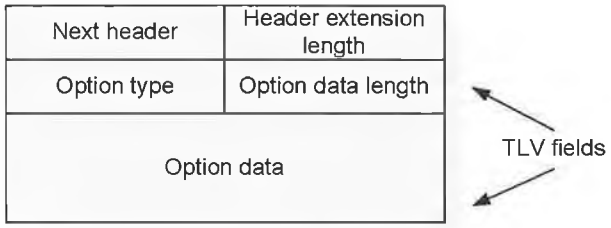

Figure 8 Hop-by-Hop Options header

Next header: This 8-bit field identifies the type of header that follows the Hop-by-Hop options header, for example TCP (6), UDP (17), or another type of extension header.

Header extension length: This is the length of the Hop-by-Hop options header in 8-octet units and does not include the first 8 octets.

Option type/length/data: These fields contain information that routers at each hop along the path of the packet need to process.

Routing Header

A source host uses this field to define a path that the datagram must take. In this case the destination IP address is not that of the ultimate destination host, but that of the next node that must be visited. The format of the options header is shown in figure 9 below.

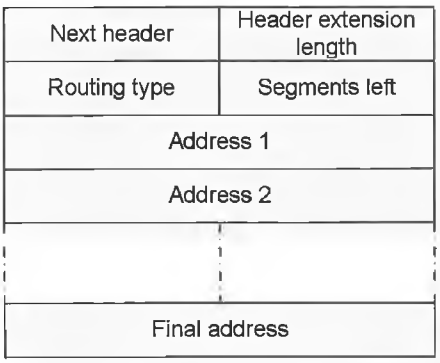

Figure 9 Routing header

Next header: This 8-bit field identifies the type of header that follows the Hop-by-Hop options header, for example TCP (6), UDP (17), or another type of extension header.

Header extension length: This is the length of the Hop-by-Hop options header in 8-octet units and does not include the first 8 octets.

Routing type: This 8-bit field indicates the type of routing header. At the moment only routing type 0 has been defined.

Segments left: This 8-bit field indicates the number of nodes left to be visited before arriving at the final destination. This directly corresponds to the number of address fields after the "Segment Left" field.

Address fields: These fields specify the address of the remaining nodes that must be visited.

Fragment Header

Unlike IPv4, fragmentation does not occur at the routers along the packet path. Instead, an IPv6 host uses Path MTU Discovery as previously described to determine the appropriate packet size and fragments if necessary. Fragmentation occurs only at the source host.

The Fragment header is shown in figure 10.

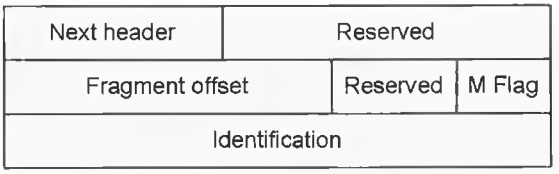

Figure 10 Fragment Header

Next header: This 8-bit field identifies the type of header that follows the Hop-by-Hop options header, for example TCP (6), UDP (17), or another type of extension header.

Reserved: This 8-bit field is not used.

Fragment offset: As in IPv4, this 13-bit field indicates the offset in 8-byte units of the data in this packet from the start of the original data.

Reserved: This 2-bit field is not used and is set to zero.

M Flag: This 1-bit field indicates whether this is the last fragment or if more fragments follow.

Identification: This 32-bit field has the same functionality as in IPv4. All fragments belonging to same original packet will have the same identification.

Destination Options Header

This field carries optional information that only destination hosts need examine. The header format is shown below in figure 11.

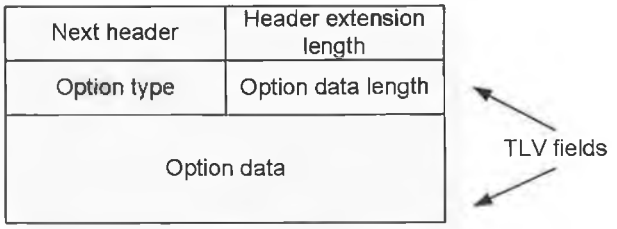

Figure 11 Destination Options Header

Next header:	This 8-bit field identifies the type of header that follows the Hop-by-Hop options header, for example TCP (6), UDP (17), or another type of extension header.
Header Ext Length:	This 8-bit field indicates the length of the options header in 8-byte units, but not including the first 8 bytes.
Options:	This variable length field is used in the same way as the hop-by-hop options header discussed earlier.

The above extension headers are described in RFC 2460.

IPv6 Addressing

The IPv6 addressing architecture (RFC3513) is in fact a radical deviation from the IPv4 address architecture that we are familiar with. The most obvious change is that IPv6 uses a 128 bit address as opposed to IPv4's 32 bit addressing scheme, thus yielding a significant increase in address space. To put this in perspective, consider each IP address as being represented by a distance of 1mm, then, there are sufficient IPv6 addresses to span the circumference of the Galaxy several times. Extending the address space in this way also allows for routing table optimisation, which will be looked at later.
The general format for IPv6 addresses is shown in figure 12.

Figure 12 IPv6 address format

IPv6 Address Types

Broadcast addressing is no longer supported in IPv6 and is replaced with multicast addressing instead. The supported categories are listed below.

Unicast

As in IPv4, the unicast address is the individual address of a specific interface. However, multiple interfaces on the same node can have the same unicast address. This is used for load balancing across multiple interfaces.

Multicast

As in IPv4, multiple nodes can share the same address. A packet sent to the multicast address is received by all nodes identified by that address.

Anycast

Anycast is new in IPv6, and is an identifier for a set of interfaces belonging to different nodes. A packet sent to an anycast address will be received only by one node and will generally be the node closest to the source. Consider, for example, an organisation's WAN that has more than one router with access to the Internet. All of these routers could be identified by the same anycast address. A host could then be configured with the anycast address as its default gateway. A packet sent to this address will be received by the router that is closest, in routing terms, to the source host.

Address Notation

An IPv6 128 bit address is represented in hexadecimal format. Each group of four hexadecimal characters are separated by a colon as shown below:

21ad:e53f:9b00:f019:0240:05ff:eo85:812b

While this syntax does look rather cumbersome, many addresses will contain a large number of zeros. Using zero compression, contiguous zeros can be replaced by a double colon, for example, the following address:-

fe80:0000:0000:0000:0240:05ff:e085:812b

Can be replaced by:-

fe80::0240:05ff:e085:812b

Caution must be taken to ensure that zero compression only occurs in one place as a host uses the full 128 bit address. For example the following address:-

fe80:0000:0000:0000:0240:05ff:0000:812b

can not be abbreviated as fe80::0240:05ff::812b since the host will not know how many zeros to insert between each double colon.

However, leading zeros can be compressed so that the above address can now be represented as:-

fe80:**0:0:0:0**:0240:05ff:**0**:812b or fe80**::**0240:05ff:**0**:812b

Text Representation of Address Prefix

IPv6 uses CIDR notation in a manner similar to that used in IPv4 for indicating the subnet prefix. In IPv6 the length of the prefix is written as the number of bits, in decimal, preceded by a slash, that is, *IPv6 address/prefix length.*

For example fe80:0:0:0:0:0240:05ff:0:812b/10 means that the leftmost 10 bits are the subnet prefix and can also be written as fe80::/10 to indicate the subnet address.

Format Prefixes

Format prefixes are prefixes that are used to identify special address types. RFC3513 "IP Addressing Architecture" which obsoletes RFC2373 specifies the following assigned prefixes.

Address Type	Binary Prefix	IPv6 Notation
Unspecified	0000…0 (128 bits)	::/128
Loopback	000…1 (128 bits)	::1/128
Multicast	11111111	ff00::/8
Link-local unicast	1111111010	fe80::/10
Site-local unicast	1111111011	fec0::/10
Global unicast	(everything else)	

Table 2 Assigned prefixes

As can be seen from table 2, there is no reserved space for an anycast address. Anycast addresses are taken from the unicast address space. Interfaces using anycast addressing must be configured so that they know this address is an anycast address.

Unspecified

The unspecified address of all zeros may not be assigned to a host. A typical use for this address is during auto configuration where the host does not yet know its own IP address.

Loopback

The loopback address is used for testing the protocol stack as in IPv4's 127.0.0.1. This address may not be assigned to an interface. Any packet received by a host with this destination address will be discarded.

Multicast

The multicast address has already been explained. All multicast addresses must begin with ff00.

Link-local unicast

In IPv6 a link refers to a network segment or subnet. A link-local unicast address is an address that is unique to that subnet/segment only. Link-local addresses cannot be forwarded by routers and, as a result, cannot connect to other networks. They are used on single segment lans only.

Site-local unicast

In IPv6 a site refers to an autonomous network. Site-local addresses can be routed within an autonomous network but routers must not forward these packets out onto the Internet.
The format for Link-local and Site-local addresses is shown in figure 13.

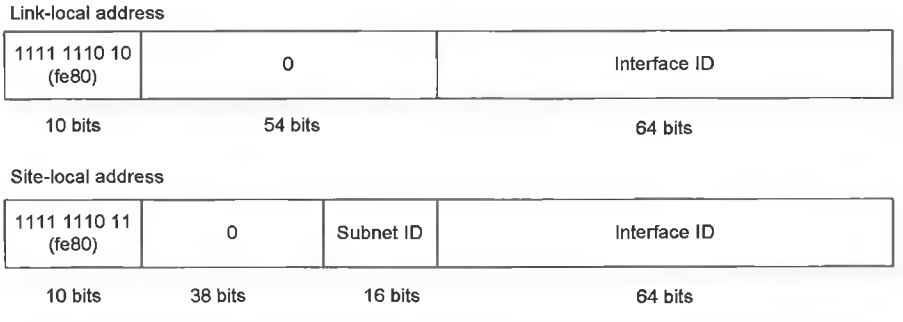

Figure 13 Link & Site local address format

Global unicast

This is a globally unique address. It was formerly known as "Aggregatable Global Unicast Address" but has been slightly redefined in RFC3587 in August 2003 with the focus being on routing optimisation. All unicast addresses except those beginning with binary 000 must adhere to the format shown in figure 14 below.

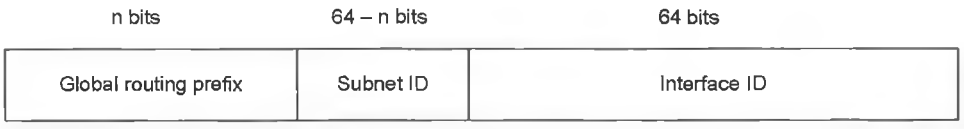

Figure 14 Global Unicast address format

The global routing prefix is designed to be structured hierarchically by the Regional Internet Registries and is the value assigned to a site. The subnet field is designed to be structured hierarchically by the site administrators and identifies a subnet within a site. The interface ID must be constructed in Modified EUI64 format. Such an address is constructed from the hosts MAC address in the following manner:-

First the hex digits 0xfffe are inserted into the middle of the MAC address between bytes three and four. The second low order bit of the first byte of the MAC address, called the universal/local bit, is inverted. This is best explained by example.

The MAC address of my PC is **00-40-05-85-81-2b.** To construct the Modified EUI-64 ID, we take this address and insert fffe between bytes three and four. The address now becomes 00-40-05-**ff-fe**-85-81-2b. We must now invert the second low order bit of the first byte. The first byte in this case is 0x00 so this now becomes 0x02. The 64 bit interface ID now looks as follows:- 0**2**-40-05-**ff-fe**-85-81-2b.

While looking at Global Unicast addressing it is best to look at the way in which IPv6 addresses are to be allocated.

IPv6 Address Allocation

Figure 15 shows the IPv6 address allocation hierarchy. IANA allocate address space to the Regional Internet Registries. Our RIR in Europe is RIPE NCC (Reseaux IP Europeens Network Coordination Centre). The RIRs in turn allocate address space to the Local Internet Registries (LIR). LIRs are essentially ISPs. Some of the main LIRs in Ireland include Eircom, Esat Net, HEAnet and UUnet to name but a few. An up to date list can be found on the RIPE website [3].

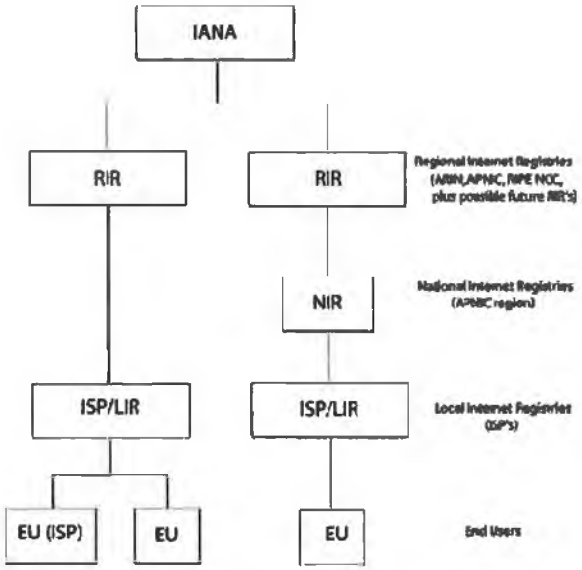

Figure 15 Address allocation hierarchy [RIPE]

Recommendations on IPv6 Address Allocations to Sites

Unlike IPv4, variable length subnet masking is not used, which means that for global unicast addresses, the host ID will always be 64 bits long using the Modified EUI-64 format. RFC3177 sets out the recommendations to the addressing registries (APNIC, ARIN AND RIPE-NCC) on polices for assigning IPv6 addresses.

The recommendation sets out that a /48 bit prefix be allocated in the general case. This means that if an organisation has been assigned a /48 prefix that it can further subnet the network into 2^{16} subnets. Interestingly, the recommendation also proposes that this prefix be allocated to homes. This does seem wasteful but there is good justification. At present, all global

unicast addresses begin with 001 in the first three bits, therefore there are 45 bits remaining in the prefix. That is to say, the number of available /48 prefixes is 2^{45} which approximately equals 35 trillion. Another reason for this is to do with site renumbering. During site renumbering from one prefix to another the whole process is greatly complicated if the prefixes have different lengths. Site renumbering and route optimisation will be looked at later.

A /64 prefix is to be issued to a site when it is known that only one subnet is needed. A /128 prefix will be assigned when it is absolutely known that one and only one device is being used.

In the case of a very large organisation multiple /48 prefixes or smaller can be allocated. If an ISP can show that it has plans to make 200 /48 assignments to other organisations within two years then it can be allocated /32 prefix. [4]

Address Assignment examples

At the moment, for global unicast addressing, IANA is issuing 2001::/16 to the RIRs. Therefore all addresses issued to RIPE, ARIN, APNIC, etc., will begin with 2001. RIPE have been allocated the following addresses:

2001:0600::/23
2001:0800::/23
2001:1400::/23 up to and including 2001:3A00::/23, but excluding 2001:1800::/23
2001:4000::/23

Referring back to figure 14 it can be seen that RIPE can now further allocate up to 41 bits to ISPs. So, for example, RIPE has allocated 2001:0bb0::/32 to Eircom and HEAnet has been allocated the following:

2001:0770::/35
2001:0770:2000::/35
2001:0770:4000::/34
2001:0770:8000::/33

HEAnet in turn are allocating /48 prefixes to Universities and Institutes of Technology.

Chapter 3 – Border Gateway Protocol

In order to gain an appreciation of route optimisation and related issues, it is first necessary to provide a background on the Border Gateway Protocol (BGP). It is not intended to provide a detailed outline of BGP, but rather to highlight the features of BGP that are relevant to this discussion. The working details of BGP4 are defined in RFC1771.

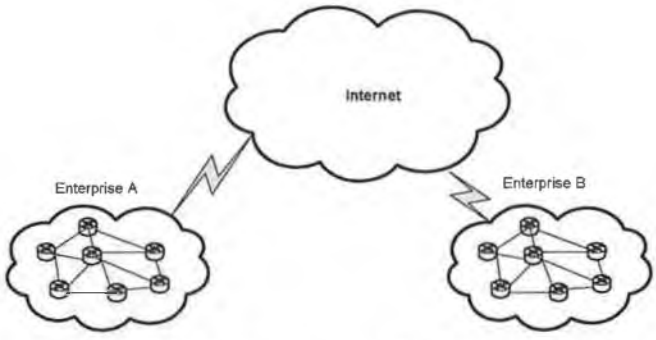

Figure 16 Internet Routing

Observing figure 16 above, it is obvious that every router in Enterprise A needs to know the route to every other router within that enterprise. The larger the enterprise, the larger the routing tables will be. Enterprise A may never need to communicate with enterprise B, however, should enterprise A need to communicate with enterprise B, then routers in both enterprises would need to know how to reach each other. The more enterprises connected to the Internet therefore, the larger the routing tables will be. This situation is very undesirable for the following reasons. Firstly all routers would require massive memory to support the huge amount of routing entries. Secondly, all routers would require substantial processing power in order to index through their massive routing tables. Thirdly, convergence would take an inordinate amount of time due to the large amount of routers involved.

Observing figure 16 again, it can be seen that only one router from each enterprise is connected to the outside world. We will call this the *border router*. It would make more sense that the border routers in each enterprise know about each other's existence. These

border routers can summarise all of the networks in their own enterprise (autonomous system) in to one *Supernet* therefore, significantly reducing the size of the routing table. Routers within an Autonomous System (AS) will run an Interior Gateway Protocol (IGP) like RIP or OSPF. An IGP router only exchanges routing tables with other IGP routers within the same AS. The border routers, on the other hand, will use an *External Gateway Protocol (EGP)*. EGP routers summarise the routes belonging to their own AS. These route summaries are then advertised to other EGP routers.

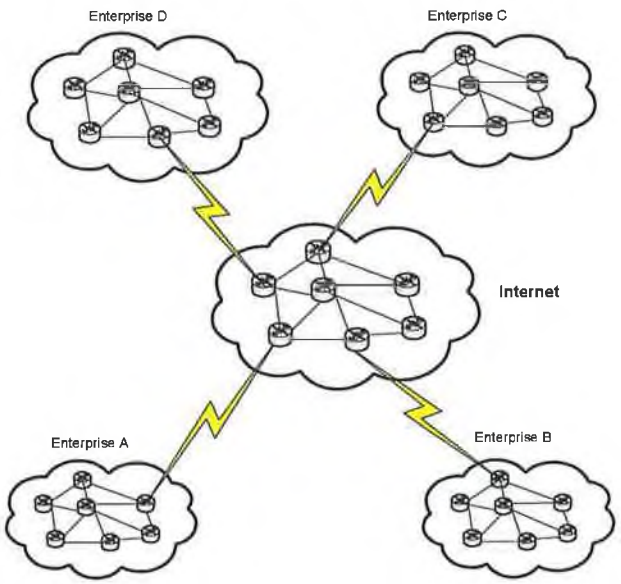

Figure 17 Internet Backbone

From figure 17 above it can be seen that only four networks now need to be advertised as opposed to a minimum of forty networks (each enterprise has ten networks) without using an exterior gateway protocol.

The EGP that is used is BGP4. At the moment, as is evident from figure 21, there are nearly 200,000 routes advertised by BGP routers. BGP needs to be quite scaleable. To achieve this level of scalability BGP uses routing policies called attributes. These attributes are explained below. Refer to Cisco on-line [5] for more detailed explanations.

Local Preference

Figure 18 below shows an autonomous system with two exit/entry points. The Local Preference, which is advertised throughout an AS, is used to select the preferred exit point from the local AS.

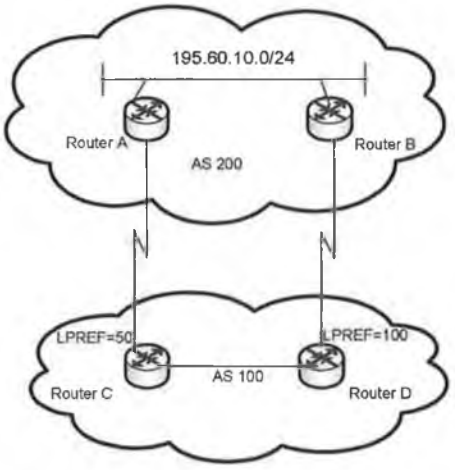

Figure 18 Local Preference

Since router D has the highest local preference it will be used to route packets from AS 100 to AS 200.

In the figure above routers A and B are said to be internal peers, as are routers C and D. Routers A and C, on the other hand, are external peers, as are routers B and D. Although all routers are configured using the same BGP4 protocol, internal peers are considered as using internal BGP (iBGP) while external peers are considered as using external BGP (eBGP)

Multi-Exit Discriminator (MED)

This attribute is used to suggest to routers in an external AS which entry point to use to gain access to the local AS. It can be thought of as the opposite or reverse of the local preference attribute.

Origin Attribute

This attribute indicates how BGP learned about a particular route. It has three possible values:

- IGP- The router learned this route from another BGP within the same AS.
- EGP- The router learned about this router from another BGP in a different AS.
- Incomplete- The origin of the route is unknown.

AS Path Attribute

Every autonomous system connected to the Internet must use a unique AS number which is issued by IANA.

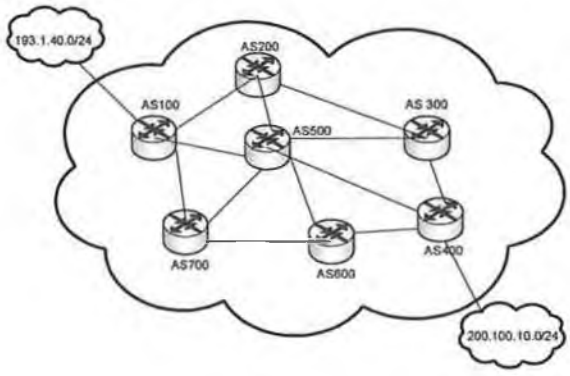

Figure 19 AS Path Attribute

In figure 19 above it can be seen that a number of ASes must be traversed in order to get from network 193.1.40.0/24 to network 200.100.10.0/24. So, the AS Path Attribute, for example, could be {100,500,400}.

Next Hop Attribute.

This attribute specifies the address that is used to reach the advertising router. In figure 20 router A advertises network 195.60.10.0/24 with a next hop address of 193.1.40.1

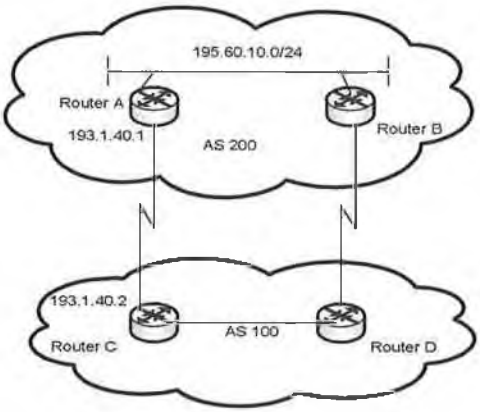

Figure 20 Next Hop Attribute

Community Attribute

The community attribute groups destinations into groups called communities to which preferences can be applied. Predefined community attributes are:-

- **No Export** Do not advertise this route to EBGP peers
- **No Advertise** Do not advertise this route to any peer
- **Internet** Advertise this route to the entire Internet community

The above attributes are what make BGP4 the ideal routing protocol for the Internet backbone. Looking again at figure 19, for example, AS 500 can be considered as a transit network in that it can be used to carry packets from AS100 to AS400. However, this could be a corporate network that may not wish to carry packets for another corporation. BGP4 allows routers to be configured to act as transit provider for other ASes or even a selection of ASes but not for others. BGP4 can also be used for load balancing as in figure 20 and for route filtering as is demonstrated later.

Chapter 4 - Route Scaling Techniques in IPv4

Chapter 2 outlined the IPv6 addressing scheme. The main driving force behind the introduction of IPv6 was to introduce a larger address space and to provide route optimisation for routing tables on the Internet. This chapter looks at existing methods put in place with IPv4 to alleviate routing and address problems.

Route Scaling

Figure 21 below shows a historical track of BGP entries on AS1221. This router has a default-free routing table which essentially means that it can see all routes advertised on the Internet. This graph and other related data are updated daily on Potaroo website [6]

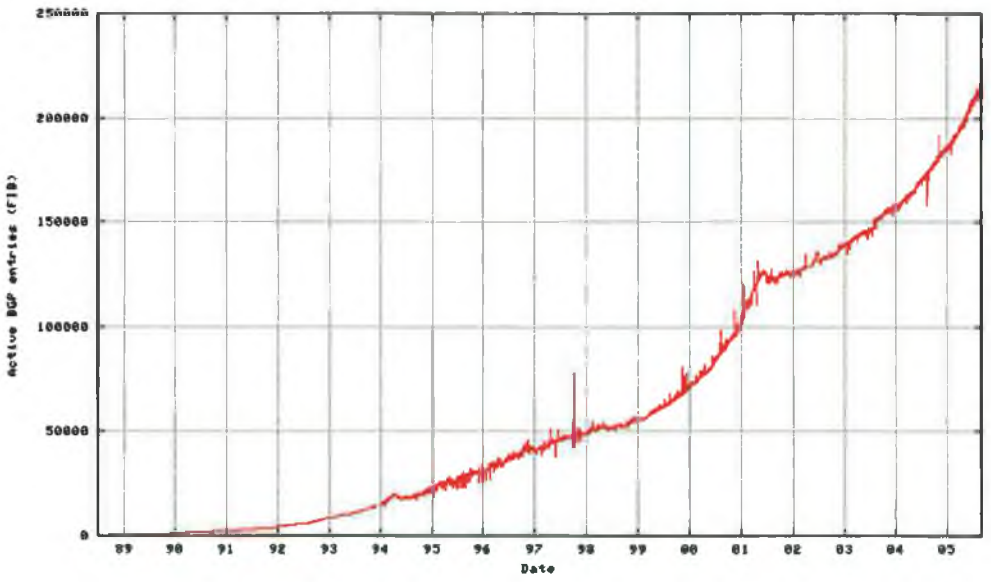

Figure 21 Active BGP entries Aug 2004 [potaroo.net]

It can be seen that from 1988 to the first quarter of 1994 routing table sizes were growing at an exponential rate. The obvious reason for such growth was that the Internet was simply becoming more and more popular. This popularity could largely be attributed to developments in email and web browsing technology. However, there was also a very valid

technical reason behind this growth which was largely to do with the inefficiencies of the addressing scheme at that time.

Pre-1992 there were essentially three network sizes. The Class A network supported 16.7 million hosts. The Class B network supported 65,536 and the Class C network supported 256 hosts. A typical network at the time (1000 to 2000 hosts) did not fit easily into any of the three networks. To issue a Class B address was completely wasteful so the alternative was to issue multiple Class C addresses. The assignment of multiple Class C networks combined with the rapid growth in the Internet meant that routing tables were going to grow so large that firstly, the existing memory and hardware in routers wouldn't be able to cope, secondly the convergence time would significantly increase, and thirdly, the address space would eventually run out. It was apparent that this situation could not continue to go unchecked.

Classless Interdomain Routing (CIDR)

In June 1992, RFC1338 *"Supernetting: an Address Assignment and Aggregation Strategy"* was published. Its focus was not on the eventual demise of addresses, but on the more immediate problem of route scalability.

The proposed solution was "*to hierarchically allocate future IP address assignment, by delegating control of segments of the IP address space to the various network service providers.*" The main goal was to allocate an address block to an ISP, who in turn would further allocate addresses from this block to their clients. The ISP could then advertise just one aggregate route for all client networks derived from the same block. This, of course, took some time to implement as new routing protocols were also required in order to support Supernetting. RFC1338 was soon to be replaced by RFC1519 *"Classless Inter-Domain Routing (CIDR)"* in 1993.

From figure 21 it can be seen that the combination of CIDR and hierarchical routing did indeed stem the rapid growth of routing table entries on the Internet. From 1994 to 1998 there is almost a linear, as opposed to exponential growth. Routers could now easily cope with this growth as advances were made in router hardware and memory. However, the pattern of

exponential growth resumes again in 1998. According to Huston (2001), the main contributing factor to this exponential growth rate is site multihoming.

Route Aggregation

Multihoming is when an end-user network, or even indeed an ISP is connected to more than one outside network. This practice multiplies the number of routes advertised on the Internet, disrupts the CIDR address hierarchy, and punches holes in the route aggregation process. The main benefit to customers and ISPs, however, is resilience and load balancing. Figure 22 helps to illustrate this.

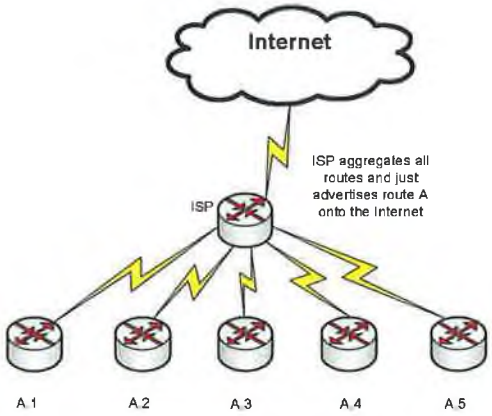

Figure 22 Route Aggregation

In the illustration above the ISP assigns networks A.1 through A.5 to its clients. The ISP then aggregates and summarises these routes into just one advertisement that it pushes out onto the Internet. Any return packet on the Internet with a destination address beginning with "A" is then sent to the ISP router.

Closer to home, HEAnet, which is a service provider for the Universities and Institutes of Technology has been assigned 193.1.0.0/16 from RIPE. This way only one aggregated route needs to be advertised on behalf of all third level organisations. In turn the Institutes, for example, have been assigned a /21 prefix from this address space. DKIT has been assigned 193.1.40.0/21, which gives DKIT eight Class C networks from 193.1.40.0 to 193.1.40.7 inclusive. All of these networks can be aggregated in to just one route (193.1.40.0/21). From

this it can be seen that CIDR and route aggregation have the potential to significantly reduce the size of routing tables on the Internet.

Multihoming

As mentioned earlier CIDR did indeed curb the exponential growth of routing tables and this was the case for a few years until ISPs and end users started multihoming their networks. The effect is illustrated in figure 23 below.

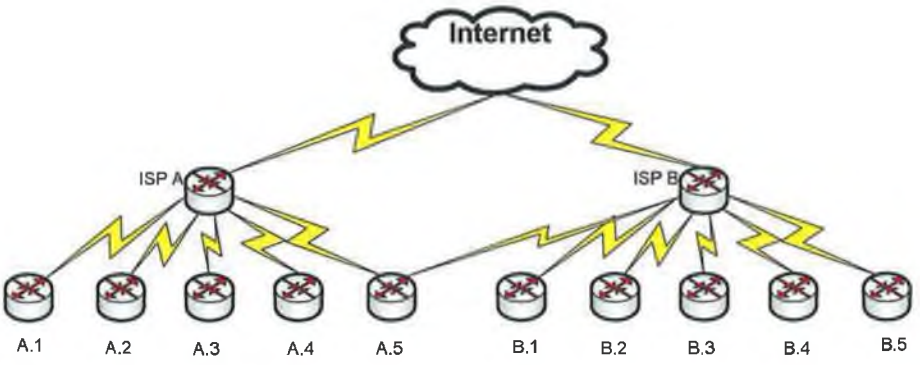

Figure 23 Multihoming

From the diagram it can be seen that network A.5 is multihomed with ISP A and ISP B. At least one of the ISPs will now have to advertise an additional route, which is outside of their own contiguous scope. However, this additional route is already part of another ISP's aggregate prefix that is already being advertised. This is referred to as a *more specific prefix*. This additional route advertisement essentially punches a hole in the contiguous routing advertisement of the ISP. It is obvious therefore that multihoming can have a very serious and detrimental impact on the number of routes advertised.

Referring to Potaroo.net [6]; of the present 215,000 bgp routing entries on the Internet, a staggering 62% or 133,000 of these entries are specific advertisements. A *specific advertisement* is an advertisement that is associated with a sub-span of address space of another advertisement. For example, in figure 23, network A.5 will be advertised by ISP A under the Supernet of 'A'. ISP B, on the other hand, will have to advertise the more specific network of A.5 in addition to its own aggregate advertisement for Supernet B.

CIDR and route aggregation did indeed work well for a few years but are no longer able to slow down the growth of route table entries. If the issues around multihoming were addressed, then perhaps the growth rate of routing tables could be curbed. These issues are examined in the next section.

Existing Multihoming Techniques

End user sites have good reasons to use multihoming, which include redundancy, load sharing and performance. This option has been made even more attractive by the opening up of Telecom markets. The practice will continue to grow until there is some disincentive for the end user not to do so. What is needed is a method that will support multihoming without having a major impact on the growth of BGP routing entries. RFC 2260 *"Scaleable Support for Multi-homed Multi-provider Connectivity"* describes various multi-homing methods that can be used with both IPv4 and IPv6. These methods are summarised below.

Auto Route Injection BGP-4 [RFC2260]

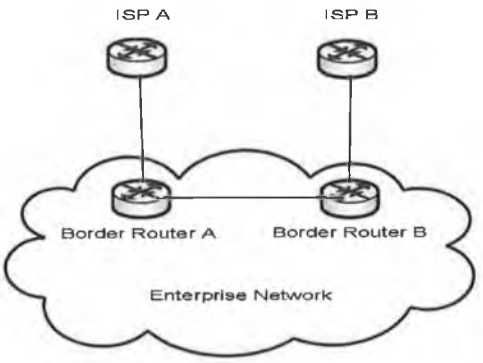

Figure 24 Internal Peering

Figure 24 shows an enterprise network with two border routers. Each router is connected to a different ISP. Each ISP allocates a subnet of its own address prefix to each site, so that site A will have prefix A and site B will have prefix B. Under normal conditions (both ISPs are up) each of the enterprise's border routers will only advertise the appropriate prefix. For example, border router A advertises prefix A, while border router B advertises prefix B. Networks A and B are advertised under the aggregate prefix associated with each ISP so that there is no increase in route advertisements on the Internet. Should a link to one of the ISPs

fail, the other border router will advertise the prefix associated with the failed link. For example, suppose ISP B goes down. Border router A will advertise reachability to the address prefix allocated by ISP B. This will increase the size of the BGP tables on the Internet since ISP A is now advertising an address prefix outside of its own aggregate prefix. However, this will only occur during an "ISP down" scenario. In fact the probability of all multihomed sites loosing a connection to one of their ISPs at the same time is quite small. Therefore the resulting increase in advertised routes using the above mechanism will only be a fraction of the total number of multihomed sites.

The above scenario depends firstly on the ability of the border routers to determine that the connection to the other ISP has failed and, secondly, that the address prefix of the other network is known. BGP4 provides a solution to this through the use of peering. Peering is the association between adjacent BGP routers. When BGP is running between two or more routers that belong to the same AS, it is referred to as internal BGP (iBGP). When BGP is running between two routers belonging to different ASes then it is referred to as external BGP (eBGP).

Non Direct eBGP Peering BGP-4 [RFC2260]

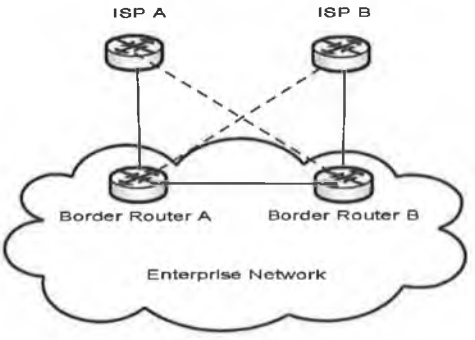

Figure 25 Indirect Peering

Another solution that can potentially eliminate the advertisement of additional routes is shown in figure 25.

In figure 25 both border routers maintain peering with both ISPs. Border router A advertises prefix A to ISP A and prefix B to ISP B. Likewise Border router B advertises prefix B to ISP B and prefix A to ISP A. Under normal "up" conditions, packets on the Internet destined for hosts with prefix A are handled by ISP A. Likewise, ISP B handles packets destined for prefix B. In the event that the link between ISP B and Border router B fails, ISP B will still receive prefix B packets since it is still advertising the larger aggregate prefix of B. However, when prefix B packets are received, ISP B will encapsulate the packets and forward them to Border router A. Reachability is still maintained without the overhead of additional route advertisements. The drawback of this configuration is that packets will be using non-optimal routes.

Solution 3 Provider Independent Addressing [RFC1518]

Another recommended method is for a multihomed enterprise to use provider independent addressing. This allows an enterprise to summarise all reachable addresses within the enterprise into one prefix. The problem with this solution is that the upstream providers will now need to advertise an address prefix which is not associated with their own address space. Despite these obvious problems this is the most commonly used method for multihoming.

Solution 4 Prefix Filtering [RFC1518]

There are of course other recommended solutions, but prefix filtering is the solution with the greatest potential impact. ISPs and network operators are advised to configure their routers to filter out certain IP addresses. Routers should not advertise routes to the addresses specified in table 3.

Default/broadcast	0.0.0.0/8
Loopback address	127.0.0.0/8
Private addresses	10.0.0.0/8 172.16.0.0/12 192.168.0.0/16
Class D&E	224.0.0.0/3
Auto-configuration	169.254.0.0/16
Test network	192.0.2.0/24
Exchange points	192.41.177.0/24 192.157.69.0/24 198.32.0.0/16 206.220.243.0/24
IANA reserved	128.0.0.0/16

Table 3 Reserved Addresses

For example 127.0.0.0/8 is a loopback address and is used for testing the protocol stack on an IP host. The network 192.168.0.0/16 is used behind NAT routers and again should not be advertised onto the global Internet. Operators can also filter prefixes of length greater than /24, and prefixes greater than /16 for Class B addresses. In order to reduce the size of BGP tables the RIRs issue guidelines for prefix filtering. Table 4 below summarises the filtering recommendations issued by RIPE [7].

IPv4 CIDR block	Smallest RIPE NCC Allocation	Smallest RIPE NCC Assignment
62/8	/19	/19
80/8	/20	/20
81/8	/20	/20
82/8	/20	/20
83/8	/21	/21
84/8	/21	/21
85/8	/21	/21
86/8	/21	/21
87/8	/21	/21
88/8	/21	/21
193/8	/19	/29
194/8	/19	/29
195/8	/19	/29
196.200/13	/20	/24
212/8	/19	/19
213/8	/19	/19
217/8	/20	/20

Table 4 RIPE Guidelines for Filtered Prefixes

The table shows, for example, that RIPE can issue 193.1.0.0/19 to an ISP. In fact this is the allocation to HEAnet (HEAnet have /16 as the allocation was pre 1994). The ISP, HEAnet in this case, can then allocate a portion of this address space to its clients. In the case of HEAnet, DKIT are assigned 193.1.40.0/21.

Table 4 shows, however, that the ISP should not allocate a prefix length longer than /29 in this part of the address space. In other words, other network operators may filter an advertisement from any network beginning with 193.0.0.0/8 and having a prefix length greater than /29, such a network would then become unreachable.

If these guidelines are strictly adhered to, then the size of BGP tables will not only be significantly reduced, but the growth rate of the tables will also be stemmed. In a paper published by Bellovin *et al.*[8], it is shown that not all BGP routers are using filtering to the extent outlined above. On the Telstra BGP router there were almost 6,000 entries of prefix

length greater than /24. Also if filter rules were to be applied for Class A and C addresses as recommended by the RIRs (table 4 only shows RIPE recommendations) a further 16,000 BGP entries would be removed from the Telstra BGP router. Unfortunately this paper is now more than three years old, so it is important to draw on current information. Up to date BGP table information is available from Potaroo [2]. The Telstra router in the "default free" zone is the router that will be used throughout this text for various measurements.

Looking at today's figures for Telstra [6]; there are 29,844 routes of prefix length greater than /24. That is just over 17% of all BGP entries. This figure, however, doesn't take into account that some prefixes are allowed to be greater than /24, for example, as in table 4 above. Looking again at the Telstra BGP report [6]; what is interesting is the number of /32 addresses. A staggering 12,624 or 7.28% of total BGP entries have /32 prefixes.

If stringent filtering were to be applied, then would that not render many parts of the Internet unreachable? Not necessarily. For example, if HEAnet were to allocate a /30 prefix in the 193.0.0.0/8 address space and if such a site was multihomed, then the likelihood is that the /30 will be filtered by other network operators. However, reachability to the /30 network would still by advertised through the larger prefix of /19. Bellovin *et al.* [8], suggest that worst case, a filtered prefix would be covered by another aggregate prefix leaving only about 0.3% of the address space uncovered by other prefixes. This theory is not conclusive and more testing would be required. An example might better illustrate the issues raised. Figure 26 shows a small sample of the BGP routing table taken form the Telstra default free router.

This portion of the routing table illustrates very well all the issues raised in the previous paragraphs. The left hand column indicates the destination network. It is clear from this section of the table that there are different networks that have same AS path. Using RWhois [9] and searching for address 139.191.0.0 it can be seen that this address is assigned to MCI EMEA in Europe. In fact the address assigned by RIPE to MCI is 139.191.0.0/16. It is possible that the addresses listed in figure 26 could be advertised under the aggregate 139.191.0.0/16. By examining the BGP entries for MCI assigned addresses, there are 653 advertised routes, 115 of which could be potentially withdrawn and advertised under a larger aggregate, thus leading to a 17.6% reduction in advertised routes. It is important to note that

the above aggregation possibility does not take into account issues such as traffic engineering requirements or policies local to MCI.

Network	Next Hop	Metric	LocPrf	Weight	Path
139.191.0.0/18	202.84.219.193		55	0	4637 701 702 i
* i	134.159.127.248	0	55	0	4637 701 702 i
*> 139.191.0.0	202.84.219.193		55	0	4637 701 702 i
* i	134.159.127.248	0	55	0	4637 701 702 i
*> 139.191.64.0/19	202.84.219.193		55	0	4637 701 702 i
* i	134.159.127.248	0	55	0	4637 701 702 i
*> 139.191.96.0/20	202.84.219.193		55	0	4637 701 702 i
* i	134.159.127.248	0	55	0	4637 701 702 i
*> 139.191.112.0/20	202.84.219.193		55	0	4637 5511 2611 i
* i	134.159.127.248	0	55	0	4637 5511 2611 i
*> 139.191.128.0/20	202.84.219.193		55	0	4637 701 702 i
* i	134.159.127.248	0	55	0	4637 701 702 i
*> 139.191.144.0/21	202.84.219.193		55	0	4637 701 702 i
* i	134.159.127.248	0	55	0	4637 701 702 i
*> 139.191.152.0/21	202.84.219.193		55	0	4637 1239 1299 766 766 766 766 766 i
* i	134.159.127.248	0	55	0	4637 1239 1299 766 766 766 766 766 i
*> 139.191.160.0/20	202.84.219.193		55	0	4637 701 702 i
* i	134.159.127.248	0	55	0	4637 701 702 i
*> 139.191.176.0/21	202.84.219.193		55	0	4637 701 702 i
* i	134.159.127.248	0	55	0	4637 701 702 i
*> 139.191.184.0/21	202.84.219.193		55	0	4637 3549 680 680 680 680 680 680 680 ?
* i	134.159.127.248	0	55	0	4637 3549 680 680 680 680 680 680 680 ?
*> 139.191.192.0/20	202.84.219.193		55	0	4637 701 702 i
* i	134.159.127.248	0	55	0	4637 701 702 i
*> 139.191.208.0/20	202.84.219.193		55	0	4637 701 702 i
* i	134.159.127.248	0	55	0	4637 701 702 i
*> 139.191.224.0/19	202.84.219.193		55	0	4637 701 702 i

Figure 26 Portion of BGP routing table

Address Space Fragmentation

Multihoming is not the only reason for increased growth in the BGP routing tables. Suppose, for example, that a customer changes ISP and takes with them their previously assigned address prefix space. This is desirable from the customers point of view as it means that the site would not need to be completely renumbered, which can be both expensive and time consuming. By taking their address space with them, they leave behind a hole in the contiguous address space of their former ISP, not only that, but now the new ISP must

advertise connectivity to a network that is not covered under its own prefix, which in turn adds to the BGP table size in the default-free zone. Handing back addresses to the original ISP would solve this problem, however, this means that the enterprise would have to completely renumber. An alternative solution to this is to use NAT which will be looked at shortly.

Network Address Translation also offers some interesting and practical ways for reducing the number of routes advertised by multihomed sites. The next chapter presents a brief overview of NAT itself and then discusses how NAT can be used on multihomed sites.

Chapter 5 - Address Scaling Techniques in IPv4

By the mid nineties it was very clear that the growth rate of the Internet would eventually deplete all of the available IPv4 address space. The previous chapter examined issues relating to route scaling as a result of this growth. This chapter focuses on the more obvious issue relating to Internet growth, which is the eventual depletion of the IPv4 address space. Moves were already afoot to design and introduce a new IP protocol, which is now known as IPv6. A more immediate solution, however, was required in order to retard the consumption rate of IPv4 addresses. The next section describes multihoming with NAT. A brief introduction to NAT is provided below.

Network Address Translation

Network Address Translation or NAT was first outlined in RFC 1631 and subsequently updated by RFC 3022. The main concept behind NAT was simply to extend the lifetime of the IPv4 address space without replacing IPv4 itself. Since NAT has been around for some time now, only a brief overview will be outlined in this section so that issues concerning NAT and multihoming will be better understood.

NAT is simply a service that a router provides by translating a private IP address into a public IP address. Every IP host on the Internet must have a unique publicly assigned address. No two hosts on the Internet can have the same address. This is referred to as a 'public address'. A 'private address', on the other hand, is an address that is only used within a private network and cannot be transmitted out on the Internet. In the past, if a private network was required to connect to the Internet, then renumbering of the network to publicly assigned IP addresses was essential. With NAT, however, renumbering is not required, instead, NAT translates the internal private addresses into external public addresses as shown in figure 27 below.
The figure shows a private network with internal private addresses using 192.168.2.0/24. The host 192.168.2.2 wishes to communicate with 200.100.10.1, the source and destination addresses are therefore 192.168.2.2 and 200.100.10.1 respectively. The NAT device, however, replaces or translates the source address of 192.168.2.2 with the public address 195.60.10.1 and keeps a record of this translation in its NAT table. When the host on the

remote network sends a packet back to the originating host it will use the destination address 195.60.10.1.

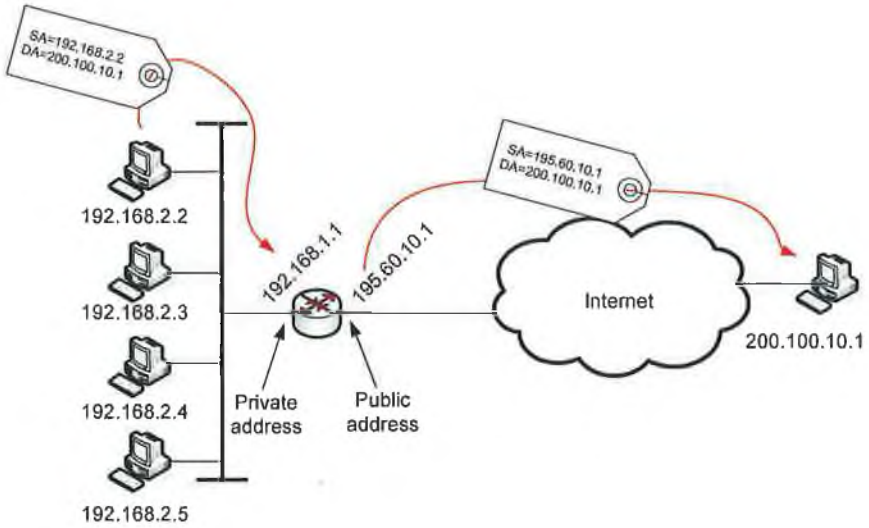

Figure 27 Network Address Translation

Upon receiving the packet on its return path, the NAT device will search its table and translate the address 195.60.10.1 into 192.168.2.2 and then forward the packet to this host.

NAT can be configured so that a pool of internal private addresses can be translated into a pool of publicly assigned addresses on a one to one basis. This does not mean, however, that the number of public addresses must be the same as the number of private addresses as not all internal hosts, as a general rule, will wish to connect to the Internet at the same time. Using Network Address Protocol Translation (NAPT) many hosts can be translated into the same public address at the same time. NAPT works by translating both the protocol port number and IP address. Referring to figure 27 again, assume that host 200.100.10.1 is a web server. When host 192.168.2.2 initiates a TCP session with the WEB server it will use the well known port number of 80 as the destination port. The operating system in host 192.168.2.2 will assign a source port greater than 1023. Let's assume that the operating system assigns port number 1025 as the source port to the packet. The originating packet will now have a source IP address and TCP port number of 192.168.2.2 and 1025 respectively. The destination IP address and TCP port number will be 200.100.10.1 and 80 respectively. The NAT device will now translate the source IP address into 195.60.10.1 as before but will also

translate the source TCP port number, for example, into 1500. The NAT device will keep a record of this translation in its NAT table as shown in figure 28. When the Web server replies, the destination address and TCP port number will be 195.60.10.1 and 1500 respectively.

Inside host IP address	Original TCP source port	Public IP address	TCP source port
192.168.2.2	1025	195.60.10.1	1500
192.168.2.3	1025	195.60.10.1	1501
192.168.2.4	1030	195.60.10.1	1502
192.168.2.5	1026	195.60.10.1	1503

Figure 28 NAT Table

As can be seen from figure 28, all inside hosts can communicate simultaneously with the outside world using the same public IP address. The translation table keeps track of which port was assigned to which host. So, for example, when the NAT device receives a return packet from the Internet, the destination address and TCP port number of 195.60.10.1 and 1502 respectively will be translated into 192.168.2.4 and 1030. To the end host, this is a completely transparent process.

The obvious advantages of NAT are the conservation of address space. In the above example it can be seen that an entire network of hosts can communicate on the Internet by using a single publicly assigned IP address. The other advantage is security. The NAT device acts as a sort of firewall by not advertising the original source address on the Internet. The furthest, therefore, that anyone could hack into is the NAT device itself.

But NAT is not without its drawbacks. NAT only alters the IP address in the IP header and in the case of NAPT (for the rest of the discussion NAT will refer to both NAT & NAPT unless otherwise stated), the port number in the TCP header. If, however, the payload portion of a packet carries addressing information also, then NAT will not be able to translate this as the payload portion of a packet is not examined and as such the "end to end" identity is lost. RFC 3027 provides a detailed list of some of the applications that have difficulties using NAT and will not be looked at in this discussion.

Chapter 6 - Multihoming with NAT

In order for multihoming to work with NAT, it is important that NAT can translate both the source and destination IP address. To illustrate this observe figure 29 below.

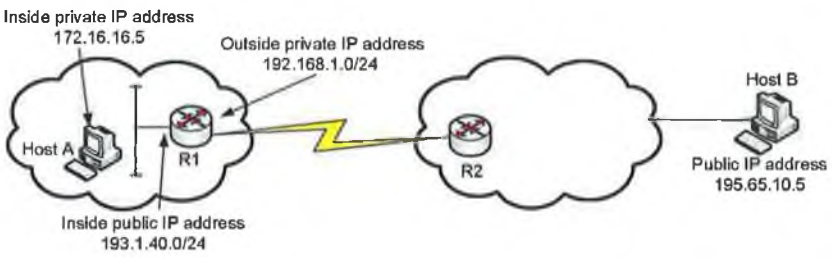

Figure 29 Bidirectional NAT

The network on the left has both private and public IP address assignments. To simplify matters, assume that host A has already connected to host B. When host A wishes to transmit packets to host B, the IP packet will have a source IP address of 172.16.16.5 and a destination IP address of say 192.168.1.5. When router R1 receives this packet it will search through its NAT table for the entry 172.16.16.5 in order to determine the corresponding public address is for this entry. Router R1 will now change the source address of the packet to the corresponding public IP address. In this case, for example, we will say that the public address is 193.1.40.5.

In the case where host A has never communicated outside of its own network before, then the router will create an entry for the source private IP address and assign it a corresponding public IP address, Cisco (1997). Next the router will search its translation table for the private destination IP address, which in this case is 192.168.1.5. There will already be a corresponding entry in the translation table of R1 since host A had previously established communication with host B. The NAT router will, therefore, change the destination IP address to 195.65.10.5.

To summarise, when the IP packet is on its way from host A to R1, the source and destination IP addresses are 172.16.16.5 and 192.168.1.5 respectively, that is, both addresses are private

IP addresses. However, as the packet leaves R1 on its way to host B, the source and destination addresses will be 193.1.40.5 and 195.65.10.5 respectively.

When host B wishes to return a packet to host A it will use only public IP addresses. The source and destination IP addresses will be 195.65.10.5 and 193.1.40.5 respectively. When router R1 receives this packet it will look up its translation table for the private address entry corresponding to 195.65.10.5, which in this case will be 192.168.1.5. Next router R1 will look up its translation table for the corresponding entry of 193.1.40.5, which in this case will be 172.16.16.5. R1 will assign this as the destination address in the packet. The new source and destination addresses will now be 192.168.1.5 and 172.16.16.5 respectively.

To summarise, only private source and destination addresses are used by hosts behind a NAT device. Packets entering the enterprise's NAT router from the outside will only have public source and destination IP addressees.

DNS with Bidirectional NAT

The above technique is also used for DNS queries and responses. Assume for example that Host A issues a DNS query to translate the DNS name of Host B into an IP address. If the DNS response comes back from outside the enterprise, the NAT router R1 will look up its translation table to check if it already has an entry for this pubic IP address. If it has, then it will replace this public address with its corresponding private address entry from the table, for example 192.168.1.5 as in the previous examples. When Host A wishes to communicate with Host B it will use 192.168.1.5 as the destination address and translation takes place as already described above. If, however, there is no entry in the NAT router's table, then a new private address will be assigned from its pool.

Externally Originated Connection

The following section describes how a host using a public IP address on the outside can communicate with a host that is using a private IP address behind a NAT device, Cisco (1997). Observe figure 30 below.

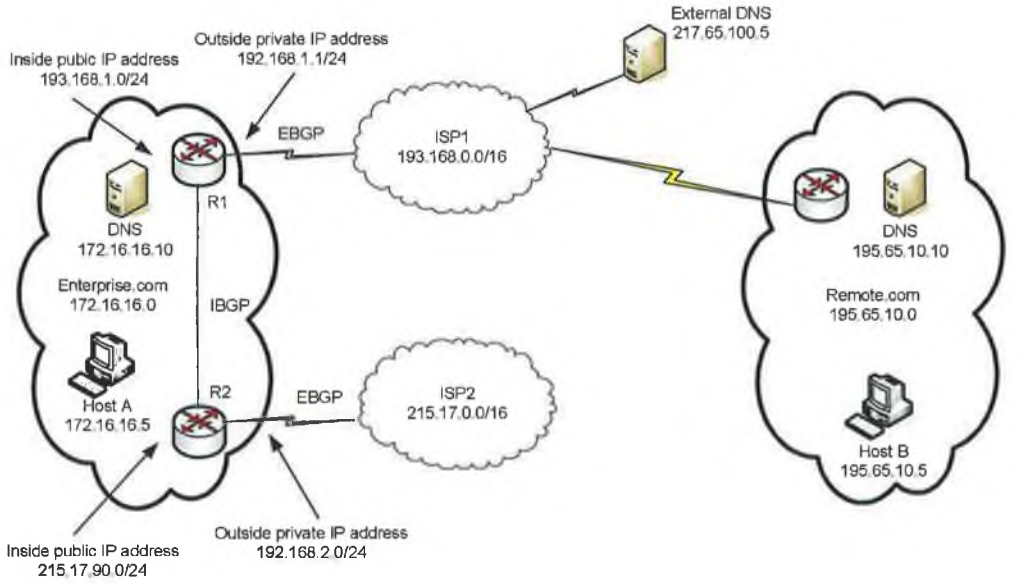

Figure 30 Multihoming with NAT and DNS

Assume that Host B wishes to communicate with Host A. Host B will send a DNS query to its own DNS server (195.65.10.10), which in turn will query the .com server to establish the authoritative server for "enterprise.com". The authoritative server will respond with two addresses, for example 193.168.1.10 and 215.17.90.10. The DNS server for Host B (195.65.10.10) will now send a query to one of these addresses, say 193.168.1.10. This query will then be forwarded to R1. When R1 receives the packet it will look in its translation table for a private address corresponding to the entry for 193.168.1.10. In this case it will be 172.16.16.10. R1 will therefore change the destination address to 172.16.16.10. The NAT router R1 now looks at the source address of the DNS query, which is 195.65.10.10. The router will look in its routing table for a private address corresponding to the entry 195.65.10.10. If there is no such entry then it will assign an address from its private address pool, in this case, 192.168.1.10. The DNS query is now forwarded to the "enterprise.com" DNS server with a source address of 192.168.1.10 and a destination address of 172.16.16.10.

When the DNS server receives the query it will generate a response indicating the IP address of Host, A which in this case is 172.16.16.5. This response is now forwarded to R1. The router will now look up its translation table for a public address corresponding to the entry for 172.16.16.10, which in this case is 193.168.1.10 and will change the source address of the

DNS response to this public address. The router will also look in its translation table for the public address corresponding to the entry for 192.168.1.10. In this case the public address is 195.65.10.10. The destination address of the DNS response is changed to 195.65.10.10. So now the DNS response has a source address of 193.68.1.10 and a destination address of 195.65.10.10. The NAT router will also have to change the entry in the DNS response.

Remember that the entry in the DNS response will have an IP address of 172.16.16.5 for Host A. Again the NAT router, R1 will look up its table for a public address entry corresponding to the private address entry for 172.16.16.5. If there is no such entry then it will allocate a public address from its public address pool 193.168.1.0, let's say 193.168.1.5. The NAT device will now change the entry in the DNS response from 172.16.16.5 for Host A to 193.168.1.5. This response is now returned to the DNS server for "remote.com" which in turn will forward it to Host B. Host B now can communicate with Host A using the public address 193.168.1.5. Translation of addresses at router R1 will now take place as previously outlined.

As can be seen from the above example even though a host in an enterprise is using NAT, an outside host can still initiate communications with that host.

Conclusion

The BGP methods described in chapter four and RFC2260 *(Scalable support for Multi-homed Multi-provider Connectivity)* can now be combined with the above NAT methods. Unfortunately, this solution still doesn't address the issues of injecting more specific prefixes into the "default free" zone. This method does overcome some of the shortcomings associated with the methods described in RFC2260. For example, because the enterprise is connected to multiple ISPs it will have multiple address prefixes. Site renumbering will be required should the enterprise wish to change any of its ISPs. An enterprise using NAT in this way can change ISP without the need for renumbering. All that will be required is for the internal public address pool on the NAT router to be reconfigured.

From an enterprise's point of view, provider independent addressing could be used. This way the address prefix stays with the customer. This is good news for the customer but it still means that the ISP is advertising and address prefix outside of its own contiguous address space. In addition, the use of Provider Independent addresses is discouraged by IANA and the RIRs.

Chapter 7 - Financial Incentives for Route Aggregation

Previous chapters have examined how multihoming can punch holes in the routing tables of routers in the "default free zone" and the methods that can be used to reduce the need for routers having to advertise more specific routes. Multihoming, however, is not the only reason why more specific networks are advertised. Part of the problem is also to do with site renumbering.

Take, for example, a site that has been allocated a certain number of IP addresses. Should additional addresses be required by that site in the future, it is conceivable that the next contiguous block of addresses would have already been assigned to another site. In its simplest form a new block of non-contiguous addresses can be allocated to the site. The main disadvantage with this scheme is that now an additional route has to be advertised which effectively breaches the route aggregation policy. We have seen that NAT can be used to good effect in this scenario.

Another method would be to simply allocate the site a new block of contiguous addresses that would satisfy the entire site's requirements. This is good for the Internet as a whole, in that no additional routes need to be advertised. On the other hand it is bad news for the site as the entire site now has to be renumbered which will obviously involve down time, disruption and no doubt will have some financial implications as well.

A third method would be for the ISP to reserve a block of contiguous addresses that would satisfy both the immediate and future requirements of the site. This way a site would not need to renumber if additional addresses are required, nor would additional routes need to be advertised. The drawback with this scheme is that these reserved addresses may never be used and would therefore be effectively wasted. This, however, is the method used by RIPE. RIPE allocate a /19 prefix and reserve the next contiguous /19 addresses effectively creating a /18 prefix, i.e., in the binary numbering scheme the binary bit to the left has twice the significance of the bit to its right. Therefore two consecutive allocations of /19 addresses is the same a /18 allocation.

This address allocation policy is an actual over simplification of what's actually used by RIPE. RIPE in fact use the notion of an "Assignment Window" to allocate additional addresses when required. More information regarding RIPE's address allocation policy can be found in the document RIPE-324 [10].

Bellovin *et al.* (1997) put forward the notion of sites having to pay for additional route advertisements is put forward. I can certainly agree with their motives here. For example, as explained in the first method outlined above, if additional addresses are required then site renumbering is essential. This way the site puts up with the inconvenience of renumbering but the Internet on the whole is unaffected.

In their document the idea of "Bilateral Agreements" is put forward for charging for route advertisements. Consider, for example, figure 31 below.

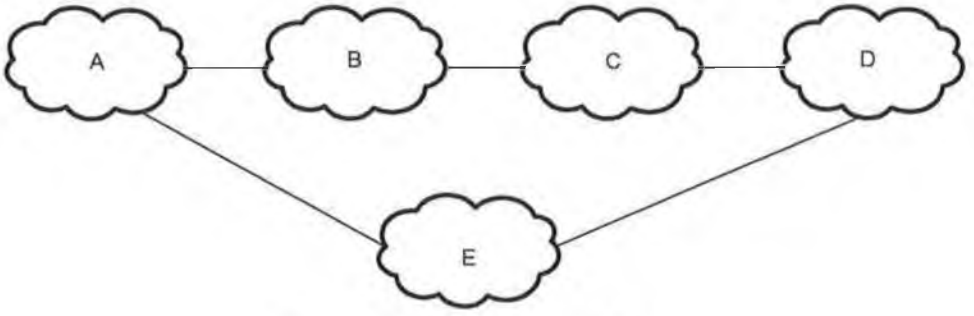

Figure 31 Dissemination of Routes

In the above diagram B advertises to C the availability of A and offers to forward transit traffic from C to A. C in turn, advertises the availability of A to D and so on. In the best case scenario all of these routes could be aggregated into just one route, if not, then separate networks need to be advertised. D might be aware of A as a result of the advertisement from C. However, D might choose to use E to get to A even though that particular route might not have been advertised. The document refers to this as a "Pull" route. In either case only the end user benefits and the cost of additional advertisements is incurred by the routers along the path. The idea, therefore, is to have an agreement between peers. A financial fee is then agreed between the parties.

Basically the bottom line in their proposal is quite simple. That is, if a site chooses to cause more routes to be advertised then there should be a financial cost involved. This would certainly be an incentive for sites to think twice about, for instance renumbering or advertising an additional route. Larger corporates on the other hand could feel that the cost implications might well be justified and would be prepared to pay the price. In terms of better route selection as in figure 31, then the decision would be based on whether a site would be happy to have basic connectivity as in the connection between site A and site D, or whether a more optimal route between sites A and D is desirable. Once again the costs involved might be worth it for some of the larger corporates.

Really what is needed is a new approach to route fragmentation. IPv6 because of its huge address space can resolve the issue of contiguous addresses being reserved for future use for each site. This couldn't possibly be a problem given the address space involved, i.e., we could still assign a unique IP address to each grain of sand on the planet and still have billions of unused addresses. But again the large amount of IPv6 addresses doesn't necessarily mean that the problem of exponential growth of route advertisements is going to go away.

The next chapter looks at what IPv6 has to offer in terms of scalable routing in the Internet, again with particular emphasis on multihoming.

Chapter 8 - Multihoming in IPv6

The methods described thus far can all be used with IPv6. Note, however, that the multihoming techniques illustrated here are more of a function or feature of BGP4 than IPv4 itself. There are still two main drawbacks with using multihoming in a non NAT environment. The first drawback is the lack of scalability, i.e., every multihomed site injects at least another route into the default free zone. The second drawback is transport layer survivability. For example, if a local host is communicating with a remote host over a multihomed link, then, should that link fail, the other multihomed link will be used in its place. This means that routing changes have to occur, which would inevitably result in indeterminate delays that in turn could cause transport layer sessions to time out.

IPv6 meets the need for expanded address space, increased security and better mobility, however, it does not, in its present form, meet the needs of better scalability in multihomed environments. This does not mean, however, that IPv6 won't be part of the solution for scalable routing and multihoming in the future. Indeed multihoming issues were never even a requirement to be addressed by IPv6. Again, in terms of addressing, the large address space offered by IPv6 could be considered as a cure for the symptom only and not really a cure for the problem itself. The addressing structure in IPv6 isn't radically different form that of Ipv4. It still holds on to the network and host semantic. Perhaps a radical departure from this structure could be helpful. This will be looked at shortly when discussing "locator/identity" in IP addressing.

Another part of the solution could be in the way addresses are allocated. Because IPv6 is still in its early deployment stage it might be possible to change the way that addresses are allocated. Because IPv6 is such an extensible protocol it might be possible to make some changes at this early stage without having a significant impact on existing or future users.

Let's take, for example, a new address allocation policy. One of the nice features of the IPv6 addressing scheme is that it is very much hierarchical and indeed very aggregatable. For example the new addressing architecture has a global routing prefix. Because of its early deployment couldn't we say, for instance, that RIPE would have a particular set of prefixes

(as they do) from which subnets could be allocated. In theory this could mean that Ireland could be allocated one or more consecutive prefixes and that all routes within Ireland can be aggregated into one route. A corporate on the other hand might wish to be multihomed with an Irish and British ISP. This still might not inject additional routes into the system since all European ISPs are allocated from the same "Global routing prefix". In theory this could have also worked with IPv4 but you need to remember that CIDR was introduced "after the horse had bolted" and an address allocation policy had already been in place based on classful addressing. With this solution we can put the new policy in place before the problem are arises. Admittedly this is an over simplification that on its own may not work but it should at least form part of the solution.

The IETF recognise that IPv6 presents new opportunities to tackle the multihoming issue in new ways not possible with IPv4. Although no new protocols or solutions have yet been produced, a new IETF Working Group (WG) has been set up. The main objective of the working group is to seek *"alternative approaches with better scaling properties. Specifically, the WG will prefer multihoming solutions that tend to minimise adverse impacts on the end-to-end routing system and limit the number of prefixes that need to be advertised in the Default-Free Zone (DFZ)"*. Documents produced by the working group can be obtained o the IETF website [11]. RFC 3582 sets out the goals for IPv6 Site-Multihoming Architectures.

One issue being looked at, at the moment is the split between identity and locator. Let's take a real world example for the moment. When conventional postal mail is addressed to a person, the address comprises two main parts, the person's name (identity) and the person's location (locator). A person's identity will never change, their name will always be the same. However, as we all know, a person's address can change. On the other hand, in the IP world, a host's identity is its locator and vice versa. A host's identity and its location are actually combined into a single protocol element, the IP address. The assumption here is that the network topology is static. As long as the topology remains static there is no issue. Today, however, this is not the case. Mobile telephony is a key example. A mobile user is free to roam without ever having to change the mobile number. In this case the identity remains fixed while the locator changes as the user moves from cell to cell. This is also the case with mobile IP hosts on wireless networks. In mobile IP, a host is allocated a static "Home Address". As a host moves around and out of its home locality a "Care of Address" (CoA) is

used. This CoA can be likened to a forwarding address used in the conventional postal system. Mobile IPv6 protocol is specified in RFC3775.

The Interesting point here is that mobile IP is not radically different to multihoming, in terms of addressing issues. For example, when a multihomed link fails and another link is selected, then the address associated with the new link must be used. The ability to dynamically change locators while maintaining a constant identifier is common to both scenarios. Huston 2005 investigates this possibility of using some features of MIPv6 for multihoming.

In a mobile scenario a fixed Home Address (HoA) is used. The HoA will act as both the identity and the locator for the host as long as the host remains attached to its home network. When a host moves from its home network to a foreign network, the host will adapt a new IP address that is part of the foreign network. This address is referred to as the Care of Address (CoA). The host must now inform an agent on its home network of the new CoA. In this situation, the HoA acts as the identity and the CoA acts as the locator. The HoA will always be used by a remote host to communicate with a mobile host. The agent on the home network acts as a forwarding agent for the mobile host so that all packets addressed to the HoA will be forwarded to the CoA when the mobile host is not attached to its home network.

This same principle could also be used in a multihomed environment. In this situation one of the multihomed addresses is used as the HoA. This address will continue to be used for as long as the associated multihomed link remains up. Should this link fail then one of the other multihomed addresses can be used as a CoA.

Identity/Locator Split.

One of the issues with multihoming is the inability of a remote node to realise that it is communicating with a multihomed host. Although a multihomed host will have more than one IP address, only one address can by used in any one communication session. All communications between a local multihomed host and a remote host must pass through the ISP associated with the chosen address prefix. If prefix A is selected, then all traffic must be routed through ISP A. However, in the event that the link between the local multihomed host and ISP A fails, reachability to the network will still be advertised under ISP A's larger prefix advertisement. This means that the remote host will not be aware of the link failure and will

still forward packets to ISP A. These packets, however, will be dropped upon reaching ISP A.

Another concern is the address that a multihomed host uses to establish a connection with a remote host? If prefix A is selected and the local router forwards the packet to ISP B, then it is most likely that these packets will be dropped by ISP B's ingress filters. Ingress filters or "Reverse Path Forwarding Filters" are used by ISPs as a means of mitigating IP address spoofing. In the example above ISP B's router will see a packet with a prefix of A arriving at its interface. If this is a genuine packet then the router will have an entry in its table linking this address with that particular interface. If there is no such association then the packet will be dropped because it suggests that the IP address is being spoofed. Miller [12] provides a brief tutorial on ingress filtering.

Again the problem is that there is no separation between identity and location. The ideal solution is to separate identity and locator. In the ideal world this would mean that a multihomed host will always be addressed by a fixed and constant identity but that the locator can be dynamically changed as multihomed links go up and down. This solution would indeed be very scalable since no additional routes would be advertised. Houston (2001) explores such possibilities.

In his document, Houston puts forward the idea of introducing a new protocol element into the TCP/IP stack. This new element would present a fixed end point identity to the upper layer protocols. The presentation to the lower stack elements would be in the form of a locator, which can change as the multihomed links go up and down. Houston also explores the possibility of modifying an existing protocol element and even perhaps modifying DNS services to allow remote sites determine which specific address to use.

Chapter 9 - Comparison of Multihoming Techniques

Auto Route Injection

Chapter 4 presented a number of methods that can be used to support scalable multihoming. Auto route injection as proposed in RFC 2260 was briefly outlined. One problem with this approach is to do with prefix filtering.

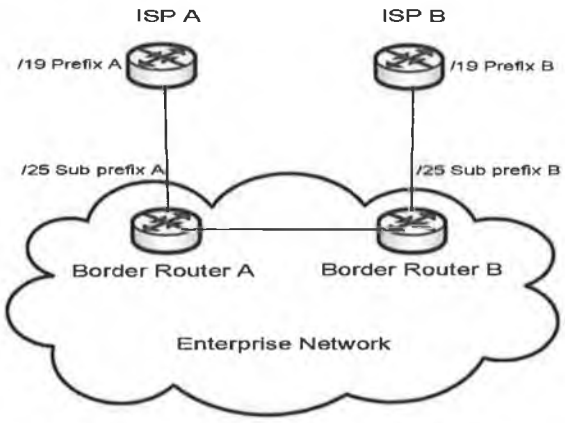

Figure 32 Auto Route Injection

Assume as in figure 32, that ISPs A and B have been assigned a /19 prefix from appropriate RIRs. Both ISPs could in turn assign a /25 prefix from their larger address prefixes to the enterprise network. Border router A will advertise the '/25 Sub prefix A' to ISP A. Likewise border router B will advertise '/25 Sub prefix B' to ISP B. Neither ISP will advertise the /25 prefixes since these networks will be advertised under the larger /19 prefix. As outlined in chapter 4, however, border router A will advertise '/25 Sub prefix B' in the event that the link to ISP B fails. Not only does this introduce an additional advertisement into the default free zone but it is also likely that routers will filter this larger prefix as per the guidelines outlined in chapter 4. This means that any remote host that was communicating with the enterprise using the '/25 Sub prefix B' would loose that connection since this route would have been filtered out. Using this configuration, therefore, would suggest that multihoming can't offer any benefits in terms of resilience but would still offer load balancing and traffic engineering.

This, however, is not strictly the case. Referring back to figure 26, it is quite evident that stringent filtering rules are not being applied. It is therefore most likely that the '/25 Sub prefix B' would not be filtered and that this network would remain reachable.
On the other hand, in the document by Bellovin *et al.* 2001, it was calculated that only about 0.3% of routes would be deemed as unreachable if recommended filtering rules were applied. This, however, needs to be taken in context. The calculations in that document can not really be applied here because specific prefixes advertised as a result of multihoming are different to specific prefixes advertised as a result of corporate traffic policies. In order to explain this we will assume, for example, that there is a filter rule that will block any prefix larger than /18 in address range 139.191.0.0. This means, therefore, that an advertisement for 139.191.152.0/21 will be blocked. Such a filer will not present a problem, however, since reachability would still be advertised through 139.191.0.0/18.

The outcome is quite different when this addressing scheme is applied in a multihomed environment as shown in figure 33.

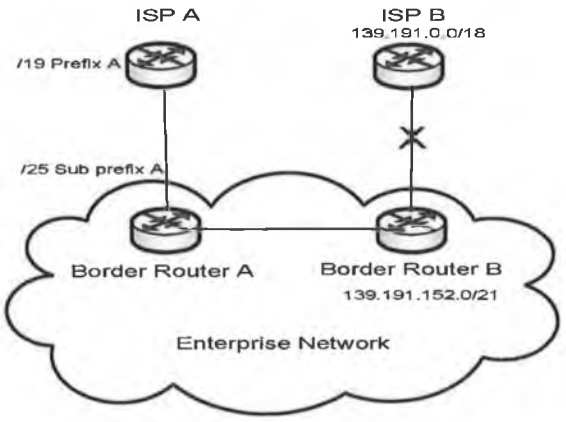

Figure 33 BGP Link Failure

ISP A will advertise reachability to 139.191.152.0/21 when the link between ISP B and border router B fails. Because this is a more specific route than 139.191.0.0/18, remote routers will select ISP A as the optimum route. Filter rules as described above, however, will prevent this route from being advertised on the default free zone. Therefore, all packets destined for 139.191.152.0/21 will follow the route advertised by ISP B under the larger

prefix advertisement of 139.191.0.0. Network 139.191.152.0/21 will remain unreachable, therefore, since the link between ISP B and border router B has failed.

Non Direct Peering

Auto route injection can therefore be somewhat unreliable and is largely dependant upon the absence of filtering rules. It is worth noting, however, that stringent filtering rules are not being applied as is evident from the portion of the BGP routing table shown in figure 26.

The non direct peering option specified in the same RFC (RFC 2260) is a marked improvement on Auto route injection. The most significant benefit is that it doesn't cause any more additional routes to be advertised into the default free zone. There are, however, three significant disadvantages that are worth mentioning. Firstly, as with auto route injection, significant additional loading is placed on the other router. As in the previous scenario, when the link to ISP B fails, border router A will be used to route traffic into and out of the enterprise in such a way that it will assume part of the workload of border router B. However, since border router A is not advertising a new prefix to ISP the filtering issues previously outlined are of no consequence.

The second disadvantage with this configuration is that ISP B will continue to advertise prefix B under its own larger address prefix. Therefore all traffic destined for border router B is still directed to ISP B, which is now a non optimal route since ISP B must now forward the traffic to border router A. The optimal route, in this case, for prefix B traffic is actually through ISP A. However, since border router A does not advertise prefix B to ISP A, all prefix B traffic will go through ISP B.

The third disadvantage with non direct peering is that resilience is only maintained in the event of a link failure between the border router and its ISP. Should ISP B, for example, suffer from a catastrophic failure, then all prefix B hosts would be completely unreachable since ISP B will no longer be a position to advertise prefix B routes. This is not the case with auto route injection as ISP A will advertise prefix B routes. On the other hand, the probability of an ISP experiencing a catastrophic failure is extremely small.

Network Address Translation

The NAT solution put forward by Cisco and summarised in chapter 6 has a little bit more to offer. The obvious advantages are the conservation of address space and the fact that an enterprise is free to change ISPs without ever having to renumber. The only address that would need to be changed is the outside address on the border router.

Another important feature of multihoming with NAT is that it can be used with both auto route injection and non direct peering. The significant difference, however, is that the entire enterprise network can be allocated addresses from one internal private IP address block. In the previous method the enterprise numbering was based on prefixes assigned by each ISP. This means that all hosts would either have multiple IP addresses, i.e., an address assignment from each ISP that the enterprise is multihomed to, or each part of the enterprise has a different IP numbering scheme based on the host's proximity to a border router. For example hosts in figure 33 that are closest to border router A would have a prefix A assignment and hosts closer to border router B would have a prefix B assignment. NAT alleviates this complexity by allowing a single addressing scheme to be used.

The outside public address of each border router still has to be assigned by the ISP that the router connects to. This means that the disadvantages associated with non direct peering still apply. However, using NAT with non direct peering is more immune to catastrophic ISP failure. Again, assuming that, for instance, ISP B experiences a complete failure then prefix B would be unreachable as previously explained. With NAT, however, an alternative default gateway can automatically be selected by the hosts on the enterprise. This would now route all packets that previously would have been using ISP B to ISPA. This is made possible by the fact that all hosts in the enterprise belong to the same network so that border routers A and B are also on the same IP network. Without NAT, the border routers will be on different IP networks.

NAT, however, does have its limitations and these are outlined in RFC 3027. As expected there are solutions for most of the complications that arise with NAT. This is evident from the fact that large organisations including Local Government make widespread use of NAT.

Every County Council in Ireland, for example, is allocated IP address blocks from the 10.0.0.0/8 address space.

With the prolific use of VOIP applications concerns could arise regarding the ability of outside hosts initiating VOIP calls to hosts that are behind a NAT device. Common applications that use NAT require the inside host to establish a connection to an external host. For example a web client behind a NAT device establishes the connection to an outside web server. The translation address assignment is made on the outgoing data so that the destination address on return packets can be translated back into private addresses as they enter the router.

The issue becomes a little bit more complicated when a remote host wishes to establish a connection with a local host that is behind a NAT device. Since private addresses are not revealed by the NAT host remote devices have no means for establishing the local host's IP address. Cisco has addressed this issue in an article entitled '*Enabling Enterprise Multihoming with Cisco IOS NAT'* which has already been outlined in chapter 6.

Many articles have been written about VOIP and NAT traversal. Suffice it say that there are plenty of solutions at hand. These solutions are also being driven to a large extent by the fact that more and more domestic users are now using DSL broadband. One of the major advantages of using DSL broadband is the ability to use VOIP, but most DSL modems incorporate NAT. Juniper Networks in a white paper entitled '*Hosted NAT Traversal Unlocks VOIP Service Provider Offerings'* describe how remote hosts can initiate VOIP call to hosts using private IP addresses behind a NAT device. This solution makes use of a 'Session Border Controller' which is placed behind the edge router at the VOIP service provider's site. The semantics of any solution, however, are not important here. The main issue is that there is no longer any concern about establishing a VOIP call to a host behind NAT device or firewall.

Financial Incentives for Route Aggregation

Of all the options presented this is the least favourable. This is merely a cure for the symptom and not a cure for the cause. At first glance the idea of imposing a charge or penalty does seem to be attractive. Making this option work, however, is another thing. In their document,

Rekhter *et al.* suggest "Bilateral agreements" between ISP's where a fee is agreed upon. This fee would then presumably be passed on to the end user. The main issue with this is how much should that fee be?, and would there be any level of global consistency? Closer to home, for example, there could be significant disparity between ISPs in the Republic and in the North of Ireland. This could well encourage users to make multihomed connections with two different ISPs, one in the north and the other in the south.

The other concern with this proposition is that if the financial fee is set too low then it could well encourage users to multihome. In other parts of the world while the fee might be set too high, multinationals could simply feel the return on investment is worth it and that resilience is necessary at any cost.

The notion of such fees also raises the age old argument of information rich and information poor societies. At the end of the day there is nothing wrong with multihoming and its use should not be discouraged just because technical solutions for its associated scaling issues are difficult to address. Just as we are not going to solve the green house effect by increasing fuel prices, we are not going to solve route scaling issues by imposing additional charges. We need to look at the root of the problem and find a cure for that cause.

Multihoming in IPv6

Unfortunately IPv6 has no more multihoming capabilities than IPv4 and although the BGP solutions described in this document will also work with IPv6, the exponential growth in BGP tables will still continue. The problem could even be exacerbated as a result of the almost endless amount of IPv6 addresses available. Prevention is always better than cure and IPv6 presents us with the opportunity to address multihoming capabilities in advance of its widespread deployment throughout the Internet.

Potential areas where IPv6 could be modified to support multihoming are addressed by Houston in RFC 4177 and some of these were outlined in chapter 8. Houston likens multihoming to mobile IP. For example, when a mobile host moves out of its home network a new 'care of address' is used. The same is also true when a multihomed site looses one of its connections. When this happens a new prefix needs to be advertised. It would make sense

therefore that Mobile IP should be modified or adopted for multihoming. Houston, however, acknowledges that there are issues with this. For example, when a mobile node moves away from its home location it will register its new 'care of address' (CoA) with an agent on its home network. The home agent will pass traffic addressed to the node's 'Home Address' (HoA) to the new CoA thus maintaining an uninterrupted link. The mobile node can also optionally pass its new CoA directly to the remote node that it was communicating with. The remote node needs to be able to validate that this information is indeed coming from the original mobile node and not some 'spoofed' node. To this end, the remote node will pass two different secrets to both the HoA and the new CoA. If the mobile node receives both secrets then this verifies to the remote node that the mobile host is authentic. In order to prevent playback attacks and man-in-the- middle attacks, such an association will time out after seven minutes. In a multihomed environment this binding needs to be indefinite or at least last as long as it takes for the failed link to be restored.

Houston explores other alternatives where IPv6 could be modified to accommodate multihoming. All suggestions however raise security concerns and are well documented in his paper.

Conclusion

This paper examined issues relating to route scaling within BGP routing tables in the Internet's 'default free zone', or backbone. The exponential growth in route entries is contributed to by both IP address fragmentation and 'specific route advertisements' as a result of multihoming. The later being the most significant contributor to BGP table growth.

The IPv6 protocol was examined in detail with particular emphasis being placed on its addressing architecture. Unfortunately multihoming support was never a design specification for this new protocol. This does not mean, however, that IPv6 won't be part of the solution in curbing the exponential growth of routing tables. The fact that IPv6 hasn't yet seen widespread deployment throughout the Internet means that modifications are possible without significant disruption.

Current solutions and practises were examined and their strengths and weaknesses highlighted. As is evident from the graph shown in figure 21, either these solutions are not being employed, or are simply not working, as we continue to observe an alarming growth rate in the number of routes being advertised.

The problem of route scaling today is as significant as the shortage of IP address space was in the eighties. No doubt a solution will be found and this will be in the form of a modification to the IPv6 protocol stack.

Bibliography

Bellovin, S., Rekhter, M., Resnick, P., (1997). Financial Incentives for Route Aggregation and Efficient Address Utilization in the Internet. Coordination the Internet. MIT Press. 1997.

Borthick, S., 2001. Today's Internet Can't Scale, Business Communications Review, May 2001, 28-33.

Cisco, 1997. Enabling Enterprise Multihoming with Cisco IOS Network Address Translation (NAT)

Hagan, S., 2002. IPv6 Essentials. O' Reilly & Associates.

Huston, G., 2001. Analyzing the Internet BGP Routing Table. The Internet Protocol Journal, 4 (1), 2-15.

Huston, G., 2003. The Mythology of IP Version 6. The Internet Protocol Journal, 6 (2), 23-29.

RFC 1287, Clarke, D., Chapin, L., Cerf, V., Braden, R., Hobby, R., "Towards the Future Internet Architecture", December 1991.

RFC 1338, Fuller, V., Li, T., Yu, J., Varadhan, K., "Supernetting: an Address Assignment and Aggregation Strategy", June 1992.

RFC 1518, Rekhter, Y., Li, T., "An Architecture for IP Address Allocation with CIDR", September 1993.

RFC 1519, Fuller, V., Li, T., Yu, J., Varadhan, K., "Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy", September 1993.

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

RFC 2260, Bates, T., Rekhter, Y., "Scalable Support for Multi-homed Multi-provider Connectivity", January 1998.

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

RFC 2474, Nichols, K., Blake, S., Baker, F., Black, D., "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", December 1998.

RFC 3177, IAB, IESG, "IAB/IESG Recommendations on IPv6 Address Allocation to Sites", September 2001.

RFC 3513, Deering, S., Hinden, R., "Internet Protocol Version 6 (IPv6) Addressing Architecture", April 2003.

RFC 3582, Abley, J., Black, B., Gill, V. "Goals for IPv6 Site Multihoming Architecture", August 2003

RFC 3587, Deering, S., Hinden, R., Nordmark, E., "IPv6 Global Unicast Address Format", August 2003.

RFC 3775, Johnson, D., Perkins, C., Arkoo, J. "Mobility Support in IPv6", June 2004

RFC 4177, Houston, G., "Architectural Approaches to Multi-homing for IPv6", September 2005

[1] LI, T., 2001. Hardware Implications of Internet Routing Table Growth [online]. NANOG. Available from: http://www.nanog.org/mtg-0102/ppt/li/

[2] POTAROO. BGP Reports [online].
Available from: http://bgp.potaroo.net/index-bgp.html

[3] RIPE NCC. Local Internet Registries Offering Service in Ireland [online].
Available from: http://www.ripe.net/membership/indices/IE.html

[4] RIPE NCC. Local Internet Registries Training Course [online]
Available from: http://www.ripe.net/training/lir/

[5] CISCO. Border Gateway Protocol, Internetworking Technologies Handbook [online]
Available from: http://www.cisco.com/univercd/cc/td/doc/cisintwk/ito_doc/bgp.pdf

[6] POTAROO. AS1221 BGP Table Data [online]
Available from: http://bgp.potaroo.net/as1221/bgp-active.html

[7] RIPE NCC. RIPE-326, RIPE Document Store [online]
Available from: http://www.ripe.net/docs/alltitle.html

[8] BELLIVON. AT&L Labs Research [online]
Available from: http://www.research.att.com/~jrex/papers/filter.pdf

[9] RWhois. RWhois web interface [online]
Available from: http://www.rwhois.net/rwhois/prwhois.html

[10] RIPE NCC. RIPE-324, RIPE Document Store [online]
Available from: http://www.ripe.net/docs/alltitle.html

[11] IETF. Site Multihoming in IPv6 [online]
Available from: www.ietf.org/html.charters/multi6-charter.html

[12] Miller. Carnigie Mellon, Three Practical Ways to Improve Your Network
[online]. Available at: http://www.net.cmu.edu/pres/lisa03/tpi.ppt